Selective Dyna-Style Planning Under Limited Model Capacity

- Zaheer Abbas, Samuel Sokota, Erin Talvitie, Martha White
- abstract: In model-based reinforcement learning, planning with an imperfect model of the environment has the potential to harm learning progress. But even when a model is imperfect, it may still contain information that is useful for planning. In this paper, we investigate the idea of using an imperfect model selectively. The agent should plan in parts of the state space where the model would be helpful but refrain from using the model where it would be harmful. An effective selective planning mechanism requires estimating predictive uncertainty, which arises out of aleatoric uncertainty, parameter uncertainty, and model inadequacy, among other sources. Prior work has focused on parameter uncertainty for selective planning. In this work, we emphasize the importance of model inadequacy. We show that heteroscedastic regression can signal predictive uncertainty arising from model inadequacy that is complementary to that which is detected by methods designed for parameter uncertainty, indicating that considering both parameter uncertainty and model inadequacy may be a more promising direction for effective selective planning than either in isolation.

A distributional view on multi-objective policy optimization

- Abbas Abdolmaleki, Sandy Huang, Leonard Hasenclever, Michael Neunert, Francis Song, Martina Zambelli, Murilo Martins, Nicolas Heess, Raia Hadsell, Martin Riedmiller
- abstract: Many real-world problems require trading off multiple competing objectives. However, these objectives are often in different units and/or scales, which can make it challenging for practitioners to express numerical preferences over objectives in their native units. In this paper we propose a novel algorithm for multi-objective reinforcement learning that enables setting desired preferences for objectives in a scale-invariant way. We propose to learn an action distribution for each objective, and we use supervised learning to fit a parametric policy to a combination of these distributions. We demonstrate the effectiveness of our approach on challenging high-dimensional real and simulated robotics tasks, and show that setting different preferences in our framework allows us to trace out the space of nondominated solutions.

Efficient Optimistic Exploration in Linear-Quadratic Regulators via Lagrangian Relaxation

- Marc Abeille, Alessandro Lazaric
- abstract: We study the exploration-exploitation dilemma in the linear quadratic regulator (LQR) setting. Inspired by the extended value iteration algorithm used in optimistic algorithms for finite MDPs, we propose to relax the optimistic optimization of \ofulq and cast it into a constrained \emph{extended} LQR problem, where an additional control variable implicitly selects the system dynamics within a confidence interval. We then move to the corresponding Lagrangian formulation for which we prove strong duality. As a result, we show that an ϵ-optimistic controller can be computed efficiently by solving at most $O\big(\log(1/\epsilon)\big)$ Riccati equations. Finally, we prove that relaxing the original \ofu problem does not impact the learning performance, thus recovering the $\wt O(\sqrt{T})$ regret of \ofulq. To the best of our knowledge, this is the first computationally efficient confidence-based algorithm for LQR with worst-case optimal regret guarantees.

Super-efficiency of automatic differentiation for functions defined as a minimum

- Pierre Ablin, Gabriel Peyré, Thomas Moreau
- abstract: In min-min optimization or max-min optimization, one has to compute the gradient of a function defined as a minimum. In most cases, the minimum has no closed-form, and an approximation is obtained via an iterative algorithm. There are two usual ways of estimating the gradient of the function: using either an analytic formula obtained by assuming exactness of the approximation, or automatic differentiation through the algorithm. In this paper, we study the asymptotic error made by these estimators as a function of the optimization error. We find that the error of the automatic estimator is close to the square of the error of the analytic estimator, reflecting a super-efficiency phenomenon. The convergence of the automatic estimator greatly depends on the convergence of the Jacobian of the algorithm. We analyze it for gradient descent and stochastic gradient descent and derive convergence rates for the estimators in these cases. Our analysis is backed by numerical experiments on toy problems and on Wasserstein barycenter computation. Finally, we discuss the computational complexity of these estimators and give practical guidelines to chose between them.

A Geometric Approach to Archetypal Analysis via Sparse Projections

- Vinayak Abrol, Pulkit Sharma
- abstract: Archetypal analysis (AA) aims to extract patterns using self-expressive decomposition of data as convex combinations of extremal points (on the convex hull) of the data. This work presents a computationally efficient greedy AA (GAA) algorithm. GAA leverages the underlying geometry of AA, is scalable to larger datasets, and has significantly faster convergence rate. To achieve this, archetypes are learned via sparse projection of data. In the transformed space, GAA employs an iterative subset selection approach to identify archetypes based on the sparsity of convex representations. The work further presents the use of GAA algorithm for extended AA models such as robust and kernel AA. Experimental results show that GAA is considerably faster while performing comparable to existing methods for tasks such as classification, data visualization/categorization.

Context Aware Local Differential Privacy

- Jayadev Acharya, Kallista Bonawitz, Peter Kairouz, Daniel Ramage, Ziteng Sun
- abstract: Local differential privacy (LDP) is a strong notion of privacy that often leads to a significant drop in utility. The original definition of LDP assumes that all the elements in the data domain are equally sensitive. However, in many real-life applications, some elements are more sensitive than others. We propose a context-aware framework for LDP that allows the privacy level to vary across the data domain, enabling system designers to place privacy constraints where they matter without paying the cost where they do not. For binary data domains, we provide a universally optimal privatization scheme and highlight its connections to Warner's randomized response and Mangat's improved response. Motivated by geo-location and web search applications, for k-ary data domains, we consider two special cases of context-aware LDP: block-structured LDP and high-low LDP. We study minimax discrete distribution estimation under both cases and provide communication-efficient, sample-optimal schemes, and information-theoretic lower bounds. We show, using worst-case analyses and experiments on Gowalla's 3.6 million check-ins to 43,750 locations, that context-aware LDP achieves a far better accuracy under the same number of samples.

Efficient Intervention Design for Causal Discovery with Latents

- Raghavendra Addanki, Shiva Kasiviswanathan, Andrew Mcgregor, Cameron Musco
- abstract: We consider recovering a causal graph in presence of latent variables, where we seek to minimize the cost of interventions used in the recovery process. We consider two intervention cost models: (1) a linear cost model where the cost of an intervention on a subset of variables has a linear form, and (2) an identity cost model where the cost of an intervention is the same, regardless of what variables it is on, i.e., the goal is just to minimize the number of interventions. Under the linear cost model, we give an algorithm to identify the ancestral relations of the underlying causal graph, achieving within a 2-factor of the optimal intervention cost. This approximation factor can be improved to $1+\eps$ for any $\eps > 0$ under some mild restrictions. Under the identity cost model, we bound the number of interventions needed to recover the entire causal graph, including the latent variables, using a parameterization of the causal graph through a special type of colliders. In particular, we introduce the notion of p-colliders, that are colliders between pair of nodes arising from a specific type of conditioning in the causal graph, and provide an upper bound on the number of interventions as a function of the maximum number of p-colliders between any two nodes in the causal graph.

The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization

- Ben Adlam, Jeffrey Pennington
- abstract: Modern deep learning models employ considerably more parameters than required to fit the training data. Whereas conventional statistical wisdom suggests such models should drastically overfit, in practice these models generalize remarkably well. An emerging paradigm for describing this unexpected behavior is in terms of a \emph{double descent} curve, in which increasing a model's capacity causes its test error to first decrease, then increase to a maximum near the interpolation threshold, and then decrease again in the overparameterized regime. Recent efforts to explain this phenomenon theoretically have focused on simple settings, such as linear regression or kernel regression with unstructured random features, which we argue are too coarse to reveal important nuances of actual neural networks. We provide a precise high-dimensional asymptotic analysis of generalization under kernel regression with the Neural Tangent Kernel, which characterizes the behavior of wide neural networks optimized with gradient descent. Our results reveal that the test error has nonmonotonic behavior deep in the overparameterized regime and can even exhibit additional peaks and descents when the number of parameters scales quadratically with the dataset size.

Rank Aggregation from Pairwise Comparisons in the Presence of Adversarial Corruptions

- Arpit Agarwal, Shivani Agarwal, Sanjeev Khanna, Prathamesh Patil
- abstract: Rank aggregation from pairwise preferences has widespread applications in recommendation systems and information retrieval. Given the enormous economic and societal impact of these applications, and the consequent incentives for malicious players to manipulate ranking outcomes in their favor, an important challenge is to make rank aggregation algorithms robust to adversarial manipulations in data. In this paper, we initiate the study of robustness in rank aggregation under the popular Bradley-Terry-Luce (BTL) model for pairwise comparisons. We consider a setting where pairwise comparisons are initially generated according to a BTL model, but a fraction of these comparisons are corrupted by an adversary prior to being reported to us. We consider a strong contamination model, where an adversary having complete knowledge of the initial truthful data and the underlying true BTL parameters, can subsequently corrupt the truthful data by inserting, deleting, or changing data points. The goal is to estimate the true score/weight of each item under the BTL model, even in the presence of these corruptions. We characterize the extent of adversarial corruption under which the true BTL parameters are uniquely identifiable. We also provide a novel pruning algorithm that provably cleans the data of adversarial corruption under reasonable conditions on data generation and corruption. We corroborate our theory with experiments on both synthetic as well as real data showing that previous algorithms are vulnerable to even small amounts of corruption, whereas our algorithm can clean a reasonably high amount of corruption.

Boosting for Control of Dynamical Systems

- Naman Agarwal, Nataly Brukhim, Elad Hazan, Zhou Lu
- abstract: We study the question of how to aggregate controllers for dynamical systems in order to improve their performance. To this end, we propose a framework of boosting for online control. Our main result is an efficient boosting algorithm that combines weak controllers into a provably more accurate one. Empirical evaluation on a host of control settings supports our theoretical findings.

An Optimistic Perspective on Offline Reinforcement Learning

- Rishabh Agarwal, Dale Schuurmans, Mohammad Norouzi
- abstract: Off-policy reinforcement learning (RL) using a fixed offline dataset of logged interactions is an important consideration in real world applications. This paper studies offline RL using the DQN replay dataset comprising the entire replay experience of a DQN agent on 60 Atari 2600 games. We demonstrate that recent off-policy deep RL algorithms, even when trained solely on this fixed dataset, outperform the fully trained DQN agent. To enhance generalization in the offline setting, we present Random Ensemble Mixture (REM), a robust Q-learning algorithm that enforces optimal Bellman consistency on random convex combinations of multiple Q-value estimates. Offline REM trained on the DQN replay dataset surpasses strong RL baselines. Ablation studies highlight the role of offline dataset size and diversity as well as the algorithm choice in our positive results. Overall, the results here present an optimistic view that robust RL algorithms trained on sufficiently large and diverse offline datasets can lead to high quality policies. The DQN replay dataset can serve as an offline RL benchmark and is open-sourced.

Optimal Bounds between f-Divergences and Integral Probability Metrics

- Rohit Agrawal, Thibaut Horel
- abstract: The families of f-divergences (e.g. the Kullback-Leibler divergence) and Integral Probability Metrics (e.g. total variation distance or maximum mean discrepancies) are commonly used in optimization and estimation. In this work, we systematically study the relationship between these two families from the perspective of convex duality. Starting from a tight variational representation of the f-divergence, we derive a generalization of the moment generating function, which we show exactly characterizes the best lower bound of the f-divergence as a function of a given IPM. Using this characterization, we obtain new bounds on IPMs defined by classes of unbounded functions, while also recovering in a unified manner well-known results for bounded and subgaussian functions (e.g. Pinsker's inequality and Hoeffding's lemma).

LazyIter: A Fast Algorithm for Counting Markov Equivalent DAGs and Designing Experiments

- Ali Ahmaditeshnizi, Saber Salehkaleybar, Negar Kiyavash
- abstract: The causal relationships among a set of random variables are commonly represented by a Directed Acyclic Graph (DAG), where there is a directed edge from variable X to variable Y if X is a direct cause of Y. From the purely observational data, the true causal graph can be identified up to a Markov Equivalence Class (MEC), which is a set of DAGs with the same conditional independencies between the variables. The size of an MEC is a measure of complexity for recovering the true causal graph by performing interventions. We propose a method for efficient iteration over possible MECs given intervention results. We utilize the proposed method for computing MEC sizes and experiment design in active and passive learning settings. Compared to previous work for computing the size of MEC, our proposed algorithm reduces the time complexity by a factor of $O(n)$ for sparse graphs where n is the number of variables in the system. Additionally, integrating our approach with dynamic programming, we design an optimal algorithm for passive experiment design. Experimental results show that our proposed algorithms for both computing the size of MEC and experiment design outperform the state of the art.

Learning What to Defer for Maximum Independent Sets

- Sungsoo Ahn, Younggyo Seo, Jinwoo Shin
- abstract: Designing efficient algorithms for combinatorial optimization appears ubiquitously in various scientific fields. Recently, deep reinforcement learning (DRL) frameworks have gained considerable attention as a new approach: they can automate the design of a solver while relying less on sophisticated domain knowledge of the target problem. However, the existing DRL solvers determine the solution using a number of stages proportional to the number of elements in the solution, which severely limits their applicability to large-scale graphs. In this paper, we seek to resolve this issue by proposing a novel DRL scheme, coined learning what to defer (LwD), where the agent adaptively shrinks or stretch the number of stages by learning to distribute the element-wise decisions of the solution at each stage. We apply the proposed framework to the maximum independent set (MIS) problem, and demonstrate its significant improvement over the current state-of-the-art DRL scheme. We also show that LwD can outperform the conventional MIS solvers on large-scale graphs having millions of vertices, under a limited time budget.

Invariant Risk Minimization Games

- Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, Amit Dhurandhar
- abstract: The standard risk minimization paradigm of machine learning is brittle when operating in environments whose test distributions are different from the training distribution due to spurious correlations. Training on data from many environments and finding invariant predictors reduces the effect of spurious features by concentrating models on features that have a causal relationship with the outcome. In this work, we pose such invariant risk minimization as finding the Nash equilibrium of an ensemble game among several environments. By doing so, we develop a simple training algorithm that uses best response dynamics and, in our experiments, yields similar or better empirical accuracy with much lower variance than the challenging bi-level optimization problem of Arjovsky et al. (2019). One key theoretical contribution is showing that the set of Nash equilibria for the proposed game are equivalent to the set of invariant predictors for any finite number of environments, even with nonlinear classifiers and transformations. As a result, our method also retains the generalization guarantees to a large set of environments shown in Arjovsky et al. (2019). The proposed algorithm adds to the collection of successful game-theoretic machine learning algorithms such as generative adversarial networks.

Why bigger is not always better: on finite and infinite neural networks

- Laurence Aitchison
- abstract: Recent work has argued that neural networks can be understood theoretically by taking the number of channels to infinity, at which point the outputs become Gaussian process (GP) distributed. However, we note that infinite Bayesian neural networks lack a key facet of the behaviour of real neural networks: the fixed kernel, determined only by network hyperparameters, implies that they cannot do any form of representation learning. The lack of representation or equivalently kernel learning leads to less flexibility and hence worse performance, giving a potential explanation for the inferior performance of infinite networks observed in the literature (e.g. Novak et al. 2019). We give analytic results characterising the prior over representations and representation learning in finite deep linear networks. We show empirically that the representations in SOTA architectures such as ResNets trained with SGD are much closer to those suggested by our deep linear results than by the corresponding infinite network. This motivates the introduction of a new class of network: infinite networks with bottlenecks, which inherit the theoretical tractability of infinite networks while at the same time allowing representation learning.

Discriminative Jackknife: Quantifying Uncertainty in Deep Learning via Higher-Order Influence Functions

- Ahmed Alaa, Mihaela Van Der Schaar
- abstract: Deep learning models achieve high predictive accuracy across a broad spectrum of tasks, but rigorously quantifying their predictive uncertainty remains challenging. Usable estimates of predictive uncertainty should (1) cover the true prediction targets with high probability, and (2) discriminate between high- and low confidence prediction instances. Existing methods for uncertainty quantification are based predominantly on Bayesian neural networks; these may fall short of (1) and (2) {—} i.e., Bayesian credible intervals do not guarantee frequentist coverage, and approximate posterior inference undermines discriminative accuracy. In this paper, we develop the discriminative jackknife (DJ), a frequentist procedure that utilizes influence functions of a model's loss functional to construct a jackknife (or leave one-out) estimator of predictive confidence intervals. The DJ satisfies (1) and (2), is applicable to a wide range of deep learning models, is easy to implement, and can be applied in a post-hoc fashion without interfering with model training or compromising its accuracy. Experiments demonstrate that DJ performs competitively compared to existing Bayesian and non-Bayesian regression baselines.

Frequentist Uncertainty in Recurrent Neural Networks via Blockwise Influence Functions

- Ahmed Alaa, Mihaela Van Der Schaar
- abstract: Recurrent neural networks (RNNs) are instrumental in modelling sequential and time-series data. Yet, when using RNNs to inform decision-making, predictions by themselves are not sufficient {—} we also need estimates of predictive uncertainty. Existing approaches for uncertainty quantification in RNNs are based predominantly on Bayesian methods; these are computationally prohibitive, and require major alterations to the RNN architecture and training. Capitalizing on ideas from classical jackknife resampling, we develop a frequentist alternative that: (a) does not interfere with model training or compromise its accuracy, (b) applies to any RNN architecture, and (c) provides theoretical coverage guarantees on the estimated uncertainty intervals. Our method derives predictive uncertainty from the variability of the (jackknife) sampling distribution of the RNN outputs, which is estimated by repeatedly deleting "blocks" of (temporally-correlated) training data, and collecting the predictions of the RNN re-trained on the remaining data. To avoid exhaustive re-training, we utilize influence functions to estimate the effect of removing training data blocks on the learned RNN parameters. Using data from a critical care setting, we demonstrate the utility of uncertainty quantification in sequential decision-making.

Random extrapolation for primal-dual coordinate descent

- Ahmet Alacaoglu, Olivier Fercoq, Volkan Cevher
- abstract: We introduce a randomly extrapolated primal-dual coordinate descent method that adapts to sparsity of the data matrix and the favorable structures of the objective function. Our method updates only a subset of primal and dual variables with sparse data, and it uses large step sizes with dense data, retaining the benefits of the specific methods designed for each case. In addition to adapting to sparsity, our method attains fast convergence guarantees in favorable cases \emph{without any modifications}. In particular, we prove linear convergence under metric subregularity, which applies to strongly convex-strongly concave problems and piecewise linear quadratic functions. We show almost sure convergence of the sequence and optimal sublinear convergence rates for the primal-dual gap and objective values, in the general convex-concave case. Numerical evidence demonstrates the state-of-the-art empirical performance of our method in sparse and dense settings, matching and improving the existing methods.

A new regret analysis for Adam-type algorithms

- Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, Volkan Cevher
- abstract: In this paper, we focus on a theory-practice gap for Adam and its variants (AMSGrad, AdamNC, etc.). In practice, these algorithms are used with a constant first-order moment parameter β_{1} (typically between 0.9 and 0.99). In theory, regret guarantees for online convex optimization require a rapidly decaying $\beta_{1}\to0$ schedule. We show that this is an artifact of the standard analysis, and we propose a novel framework that allows us to derive optimal, data-dependent regret bounds with a constant β_{1}, without further assumptions. We also demonstrate the flexibility of our analysis on a wide range of different algorithms and settings.

Restarted Bayesian Online Change-point Detector achieves Optimal Detection Delay

- Reda Alami, Odalric Maillard, Raphael Feraud
- abstract: we consider the problem of sequential change-point detection where both the change-points and the distributions before and after the change are assumed to be unknown. For this problem of primary importance in statistical and sequential learning theory, we derive a variant of the Bayesian Online Change Point Detector proposed by \cite{fearnhead2007line} which is easier to analyze than the original version while keeping its powerful message-passing algorithm. We provide a non-asymptotic analysis of the false-alarm rate and the detection delay that matches the existing lower-bound. We further provide the first explicit high-probability control of the detection delay for such approach. Experiments on synthetic and real-world data show that this

proposal outperforms the state-of-art change-point detection strategy, namely the Improved Generalized Likelihood Ratio (Improved GLR) while compares favorably with the original Bayesian Online Change Point Detection strategy.

Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation

- Amr Alexandari, Anshul Kundaje, Avanti Shrikumar
- abstract: Label shift refers to the phenomenon where the prior class probability p(y) changes between the training and test distributions, while the conditional probability p(xly) stays fixed. Label shift arises in settings like medical diagnosis, where a classifier trained to predict disease given symptoms must be adapted to scenarios where the baseline prevalence of the disease is different. Given estimates of p(ylx) from a predictive model, Saerens et al. proposed an efficient maximum likelihood algorithm to correct for label shift that does not require model retraining, but a limiting assumption of this algorithm is that p(ylx) is calibrated, which is not true of modern neural networks. Recently, Black Box Shift Learning (BBSL) and Regularized Learning under Label Shifts (RLLS) have emerged as state-of-the-art techniques to cope with label shift when a classifier does not output calibrated probabilities, but both methods require model retraining with importance weights and neither has been benchmarked against maximum likelihood. Here we (1) show that combining maximum likelihood with a type of calibration we call bias-corrected calibration outperforms both BBSL and RLLS across diverse datasets and distribution shifts, (2) prove that the maximum likelihood objective is concave, and (3) introduce a principled strategy for estimating source-domain priors that improves robustness to poor calibration. This work demonstrates that the maximum likelihood with appropriate calibration is a formidable and efficient baseline for label shift adaptation; notebooks reproducing experiments available at https://github.com/kundajelab/labelshiftexperiments , video: https://youtu.be/ZBXjE9QTruE , blogpost: https://bit.ly/3kTds7J

The Implicit Regularization of Stochastic Gradient Flow for Least Squares

- Alnur Ali, Edgar Dobriban, Ryan Tibshirani
- abstract: We study the implicit regularization of mini-batch stochastic gradient descent, when applied to the fundamental problem of least squares regression. We leverage a continuous-time stochastic differential equation having the same moments as stochastic gradient descent, which we call stochastic gradient flow. We give a bound on the excess risk of stochastic gradient flow at time t, over ridge regression with tuning parameter $\lambda = 1/t$. The bound may be computed from explicit constants (e.g., the mini-batch size, step size, number of iterations), revealing precisely how these quantities drive the excess risk. Numerical examples show the bound can be small, indicating a tight relationship between the two estimators. We give a similar result relating the coefficients of stochastic gradient flow and ridge. These results hold under no conditions on the data matrix X, and across the entire optimization path (not just at convergence).

Structural Language Models of Code

- Uri Alon, Roy Sadaka, Omer Levy, Eran Yahav
- abstract: We address the problem of any-code completion - generating a missing piece of source code in a given program without any restriction on the vocabulary or structure. We introduce a new approach to any-code completion that leverages the strict syntax of programming languages to model a code snippet as a tree - structural language modeling (SLM). SLM estimates the probability of the program's abstract syntax tree (AST) by decomposing it into a product of conditional probabilities over its nodes. We present a neural model that computes these conditional probabilities by considering all AST paths leading to a target node. Unlike previous techniques that have severely restricted the kinds of expressions that can be generated in this task, our approach can generate arbitrary code in any programming language. Our model significantly outperforms both seq2seq and a variety of structured approaches in generating Java and C# code. Our code, data, and trained models are available at http://github.com/tech-srl/slm-code-generation/. An online demo is available at http://AnyCodeGen.org.

LowFER: Low-rank Bilinear Pooling for Link Prediction

- Saadullah Amin, Stalin Varanasi, Katherine Ann Dunfield, Günter Neumann
- abstract: Knowledge graphs are incomplete by nature, with only a limited number of observed facts from the world knowledge being represented as structured relations between entities. To partly address this issue, an important task in statistical relational learning is that of link prediction or knowledge graph completion. Both linear and non-linear models have been proposed to solve the problem. Bilinear models, while expressive, are prone to overfitting and lead to quadratic growth of parameters in number of relations. Simpler models have become more standard, with certain constraints on bilinear map as relation parameters. In this work, we propose a factorized bilinear pooling model, commonly used in multi-modal learning, for better fusion of entities and relations, leading to an efficient and constraint-free model. We prove that our model is fully expressive, providing bounds on the embedding dimensionality and factorization rank. Our model naturally generalizes Tucker decomposition based TuckER model, which has been shown to generalize other models, as efficient low-rank approximation without substantially compromising the performance. Due to low-rank approximation, the model complexity can be controlled by the factorization rank, avoiding the possible cubic growth of TuckER. Empirically, we evaluate on real-world datasets, reaching on par or state-of-the-art performance. At extreme low-ranks, model preserves the performance while staying parameter efficient.

Discount Factor as a Regularizer in Reinforcement Learning

- Ron Amit, Ron Meir, Kamil Ciosek
- abstract: Specifying a Reinforcement Learning (RL) task involves choosing a suitable planning horizon, which is typically modeled by a discount factor. It is known that applying RL algorithms with a lower discount factor can act as a regularizer, improving performance in the limited data regime. Yet the exact nature of this regularizer has not been investigated. In this work, we fill in this gap. For several Temporal-Difference (TD) learning methods, we show an explicit equivalence between using a reduced discount factor and adding an explicit regularization term to the algorithm's loss. Motivated by the equivalence, we empirically study this technique compared to standard L2 regularization by extensive experiments in discrete and continuous domains, using tabular and functional representations. Our experiments suggest the regularization effectiveness is strongly related to properties of the available data, such as size, distribution, and mixing rate.

Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning"

- Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, Kazuhito Koishida
- abstract: Visual reasoning tasks such as visual question answering (VQA) require an interplay of visual perception with reasoning about the question semantics grounded in perception. However, recent advances in this area are still primarily driven by perception improvements (e.g. scene graph generation) rather than reasoning. Neuro-symbolic models such as Neural Module Networks bring the benefits of compositional reasoning to VQA, but they are still entangled with visual representation learning, and thus neural reasoning is hard to improve and assess on its own. To address this, we propose (1) a framework to isolate and evaluate the reasoning aspect of VQA separately from its perception, and (2) a novel top-down calibration technique that allows the model to answer reasoning questions even with imperfect perception. To this end, we introduce a Differentiable First-Order Logic formalism for VQA that explicitly decouples question answering from visual perception. On the challenging GQA dataset, this framework is used to perform in-depth, disentangled comparisons between well-known VQA models leading to informative insights regarding the participating models as well as the task.

The Differentiable Cross-Entropy Method

- Brandon Amos, Denis Yarats
- abstract: We study the Cross-Entropy Method (CEM) for the non-convex optimization of a continuous and parameterized objective function and introduce a differentiable variant that enables us to differentiate the output of CEM with respect to the objective function's parameters. In the machine learning setting this brings CEM inside of the end-to-end learning pipeline where this has otherwise been impossible. We show applications in a synthetic energy-based structured prediction task and in non-convex continuous control. In the control setting we show how to embed optimal action sequences into a lower-dimensional space. This enables us to use policy optimization to fine-tune modeling components by differentiating through the CEM-based controller.

Customizing ML Predictions for Online Algorithms

- Keerti Anand, Rong Ge, Debmalya Panigrahi
- abstract: A popular line of recent research incorporates ML advice in the design of online algorithms to improve their performance in typical instances. These papers treat the ML algorithm as a black-box, and redesign online algorithms to take advantage of ML predictions. In this paper, we ask the complementary question: can we redesign ML algorithms to provide better predictions for online algorithms? We explore this question in the context of the classic rent-or-buy problem, and show that incorporating optimization benchmarks in ML loss functions leads to significantly better performance, while maintaining a worst-case adversarial result when the advice is completely wrong. We support this finding both through theoretical bounds and numerical simulations.

Fairwashing explanations with off-manifold detergent

- Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, Pan Kessel
- abstract: Explanation methods promise to make black-box classifiers more transparent. As a result, it is hoped that they can act as proof for a sensible, fair and trustworthy decision-making process of the algorithm and thereby increase its acceptance by the end-users. In this paper, we show both theoretically and experimentally that these hopes are presently unfounded. Specifically, we show that, for any classifier g, one can always construct another classifier $\tilde{g}$ which has the same behavior on the data (same train, validation, and test error) but has arbitrarily manipulated explanation maps. We derive this statement theoretically using differential geometry and demonstrate it experimentally for various explanation methods, architectures, and datasets. Motivated by our theoretical insights, we then propose a modification of existing explanation methods which makes them significantly more robust.

Population-Based Black-Box Optimization for Biological Sequence Design

- Christof Angermueller, David Belanger, Andreea Gane, Zelda Mariet, David Dohan, Kevin Murphy, Lucy Colwell, D Sculley
- abstract: The use of black-box optimization for the design of new biological sequences is an emerging research area with potentially revolutionary impact. The cost and latency of wet-lab experiments requires methods that find good sequences in few experimental rounds of large batches of sequences — a setting that off-the-shelf black-box optimization methods are ill-equipped to handle. We find that the performance of existing methods varies drastically across optimization tasks, posing a significant obstacle to real-world applications. To improve robustness, we propose Population-Based Black-Box Optimization (P3BO), which generates batches of sequences by sampling from an ensemble of methods. The number of sequences sampled from any method is proportional to the quality of sequences it previously proposed, allowing P3BO to combine the strengths of individual methods while hedging against their innate brittleness. Adapting the hyper-parameters of each of the methods online using evolutionary optimization further improves performance. Through extensive experiments on in-silico optimization tasks, we show that P3BO outperforms any single method in its population, proposing higher quality sequences as well as more diverse batches. As such, P3BO and Adaptive-P3BO are a crucial step towards deploying ML to real-world sequence design.

Low-loss connection of weight vectors: distribution-based approaches

- Ivan Anokhin, Dmitry Yarotsky
- abstract: Recent research shows that sublevel sets of the loss surfaces of overparameterized networks are connected, exactly or approximately. We describe and compare experimentally a panel of methods used to connect two low-loss points by a low-loss curve on this surface. Our methods vary in accuracy and complexity. Most of our methods are based on "macroscopic" distributional assumptions and are insensitive to the detailed properties of the points to be connected. Some methods require a prior training of a "global connection model" which can then be applied to any pair of points. The accuracy of the method generally correlates with its complexity and sensitivity to the endpoint detail.

Online metric algorithms with untrusted predictions

- Antonios Antoniadis, Christian Coester, Marek Elias, Adam Polak, Bertrand Simon
- abstract: Machine-learned predictors, although achieving very good results for inputs resembling training data, cannot possibly provide perfect predictions in all situations. Still, decision-making systems that are based on such predictors need not only to benefit from good predictions but also to achieve a decent performance when the predictions are inadequate. In this paper, we propose a prediction setup for arbitrary metrical task systems (MTS) (e.g., caching, k-server and convex body chasing) and online matching on the line. We utilize results from the theory of online algorithms to show how to make the setup robust. Specifically for caching, we present an algorithm whose performance, as a function of the prediction error, is exponentially better than what is achievable for general MTS. Finally, we present an empirical evaluation of our methods on real world datasets, which suggests practicality.

NADS: Neural Architecture Distribution Search for Uncertainty Awareness

- Randy Ardywibowo, Shahin Boluki, Xinyu Gong, Zhangyang Wang, Xiaoning Qian
- abstract: Machine learning (ML) systems often encounter Out-of-Distribution (OoD) errors when dealing with testing data coming from a distribution different from training data. It becomes important for ML systems in critical applications to accurately quantify its predictive uncertainty and screen out these anomalous inputs. However, existing OoD detection approaches are prone to errors and even sometimes assign higher likelihoods to OoD samples. Unlike standard learning tasks, there is currently no well established guiding principle for designing OoD detection architectures that can accurately quantify uncertainty. To address these problems, we first seek to identify guiding principles for designing uncertainty-aware architectures, by proposing Neural Architecture Distribution Search (NADS). NADS searches for a distribution of architectures that perform well on a given task, allowing us to identify common building blocks among all uncertainty-aware architectures. With this formulation, we are able to optimize a stochastic OoD detection objective and construct an ensemble of models to perform OoD detection. We perform multiple OoD detection experiments and observe that our NADS performs favorably, with up to 57% improvement in accuracy compared to state-of-the-art methods among 15 different testing configurations.

Provable Representation Learning for Imitation Learning via Bi-level Optimization

- Sanjeev Arora, Simon Du, Sham Kakade, Yuping Luo, Nikunj Saunshi
- abstract: A common strategy in modern learning systems is to learn a representation that is useful for many tasks, a.k.a. representation learning. We study this strategy in the imitation learning setting for Markov decision processes (MDPs) where multiple experts' trajectories are available. We formulate representation learning as a bi-level optimization problem where the "outer" optimization tries to learn the joint representation and the "inner" optimization encodes the imitation learning setup and tries to learn task-specific parameters. We instantiate this framework for the imitation learning

settings of behavior cloning and observation-alone. Theoretically, we show using our framework that representation learning can provide sample complexity benefits for imitation learning in both settings. We also provide proof-of-concept experiments to verify our theory.

Quantum Boosting

- Srinivasan Arunachalam, Reevu Maity
- abstract: Boosting is a technique that boosts a weak and inaccurate machine learning algorithm into a strong accurate learning algorithm. The AdaBoost algorithm by Freund and Schapire (for which they were awarded the G{ö}del prize in 2003) is one of the widely used boosting algorithms, with many applications in theory and practice. Suppose we have a gamma-weak learner for a Boolean concept class C that takes time R(C), then the time complexity of AdaBoost scales as VC(C)poly(R(C), 1/gamma), where VC(C) is the VC-dimension of C. In this paper, we show how quantum techniques can improve the time complexity of classical AdaBoost. To this end, suppose we have a gamma-weak quantum learning algorithm for a Boolean concept class C that takes time Q(C), we introduce a quantum boosting algorithm whose complexity scales as sqrt{VC(C)}poly(Q(C),1/gamma); thereby achieving quadratic quantum improvement over classical AdaBoost in terms of VC(C).

Black-box Certification and Learning under Adversarial Perturbations

- Hassan Ashtiani, Vinayak Pathak, Ruth Urner
- abstract: We formally study the problem of classification under adversarial perturbations from a learner's perspective as well as a third-party who aims at certifying the robustness of a given black-box classifier. We analyze a PAC-type framework of semi-supervised learning and identify possibility and impossibility results for proper learning of VC-classes in this setting. We further introduce a new setting of black-box certification under limited query budget, and analyze this for various classes of predictors and perturbation. We also consider the viewpoint of a black-box adversary that aims at finding adversarial examples, showing that the existence of an adversary with polynomial query complexity can imply the existence of a sample efficient robust learner.

Invertible generative models for inverse problems: mitigating representation error and dataset bias

- Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, Paul Hand
- abstract: Trained generative models have shown remarkable performance as priors for inverse problems in imaging – for example, Generative Adversarial Network priors permit recovery of test images from 5-10x fewer measurements than sparsity priors. Unfortunately, these models may be unable to represent any particular image because of architectural choices, mode collapse, and bias in the training dataset. In this paper, we demonstrate that invertible neural networks, which have zero representation error by design, can be effective natural signal priors at inverse problems such as denoising, compressive sensing, and inpainting. Given a trained generative model, we study the empirical risk formulation of the desired inverse problem under a regularization that promotes high likelihood images, either directly by penalization or algorithmically by initialization. For compressive sensing, invertible priors can yield higher accuracy than sparsity priors across almost all undersampling ratios, and due to their lack of representation error, invertible priors can yield better reconstructions than GAN priors for images that have rare features of variation within the biased training set, including out-of-distribution natural images. We additionally compare performance for compressive sensing to unlearned methods, such as the deep decoder, and we establish theoretical bounds on expected recovery error in the case of a linear invertible model.

On the Convergence of Nesterov's Accelerated Gradient Method in Stochastic Settings

- Mahmoud Assran, Mike Rabbat
- abstract: We study Nesterov's accelerated gradient method with constant step-size and momentum parameters in the stochastic approximation setting (unbiased gradients with bounded variance) and the finite-sum setting (where randomness is due to sampling mini-batches). To build better insight into the behavior of Nesterov's method in stochastic settings, we focus throughout on objectives that are smooth, strongly-convex, and twice continuously differentiable. In the stochastic approximation setting, Nesterov's method converges to a neighborhood of the optimal point at the same accelerated rate as in the deterministic setting. Perhaps surprisingly, in the finite-sum setting, we prove that Nesterov's method may diverge with the usual choice of step-size and momentum, unless additional conditions on the problem related to conditioning and data coherence are satisfied. Our results shed light as to why Nesterov's method may fail to converge or achieve acceleration in the finite-sum setting.

Safe screening rules for L0-regression from Perspective Relaxations

- Alper Atamturk, Andres Gomez
- abstract: We give safe screening rules to eliminate variables from regression with ℓ_0 regularization or cardinality constraint. These rules are based on guarantees that a feature may or may not be selected in an optimal solution. The screening rules can be computed from a convex relaxation solution in linear time, without solving the L0-optimization problem. Thus, they can be used in a preprocessing step to safely remove variables from consideration apriori. Numerical experiments on real and synthetic data indicate that a significant number of the variables can be removed quickly, hence reducing the computational burden for optimization substantially. Therefore, the proposed fast and effective screening rules extend the scope of algorithms for L0-regression to larger data sets.

Adversarial Learning Guarantees for Linear Hypotheses and Neural Networks

- Pranjal Awasthi, Natalie Frank, Mehryar Mohri
- abstract: Adversarial or test time robustness measures the susceptibility of a classifier to perturbations to the test input. While there has been a flurry of recent work on designing defenses against such perturbations, the theory of adversarial robustness is not well understood. In order to make progress on this, we focus on the problem of understanding generalization in adversarial settings, via the lens of Rademacher complexity. We give upper and lower bounds for the adversarial empirical Rademacher complexity of linear hypotheses with adversarial perturbations measured in l_r-norm for an arbitrary $r \geq 1$. We then extend our analysis to provide Rademacher complexity lower and upper bounds for a single ReLU unit. Finally, we give adversarial Rademacher complexity bounds for feed-forward neural networks with one hidden layer.

Sample Amplification: Increasing Dataset Size even when Learning is Impossible

- Brian Axelrod, Shivam Garg, Vatsal Sharan, Gregory Valiant
- abstract: Given data drawn from an unknown distribution, D, to what extent is it possible to "amplify" this dataset and faithfully output an even larger set of samples that appear to have been drawn from D? We formalize this question as follows: an (n,m) amplification procedure takes as input n independent draws from an unknown distribution D, and outputs a set of m > n "samples" which must be indistinguishable from m samples drawn iid from D. We consider this sample amplification problem in two fundamental settings: the case where D is an arbitrary discrete distribution supported on k elements, and the case where D is a d-dimensional Gaussian with unknown mean, and fixed covariance matrix. Perhaps surprisingly, we show a valid amplification procedure exists for both of these settings, even in the regime where the size of the input dataset, n, is significantly less than what would be necessary to learn distribution D to non-trivial accuracy. We also show that our procedures are optimal up to constant factors. Beyond these results, we describe potential applications of such data amplification, and formalize a number of curious directions for future research along this vein.

Sparse Convex Optimization via Adaptively Regularized Hard Thresholding

- Kyriakos Axiotis, Maxim Sviridenko
- abstract: The goal of Sparse Convex Optimization is to optimize a convex function f under a sparsity constraint $s\leq s^\gamma$, where $s^$ is the target number of non-zero entries in a feasible solution (sparsity) and $\gamma\geq 1$ is an approximation factor. There has been a lot of work to analyze the sparsity guarantees of various algorithms (LASSO, Orthogonal Matching Pursuit (OMP), Iterative Hard Thresholding (IHT)) in terms of the Restricted Condition Number κ. The best known algorithms guarantee to find an approximate solution of value $f(x^)+\epsilon$ with the sparsity bound of $\gamma = O\left(\kappa\min\left{\log \frac{f(x^0)-f(x^)}{\epsilon}, \kappa\right}\right)$, where $x^$ is the target solution. We present a new Adaptively Regularized Hard Thresholding (ARHT) algorithm that makes significant progress on this problem by bringing the bound down to $\gamma=O(\kappa)$, which has been shown to be tight for a general class of algorithms including LASSO, OMP, and IHT. This is achieved without significant sacrifice in the runtime efficiency compared to the fastest known algorithms. We also provide a new analysis of OMP with Replacement (OMPR) for general f, under the condition $s > s^ \frac{\kappa^2}{4}$, which yields Compressed Sensing bounds under the Restricted Isometry Property (RIP). When compared to other Compressed Sensing approaches, it has the advantage of providing a strong tradeoff between the RIP condition and the solution sparsity, while working for any general function f that meets the RIP condition.

Model-Based Reinforcement Learning with Value-Targeted Regression

- Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, Lin Yang
- abstract: This paper studies model-based reinforcement learning (RL) for regret minimization. We focus on finite-horizon episodic RL where the transition model P belongs to a known family of models $\mathcal{P}$, a special case of which is when models in $\mathcal{P}$ take the form of linear mixtures: $P_{\theta} = \sum_{i=1}^{d} \theta_{i}P_{i}$. We propose a model based RL algorithm that is based on the optimism principle: In each episode, the set of models that are ‘consistent’ with the data collected is constructed. The criterion of consistency is based on the total squared error that the model incurs on the task of predicting \emph{state values} as determined by the last value estimate along the transitions. The next value function is then chosen by solving the optimistic planning problem with the constructed set of models. We derive a bound on the regret, which, in the special case of linear mixtures, takes the form $\tilde{\mathcal{O}}(d\sqrt{H^{3}T})$, where H, T and d are the horizon, the total number of steps and the dimension of θ, respectively. In particular, this regret bound is independent of the total number of states or actions, and is close to a lower bound $\Omega(\sqrt{HdT})$. For a general model family $\mathcal{P}$, the regret bound is derived based on the Eluder dimension.

Forecasting Sequential Data Using Consistent Koopman Autoencoders

- Omri Azencot, N. Benjamin Erichson, Vanessa Lin, Michael Mahoney
- abstract: Recurrent neural networks are widely used on time series data, yet such models often ignore the underlying physical structures in such sequences. A new class of physics-based methods related to Koopman theory has been introduced, offering an alternative for processing nonlinear dynamical systems. In this work, we propose a novel Consistent Koopman Autoencoder model which, unlike the majority of existing work, leverages the forward and backward dynamics. Key to our approach is a new analysis which explores the interplay between consistent dynamics and their associated Koopman operators. Our network is directly related to the derived analysis, and its computational requirements are comparable to other baselines. We evaluate our method on a wide range of high-dimensional and short-term dependent problems, and it achieves accurate estimates for significant prediction horizons, while also being robust to noise.

Constant Curvature Graph Convolutional Networks

- Gregor Bachmann, Gary Becigneul, Octavian Ganea
- abstract: Interest has been rising lately towards methods representing data in non-Euclidean spaces, e.g. hyperbolic or spherical that provide specific inductive biases useful for certain real-world data properties, e.g. scale-free, hierarchical or cyclical. However, the popular graph neural networks are currently limited in modeling data only via Euclidean geometry and associated vector space operations. Here, we bridge this gap by proposing mathematically grounded generalizations of graph convolutional networks (GCN) to (products of) constant curvature spaces. We do this by i) introducing a unified formalism permitting a differentiable interpolation between all geometries of constant curvature irrespective of their sign, ii) leveraging gyro-barycentric coordinates that generalize the classic Euclidean concept of the center of mass. Our class of models smoothly recover their Euclidean counterparts when the curvature goes to zero from either side. Empirically, we outperform Euclidean GCNs in the tasks of node classification and distortion minimization for symbolic data exhibiting non-Euclidean behavior, according to their discrete curvature.

Scalable Nearest Neighbor Search for Optimal Transport

- Arturs Backurs, Yihe Dong, Piotr Indyk, Ilya Razenshteyn, Tal Wagner
- abstract: The Optimal Transport (a.k.a. Wasserstein) distance is an increasingly popular similarity measure for rich data domains, such as images or text documents. This raises the necessity for fast nearest neighbor search algorithms according to this distance, which poses a substantial computational bottleneck on massive datasets. In this work we introduce Flowtree, a fast and accurate approximation algorithm for the Wasserstein-1 distance. We formally analyze its approximation factor and running time. We perform extensive experimental evaluation of nearest neighbor search algorithms in the W_1 distance on real-world dataset. Our results show that compared to previous state of the art, Flowtree achieves up to 7.4 times faster running time.

Agent57: Outperforming the Atari Human Benchmark

- Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, Charles Blundell
- abstract: Atari games have been a long-standing benchmark in the reinforcement learning (RL) community for the past decade. This benchmark was proposed to test general competency of RL algorithms. Previous work has achieved good average performance by doing outstandingly well on many games of the set, but very poorly in several of the most challenging games. We propose Agent57, the first deep RL agent that outperforms the standard human benchmark on all 57 Atari games. To achieve this result, we train a neural network which parameterizes a family of policies ranging from very exploratory to purely exploitative. We propose an adaptive mechanism to choose which policy to prioritize throughout the training process. Additionally, we utilize a novel parameterization of the architecture that allows for more consistent and stable learning.

Fiduciary Bandits

- Gal Bahar, Omer Ben-Porat, Kevin Leyton-Brown, Moshe Tennenholtz
- abstract: Recommendation systems often face exploration-exploitation tradeoffs: the system can only learn about the desirability of new options by recommending them to some user. Such systems can thus be modeled as multi-armed bandit settings; however, users are self-interested and cannot be made to follow recommendations. We ask whether exploration can nevertheless be performed in a way that scrupulously respects agents’ interests—i.e., by a system that acts as a fiduciary. More formally, we introduce a model in which a recommendation system faces an exploration-exploitation tradeoff under the constraint that it can never recommend any action that it knows yields lower reward in expectation than an agent would achieve if it acted alone. Our main contribution is a positive result: an asymptotically optimal, incentive compatible, and ex-ante individually rational recommendation algorithm.

Learning De-biased Representations with Biased Representations

- Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, Seong Joon Oh
- abstract: Many machine learning algorithms are trained and evaluated by splitting data from a single source into training and test sets. While such focus on in-distribution learning scenarios has led to interesting advancement, it has not been able to tell if models are relying on dataset biases as shortcuts for successful prediction (e.g., using snow cues for recognising snowmobiles), resulting in biased models that fail to generalise when the bias shifts to a different class. The cross-bias generalisation problem has been addressed by de-biasing training data through augmentation or re-sampling, which are often prohibitive due to the data collection cost (e.g., collecting images of a snowmobile on a desert) and the difficulty of quantifying or expressing biases in the first place. In this work, we propose a novel framework to train a de-biased representation by encouraging it to be different from a set of representations that are biased by design. This tactic is feasible in many scenarios where it is much easier to define a set of biased representations than to define and quantify bias. We demonstrate the efficacy of our method across a variety of synthetic and real-world biases; our experiments show that the method discourages models from taking bias shortcuts, resulting in improved generalisation. Source code is available at https://github.com/clovaai/rebias.

Deep k-NN for Noisy Labels

- Dara Bahri, Heinrich Jiang, Maya Gupta
- abstract: Modern machine learning models are often trained on examples with noisy labels that hurt performance and are hard to identify. In this paper, we provide an empirical study showing that a simple k-nearest neighbor-based filtering approach on the logit layer of a preliminary model can remove mislabeled training data and produce more accurate models than many recently proposed methods. We also provide new statistical guarantees into its efficacy.

Provable Self-Play Algorithms for Competitive Reinforcement Learning

- Yu Bai, Chi Jin
- abstract: Self-play, where the algorithm learns by playing against itself without requiring any direct supervision, has become the new weapon in modern Reinforcement Learning (RL) for achieving superhuman performance in practice. However, the majority of exisiting theory in reinforcement learning only applies to the setting where the agent plays against a fixed environment; it remains largely open whether self-play algorithms can be provably effective, especially when it is necessary to manage the exploration/exploitation tradeoff. We study self-play in competitive reinforcement learning under the setting of Markov games, a generalization of Markov decision processes to the two-player case. We introduce a self-play algorithm—Value Iteration with Upper/Lower Confidence Bound (VI-ULCB)—and show that it achieves regret $\mathcal{\tilde{O}}(\sqrt{T})$ after playing T steps of the game, where the regret is measured by the agent's performance against a fully adversarial opponent who can exploit the agent's strategy at any step. We also introduce an explore-then-exploit style algorithm, which achieves a slightly worse regret of $\mathcal{\tilde{O}}(T^{2/3})$, but is guaranteed to run in polynomial time even in the worst case. To the best of our knowledge, our work presents the first line of provably sample-efficient self-play algorithms for competitive reinforcement learning.

Sparse Subspace Clustering with Entropy-Norm

- Liang Bai, Jiye Liang
- abstract: In this paper, we provide an explicit theoretical connection between Sparse subspace clustering (SSC) and spectral clustering (SC) from the perspective of learning a data similarity matrix. We show that spectral clustering with Gaussian kernel can be viewed as sparse subspace clustering with entropy-norm (SSC+E). Compared to SSC, SSC+E can obtain an analytical, symmetrical, nonnegative and nonlinearly-representational similarity matrix. Besides, SSC+E makes use of Gaussian kernel to compute the sparse similarity matrix of objects, which can avoid the complex computation of the sparse optimization program of SSC. Finally, we provide the experimental analysis to compare the efficiency and effectiveness of sparse subspace clustering and spectral clustering on ten benchmark data sets. The theoretical and experimental analysis can well help users for the selection of high-dimensional data clustering algorithms.

Coresets for Clustering in Graphs of Bounded Treewidth

- Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, Xuan Wu
- abstract: We initiate the study of coresets for clustering in graph metrics, i.e., the shortest-path metric of edge-weighted graphs. Such clustering problems are essential to data analysis and used for example in road networks and data visualization. A coreset is a compact summary of the data that approximately preserves the clustering objective for every possible center set, and it offers significant efficiency improvements in terms of running time, storage, and communication, including in streaming and distributed settings. Our main result is a near-linear time construction of a coreset for k-Median in a general graph G, with size $O_{\epsilon, k}(\mathrm{tw}(G))$ where $\mathrm{tw}(G)$ is the treewidth of G, and we complement the construction with a nearly-tight size lower bound. The construction is based on the framework of Feldman and Langberg [STOC 2011], and our main technical contribution, as required by this framework, is a uniform bound of $O(\mathrm{tw}(G))$ on the shattering dimension under any point weights. We validate our coreset on real-world road networks, and our scalable algorithm constructs tiny coresets with high accuracy, which translates to a massive speedup of existing approximation algorithms such as local search for graph k-Median.

Refined bounds for algorithm configuration: The knife-edge of dual class approximability

- Maria-Florina Balcan, Tuomas Sandholm, Ellen Vitercik
- abstract: Automating algorithm configuration is growing increasingly necessary as algorithms come with more and more tunable parameters. It is common to tune parameters using machine learning, optimizing algorithmic performance (runtime or solution quality, for example) using a training set of problem instances from the specific domain at hand. We investigate a fundamental question about these techniques: how large should the training set be to ensure that a parameter's average empirical performance over the training set is close to its expected, future performance? We answer this question for algorithm configuration problems that exhibit a widely-applicable structure: the algorithm's performance as a function of its parameters can be approximated by a "simple" function. We show that if this approximation holds under the L∞-norm, we can provide strong sample complexity bounds, but if the approximation holds only under the Lp-norm for p < ∞, it is not possible to provide meaningful sample complexity bounds in the worst case. We empirically evaluate our bounds in the context of integer programming, obtaining sample complexity bounds that are up to 700 times smaller than the previously best-known bounds.

Ready Policy One: World Building Through Active Learning

- Philip Ball, Jack Parker-Holder, Aldo Pacchiano, Krzysztof Choromanski, Stephen Roberts
- abstract: Model-Based Reinforcement Learning (MBRL) offers a promising direction for sample efficient learning, often achieving state of the art results for continuous control tasks. However many existing MBRL methods rely on combining greedy policies with exploration heuristics, and even those which utilize principled exploration bonuses construct dual objectives in an ad hoc fashion. In this paper we introduce Ready Policy One (RP1), a framework that views MBRL as an active learning problem, where we aim to improve the world model in the fewest samples possible. RP1 achieves this by utilizing a hybrid objective function, which crucially adapts during optimization, allowing the algorithm to trade off reward v.s. exploration at different stages of learning. In addition, we introduce a principled mechanism to terminate sample collection once we have a rich enough trajectory batch to improve the model. We rigorously evaluate our method on a variety of continuous control tasks, and demonstrate statistically significant gains over existing approaches.

Stochastic Optimization for Regularized Wasserstein Estimators

- Marin Ballu, Quentin Berthet, Francis Bach
- abstract: Optimal transport is a foundational problem in optimization, that allows to compare probability distributions while taking into account geometric aspects. Its optimal objective value, the Wasserstein distance, provides an important loss between distributions that has been used in many applications throughout machine learning and statistics. Recent algorithmic progress on this problem and its regularized versions have made these tools increasingly popular. However, existing techniques require solving an optimization problem to obtain a single gradient of the loss, thus slowing down first-order methods to minimize the sum of losses, that require many such gradient computations. In this work, we introduce an algorithm to solve a regularized version of this problem of Wasserstein estimators, with a time per step which is sublinear in the natural dimensions of the problem. We introduce a dual formulation, and optimize it with stochastic gradient steps that can be computed directly from samples, without solving additional optimization problems at each step. Doing so, the estimation and computation tasks are performed jointly. We show that this algorithm can be extended to other tasks, including estimation of Wasserstein barycenters. We provide theoretical guarantees and illustrate the performance of our algorithm with experiments on synthetic data.

Dual Mirror Descent for Online Allocation Problems

- Santiago Balseiro, Haihao Lu, Vahab Mirrokni
- abstract: We consider online allocation problems with concave revenue functions and resource constraints, which are central problems in revenue management and online advertising. In these settings, requests arrive sequentially during a finite horizon and, for each request, a decision maker needs to choose an action that consumes a certain amount of resources and generates revenue. The revenue function and resource consumption of each request are drawn independently and at random from a probability distribution that is unknown to the decision maker. The objective is to maximize cumulative revenues subject to a constraint on the total consumption of resources. We design a general class of algorithms that achieve sub-linear expected regret compared to the hindsight optimal allocation. Our algorithms operate in the Lagrangian dual space: they maintain a dual multiplier for each resource that is updated using online mirror descent. By choosing the reference function accordingly, we recover dual sub-gradient descent and dual exponential weights algorithm. The resulting algorithms are simple, efficient, and shown to attain the optimal order of regret when the length of the horizon and the initial number of resources are scaled proportionally. We discuss applications to online bidding in repeated auctions with budget constraints and online proportional matching with high entropy.

Inductive-bias-driven Reinforcement Learning For Efficient Schedules in Heterogeneous Clusters

- Subho Banerjee, Saurabh Jha, Zbigniew Kalbarczyk, Ravishankar Iyer
- abstract: The problem of scheduling of workloads onto heterogeneous processors (e.g., CPUs, GPUs, FPGAs) is of fundamental importance in modern data centers. Current system schedulers rely on application/system-specific heuristics that have to be built on a case-by-case basis. Recent work has demonstrated ML techniques for automating the heuristic search by using black-box approaches which require significant training data and time, which make them challenging to use in practice. This paper presents Symphony, a scheduling framework that addresses the challenge in two ways: (i) a domain-driven Bayesian reinforcement learning (RL) model for scheduling, which inherently models the resource dependencies identified from the system architecture; and (ii) a sampling-based technique to compute the gradients of a Bayesian model without performing full probabilistic inference. Together, these techniques reduce both the amount of training data and the time required to produce scheduling policies that significantly outperform black-box approaches by up to 2.2{\texttimes}.

UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training

- Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, Hsiao-Wuen Hon
- abstract: We propose to pre-train a unified language model for both autoencoding and partially autoregressive language modeling tasks using a novel training procedure, referred to as a pseudo-masked language model (PMLM). Given an input text with masked tokens, we rely on conventional masks to learn inter-relations between corrupted tokens and context via autoencoding, and pseudo masks to learn intra-relations between masked spans via partially autoregressive modeling. With well-designed position embeddings and self-attention masks, the context encodings are reused to avoid redundant computation. Moreover, conventional masks used for autoencoding provide global masking information, so that all the position embeddings are accessible in partially autoregressive language modeling. In addition, the two tasks pre-train a unified language model as a bidirectional encoder and a sequence-to-sequence decoder, respectively. Our experiments show that the unified language models pre-trained using PMLM achieve new state-of-the-art results on a wide range of language understanding and generation tasks across several widely used benchmarks. The code and pre-trained models are available at https://github.com/microsoft/unilm.

Fast OSCAR and OWL Regression via Safe Screening Rules

- Runxue Bao, Bin Gu, Heng Huang
- abstract: Ordered Weighted L_{1} (OWL) regularized regression is a new regression analysis for high-dimensional sparse learning. Proximal gradient methods are used as standard approaches to solve OWL regression. However, it is still a burning issue to solve OWL regression due to considerable computational cost and memory usage when the feature or sample size is large. In this paper, we propose the first safe screening rule for OWL regression by exploring the order of the primal solution with the unknown order structure via an iterative strategy, which overcomes the difficulties of tackling the non-separable regularizer. It effectively avoids the updates of the parameters whose coefficients must be zero during the learning process. More importantly, the proposed screening rule can be easily applied to standard and stochastic proximal gradient methods. Moreover, we prove that the algorithms with our screening rule are guaranteed to have identical results with the original algorithms. Experimental results on a variety of datasets show that our screening rule leads to a significant computational gain without any loss of accuracy, compared to existing competitive algorithms.

Option Discovery in the Absence of Rewards with Manifold Analysis

- Amitay Bar, Ronen Talmon, Ron Meir
- abstract: Options have been shown to be an effective tool in reinforcement learning, facilitating improved exploration and learning. In this paper, we present an approach based on spectral graph theory and derive an algorithm that systematically discovers options without access to a specific reward or task assignment. As opposed to the common practice used in previous methods, our algorithm makes full use of the spectrum of the graph Laplacian. Incorporating modes associated with higher graph frequencies unravels domain subtleties, which are shown to be useful for option discovery. Using geometric and manifold-based analysis, we present a theoretical justification for the algorithm. In addition, we showcase its performance in several domains, demonstrating clear improvements compared to competing methods.

Learning the piece-wise constant graph structure of a varying Ising model

- Batiste Le Bars, Pierre Humbert, Argyris Kalogeratos, Nicolas Vayatis
- abstract: This work focuses on the estimation of multiple change-points in a time-varying Ising model that evolves piece-wise constantly. The aim is to identify both the moments at which significant changes occur in the Ising model, as well as the underlying graph structures. For this purpose, we propose to estimate the neighborhood of each node by maximizing a penalized version of its conditional log-likelihood. The objective of the penalization is

twofold: it imposes sparsity in the learned graphs and, thanks to a fused-type penalty, it also enforces them to evolve piece-wise constantly. Using few assumptions, we provide two change-points consistency theorems. Those are the first in the context of unknown number of change-points detection in time-varying Ising model. Finally, experimental results on several synthetic datasets and a real-world dataset demonstrate the performance of our method.

Frequency Bias in Neural Networks for Input of Non-Uniform Density

- Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, Shira Kritchman
- abstract: Recent works have partly attributed the generalization ability of over-parameterized neural networks to frequency bias – networks trained with gradient descent on data drawn from a uniform distribution find a low frequency fit before high frequency ones. As realistic training sets are not drawn from a uniform distribution, we here use the Neural Tangent Kernel (NTK) model to explore the effect of variable density on training dynamics. Our results, which combine analytic and empirical observations, show that when learning a pure harmonic function of frequency κ, convergence at a point $x \in \S^{d-1}$ occurs in time $O(\kappa^d/p(x))$ where $p(x)$ denotes the local density at x. Specifically, for data in $\S^1$ we analytically derive the eigenfunctions of the kernel associated with the NTK for two-layer networks. We further prove convergence results for deep, fully connected networks with respect to the spectral decomposition of the NTK. Our empirical study highlights similarities and differences between deep and shallow networks in this model.

Private Query Release Assisted by Public Data

- Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan Ullman, Steven Wu
- abstract: We study the problem of differentially private query release assisted by access to public data. In this problem, the goal is to answer a large class $\mathcal{H}$ of statistical queries with error no more than α using a combination of public and private samples. The algorithm is required to satisfy differential privacy only with respect to the private samples. We study the limits of this task in terms of the private and public sample complexities. Our upper and lower bounds on the private sample complexity have matching dependence on the dual VC-dimension of $\mathcal{H}$. For a large category of query classes, our bounds on the public sample complexity have matching dependence on α.

ECLIPSE: An Extreme-Scale Linear Program Solver for Web-Applications

- Kinjal Basu, Amol Ghoting, Rahul Mazumder, Yao Pan
- abstract: Key problems arising in web applications (with millions of users and thousands of items) can be formulated as linear programs involving billions to trillions of decision variables and constraints. Despite the appeal of linear program (LP) formulations, solving problems at these scales appear to be well beyond the capabilities of existing LP solvers. Often ad-hoc decomposition rules are used to approximately solve these LPs, which have limited optimality guarantees and may lead to sub-optimal performance in practice. In this work, we propose a distributed solver that solves a perturbation of the LP problems at scale via a gradient-based algorithm on the smooth dual of the perturbed LP. The main workhorses of our algorithm are distributed matrix-vector multiplications (with load balancing) and efficient projection operations on distributed machines. Experiments on real-world data show that our proposed LP solver, ECLIPSE, can solve problems with 10^{12} decision variables – well beyond the capabilities of current solvers.

On Second-Order Group Influence Functions for Black-Box Predictions

- Samyadeep Basu, Xuchen You, Soheil Feizi
- abstract: With the rapid adoption of machine learning systems in sensitive applications, there is an increasing need to make black-box models explainable. Often we want to identify an influential group of training samples in a particular test prediction for a given machine learning model. Existing influence functions tackle this problem by using first-order approximations of the effect of removing a sample from the training set on model parameters. To compute the influence of a group of training samples (rather than an individual point) in model predictions, the change in optimal model parameters after removing that group from the training set can be large. Thus, in such cases, the first-order approximation can be loose. In this paper, we address this issue and propose second-order influence functions for identifying influential groups in test-time predictions. For linear models, across different sizes and types of groups, we show that using the proposed second-order influence function improves the correlation between the computed influence values and the ground truth ones. We also show that second-order influence functions could be used with optimization techniques to improve the selection of the most influential group for a test-sample.

Kernel interpolation with continuous volume sampling

- Ayoub Belhadji, Rémi Bardenet, Pierre Chainais
- abstract: A fundamental task in kernel methods is to pick nodes and weights, so as to approximate a given function from an RKHS by the weighted sum of kernel translates located at the nodes. This is the crux of kernel density estimation, kernel quadrature, or interpolation from discrete samples. Furthermore, RKHSs offer a convenient mathematical and computational framework. We introduce and analyse continuous volume sampling (VS), the continuous counterpart -for choosing node locations- of a discrete distribution introduced in (Deshpande & Vempala, 2006). Our contribution is theoretical: we prove almost optimal bounds for interpolation and quadrature under VS. While similar bounds already exist for some specific RKHSs using ad-hoc node constructions, VS offers bounds that apply to any Mercer kernel and depend on the spectrum of the associated integration operator. We emphasize that, unlike previous randomized approaches that rely on regularized leverage scores or determinantal point processes, evaluating the pdf of VS only requires pointwise evaluations of the kernel. VS is thus naturally amenable to MCMC samplers.

Decoupled Greedy Learning of CNNs

- Eugene Belilovsky, Michael Eickenberg, Edouard Oyallon
- abstract: A commonly cited inefficiency of neural network training by back-propagation is the update locking problem: each layer must wait for the signal to propagate through the network before updating. In recent years multiple authors have considered alternatives that can alleviate this issue. In this context, we consider a simpler, but more effective, substitute that uses minimal feedback, which we call Decoupled Greedy Learning (DGL). It is based on a greedy relaxation of the joint training objective, recently shown to be effective in the context of Convolutional Neural Networks (CNNs) on large-scale image classification. We consider an optimization of this objective that permits us to decouple the layer training, allowing for layers or modules in networks to be trained with a potentially linear parallelization in layers. We show theoretically and empirically that this approach converges. Then, we empirically find that it can lead to better generalization than sequential greedy optimization and sometimes end-to-end back-propagation. We show an extension of this approach to asynchronous settings, where modules can operate with large communication delays, is possible with the use of a replay buffer. We demonstrate the effectiveness of DGL on the CIFAR-10 dataset against alternatives and on the large-scale ImageNet dataset.

The Cost-free Nature of Optimally Tuning Tikhonov Regularizers and Other Ordered Smoothers

- Pierre Bellec, Dana Yang
- abstract: We consider the problem of selecting the best estimator among a family of Tikhonov regularized estimators, or, alternatively, to select a linear combination of these regularizers that is as good as the best regularizer in the family. Our theory reveals that if the Tikhonov regularizers share the same penalty matrix with different tuning parameters, a convex procedure based on Q-aggregation achieves the mean square error of the best estimator, up to a small error term no larger than $C\sigma^2$, where σ^2 is the noise level and $C>0$ is an absolute constant. Remarkably, the error term does

not depend on the penalty matrix or the number of estimators as long as they share the same penalty matrix, i.e., it applies to any grid of tuning parameters, no matter how large the cardinality of the grid is. This reveals the surprising "cost-free" nature of optimally tuning Tikhonov regularizers, in striking contrast with the existing literature on aggregation of estimators where one typically has to pay a cost of $\sigma^2\log(M)$ where M is the number of estimators in the family. The result holds, more generally, for any family of ordered linear smoothers; this encompasses Ridge regression as well as Principal Component Regression. The result is extended to the problem of tuning Tikhonov regularizers with different penalty matrices.

Defense Through Diverse Directions

- Christopher Bender, Yang Li, Yifeng Shi, Michael K. Reiter, Junier Oliva
- abstract: In this work we develop a novel Bayesian neural network methodology to achieve strong adversarial robustness without the need for online adversarial training. Unlike previous efforts in this direction, we do not rely solely on the stochasticity of network weights by minimizing the divergence between the learned parameter distribution and a prior. Instead, we additionally require that the model maintain some expected uncertainty with respect to all input covariates. We demonstrate that by encouraging the network to distribute evenly across inputs, the network becomes less susceptible to localized, brittle features which imparts a natural robustness to targeted perturbations. We show empirical robustness on several benchmark datasets.

Interference and Generalization in Temporal Difference Learning

- Emmanuel Bengio, Joelle Pineau, Doina Precup
- abstract: We study the link between generalization and interference in temporal-difference (TD) learning. Interference is defined as the inner product of two different gradients, representing their alignment; this quantity emerges as being of interest from a variety of observations about neural networks, parameter sharing and the dynamics of learning. We find that TD easily leads to low-interference, under-generalizing parameters, while the effect seems reversed in supervised learning. We hypothesize that the cause can be traced back to the interplay between the dynamics of interference and bootstrapping. This is supported empirically by several observations: the negative relationship between the generalization gap and interference in TD, the negative effect of bootstrapping on interference and the local coherence of targets, and the contrast between the propagation rate of information in TD(0) versus TD(λ) and regression tasks such as Monte-Carlo policy evaluation. We hope that these new findings can guide the future discovery of better bootstrapping methods.

Preselection Bandits

- Viktor Bengs, Eyke Hüllermeier
- abstract: In this paper, we introduce the Preselection Bandit problem, in which the learner preselects a subset of arms (choice alternatives) for a user, which then chooses the final arm from this subset. The learner is not aware of the user's preferences, but can learn them from observed choices. In our concrete setting, we allow these choices to be stochastic and model the user's actions by means of the Plackett-Luce model. The learner's main task is to preselect subsets that eventually lead to highly preferred choices. To formalize this goal, we introduce a reasonable notion of regret and derive lower bounds on the expected regret. Moreover, we propose algorithms for which the upper bound on expected regret matches the lower bound up to a logarithmic term of the time horizon.

Efficient Policy Learning from Surrogate-Loss Classification Reductions

- Andrew Bennett, Nathan Kallus
- abstract: Recent work on policy learning from observational data has highlighted the importance of efficient policy evaluation and has proposed reductions to weighted (cost-sensitive) classification. But, efficient policy evaluation need not yield efficient estimation of policy parameters. We consider the estimation problem given by a weighted surrogate-loss classification with any score function, either direct, inverse-propensity-weighted, or doubly robust. We show that, under a correct specification assumption, the weighted classification formulation need not be efficient for policy parameters. We draw a contrast to actual (possibly weighted) binary classification, where correct specification implies a parametric model, while for policy learning it only implies a semi-parametric model. In light of this, we instead propose an estimation approach based on generalized method of moments, which is efficient for the policy parameters. We propose a particular method based on recent developments on solving moment problems using neural networks and demonstrate the efficiency and regret benefits of this method empirically.

Training Neural Networks for and by Interpolation

- Leonard Berrada, Andrew Zisserman, M. Pawan Kumar
- abstract: In modern supervised learning, many deep neural networks are able to interpolate the data: the empirical loss can be driven to near zero on all samples simultaneously. In this work, we explicitly exploit this interpolation property for the design of a new optimization algorithm for deep learning, which we term Adaptive Learning-rates for Interpolation with Gradients (ALI-G). ALI-G retains the two main advantages of Stochastic Gradient Descent (SGD), which are (i) a low computational cost per iteration and (ii) good generalization performance in practice. At each iteration, ALI-G exploits the interpolation property to compute an adaptive learning-rate in closed form. In addition, ALI-G clips the learning-rate to a maximal value, which we prove to be helpful for non-convex problems. Crucially, in contrast to the learning-rate of SGD, the maximal learning-rate of ALI-G does not require a decay schedule. This makes ALI-G considerably easier to tune than SGD. We prove the convergence of ALI-G in various stochastic settings. Notably, we tackle the realistic case where the interpolation property is satisfied up to some tolerance. We also provide experiments on a variety of deep learning architectures and tasks: (i) learning a differentiable neural computer; (ii) training a wide residual network on the SVHN data set; (iii) training a Bi-LSTM on the SNLI data set; and (iv) training wide residual networks and densely connected networks on the CIFAR data sets. ALI-G produces state-of-the-art results among adaptive methods, and even yields comparable performance with SGD, which requires manually tuned learning-rate schedules. Furthermore, ALI-G is simple to implement in any standard deep learning framework and can be used as a drop-in replacement in existing code.

Implicit differentiation of Lasso-type models for hyperparameter optimization

- Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, Joseph Salmon
- abstract: Setting regularization parameters for Lasso-type estimators is notoriously difficult, though crucial for obtaining the best accuracy. The most popular hyperparameter optimization approach is grid-search on a held-out dataset. However, grid-search requires to choose a predefined grid of parameters and scales exponentially in the number of parameters. Another class of approaches casts hyperparameter optimization as a bi-level optimization problem, typically solved by gradient descent. The key challenge for these approaches is the estimation of the gradient w.r.t. the hyperparameters. Computing that gradient via forward or backward automatic differentiation usually suffers from high memory consumption, while implicit differentiation typically involves solving a linear system which can be prohibitive and numerically unstable. In addition, implicit differentiation usually assumes smooth loss functions, which is not the case of Lasso-type problems. This work introduces an efficient implicit differentiation algorithm, without matrix inversion, tailored for Lasso-type problems. Our proposal scales to high-dimensional data by leveraging the sparsity of the solutions. Empirically, we demonstrate that the proposed method outperforms a large number of standard methods for hyperparameter optimization.

Online Learning with Imperfect Hints

- Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, Manish Purohit

- abstract: We consider a variant of the classical online linear optimization problem in which at every step, the online player receives a “hint” vector before choosing the action for that round. Rather surprisingly, it was shown that if the hint vector is guaranteed to have a positive correlation with the cost vector, then the online player can achieve a regret of $O(\log T)$, thus significantly improving over the $O(\sqrt{T})$ regret in the general setting. However, the result and analysis require the correlation property at \emph{all} time steps, thus raising the natural question: can we design online learning algorithms that are resilient to bad hints? In this paper we develop algorithms and nearly matching lower bounds for online learning with imperfect hints. Our algorithms are oblivious to the quality of the hints, and the regret bounds interpolate between the always-correlated hints case and the no-hints case. Our results also generalize, simplify, and improve upon previous results on optimistic regret bounds, which can be viewed as an additive version of hints.

When are Non-Parametric Methods Robust?

- Robi Bhattacharjee, Kamalika Chaudhuri
- abstract: A growing body of research has shown that many classifiers are susceptible to adversarial examples – small strategic modifications to test inputs that lead to misclassification. In this work, we study general non-parametric methods, with a view towards understanding when they are robust to these modifications. We establish general conditions under which non-parametric methods are r-consistent – in the sense that they converge to optimally robust and accurate classifiers in the large sample limit. Concretely, our results show that when data is well-separated, nearest neighbors and kernel classifiers are r-consistent, while histograms are not. For general data distributions, we prove that preprocessing by Adversarial Pruning (Yang et. al., 2019)– that makes data well-separated – followed by nearest neighbors or kernel classifiers also leads to r-consistency.

Learning and Sampling of Atomic Interventions from Observations

- Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, Ashwin Maran, Vinodchandran N. Variyam
- abstract: We study the problem of efficiently estimating the effect of an intervention on a single variable using observational samples. Our goal is to give algorithms with polynomial time and sample complexity in a non-parametric setting. Tian and Pearl (AAAI ’02) have exactly characterized the class of causal graphs for which causal effects of atomic interventions can be identified from observational data. We make their result quantitative. Suppose $\mathscr{P}$ is a causal model on a set V of n observable variables with respect to a given causal graph G, and let do(x) be an identifiable intervention on a variable X. We show that assuming that G has bounded in-degree and bounded c-components (k) and that the observational distribution satisfies a strong positivity condition: (i) [Evaluation] There is an algorithm that outputs with probability 2/3 an evaluator for a distribution P^ that satisfies TV(P(V | do(x)), P^(V)) < eps using m=O (n/eps^2) samples from P and O(mn) time. The evaluator can return in O(n) time the probability P^(v) for any assignment v to V. (ii) [Sampling] There is an algorithm that outputs with probability 2/3 a sampler for a distribution P^ that satisfies TV(P(V | do(x)), P^(V)) < eps using m=O (n/eps^2) samples from P and O(mn) time. The sampler returns an iid sample from P^ with probability 1 in O(n) time. We extend our techniques to estimate P(Y | do(x)) for a subset Y of variables of interest. We also show lower bounds for the sample complexity, demonstrating that our sample complexity has optimal dependence on the parameters n and eps, as well as if k=1 on the strong positivity parameter.

Near-optimal sample complexity bounds for learning Latent $k-$polytopes and applications to Ad-Mixtures

- Chiranjib Bhattacharyya, Ravindran Kannan
- abstract: Deriving Optimal bounds on Sample Complexity of Latent Variable models is an active area of research. Recently such bounds were obtained for Mixture of Gaussians \cite{HSNCAY18}, no such results are known for Ad-mixtures, a generalization of Mixture distributions. In this paper we show that $O^*(dk/m)$ *samples are sufficient to learn each of $k-$ topic vectors of LDA, a popular Ad-mixture model, with vocabulary size d and $m\in \Omega(1)$ words per document, to any constant error in L_1 norm. The result is a corollary of the major contribution of this paper: the first sample complexity upper bound for the problem (introduced in \cite{BK20}) of learning the vertices of a Latent $k-$ Polytope in $\RR^d$, given perturbed points from it. The bound, $O*$(dk/\beta)$, is optimal and linear in number of parameters. It applies to many stochastic models including a broad class Ad-mixtures. To demonstrate the generality of the approach we specialize the setting to Mixed Membership Stochastic Block Models(MMSB) and show for the first time that if an MMSB has k blocks, the sample complexity is $O^*(k^2)$ under usual assumptions.

Low-Rank Bottleneck in Multi-head Attention Models

- Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, Sanjiv Kumar
- abstract: Attention based Transformer architecture has enabled significant advances in the field of natural language processing. In addition to new pre-training techniques, recent improvements crucially rely on working with a relatively larger embedding dimension for tokens. Unfortunately, this leads to models that are prohibitively large to be employed in the downstream tasks. In this paper we identify one of the important factors contributing to the large embedding size requirement. In particular, our analysis highlights that the scaling between the number of heads and the size of each head in the current architecture gives rise to a low-rank bottleneck in attention heads, causing this limitation. We further validate this in our experiments. As a solution we propose to set the head size of an attention unit to input sequence length, and independent of the number of heads, resulting in multi-head attention layers with provably more expressive power. We empirically show that this allows us to train models with a relatively smaller embedding dimension and with better performance scaling.

Spectral Clustering with Graph Neural Networks for Graph Pooling

- Filippo Maria Bianchi, Daniele Grattarola, Cesare Alippi
- abstract: Spectral clustering (SC) is a popular clustering technique to find strongly connected communities on a graph. SC can be used in Graph Neural Networks (GNNs) to implement pooling operations that aggregate nodes belonging to the same cluster. However, the eigendecomposition of the Laplacian is expensive and, since clustering results are graph-specific, pooling methods based on SC must perform a new optimization for each new sample. In this paper, we propose a graph clustering approach that addresses these limitations of SC. We formulate a continuous relaxation of the normalized minCUT problem and train a GNN to compute cluster assignments that minimize this objective. Our GNN-based implementation is differentiable, does not require to compute the spectral decomposition, and learns a clustering function that can be quickly evaluated on out-of-sample graphs. From the proposed clustering method, we design a graph pooling operator that overcomes some important limitations of state-of-the-art graph pooling techniques and achieves the best performance in several supervised and unsupervised tasks.

Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders

- Ioana Bica, Ahmed Alaa, Mihaela Van Der Schaar
- abstract: The estimation of treatment effects is a pervasive problem in medicine. Existing methods for estimating treatment effects from longitudinal observational data assume that there are no hidden confounders, an assumption that is not testable in practice and, if it does not hold, leads to biased estimates. In this paper, we develop the Time Series Deconfounder, a method that leverages the assignment of multiple treatments over time to enable the estimation of treatment effects in the presence of multi-cause hidden confounders. The Time Series Deconfounder uses a novel recurrent neural network architecture with multitask output to build a factor model over time and infer latent variables that render the assigned treatments conditionally independent; then, it performs causal inference using these latent variables that act as substitutes for the multi-cause unobserved confounders. We provide a theoretical analysis for obtaining unbiased causal effects of time-varying exposures using the Time Series Deconfounder. Using both simulated and real data we show the effectiveness of our method in deconfounding the estimation of treatment responses over time.

Adversarial Robustness for Code

- Pavol Bielik, Martin Vechev
- abstract: Machine learning and deep learning in particular has been recently used to successfully address many tasks in the domain of code such as finding and fixing bugs, code completion, decompilation, type inference and many others. However, the issue of adversarial robustness of models for code has gone largely unnoticed. In this work, we explore this issue by: (i) instantiating adversarial attacks for code (a domain with discrete and highly structured inputs), (ii) showing that, similar to other domains, neural models for code are vulnerable to adversarial attacks, and (iii) combining existing and novel techniques to improve robustness while preserving high accuracy.

The Boomerang Sampler

- Joris Bierkens, Sebastiano Grazzi, Kengo Kamatani, Gareth Roberts
- abstract: This paper introduces the boomerang sampler as a novel class of continuous-time non-reversible Markov chain Monte Carlo algorithms. The methodology begins by representing the target density as a density, e^{-U}, with respect to a prescribed (usually) Gaussian measure and constructs a continuous trajectory consisting of a piecewise circular path. The method moves from one circular orbit to another according to a rate function which can be written in terms of U. We demonstrate that the method is easy to implement and demonstrate empirically that it can out-perform existing benchmark piecewise deterministic Markov processes such as the bouncy particle sampler and the Zig-Zag. In the Bayesian statistics context, these competitor algorithms are of substantial interest in the large data context due to the fact that they can adopt data subsampling techniques which are exact (ie induce no error in the stationary distribution). We demonstrate theoretically and empirically that we can also construct a control-variate subsampling boomerang sampler which is also exact, and which possesses remarkable scaling properties in the large data limit. We furthermore illustrate a factorised version on the simulation of diffusion bridges.

Tight Bounds on Minimax Regret under Logarithmic Loss via Self-Concordance

- Blair Bilodeau, Dylan Foster, Daniel Roy
- abstract: We consider the classical problem of sequential probability assignment under logarithmic loss while competing against an arbitrary, potentially nonparametric class of experts. We obtain tight bounds on the minimax regret via a new approach that exploits the self-concordance property of the logarithmic loss. We show that for any expert class with (sequential) metric entropy $\mathcal{O}(\gamma^{-p})$ at scale γ, the minimax regret is $\mathcal{O}(n^{\frac{p}{p+1}})$, and that this rate cannot be improved without additional assumptions on the expert class under consideration. As an application of our techniques, we resolve the minimax regret for nonparametric Lipschitz classes of experts.

My Fair Bandit: Distributed Learning of Max-Min Fairness with Multi-player Bandits

- Ilai Bistritz, Tavor Baharav, Amir Leshem, Nicholas Bambos
- abstract: Consider N cooperative but non-communicating players where each plays one out of M arms for T turns. Players have different utilities for each arm, representable as an NxM matrix. These utilities are unknown to the players. In each turn players receive noisy observations of their utility for their selected arm. However, if any other players selected the same arm that turn, they will all receive zero utility due to the conflict. No other communication or coordination between the players is possible. Our goal is to design a distributed algorithm that learns the matching between players and arms that achieves max-min fairness while minimizing the regret. We present an algorithm and prove that it is regret optimal up to a \log\log T factor. This is the first max-min fairness multi-player bandit algorithm with (near) order optimal regret.

Provable guarantees for decision tree induction: the agnostic setting

- Guy Blanc, Jane Lange, Li-Yang Tan
- abstract: We give strengthened provable guarantees on the performance of widely employed and empirically successful {\sl top-down decision tree learning heuristics}. While prior works have focused on the realizable setting, we consider the more realistic and challenging {\sl agnostic} setting. We show that for all monotone functions f and $s\in \mathbb{N}$, these heuristics construct a decision tree of size $s^{\tilde{O}((\log s)/\varepsilon^2)}$ that achieves error $\le \mathsf{opt}_s + \varepsilon$, where opt_s denotes the error of the optimal size-s decision tree for f. Previously such a guarantee was not known to be achievable by any algorithm, even one that is not based on top-down heuristics. We complement our algorithmic guarantee with a near-matching $s^{\tilde{\Omega}(\log s)}$ lower bound.

Fast Differentiable Sorting and Ranking

- Mathieu Blondel, Olivier Teboul, Quentin Berthet, Josip Djolonga
- abstract: The sorting operation is one of the most commonly used building blocks in computer programming. In machine learning, it is often used for robust statistics. However, seen as a function, it is piecewise linear and as a result includes many kinks where it is non-differentiable. More problematic is the related ranking operator, often used for order statistics and ranking metrics. It is a piecewise constant function, meaning that its derivatives are null or undefined. While numerous works have proposed differentiable proxies to sorting and ranking, they do not achieve the $O(n \log n)$ time complexity one would expect from sorting and ranking operations. In this paper, we propose the first differentiable sorting and ranking operators with $O(n \log n)$ time and $O(n)$ space complexity. Our proposal in addition enjoys exact computation and differentiation. We achieve this feat by constructing differentiable operators as projections onto the permutahedron, the convex hull of permutations, and using a reduction to isotonic optimization. Empirically, we confirm that our approach is an order of magnitude faster than existing approaches and showcase two novel applications: differentiable Spearman's rank correlation coefficient and least trimmed squares.

Beyond Signal Propagation: Is Feature Diversity Necessary in Deep Neural Network Initialization?

- Yaniv Blumenfeld, Dar Gilboa, Daniel Soudry
- abstract: Deep neural networks are typically initialized with random weights, with variances chosen to facilitate signal propagation and stable gradients. It is also believed that diversity of features is an important property of these initializations. We construct a deep convolutional network with identical features by initializing almost all the weights to 0. The architecture also enables perfect signal propagation and stable gradients, and achieves high accuracy on standard benchmarks. This indicates that random, diverse initializations are \emph{not} necessary for training neural networks. An essential element in training this network is a mechanism of symmetry breaking; we study this phenomenon and find that standard GPU operations, which are non-deterministic, can serve as a sufficient source of symmetry breaking to enable training.

Modulating Surrogates for Bayesian Optimization

- Erik Bodin, Markus Kaiser, Ieva Kazlauskaite, Zhenwen Dai, Neill Campbell, Carl Henrik Ek
- abstract: Bayesian optimization (BO) methods often rely on the assumption that the objective function is well-behaved, but in practice, this is seldom true for real-world objectives even if noise-free observations can be collected. Common approaches, which try to model the objective as precisely as possible, often fail to make progress by spending too many evaluations modeling irrelevant details. We address this issue by proposing surrogate models that focus on the well-behaved structure in the objective function, which is informative for search, while ignoring detrimental structure that is challenging to model

from few observations. First, we demonstrate that surrogate models with appropriate noise distributions can absorb challenging structures in the objective function by treating them as irreducible uncertainty. Secondly, we show that a latent Gaussian process is an excellent surrogate for this purpose, comparing with Gaussian processes with standard noise distributions. We perform numerous experiments on a range of BO benchmarks and find that our approach improves reliability and performance when faced with challenging objective functions.

Deep Coordination Graphs

- Wendelin Boehmer, Vitaly Kurin, Shimon Whiteson
- abstract: This paper introduces the deep coordination graph (DCG) for collaborative multi-agent reinforcement learning. DCG strikes a flexible trade-off between representational capacity and generalization by factoring the joint value function of all agents according to a coordination graph into payoffs between pairs of agents. The value can be maximized by local message passing along the graph, which allows training of the value function end-to-end with Q-learning. Payoff functions are approximated with deep neural networks that employ parameter sharing and low-rank approximations to significantly improve sample efficiency. We show that DCG can solve predator-prey tasks that highlight the relative overgeneralization pathology, as well as challenging StarCraft II micromanagement tasks.

Lorentz Group Equivariant Neural Network for Particle Physics

- Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, Risi Kondor
- abstract: We present a neural network architecture that is fully equivariant with respect to transformations under the Lorentz group, a fundamental symmetry of space and time in physics. The architecture is based on the theory of the finite-dimensional representations of the Lorentz group and the equivariant nonlinearity involves the tensor product. For classification tasks in particle physics, we show that such an equivariant architecture leads to drastically simpler models that have relatively few learnable parameters and are much more physically interpretable than leading approaches that use CNNs and point cloud approaches. The performance of the network is tested on a public classification dataset [https://zenodo.org/record/2603256] for tagging top quark decays given energy-momenta of jet constituents produced in proton-proton collisions.

Efficient Robustness Certificates for Discrete Data: Sparsity-Aware Randomized Smoothing for Graphs, Images and More

- Aleksandar Bojchevski, Johannes Gasteiger, Stephan Günnemann
- abstract: Existing techniques for certifying the robustness of models for discrete data either work only for a small class of models or are general at the expense of efficiency or tightness. Moreover, they do not account for sparsity in the input which, as our findings show, is often essential for obtaining non-trivial guarantees. We propose a model-agnostic certificate based on the randomized smoothing framework which subsumes earlier work and is tight, efficient, and sparsity-aware. Its computational complexity does not depend on the number of discrete categories or the dimension of the input (e.g. the graph size), making it highly scalable. We show the effectiveness of our approach on a wide variety of models, datasets, and tasks – specifically highlighting its use for Graph Neural Networks. So far, obtaining provable guarantees for GNNs has been difficult due to the discrete and non-i.i.d. nature of graph data. Our method can certify any GNN and handles perturbations to both the graph structure and the node attributes.

Proper Network Interpretability Helps Adversarial Robustness in Classification

- Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, Luca Daniel
- abstract: Recent works have empirically shown that there exist adversarial examples that can be hidden from neural network interpretability (namely, making network interpretation maps visually similar), or interpretability is itself susceptible to adversarial attacks. In this paper, we theoretically show that with a proper measurement of interpretation, it is actually difficult to prevent prediction-evasion adversarial attacks from causing interpretation discrepancy, as confirmed by experiments on MNIST, CIFAR-10 and Restricted ImageNet. Spurred by that, we develop an interpretability-aware defensive scheme built only on promoting robust interpretation (without the need for resorting to adversarial loss minimization). We show that our defense achieves both robust classification and robust interpretation, outperforming state-of-the-art adversarial training methods against attacks of large perturbation in particular.

Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks

- Blake Bordelon, Abdulkadir Canatar, Cengiz Pehlevan
- abstract: We derive analytical expressions for the generalization performance of kernel regression as a function of the number of training samples using theoretical methods from Gaussian processes and statistical physics. Our expressions apply to wide neural networks due to an equivalence between training them and kernel regression with the Neural Tangent Kernel (NTK). By computing the decomposition of the total generalization error due to different spectral components of the kernel, we identify a new spectral principle: as the size of the training set grows, kernel machines and neural networks fit successively higher spectral modes of the target function. When data are sampled from a uniform distribution on a high-dimensional hypersphere, dot product kernels, including NTK, exhibit learning stages where different frequency modes of the target function are learned. We verify our theory with simulations on synthetic data and MNIST dataset.

Small Data, Big Decisions: Model Selection in the Small-Data Regime

- Jorg Bornschein, Francesco Visin, Simon Osindero
- abstract: Highly overparametrized neural networks can display curiously strong generalization performance – a phenomenon that has recently garnered a wealth of theoretical and empirical research in order to better understand it. In contrast to most previous work, which typically considers the performance as a function of the model size, in this paper we empirically study the generalization performance as the size of the training set varies over multiple orders of magnitude. These systematic experiments lead to some interesting and potentially very useful observations; perhaps most notably that training on smaller subsets of the data can lead to more reliable model selection decisions whilst simultaneously enjoying smaller computational overheads. Our experiments furthermore allow us to estimate Minimum Description Lengths for common datasets given modern neural network architectures, thereby paving the way for principled model selection taking into account Occams-razor.

Latent Variable Modelling with Hyperbolic Normalizing Flows

- Joey Bose, Ariella Smofsky, Renjie Liao, Prakash Panangaden, Will Hamilton
- abstract: The choice of approximate posterior distributions plays a central role in stochastic variational inference (SVI). One effective solution is the use of normalizing flows \cut{defined on Euclidean spaces} to construct flexible posterior distributions. However, one key limitation of existing normalizing flows is that they are restricted to the Euclidean space and are ill-equipped to model data with an underlying hierarchical structure. To address this fundamental limitation, we present the first extension of normalizing flows to hyperbolic spaces. We first elevate normalizing flows to hyperbolic spaces using coupling transforms defined on the tangent bundle, termed Tangent Coupling ($\mathcal{TC}$). We further introduce Wrapped Hyperboloid Coupling ($\mathcal{W}\mathbb{H}C$), a fully invertible and learnable transformation that explicitly utilizes the geometric structure of hyperbolic spaces, allowing for expressive posteriors while being efficient to sample from. We demonstrate the efficacy of our novel normalizing flow over hyperbolic VAEs and Euclidean normalizing flows. Our approach achieves improved performance on density estimation, as well as reconstruction of real-

world graph data, which exhibit a hierarchical structure. Finally, we show that our approach can be used to power a generative model over hierarchical data using hyperbolic latent variables.

Tightening Exploration in Upper Confidence Reinforcement Learning

- Hippolyte Bourel, Odalric Maillard, Mohammad Sadegh Talebi
- abstract: The upper confidence reinforcement learning (UCRL2) algorithm introduced in \citep{jaksch2010near} is a popular method to perform regret minimization in unknown discrete Markov Decision Processes under the average-reward criterion. Despite its nice and generic theoretical regret guarantees, this algorithm and its variants have remained until now mostly theoretical as numerical experiments in simple environments exhibit long burn-in phases before the learning takes place. In pursuit of practical efficiency, we present UCRL3, following the lines of UCRL2, but with two key modifications: First, it uses state-of-the-art time-uniform concentration inequalities to compute confidence sets on the reward and (component-wise) transition distributions for each state-action pair. Furthermore, to tighten exploration, it uses an adaptive computation of the support of each transition distribution, which in turn enables us to revisit the extended value iteration procedure of UCRL2 to optimize over distributions with reduced support by disregarding low probability transitions, while still ensuring near-optimism. We demonstrate, through numerical experiments in standard environments, that reducing exploration this way yields a substantial numerical improvement compared to UCRL2 and its variants. On the theoretical side, these key modifications enable us to derive a regret bound for UCRL3 improving on UCRL2, that for the first time makes appear notions of local diameter and local effective support, thanks to variance-aware concentration bounds.

Preference Modeling with Context-Dependent Salient Features

- Amanda Bower, Laura Balzano
- abstract: We consider the problem of estimating a ranking on a set of items from noisy pairwise comparisons given item features. We address the fact that pairwise comparison data often reflects irrational choice, e.g. intransitivity. Our key observation is that two items compared in isolation from other items may be compared based on only a salient subset of features. Formalizing this framework, we propose the salient feature preference model and prove a finite sample complexity result for learning the parameters of our model and the underlying ranking with maximum likelihood estimation. We also provide empirical results that support our theoretical bounds and illustrate how our model explains systematic intransitivity. Finally we demonstrate strong performance of maximum likelihood estimation of our model on both synthetic data and two real data sets: the UT Zappos50K data set and comparison data about the compactness of legislative districts in the US.

Adversarial Filters of Dataset Biases

- Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, Yejin Choi
- abstract: Large neural models have demonstrated human-level performance on language and vision benchmarks, while their performance degrades considerably on adversarial or out-of-distribution samples. This raises the question of whether these models have learned to solve a dataset rather than the underlying task by overfitting to spurious dataset biases. We investigate one recently proposed approach, AFLITE, which adversarially filters such dataset biases, as a means to mitigate the prevalent overestimation of machine performance. We provide a theoretical understanding for AFLITE, by situating it in the generalized framework for optimum bias reduction. We present extensive supporting evidence that AFLITE is broadly applicable for reduction of measurable dataset biases, and that models trained on the filtered datasets yield better generalization to out-of-distribution tasks. Finally, filtering results in a large drop in model performance (e.g., from 92% to 62% for SNLI), while human performance still remains high. Our work thus shows that such filtered datasets can pose new research challenges for robust generalization by serving as upgraded benchmarks.

Calibration, Entropy Rates, and Memory in Language Models

- Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, Yi Zhang
- abstract: Building accurate language models that capture meaningful long-term dependencies is a core challenge in natural language processing. Towards this end, we present a calibration-based approach to measure long-term discrepancies between a generative sequence model and the true distribution, and use these discrepancies to improve the model. Empirically, we show that state-of-the-art language models, including LSTMs and Transformers, are miscalibrated: the entropy rates of their generations drift dramatically upward over time. We then provide provable methods to mitigate this phenomenon. Furthermore, we show how this calibration-based approach can also be used to measure the amount of memory that language models use for prediction.

Schatten Norms in Matrix Streams: Hello Sparsity, Goodbye Dimension

- Vladimir Braverman, Robert Krauthgamer, Aditya Krishnan, Roi Sinoff
- abstract: Spectral functions of large matrices contains important structural information about the underlying data, and is thus becoming increasingly important. Many times, large matrices representing real-world data are sparse or doubly sparse (i.e., sparse in both rows and columns), and are accessed as a stream of updates, typically organized in row-order. In this setting, where space (memory) is the limiting resource, all known algorithms require space that is polynomial in the dimension of the matrix, even for sparse matrices. We address this challenge by providing the first algorithms whose space requirement is independent of the matrix dimension, assuming the matrix is doubly-sparse and presented in row-order. Our algorithms approximate the Schatten p-norms, which we use in turn to approximate other spectral functions, such as logarithm of the determinant, trace of matrix inverse, and Estrada index. We validate these theoretical performance bounds by numerical experiments on real-world matrices representing social networks. We further prove that multiple passes are unavoidable in this setting, and show extensions of our primary technique, including a trade-off between space requirements and number of passes.

All in the Exponential Family: Bregman Duality in Thermodynamic Variational Inference

- Rob Brekelmans, Vaden Masrani, Frank Wood, Greg Ver Steeg, Aram Galstyan
- abstract: The recently proposed Thermodynamic Variational Objective (TVO) leverages thermodynamic integration to provide a family of variational inference objectives, which both tighten and generalize the ubiquitous Evidence Lower Bound (ELBO). However, the tightness of TVO bounds was not previously known, an expensive grid search was used to choose a “schedule” of intermediate distributions, and model learning suffered with ostensibly tighter bounds. In this work, we propose an exponential family interpretation of the geometric mixture curve underlying the TVO and various path sampling methods, which allows us to characterize the gap in TVO likelihood bounds as a sum of KL divergences. We propose to choose intermediate distributions using equal spacing in the moment parameters of our exponential family, which matches grid search performance and allows the schedule to adaptively update over the course of training. Finally, we derive a doubly reparameterized gradient estimator which improves model learning and allows the TVO to benefit from more refined bounds. To further contextualize our contributions, we provide a unified framework for understanding thermodynamic integration and the TVO using Taylor series remainders.

Estimating the Number and Effect Sizes of Non-null Hypotheses

- Jennifer Brennan, Ramya Korlakai Vinayak, Kevin Jamieson
- abstract: We study the problem of estimating the distribution of effect sizes (the mean of the test statistic under the alternate hypothesis) in a multiple testing setting. Knowing this distribution allows us to calculate the power (type II error) of any experimental design. We show that it is possible to

estimate this distribution using an inexpensive pilot experiment, which takes significantly fewer samples than would be required by an experiment that identified the discoveries. Our estimator can be used to guarantee the number of discoveries that will be made using a given experimental design in a future experiment. We prove that this simple and computationally efficient estimator enjoys a number of favorable theoretical properties, and demonstrate its effectiveness on data from a gene knockout experiment on influenza inhibition in Drosophila.

The FAST Algorithm for Submodular Maximization

- Adam Breuer, Eric Balkanski, Yaron Singer
- abstract: In this paper we describe a new parallel algorithm called Fast Adaptive Sequencing Technique (FAST) for maximizing a monotone submodular function under a cardinality constraint k. This algorithm achieves the optimal 1-1/e approximation guarantee and is orders of magnitude faster than the state-of-the-art on a variety of experiments over real-world data sets. Following recent work by Balkanski and Singer (2018), there has been a great deal of research on algorithms whose theoretical parallel runtime is exponentially faster than algorithms used for submodular maximization over the past 40 years. However, while these new algorithms are fast in terms of asymptotic worst-case guarantees, it is computationally infeasible to use them in practice even on small data sets because the number of rounds and queries they require depend on large constants and high-degree polynomials in terms of precision and confidence. The design principles behind the FAST algorithm we present here are a significant departure from those of recent theoretically fast algorithms. Rather than optimize for asymptotic theoretical guarantees, the design of FAST introduces several new techniques that achieve remarkable practical and theoretical parallel runtimes. The approximation guarantee obtained by FAST is arbitrarily close to 1 - 1/e, and its asymptotic parallel runtime (adaptivity) is O(log(n) log^2(log k)) using O(n log log(k)) total queries. We show that FAST is orders of magnitude faster than any algorithm for submodular maximization we are aware of, including hyper-optimized parallel versions of state-of-the-art serial algorithms, by running experiments on large data sets.

GNN-FiLM: Graph Neural Networks with Feature-wise Linear Modulation

- Marc Brockschmidt
- abstract: This paper presents a new Graph Neural Network (GNN) type using feature-wise linear modulation (FiLM). Many standard GNN variants propagate information along the edges of a graph by computing messages based only on the representation of the source of each edge. In GNN-FiLM, the representation of the target node of an edge is used to compute a transformation that can be applied to all incoming messages, allowing feature-wise modulation of the passed information. Different GNN architectures are compared in extensive experiments on three tasks from the literature, using re-implementations of many baseline methods. Hyperparameters for all methods were found using extensive search, yielding somewhat surprising results: differences between state of the art models are much smaller than reported in the literature and well-known simple baselines that are often not compared to perform better than recently proposed GNN variants. Nonetheless, GNN-FiLM outperforms these methods on a regression task on molecular graphs and performs competitively on other tasks.

TaskNorm: Rethinking Batch Normalization for Meta-Learning

- John Bronskill, Jonathan Gordon, James Requeima, Sebastian Nowozin, Richard Turner
- abstract: Modern meta-learning approaches for image classification rely on increasingly deep networks to achieve state-of-the-art performance, making batch normalization an essential component of meta-learning pipelines. However, the hierarchical nature of the meta-learning setting presents several challenges that can render conventional batch normalization ineffective, giving rise to the need to rethink normalization in this setting. We evaluate a range of approaches to batch normalization for meta-learning scenarios, and develop a novel approach that we call TaskNorm. Experiments on fourteen datasets demonstrate that the choice of batch normalization has a dramatic effect on both classification accuracy and training time for both gradient based- and gradient-free meta-learning approaches. Importantly, TaskNorm is found to consistently improve performance. Finally, we provide a set of best practices for normalization that will allow fair comparison of meta-learning algorithms.

Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences

- Daniel Brown, Russell Coleman, Ravi Srinivasan, Scott Niekum
- abstract: Bayesian reward learning from demonstrations enables rigorous safety and uncertainty analysis when performing imitation learning. However, Bayesian reward learning methods are typically computationally intractable for complex control problems. We propose Bayesian Reward Extrapolation (Bayesian REX), a highly efficient Bayesian reward learning algorithm that scales to high-dimensional imitation learning problems by pre-training a low-dimensional feature encoding via self-supervised tasks and then leveraging preferences over demonstrations to perform fast Bayesian inference. Bayesian REX can learn to play Atari games from demonstrations, without access to the game score and can generate 100,000 samples from the posterior over reward functions in only 5 minutes on a personal laptop. Bayesian REX also results in imitation learning performance that is competitive with or better than state-of-the-art methods that only learn point estimates of the reward function. Finally, Bayesian REX enables efficient high-confidence policy evaluation without having access to samples of the reward function. These high-confidence performance bounds can be used to rank the performance and risk of a variety of evaluation policies and provide a way to detect reward hacking behaviors.

A Pairwise Fair and Community-preserving Approach to k-Center Clustering

- Brian Brubach, Darshan Chakrabarti, John Dickerson, Samir Khuller, Aravind Srinivasan, Leonidas Tsepenekas
- abstract: Clustering is a foundational problem in machine learning with numerous applications. As machine learning increases in ubiquity as a backend for automated systems, concerns about fairness arise. Much of the current literature on fairness deals with discrimination against protected classes in supervised learning (group fairness). We define a different notion of fair clustering wherein the probability that two points (or a community of points) become separated is bounded by an increasing function of their pairwise distance (or community diameter). We capture the situation where data points represent people who gain some benefit from being clustered together. Unfairness arises when certain points are deterministically separated, either arbitrarily or by someone who intends to harm them as in the case of gerrymandering election districts. In response, we formally define two new types of fairness in the clustering setting, pairwise fairness and community preservation. To explore the practicality of our fairness goals, we devise an approach for extending existing k-center algorithms to satisfy these fairness constraints. Analysis of this approach proves that reasonable approximations can be achieved while maintaining fairness. In experiments, we compare the effectiveness of our approach to classical k-center algorithms/heuristics and explore the tradeoff between optimal clustering and fairness.

Scalable Exact Inference in Multi-Output Gaussian Processes

- Wessel Bruinsma, Eric Perim, William Tebbutt, Scott Hosking, Arno Solin, Richard Turner
- abstract: Multi-output Gaussian processes (MOGPs) leverage the flexibility and interpretability of GPs while capturing structure across outputs, which is desirable, for example, in spatio-temporal modelling. The key problem with MOGPs is their computational scaling $O(n^3 p^3)$, which is cubic in the number of both inputs n (e.g., time points or locations) and outputs p. For this reason, a popular class of MOGPs assumes that the data live around a low-dimensional linear subspace, reducing the complexity to $O(n^3 m^3)$. However, this cost is still cubic in the dimensionality of the subspace m, which is still prohibitively expensive for many applications. We propose the use of a sufficient statistic of the data to accelerate inference and learning in MOGPs with orthogonal bases. The method achieves linear scaling in m in practice, allowing these models to scale to large m without sacrificing significant expressivity or requiring approximation. This advance opens up a wide range of real-world tasks and can be combined with existing GP approximations in a plug-and-play way. We demonstrate the efficacy of the method on various synthetic and real-world data sets.

Online Pricing with Offline Data: Phase Transition and Inverse Square Law

- Jinzhi Bu, David Simchi-Levi, Yunzong Xu
- abstract: This paper investigates the impact of pre-existing offline data on online learning, in the context of dynamic pricing. We study a single-product dynamic pricing problem over a selling horizon of T periods. The demand in each period is determined by the price of the product according to a linear demand model with unknown parameters. We assume that the seller already has some pre-existing offline data before the start of the selling horizon. The seller wants to utilize both the pre-existing offline data and the sequential online data to minimize the regret of the online learning process. We characterize the joint effect of the size, location and dispersion of the offline data on the optimal regret of the online learning process. Our results reveal surprising transformations of the optimal regret rate with respect to the size of the offline data, which we refer to as phase transitions. In addition, our results demonstrate that the location and dispersion of the offline data also have an intrinsic effect on the optimal regret, and we quantify this effect via the inverse-square law.

Empirical Study of the Benefits of Overparameterization in Learning Latent Variable Models

- Rares-Darius Buhai, Yoni Halpern, Yoon Kim, Andrej Risteski, David Sontag
- abstract: One of the most surprising and exciting discoveries in supervised learning was the benefit of overparameterization (i.e. training a very large model) to improving the optimization landscape of a problem, with minimal effect on statistical performance (i.e. generalization). In contrast, unsupervised settings have been under-explored, despite the fact that it was observed that overparameterization can be helpful as early as Dasgupta & Schulman (2007). We perform an empirical study of different aspects of overparameterization in unsupervised learning of latent variable models via synthetic and semi-synthetic experiments. We discuss benefits to different metrics of success (recovering the parameters of the ground-truth model, held-out log-likelihood), sensitivity to variations of the training algorithm, and behavior as the amount of overparameterization increases. We find that across a variety of models (noisy-OR networks, sparse coding, probabilistic context-free grammars) and training algorithms (variational inference, alternating minimization, expectation-maximization), overparameterization can significantly increase the number of ground truth latent variables recovered.

DeBayes: a Bayesian Method for Debiasing Network Embeddings

- Maarten Buyl, Tijl De Bie
- abstract: As machine learning algorithms are increasingly deployed for high-impact automated decision making, ethical and increasingly also legal standards demand that they treat all individuals fairly, without discrimination based on their age, gender, race or other sensitive traits. In recent years much progress has been made on ensuring fairness and reducing bias in standard machine learning settings. Yet, for network embedding, with applications in vulnerable domains ranging from social network analysis to recommender systems, current options remain limited both in number and performance. We thus propose DeBayes: a conceptually elegant Bayesian method that is capable of learning debiased embeddings by using a biased prior. Our experiments show that these representations can then be used to perform link prediction that is significantly more fair in terms of popular metrics such as demographic parity and equalized opportunity.

Structured Prediction with Partial Labelling through the Infimum Loss

- Vivien Cabannnes, Alessandro Rudi, Francis Bach
- abstract: Annotating datasets is one of the main costs in nowadays supervised learning. The goal of weak supervision is to enable models to learn using only forms of labelling which are cheaper to collect, as partial labelling. This is a type of incomplete annotation where, for each datapoint, supervision is cast as a set of labels containing the real one. The problem of supervised learning with partial labelling has been studied for specific instances such as classification, multi-label, ranking or segmentation, but a general framework is still missing. This paper provides a unified framework based on structured prediction and on the concept of \emph{infimum loss} to deal with partial labelling over a wide family of learning problems and loss functions. The framework leads naturally to explicit algorithms that can be easily implemented and for which proved statistical consistency and learning rates. Experiments confirm the superiority of the proposed approach over commonly used baselines.

Online Learned Continual Compression with Adaptive Quantization Modules

- Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Joelle Pineau
- abstract: We introduce and study the problem of Online Continual Compression, where one attempts to simultaneously learn to compress and store a representative dataset from a non i.i.d data stream, while only observing each sample once. A naive application of auto-encoder in this setting encounters a major challenge: representations derived from earlier encoder states must be usable by later decoder states. We show how to use discrete auto-encoders to effectively address this challenge and introduce Adaptive Quantization Modules (AQM) to control variation in the compression ability of the module at any given stage of learning. This enables selecting an appropriate compression for incoming samples, while taking into account overall memory constraints and current progress of the learned compression. Unlike previous methods, our approach does not require any pretraining, even on challenging datasets. We show that using AQM to replace standard episodic memory in continual learning settings leads to significant gains on continual learning benchmarks with images, LiDAR, and reinforcement learning agents.

Boosted Histogram Transform for Regression

- Yuchao Cai, Hanyuan Hang, Hanfang Yang, Zhouchen Lin
- abstract: In this paper, we propose a boosting algorithm for regression problems called \emph{boosted histogram transform for regression} (BHTR) based on histogram transforms composed of random rotations, stretchings, and translations. From the theoretical perspective, we first prove fast convergence rates for BHTR under the assumption that the target function lies in the spaces $C^{0,\alpha}$. Moreover, if the target function resides in the subspace $C^{1,\alpha}$, by establishing the upper bound of the convergence rate for the boosted regressor, i.e. BHTR, and the lower bound for base regressors, i.e. histogram transform regressors (HTR), we manage to explain the benefits of the boosting procedure. In the experiments, compared with other state-of-the-art algorithms such as gradient boosted regression tree (GBRT), Breiman's forest, and kernel-based methods, our BHTR algorithm shows promising performance on both synthetic and real datasets.

On Validation and Planning of An Optimal Decision Rule with Application in Healthcare Studies

- Hengrui Cai, Wenbin Lu, Rui Song
- abstract: In the current era of personalized recommendation, one major interest is to develop an optimal individualized decision rule that assigns individuals with the best treatment option according to their covariates. Estimation of optimal decision rules (ODR) has been extensively investigated recently, however, at present, no testing procedure is proposed to verify whether these ODRs are significantly better than the naive decision rule that always assigning individuals to a fixed treatment option. In this paper, we propose a testing procedure for detecting the existence of an ODR that is better than the naive decision rule under the randomized trials. We construct the proposed test based on the difference of estimated value functions using the augmented inverse probability weighted method. The asymptotic distributions of the proposed test statistic under the null and local alternative hypotheses are established. Based on the established asymptotic distributions, we further develop a sample size calculation formula for testing the existence of an ODR in designing A/B tests. Extensive simulations and a real data application to a schizophrenia clinical trial data are conducted to demonstrate the empirical validity of the proposed methods.

Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality

- Changxiao Cai, H. Vincent Poor, Yuxin Chen
- abstract: We study the distribution and uncertainty of nonconvex optimization for noisy tensor completion — the problem of estimating a low-rank tensor given incomplete and corrupted observations of its entries. Focusing on a two-stage nonconvex estimation algorithm proposed by (Cai et al., 2019), we characterize the distribution of this estimator down to fine scales. This distributional theory in turn allows one to construct valid and short confidence intervals for both the unseen tensor entries and its underlying tensor factors. The proposed inferential procedure enjoys several important features: (1) it is fully adaptive to noise heteroscedasticity, and (2) it is data-driven and adapts automatically to unknown noise distributions. Furthermore, our findings unveil the statistical optimality of nonconvex tensor completion: it attains un-improvable estimation accuracy — including both the rates and the pre-constants — under i.i.d. Gaussian noise.

Provably Efficient Exploration in Policy Optimization

- Qi Cai, Zhuoran Yang, Chi Jin, Zhaoran Wang
- abstract: While policy-based reinforcement learning (RL) achieves tremendous successes in practice, it is significantly less understood in theory, especially compared with value-based RL. In particular, it remains elusive how to design a provably efficient policy optimization algorithm that incorporates exploration. To bridge such a gap, this paper proposes an Optimistic variant of the Proximal Policy Optimization algorithm (OPPO), which follows an “optimistic version” of the policy gradient direction. This paper proves that, in the problem of episodic Markov decision process with linear function approximation, unknown transition, and adversarial reward with full-information feedback, OPPO achieves $\tilde{O}(\sqrt{d^2 H^3 T})$ regret. Here d is the feature dimension, H is the episode horizon, and T is the total number of steps. To the best of our knowledge, OPPO is the first provably efficient policy optimization algorithm that explores.

Near-linear time Gaussian process optimization with adaptive batching and resparsification

- Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, Lorenzo Rosasco
- abstract: Gaussian processes (GP) are one of the most successful frameworks to model uncertainty. However, GP optimization (e.g., GP-UCB) suffers from major scalability issues. Experimental time grows linearly with the number of evaluations, unless candidates are selected in batches (e.g., using GP-BUCB) and evaluated in parallel. Furthermore, computational cost is often prohibitive since algorithms such as GP-BUCB require a time at least quadratic in the number of dimensions and iterations to select each batch. In this paper, we introduce BBKB (Batch Budgeted Kernel Bandits), the first no-regret GP optimization algorithm that provably runs in near-linear time and selects candidates in batches. This is obtained with a new guarantee for the tracking of the posterior variances that allows BBKB to choose increasingly larger batches, improving over GP-BUCB. Moreover, we show that the same bound can be used to adaptively delay costly updates to the sparse GP approximation used by BBKB, achieving a near-constant per-step amortized cost. These findings are then confirmed in several experiments, where BBKB is much faster than state-of-the-art methods.

Poisson Learning: Graph Based Semi-Supervised Learning At Very Low Label Rates

- Jeff Calder, Brendan Cook, Matthew Thorpe, Dejan Slepcev
- abstract: We propose a new framework, called Poisson learning, for graph based semi-supervised learning at very low label rates. Poisson learning is motivated by the need to address the degeneracy of Laplacian semi-supervised learning in this regime. The method replaces the assignment of label values at training points with the placement of sources and sinks, and solves the resulting Poisson equation on the graph. The outcomes are provably more stable and informative than those of Laplacian learning. Poisson learning is efficient and simple to implement, and we present numerical experiments showing the method is superior to other recent approaches to semi-supervised learning at low label rates on MNIST, FashionMNIST, and Cifar-10. We also propose a graph-cut enhancement of Poisson learning, called Poisson MBO, that gives higher accuracy and can incorporate prior knowledge of relative class sizes.

Explore, Discover and Learn: Unsupervised Discovery of State-Covering Skills

- Victor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giro-I-Nieto, Jordi Torres
- abstract: Acquiring abilities in the absence of a task-oriented reward function is at the frontier of reinforcement learning research. This problem has been studied through the lens of empowerment, which draws a connection between option discovery and information theory. Information-theoretic skill discovery methods have garnered much interest from the community, but little research has been conducted in understanding their limitations. Through theoretical analysis and empirical evidence, we show that existing algorithms suffer from a common limitation – they discover options that provide a poor coverage of the state space. In light of this, we propose Explore, Discover and Learn (EDL), an alternative approach to information-theoretic skill discovery. Crucially, EDL optimizes the same information-theoretic objective derived from the empowerment literature, but addresses the optimization problem using different machinery. We perform an extensive evaluation of skill discovery methods on controlled environments and show that EDL offers significant advantages, such as overcoming the coverage problem, reducing the dependence of learned skills on the initial state, and allowing the user to define a prior over which behaviors should be learned.

Logarithmic Regret for Learning Linear Quadratic Regulators Efficiently

- Asaf Cassel, Alon Cohen, Tomer Koren
- abstract: We consider the problem of learning in Linear Quadratic Control systems whose transition parameters are initially unknown. Recent results in this setting have demonstrated efficient learning algorithms with regret growing with the square root of the number of decision steps. We present new efficient algorithms that achieve, perhaps surprisingly,regret that scales only (poly-)logarithmically with the number of steps, in two scenarios: when only the state transition matrix A is unknown, and when only the state-action transition matrix B is unknown and the optimal policy satisfies a certain non-degeneracy condition. On the other hand, we give a lower bound which shows that when the latter condition is violated, square root regret is unavoidable.

Fully Parallel Hyperparameter Search: Reshaped Space-Filling

- Marie-Liesse Cauwet, Camille Couprie, Julien Dehos, Pauline Luc, Jeremy Rapin, Morgane Riviere, Fabien Teytaud, Olivier Teytaud, Nicolas Usunier
- abstract: Space-filling designs such as Low Discrepancy Sequence (LDS), Latin Hypercube Sampling (LHS) and Jittered Sampling (JS) were proposed for fully parallel hyperparameter search, and were shown to be more effective than random and grid search. We prove that LHS and JS outperform random search only by a constant factor. Consequently, we introduce a new sampling approach based on the reshaping of the search distribution, and we show both theoretically and numerically that it leads to significant gains over random search. Two methods are proposed for the reshaping: Recentering (when the distribution of the optimum is known), and Cauchy transformation (when the distribution of the optimum is unknown). The proposed methods are first validated on artificial experiments and simple real-world tests on clustering and Salmon mappings. Then we demonstrate that they drive performance improvement in a wide range of expensive artificial intelligence tasks, namely attend/infer/repeat, video next frame segmentation forecasting and progressive generative adversarial networks.

Data preprocessing to mitigate bias: A maximum entropy based approach

- L. Elisa Celis, Vijay Keswani, Nisheeth Vishnoi

- abstract: Data containing human or social attributes may over- or under-represent groups with respect to salient social attributes such as gender or race, which can lead to biases in downstream applications. This paper presents an algorithmic framework that can be used as a data preprocessing method towards mitigating such bias. Unlike prior work, it can efficiently learn distributions over large domains, controllably adjust the representation rates of protected groups and achieve target fairness metrics such as statistical parity, yet remains close to the empirical distribution induced by the given dataset. Our approach leverages the principle of maximum entropy {–} amongst all distributions satisfying a given set of constraints, we should choose the one closest in KL-divergence to a given prior. While maximum entropy distributions can succinctly encode distributions over large domains, they can be difficult to compute. Our main contribution is an instantiation of this framework for our set of constraints and priors, which encode our bias mitigation goals, and that runs in time polynomial in the dimension of the data. Empirically, we observe that samples from the learned distribution have desired representation rates and statistical rates, and when used for training a classifier incurs only a slight loss in accuracy while maintaining fairness properties.

Meta-learning with Stochastic Linear Bandits

- Leonardo Cella, Alessandro Lazaric, Massimiliano Pontil
- abstract: We investigate meta-learning procedures in the setting of stochastic linear bandits tasks. The goal is to select a learning algorithm which works well on average over a class of bandits tasks, that are sampled from a task-distribution. Inspired by recent work on learning-to-learn linear regression, we consider a class of bandit algorithms that implement a regularized version of the well-known OFUL algorithm, where the regularization is a square euclidean distance to a bias vector. We first study the benefit of the biased OFUL algorithm in terms of regret minimization. We then propose two strategies to estimate the bias within the learning-to-learn setting. We show both theoretically and experimentally, that when the number of tasks grows and the variance of the task-distribution is small, our strategies have a significant advantage over learning the tasks in isolation.

Description Based Text Classification with Reinforcement Learning

- Duo Chai, Wei Wu, Qinghong Han, Fei Wu, Jiwei Li
- abstract: The task of text classification is usually divided into two stages: text feature extraction and classification. In this standard formalization, categories are merely represented as indexes in the label vocabulary, and the model lacks for explicit instructions on what to classify. Inspired by the current trend of formalizing NLP problems as question answering tasks, we propose a new framework for text classification, in which each category label is associated with a category description. Descriptions are generated by hand-crafted templates or using abstractive/extractive models from reinforcement learning. The concatenation of the description and the text is fed to the classifier to decide whether or not the current label should be assigned to the text. The proposed strategy forces the model to attend to the most salient texts with respect to the label, which can be regarded as a hard version of attention, leading to better performances. We observe significant performance boosts over strong baselines on a wide range of text classification tasks including single-label classification, multi-label classification and multi-aspect sentiment analysis.

Concise Explanations of Neural Networks using Adversarial Training

- Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, Somesh Jha
- abstract: We show new connections between adversarial learning and explainability for deep neural networks (DNNs). One form of explanation of the output of a neural network model in terms of its input features, is a vector of feature-attributions, which can be generated by various techniques such as Integrated Gradients (IG), DeepSHAP, LIME, and CXPlain. Two desirable characteristics of an attribution-based explanation are: (1) \emph{sparseness}: the attributions of irrelevant or weakly relevant features should be negligible, thus resulting in \emph{concise} explanations in terms of the significant features, and (2) \emph{stability}: it should not vary significantly within a small local neighborhood of the input. Our first contribution is a theoretical exploration of how these two properties (when using IG-based attributions) are related to adversarial training, for a class of 1-layer networks (which includes logistic regression models for binary and multi-class classification); for these networks we show that (a) adversarial training using an ℓ_∞-bounded adversary produces models with sparse attribution vectors, and (b) natural model-training while encouraging stable explanations (via an extra term in the loss function), is equivalent to adversarial training. Our second contribution is an empirical verification of phenomenon (a), which we show, somewhat surprisingly, occurs \emph{not only in 1-layer networks, but also DNNs trained on standard image datasets}, and extends beyond IG-based attributions, to those based on DeepSHAP: adversarial training with $\linf$-bounded perturbations yields significantly sparser attribution vectors, with little degradation in performance on natural test data, compared to natural training. Moreover, the sparseness of the attribution vectors is significantly better than that achievable via ℓ_1-regularized natural training.

Unlabelled Data Improves Bayesian Uncertainty Calibration under Covariate Shift

- Alex Chan, Ahmed Alaa, Zhaozhi Qian, Mihaela Van Der Schaar
- abstract: Modern neural networks have proven to be powerful function approximators, providing state-of-the-art performance in a multitude of applications. They however fall short in their ability to quantify confidence in their predictions — this is crucial in high-stakes applications that involve critical decision-making. Bayesian neural networks (BNNs) aim at solving this problem by placing a prior distribution over the network’s parameters, thereby inducing a posterior distribution that encapsulates predictive uncertainty. While existing variants of BNNs based on Monte Carlo dropout produce reliable (albeit approximate) uncertainty estimates over in-distribution data, they tend to exhibit over-confidence in predictions made on target data whose feature distribution differs from the training data, i.e., the covariate shift setup. In this paper, we develop an approximate Bayesian inference scheme based on posterior regularisation, wherein unlabelled target data are used as “pseudo-labels” of model confidence that are used to regularise the model’s loss on labelled source data. We show that this approach significantly improves the accuracy of uncertainty quantification on covariate-shifted data sets, with minimal modification to the underlying model architecture. We demonstrate the utility of our method in the context of transferring prognostic models of prostate cancer across globally diverse populations.

Imputer: Sequence Modelling via Imputation and Dynamic Programming

- William Chan, Chitwan Saharia, Geoffrey Hinton, Mohammad Norouzi, Navdeep Jaitly
- abstract: This paper presents the Imputer, a neural sequence model that generates output sequences iteratively via imputations. The Imputer is an iterative generation model, requiring only a constant number of generation steps independent of the number of input or output tokens. The Imputer can be trained to approximately marginalize over all possible alignments between the input and output sequences, and all possible generation orders. We present a tractable dynamic programming training algorithm, which yields a lower bound on the log marginal likelihood. When applied to end-to-end speech recognition, the Imputer outperforms prior non-autoregressive models and achieves competitive results to autoregressive models. On LibriSpeech test-other, the Imputer achieves 11.1 WER, outperforming CTC at 13.0 WER and seq2seq at 12.5 WER.

Optimizing for the Future in Non-Stationary MDPs

- Yash Chandak, Georgios Theocharous, Shiv Shankar, Martha White, Sridhar Mahadevan, Philip Thomas
- abstract: Most reinforcement learning methods are based upon the key assumption that the transition dynamics and reward functions are fixed, that is, the underlying Markov decision process is stationary. However, in many real-world applications, this assumption is violated, and using existing algorithms may result in a performance lag. To proactively search for a good future policy, we present a policy gradient algorithm that maximizes a forecast of future performance. This forecast is obtained by fitting a curve to the counter-factual estimates of policy performance over time, without explicitly modeling the underlying non-stationarity. The resulting algorithm amounts to a non-uniform reweighting of past data, and we observe that minimizing performance over some of the data from past episodes can be beneficial when searching for a policy that maximizes future performance. We show that our algorithm,

called Prognosticator, is more robust to non-stationarity than two online adaptation techniques, on three simulated problems motivated by real-world applications.

Learning to Simulate and Design for Structural Engineering

- Kai-Hung Chang, Chin-Yi Cheng
- abstract: The structural design process for buildings is time-consuming and laborious. To automate this process, structural engineers combine optimization methods with simulation tools to find an optimal design with minimal building mass subject to building regulations. However, structural engineers in practice often avoid optimization and compromise on a suboptimal design for the majority of buildings, due to the large size of the design space, the iterative nature of the optimization methods, and the slow simulation tools. In this work, we formulate the building structures as graphs and create an end-to-end pipeline that can learn to propose the optimal cross-sections of columns and beams by training together with a pre-trained differentiable structural simulator. The performance of the proposed structural designs is comparable to the ones optimized by genetic algorithm (GA), with all the constraints satisfied. The optimal structural design with the reduced the building mass can not only lower the material cost, but also decrease the carbon footprint.

Decentralized Reinforcement Learning: Global Decision-Making via Local Economic Transactions

- Michael Chang, Sid Kaushik, S. Matthew Weinberg, Tom Griffiths, Sergey Levine
- abstract: This paper seeks to establish a framework for directing a society of simple, specialized, self-interested agents to solve what traditionally are posed as monolithic single-agent sequential decision problems. What makes it challenging to use a decentralized approach to collectively optimize a central objective is the difficulty in characterizing the equilibrium strategy profile of non-cooperative games. To overcome this challenge, we design a mechanism for defining the learning environment of each agent for which we know that the optimal solution for the global objective coincides with a Nash equilibrium strategy profile of the agents optimizing their own local objectives. The society functions as an economy of agents that learn the credit assignment process itself by buying and selling to each other the right to operate on the environment state. We derive a class of decentralized reinforcement learning algorithms that are broadly applicable not only to standard reinforcement learning but also for selecting options in semi-MDPs and dynamically composing computation graphs. Lastly, we demonstrate the potential advantages of a society's inherent modular structure for more efficient transfer learning.

Invariant Rationalization

- Shiyu Chang, Yang Zhang, Mo Yu, Tommi Jaakkola
- abstract: Selective rationalization improves neural network interpretability by identifying a small subset of input features {—} the rationale {—} that best explains or supports the prediction. A typical rationalization criterion, i.e. maximum mutual information (MMI), finds the rationale that maximizes the prediction performance based only on the rationale. However, MMI can be problematic because it picks up spurious correlations between the input features and the output. Instead, we introduce a game-theoretic invariant rationalization criterion where the rationales are constrained to enable the same predictor to be optimal across different environments. We show both theoretically and empirically that the proposed rationales can rule out spurious correlations and generalize better to different test scenarios. The resulting explanations also align better with human judgments. Our implementations are publicly available at https://github.com/code-terminator/invariant_rationalization.

Circuit-Based Intrinsic Methods to Detect Overfitting

- Satrajit Chatterjee, Alan Mishchenko
- abstract: The focus of this paper is on intrinsic methods to detect overfitting. By intrinsic methods, we mean methods that rely only on the model and the training data, as opposed to traditional methods (we call them extrinsic methods) that rely on performance on a test set or on bounds from model complexity. We propose a family of intrinsic methods called Counterfactual Simulation (CFS) which analyze the flow of training examples through the model by identifying and perturbing rare patterns. By applying CFS to logic circuits we get a method that has no hyper-parameters and works uniformly across different types of models such as neural networks, random forests and lookup tables. Experimentally, CFS can separate models with different levels of overfit using only their logic circuit representations without any access to the high level structure. By comparing lookup tables, neural networks, and random forests using CFS, we get insight into why neural networks generalize. In particular, we find that stochastic gradient descent in neural nets does not lead to "brute force" memorization, but finds common patterns (whether we train with actual or randomized labels), and neural networks are not unlike forests in this regard. Finally, we identify a limitation with our proposal that makes it unsuitable in an adversarial setting, but points the way to future work on robust intrinsic methods.

Better depth-width trade-offs for neural networks through the lens of dynamical systems

- Vaggos Chatziafratis, Sai Ganesh Nagarajan, Ioannis Panageas
- abstract: The expressivity of neural networks as a function of their depth, width and type of activation units has been an important question in deep learning theory. Recently, depth separation results for ReLU networks were obtained via a new connection with dynamical systems, using a generalized notion of fixed points of a continuous map f, called periodic points. In this work, we strengthen the connection with dynamical systems and we improve the existing width lower bounds along several aspects. Our first main result is period-specific width lower bounds that hold under the stronger notion of L^1-approximation error, instead of the weaker classification error. Our second contribution is that we provide sharper width lower bounds, still yielding meaningful exponential depth-width separations, in regimes where previous results wouldn't apply. A byproduct of our results is that there exists a universal constant characterizing the depth-width trade-offs, as long as f has odd periods. Technically, our results follow by unveiling a tighter connection between the following three quantities of a given function: its period, its Lipschitz constant and the growth rate of the number of oscillations arising under compositions of the function f with itself.

Explainable and Discourse Topic-aware Neural Language Understanding

- Yatin Chaudhary, Hinrich Schuetze, Pankaj Gupta
- abstract: Marrying topic models and language models exposes language understanding to a broader source of document-level context beyond sentences via topics. While introducing topical semantics in language models, existing approaches incorporate latent document topic proportions and ignore topical discourse in sentences of the document. This work extends the line of research by additionally introducing an explainable topic representation in language understanding, obtained from a set of key terms correspondingly for each latent topic of the proportion. Moreover, we retain sentence-topic association along with document-topic association by modeling topical discourse for every sentence in the document. We present a novel neural composite language modeling (NCLM) framework that exploits both the latent and explainable topics along with topical discourse at sentence-level in a joint learning framework of topic and language models. Experiments over a range of tasks such as language modeling, word sense disambiguation, document classification, retrieval and text generation demonstrate ability of the proposed model in improving language understanding.

Uncertainty-Aware Lookahead Factor Models for Quantitative Investing

- Lakshay Chauhan, John Alberg, Zachary Lipton

- abstract: On a periodic basis, publicly traded companies report fundamentals, financial data including revenue, earnings, debt, among others. Quantitative finance research has identified several factors, functions of the reported data that historically correlate with stock market performance. In this paper, we first show through simulation that if we could select stocks via factors calculated on future fundamentals (via oracle), that our portfolios would far outperform standard factor models. Motivated by this insight, we train deep nets to forecast future fundamentals from a trailing 5-year history. We propose lookahead factor models which plug these predicted future fundamentals into traditional factors. Finally, we incorporate uncertainty estimates from both neural heteroscedastic regression and a dropout-based heuristic, improving performance by adjusting our portfolios to avert risk. In retrospective analysis, we leverage an industry-grade portfolio simulator (backtester) to show simultaneous improvement in annualized return and Sharpe ratio. Specifically, the simulated annualized return for the uncertainty-aware model is 17.7% (vs 14.0% for a standard factor model) and the Sharpe ratio is 0.84 (vs 0.52).

Deep Reasoning Networks for Unsupervised Pattern De-mixing with Constraint Reasoning

- Di Chen, Yiwei Bai, Wenting Zhao, Sebastian Ament, John Gregoire, Carla Gomes
- abstract: We introduce Deep Reasoning Networks (DRNets), an end-to-end framework that combines deep learning with constraint reasoning for solving pattern de-mixing problems, typically in an unsupervised or very-weakly-supervised setting. DRNets exploit problem structure and prior knowledge by tightly combining constraint reasoning with stochastic-gradient-based neural network optimization. Our motivating task is from materials discovery and concerns inferring crystal structures of materials from X-ray diffraction data (Crystal-Structure-Phase-Mapping). Given the complexity of its underlying scientific domain, we start by introducing DRNets on an analogous but much simpler task: de-mixing overlapping hand-written Sudokus (Multi-MNIST-Sudoku). On Multi-MNIST-Sudoku, DRNets almost perfectly recovered the mixed Sudokus' digits, with 100% digit accuracy, outperforming the supervised state-of-the-art MNIST de-mixing models. On Crystal-Structure-Phase-Mapping, DRNets significantly outperform the state of the art and experts' capabilities, recovering more precise and physically meaningful crystal structures.

Self-PU: Self Boosted and Calibrated Positive-Unlabeled Training

- Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, Zhangyang Wang
- abstract: Many real-world applications have to tackle the Positive-Unlabeled (PU) learning problem, i.e., learning binary classifiers from a large amount of unlabeled data and a few labeled positive examples. While current state-of-the-art methods employ importance reweighting to design various biased or unbiased risk estimators, they completely ignored the learning capability of the model itself, which could provide reliable supervision. This motivates us to propose a novel Self-PU learning framework, which seamlessly integrates PU learning and self-training. Self-PU highlights three "self"-oriented building blocks: a self-paced training algorithm that adaptively discovers and augments confident positive/negative examples as the training proceeds; a self-reweighted, instance-aware loss; and a self-distillation scheme that introduces teacher-students learning as an effective regularization for PU learning. We demonstrate the state-of-the-art performance of Self-PU on common PU learning benchmarks (MNIST and CIFAR10), which compare favorably against the latest competitors. Moreover, we study a real-world application of PU learning, i.e., classifying brain images of Alzheimer's Disease. Self-PU obtains significantly improved results on the renowned Alzheimer's Disease Neuroimaging Initiative (ADNI) database over existing methods.

Learning To Stop While Learning To Predict

- Xinshi Chen, Hanjun Dai, Yu Li, Xin Gao, Le Song
- abstract: There is a recent surge of interest in designing deep architectures based on the update steps in traditional algorithms, or learning neural networks to improve and replace traditional algorithms. While traditional algorithms have certain stopping criteria for outputting results at different iterations, many algorithm-inspired deep models are restricted to a "fixed-depth" for all inputs. Similar to algorithms, the optimal depth of a deep architecture may be different for different input instances, either to avoid "over-thinking", or because we want to compute less for operations converged already. In this paper, we tackle this varying depth problem using a steerable architecture, where a feed-forward deep model and a variational stopping policy are learned together to sequentially determine the optimal number of layers for each input instance. Training such architecture is very challenging. We provide a variational Bayes perspective and design a novel and effective training procedure which decomposes the task into an oracle model learning stage and an imitation stage. Experimentally, we show that the learned deep model along with the stopping policy improves the performances on a diverse set of tasks, including learning sparse recovery, few-shot meta learning, and computer vision tasks.

Combinatorial Pure Exploration for Dueling Bandit

- Wei Chen, Yihan Du, Longbo Huang, Haoyu Zhao
- abstract: In this paper, we study combinatorial pure exploration for dueling bandits (CPE-DB): we have multiple candidates for multiple positions as modeled by a bipartite graph, and in each round we sample a duel of two candidates on one position and observe who wins in the duel, with the goal of finding the best candidate-position matching with high probability after multiple rounds of samples. CPE-DB is an adaptation of the original combinatorial pure exploration for multi-armed bandit (CPE-MAB) problem to the dueling bandit setting. We consider both the Borda winner and the Condorcet winner cases. For Borda winner, we establish a reduction of the problem to the original CPE-MAB setting and design PAC and exact algorithms that achieve both the sample complexity similar to that in the CPE-MAB setting (which is nearly optimal for a subclass of problems) and polynomial running time per round. For Condorcet winner, we first design a fully polynomial time approximation scheme (FPTAS) for the offline problem of finding the Condorcet winner with known winning probabilities, and then use the FPTAS as an oracle to design a novel pure exploration algorithm CAR-Cond with sample complexity analysis. CAR-Cond is the first algorithm with polynomial running time per round for identifying the Condorcet winner in CPE-DB.

Graph Optimal Transport for Cross-Domain Alignment

- Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, Jingjing Liu
- abstract: Cross-domain alignment between two sets of entities (e.g., objects in an image, words in a sentence) is fundamental to both computer vision and natural language processing. Existing methods mainly focus on designing advanced attention mechanisms to simulate soft alignment, where no training signals are provided to explicitly encourage alignment. Plus, the learned attention matrices are often dense and difficult to interpret. We propose Graph Optimal Transport (GOT), a principled framework that builds upon recent advances in Optimal Transport (OT). In GOT, cross-domain alignment is formulated as a graph matching problem, by representing entities as a dynamically-constructed graph. Two types of OT distances are considered: (i) Wasserstein distance (WD) for node (entity) matching; and (ii) Gromov-Wasserstein distance (GWD) for edge (structure) matching. Both WD and GWD can be incorporated into existing neural network models, effectively acting as a drop-in regularizer. The inferred transport plan also yields sparse and self-normalized alignment, enhancing the interpretability of the learned model. Experiments show consistent outperformance of GOT over baselines across a wide range of tasks, including image-text retrieval, visual question answering, image captioning, machine translation, and text summarization.

Stabilizing Differentiable Architecture Search via Perturbation-based Regularization

- Xiangning Chen, Cho-Jui Hsieh
- abstract: Differentiable architecture search (DARTS) is a prevailing NAS solution to identify architectures. Based on the continuous relaxation of the architecture space, DARTS learns a differentiable architecture weight and largely reduces the search cost. However, its stability has been challenged for yielding deteriorating architectures as the search proceeds. We find that the precipitous validation loss landscape, which leads to a dramatic performance drop when distilling the final architecture, is an essential factor that causes instability. Based on this observation, we propose a perturbation-based regularization - SmoothDARTS (SDARTS), to smooth the loss landscape and improve the generalizability of DARTS-based methods. In particular, our new formulations stabilize DARTS-based methods by either random smoothing or adversarial attack. The search trajectory on NAS-Bench-1Shot1

demonstrates the effectiveness of our approach and due to the improved stability, we achieve performance gain across various search spaces on 4 datasets. Furthermore, we mathematically show that SDARTS implicitly regularizes the Hessian norm of the validation loss, which accounts for a smoother loss landscape and improved performance.

Mapping natural-language problems to formal-language solutions using structured neural representations

- Kezhen Chen, Qiuyuan Huang, Hamid Palangi, Paul Smolensky, Ken Forbus, Jianfeng Gao
- abstract: Generating formal-language programs represented by relational tuples, such as Lisp programs or mathematical operations, to solve problems stated in natural language is a challenging task because it requires explicitly capturing discrete symbolic structural information implicit in the input. However, most general neural sequence models do not explicitly capture such structural information, limiting their performance on these tasks. In this paper, we propose a new encoder-decoder model based on a structured neural representation, Tensor Product Representations (TPRs), for mapping Natural-language problems to Formal-language solutions, called TP-N2F. The encoder of TP-N2F employs TPR 'binding' to encode natural-language symbolic structure in vector space and the decoder uses TPR 'unbinding' to generate, in symbolic space, a sequential program represented by relational tuples, each consisting of a relation (or operation) and a number of arguments. TP-N2F considerably outperforms LSTM-based seq2seq models on two benchmarks and creates new state-of-the-art results. Ablation studies show that improvements can be attributed to the use of structured TPRs explicitly in both the encoder and decoder. Analysis of the learned structures shows how TPRs enhance the interpretability of TP-N2F.

Convolutional Kernel Networks for Graph-Structured Data

- Dexiong Chen, Laurent Jacob, Julien Mairal
- abstract: We introduce a family of multilayer graph kernels and establish new links between graph convolutional neural networks and kernel methods. Our approach generalizes convolutional kernel networks to graph-structured data, by representing graphs as a sequence of kernel feature maps, where each node carries information about local graph substructures. On the one hand, the kernel point of view offers an unsupervised, expressive, and easy-to-regularize data representation, which is useful when limited samples are available. On the other hand, our model can also be trained end-to-end on large-scale data, leading to new types of graph convolutional neural networks. We show that our method achieves competitive performance on several graph classification benchmarks, while offering simple model interpretation. Our code is freely available at https://github.com/claying/GCKN.

Learning Flat Latent Manifolds with VAEs

- Nutan Chen, Alexej Klushyn, Francesco Ferroni, Justin Bayer, Patrick Van Der Smagt
- abstract: Measuring the similarity between data points often requires domain knowledge, which can in parts be compensated by relying on unsupervised methods such as latent-variable models, where similarity/distance is estimated in a more compact latent space. Prevalent is the use of the Euclidean metric, which has the drawback of ignoring information about similarity of data stored in the decoder, as captured by the framework of Riemannian geometry. We propose an extension to the framework of variational auto-encoders allows learning flat latent manifolds, where the Euclidean metric is a proxy for the similarity between data points. This is achieved by defining the latent space as a Riemannian manifold and by regularising the metric tensor to be a scaled identity matrix. Additionally, we replace the compact prior typically used in variational auto-encoders with a recently presented, more expressive hierarchical one—and formulate the learning problem as a constrained optimisation problem. We evaluate our method on a range of data-sets, including a video-tracking benchmark, where the performance of our unsupervised approach nears that of state-of-the-art supervised approaches, while retaining the computational efficiency of straight-line-based approaches.

A Simple Framework for Contrastive Learning of Visual Representations

- Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton
- abstract: This paper presents SimCLR: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 85.8% top-5 accuracy, outperforming AlexNet with 100X fewer labels.

Retro*: Learning Retrosynthetic Planning with Neural Guided A* Search

- Binghong Chen, Chengtao Li, Hanjun Dai, Le Song
- abstract: Retrosynthetic planning is a critical task in organic chemistry which identifies a series of reactions that can lead to the synthesis of a target product. The vast number of possible chemical transformations makes the size of the search space very big, and retrosynthetic planning is challenging even for experienced chemists. However, existing methods either require expensive return estimation by rollout with high variance, or optimize for search speed rather than the quality. In this paper, we propose Retro*, *a neural-based A*-like algorithm that finds high-quality synthetic routes efficiently. It maintains the search as an AND-OR tree, and learns a neural search bias with off-policy data. Then guided by this neural network, it performs best-first search efficiently during new planning episodes. Experiments on benchmark USPTO datasets show that, our proposed method outperforms existing state-of-the-art with respect to both the success rate and solution quality, while being more efficient at the same time.

Differentiable Product Quantization for End-to-End Embedding Compression

- Ting Chen, Lala Li, Yizhou Sun
- abstract: Embedding layers are commonly used to map discrete symbols into continuous embedding vectors that reflect their semantic meanings. Despite their effectiveness, the number of parameters in an embedding layer increases linearly with the number of symbols and poses a critical challenge on memory and storage constraints. In this work, we propose a generic and end-to-end learnable compression framework termed differentiable product quantization (DPQ). We present two instantiations of DPQ that leverage different approximation techniques to enable differentiability in end-to-end learning. Our method can readily serve as a drop-in alternative for any existing embedding layer. Empirically, DPQ offers significant compression ratios (14-238X) at negligible or no performance cost on 10 datasets across three different language tasks.

On Efficient Constructions of Checkpoints

- Yu Chen, Zhenming Liu, Bin Ren, Xin Jin
- abstract: Efficient construction of checkpoints/snapshots is a critical tool for training and diagnosing deep learning models. In this paper, we propose a lossy compression scheme for checkpoint constructions (called LC-Checkpoint). LC-Checkpoint simultaneously maximizes the compression rate and optimizes the recovery speed, under the assumption that SGD is used to train the model. LC-Checkpoint uses quantization and priority promotion to store the most crucial information for SGD to recover, and then uses a Huffman coding to leverage the non-uniform distribution of the gradient scales. Our

extensive experiments show that LC-Checkpoint achieves a compression rate up to 28{\texttimes} and recovery speedup up to 5.77{\texttimes} over a state-of-the-art algorithm (SCAR).

Angular Visual Hardness

- Beidi Chen, Weiyang Liu, Zhiding Yu, Jan Kautz, Anshumali Shrivastava, Animesh Garg, Animashree Anandkumar
- abstract: Recent convolutional neural networks (CNNs) have led to impressive performance but often suffer from poor calibration. They tend to be overconfident, with the model confidence not always reflecting the underlying true ambiguity and hardness. In this paper, we propose angular visual hardness (AVH), a score given by the normalized angular distance between the sample feature embedding and the target classifier to measure sample hardness. We validate this score with an in-depth and extensive scientific study, and observe that CNN models with the highest accuracy also have the best AVH scores. This agrees with an earlier finding that state-of-art models improve on the classification of harder examples. We observe that the training dynamics of AVH is vastly different compared to the training loss. Specifically, AVH quickly reaches a plateau for all samples even though the training loss keeps improving. This suggests the need for designing better loss functions that can target harder examples more effectively. We also find that AVH has a statistically significant correlation with human visual hardness. Finally, we demonstrate the benefit of AVH to a variety of applications such as self-training for domain adaptation and domain generalization.

Estimating the Error of Randomized Newton Methods: A Bootstrap Approach

- Jessie X.T. Chen, Miles Lopes
- abstract: Randomized Newton methods have recently become the focus of intense research activity in large-scale and distributed optimization. In general, these methods are based on a “computation-accuracy trade-off”, which allows the user to gain scalability in exchange for error in the solution. However, the user does not know how much error is created by the randomized approximation, which can be detrimental in two ways: On one hand, the user may try to assess the unknown error with theoretical worst-case error bounds, but this approach is impractical when the bounds involve unknown constants, and it often leads to excessive computation. On the other hand, the user may select the “sketch size” and stopping criteria in a heuristic manner, but this can lead to unreliable results. Motivated by these difficulties, we show how bootstrapping can be used to directly estimate the unknown error, which prevents excessive computation, and offers more confidence about the quality of a randomized solution.

VFlow: More Expressive Generative Flows with Variational Data Augmentation

- Jianfei Chen, Cheng Lu, Biqi Chenli, Jun Zhu, Tian Tian
- abstract: Generative flows are promising tractable models for density modeling that define probabilistic distributions with invertible transformations. However, tractability imposes architectural constraints on generative flows. In this work, we study a previously overlooked constraint that all the intermediate representations must have the same dimensionality with the data due to invertibility, limiting the width of the network. We propose VFlow to tackle this constraint on dimensionality. VFlow augments the data with extra dimensions and defines a maximum evidence lower bound (ELBO) objective for estimating the distribution of augmented data jointly with the variational data augmentation distribution. Under mild assumptions, we show that the maximum ELBO solution of VFlow is always better than the original maximum likelihood solution. For image density modeling on the CIFAR-10 dataset, VFlow achieves a new state-of-the-art 2.98 bits per dimension.

More Data Can Expand The Generalization Gap Between Adversarially Robust and Standard Models

- Lin Chen, Yifei Min, Mingrui Zhang, Amin Karbasi
- abstract: Despite remarkable success in practice, modern machine learning models have been found to be susceptible to adversarial attacks that make human-imperceptible perturbations to the data, but result in serious and potentially dangerous prediction errors. To address this issue, practitioners often use adversarial training to learn models that are robust against such attacks at the cost of higher generalization error on unperturbed test sets. The conventional wisdom is that more training data should shrink the gap between the generalization error of adversarially-trained models and standard models. However, we study the training of robust classifiers for both Gaussian and Bernoulli models under ℓ_∞ attacks, and we prove that more data may actually increase this gap. Furthermore, our theoretical results identify if and when additional data will finally begin to shrink the gap. Lastly, we experimentally demonstrate that our results also hold for linear regression models, which may indicate that this phenomenon occurs more broadly.

An Accelerated DFO Algorithm for Finite-sum Convex Functions

- Yuwen Chen, Antonio Orvieto, Aurelien Lucchi
- abstract: Derivative-free optimization (DFO) has recently gained a lot of momentum in machine learning, spawning interest in the community to design faster methods for problems where gradients are not accessible. While some attention has been given to the concept of acceleration in the DFO literature, existing stochastic algorithms for objective functions with a finite-sum structure have not been shown theoretically to achieve an accelerated rate of convergence. Algorithms that use acceleration in such a setting are prone to instabilities, making it difficult to reach convergence. In this work, we exploit the finite-sum structure of the objective in order to design a variance-reduced DFO algorithm that provably yields acceleration. We prove rates of convergence for both smooth convex and strongly-convex finite-sum objective functions. Finally, we validate our theoretical results empirically on several tasks and datasets.

Generative Pretraining From Pixels

- Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, Ilya Sutskever
- abstract: Inspired by progress in unsupervised representation learning for natural language, we examine whether similar models can learn useful representations for images. We train a sequence Transformer to auto-regressively predict pixels, without incorporating knowledge of the 2D input structure. Despite training on low-resolution ImageNet without labels, we find that a GPT-2 scale model learns strong image representations as measured by linear probing, fine-tuning, and low-data classification. On CIFAR-10, we achieve 96.3% accuracy with a linear probe, outperforming a supervised Wide ResNet, and 99.0% accuracy with full fine-tuning, matching the top supervised pre-trained models. We are also competitive with self-supervised benchmarks on ImageNet when substituting pixels for a VQVAE encoding, achieving 69.0% top-1 accuracy on a linear probe of our features.

Negative Sampling in Semi-Supervised learning

- John Chen, Vatsal Shah, Anastasios Kyrillidis
- abstract: We introduce Negative Sampling in Semi-Supervised Learning (NS^3L), a simple, fast, easy to tune algorithm for semi-supervised learning (SSL). NS^3L is motivated by the success of negative sampling/contrastive estimation. We demonstrate that adding the NS^3L loss to state-of-the-art SSL algorithms, such as the Virtual Adversarial Training (VAT), significantly improves upon vanilla VAT and its variant, VAT with Entropy Minimization. By adding the NS^3L loss to MixMatch, the current state-of-the-art approach on semi-supervised tasks, we observe significant improvements over vanilla MixMatch. We conduct extensive experiments on the CIFAR10, CIFAR100, SVHN and STL10 benchmark datasets. Finally, we perform an ablation study for NS3L regarding its hyperparameter tuning.

Optimization from Structured Samples for Coverage Functions

- Wei Chen, Xiaoming Sun, Jialin Zhang, Zhijie Zhang
- abstract: We revisit the optimization from samples (OPS) model, which studies the problem of optimizing objective functions directly from the sample data. Previous results showed that we cannot obtain a constant approximation ratio for the maximum coverage problem using polynomially many independent samples of the form ${S_i, f(S_i)}_{i=1}^t$ (Balkanski et al., 2017), even if coverage functions are $(1 - \epsilon)$-PMAC learnable using these samples (Badanidiyuru et al., 2012), which means most of the function values can be approximately learned very well with high probability. In this work, to circumvent the impossibility result of OPS, we propose a stronger model called optimization from structured samples (OPSS) for coverage functions, where the data samples encode the structural information of the functions. We show that under three general assumptions on the sample distributions, we can design efficient OPSS algorithms that achieve a constant approximation for the maximum coverage problem. We further prove a constant lower bound under these assumptions, which is tight when not considering computational efficiency. Moreover, we also show that if we remove any one of the three assumptions, OPSS for the maximum coverage problem has no constant approximation.

Simple and Deep Graph Convolutional Networks

- Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, Yaliang Li
- abstract: Graph convolutional networks (GCNs) are a powerful deep learning approach for graph-structured data. Recently, GCNs and subsequent variants have shown superior performance in various application areas on real-world datasets. Despite their success, most of the current GCN models are shallow, due to the \emph{over-smoothing} problem. In this paper, we study the problem of designing and analyzing deep graph convolutional networks. We propose the GCNII, an extension of the vanilla GCN model with two simple yet effective techniques: \emph{Initial residual} and \emph{Identity mapping}. We provide theoretical and empirical evidence that the two techniques effectively relieves the problem of over-smoothing. Our experiments show that the deep GCNII model outperforms the state-of-the-art methods on various semi- and full-supervised tasks.

On Breaking Deep Generative Model-based Defenses and Beyond

- Yanzhi Chen, Renjie Xie, Zhanxing Zhu
- abstract: Deep neural networks have been proven to be vulnerable to the so-called adversarial attacks. Recently there have been efforts to defend such attacks with deep generative models. These defenses often predict by inverting the deep generative models rather than simple feedforward propagation. Such defenses are difficult to attack due to the obfuscated gradients caused by inversion. In this work, we propose a new white-box attack to break these defenses. The idea is to view the inversion phase as a dynamical system, through which we extract the gradient w.r.t the image by backtracking its trajectory. An amortized strategy is also developed to accelerate the attack. Experiments show that our attack better breaks state-of-the-art defenses (e.g DefenseGAN, ABS) than other attacks (e.g BPDA). Additionally, our empirical results provide insights for understanding the weaknesses of deep generative model defenses.

Automated Synthetic-to-Real Generalization

- Wuyang Chen, Zhiding Yu, Zhangyang Wang, Animashree Anandkumar
- abstract: Models trained on synthetic images often face degraded generalization to real data. As a convention, these models are often initialized with ImageNet pretrained representation. Yet the role of ImageNet knowledge is seldom discussed despite common practices that leverage this knowledge to maintain the generalization ability. An example is the careful hand-tuning of early stopping and layer-wise learning rates, which is shown to improve synthetic-to-real generalization but is also laborious and heuristic. In this work, we explicitly encourage the synthetically trained model to maintain similar representations with the ImageNet pretrained model, and propose a \emph{learning-to-optimize (L2O)} strategy to automate the selection of layer-wise learning rates. We demonstrate that the proposed framework can significantly improve the synthetic-to-real generalization performance without seeing and training on real data, while also benefiting downstream tasks such as domain adaptation. Code is available at: https://github.com/NVlabs/ASG.

(Locally) Differentially Private Combinatorial Semi-Bandits

- Xiaoyu Chen, Kai Zheng, Zixin Zhou, Yunchang Yang, Wei Chen, Liwei Wang
- abstract: In this paper, we study Combinatorial Semi-Bandits (CSB) that is an extension of classic Multi-Armed Bandits (MAB) under Differential Privacy (DP) and stronger Local Differential Privacy (LDP) setting. Since the server receives more information from users in CSB, it usually causes additional dependence on the dimension of data, which is a notorious side-effect for privacy preserving learning. However for CSB under two common smoothness assumptions, we show it is possible to remove this side-effect. In detail, for B_{∞}-bounded smooth CSB under either ε-LDP or ε-DP, we prove the optimal regret bound is $\Theta(\frac{mB^2_{\infty}\ln T } {\Delta\varepsilon^2})$ or $\tilde{\Theta}(\frac{mB^2_{\infty}\ln T} { \Delta\varepsilon})$ respectively, where T is time period, Δ is the gap of rewards and m is the number of base arms, by proposing novel algorithms and matching lower bounds. For B_1-bounded smooth CSB under ε-DP, we also prove the optimal regret bound is $\tilde{\Theta}(\frac{mKB^2_1\ln T} {\Delta\varepsilon})$ with both upper bound and lower bound, where K is the maximum number of feedback in each round. All above results nearly match corresponding non-private optimal rates, which imply there is no additional price for (locally) differentially private CSB in above common settings.

High-dimensional Robust Mean Estimation via Gradient Descent

- Yu Cheng, Ilias Diakonikolas, Rong Ge, Mahdi Soltanolkotabi
- abstract: We study the problem of high-dimensional robust mean estimation in the presence of a constant fraction of adversarial outliers. A recent line of work has provided sophisticated polynomial-time algorithms for this problem with dimension-independent error guarantees for a range of natural distribution families. In this work, we show that a natural non-convex formulation of the problem can be solved directly by gradient descent. Our approach leverages a novel structural lemma, roughly showing that any approximate stationary point of our non-convex objective gives a near-optimal solution to the underlying robust estimation task. Our work establishes an intriguing connection between algorithmic high-dimensional robust statistics and non-convex optimization, which may have broader applications to other robust estimation tasks.

CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information

- Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, Lawrence Carin
- abstract: Mutual information (MI) minimization has gained considerable interests in various machine learning tasks. However, estimating and minimizing MI in high-dimensional spaces remains a challenging problem, especially when only samples, rather than distribution forms, are accessible. Previous works mainly focus on MI lower bound approximation, which is not applicable to MI minimization problems. In this paper, we propose a novel Contrastive Log-ratio Upper Bound (CLUB) of mutual information. We provide a theoretical analysis of the properties of CLUB and its variational approximation. Based on this upper bound, we introduce a MI minimization training scheme and further accelerate it with a negative sampling strategy. Simulation studies on Gaussian distributions show the reliable estimation ability of CLUB. Real-world MI minimization experiments, including domain adaptation and information bottleneck, demonstrate the effectiveness of the proposed method. The code is at https://github.com/Linear95/CLUB.

Learning with Bounded Instance and Label-dependent Label Noise

- Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, Dacheng Tao

- abstract: Instance- and Label-dependent label Noise (ILN) widely exists in real-world datasets but has been rarely studied. In this paper, we focus on Bounded Instance- and Label-dependent label Noise (BILN), a particular case of ILN where the label noise rates—the probabilities that the true labels of examples flip into the corrupted ones—have upper bound less than 1. Specifically, we introduce the concept of distilled examples, i.e. examples whose labels are identical with the labels assigned for them by the Bayes optimal classifier, and prove that under certain conditions classifiers learnt on distilled examples will converge to the Bayes optimal classifier. Inspired by the idea of learning with distilled examples, we then propose a learning algorithm with theoretical guarantees for its robustness to BILN. At last, empirical evaluations on both synthetic and real-world datasets show effectiveness of our algorithm in learning with BILN.

Mutual Transfer Learning for Massive Data

- Ching-Wei Cheng, Xingye Qiao, Guang Cheng
- abstract: In the transfer learning problem, the target and the source data domains are typically known. In this article, we study a new paradigm called mutual transfer learning where among many heterogeneous data domains, every data domain could potentially be the target of interest, and it could also be a useful source to help the learning in other data domains. However, it is important to note that given a target not every data domain can be a successful source; only data sets that are similar enough to be thought as from the same population can be useful sources for each other. Under this mutual learnability assumption, a confidence distribution fusion approach is proposed to recover the mutual learnability relation in the transfer learning regime. Our proposed method achieves the same oracle statistical inferential accuracy as if the true learnability structure were known. It can be implemented in an efficient parallel fashion to deal with large-scale data. Simulated and real examples are analyzed to illustrate the usefulness of the proposed method.

Stochastic Gradient and Langevin Processes

- Xiang Cheng, Dong Yin, Peter Bartlett, Michael Jordan
- abstract: We prove quantitative convergence rates at which discrete Langevin-like processes converge to the invariant distribution of a related stochastic differential equation. We study the setup where the additive noise can be non-Gaussian and state-dependent and the potential function can be non-convex. We show that the key properties of these processes depend on the potential function and the second moment of the additive noise. We apply our theoretical findings to studying the convergence of Stochastic Gradient Descent (SGD) for non-convex problems and corroborate them with experiments using SGD to train deep neural networks on the CIFAR-10 dataset.

Representation Learning via Adversarially-Contrastive Optimal Transport

- Anoop Cherian, Shuchin Aeron
- abstract: In this paper, we study the problem of learning compact (low-dimensional) representations for sequential data that captures its implicit spatio-temporal cues. To maximize extraction of such informative cues from the data, we set the problem within the context of contrastive representation learning and to that end propose a novel objective via optimal transport. Specifically, our formulation seeks a low-dimensional subspace representation of the data that jointly (i) maximizes the distance of the data (embedded in this subspace) from an adversarial data distribution under the optimal transport, a.k.a. the Wasserstein distance, (ii) captures the temporal order, and (iii) minimizes the data distortion. To generate the adversarial distribution, we propose a novel framework connecting Wasserstein GANs with a classifier, allowing a principled mechanism for producing good negative distributions for contrastive learning, which is currently a challenging problem. Our full objective is cast as a subspace learning problem on the Grassmann manifold and solved via Riemannian optimization. To empirically study our formulation, we provide experiments on the task of human action recognition in video sequences. Our results demonstrate competitive performance against challenging baselines.

Convergence Rates of Variational Inference in Sparse Deep Learning

- Badr-Eddine Chérief-Abdellatif
- abstract: Variational inference is becoming more and more popular for approximating intractable posterior distributions in Bayesian statistics and machine learning. Meanwhile, a few recent works have provided theoretical justification and new insights on deep neural networks for estimating smooth functions in usual settings such as nonparametric regression. In this paper, we show that variational inference for sparse deep learning retains precisely the same generalization properties than exact Bayesian inference. In particular, we show that a wise choice of the neural network architecture leads to near-minimax rates of convergence for Hölder smooth functions. Additionally, we show that the model selection framework over the architecture of the network via ELBO maximization does not overfit and adaptively achieves the optimal rate of convergence.

Reinforcement Learning for Non-Stationary Markov Decision Processes: The Blessing of (More) Optimism

- Wang Chi Cheung, David Simchi-Levi, Ruihao Zhu
- abstract: We consider un-discounted reinforcement learning (RL) in Markov decision processes (MDPs) under drifting non-stationarity, \ie, both the reward and state transition distributions are allowed to evolve over time, as long as their respective total variations, quantified by suitable metrics, do not exceed certain \emph{variation budgets}. We first develop the Sliding Window Upper-Confidence bound for Reinforcement Learning with Confidence Widening (\texttt{SWUCRL2-CW}) algorithm, and establish its dynamic regret bound when the variation budgets are known. In addition, we propose the Bandit-over-Reinforcement Learning (\texttt{BORL}) algorithm to adaptively tune the \sw to achieve the same dynamic regret bound, but in a \emph{parameter-free} manner, \ie, without knowing the variation budgets. Notably, learning drifting MDPs via conventional optimistic exploration presents a unique challenge absent in existing (non-stationary) bandit learning settings. We overcome the challenge by a novel confidence widening technique that incorporates additional optimism.

Streaming Coresets for Symmetric Tensor Factorization

- Rachit Chhaya, Jayesh Choudhari, Anirban Dasgupta, Supratim Shit
- abstract: Factorizing tensors has recently become an important optimization module in a number of machine learning pipelines, especially in latent variable models. We show how to do this efficiently in the streaming setting. Given a set of n vectors, each in $\mathbb{R}^d$, we present algorithms to select a sublinear number of these vectors as coreset, while guaranteeing that the CP decomposition of the p-moment tensor of the coreset approximates the corresponding decomposition of the p-moment tensor computed from the full data. We introduce two novel algorithmic techniques: online filtering and kernelization. Using these two, we present four algorithms that achieve different tradeoffs of coreset size, update time and working space, beating or matching various state of the art algorithms. In the case of matrices (2-ordered tensor), our online row sampling algorithm guarantees $(1 \pm \epsilon)$ relative error spectral approximation. We show applications of our algorithms in learning single topic modeling.

On Coresets for Regularized Regression

- Rachit Chhaya, Anirban Dasgupta, Supratim Shit
- abstract: We study the effect of norm based regularization on the size of coresets for regression problems. Specifically, given a matrix $ \mathbf{A} \in {\mathbb{R}}^{n \times d}$ with $n\gg d$ and a vector $\mathbf{b} \in \mathbb{R} ^ n $ and $\lambda > 0$, we analyze the size of coresets for regularized versions of regression of the form $\|\mathbf{Ax}-\mathbf{b}\|_p^r + \lambda\|{\mathbf{x}}\|_q^s$. Prior work has shown that for ridge regression (where $p,q,r,s=2$) we can obtain a coreset that is smaller than the coreset for the unregularized counterpart i.e. least squares

regression \cite{avron2017sharper}. We show that when $r \neq s$, no coreset for regularized regression can have size smaller than the optimal coreset of the unregularized version. The well known lasso problem falls under this category and hence does not allow a coreset smaller than the one for least squares regression. We propose a modified version of the lasso problem and obtain for it a coreset of size smaller than the least square regression. We empirically show that the modified version of lasso also induces sparsity in solution, similar to the original lasso. We also obtain smaller coresets for ℓ_p regression with ℓ_p regularization. We extend our methods to multi response regularized regression. Finally, we empirically demonstrate the coreset performance for the modified lasso and the ℓ_1 regression with ℓ_1 regularization.

How to Solve Fair k-Center in Massive Data Models

- Ashish Chiplunkar, Sagar Kale, Sivaramakrishnan Natarajan Ramamoorthy
- abstract: Fueled by massive data, important decision making is being automated with the help of algorithms, therefore, fairness in algorithms has become an especially important research topic. In this work, we design new streaming and distributed algorithms for the fair k-center problem that models fair data summarization. The streaming and distributed models of computation have an attractive feature of being able to handle massive data sets that do not fit into main memory. Our main contributions are: (a) the first distributed algorithm; which has provably constant approximation ratio and is extremely parallelizable, and (b) a two-pass streaming algorithm with a provable approximation guarantee matching the best known algorithm (which is not a streaming algorithm). Our algorithms have the advantages of being easy to implement in practice, being fast with linear running times, having very small working memory and communication, and outperforming existing algorithms on several real and synthetic data sets. To complement our distributed algorithm, we also give a hardness result for natural distributed algorithms, which holds for even the special case of k-center.

Fair Generative Modeling via Weak Supervision

- Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, Stefano Ermon
- abstract: Real-world datasets are often biased with respect to key demographic factors such as race and gender. Due to the latent nature of the underlying factors, detecting and mitigating bias is especially challenging for unsupervised machine learning. We present a weakly supervised algorithm for overcoming dataset bias for deep generative models. Our approach requires access to an additional small, unlabeled reference dataset as the supervision signal, thus sidestepping the need for explicit labels on the underlying bias factors. Using this supplementary dataset, we detect the bias in existing datasets via a density ratio technique and learn generative models which efficiently achieve the twin goals of: 1) data efficiency by using training examples from both biased and reference datasets for learning; and 2) data generation close in distribution to the reference dataset at test time. Empirically, we demonstrate the efficacy of our approach which reduces bias w.r.t. latent factors by an average of up to 34.6% over baselines for comparable image generation using generative adversarial networks.

Encoding Musical Style with Transformer Autoencoders

- Kristy Choi, Curtis Hawthorne, Ian Simon, Monica Dinculescu, Jesse Engel
- abstract: We consider the problem of learning high-level controls over the global structure of generated sequences, particularly in the context of symbolic music generation with complex language models. In this work, we present the Transformer autoencoder, which aggregates encodings of the input data across time to obtain a global representation of style from a given performance. We show it is possible to combine this global representation with other temporally distributed embeddings, enabling improved control over the separate aspects of performance style and melody. Empirically, we demonstrate the effectiveness of our method on various music generation tasks on the MAESTRO dataset and a YouTube dataset with 10,000+ hours of piano performances, where we achieve improvements in terms of log-likelihood and mean listening scores as compared to baselines.

k-means++: few more steps yield constant approximation

- Davin Choo, Christoph Grunau, Julian Portmann, Vaclav Rozhon
- abstract: The k-means++ algorithm of Arthur and Vassilvitskii (SODA 2007) is a state-of-the-art algorithm for solving the k-means clustering problem and is known to give an O(log k) approximation. Recently, Lattanzi and Sohler (ICML 2019) proposed augmenting k-means++ with O(k log log k) local search steps to yield a constant approximation (in expectation) to the k-means clustering problem. In this paper, we improve their analysis to show that, for any arbitrarily small constant epsilon > 0, with only epsilon * k additional local search steps, one can achieve a constant approximation guarantee (with high probability in k), resolving an open problem in their paper.

Stochastic Flows and Geometric Optimization on the Orthogonal Group

- Krzysztof Choromanski, David Cheikhi, Jared Davis, Valerii Likhosherstov, Achille Nazaret, Achraf Bahamou, Xingyou Song, Mrugank Akarte, Jack Parker-Holder, Jacob Bergquist, Yuan Gao, Aldo Pacchiano, Tamas Sarlos, Adrian Weller, Vikas Sindhwani
- abstract: We present a new class of stochastic, geometrically-driven optimization algorithms on the orthogonal group O(d) and naturally reductive homogeneous manifolds obtained from the action of the rotation group SO(d). We theoretically and experimentally demonstrate that our methods can be applied in various fields of machine learning including deep, convolutional and recurrent neural networks, reinforcement learning, normalizing flows and metric learning. We show an intriguing connection between efficient stochastic optimization on the orthogonal group and graph theory (e.g. matching problem, partition functions over graphs, graph-coloring). We leverage the theory of Lie groups and provide theoretical results for the designed class of algorithms. We demonstrate broad applicability of our methods by showing strong performance on the seemingly unrelated tasks of learning world models to obtain stable policies for the most difficult Humanoid agent from OpenAI Gym and improving convolutional neural networks.

Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels

- Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, Masashi Sugiyama
- abstract: In weakly supervised learning, unbiased risk estimator(URE) is a powerful tool for training classifiers when training and test data are drawn from different distributions. Nevertheless, UREs lead to overfitting in many problem settings when the models are complex like deep networks. In this paper, we investigate reasons for such overfitting by studying a weakly supervised problem called learning with complementary labels. We argue the quality of gradient estimation matters more in risk minimization. Theoretically, we show that a URE gives an unbiased gradient estimator(UGE). Practically, however, UGEs may suffer from huge variance, which causes empirical gradients to be usually far away from true gradients during minimization. To this end, we propose a novel surrogate complementary loss(SCL) framework that trades zero bias with reduced variance and makes empirical gradients more aligned with true gradients in the direction. Thanks to this characteristic, SCL successfully mitigates the overfitting issue and improves URE-based methods.

Data-Dependent Differentially Private Parameter Learning for Directed Graphical Models

- Amrita Roy Chowdhury, Theodoros Rekatsinas, Somesh Jha
- abstract: Directed graphical models (DGMs) are a class of probabilistic models that are widely used for predictive analysis in sensitive domains such as medical diagnostics. In this paper, we present an algorithm for differentially-private learning of the parameters of a DGM. Our solution optimizes for the utility of inference queries over the DGM and \emph{adds noise that is customized to the properties of the private input dataset and the graph structure of the DGM}. To the best of our knowledge, this is the first explicit data-dependent privacy budget allocation algorithm in the context of DGMs. We

compare our algorithm with a standard data-independent approach over a diverse suite of benchmarks and demonstrate that our solution requires a privacy budget that is roughly $3\times$ smaller to obtain the same or higher utility.

Online Continual Learning from Imbalanced Data

- Aristotelis Chrysakis, Marie-Francine Moens
- abstract: A well-documented weakness of neural networks is the fact that they suffer from catastrophic forgetting when trained on data provided by a non-stationary distribution. Recent work in the field of continual learning attempts to understand and overcome this issue. Unfortunately, the majority of relevant work embraces the implicit assumption that the distribution of observed data is perfectly balanced, despite the fact that, in the real world, humans and animals learn from observations that are temporally correlated and severely imbalanced. Motivated by this remark, we aim to evaluate memory population methods that are used in online continual learning, when dealing with highly imbalanced and temporally correlated streams of data. More importantly, we introduce a new memory population approach, which we call class-balancing reservoir sampling (CBRS). We demonstrate that CBRS outperforms the state-of-the-art memory population algorithms in a considerably challenging learning setting, over a range of different datasets, and for multiple architectures.

Distance Metric Learning with Joint Representation Diversification

- Xu Chu, Yang Lin, Yasha Wang, Xiting Wang, Hailong Yu, Xin Gao, Qi Tong
- abstract: Distance metric learning (DML) is to learn a representation space equipped with a metric, such that similar examples are closer than dissimilar examples concerning the metric. The recent success of DNNs motivates many DML losses that encourage the intra-class compactness and inter-class separability. The trade-off between inter-class compactness and inter-class separability shapes the DML representation space by determining how much information of the original inputs to retain. In this paper, we propose a Distance Metric Learning with Joint Representation Diversification (JRD) that allows a better balancing point between intra-class compactness and inter-class separability. Specifically, we propose a Joint Representation Similarity regularizer that captures different abstract levels of invariant features and diversifies the joint distributions of representations across multiple layers. Experiments on three deep DML benchmark datasets demonstrate the effectiveness of the proposed approach.

Semismooth Newton Algorithm for Efficient Projections onto $\ell_1, ∞$-norm Ball

- Dejun Chu, Changshui Zhang, Shiliang Sun, Qing Tao
- abstract: The structured sparsity-inducing $\ell_{1, \infty}$-norm, as a generalization of the classical ℓ_1-norm, plays an important role in jointly sparse models which select or remove simultaneously all the variables forming a group. However, its resulting problem is more difficult to solve than the conventional ℓ_1-norm constrained problem. In this paper, we propose an efficient algorithm for Euclidean projection onto $\ell_{1, \infty}$-norm ball. We tackle the projection problem via semismooth Newton algorithm to solve the system of semismooth equations. Meanwhile, exploiting the structure of the Jacobian matrix via LU decomposition yields an equivalent algorithm which is proved to terminate after a finite number of iterations. Empirical studies demonstrate that our proposed algorithm outperforms the existing state-of-the-art solver and is promising for the optimization of learning problems with the $\ell_{1, \infty}$-norm ball constraint.

Estimating Generalization under Distribution Shifts via Domain-Invariant Representations

- Ching-Yao Chuang, Antonio Torralba, Stefanie Jegelka
- abstract: When machine learning models are deployed on a test distribution different from the training distribution, they can perform poorly, but overestimate their performance. In this work, we aim to better estimate a model’s performance under distribution shift, without supervision. To do so, we use a set of domain-invariant predictors as a proxy for the unknown, true target labels. Since the error of the resulting risk estimate depends on the target risk of the proxy model, we study generalization of domain-invariant representations and show that the complexity of the latent representation has a significant influence on the target risk. Empirically, our approach (1) enables self-tuning of domain adaptation models, and (2) accurately estimates the target error of given models under distribution shift. Other applications include model selection, deciding early stopping and error detection.

Scalable and Efficient Comparison-based Search without Features

- Daniyar Chumbalov, Lucas Maystre, Matthias Grossglauser
- abstract: We consider the problem of finding a target object t using pairwise comparisons, by asking an oracle questions of the form “Which object from the pair (i,j) is more similar to t?”. Objects live in a space of latent features, from which the oracle generates noisy answers. First, we consider the non-blind setting where these features are accessible. We propose a new Bayesian comparison-based search algorithm with noisy answers; it has low computational complexity yet is efficient in the number of queries. We provide theoretical guarantees, deriving the form of the optimal query and proving almost sure convergence to the target t. Second, we consider the blind setting, where the object features are hidden from the search algorithm. In this setting, we combine our search method and a new distributional triplet embedding algorithm into one scalable learning framework called Learn2Search. We show that the query complexity of our approach on two real-world datasets is on par with the non-blind setting, which is not achievable using any of the current state-of-the-art embedding methods. Finally, we demonstrate the efficacy of our framework by conducting a movie actors search experiment with real users.

Feature-map-level Online Adversarial Knowledge Distillation

- Inseop Chung, Seonguk Park, Jangho Kim, Nojun Kwak
- abstract: Feature maps contain rich information about image intensity and spatial correlation. However, previous online knowledge distillation methods only utilize the class probabilities. Thus in this paper, we propose an online knowledge distillation method that transfers not only the knowledge of the class probabilities but also that of the feature map using the adversarial training framework. We train multiple networks simultaneously by employing discriminators to distinguish the feature map distributions of different networks. Each network has its corresponding discriminator which discriminates the feature map from its own as fake while classifying that of the other network as real. By training a network to fool the corresponding discriminator, it can learn the other network’s feature map distribution. We show that our method performs better than the conventional direct alignment method such as L1 and is more suitable for online distillation. Also, we propose a novel cyclic learning scheme for training more than two networks together. We have applied our method to various network architectures on the classification task and discovered a significant improvement of performance especially in the case of training a pair of a small network and a large one.

Teaching with Limited Information on the Learner’s Behaviour

- Ferdinando Cicalese, Sergio Filho, Eduardo Laber, Marco Molinaro
- abstract: Machine Teaching studies how efficiently a Teacher can guide a Learner to a target hypothesis. We focus on the model of Machine Teaching with a black box learner introduced in [Dasgupta et al., ICML 2019], where the teaching is done interactively without having any knowledge of the Learner’s algorithm and class of hypotheses, apart from the fact that it contains the target hypothesis $h^$. *We first refine some existing results for this model and, then, we study new variants of it. Motivated by the realistic possibility that $h^$* is not available to the learner, we consider the case where the teacher can only aim at having the learner converge to a best available approximation of h^*. We also consider weaker black box learners, where, in each round, the

choice of the consistent hypothesis returned to the Teacher is not adversarial, and in particular, we show that better provable bounds can be obtained for a type of Learner that moves to the next hypothesis smoothly, preferring hypotheses that are close to the current one; and for another type of Learner that can provide to the Teacher hypotheses chosen at random among those consistent with the examples received so far. Finally, we present an empirical evaluation of our basic interactive teacher on real datasets.

Deep Divergence Learning

- Hatice Kubra Cilingir, Rachel Manzelli, Brian Kulis
- abstract: Classical linear metric learning methods have recently been extended along two distinct lines: deep metric learning methods for learning embeddings of the data using neural networks, and Bregman divergence learning approaches for extending learning Euclidean distances to more general divergence measures such as divergences over distributions. In this paper, we introduce deep Bregman divergences, which are based on learning and parameterizing functional Bregman divergences using neural networks, and which unify and extend these existing lines of work. We show in particular how deep metric learning formulations, kernel metric learning, Mahalanobis metric learning, and moment-matching functions for comparing distributions arise as special cases of these divergences in the symmetric setting. We then describe a deep learning framework for learning general functional Bregman divergences, and show in experiments that this method yields superior performance on benchmark datasets as compared to existing deep metric learning approaches. We also discuss novel applications, including a semi-supervised distributional clustering problem, and a new loss function for unsupervised data generation.

Model Fusion with Kullback-Leibler Divergence

- Sebastian Claici, Mikhail Yurochkin, Soumya Ghosh, Justin Solomon
- abstract: We propose a method to fuse posterior distributions learned from heterogeneous datasets. Our algorithm relies on a mean field assumption for both the fused model and the individual dataset posteriors and proceeds using a simple assign-and-average approach. The components of the dataset posteriors are assigned to the proposed global model components by solving a regularized variant of the assignment problem. The global components are then updated based on these assignments by their mean under a KL divergence. For exponential family variational distributions, our formulation leads to an efficient non-parametric algorithm for computing the fused model. Our algorithm is easy to describe and implement, efficient, and competitive with state-of-the-art on motion capture analysis, topic modeling, and federated learning of Bayesian neural networks.

Leveraging Procedural Generation to Benchmark Reinforcement Learning

- Karl Cobbe, Chris Hesse, Jacob Hilton, John Schulman
- abstract: We introduce Procgen Benchmark, a suite of 16 procedurally generated game-like environments designed to benchmark both sample efficiency and generalization in reinforcement learning. We believe that the community will benefit from increased access to high quality training environments, and we provide detailed experimental protocols for using this benchmark. We empirically demonstrate that diverse environment distributions are essential to adequately train and evaluate RL agents, thereby motivating the extensive use of procedural content generation. We then use this benchmark to investigate the effects of scaling model size, finding that larger models significantly improve both sample efficiency and generalization.

Composable Sketches for Functions of Frequencies: Beyond the Worst Case

- Edith Cohen, Ofir Geri, Rasmus Pagh
- abstract: Recently there has been increased interest in using machine learning techniques to improve classical algorithms. In this paper we study when it is possible to construct compact, composable sketches for weighted sampling and statistics estimation according to functions of data frequencies. Such structures are now central components of large-scale data analytics and machine learning pipelines. However, many common functions, such as thresholds and pth frequency moments with $p>2$, are known to require polynomial size sketches in the worst case. We explore performance beyond the worst case under two different types of assumptions. The first is having access to noisy \emph{advice} on item frequencies. This continues the line of work of Hsu et al. (ICLR 2019), who assume predictions are provided by a machine learning model. The second is providing guaranteed performance on a restricted class of input frequency distributions that are better aligned with what is observed in practice. This extends the work on heavy hitters under Zipfian distributions in a seminal paper of Charikar et al. (ICALP 2002). Surprisingly, we show analytically and empirically that "in practice" small polylogarithmic-size sketches provide accuracy for "hard" functions.

Healing Products of Gaussian Process Experts

- Samuel Cohen, Rendani Mbuvha, Tshilidzi Marwala, Marc Deisenroth
- abstract: Gaussian processes (GPs) are nonparametric Bayesian models that have been applied to regression and classification problems. One of the approaches to alleviate their cubic training cost is the use of local GP experts trained on subsets of the data. In particular, product-of-expert models combine the predictive distributions of local experts through a tractable product operation. While these expert models allow for massively distributed computation, their predictions typically suffer from erratic behaviour of the mean or uncalibrated uncertainty quantification. By calibrating predictions via a tempered softmax weighting, we provide a solution to these problems for multiple product-of-expert models, including the generalised product of experts and the robust Bayesian committee machine. Furthermore, we leverage the optimal transport literature and propose a new product-of-expert model that combines predictions of local experts by computing their Wasserstein barycenter, which can be applied to both regression and classification.

On Efficient Low Distortion Ultrametric Embedding

- Vincent Cohen-Addad, Karthik C. S., Guillaume Lagarde
- abstract: A classic problem in unsupervised learning and data analysis is to find simpler and easy-to-visualize representations of the data that preserve its essential properties. A widely-used method to preserve the underlying hierarchical structure of the data while reducing its complexity is to find an embedding of the data into a tree or an ultrametric, but computing such an embedding on a data set of n points in $\Omega(\log n)$ dimensions incurs a quite prohibitive running time of $\Theta(n^2)$. In this paper, we provide a new algorithm which takes as input a set of points P in $\R^d$, and for every $c\ge 1$, runs in time $n^{1+\frac{\rho}{c^2}}$ (for some universal constant $\rho>1$) to output an ultrametric Δ such that for any two points u,v in P, we have $\Delta(u,v)$ is within a multiplicative factor of $5c$ to the distance between u and v in the best ultrametric representation of P. Here, the best ultrametric is the ultrametric $\tilde\Delta$ that minimizes the maximum distance distortion with respect to the ℓ_2 distance, namely that minimizes $\underset{u,v \in P}{\max} \nicefrac{\tilde\Delta(u,v)}{\|u-v\|_2}$. *We complement the above result by showing that under popular complexity theoretic assumptions, for every constant $\varepsilon>0$, no algorithm with running time $n^{2-\varepsilon}$ can distinguish between inputs in $\ell*\infty$-metric that admit isometric embedding and those that incur a distortion of $\nicefrac{3}{2}$. Finally, we present empirical evaluation on classic machine learning datasets and show that the output of our algorithm is comparable to the output of the linkage algorithms while achieving a much faster running time.

Sub-linear Memory Sketches for Near Neighbor Search on Streaming Data

- Benjamin Coleman, Richard Baraniuk, Anshumali Shrivastava

- abstract: We present the first sublinear memory sketch that can be queried to find the nearest neighbors in a dataset. Our online sketching algorithm compresses an N element dataset to a sketch of size $O(N^b \log^3 N)$ in $O(N^{(b+1)} \log^3 N)$ time, where $b < 1$. This sketch can correctly report the nearest neighbors of any query that satisfies a stability condition parameterized by b. We achieve sublinear memory performance on stable queries by combining recent advances in locality sensitive hash (LSH)-based estimators, online kernel density estimation, and compressed sensing. Our theoretical results shed new light on the memory-accuracy tradeoff for nearest neighbor search, and our sketch, which consists entirely of short integer arrays, has a variety of attractive features in practice. We evaluate the memory-recall tradeoff of our method on a friend recommendation task in the Google plus social media network. We obtain orders of magnitude better compression than the random projection based alternative while retaining the ability to report the nearest neighbors of practical queries.

Word-Level Speech Recognition With a Letter to Word Encoder

- Ronan Collobert, Awni Hannun, Gabriel Synnaeve
- abstract: We propose a direct-to-word sequence model which uses a word network to learn word embeddings from letters. The word network can be integrated seamlessly with arbitrary sequence models including Connectionist Temporal Classification and encoder-decoder models with attention. We show our direct-to-word model can achieve word error rate gains over sub-word level models for speech recognition. We also show that our direct-to-word approach retains the ability to predict words not seen at training time without any retraining. Finally, we demonstrate that a word-level model can use a larger stride than a sub-word level model while maintaining accuracy. This makes the model more efficient both for training and inference.

Boosting Frank-Wolfe by Chasing Gradients

- Cyrille Combettes, Sebastian Pokutta
- abstract: The Frank-Wolfe algorithm has become a popular first-order optimization algorithm for it is simple and projection-free, and it has been successfully applied to a variety of real-world problems. Its main drawback however lies in its convergence rate, which can be excessively slow due to naive descent directions. We propose to speed up the Frank-Wolfe algorithm by better aligning the descent direction with that of the negative gradient via a subroutine. This subroutine chases the negative gradient direction in a matching pursuit-style while still preserving the projection-free property. Although the approach is reasonably natural, it produces very significant results. We derive convergence rates $\mathcal{O}(1/t)$ to $\mathcal{O}(e^{-\omega t})$ of our method and we demonstrate its competitive advantage both per iteration and in CPU time over the state-of-the-art in a series of computational experiments.

Learning Opinions in Social Networks

- Vincent Conitzer, Debmalya Panigrahi, Hanrui Zhang
- abstract: We study the problem of learning opinions in social networks. The learner observes the states of some sample nodes from a social network, and tries to infer the states of other nodes, based on the structure of the network. We show that sample-efficient learning is impossible when the network exhibits strong noise, and give a polynomial-time algorithm for the problem with nearly optimal sample complexity when the network is sufficiently stable.

Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows

- Rob Cornish, Anthony Caterini, George Deligiannidis, Arnaud Doucet
- abstract: We show that normalising flows become pathological when used to model targets whose supports have complicated topologies. In this scenario, we prove that a flow must become arbitrarily numerically noninvertible in order to approximate the target closely. This result has implications for all flow-based models, and especially residual flows (ResFlows), which explicitly control the Lipschitz constant of the bijection used. To address this, we propose continuously indexed flows (CIFs), which replace the single bijection used by normalising flows with a continuously indexed family of bijections, and which can intuitively "clean up" mass that would otherwise be misplaced by a single bijection. We show theoretically that CIFs are not subject to the same topological limitations as normalising flows, and obtain better empirical performance on a variety of models and benchmarks.

Adaptive Region-Based Active Learning

- Corinna Cortes, Giulia Desalvo, Claudio Gentile, Mehryar Mohri, Ningshan Zhang
- abstract: We present a new active learning algorithm that adaptively partitions the input space into a finite number of regions, and subsequently seeks a distinct predictor for each region, while actively requesting labels. We prove theoretical guarantees for both the generalization error and the label complexity of our algorithm, and analyze the number of regions defined by the algorithm under some mild assumptions. We also report the results of an extensive suite of experiments on several real-world datasets demonstrating substantial empirical benefits over existing single-region and non-adaptive region-based active learning baselines.

Online Learning with Dependent Stochastic Feedback Graphs

- Corinna Cortes, Giulia Desalvo, Claudio Gentile, Mehryar Mohri, Ningshan Zhang
- abstract: A general framework for online learning with partial information is one where feedback graphs specify which losses can be observed by the learner. We study a challenging scenario where feedback graphs vary stochastically with time and, more importantly, where graphs and losses are dependent. This scenario appears in several real-world applications that we describe where the outcome of actions are correlated. We devise a new algorithm for this setting that exploits the stochastic properties of the graphs and that benefits from favorable regret guarantees. We present a detailed theoretical analysis of this algorithm, and also report the result of a series of experiments on real-world datasets, which show that our algorithm outperforms standard baselines for online learning with feedback graphs.

Learnable Group Transform For Time-Series

- Romain Cosentino, Behnaam Aazhang
- abstract: We propose a novel approach to filter bank learning for time-series by considering spectral decompositions of signals defined as a Group Transform. This framework allows us to generalize classical time-frequency transformations such as the Wavelet Transform, and to efficiently learn the representation of signals. While the creation of the wavelet transform filter-bank relies on affine transformations of a mother filter, our approach allows for non-linear transformations. The transformations induced by such maps enable us to span a larger class of signal representations, from wavelet to chirplet-like filters. We propose a parameterization of such a non-linear map such that its sampling can be optimized for a specific task and signal. The Learnable Group Transform can be cast into a Deep Neural Network. The experiments on diverse time-series datasets demonstrate the expressivity of this framework, which competes with state-of-the-art performances.

DINO: Distributed Newton-Type Optimization Method

- Rixon Crane, Fred Roosta

- abstract: We present a novel communication-efficient Newton-type algorithm for finite-sum optimization over a distributed computing environment. Our method, named DINO, overcomes both theoretical and practical shortcomings of similar existing methods. Under minimal assumptions, we guarantee global sub-linear convergence of DINO to a first-order stationary point for general non-convex functions and arbitrary data distribution over the network. Furthermore, for functions satisfying Polyak-Lojasiewicz (PL) inequality, we show that DINO enjoys a linear convergence rate. Our proposed algorithm is practically parameter free, in that it will converge regardless of the selected hyper-parameters, which are easy to tune. Additionally, its sub-problems are simple linear least-squares, for which efficient solvers exist, and numerical simulations demonstrate the efficiency of DINO as compared with similar alternatives.

Causal Modeling for Fairness In Dynamical Systems

- Elliot Creager, David Madras, Toniann Pitassi, Richard Zemel
- abstract: In many applications areas—lending, education, and online recommenders, for example—fairness and equity concerns emerge when a machine learning system interacts with a dynamically changing environment to produce both immediate and long-term effects for individuals and demographic groups. We discuss causal directed acyclic graphs (DAGs) as a unifying framework for the recent literature on fairness in such dynamical systems. We show that this formulation affords several new directions of inquiry to the modeler, where sound causal assumptions can be expressed and manipulated. We emphasize the importance of computing interventional quantities in the dynamical fairness setting, and show how causal assumptions enable simulation (when environment dynamics are known) and estimation by adjustment (when dynamics are unknown) of intervention on short- and long-term outcomes, at both the group and individual levels.

Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack

- Francesco Croce, Matthias Hein
- abstract: The evaluation of robustness against adversarial manipulation of neural networks-based classifiers is mainly tested with empirical attacks as methods for the exact computation, even when available, do not scale to large networks. We propose in this paper a new white-box adversarial attack wrt the l_p-norms for $p \in {1,2,\infty}$ aiming at finding the minimal perturbation necessary to change the class of a given input. It has an intuitive geometric meaning, yields quickly high quality results, minimizes the size of the perturbation (so that it returns the robust accuracy at every threshold with a single run). It performs better or similar to state-of-the-art attacks which are partially specialized to one l_p-norm, and is robust to the phenomenon of gradient obfuscation.

Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks

- Francesco Croce, Matthias Hein
- abstract: The field of defense strategies against adversarial attacks has significantly grown over the last years, but progress is hampered as the evaluation of adversarial defenses is often insufficient and thus gives a wrong impression of robustness. Many promising defenses could be broken later on, making it difficult to identify the state-of-the-art. Frequent pitfalls in the evaluation are improper tuning of hyperparameters of the attacks, gradient obfuscation or masking. In this paper we first propose two extensions of the PGD-attack overcoming failures due to suboptimal step size and problems of the objective function. We then combine our novel attacks with two complementary existing ones to form a parameter-free, computationally affordable and user-independent ensemble of attacks to test adversarial robustness. We apply our ensemble to over 50 models from papers published at recent top machine learning and computer vision venues. In all except one of the cases we achieve lower robust test accuracy than reported in these papers, often by more than $10%$, identifying several broken defenses.

Real-Time Optimisation for Online Learning in Auctions

- Lorenzo Croissant, Marc Abeille, Clement Calauzenes
- abstract: In display advertising, a small group of sellers and bidders face each other in up to 10^{12} auctions a day. In this context, revenue maximisation via monopoly price learning is a high-value problem for sellers. By nature, these auctions are online and produce a very high frequency stream of data. This results in a computational strain that requires algorithms be real-time. Unfortunately, existing methods inherited from the batch setting suffer $O(\sqrt{t})$ time/memory complexity at each update, prohibiting their use. In this paper, we provide the first algorithm for online learning of monopoly prices in online auctions whose update is constant in time and memory.

Privately detecting changes in unknown distributions

- Rachel Cummings, Sara Krehbiel, Yuliia Lut, Wanrong Zhang
- abstract: The change-point detection problem seeks to identify distributional changes in streams of data. Increasingly, tools for change-point detection are applied in settings where data may be highly sensitive and formal privacy guarantees are required, such as identifying disease outbreaks based on hospital records, or IoT devices detecting activity within a home. Differential privacy has emerged as a powerful technique for enabling data analysis while preventing information leakage about individuals. Much of the prior work on change-point detection{—}including the only private algorithms for this problem{—}requires complete knowledge of the pre-change and post-change distributions, which is an unrealistic assumption for many practical applications of interest. This work develops differentially private algorithms for solving the change-point detection problem when the data distributions are unknown. Additionally, the data may be sampled from distributions that change smoothly over time, rather than fixed pre-change and post-change distributions. We apply our algorithms to detect changes in the linear trends of such data streams. Finally, we also provide experimental results to empirically validate the performance of our algorithms.

Flexible and Efficient Long-Range Planning Through Curious Exploration

- Aidan Curtis, Minjian Xin, Dilip Arumugam, Kevin Feigelis, Daniel Yamins
- abstract: Identifying algorithms that flexibly and efficiently discover temporally-extended multi-phase plans is an essential step for the advancement of robotics and model-based reinforcement learning. The core problem of long-range planning is finding an efficient way to search through the tree of possible action sequences. Existing non-learned planning solutions from the Task and Motion Planning (TAMP) literature rely on the existence of logical descriptions for the effects and preconditions for actions. This constraint allows TAMP methods to efficiently reduce the tree search problem but limits their ability to generalize to unseen and complex physical environments. In contrast, deep reinforcement learning (DRL) methods use flexible neural-network-based function approximators to discover policies that generalize naturally to unseen circumstances. However, DRL methods struggle to handle the very sparse reward landscapes inherent to long-range multi-step planning situations. Here, we propose the Curious Sample Planner (CSP), which fuses elements of TAMP and DRL by combining a curiosity-guided sampling strategy with imitation learning to accelerate planning. We show that CSP can efficiently discover interesting and complex temporally-extended plans for solving a wide range of physically realistic 3D tasks. In contrast, standard planning and learning methods often fail to solve these tasks at all or do so only with a huge and highly variable number of training samples. We explore the use of a variety of curiosity metrics with CSP and analyze the types of solutions that CSP discovers. Finally, we show that CSP supports task transfer so that the exploration policies learned during experience with one task can help improve efficiency on related tasks.

Parameter-free, Dynamic, and Strongly-Adaptive Online Learning

- Ashok Cutkosky
- abstract: We provide a new online learning algorithm that for the first time combines several disparate notions of adaptivity. First, our algorithm obtains a “parameter-free” regret bound that adapts to the norm of the comparator and the squared norm of the size of the gradients it observes. Second, it obtains a “strongly-adaptive” regret bound, so that for any given interval of length N, the regret over the interval is $\tilde O(\sqrt{N})$. Finally, our algorithm obtains an optimal “dynamic” regret bound: for any sequence of comparators with path-length P, our algorithm obtains regret $\tilde O(\sqrt{PN})$ over intervals of length N. Our primary technique for achieving these goals is a new method of combining constrained online learning regret bounds that does not rely on an expert meta-algorithm to aggregate learners.

Momentum Improves Normalized SGD

- Ashok Cutkosky, Harsh Mehta
- abstract: We provide an improved analysis of normalized SGD showing that adding momentum provably removes the need for large batch sizes on non-convex objectives. Then, we consider the case of objectives with bounded second derivative and show that in this case a small tweak to the momentum formula allows normalized SGD with momentum to find an ϵ-critical point in $O(1/\epsilon^{3.5})$ iterations, matching the best-known rates without accruing any logarithmic factors or dependence on dimension. We provide an adaptive learning rate schedule that automatically improves convergence rates when the variance in the gradients is small. Finally, we show that our method is effective when employed on popular large scale tasks such as ResNet-50 and BERT pretraining, matching the performance of the disparate methods used to get state-of-the-art results on both tasks.

Supervised Quantile Normalization for Low Rank Matrix Factorization

- Marco Cuturi, Olivier Teboul, Jonathan Niles-Weed, Jean-Philippe Vert
- abstract: Low rank matrix factorization is a fundamental building block in machine learning, used for instance to summarize gene expression profile data or word-document counts. To be robust to outliers and differences in scale across features, a matrix factorization step is usually preceded by ad-hoc feature normalization steps, such as tf-idf scaling or data whitening. We propose in this work to learn these normalization operators jointly with the factorization itself. More precisely, given a $d\times n$ matrix X of d features measured on n individuals, we propose to learn the parameters of quantile normalization operators that can operate row-wise on the values of X and/or of its factorization UV to improve the quality of the low-rank representation of X itself. This optimization is facilitated by the introduction of a new differentiable quantile normalization operator built using optimal transport, providing new results on top of existing work by Cuturi et al. (2019). We demonstrate the applicability of these techniques on synthetic and genomics datasets.

Double Trouble in Double Descent: Bias and Variance(s) in the Lazy Regime

- Stéphane D’Ascoli, Maria Refinetti, Giulio Biroli, Florent Krzakala
- abstract: Deep neural networks can achieve remarkable generalization performances while interpolating the training data. Rather than the U-curve emblematic of the bias-variance trade-off, their test error often follows a "double descent"—a mark of the beneficial role of overparametrization. In this work, we develop a quantitative theory for this phenomenon in the so-called lazy learning regime of neural networks, by considering the problem of learning a high-dimensional function with random features regression. We obtain a precise asymptotic expression for the bias-variance decomposition of the test error, and show that the bias displays a phase transition at the interpolation threshold, beyond it which it remains constant. We disentangle the variances stemming from the sampling of the dataset, from the additive noise corrupting the labels, and from the initialization of the weights. We demonstrate that the latter two contributions are the crux of the double descent: they lead to the overfitting peak at the interpolation threshold and to the decay of the test error upon overparametrization. We quantify how they are suppressed by ensembling the outputs of K independently initialized estimators. For $K\rightarrow \infty$, the test error is monotonously decreasing and remains constant beyond the interpolation threshold. We further compare the effects of overparametrizing, ensembling and regularizing. Finally, we present numerical experiments on classic deep learning setups to show that our results hold qualitatively in realistic lazy learning scenarios.

R2-B2: Recursive Reasoning-Based Bayesian Optimization for No-Regret Learning in Games

- Zhongxiang Dai, Yizhou Chen, Bryan Kian Hsiang Low, Patrick Jaillet, Teck-Hua Ho
- abstract: This paper presents a recursive reasoning formalism of Bayesian optimization (BO) to model the reasoning process in the interactions between boundedly rational, self-interested agents with unknown, complex, and costly-to-evaluate payoff functions in repeated games, which we call Recursive Reasoning-Based BO (R2-B2). Our R2-B2 algorithm is general in that it does not constrain the relationship among the payoff functions of different agents and can thus be applied to various types of games such as constant-sum, general-sum, and common-payoff games. We prove that by reasoning at level 2 or more and at one level higher than the other agents, our R2-B2 agent can achieve faster asymptotic convergence to no regret than that without utilizing recursive reasoning. We also propose a computationally cheaper variant of R2-B2 called R2-B2-Lite at the expense of a weaker convergence guarantee. The performance and generality of our R2-B2 algorithm are empirically demonstrated using synthetic games, adversarial machine learning, and multi-agent reinforcement learning.

Scalable Deep Generative Modeling for Sparse Graphs

- Hanjun Dai, Azade Nazi, Yujia Li, Bo Dai, Dale Schuurmans
- abstract: Learning graph generative models is a challenging task for deep learning and has wide applicability to a range of domains like chemistry, biology and social science. However current deep neural methods suffer from limited scalability: for a graph with n nodes and m edges, existing deep neural methods require Omega(n^2) complexity by building up the adjacency matrix. On the other hand, many real world graphs are actually sparse in the sense that m << n^2. Based on this, we develop a novel autoregressive model, named BiGG, that utilizes this sparsity to avoid generating the full adjacency matrix, and importantly reduces the graph generation time complexity to O((n + m) log n). Furthermore, during training this autoregressive model can be parallelized with O(log n) synchronization stages, which makes it much more efficient than other autoregressive models that require Omega(n). Experiments on several benchmarks show that the proposed approach not only scales to orders of magnitude larger graphs than previously possible with deep autoregressive graph generative models, but also yields better graph generation quality.

The Usual Suspects? Reassessing Blame for VAE Posterior Collapse

- Bin Dai, Ziyu Wang, David Wipf
- abstract: In narrow asymptotic settings Gaussian VAE models of continuous data have been shown to possess global optima aligned with ground-truth distributions. Even so, it is well known that poor solutions whereby the latent posterior collapses to an uninformative prior are sometimes obtained in practice. However, contrary to conventional wisdom that largely assigns blame for this phenomena on the undue influence of KL-divergence regularization, we will argue that posterior collapse is, at least in part, a direct consequence of bad local minima inherent to the loss surface of deep autoencoder networks. In particular, we prove that even small nonlinear perturbations of affine VAE decoder models can produce such minima, and in deeper models, analogous minima can force the VAE to behave like an aggressive truncation operator, provably discarding information along all latent dimensions in certain circumstances. Regardless, the underlying message here is not meant to undercut valuable existing explanations of posterior collapse, but rather, to refine the discussion and elucidate alternative risk factors that may have been previously underappreciated.

Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting

- Niccolo Dalmasso, Rafael Izbicki, Ann Lee
- abstract: Parameter estimation, statistical tests and confidence sets are the cornerstones of classical statistics that allow scientists to make inferences about the underlying process that generated the observed data. A key question is whether one can still construct hypothesis tests and confidence sets with proper coverage and high power in a so-called likelihood-free inference (LFI) setting; that is, a setting where the likelihood is not explicitly known but one can forward-simulate observable data according to a stochastic model. In this paper, we present ACORE (Approximate Computation via Odds Ratio Estimation), a frequentist approach to LFI that first formulates the classical likelihood ratio test (LRT) as a parametrized classification problem, and then uses the equivalence of tests and confidence sets to build confidence regions for parameters of interest. We also present a goodness-of-fit procedure for checking whether the constructed tests and confidence regions are valid. ACORE is based on the key observation that the LRT statistic, the rejection probability of the test, and the coverage of the confidence set are conditional distribution functions which often vary smoothly as a function of the parameters of interest. Hence, instead of relying solely on samples simulated at fixed parameter settings (as is the convention in standard Monte Carlo solutions), one can leverage machine learning tools and data simulated in the neighborhood of a parameter to improve estimates of quantities of interest. We demonstrate the efficacy of ACORE with both theoretical and empirical results. Our implementation is available on Github.

Goodness-of-Fit Tests for Inhomogeneous Random Graphs

- Soham Dan, Bhaswar B. Bhattacharya
- abstract: Hypothesis testing of random networks is an emerging area of modern research, especially in the high-dimensional regime, where the number of samples is smaller or comparable to the size of the graph. In this paper we consider the goodness-of-fit testing problem for large inhomogeneous random (IER) graphs, where given a (known) reference symmetric matrix $Q \in [0, 1]^{n \times n}$ and m independent samples from an IER graph given by an unknown symmetric matrix $P \in [0, 1]^{n \times n}$, the goal is to test the hypothesis $P=Q$ versus $||P-Q|| \geq \varepsilon$, where $||\cdot||$ is some specified norm on symmetric matrices. Building on recent related work on two-sample testing for IER graphs, we derive the optimal minimax sample complexities for the goodness-of-fit problem in various natural norms, such as the Frobenius norm and the operator norm. We also propose practical implementations of natural test statistics, using their asymptotic distributions and through the parametric bootstrap. We compare the performances of the different tests in simulations, and show that the proposed tests outperform the baseline tests across various natural random graphs models.

Sharp Statistical Guaratees for Adversarially Robust Gaussian Classification

- Chen Dan, Yuting Wei, Pradeep Ravikumar
- abstract: Adversarial robustness has become a fundamental requirement in modern machine learning applications. Yet, there has been surprisingly little statistical understanding so far. In this paper, we provide the first result of the \emph{optimal} minimax guarantees for the excess risk for adversarially robust classification, under Gaussian mixture model proposed by \cite{schmidt2018adversarially}. The results are stated in terms of the \emph{Adversarial Signal-to-Noise Ratio (AdvSNR)}, which generalizes a similar notion for standard linear classification to the adversarial setting. For the Gaussian mixtures with AdvSNR value of r, we prove an excess risk lower bound of order $\Theta(e^{-(\frac{1}{2}+o(1)) r^2} \frac{d}{n})$ and design a computationally efficient estimator that achieves this optimal rate. Our results built upon minimal assumptions while cover a wide spectrum of adversarial perturbations including ℓ_p balls for any $p \ge 1$.

Adversarial Attacks on Probabilistic Autoregressive Forecasting Models

- Raphaël Dang-Nhu, Gagandeep Singh, Pavol Bielik, Martin Vechev
- abstract: We develop an effective generation of adversarial attacks on neural models that output a sequence of probability distributions rather than a sequence of single values. This setting includes the recently proposed deep probabilistic autoregressive forecasting models that estimate the probability distribution of a time series given its past and achieve state-of-the-art results in a diverse set of application domains. The key technical challenge we address is how to effectively differentiate through the Monte-Carlo estimation of statistics of the output sequence joint distribution. Additionally, we extend prior work on probabilistic forecasting to the Bayesian setting which allows conditioning on future observations, instead of only on past observations. We demonstrate that our approach can successfully generate attacks with small input perturbations in two challenging tasks where robust decision making is crucial – stock market trading and prediction of electricity consumption.

Subspace Fitting Meets Regression: The Effects of Supervision and Orthonormality Constraints on Double Descent of Generalization Errors

- Yehuda Dar, Paul Mayer, Lorenzo Luzi, Richard Baraniuk
- abstract: We study the linear subspace fitting problem in the overparameterized setting, where the estimated subspace can perfectly interpolate the training examples. Our scope includes the least-squares solutions to subspace fitting tasks with varying levels of supervision in the training data (i.e., the proportion of input-output examples of the desired low-dimensional mapping) and orthonormality of the vectors defining the learned operator. This flexible family of problems connects standard, unsupervised subspace fitting that enforces strict orthonormality with a corresponding regression task that is fully supervised and does not constrain the linear operator structure. This class of problems is defined over a supervision-orthonormality plane, where each coordinate induces a problem instance with a unique pair of supervision level and softness of orthonormality constraints. We explore this plane and show that the generalization errors of the corresponding subspace fitting problems follow double descent trends as the settings become more supervised and less orthonormally constrained.

Probing Emergent Semantics in Predictive Agents via Question Answering

- Abhishek Das, Federico Carnevale, Hamza Merzic, Laura Rimell, Rosalia Schneider, Josh Abramson, Alden Hung, Arun Ahuja, Stephen Clark, Greg Wayne, Felix Hill
- abstract: Recent work has shown how predictive modeling can endow agents with rich knowledge of their surroundings, improving their ability to act in complex environments. We propose question-answering as a general paradigm to decode and understand the representations that such agents develop, applying our method to two recent approaches to predictive modelling - action-conditional CPC (Guo et al., 2018) and SimCore (Gregor et al., 2019). After training agents with these predictive objectives in a visually-rich, 3D environment with an assortment of objects, colors, shapes, and spatial configurations, we probe their internal state representations with a host of synthetic (English) questions, without backpropagating gradients from the question-answering decoder into the agent. The performance of different agents when probed in this way reveals that they learn to encode factual, and seemingly compositional, information about objects, properties and spatial relations from their physical environment. Our approach is intuitive, i.e. humans can easily interpret the responses of the model as opposed to inspecting continuous vectors, and model-agnostic, i.e. applicable to any modeling approach. By revealing the implicit knowledge of objects, quantities, properties and relations acquired by agents as they learn, question-conditional agent probing can stimulate the design and development of stronger predictive learning objectives.

Low-Variance and Zero-Variance Baselines for Extensive-Form Games

- Trevor Davis, Martin Schmid, Michael Bowling

- abstract: Extensive-form games (EFGs) are a common model of multi-agent interactions with imperfect information. State-of-the-art algorithms for solving these games typically perform full walks of the game tree that can prove prohibitively slow in large games. Alternatively, sampling-based methods such as Monte Carlo Counterfactual Regret Minimization walk one or more trajectories through the tree, touching only a fraction of the nodes on each iteration, at the expense of requiring more iterations to converge due to the variance of sampled values. In this paper, we extend recent work that uses baseline estimates to reduce this variance. We introduce a framework of baseline-corrected values in EFGs that generalizes the previous work. Within our framework, we propose new baseline functions that result in significantly reduced variance compared to existing techniques. We show that one particular choice of such a function — predictive baseline — is provably optimal under certain sampling schemes. This allows for efficient computation of zero-variance value estimates even along sampled trajectories.

Combining Differentiable PDE Solvers and Graph Neural Networks for Fluid Flow Prediction

- Filipe De Avila Belbute-Peres, Thomas Economon, Zico Kolter
- abstract: Solving large complex partial differential equations (PDEs), such as those that arise in computational fluid dynamics (CFD), is a computationally expensive process. This has motivated the use of deep learning approaches to approximate the PDE solutions, yet the simulation results predicted from these approaches typically do not generalize well to truly novel scenarios. In this work, we develop a hybrid (graph) neural network that combines a traditional graph convolutional network with an embedded differentiable fluid dynamics simulator inside the network itself. By combining an actual CFD simulator (run on a much coarser resolution representation of the problem) with the graph network, we show that we can both generalize well to new situations and benefit from the substantial speedup of neural network CFD predictions, while also substantially outperforming the coarse CFD simulation alone.

Representing Unordered Data Using Complex-Weighted Multiset Automata

- Justin DeBenedetto, David Chiang
- abstract: Unordered, variable-sized inputs arise in many settings across multiple fields. The ability for set- and multiset-oriented neural networks to handle this type of input has been the focus of much work in recent years. We propose to represent multisets using complex-weighted multiset automata and show how the multiset representations of certain existing neural architectures can be viewed as special cases of ours. Namely, (1) we provide a new theoretical and intuitive justification for the Transformer model's representation of positions using sinusoidal functions, and (2) we extend the DeepSets model to use complex numbers, enabling it to outperform the existing model on an extension of one of their tasks.

An end-to-end Differentially Private Latent Dirichlet Allocation Using a Spectral Algorithm

- Chris Decarolis, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, Furong Huang
- abstract: We provide an end-to-end differentially private spectral algorithm for learning LDA, based on matrix/tensor decompositions, and establish theoretical guarantees on utility/consistency of the estimated model parameters. We represent the spectral algorithm as a computational graph. Noise can be injected along the edges of this graph to obtain differential privacy. We identify subsets of edges, named "configurations", such that adding noise to all edges in such a subset guarantees differential privacy of the end-to-end spectral algorithm. We characterize the sensitivity of the edges with respect to the input and thus estimate the amount of noise to be added to each edge for any required privacy level. We then characterize the utility loss for each configuration as a function of injected noise. Overall, by combining the sensitivity and utility characterization, we obtain an end-to-end differentially private spectral algorithm for LDA and identify which configurations outperform others under specific regimes. We are the first to achieve utility guarantees under a required level of differential privacy for learning in LDA. We additionally show that our method systematically outperforms differentially private variational inference.

Gamification of Pure Exploration for Linear Bandits

- Rémy Degenne, Pierre Menard, Xuedong Shang, Michal Valko
- abstract: We investigate an active \emph{pure-exploration} setting, that includes \emph{best-arm identification}, in the context of \emph{linear stochastic bandits}. While asymptotically optimal algorithms exist for standard \emph{multi-armed bandits}, the existence of such algorithms for the best-arm identification in linear bandits has been elusive despite several attempts to address it. First, we provide a thorough comparison and new insight over different notions of optimality in the linear case, including G-optimality, transductive optimality from optimal experimental design and asymptotic optimality. Second, we design the first asymptotically optimal algorithm for fixed-confidence pure exploration in linear bandits. As a consequence, our algorithm naturally bypasses the pitfall caused by a simple but difficult instance, that most prior algorithms had to be engineered to deal with explicitly. Finally, we avoid the need to fully solve an optimal design problem by providing an approach that entails an efficient implementation.

Structure Adaptive Algorithms for Stochastic Bandits

- Rémy Degenne, Han Shao, Wouter Koolen
- abstract: We study reward maximisation in a wide class of structured stochastic multi-armed bandit problems, where the mean rewards of arms satisfy some given structural constraints, e.g. linear, unimodal, sparse, etc. Our aim is to develop methods that are \emph{flexible} (in that they easily adapt to different structures), \emph{powerful} (in that they perform well empirically and/or provably match instance-dependent lower bounds) and \emph{efficient} in that the per-round computational burden is small. We develop asymptotically optimal algorithms from instance-dependent lower-bounds using iterative saddle-point solvers. Our approach generalises recent iterative methods for pure exploration to reward maximisation, where a major challenge arises from the estimation of the sub-optimality gaps and their reciprocals. Still we manage to achieve all the above desiderata. Notably, our technique avoids the computational cost of the full-blown saddle point oracle employed by previous work, while at the same time enabling finite-time regret bounds. Our experiments reveal that our method successfully leverages the structural assumptions, while its regret is at worst comparable to that of vanilla UCB.

Randomly Projected Additive Gaussian Processes for Regression

- Ian Delbridge, David Bindel, Andrew Gordon Wilson
- abstract: Gaussian processes (GPs) provide flexible distributions over functions, with inductive biases controlled by a kernel. However, in many applications Gaussian processes can struggle with even moderate input dimensionality. Learning a low dimensional projection can help alleviate this curse of dimensionality, but introduces many trainable hyperparameters, which can be cumbersome, especially in the small data regime. We use additive sums of kernels for GP regression, where each kernel operates on a different random projection of its inputs. Surprisingly, we find that as the number of random projections increases, the predictive performance of this approach quickly converges to the performance of a kernel operating on the original full dimensional inputs, over a wide range of data sets, even if we are projecting into a single dimension. As a consequence, many problems can remarkably be reduced to one dimensional input spaces, without learning a transformation. We prove this convergence and its rate, and additionally propose a deterministic approach that converges more quickly than purely random projections. Moreover, we demonstrate our approach can achieve faster inference and improved predictive accuracy for high-dimensional inputs compared to kernels in the original input space.

Interpreting Robust Optimization via Adversarial Influence Functions

- Zhun Deng, Cynthia Dwork, Jialiang Wang, Linjun Zhang
- abstract: Robust optimization has been widely used in nowadays data science, especially in adversarial training. However, little research has been done to quantify how robust optimization changes the optimizers and the prediction losses comparing to standard training. In this paper, inspired by the influence function in robust statistics, we introduce the Adversarial Influence Function (AIF) as a tool to investigate the solution produced by robust optimization. The proposed AIF enjoys a closed-form and can be calculated efficiently. To illustrate the usage of AIF, we apply it to study model sensitivity — a quantity defined to capture the change of prediction losses on the natural data after implementing robust optimization. We use AIF to analyze how model complexity and randomized smoothing affect the model sensitivity with respect to specific models. We further derive AIF for kernel regressions, with a particular application to neural tangent kernels, and experimentally demonstrate the effectiveness of the proposed AIF. Lastly, the theories of AIF will be extended to distributional robust optimization.

Non-convex Learning via Replica Exchange Stochastic Gradient MCMC

- Wei Deng, Qi Feng, Liyao Gao, Faming Liang, Guang Lin
- abstract: Replica exchange Monte Carlo (reMC), also known as parallel tempering, is an important technique for accelerating the convergence of the conventional Markov Chain Monte Carlo (MCMC) algorithms. However, such a method requires the evaluation of the energy function based on the full dataset and is not scalable to big data. The naïve implementation of reMC in mini-batch settings introduces large biases, which cannot be directly extended to the stochastic gradient MCMC (SGMCMC), the standard sampling method for simulating from deep neural networks (DNNs). In this paper, we propose an adaptive replica exchange SGMCMC (reSGMCMC) to automatically correct the bias and study the corresponding properties. The analysis implies an acceleration-accuracy trade-off in the numerical discretization of a Markov jump process in a stochastic environment. Empirically, we test the algorithm through extensive experiments on various setups and obtain the state-of-the-art results on CIFAR10, CIFAR100, and SVHN in both supervised learning and semi-supervised learning tasks.

Towards Understanding the Dynamics of the First-Order Adversaries

- Zhun Deng, Hangfeng He, Jiaoyang Huang, Weijie Su
- abstract: An acknowledged weakness of neural networks is their vulnerability to adversarial perturbations to the inputs. To improve the robustness of these models, one of the most popular defense mechanisms is to alternatively maximize the loss over the constrained perturbations (or called adversaries) on the inputs using projected gradient ascent and minimize over weights. In this paper, we analyze the dynamics of the maximization step towards understanding the experimentally observed effectiveness of this defense mechanism. Specifically, we investigate the non-concave landscape of the adversaries for a two-layer neural network with a quadratic loss. Our main result proves that projected gradient ascent finds a local maximum of this non-concave problem in a polynomial number of iterations with high probability. To our knowledge, this is the first work that provides a convergence analysis of the first-order adversaries. Moreover, our analysis demonstrates that, in the initial phase of adversarial training, the scale of the inputs matters in the sense that a smaller input scale leads to faster convergence of adversarial training and a “more regular” landscape. Finally, we show that these theoretical findings are in excellent agreement with a series of experiments.

Robust Pricing in Dynamic Mechanism Design

- Yuan Deng, Sebastien Lahaie, Vahab Mirrokni
- abstract: Motivated by the repeated sale of online ads via auctions, optimal pricing in repeated auctions has attracted a large body of research. While dynamic mechanisms offer powerful techniques to improve on both revenue and efficiency by optimizing auctions across different items, their reliance on exact distributional information of buyers’ valuations (present and future) limits their use in practice. In this paper, we propose robust dynamic mechanism design. We develop a new framework to design dynamic mechanisms that are robust to both estimation errors in value distributions and strategic behavior. We apply the framework in learning environments, leading to the first policy that achieves provably low regret against the optimal dynamic mechanism in contextual auctions, where the dynamic benchmark has full and accurate distributional information.

A Swiss Army Knife for Minimax Optimal Transport

- Sofien Dhouib, Ievgen Redko, Tanguy Kerdoncuff, Rémi Emonet, Marc Sebban
- abstract: The Optimal transport (OT) problem and its associated Wasserstein distance have recently become a topic of great interest in the machine learning community. However, the underlying optimization problem is known to have two major restrictions: (i) it largely depends on the choice of the cost function and (ii) its sample complexity scales exponentially with the dimension. In this paper, we propose a general formulation of a minimax OT problem that can tackle these restrictions by jointly optimizing the cost matrix and the transport plan, allowing us to define a robust distance between distributions. We propose to use a cutting-set method to solve this general problem and show its links and advantages compared to other existing minimax OT approaches. Additionally, we use this method to define a notion of stability allowing us to select the most robust cost matrix. Finally, we provide an experimental study highlighting the efficiency of our approach.

Margin-aware Adversarial Domain Adaptation with Optimal Transport

- Sofien Dhouib, Ievgen Redko, Carole Lartizien
- abstract: In this paper, we propose a new theoretical analysis of unsupervised domain adaptation that relates notions of large margin separation, adversarial learning and optimal transport. This analysis generalizes previous work on the subject by providing a bound on the target margin violation rate, thus reflecting a better control of the quality of separation between classes in the target domain than bounding the misclassification rate. The bound also highlights the benefit of a large margin separation on the source domain for adaptation and introduces an optimal transport (OT) based distance between domains that has the virtue of being task-dependent, contrary to other approaches. From the obtained theoretical results, we derive a novel algorithmic solution for domain adaptation that introduces a novel shallow OT-based adversarial approach and outperforms other OT-based DA baselines on several simulated and real-world classification tasks.

Enhancing Simple Models by Exploiting What They Already Know

- Amit Dhurandhar, Karthikeyan Shanmugam, Ronny Luss
- abstract: There has been recent interest in improving performance of simple models for multiple reasons such as interpretability, robust learning from small data, deployment in memory constrained settings as well as environmental considerations. In this paper, we propose a novel method SRatio that can utilize information from high performing complex models (viz. deep neural networks, boosted trees, random forests) to reweight a training dataset for a potentially low performing simple model of much lower complexity such as a decision tree or a shallow network enhancing its performance. Our method also leverages the per sample hardness estimate of the simple model which is not the case with the prior works which primarily consider the complex model’s confidences/predictions and is thus conceptually novel. Moreover, we generalize and formalize the concept of attaching probes to intermediate layers of a neural network to other commonly used classifiers and incorporate this into our method. The benefit of these contributions is witnessed in the experiments where on 6 UCI datasets and CIFAR-10 we outperform competitors in a majority (16 out of 27) of the cases and tie for best performance in the remaining cases. In fact, in a couple of cases, we even approach the complex model’s performance. We also conduct further experiments to validate assertions and intuitively understand why our method works. Theoretically, we motivate our approach by showing that the weighted loss minimized by simple models using our weighting upper bounds the loss of the complex model.

Spectral Frank-Wolfe Algorithm: Strict Complementarity and Linear Convergence

- Lijun Ding, Yingjie Fei, Qiantong Xu, Chengrun Yang
- abstract: We develop a novel variant of the classical Frank-Wolfe algorithm, which we call spectral Frank-Wolfe, for convex optimization over a spectrahedron. The spectral Frank-Wolfe algorithm has a novel ingredient: it computes a few eigenvectors of the gradient and solves a small-scale subproblem in each iteration. Such a procedure overcomes the slow convergence of the classical Frank-Wolfe algorithm due to ignoring eigenvalue coalescence. We demonstrate that strict complementarity of the optimization problem is key to proving linear convergence of various algorithms, such as the spectral Frank-Wolfe algorithm as well as the projected gradient method and its accelerated version. We showcase that the strict complementarity is equivalent to the eigengap assumption on the gradient at the optimal solution considered in the literature. As a byproduct of this observation, we also develop a generalized block Frank-Wolfe algorithm and prove its linear convergence.

Generalization Guarantees for Sparse Kernel Approximation with Entropic Optimal Features

- Liang Ding, Rui Tuo, Shahin Shahrampour
- abstract: Despite their success, kernel methods suffer from a massive computational cost in practice. In this paper, in lieu of commonly used kernel expansion with respect to N inputs, we develop a novel optimal design maximizing the entropy among kernel features. This procedure results in a kernel expansion with respect to entropic optimal features (EOF), improving the data representation dramatically due to features dissimilarity. Under mild technical assumptions, our generalization bound shows that with only $O(N^{\frac{1}{4}})$ features (disregarding logarithmic factors), we can achieve the optimal statistical accuracy (i.e., $O(1/\sqrt{N})$). The salient feature of our design is its sparsity that significantly reduces the time and space costs. Our numerical experiments on benchmark datasets verify the superiority of EOF over the state-of-the-art in kernel approximation.

Layered Sampling for Robust Optimization Problems

- Hu Ding, Zixiu Wang
- abstract: In real world, our datasets often contain outliers. Most existing algorithms for handling outliers take high time complexities (\emph{e.g.} quadratic or cubic complexity). \emph{Coreset} is a popular approach for compressing data so as to speed up the optimization algorithms. However, the current coreset methods cannot be easily extended to handle the case with outliers. In this paper, we propose a new variant of coreset technique, \emph{layered sampling}, to deal with two fundamental robust optimization problems: \emph{k-median/means clustering with outliers} and \emph{linear regression with outliers}. This new coreset method is in particular suitable to speed up the iterative algorithms (which often improve the solution within a local range) for those robust optimization problems.

Growing Adaptive Multi-hyperplane Machines

- Nemanja Djuric, Zhuang Wang, Slobodan Vucetic
- abstract: Adaptive Multi-hyperplane Machine (AMM) is an online algorithm for learning Multi-hyperplane Machine (MM), a classification model which allows multiple hyperplanes per class. AMM is based on Stochastic Gradient Descent (SGD), with training time comparable to linear Support Vector Machine (SVM) and significantly higher accuracy. On the other hand, empirical results indicate there is a large accuracy gap between AMM and non-linear SVMs. In this paper we show that this performance gap is not due to limited representability of the MM model, as it can represent arbitrary concepts. We set to explain the connection between the AMM and Learning Vector Quantization (LVQ) algorithms, and introduce a novel Growing AMM (GAMM) classifier motivated by Growing LVQ, that imputes duplicate hyperplanes into the MM model during SGD training. We provide theoretical results showing that GAMM has favorable convergence properties, and analyze the generalization bound of the MM models. Experiments indicate that GAMM achieves significantly improved accuracy on non-linear problems, with only slightly slower training compared to AMM. On some tasks GAMM comes close to non-linear SVM, and outperforms other popular classifiers such as Neural Networks and Random Forests.

Inexact Tensor Methods with Dynamic Accuracies

- Nikita Doikov, Yurii Nesterov
- abstract: In this paper, we study inexact high-order Tensor Methods for solving convex optimization problems with composite objective. At every step of such methods, we use approximate solution of the auxiliary problem, defined by the bound for the residual in function value. We propose two dynamic strategies for choosing the inner accuracy: the first one is decreasing as $1/k^{p + 1}$, where $p \geq 1$ is the order of the method and k is the iteration counter, and the second approach is using for the inner accuracy the last progress in the target objective. We show that inexact Tensor Methods with these strategies achieve the same global convergence rate as in the error-free case. For the second approach we also establish local superlinear rates (for $p \geq 2$), and propose the accelerated scheme. Lastly, we present computational results on a variety of machine learning problems for several methods and different accuracy policies.

Provable Smoothness Guarantees for Black-Box Variational Inference

- Justin Domke
- abstract: Black-box variational inference tries to approximate a complex target distribution through a gradient-based optimization of the parameters of a simpler distribution. Provable convergence guarantees require structural properties of the objective. This paper shows that for location-scale family approximations, if the target is M-Lipschitz smooth, then so is the “energy” part of the variational objective. The key proof idea is to describe gradients in a certain inner-product space, thus permitting the use of Bessel’s inequality. This result gives bounds on the location of the optimal parameters, and is a key ingredient for convergence guarantees.

Optimal Differential Privacy Composition for Exponential Mechanisms

- Jinshuo Dong, David Durfee, Ryan Rogers
- abstract: Composition is one of the most important properties of differential privacy (DP), as it allows algorithm designers to build complex private algorithms from DP primitives. We consider precise composition bounds of the overall privacy loss for exponential mechanisms, one of the fundamental classes of mechanisms in DP. Exponential mechanism has also become a fundamental building block in private machine learning, e.g. private PCA and hyper-parameter selection. We give explicit formulations of the optimal privacy loss for both the adaptive and non-adaptive composition of exponential mechanism. For the non-adaptive setting in which each mechanism has the same privacy parameter, we give an efficiently computable formulation of the optimal privacy loss. In the adaptive case, we derive a recursive formula and an efficiently computable upper bound. These precise understandings about the problem lead to a 40% saving of the privacy budget in a practical application. Furthermore, the algorithm-specific analysis shows a difference in privacy parameters of adaptive and non-adaptive composition, which was widely believed to not exist based on the evidence from general analysis.

Multinomial Logit Bandit with Low Switching Cost

- Kefan Dong, Yingkai Li, Qin Zhang, Yuan Zhou
- abstract: We study multinomial logit bandit with limited adaptivity, where the algorithms change their exploration actions as infrequently as possible when achieving almost optimal minimax regret. We propose two measures of adaptivity: the assortment switching cost and the more fine-grained item switching

cost. We present an anytime algorithm (AT-DUCB) with $O(N \log T)$ assortment switches, almost matching the lower bound $\Omega(\frac{N \log T}{ \log \log T})$. In the fixed-horizon setting, our algorithm FH-DUCB incurs $O(N \log \log T)$ assortment switches, matching the asymptotic lower bound. We also present the ESUCB algorithm with item switching cost $O(N \log^2 T)$.

Towards Adaptive Residual Network Training: A Neural-ODE Perspective

- Chengyu Dong, Liyuan Liu, Zichao Li, Jingbo Shang
- abstract: In pursuit of resource-economical machine learning, attempts have been made to dynamically adjust computation workloads in different training stages, i.e., starting with a shallow network and gradually increasing the model depth (and computation workloads) during training. However, there is neither guarantee nor guidance on designing such network grow, due to the lack of its theoretical underpinnings. In this work, to explore the theory behind, we conduct theoretical analyses from an ordinary differential equation perspective. Specifically, we illustrate the dynamics of network growth and propose a novel performance measure specific to the depth increase. Illuminated by our analyses, we move towards theoretically sound growing operations and schedulers, giving rise to an adaptive training algorithm for residual networks, LipGrow, which automatically increases network depth thus accelerates training. In our experiments, it achieves comparable performance while reducing ~ 50% of training time.

On the Expressivity of Neural Networks for Deep Reinforcement Learning

- Kefan Dong, Yuping Luo, Tianhe Yu, Chelsea Finn, Tengyu Ma
- abstract: We compare the model-free reinforcement learning with the model-based approaches through the lens of the expressive power of neural networks for policies, Q-functions, and dynamics. We show, theoretically and empirically, that even for one-dimensional continuous state space, there are many MDPs whose optimal Q-functions and policies are much more complex than the dynamics. For these MDPs, model-based planning is a favorable algorithm, because the resulting policies can approximate the optimal policy significantly better than a neural network parameterization can, and model-free or model-based policy optimization rely on policy parameterization. Motivated by the theory, we apply a simple multi-step model-based bootstrapping planner (BOOTS) to bootstrap a weak Q-function into a stronger policy. Empirical results show that applying BOOTS on top of model-based or model-free policy optimization algorithms at the test time improves the performance on benchmark tasks.

Collapsed Amortized Variational Inference for Switching Nonlinear Dynamical Systems

- Zhe Dong, Bryan Seybold, Kevin Murphy, Hung Bui
- abstract: We propose an efficient inference method for switching nonlinear dynamical systems. The key idea is to learn an inference network which can be used as a proposal distribution for the continuous latent variables, while performing exact marginalization of the discrete latent variables. This allows us to use the reparameterization trick, and apply end-to-end training with stochastic gradient descent. We show that the proposed method can successfully segment time series data, including videos and 3D human pose, into meaningful “regimes” by using the piece-wise nonlinear dynamics.

Expert Learning through Generalized Inverse Multiobjective Optimization: Models, Insights, and Algorithms

- Chaosheng Dong, Bo Zeng
- abstract: We consider a new unsupervised learning task of inferring parameters of a multiobjective decision making model, based on a set of observed decisions from the human expert. This setting is important in applications (such as the task of portfolio management) where it may be difficult to obtain the human expert’s intrinsic decision making model. We formulate such a learning problem as an inverse multiobjective optimization problem (IMOP) and propose its first sophisticated model with statistical guarantees. Then, we reveal several fundamental connections between IMOP, K-means clustering, and manifold learning. Leveraging these critical insights and connections, we propose two algorithms to solve IMOP through manifold learning and clustering. Numerical results confirm the effectiveness of our model and the computational efficacy of algorithms.

The Complexity of Finding Stationary Points with Stochastic Gradient Descent

- Yoel Drori, Ohad Shamir
- abstract: We study the iteration complexity of stochastic gradient descent (SGD) for minimizing the gradient norm of smooth, possibly nonconvex functions. We provide several results, implying that the classical $\mathcal{O}(\epsilon^{-4})$ upper bound (for making the average gradient norm less than ϵ) cannot be improved upon, unless a combination of additional assumptions is made. Notably, this holds even if we limit ourselves to convex quadratic functions. We also show that for nonconvex functions, the feasibility of minimizing gradients with SGD is surprisingly sensitive to the choice of optimality criteria.

Optimal Non-parametric Learning in Repeated Contextual Auctions with Strategic Buyer

- Alexey Drutsa
- abstract: We study learning algorithms that optimize revenue in repeated contextual posted-price auctions where a seller interacts with a single strategic buyer that seeks to maximize his cumulative discounted surplus. The buyer’s valuation of a good is a fixed private function of a d-dimensional context (feature) vector that describes the good being sold. In contrast to existing studies on repeated contextual auctions with strategic buyer, in our work, the seller is not assumed to know the parametric model that underlies this valuation function. We introduce a novel non-parametric learning algorithm that is horizon-independent and has tight strategic regret upper bound of $\Theta(T^{d/(d+1)})$. We also non-trivially generalize several value-localization techniques of non-contextual repeated auctions to make them effective in the considered contextual non-parametric learning of the buyer valuation function.

Reserve Pricing in Repeated Second-Price Auctions with Strategic Bidders

- Alexey Drutsa
- abstract: We study revenue optimization learning algorithms for repeated second-price auctions with reserve where a seller interacts with multiple strategic bidders each of which holds a fixed private valuation for a good and seeks to maximize his expected future cumulative discounted surplus. We propose a novel algorithm that has strategic regret upper bound of $O(\log\log T)$ for worst-case valuations. This pricing is based on our novel transformation that upgrades an algorithm designed for the setup with a single buyer to the multi-buyer case. We provide theoretical guarantees on the ability of a transformed algorithm to learn the valuation of a strategic buyer, which has uncertainty about the future due to the presence of rivals.

NGBoost: Natural Gradient Boosting for Probabilistic Prediction

- Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, Alejandro Schuler
- abstract: We present Natural Gradient Boosting (NGBoost), an algorithm for generic probabilistic prediction via gradient boosting. Typical regression models return a point estimate, conditional on covariates, but probabilistic regression models output a full probability distribution over the outcome space, conditional on the covariates. This allows for predictive uncertainty estimation - crucial in applications like healthcare and weather forecasting. NGBoost generalizes gradient boosting to probabilistic regression by treating the parameters of the conditional distribution as targets for a multiparameter boosting algorithm. Furthermore, we show how the Natural Gradient is required to correct the training dynamics of our multiparameter boosting approach.

NGBoost can be used with any base learner, any family of distributions with continuous parameters, and any scoring rule. NGBoost matches or exceeds the performance of existing methods for probabilistic prediction while offering additional benefits in flexibility, scalability, and usability. An open-source implementation is available at github.com/stanfordmlgroup/ngboost.

Minimax-Optimal Off-Policy Evaluation with Linear Function Approximation

- Yaqi Duan, Zeyu Jia, Mengdi Wang
- abstract: This paper studies the statistical theory of off-policy evaluation with function approximation in batch data reinforcement learning problem. We consider a regression-based fitted Q-iteration method, show that it is equivalent to a model-based method that estimates a conditional mean embedding of the transition operator, and prove that this method is information-theoretically optimal and has nearly minimal estimation error. In particular, by leveraging contraction property of Markov processes and martingale concentration, we establish a finite-sample instance-dependent error upper bound and a nearly-matching minimax lower bound. The policy evaluation error depends sharply on a restricted χ^2-divergence over the function class between the long-term distribution of target policy and the distribution of past data. This restricted χ^2-divergence characterizes the statistical limit of off-policy evaluation and is both instance-dependent and function-class-dependent. Further, we provide an easily computable confidence bound for the policy evaluator, which may be useful for optimistic planning and safe policy improvement.

Online Bayesian Moment Matching based SAT Solver Heuristics

- Haonan Duan, Saeed Nejati, George Trimponias, Pascal Poupart, Vijay Ganesh
- abstract: In this paper, we present a Bayesian Moment Matching (BMM) based method aimed at solving the initialization problem in Boolean SAT solvers. The initialization problem can be stated as follows: given a SAT formula ϕ, compute an initial order over the variables of ϕ and values/polarity for these variables such that the runtime of SAT solvers on input ϕ is minimized. At the start of a solver run, our BMM-based methods compute a posterior probability distribution for an assignment to the variables of the input formula after analyzing its clauses, which will then be used by the solver to initialize its search. We perform extensive experiments to evaluate the efficacy of our BMM-based heuristic against 4 other initialization methods (random, survey propagation, Jeroslow-Wang, and default) in state-of-the-art solvers, MapleCOMSPS and MapleLCMDistChronotBT over the SAT competition 2018 application benchmark, as well as the best-known solvers in the cryptographic category, namely, CryptoMiniSAT, Glucose, and MapleSAT. On the cryptographic benchmark, BMM-based solvers out-perform all other initialization methods. Further, the BMM-based MapleCOMSPS significantly out-perform the same solver using all other initialization methods by 12 additional instances solved and better average runtime, over the SAT 2018 competition benchmark.

Familywise Error Rate Control by Interactive Unmasking

- Boyan Duan, Aaditya Ramdas, Larry Wasserman
- abstract: We propose a method for multiple hypothesis testing with familywise error rate (FWER) control, called the i-FWER test. Most testing methods are predefined algorithms that do not allow modifications after observing the data. However, in practice, analysts tend to choose a promising algorithm after observing the data; unfortunately, this violates the validity of the conclusion. The i-FWER test allows much flexibility: a human (or a computer program acting on the human's behalf) may adaptively guide the algorithm in a data-dependent manner. We prove that our test controls FWER if the analysts adhere to a particular protocol of masking and unmasking. We demonstrate via numerical experiments the power of our test under structured non-nulls, and then explore new forms of masking.

Cooperative Multi-Agent Bandits with Heavy Tails

- Abhimanyu Dubey, Alex 'Sandy' Pentland
- abstract: We study the heavy-tailed stochastic bandit problem in the cooperative multi-agent setting, where a group of agents interact with a common bandit problem, while communicating on a network with delays. Existing algorithms for the stochastic bandit in this setting utilize confidence intervals arising from an averaging-based communication protocol known as running consensus, that does not lend itself to robust estimation for heavy-tailed settings. We propose MP-UCB, a decentralized multi-agent algorithm for the cooperative stochastic bandit that incorporates robust estimation with a message-passing protocol. We prove optimal regret bounds for MP-UCB for several problem settings, and also demonstrate its superiority to existing methods. Furthermore, we establish the first lower bounds for the cooperative bandit problem, in addition to providing efficient algorithms for robust bandit estimation of location.

Kernel Methods for Cooperative Multi-Agent Contextual Bandits

- Abhimanyu Dubey, Alex 'Sandy' Pentland
- abstract: Cooperative multi-agent decision making involves a group of agents cooperatively solving learning problems while communicating over a network with delays. In this paper, we consider the kernelised contextual bandit problem, where the reward obtained by an agent is an arbitrary linear function of the contexts' images in the related reproducing kernel Hilbert space (RKHS), and a group of agents must cooperate to collectively solve their unique decision problems. For this problem, we propose Coop-KernelUCB, an algorithm that provides near-optimal bounds on the per-agent regret, and is both computationally and communicatively efficient. For special cases of the cooperative problem, we also provide variants of Coop-KernelUCB that provides optimal per-agent regret. In addition, our algorithm generalizes several existing results in the multi-agent bandit setting. Finally, on a series of both synthetic and real-world multi-agent network benchmarks, we demonstrate that our algorithm significantly outperforms existing benchmarks.

Optimization Theory for ReLU Neural Networks Trained with Normalization Layers

- Yonatan Dukler, Quanquan Gu, Guido Montufar
- abstract: The current paradigm of deep neural networks has been successful in part due to the use of normalization layers. Normalization layers like Batch Normalization, Layer Normalization and Weight Normalization are ubiquitous in practice as they improve the generalization performance and training speed of neural networks significantly. Nonetheless, the vast majority of current deep learning theory and non-convex optimization literature focuses on the un-normalized setting. We bridge this gap by providing the first global convergence result for 2 layer non-linear neural networks with ReLU activations trained with a normalization layer, namely Weight Normalization. The analysis shows how the introduction of normalization layers changes the optimization landscape and in some settings enables faster convergence as compared with un-normalized neural networks.

Equivariant Neural Rendering

- Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, Qi Shan
- abstract: We propose a framework for learning neural scene representations directly from images, without 3D supervision. Our key insight is that 3D structure can be imposed by ensuring that the learned representation transforms like a real 3D scene. Specifically, we introduce a loss which enforces equivariance of the scene representation with respect to 3D transformations. Our formulation allows us to infer and render scenes in real time while achieving comparable results to models requiring minutes for inference. In addition, we introduce two challenging new datasets for scene representation and neural rendering, including scenes with complex lighting and backgrounds. Through experiments, we show that our model achieves compelling results on these datasets as well as on standard ShapeNet benchmarks.

On Contrastive Learning for Likelihood-free Inference

- Conor Durkan, Iain Murray, George Papamakarios
- abstract: Likelihood-free methods perform parameter inference in stochastic simulator models where evaluating the likelihood is intractable but sampling synthetic data is possible. One class of methods for this likelihood-free problem uses a classifier to distinguish between pairs of parameter-observation samples generated using the simulator and pairs sampled from some reference distribution, which implicitly learns a density ratio proportional to the likelihood. Another popular class of methods fits a conditional distribution to the parameter posterior directly, and a particular recent variant allows for the use of flexible neural density estimators for this task. In this work, we show that both of these approaches can be unified under a general contrastive learning scheme, and clarify how they should be run and compared.

Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors

- Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, Dustin Tran
- abstract: Bayesian neural networks (BNNs) demonstrate promising success in improving the robustness and uncertainty quantification of modern deep learning. However, they generally struggle with underfitting at scale and parameter efficiency. On the other hand, deep ensembles have emerged as alternatives for uncertainty quantification that, while outperforming BNNs on certain problems, also suffer from efficiency issues. It remains unclear how to combine the strengths of these two approaches and remediate their common issues. To tackle this challenge, we propose a rank-1 parameterization of BNNs, where each weight matrix involves only a distribution on a rank-1 subspace. We also revisit the use of mixture approximate posteriors to capture multiple modes, where unlike typical mixtures, this approach admits a significantly smaller memory increase (e.g., only a 0.4% increase for a ResNet-50 mixture of size 10). We perform a systematic empirical study on the choices of prior, variational posterior, and methods to improve training. For ResNet-50 on ImageNet, Wide ResNet 28-10 on CIFAR-10/100, and an RNN on MIMIC-III, rank-1 BNNs achieve state-of-the-art performance across log-likelihood, accuracy, and calibration on the test sets and out-of-distribution variants.

Sparse Gaussian Processes with Spherical Harmonic Features

- Vincent Dutordoir, Nicolas Durrande, James Hensman
- abstract: We introduce a new class of inter-domain variational Gaussian processes (GP) where data is mapped onto the unit hypersphere in order to use spherical harmonic representations. Our inference scheme is comparable to variational Fourier features, but it does not suffer from the curse of dimensionality, and leads to diagonal covariance matrices between inducing variables. This enables a speed-up in inference, because it bypasses the need to invert large covariance matrices. Our experiments show that our model is able to fit a regression model for a dataset with 6 million entries two orders of magnitude faster compared to standard sparse GPs, while retaining state of the art accuracy. We also demonstrate competitive performance on classification with non-conjugate likelihoods.

Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing

- Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, Kush Varshney
- abstract: A trade-off between accuracy and fairness is almost taken as a given in the existing literature on fairness in machine learning. Yet, it is not preordained that accuracy should decrease with increased fairness. Novel to this work, we examine fair classification through the lens of mismatched hypothesis testing: trying to find a classifier that distinguishes between two ideal distributions when given two mismatched distributions that are biased. Using Chernoff information, a tool in information theory, we theoretically demonstrate that, contrary to popular belief, there always exist ideal distributions such that optimal fairness and accuracy (with respect to the ideal distributions) are achieved simultaneously: there is no trade-off. Moreover, the same classifier yields the lack of a trade-off with respect to ideal distributions while yielding a trade-off when accuracy is measured with respect to the given (possibly biased) dataset. To complement our main result, we formulate an optimization to find ideal distributions and derive fundamental limits to explain why a trade-off exists on the given biased dataset. We also derive conditions under which active data collection can alleviate the fairness-accuracy trade-off in the real world. Our results lead us to contend that it is problematic to measure accuracy with respect to data that reflects bias, and instead, we should be considering accuracy with respect to ideal, unbiased data.

Self-Concordant Analysis of Frank-Wolfe Algorithms

- Pavel Dvurechensky, Petr Ostroukhov, Kamil Safin, Shimrit Shtern, Mathias Staudigl
- abstract: Projection-free optimization via different variants of the Frank-Wolfe (FW), a.k.a. Conditional Gradient method has become one of the cornerstones in optimization for machine learning since in many cases the linear minimization oracle is much cheaper to implement than projections and some sparsity needs to be preserved. In a number of applications, e.g. Poisson inverse problems or quantum state tomography, the loss is given by a self-concordant (SC) function having unbounded curvature, implying absence of theoretical guarantees for the existing FW methods. We use the theory of SC functions to provide a new adaptive step size for FW methods and prove global convergence rate O(1/k) after k iterations. If the problem admits a stronger local linear minimization oracle, we construct a novel FW method with linear convergence rate for SC functions.

Estimating Q(s,s') with Deep Deterministic Dynamics Gradients

- Ashley Edwards, Himanshu Sahni, Rosanne Liu, Jane Hung, Ankit Jain, Rui Wang, Adrien Ecoffet, Thomas Miconi, Charles Isbell, Jason Yosinski
- abstract: In this paper, we introduce a novel form of value function, $Q(s, s')$, that expresses the utility of transitioning from a state s to a neighboring state s' and then acting optimally thereafter. In order to derive an optimal policy, we develop a forward dynamics model that learns to make next-state predictions that maximize this value. This formulation decouples actions from values while still learning off-policy. We highlight the benefits of this approach in terms of value function transfer, learning within redundant action spaces, and learning off-policy from state observations generated by sub-optimal or completely random policies. Code and videos are available at http://sites.google.com/view/qss-paper.

Training Linear Neural Networks: Non-Local Convergence and Complexity Results

- Armin Eftekhari
- abstract: Linear networks provide valuable insights into the workings of neural networks in general. This paper identifies conditions under which the gradient flow provably trains a linear network, in spite of the non-strict saddle points present in the optimization landscape. This paper also provides the computational complexity of training linear networks with gradient flow. To achieve these results, this work develops a machinery to provably identify the stable set of gradient flow, which then enables us to improve over the state of the art in the literature of linear networks (Bah et al., 2019;Arora et al., 2018a). Crucially, our results appear to be the first to break away from the lazy training regime which has dominated the literature of neural networks. This work requires the network to have a layer with one neuron, which subsumes the networks with a scalar output, but extending the results of this theoretical work to all linear networks remains a challenging open problem.

Student-Teacher Curriculum Learning via Reinforcement Learning: Predicting Hospital Inpatient Admission Location

- Rasheed El-Bouri, David Eyre, Peter Watkinson, Tingting Zhu, David Clifton

- abstract: Accurate and reliable prediction of hospital admission location is important due to resource-constraints and space availability in a clinical setting, particularly when dealing with patients who come from the emergency department. In this work we propose a student-teacher network via reinforcement learning to deal with this specific problem. A representation of the weights of the student network is treated as the state and is fed as an input to the teacher network. The teacher network's action is to select the most appropriate batch of data to train the student network on from a training set sorted according to entropy. By validating on three datasets, not only do we show that our approach outperforms state-of-the-art methods on tabular data and performs competitively on image recognition, but also that novel curricula are learned by the teacher network. We demonstrate experimentally that the teacher network can actively learn about the student network and guide it to achieve better performance than if trained alone.

Decision Trees for Decision-Making under the Predict-then-Optimize Framework

- Adam Elmachtoub, Jason Cheuk Nam Liang, Ryan Mcnellis
- abstract: We consider the use of decision trees for decision-making problems under the predict-then-optimize framework. That is, we would like to first use a decision tree to predict unknown input parameters of an optimization problem, and then make decisions by solving the optimization problem using the predicted parameters. A natural loss function in this framework is to measure the suboptimality of the decisions induced by the predicted input parameters, as opposed to measuring loss using input parameter prediction error. This natural loss function is known in the literature as the Smart Predict-then-Optimize (SPO) loss, and we propose a tractable methodology called SPO Trees (SPOTs) for training decision trees under this loss. SPOTs benefit from the interpretability of decision trees, providing an interpretable segmentation of contextual features into groups with distinct optimal solutions to the optimization problem of interest. We conduct several numerical experiments on synthetic and real data including the prediction of travel times for shortest path problems and predicting click probabilities for news article recommendation. We demonstrate on these datasets that SPOTs simultaneously provide higher quality decisions and significantly lower model complexity than other machine learning approaches (e.g., CART) trained to minimize prediction error.

Revisiting Spatial Invariance with Low-Rank Local Connectivity

- Gamaleldin Elsayed, Prajit Ramachandran, Jonathon Shlens, Simon Kornblith
- abstract: Convolutional neural networks are among the most successful architectures in deep learning with this success at least partially attributable to the efficacy of spatial invariance as an inductive bias. Locally connected layers, which differ from convolutional layers only in their lack of spatial invariance, usually perform poorly in practice. However, these observations still leave open the possibility that some degree of relaxation of spatial invariance may yield a better inductive bias than either convolution or local connectivity. To test this hypothesis, we design a method to relax the spatial invariance of a network layer in a controlled manner; we create a \emph{low-rank} locally connected layer, where the filter bank applied at each position is constructed as a linear combination of basis set of filter banks with spatially varying combining weights. By varying the number of basis filter banks, we can control the degree of relaxation of spatial invariance. In experiments with small convolutional networks, we find that relaxing spatial invariance improves classification accuracy over both convolution and locally connected layers across MNIST, CIFAR-10, and CelebA datasets, thus suggesting that spatial invariance may be an overly restrictive prior.

Divide and Conquer: Leveraging Intermediate Feature Representations for Quantized Training of Neural Networks

- Ahmed Taha Elthakeb, Prannoy Pilligundla, Fatemeh Mireshghallah, Alexander Cloninger, Hadi Esmaeilzadeh
- abstract: The deep layers of modern neural networks extract a rather rich set of features as an input propagates through the network, this paper sets out to harvest these rich intermediate representations for quantization with minimal accuracy loss while significantly reducing the memory footprint and compute intensity of the DNN. This paper utilizes knowledge distillation through teacher-student paradigm (Hinton et al., 2015) in a novel setting that exploits the feature extraction capability of DNNs for higher accuracy quantization. As such, our algorithm logically divides a pretrained full-precision DNN to multiple sections, each of which exposes intermediate features to train a team of students independently in the quantized domain and simply stitching them afterwards. This divide and conquer strategy, makes the training of each student section possible in isolation, speeding up training by enabling parallelization. Experiments on various DNNs (AlexNet, LeNet, MobileNet, ResNet-18, ResNet-20, SVHN and VGG-11) show that, this approach{—}called DCQ (Divide and Conquer Quantization){—}on average, improves the performance of a state-of-the-art quantized training technique, DoReFa-Net (Zhou et al., 2016) by 21.6% and 9.3% for binary and ternary quantization, respectively. Additionally, we show that incorporating DCQ to existing quantized training methods leads to improved accuracies as compared to previously reported by multiple state-of-the-art quantized training methods.

Generalization Error of Generalized Linear Models in High Dimensions

- Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, Alyson Fletcher
- abstract: At the heart of machine learning lies the question of generalizability of learned rules over previously unseen data. While over-parameterized models based on neural networks are now ubiquitous in machine learning applications, our understanding of their generalization capabilities is incomplete and this task is made harder by the non-convexity of the underlying learning problems. We provide a general framework to characterize the asymptotic generalization error for single-layer neural networks (i.e., generalized linear models) with arbitrary non-linearities, making it applicable to regression as well as classification problems. This framework enables analyzing the effect of (i) over-parameterization and non-linearity during modeling; (ii) choices of loss function, initialization, and regularizer during learning; and (iii) mismatch between training and test distributions. As examples, we analyze a few special cases, namely linear regression and logistic regression. We are also able to rigorously and analytically explain the \emph{double descent} phenomenon in generalized linear models.

Parallel Algorithm for Non-Monotone DR-Submodular Maximization

- Alina Ene, Huy Nguyen
- abstract: In this work, we give a new parallel algorithm for the problem of maximizing a non-monotone diminishing returns submodular function subject to a cardinality constraint. For any desired accuracy ϵ, our algorithm achieves a $1/e - \epsilon$ approximation using $O(\log{n} \log(1/\epsilon) / \epsilon^3)$ parallel rounds of function evaluations. The approximation guarantee nearly matches the best approximation guarantee known for the problem in the sequential setting and the number of parallel rounds is nearly-optimal for any constant ϵ. Previous algorithms achieve worse approximation guarantees using $\Omega(\log^2{n})$ parallel rounds. Our experimental evaluation suggests that our algorithm obtains solutions whose objective value nearly matches the value obtained by the state of the art sequential algorithms, and it outperforms previous parallel algorithms in number of parallel rounds, iterations, and solution quality.

Continuous Time Bayesian Networks with Clocks

- Nicolai Engelmann, Dominik Linzner, Heinz Koeppl
- abstract: Structured stochastic processes evolving in continuous time present a widely adopted framework to model phenomena occurring in nature and engineering. However, such models are often chosen to satisfy the Markov property to maintain tractability. One of the more popular of such memoryless models are Continuous Time Bayesian Networks (CTBNs). In this work, we lift its restriction to exponential survival times to arbitrary distributions. Current extensions achieve this via auxiliary states, which hinder tractability. To avoid that, we introduce a set of node-wise clocks to construct a collection of graph-coupled semi-Markov chains. We provide algorithms for parameter and structure inference, which make use of local dependencies and

conduct experiments on synthetic data and a data-set generated through a benchmark tool for gene regulatory networks. In doing so, we point out advantages compared to current CTBN extensions.

Identifying Statistical Bias in Dataset Replication

- Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, Aleksander Madry
- abstract: Dataset replication is a useful tool for assessing whether improvements in test accuracy on a specific benchmark correspond to improvements in models' ability to generalize reliably. In this work, we present unintuitive yet significant ways in which standard approaches to dataset replication introduce statistical bias, skewing the resulting observations. We study ImageNet-v2, a replication of the ImageNet dataset on which models exhibit a significant (11-14%) drop in accuracy, even after controlling for selection frequency, a human-in-the-loop measure of data quality. We show that after remeasuring selection frequencies and correcting for statistical bias, only an estimated 3.6% of the original 11.7% accuracy drop remains unaccounted for. We conclude with concrete recommendations for recognizing and avoiding bias in dataset replication. Code for our study is publicly available: https://git.io/data-rep-analysis.

Distributed Online Optimization over a Heterogeneous Network with Any-Batch Mirror Descent

- Nima Eshraghi, Ben Liang
- abstract: In distributed online optimization over a computing network with heterogeneous nodes, slow nodes can adversely affect the progress of fast nodes, leading to drastic slowdown of the overall convergence process. To address this issue, we consider a new algorithm termed Distributed Any-Batch Mirror Descent (DABMD), which is based on distributed Mirror Descent but uses a fixed per-round computing time to limit the waiting by fast nodes to receive information updates from slow nodes. DABMD is characterized by varying minibatch sizes across nodes. It is applicable to a broader range of problems compared with existing distributed online optimization methods such as those based on dual averaging, and it accommodates time-varying network topology. We study two versions of DABMD, depending on whether the computing nodes average their primal variables via single or multiple consensus iterations. We show that both versions provide strong theoretical performance guarantee, by deriving upperbounds on their expected dynamic regret, which capture the variability in minibatch sizes. Our experimental results show substantial reduction in cost and acceleration in convergence compared with the known best alternative.

Rigging the Lottery: Making All Tickets Winners

- Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, Erich Elsen
- abstract: Many applications require sparse neural networks due to space or inference time restrictions. There is a large body of work on training dense networks to yield sparse networks for inference, but this limits the size of the largest trainable sparse model to that of the largest trainable dense model. In this paper we introduce a method to train sparse neural networks with a fixed parameter count and a fixed computational cost throughout training, without sacrificing accuracy relative to existing dense-to-sparse training methods. Our method updates the topology of the sparse network during training by using parameter magnitudes and infrequent gradient calculations. We show that this approach requires fewer floating-point operations (FLOPs) to achieve a given level of accuracy compared to prior techniques. We demonstrate state-of-the-art sparse training results on a variety of networks and datasets, including ResNet-50, MobileNets on Imagenet-2012, and RNNs on WikiText-103. Finally, we provide some insights into why allowing the topology to change during the optimization can overcome local minima encountered when the topology remains static.

Faster Graph Embeddings via Coarsening

- Matthew Fahrbach, Gramoz Goranci, Richard Peng, Sushant Sachdeva, Chi Wang
- abstract: Graph embeddings are a ubiquitous tool for machine learning tasks, such as node classification and link prediction, on graph-structured data. However, computing the embeddings for large-scale graphs is prohibitively inefficient even if we are interested only in a small subset of relevant vertices. To address this, we present an efficient graph coarsening approach, based on Schur complements, for computing the embedding of the relevant vertices. We prove that these embeddings are preserved exactly by the Schur complement graph that is obtained via Gaussian elimination on the non-relevant vertices. As computing Schur complements is expensive, we give a nearly-linear time algorithm that generates a coarsened graph on the relevant vertices that provably matches the Schur complement in expectation in each iteration. Our experiments involving prediction tasks on graphs demonstrate that computing embeddings on the coarsened graph, rather than the entire graph, leads to significant time savings without sacrificing accuracy.

Latent Bernoulli Autoencoder

- Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, Paolo Remagnino
- abstract: In this work, we pose the question whether it is possible to design and train an autoencoder model in an end-to-end fashion to learn representations in the multivariate Bernoulli latent space, and achieve performance comparable with the state-of-the-art variational methods. Moreover, we investigate how to generate novel samples and perform smooth interpolation and attributes modification in the binary latent space. To meet our objective, we propose a simplified, deterministic model with a straight-through gradient estimator to learn the binary latents and show its competitiveness with the latest VAE methods. Furthermore, we propose a novel method based on a random hyperplane rounding for sampling and smooth interpolation in the latent space. Our method performs on a par or better than the current state-of-the-art methods on common CelebA, CIFAR-10 and MNIST datasets.

Optimal Sequential Maximization: One Interview is Enough!

- Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati
- abstract: Maximum selection under probabilistic queries \emph{(probabilistic maximization)} is a fundamental algorithmic problem arising in numerous theoretical and practical contexts. We derive the first query-optimal sequential algorithm for probabilistic-maximization. Departing from previous assumptions, the algorithm and performance guarantees apply even for infinitely many items, hence in particular do not require a-priori knowledge of the number of items. The algorithm has linear query complexity, and is optimal also in the streaming setting. To derive these results we consider a probabilistic setting where several candidates for a position are asked multiple questions with the goal of finding who has the highest probability of answering interview questions correctly. Previous work minimized the total number of questions asked by alternating back and forth between the best performing candidates, in a sense, inviting them to multiple interviews. We show that the same order-wise selection accuracy can be achieved by querying the candidates sequentially, never returning to a previously queried candidate. Hence one interview is enough!

Spectral Graph Matching and Regularized Quadratic Relaxations: Algorithm and Theory

- Zhou Fan, Cheng Mao, Yihong Wu, Jiaming Xu
- abstract: Graph matching, also known as network alignment, aims at recovering the latent vertex correspondence between two unlabeled, edge-correlated weighted graphs. To tackle this task, we propose a spectral method, GRAph Matching by Pairwise eigen-Alignments (GRAMPA), which first constructs a similarity matrix as a weighted sum of outer products between all pairs of eigenvectors of the two graphs, and then outputs a matching by a simple rounding procedure. For a universality class of correlated Wigner models, GRAMPA achieves exact recovery of the latent matching between two graphs with edge correlation $1 - 1/\mathrm{polylog}(n)$ and average degree at least $\mathrm{polylog}(n)$. This matches the state-of-the-art guarantees for

polynomial-time algorithms established for correlated Erdős-Rényi graphs, and significantly improves over existing spectral methods. The superiority of GRAMPA is also demonstrated on a variety of synthetic and real datasets, in terms of both statistical accuracy and computational efficiency.

On hyperparameter tuning in general clustering problemsm

- Xinjie Fan, Yuguang Yue, Purnamrita Sarkar, Y. X. Rachel Wang
- abstract: Tuning hyperparameters for unsupervised learning problems is difficult in general due to the lack of ground truth for validation. However, the success of most clustering methods depends heavily on the correct choice of the involved hyperparameters. Take for example the Lagrange multipliers of penalty terms in semidefinite programming (SDP) relaxations of community detection in networks, or the bandwidth parameter needed in the Gaussian kernel used to construct similarity matrices for spectral clustering. Despite the popularity of these clustering algorithms, there are not many provable methods for tuning these hyperparameters. In this paper, we provide an overarching framework with provable guarantees for tuning hyperparameters in the above class of problems under two different models. Our framework can be augmented with a cross validation procedure to do model selection as well. In a variety of simulation and real data experiments, we show that our framework outperforms other widely used tuning procedures in a broad range of parameter settings.

Online mirror descent and dual averaging: keeping pace in the dynamic case

- Huang Fang, Nick Harvey, Victor Portella, Michael Friedlander
- abstract: Online mirror descent (OMD) and dual averaging (DA)—two fundamental algorithms for online convex optimization—are known to have very similar (and sometimes identical) performance guarantees when used with a fixed learning rate. Under dynamic learning rates, however, OMD is provably inferior to DA and suffers a linear regret, even in common settings such as prediction with expert advice. We modify the OMD algorithm through a simple technique that we call stabilization. We give essentially the same abstract regret bound for OMD with stabilization and for DA by modifying the classical OMD convergence analysis in a careful and modular way that allows for straightforward and flexible proofs. Simple corollaries of these bounds show that OMD with stabilization and DA enjoy the same performance guarantees in many applications—even under dynamic learning rates. We also shed light on the similarities between OMD and DA and show simple conditions under which stabilized-OMD and DA generate the same iterates.

Stochastic Regret Minimization in Extensive-Form Games

- Gabriele Farina, Christian Kroer, Tuomas Sandholm
- abstract: Monte-Carlo counterfactual regret minimization (MCCFR) is the state-of-the-art algorithm for solving sequential games that are too large for full tree traversals. It works by using gradient estimates that can be computed via sampling. However, stochastic methods for sequential games have not been investigated extensively beyond MCCFR. In this paper we develop a new framework for developing stochastic regret minimization methods. This framework allows us to use any regret-minimization algorithm, coupled with any gradient estimator. The MCCFR algorithm can be analyzed as a special case of our framework, and this analysis leads to significantly stronger theoretical guarantees on convergence, while simultaneously yielding a simplified proof. Our framework allows us to instantiate several new stochastic methods for solving sequential games. We show extensive experiments on five games, where some variants of our methods outperform MCCFR.

Do GANs always have Nash equilibria?

- Farzan Farnia, Asuman Ozdaglar
- abstract: Generative adversarial networks (GANs) represent a zero-sum game between two machine players, a generator and a discriminator, designed to learn the distribution of data. While GANs have achieved state-of-the-art performance in several benchmark learning tasks, GAN minimax optimization still poses great theoretical and empirical challenges. GANs trained using first-order optimization methods commonly fail to converge to a stable solution where the players cannot improve their objective, i.e., the Nash equilibrium of the underlying game. Such issues raise the question of the existence of Nash equilibria in GAN zero-sum games. In this work, we show through theoretical and numerical results that indeed GAN zero-sum games may have no Nash equilibria. To characterize an equilibrium notion applicable to GANs, we consider the equilibrium of a new zero-sum game with an objective function given by a proximal operator applied to the original objective, a solution we call the proximal equilibrium. Unlike the Nash equilibrium, the proximal equilibrium captures the sequential nature of GANs, in which the generator moves first followed by the discriminator. We prove that the optimal generative model in Wasserstein GAN problems provides a proximal equilibrium. Inspired by these results, we propose a new approach, which we call proximal training, for solving GAN problems. We perform several numerical experiments indicating the existence of proximal equilibria in GANs.

Growing Action Spaces

- Gregory Farquhar, Laura Gustafson, Zeming Lin, Shimon Whiteson, Nicolas Usunier, Gabriel Synnaeve
- abstract: In complex tasks, such as those with large combinatorial action spaces, random exploration may be too inefficient to achieve meaningful learning progress. In this work, we use a curriculum of progressively growing action spaces to accelerate learning. We assume the environment is out of our control, but that the agent may set an internal curriculum by initially restricting its action space. Our approach uses off-policy reinforcement learning to estimate optimal value functions for multiple action spaces simultaneously and efficiently transfers data, value estimates, and state representations from restricted action spaces to the full task. We show the efficacy of our approach in proof-of-concept control tasks and on challenging large-scale StarCraft micromanagement tasks with large, multi-agent action spaces.

Improved Optimistic Algorithms for Logistic Bandits

- Louis Faury, Marc Abeille, Clement Calauzenes, Olivier Fercoq
- abstract: The generalized linear bandit framework has attracted a lot of attention in recent years by extending the well-understood linear setting and allowing to model richer reward structures. It notably covers the logistic model, widely used when rewards are binary. For logistic bandits, the frequentist regret guarantees of existing algorithms are $\tilde{\mathcal{O}}(\kappa \sqrt{T})$, where κ is a problem-dependent constant. Unfortunately, κ can be arbitrarily large as it scales exponentially with the size of the decision set. This may lead to significantly loose regret bounds and poor empirical performance. In this work, we study the logistic bandit with a focus on the prohibitive dependencies introduced by κ. We propose a new optimistic algorithm based on a finer examination of the non-linearities of the reward function. We show that it enjoys a $\tilde{\mathcal{O}}(\sqrt{T})$ regret with no dependency in κ, but for a second order term. Our analysis is based on a new tail-inequality for self-normalized martingales, of independent interest.

Revisiting Fundamentals of Experience Replay

- William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, Will Dabney
- abstract: Experience replay is central to off-policy algorithms in deep reinforcement learning (RL), but there remain significant gaps in our understanding. We therefore present a systematic and extensive analysis of experience replay in Q-learning methods, focusing on two fundamental properties: the replay capacity and the ratio of learning updates to experience collected (replay ratio). Our additive and ablative studies upend conventional wisdom around experience replay {—} greater capacity is found to substantially increase the performance of certain algorithms, while leaving others unaffected. Counterintuitively we show that theoretically ungrounded, uncorrected n-step returns are uniquely beneficial while other techniques confer limited benefit

for sifting through larger memory. Separately, by directly controlling the replay ratio we contextualize previous observations in the literature and empirically measure its importance across a variety of deep RL algorithms. Finally, we conclude by testing a set of hypotheses on the nature of these performance benefits.

Learning with Multiple Complementary Labels

- Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, Masashi Sugiyama
- abstract: A complementary label (CL) simply indicates an incorrect class of an example, but learning with CLs results in multi-class classifiers that can predict the correct class. Unfortunately, the problem setting only allows a single CL for each example, which notably limits its potential since our labelers may easily identify multiple CLs (MCLs) to one example. In this paper, we propose a novel problem setting to allow MCLs for each example and two ways for learning with MCLs. In the first way, we design two wrappers that decompose MCLs into many single CLs, so that we could use any method for learning with CLs. However, the supervision information that MCLs hold is conceptually diluted after decomposition. Thus, in the second way, we derive an unbiased risk estimator; minimizing it processes each set of MCLs as a whole and possesses an estimation error bound. We further improve the second way into minimizing properly chosen upper bounds. Experiments show that the former way works well for learning with MCLs but the latter is even better.

Global Concavity and Optimization in a Class of Dynamic Discrete Choice Models

- Yiding Feng, Ekaterina Khmelnitskaya, Denis Nekipelov
- abstract: Discrete choice models with unobserved heterogeneity are commonly used Econometric models for dynamic Economic behavior which have been adopted in practice to predict behavior of individuals and firms from schooling and job choices to strategic decisions in market competition. These models feature optimizing agents who choose among a finite set of options in a sequence of periods and receive choice-specific payoffs that depend on both variables that are observed by the agent and recorded in the data and variables that are only observed by the agent but not recorded in the data. Existing work in Econometrics assumes that optimizing agents are fully rational and requires finding a functional fixed point to find the optimal policy. We show that in an important class of discrete choice models the value function is globally concave in the policy. That means that simple algorithms that do not require fixed point computation, such as the policy gradient algorithm, globally converge to the optimal policy. This finding can both be used to relax behavioral assumption regarding the optimizing agents and to facilitate Econometric analysis of dynamic behavior. In particular, we demonstrate significant computational advantages in using a simple implementation policy gradient algorithm over existing “nested fixed point” algorithms used in Econometrics.

The Intrinsic Robustness of Stochastic Bandits to Strategic Manipulation

- Zhe Feng, David Parkes, Haifeng Xu
- abstract: Motivated by economic applications such as recommender systems, we study the behavior of stochastic bandits algorithms under \emph{strategic behavior} conducted by rational actors, i.e., the arms. Each arm is a \emph{self-interested} strategic player who can modify its own reward whenever pulled, subject to a cross-period budget constraint, in order to maximize its own expected number of times of being pulled. We analyze the robustness of three popular bandit algorithms: UCB, ε-Greedy, and Thompson Sampling. We prove that all three algorithms achieve a regret upper bound $\mathcal{O}(\max { B, K\ln T})$ where B is the total budget across arms, K is the total number of arms and T is the running time of the algorithms. This regret guarantee holds for \emph{arbitrary adaptive} manipulation strategy of arms. Our second set of main results shows that this regret bound is \emph{tight}— in fact, for UCB, it is tight even when we restrict the arms’ manipulation strategies to form a \emph{Nash equilibrium}. We do so by characterizing the Nash equilibrium of the game induced by arms’ strategic manipulations and show a regret lower bound of $\Omega(\max { B, K\ln T})$ at the equilibrium.

Accountable Off-Policy Evaluation With Kernel Bellman Statistics

- Yihao Feng, Tongzheng Ren, Ziyang Tang, Qiang Liu
- abstract: We consider off-policy evaluation (OPE), which evaluates the performance of a new policy from observed data collected from previous experiments, without requiring the execution of the new policy. This finds important applications in areas with high execution cost or safety concerns, such as medical diagnosis, recommendation systems and robotics. In practice, due to the limited information from off-policy data, it is highly desirable to construct rigorous confidence intervals, not just point estimation, for the policy performance. In this work, we propose a new variational framework which reduces the problem of calculating tight confidence bounds in OPE into an optimization problem on a feasible set that catches the true state-action value function with high probability. The feasible set is constructed by leveraging statistical properties of a recently proposed kernel Bellman loss (Feng et al., 2019). We design an efficient computational approach for calculating our bounds, and extend it to perform post-hoc diagnosis and correction for existing estimators. Empirical results show that our method yields tight confidence intervals in different settings.

Kernelized Stein Discrepancy Tests of Goodness-of-fit for Time-to-Event Data

- Tamara Fernandez, Nicolas Rivera, Wenkai Xu, Arthur Gretton
- abstract: Survival Analysis and Reliability Theory are concerned with the analysis of time-to-event data, in which observations correspond to waiting times until an event of interest such as death from a particular disease or failure of a component in a mechanical system. This type of data is unique due to the presence of censoring, a type of missing data that occurs when we do not observe the actual time of the event of interest but, instead, we have access to an approximation for it given by random interval in which the observation is known to belong. Most traditional methods are not designed to deal with censoring, and thus we need to adapt them to censored time-to-event data. In this paper, we focus on non-parametric goodness-of-fit testing procedures based on combining the Stein’s method and kernelized discrepancies. While for uncensored data, there is a natural way of implementing a kernelized Stein discrepancy test, for censored data there are several options, each of them with different advantages and disadvantages. In this paper, we propose a collection of kernelized Stein discrepancy tests for time-to-event data, and we study each of them theoretically and empirically; our experimental results show that our proposed methods perform better than existing tests, including previous tests based on a kernelized maximum mean discrepancy.

Why Are Learned Indexes So Effective?

- Paolo Ferragina, Fabrizio Lillo, Giorgio Vinciguerra
- abstract: A recent trend in algorithm design consists of augmenting classic data structures with machine learning models, which are better suited to reveal and exploit patterns and trends in the input data so to achieve outstanding practical improvements in space occupancy and time efficiency. This is especially known in the context of indexing data structures where, despite few attempts in evaluating their asymptotic efficiency, theoretical results are yet missing in showing that learned indexes are provably better than classic indexes, such as B+ trees and their variants. In this paper, we present the first mathematically-grounded answer to this open problem. We obtain this result by discovering and exploiting a link between the original problem and a mean exit time problem over a proper stochastic process which, we show, is related to the space and time occupancy of those learned indexes. Our general result is then specialised to five well-known distributions: Uniform, Lognormal, Pareto, Exponential, and Gamma; and it is corroborated in precision and robustness by a large set of experiments.

Implicit Learning Dynamics in Stackelberg Games: Equilibria Characterization, Convergence Analysis, and Empirical Study

- Tanner Fiez, Benjamin Chasnov, Lillian Ratliff
- abstract: Contemporary work on learning in continuous games has commonly overlooked the hierarchical decision-making structure present in machine learning problems formulated as games, instead treating them as simultaneous play games and adopting the Nash equilibrium solution concept. We deviate from this paradigm and provide a comprehensive study of learning in Stackelberg games. This work provides insights into the optimization landscape of zero-sum games by establishing connections between Nash and Stackelberg equilibria along with the limit points of simultaneous gradient descent. We derive novel gradient-based learning dynamics emulating the natural structure of a Stackelberg game using the implicit function theorem and provide convergence analysis for deterministic and stochastic updates for zero-sum and general-sum games. Notably, in zero-sum games using deterministic updates, we show the only critical points the dynamics converge to are Stackelberg equilibria and provide a local convergence rate. Empirically, our learning dynamics mitigate rotational behavior and exhibit benefits for training generative adversarial networks compared to simultaneous gradient descent.

Can Autonomous Vehicles Identify, Recover From, and Adapt to Distribution Shifts?

- Angelos Filos, Panagiotis Tigkas, Rowan Mcallister, Nicholas Rhinehart, Sergey Levine, Yarin Gal
- abstract: Out-of-training-distribution (OOD) scenarios are a common challenge of learning agents at deployment, typically leading to arbitrary deductions and poorly-informed decisions. In principle, detection of and adaptation to OOD scenes can mitigate their adverse effects. In this paper, we highlight the limitations of current approaches to novel driving scenes and propose an epistemic uncertainty-aware planning method, called \emph{robust imitative planning} (RIP). Our method can detect and recover from some distribution shifts, reducing the overconfident and catastrophic extrapolations in OOD scenes. If the model's uncertainty is too great to suggest a safe course of action, the model can instead query the expert driver for feedback, enabling sample-efficient online adaptation, a variant of our method we term \emph{adaptive robust imitative planning} (AdaRIP). Our methods outperform current state-of-the-art approaches in the nuScenes \emph{prediction} challenge, but since no benchmark evaluating OOD detection and adaption currently exists to assess \emph{control}, we introduce an autonomous car novel-scene benchmark, \texttt{CARNOVEL}, to evaluate the robustness of driving agents to a suite of tasks with distribution shifts, where our methods outperform all the baselines.

How to Train Your Neural ODE: the World of Jacobian and Kinetic Regularization

- Chris Finlay, Joern-Henrik Jacobsen, Levon Nurbekyan, Adam Oberman
- abstract: Training neural ODEs on large datasets has not been tractable due to the necessity of allowing the adaptive numerical ODE solver to refine its step size to very small values. In practice this leads to dynamics equivalent to many hundreds or even thousands of layers. In this paper, we overcome this apparent difficulty by introducing a theoretically-grounded combination of both optimal transport and stability regularizations which encourage neural ODEs to prefer simpler dynamics out of all the dynamics that solve a problem well. Simpler dynamics lead to faster convergence and to fewer discretizations of the solver, considerably decreasing wall-clock time without loss in performance. Our approach allows us to train neural ODE-based generative models to the same performance as the unregularized dynamics, with significant reductions in training time. This brings neural ODEs closer to practical relevance in large-scale applications.

Generalizing Convolutional Neural Networks for Equivariance to Lie Groups on Arbitrary Continuous Data

- Marc Finzi, Samuel Stanton, Pavel Izmailov, Andrew Gordon Wilson
- abstract: The translation equivariance of convolutional layers enables CNNs to generalize well on image problems. While translation equivariance provides a powerful inductive bias for images, we often additionally desire equivariance to other transformations, such as rotations, especially for non-image data. We propose a general method to construct a convolutional layer that is equivariant to transformations from any specified Lie group with a surjective exponential map. Incorporating equivariance to a new group requires implementing only the group exponential and logarithm maps, enabling rapid prototyping. Showcasing the simplicity and generality of our method, we apply the same model architecture to images, ball-and-stick molecular data, and Hamiltonian dynamical systems. For Hamiltonian systems, the equivariance of our models is especially impactful, leading to exact conservation of linear and angular momentum.

Information Particle Filter Tree: An Online Algorithm for POMDPs with Belief-Based Rewards on Continuous Domains

- Johannes Fischer, Ömer Sahin Tas
- abstract: Planning in Partially Observable Markov Decision Processes (POMDPs) inherently gathers the information necessary to act optimally under uncertainties. The framework can be extended to model pure information gathering tasks by considering belief-based rewards. This allows us to use reward shaping to guide POMDP planning to informative beliefs by using a weighted combination of the original reward and the expected information gain as the objective. In this work we propose a novel online algorithm, Information Particle Filter Tree (IPFT), to solve problems with belief-dependent rewards on continuous domains. It simulates particle-based belief trajectories in a Monte Carlo Tree Search (MCTS) approach to construct a search tree in the belief space. The evaluation shows that the consideration of information gain greatly improves the performance in problems where information gathering is an essential part of the optimal policy.

Topic Modeling via Full Dependence Mixtures

- Dan Fisher, Mark Kozdoba, Shie Mannor
- abstract: In this paper we introduce a new approach to topic modelling that scales to large datasets by using a compact representation of the data and by leveraging the GPU architecture. In this approach, topics are learned directly from the co-occurrence data of the corpus. In particular, we introduce a novel mixture model which we term the Full Dependence Mixture (FDM) model. FDMs model second moment under general generative assumptions on the data. While there is previous work on topic modeling using second moments, we develop a direct stochastic optimization procedure for fitting an FDM with a single Kullback Leibler objective. Moment methods in general have the benefit that an iteration no longer needs to scale with the size of the corpus. Our approach allows us to leverage standard optimizers and GPUs for the problem of topic modeling. In particular, we evaluate the approach on two large datasets, NeurIPS papers and a Twitter corpus, with a large number of topics, and show that the approach performs comparably or better than the standard benchmarks.

Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles

- Dylan Foster, Alexander Rakhlin
- abstract: A fundamental challenge in contextual bandits is to develop flexible, general-purpose algorithms with computational requirements no worse than classical supervised learning tasks such as classification and regression. Algorithms based on regression have shown promising empirical success, but theoretical guarantees have remained elusive except in special cases. We provide the first universal and optimal reduction from contextual bandits to online regression. We show how to transform any oracle for online regression with a given value function class into an algorithm for contextual bandits with the induced policy class, with no overhead in runtime or memory requirements. We characterize the minimax rates for contextual bandits with general, potentially nonparametric function classes, and show that our algorithm is minimax optimal whenever the oracle obtains the optimal rate for regression. Compared to previous results, our algorithm requires no distributional assumptions beyond realizability, and works even when contexts are chosen adversarially.

Logarithmic Regret for Adversarial Online Control

- Dylan Foster, Max Simchowitz
- abstract: We introduce a new algorithm for online linear-quadratic control in a known system subject to adversarial disturbances. Existing regret bounds for this setting scale as $\sqrt{T}$ unless strong stochastic assumptions are imposed on the disturbance process. We give the first algorithm with logarithmic regret for arbitrary adversarial disturbance sequences, provided the state and control costs are given by known quadratic functions. Our algorithm and analysis use a characterization for the optimal offline control law to reduce the online control problem to (delayed) online learning with approximate advantage functions. Compared to previous techniques, our approach does not need to control movement costs for the iterates, leading to logarithmic regret.

p-Norm Flow Diffusion for Local Graph Clustering

- Kimon Fountoulakis, Di Wang, Shenghao Yang
- abstract: Local graph clustering and the closely related seed set expansion problem are primitives on graphs that are central to a wide range of analytic and learning tasks such as local clustering, community detection, semi-supervised learning, nodes ranking and feature inference. Prior work on local graph clustering mostly falls into two categories with numerical and combinatorial roots respectively, in this work we draw inspiration from both fields and propose a family of convex optimization formulations based on the idea of diffusion with p-norm network flow for $p\in (1,\infty)$. In the context of local clustering, we characterize the optimal solutions for these optimization problems and show their usefulness in finding low conductance cuts around input seed set. In particular, we achieve quadratic approximation of conductance in the case of $p=2$ similar to the Cheeger-type bounds of spectral methods, constant factor approximation when $p\rightarrow\infty$ similar to max-flow based methods, and a smooth transition for general p values in between. Thus, our optimization formulation can be viewed as bridging the numerical and combinatorial approaches, and we can achieve the best of both worlds in terms of speed and noise robustness. We show that the proposed problem can be solved in strongly local running time for $p\ge 2$ and conduct empirical evaluations on both synthetic and real-world graphs to illustrate our approach compares favorably with existing methods.

Stochastic Latent Residual Video Prediction

- Jean-Yves Franceschi, Edouard Delasalles, Mickael Chen, Sylvain Lamprier, Patrick Gallinari
- abstract: Designing video prediction models that account for the inherent uncertainty of the future is challenging. Most works in the literature are based on stochastic image-autoregressive recurrent networks, which raises several performance and applicability issues. An alternative is to use fully latent temporal models which untie frame synthesis and temporal dynamics. However, no such model for stochastic video prediction has been proposed in the literature yet, due to design and training difficulties. In this paper, we overcome these difficulties by introducing a novel stochastic temporal model whose dynamics are governed in a latent space by a residual update rule. This first-order scheme is motivated by discretization schemes of differential equations. It naturally models video dynamics as it allows our simpler, more interpretable, latent model to outperform prior state-of-the-art methods on challenging datasets.

Leveraging Frequency Analysis for Deep Fake Image Recognition

- Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, Thorsten Holz
- abstract: Deep neural networks can generate images that are astonishingly realistic, so much so that it is often hard for humans to distinguish them from actual photos. These achievements have been largely made possible by Generative Adversarial Networks (GANs). While deep fake images have been thoroughly investigated in the image domain{—}a classical approach from the area of image forensics{—}an analysis in the frequency domain has been missing so far. In this paper,we address this shortcoming and our results reveal that in frequency space, GAN-generated images exhibit severe artifacts that can be easily identified. We perform a comprehensive analysis, showing that these artifacts are consistent across different neural network architectures, data sets, and resolutions. In a further investigation, we demonstrate that these artifacts are caused by upsampling operations found in all current GAN architectures, indicating a structural and fundamental problem in the way images are generated via GANs. Based on this analysis, we demonstrate how the frequency representation can be used to identify deep fake images in an automated way, surpassing state-of-the-art methods.

Linear Mode Connectivity and the Lottery Ticket Hypothesis

- Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, Michael Carbin
- abstract: We study whether a neural network optimizes to the same, linearly connected minimum under different samples of SGD noise (e.g., random data order and augmentation). We find that standard vision models become stable to SGD noise in this way early in training. From then on, the outcome of optimization is determined to a linearly connected region. We use this technique to study iterative magnitude pruning (IMP), the procedure used by work on the lottery ticket hypothesis to identify subnetworks that could have trained in isolation to full accuracy. We find that these subnetworks only reach full accuracy when they are stable to SGD noise, which either occurs at initialization for small-scale settings (MNIST) or early in training for large-scale settings (ResNet-50 and Inception-v3 on ImageNet).

No-Regret and Incentive-Compatible Online Learning

- Rupert Freeman, David Pennock, Chara Podimata, Jennifer Wortman Vaughan
- abstract: We study online learning settings in which experts act strategically to maximize their influence on the learning algorithm’s predictions by potentially misreporting their beliefs about a sequence of binary events. Our goal is twofold. First, we want the learning algorithm to be no-regret with respect to the best-fixed expert in hindsight. Second, we want incentive compatibility, a guarantee that each expert’s best strategy is to report his true beliefs about the realization of each event. To achieve this goal, we build on the literature on wagering mechanisms, a type of multi-agent scoring rule. We provide algorithms that achieve no regret and incentive compatibility for myopic experts for both the full and partial information settings. In experiments on datasets from FiveThirtyEight, our algorithms have regret comparable to classic no-regret algorithms, which are not incentive-compatible. Finally, we identify an incentive-compatible algorithm for forward-looking strategic agents that exhibits diminishing regret in practice.

Fast and Three-rious: Speeding Up Weak Supervision with Triplet Methods

- Daniel Fu, Mayee Chen, Frederic Sala, Sarah Hooper, Kayvon Fatahalian, Christopher Re
- abstract: Weak supervision is a popular method for building machine learning models without relying on ground truth annotations. Instead, it generates probabilistic training labels by estimating the accuracies of multiple noisy labeling sources (e.g., heuristics, crowd workers). Existing approaches use latent variable estimation to model the noisy sources, but these methods can be computationally expensive, scaling superlinearly in the data. In this work, we show that, for a class of latent variable models highly applicable to weak supervision, we can find a closed-form solution to model parameters, obviating the need for iterative solutions like stochastic gradient descent (SGD). We use this insight to build FlyingSquid, a weak supervision framework that runs orders of magnitude faster than previous weak supervision approaches and requires fewer assumptions. In particular, we prove bounds on generalization error without assuming that the latent variable model can exactly parameterize the underlying data distribution. Empirically, we validate FlyingSquid on benchmark weak supervision datasets and find that it achieves the same or higher quality compared to previous approaches without the need to tune an SGD procedure, recovers model parameters 170 times faster on average, and enables new video analysis and online learning applications.

AutoGAN-Distiller: Searching to Compress Generative Adversarial Networks

- Yonggan Fu, Wuyang Chen, Haotao Wang, Haoran Li, Yingyan Lin, Zhangyang Wang
- abstract: The compression of Generative Adversarial Networks (GANs) has lately drawn attention, due to the increasing demand for deploying GANs into mobile devices for numerous applications such as image translation, enhancement and editing. However, compared to the substantial efforts to compressing other deep models, the research on compressing GANs (usually the generators) remains at its infancy stage. Existing GAN compression algorithms are limited to handling specific GAN architectures and losses. Inspired by the recent success of AutoML in deep compression, we introduce AutoML to GAN compression and develop an AutoGAN-Distiller (AGD) framework. Starting with a specifically designed efficient search space, AGD performs an end-to-end discovery for new efficient generators, given the target computational resource constraints. The search is guided by the original GAN model via knowledge distillation, therefore fulfilling the compression. AGD is fully automatic, standalone (i.e., needing no trained discriminators), and generically applicable to various GAN models. We evaluate AGD in two representative GAN tasks: image translation and super resolution. Without bells and whistles, AGD yields remarkably lightweight yet more competitive compressed models, that largely outperform existing alternatives. Our codes and pretrained models are available at: https://github.com/TAMU-VITA/AGD.

Don’t Waste Your Bits! Squeeze Activations and Gradients for Deep Neural Networks via TinyScript

- Fangcheng Fu, Yuzheng Hu, Yihan He, Jiawei Jiang, Yingxia Shao, Ce Zhang, Bin Cui
- abstract: Recent years have witnessed intensive research interests on training deep neural networks (DNNs) more efficiently by quantization-based compression methods, which facilitate DNNs training in two ways: (1) activations are quantized to shrink the memory consumption, and (2) gradients are quantized to decrease the communication cost. However, existing methods mostly use a uniform mechanism that quantizes the values evenly. Such a scheme may cause a large quantization variance and slow down the convergence in practice. In this work, we introduce TinyScript, which applies a non-uniform quantization algorithm to both activations and gradients. TinyScript models the original values by a family of Weibull distributions and searches for ”quantization knobs” that minimize quantization variance. We also discuss the convergence of the non-uniform quantization algorithm on DNNs with varying depths, shedding light on the number of bits required for convergence. Experiments show that TinyScript always obtains lower quantization variance, and achieves comparable model qualities against full precision training using 1-2 bits less than the uniform-based counterpart.

DessiLBI: Exploring Structural Sparsity of Deep Networks via Differential Inclusion Paths

- Yanwei Fu, Chen Liu, Donghao Li, Xinwei Sun, Jinshan Zeng, Yuan Yao
- abstract: Over-parameterization is ubiquitous nowadays in training neural networks to benefit both optimization in seeking global optima and generalization in reducing prediction error. However, compressive networks are desired in many real world applications and direct training of small networks may be trapped in local optima. In this paper, instead of pruning or distilling over-parameterized models to compressive ones, we propose a new approach based on differential inclusions of inverse scale spaces. Specifically, it generates a family of models from simple to complex ones that couples a pair of parameters to simultaneously train over-parameterized deep models and structural sparsity on weights of fully connected and convolutional layers. Such a differential inclusion scheme has a simple discretization, proposed as Deep structurally splitting Linearized Bregman Iteration (DessiLBI), whose global convergence analysis in deep learning is established that from any initializations, algorithmic iterations converge to a critical point of empirical risks. Experimental evidence shows that DessiLBI achieve comparable and even better performance than the competitive optimizers in exploring the structural sparsity of several widely used backbones on the benchmark datasets. Remarkably, with early stopping, DessiLBI unveils “winning tickets” in early epochs: the effective sparse structure with comparable test accuracy to fully trained over-parameterized models.

Approximation Guarantees of Local Search Algorithms via Localizability of Set Functions

- Kaito Fujii
- abstract: This paper proposes a new framework for providing approximation guarantees of local search algorithms. Local search is a basic algorithm design technique and is widely used for various combinatorial optimization problems. To analyze local search algorithms for set function maximization, we propose a new notion called \emph{localizability} of set functions, which measures how effective local improvement is. Moreover, we provide approximation guarantees of standard local search algorithms under various combinatorial constraints in terms of localizability. The main application of our framework is sparse optimization, for which we show that restricted strong concavity and restricted smoothness of the objective function imply localizability, and further develop accelerated versions of local search algorithms. We conduct experiments in sparse regression and structure learning of graphical models to confirm the practical efficiency of the proposed local search algorithms.

Accelerating the diffusion-based ensemble sampling by non-reversible dynamics

- Futoshi Futami, Issei Sato, Masashi Sugiyama
- abstract: Posterior distribution approximation is a central task in Bayesian inference. Stochastic gradient Langevin dynamics (SGLD) and its extensions have been practically used and theoretically studied. While SGLD updates a single particle at a time, ensemble methods that update multiple particles simultaneously have been recently gathering attention. Compared with the naive parallel-chain SGLD that updates multiple particles independently, ensemble methods update particles with their interactions. Thus, these methods are expected to be more particle-efficient than the naive parallel-chain SGLD because particles can be aware of other particles’ behavior through their interactions. Although ensemble methods numerically demonstrated their superior performance, no theoretical guarantee exists to assure such particle-efficiency and it is unclear whether those ensemble methods are really superior to the naive parallel-chain SGLD in the non-asymptotic settings. To cope with this problem, we propose a novel ensemble method that uses a non-reversible Markov chain for the interaction, and we present a non-asymptotic theoretical analysis for our method. Our analysis shows that, for the first time, the interaction causes a faster convergence rate than the naive parallel-chain SGLD in the non-asymptotic setting if the discretization error is appropriately controlled. Numerical experiments show that we can control the discretization error by tuning the interaction appropriately.

Stochastic bandits with arm-dependent delays

- Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, Michal Valko
- abstract: Significant work has been recently dedicated to the stochastic delayed bandits because of its relevance in applications. The applicability of existing algorithms is however restricted by the fact that strong assumptions are often made on the delay distributions, such as full observability, restrictive shape constraints, or uniformity over arms. In this work, we weaken them significantly and only assume that there is a bound on the tail of the delay. In particular, we cover the important case where the delay distributions vary across arms, and the case where the delays are heavy-tailed. Addressing these difficulties, we propose a simple but efficient UCB-based algorithm called the PatientBandits. We provide both problemsdependent and problems-independent bounds on the regret as well as performance lower bounds.

Abstraction Mechanisms Predict Generalization in Deep Neural Networks

- Alex Gain, Hava Siegelmann
- abstract: A longstanding problem for Deep Neural Networks (DNNs) is understanding their puzzling ability to generalize well. We approach this problem through the unconventional angle of \emph{cognitive abstraction mechanisms}, drawing inspiration from recent neuroscience work, allowing us to define the Cognitive Neural Activation metric (CNA) for DNNs, which is the correlation between information complexity (entropy) of given input and the

concentration of higher activation values in deeper layers of the network. The CNA is highly predictive of generalization ability, outperforming norm-and-sharpness-based generalization metrics on an extensive evaluation of close to 200 network instances comprising a breadth of dataset-architecture combinations, especially in cases where additive noise is present and/or training labels are corrupted. These strong empirical results show the usefulness of the CNA as a generalization metric and encourage further research on the connection between information complexity and representations in the deeper layers of networks in order to better understand the generalization capabilities of DNNs.

A Free-Energy Principle for Representation Learning

- Yansong Gao, Pratik Chaudhari
- abstract: This paper employs a formal connection of machine learning with thermodynamics to characterize the quality of learnt representations for transfer learning. We discuss how information-theoretic functionals such as rate, distortion and classification loss of a model lie on a convex, so-called equilibrium surface. We prescribe dynamical processes to traverse this surface under constraints, e.g., an iso-classification process that trades off rate and distortion to keep the classification loss unchanged. We demonstrate how this process can be used for transferring representations from a source dataset to a target dataset while keeping the classification loss constant. Experimental validation of the theoretical results is provided on standard image-classification datasets.

Can Stochastic Zeroth-Order Frank-Wolfe Method Converge Faster for Non-Convex Problems?

- Hongchang Gao, Heng Huang
- abstract: Frank-Wolfe algorithm is an efficient method for optimizing non-convex constrained problems. However, most of existing methods focus on the first-order case. In real-world applications, the gradient is not always available. To address the problem of lacking gradient in many applications, we propose two new stochastic zeroth-order Frank-Wolfe algorithms and theoretically proved that they have a faster convergence rate than existing methods for non-convex problems. Specifically, the function queries oracle of the proposed faster zeroth-order Frank-Wolfe (FZFW) method is $O(\frac{n^{1/2}d}{\epsilon^2})$ which can match the iteration complexity of the first-order counterpart approximately. As for the proposed faster zeroth-order conditional gradient sliding (FZCGS) method, its function queries oracle is improved to $O(\frac{n^{1/2}d}{\epsilon})$, indicating that its iteration complexity is even better than that of its first-order counterpart NCGS-VR. In other words, the iteration complelxity of the accelerated first-order Frank-Wolfe method NCGS-VR is suboptimal. Then, we proposed a new algorithm to improve its IFO (incremental first-order oracle) to $O(\frac{n^{1/2}}{\epsilon})$. At last, the empirical studies on benchmark datasets validate our theoretical results.

Online Convex Optimization in the Random Order Model

- Dan Garber, Gal Korcia, Kfir Levy
- abstract: Online Convex Optimization (OCO) is a powerful framework for sequential prediction, portraying the natural uncertainty inherent in data-streams as though the data were generated by an almost omniscient adversary. However, this view, which is often too pessimistic for real-world data, comes with a price. The complexity of solving many important online tasks in this adversarial framework becomes much worse than that of their offline and even stochastic counterparts. In this work we consider a natural random-order version of the OCO model, in which the adversary can choose the set of loss functions, but does not get to choose the order in which they are supplied to the learner; Instead, they are observed in uniformly random order. Focusing on two important families of online tasks, one in which the cumulative loss function is strongly convex (though individual loss functions may not even be convex), and the other being online k-PCA, we show that under standard well-conditioned-data assumptions, standard online gradient descent (OGD) methods become much more efficient in the random-order model. In particular, for the first group of tasks OGD guarantees poly-logarithmic regret. In the case of online k-PCA, OGD guarantees sublinear regret using only a rank-k SVD on each iteration and memory linear in the size of the solution.

Symbolic Network: Generalized Neural Policies for Relational MDPs

- Sankalp Garg, Aniket Bajpai, Mausam
- abstract: A Relational Markov Decision Process (RMDP) is a first-order representation to express all instances of a single probabilistic planning domain with possibly unbounded number of objects. Early work in RMDPs outputs generalized (instance-independent) first-order policies or value functions as a means to solve all instances of a domain at once. Unfortunately, this line of work met with limited success due to inherent limitations of the representation space used in such policies or value functions. Can neural models provide the missing link by easily representing more complex generalized policies, thus making them effective on all instances of a given domain? We present SymNet, the first neural approach for solving RMDPs that are expressed in the probabilistic planning language of RDDL. SymNet trains a set of shared parameters for an RDDL domain using training instances from that domain. For each instance, SymNet first converts it to an instance graph and then uses relational neural models to compute node embeddings. It then scores each ground action as a function over the first-order action symbols and node embeddings related to the action. Given a new test instance from the same domain, SymNet architecture with pre-trained parameters scores each ground action and chooses the best action. This can be accomplished in a single forward pass without any retraining on the test instance, thus implicitly representing a neural generalized policy for the whole domain. Our experiments on nine RDDL domains from IPPC demonstrate that SymNet policies are significantly better than random and sometimes even more effective than training a state-of-the-art deep reactive policy from scratch.

Predicting deliberative outcomes

- Vikas Garg, Tommi Jaakkola
- abstract: We extend structured prediction to deliberative outcomes. Specifically, we learn parameterized games that can map any inputs to equilibria as the outcomes. Standard structured prediction models rely heavily on global scoring functions and are therefore unable to model individual player preferences or how they respond to others asymmetrically. Our games take as input, e.g., UN resolution to be voted on, and map such contexts to initial strategies, player utilities, and interactions. Players are then thought to repeatedly update their strategies in response to weighted aggregates of other players' choices towards maximizing their individual utilities. The output from the game is a sample from the resulting (near) equilibrium mixed strategy profile. We characterize conditions under which players' strategies converge to an equilibrium in such games and when the game parameters can be provably recovered from observations. Empirically, we demonstrate on two real voting datasets that our games can recover interpretable strategic interactions, and predict strategies for players in new settings.

Generalization and Representational Limits of Graph Neural Networks

- Vikas Garg, Stefanie Jegelka, Tommi Jaakkola
- abstract: We address two fundamental questions about graph neural networks (GNNs). First, we prove that several important graph properties, e.g., shortest/longest cycle, diameter, or certain motifs, cannot be computed by GNNs that rely entirely on local information. Such GNNs include the standard message passing models, and more powerful spatial variants that exploit local graph structure (e.g., via relative orientation of messages, or local port ordering) to distinguish neighbors of each node. Our treatment includes a novel graph-theoretic formalism. Second, we provide the first data dependent generalization bounds for message passing GNNs. This analysis explicitly accounts for the local permutation invariance of GNNs. Our bounds are much tighter than existing VC-dimension based guarantees for GNNs, and are comparable to Rademacher bounds for recurrent neural networks.

Deep PQR: Solving Inverse Reinforcement Learning using Anchor Actions

- Sinong Geng, Houssam Nassif, Carlos Manzanares, Max Reppen, Ronnie Sircar
- abstract: We propose a reward function estimation framework for inverse reinforcement learning with deep energy-based policies. We name our method PQR, as it sequentially estimates the Policy, the Q-function, and the Reward function by deep learning. PQR does not assume that the reward solely depends on the state, instead it allows for a dependency on the choice of action. Moreover, PQR allows for stochastic state transitions. To accomplish this, we assume the existence of one anchor action whose reward is known, typically the action of doing nothing, yielding no reward. We present both estimators and algorithms for the PQR method. When the environment transition is known, we prove that the PQR reward estimator uniquely recovers the true reward. With unknown transitions, we bound the estimation error of PQR. Finally, the performance of PQR is demonstrated by synthetic and real-world datasets.

Multilinear Latent Conditioning for Generating Unseen Attribute Combinations

- Markos Georgopoulos, Grigorios Chrysos, Maja Pantic, Yannis Panagakis
- abstract: Deep generative models rely on their inductive bias to facilitate generalization, especially for problems with high dimensional data, like images. However, empirical studies have shown that variational autoencoders (VAE) and generative adversarial networks (GAN) lack the generalization ability that occurs naturally in human perception. For example, humans can visualize a woman smiling after only seeing a smiling man. On the contrary, the standard conditional VAE (cVAE) is unable to generate unseen attribute combinations. To this end, we extend cVAE by introducing a multilinear latent conditioning framework that captures the multiplicative interactions between the attributes. We implement two variants of our model and demonstrate their efficacy on MNIST, Fashion-MNIST and CelebA. Altogether, we design a novel conditioning framework that can be used with any architecture to synthesize unseen attribute combinations.

Generalisation error in learning with random features and the hidden manifold model

- Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, Lenka Zdeborova
- abstract: We study generalised linear regression and classification for a synthetically generated dataset encompassing different problems of interest, such as learning with random features, neural networks in the lazy training regime, and the hidden manifold model. We consider the high-dimensional regime and using the replica method from statistical physics, we provide a closed-form expression for the asymptotic generalisation performance in these problems, valid in both the under- and over-parametrised regimes and for a broad choice of generalised linear model loss functions. In particular, we show how to obtain analytically the so-called double descent behaviour for logistic regression with a peak at the interpolation threshold, we illustrate the superiority of orthogonal against random Gaussian projections in learning with random features, and discuss the role played by correlations in the data generated by the hidden manifold model. Beyond the interest in these particular problems, the theoretical formalism introduced in this manuscript provides a path to further extensions to more complex tasks.

Black-Box Methods for Restoring Monotonicity

- Evangelia Gergatsouli, Brendan Lucier, Christos Tzamos
- abstract: In many practical applications, heuristic or approximation algorithms are used to efficiently solve the task at hand. However their solutions frequently do not satisfy natural monotonicity properties expected to hold in the optimum. In this work we develop algorithms that are able to restore monotonicity in the parameters of interest. Specifically, given oracle access to a possibly non monotone function, we provide an algorithm that restores monotonicity while degrading the expected value of the function by at most ϵ. The number of queries required is at most logarithmic in $1/\epsilon$ and exponential in the number of parameters. We also give a lower bound showing that this exponential dependence is necessary. Finally, we obtain improved query complexity bounds for restoring the weaker property of k-marginal monotonicity. Under this property, every k-dimensional projection of the function is required to be monotone. The query complexity we obtain only scales exponentially with k and is polynomial in the number of parameters.

Online Multi-Kernel Learning with Graph-Structured Feedback

- Pouya M Ghari, Yanning Shen
- abstract: Multi-kernel learning (MKL) exhibits reliable performance in nonlinear function approximation tasks. Instead of using one kernel, it learns the optimal kernel from a pre-selected dictionary of kernels. The selection of the dictionary has crucial impact on both the performance and complexity of MKL. Specifically, inclusion of a large number of irrelevant kernels may impair the accuracy, and increase the complexity of MKL algorithms. To enhance the accuracy, and alleviate the computational burden, the present paper develops a novel scheme which actively chooses relevant kernels. The proposed framework models the pruned kernel combination as feedback collected from a graph, that is refined 'on the fly.' Leveraging the random feature approximation, we propose an online scalable multi-kernel learning approach with graph feedback, and prove that the proposed algorithm enjoys sublinear regret. Numerical tests on real datasets demonstrate the effectiveness of the novel approach.

Task-Oriented Active Perception and Planning in Environments with Partially Known Semantics

- Mahsa Ghasemi, Erdem Bulgur, Ufuk Topcu
- abstract: We consider an agent that is assigned with a temporal logic task in an environment whose semantic representation is only partially known. We represent the semantics of the environment with a set of state properties, called \emph{atomic propositions} over which, the agent holds a probabilistic belief and updates it as new sensory measurements arrive. The goal is to design a joint perception and planning strategy for the agent that realizes the task with high probability. We develop a planning strategy that takes the semantic uncertainties into account and by doing so provides probabilistic guarantees on the task success. Furthermore, as new data arrive, the belief over the atomic propositions evolves and, subsequently, the planning strategy adapts accordingly. We evaluate the proposed method on various finite-horizon tasks in planar navigation settings where the empirical results show that the proposed method provides reliable task performance that also improves as the knowledge about the environment enhances.

Characterizing Distribution Equivalence and Structure Learning for Cyclic and Acyclic Directed Graphs

- Amiremad Ghassami, Alan Yang, Negar Kiyavash, Kun Zhang
- abstract: The main approach to defining equivalence among acyclic directed causal graphical models is based on the conditional independence relationships in the distributions that the causal models can generate, in terms of the Markov equivalence. However, it is known that when cycles are allowed in the causal structure, conditional independence may not be a suitable notion for equivalence of two structures, as it does not reflect all the information in the distribution that is useful for identification of the underlying structure. In this paper, we present a general, unified notion of equivalence for linear Gaussian causal directed graphical models, whether they are cyclic or acyclic. In our proposed definition of equivalence, two structures are equivalent if they can generate the same set of data distributions. We also propose a weaker notion of equivalence called quasi-equivalence, which we show is the extent of identifiability from observational data. We propose analytic as well as graphical methods for characterizing the equivalence of two structures. Additionally, we propose a score-based method for learning the structure from observational data, which successfully deals with both acyclic and cyclic structures.

Private Counting from Anonymous Messages: Near-Optimal Accuracy with Vanishing Communication Overhead

- Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Rasmus Pagh
- abstract: Differential privacy (DP) is a formal notion for quantifying the privacy loss of algorithms. Algorithms in the central model of DP achieve high accuracy but make the strongest trust assumptions whereas those in the local DP model make the weakest trust assumptions but incur substantial accuracy loss. The shuffled DP model [Bittau et al 2017, Erlingsson et al 2019, Cheu et al 19] has recently emerged as a feasible middle ground between the central and local models, providing stronger trust assumptions than the former while promising higher accuracies than the latter. In this paper, we obtain practical communication-efficient algorithms in the shuffled DP model for two basic aggregation primitives used in machine learning: 1) binary summation, and 2) histograms over a moderate number of buckets. Our algorithms achieve accuracy that is arbitrarily close to that of central DP algorithms with an expected communication per user essentially matching what is needed without any privacy constraints! We demonstrate the practicality of our algorithms by experimentally evaluating them and comparing their performance to several widely-used protocols such as Randomized Response [Warner 1965] and RAPPOR [Erlingsson et al. 2014].

Aligned Cross Entropy for Non-Autoregressive Machine Translation

- Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, Omer Levy
- abstract: Non-autoregressive machine translation models significantly speed up decoding by allowing for parallel prediction of the entire target sequence. However, modeling word order is more challenging due to the lack of autoregressive factors in the model. This difficultly is compounded during training with cross entropy loss, which can highly penalize small shifts in word order. In this paper, we propose aligned cross entropy (AXE) as an alternative loss function for training of non-autoregressive models. AXE uses a differentiable dynamic program to assign loss based on the best possible monotonic alignment between target tokens and model predictions. AXE-based training of conditional masked language models (CMLMs) substantially improves performance on major WMT benchmarks, while setting a new state of the art for non-autoregressive models.

Gradient Temporal-Difference Learning with Regularized Corrections

- Sina Ghiassian, Andrew Patterson, Shivam Garg, Dhawal Gupta, Adam White, Martha White
- abstract: It is still common to use Q-learning and temporal difference (TD) learning{—}even though they have divergence issues and sound Gradient TD alternatives exist{—}because divergence seems rare and they typically perform well. However, recent work with large neural network learning systems reveals that instability is more common than previously thought. Practitioners face a difficult dilemma: choose an easy to use and performant TD method, or a more complex algorithm that is more sound but harder to tune and all but unexplored with non-linear function approximation or control. In this paper, we introduce a new method called TD with Regularized Corrections (TDRC), that attempts to balance ease of use, soundness, and performance. It behaves as well as TD, when TD performs well, but is sound in cases where TD diverges. We empirically investigate TDRC across a range of problems, for both prediction and control, and for both linear and non-linear function approximation, and show, potentially for the first time, that Gradient TD methods could be a better alternative to TD and Q-learning.

A Distributional Framework For Data Valuation

- Amirata Ghorbani, Michael Kim, James Zou
- abstract: Shapley value is a classic notion from game theory, historically used to quantify the contributions of individuals within groups, and more recently applied to assign values to data points when training machine learning models. Despite its foundational role, a key limitation of the data Shapley framework is that it only provides valuations for points within a fixed data set. It does not account for statistical aspects of the data and does not give a way to reason about points outside the data set. To address these limitations, we propose a novel framework – distributional Shapley– where the value of a point is defined in the context of an underlying data distribution. We prove that distributional Shapley has several desirable statistical properties; for example, the values are stable under perturbations to the data points themselves and to the underlying data distribution. We leverage these properties to develop a new algorithm for estimating values from data, which comes with formal guarantees and runs two orders of magnitude faster than state-of-the-art algorithms for computing the (non distributional) data Shapley values. We apply distributional Shapley to diverse data sets and demonstrate its utility in a data market setting.

Fractal Gaussian Networks: A sparse random graph model based on Gaussian Multiplicative Chaos

- Subhroshekhar Ghosh, Krishna Balasubramanian, Xiaochuan Yang
- abstract: We propose a novel stochastic network model, called Fractal Gaussian Network (FGN), that embodies well-defined and analytically tractable fractal structures. Such fractal structures have been empirically observed in diverse applications. FGNs interpolate continuously between the popular purely random geometric graphs (a.k.a. the Poisson Boolean network), and random graphs with increasingly fractal behavior. In fact, they form a parametric family of sparse random geometric graphs that are parametrised by a fractality parameter ν which governs the strength of the fractal structure. FGNs are driven by the latent spatial geometry of Gaussian Multiplicative Chaos (GMC), a canonical model of fractality in its own right. We explore the natural question of detecting the presence of fractality and the problem of parameter estimation based on observed network data. Finally, we explore fractality in community structures by unveiling a natural stochastic block model in the setting of FGNs.

Representations for Stable Off-Policy Reinforcement Learning

- Dibya Ghosh, Marc G. Bellemare
- abstract: Reinforcement learning with function approximation can be unstable and even divergent, especially when combined with off-policy learning and Bellman updates. In deep reinforcement learning, these issues have been dealt with empirically by adapting and regularizing the representation, in particular with auxiliary tasks. This suggests that representation learning may provide a means to guarantee stability. In this paper, we formally show that there are indeed nontrivial state representations under which the canonical SARSA algorithm is stable, even when learning off-policy. We analyze representation learning schemes that are based on the transition matrix of a policy, such as proto-value functions, along three axes: approximation error, stability, and ease of estimation. In the most general case of a defective transition matrix, we show that a Schur basis provides convergence guarantees, but is difficult to estimate from samples. For a fixed reward function, we find that an orthogonal basis of the corresponding Krylov subspace is an even better choice. We conclude by empirically demonstrating that these stable representations can be learned using stochastic gradient descent, opening the door to improved techniques for representation learning with deep networks.

Adaptive Sketching for Fast and Convergent Canonical Polyadic Decomposition

- Alex Gittens, Kareem Aggour, Bülent Yener
- abstract: This work considers the canonical polyadic decomposition (CPD) of tensors using proximally regularized sketched alternating least squares algorithms. First, it establishes a sublinear rate of convergence for proximally regularized sketched CPD algorithms under two natural conditions that are known to be satisfied by many popular forms of sketching. Second, it demonstrates that the iterative nature of CPD algorithms can be exploited algorithmically to choose more performant sketching rates. This is accomplished by introducing CPD-MWU, a proximally-regularized sketched alternating least squares algorithm that adaptively selects the sketching rate at each iteration. On both synthetic and real data we observe that for noisy tensors CPD-MWU produces decompositions of comparable accuracy to the standard CPD decomposition in less time, often half the time; for ill-

conditioned tensors, given the same time budget, CPD-MWU produces decompositions with an order-of-magnitude lower relative error. For a representative real-world dataset CPD-MWU produces residual errors on average 20% lower than CPRAND-MIX and 44% lower than SPALS, two recent sketched CPD algorithms.

One Size Fits All: Can We Train One Denoiser for All Noise Levels?

- Abhiram Gnanasambandam, Stanley Chan
- abstract: When training an estimator such as a neural network for tasks like image denoising, it is often preferred to train one estimator and apply it to all noise levels. The de facto training protocol to achieve this goal is to train the estimator with noisy samples whose noise levels are uniformly distributed across the range of interest. However, why should we allocate the samples uniformly? Can we have more training samples that are less noisy, and fewer samples that are more noisy? What is the optimal distribution? How do we obtain such a distribution? The goal of this paper is to address this training sample distribution problem from a minimax risk optimization perspective. We derive a dual ascent algorithm to determine the optimal sampling distribution of which the convergence is guaranteed as long as the set of admissible estimators is closed and convex. For estimators with non-convex admissible sets such as deep neural networks, our dual formulation converges to a solution of the convex relaxation. We discuss how the algorithm can be implemented in practice. We evaluate the algorithm on linear estimators and deep networks.

Superpolynomial Lower Bounds for Learning One-Layer Neural Networks using Gradient Descent

- Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, Adam Klivans
- abstract: We give the first superpolynomial lower bounds for learning one-layer neural networks with respect to the Gaussian distribution for a broad class of algorithms. In the regression setting, we prove that gradient descent run on any classifier with respect to square loss will fail to achieve small test error in polynomial time. Prior work held only for gradient descent run with small batch sizes and sufficiently smooth classifiers. For classification, we give a stronger result, namely that any statistical query (SQ) algorithm will fail to achieve small test error in polynomial time. Our lower bounds hold for commonly used activations such as ReLU and sigmoid. The core of our result relies on a novel construction of a simple family of neural networks that are exactly orthogonal with respect to all spherically symmetric distributions.

SimGANs: Simulator-Based Generative Adversarial Networks for ECG Synthesis to Improve Deep ECG Classification

- Tomer Golany, Kira Radinsky, Daniel Freedman
- abstract: Generating training examples for supervised tasks is a long sought after goal in AI. We study the problem of heart signal electrocardiogram (ECG) synthesis for improved heartbeat classification. ECG synthesis is challenging: the generation of training examples for such biological-physiological systems is not straightforward, due to their dynamic nature in which the various parts of the system interact in complex ways. However, an understanding of these dynamics has been developed for years in the form of mathematical process simulators. We study how to incorporate this knowledge into the generative process by leveraging a biological simulator for the task of ECG classification. Specifically, we use a system of ordinary differential equations representing heart dynamics, and incorporate this ODE system into the optimization process of a generative adversarial network to create biologically plausible ECG training examples. We perform empirical evaluation and show that heart simulation knowledge during the generation process improves ECG classification.

Unraveling Meta-Learning: Understanding Feature Representations for Few-Shot Tasks

- Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, Tom Goldstein
- abstract: Meta-learning algorithms produce feature extractors which achieve state-of-the-art performance on few-shot classification. While the literature is rich with meta-learning methods, little is known about why the resulting feature extractors perform so well. We develop a better understanding of the underlying mechanics of meta-learning and the difference between models trained using meta-learning and models which are trained classically. In doing so, we introduce and verify several hypotheses for why meta-learned models perform better. Furthermore, we develop a regularizer which boosts the performance of standard training routines for few-shot classification. In many cases, our routine outperforms meta-learning while simultaneously running an order of magnitude faster.

Towards a General Theory of Infinite-Width Limits of Neural Classifiers

- Eugene Golikov
- abstract: Obtaining theoretical guarantees for neural networks training appears to be a hard problem in a general case. Recent research has been focused on studying this problem in the limit of infinite width and two different theories have been developed: a mean-field (MF) and a constant kernel (NTK) limit theories. We propose a general framework that provides a link between these seemingly distinct theories. Our framework out of the box gives rise to a discrete-time MF limit which was not previously explored in the literature. We prove a convergence theorem for it, and show that it provides a more reasonable approximation for finite-width nets compared to the NTK limit if learning rates are not very small. Also, our framework suggests a limit model that coincides neither with the MF limit nor with the NTK one. We show that for networks with more than two hidden layers RMSProp training has a non-trivial discrete-time MF limit but GD training does not have one. Overall, our framework demonstrates that both MF and NTK limits have considerable limitations in approximating finite-sized neural nets, indicating the need for designing more accurate infinite-width approximations for them.

Differentially Private Set Union

- Sivakanth Gopi, Pankaj Gulhane, Janardhan Kulkarni, Judy Hanwen Shen, Milad Shokouhi, Sergey Yekhanin
- abstract: We study the basic operation of set union in the global model of differential privacy. In this problem, we are given a universe U of items, possibly of infinite size, and a database D of users. Each user i contributes a subset $W_i \subseteq U$ of items. We want an (ϵ,δ)-differentially private Algorithm which outputs a subset $S \subset \cup_i W_i$ such that the size of S is as large as possible. The problem arises in countless real world applications, and is particularly ubiquitous in natural language processing (NLP) applications. For example, discovering words, sentences, n-grams etc., from private text data belonging to users is an instance of the set union problem. In this paper we design new algorithms for this problem that significantly outperform the best known algorithms.

The continuous categorical: a novel simplex-valued exponential family

- Elliott Gordon-Rodriguez, Gabriel Loaiza-Ganem, John Cunningham
- abstract: Simplex-valued data appear throughout statistics and machine learning, for example in the context of transfer learning and compression of deep networks. Existing models for this class of data rely on the Dirichlet distribution or other related loss functions; here we show these standard choices suffer systematically from a number of limitations, including bias and numerical issues that frustrate the use of flexible network models upstream of these distributions. We resolve these limitations by introducing a novel exponential family of distributions for modeling simplex-valued data {–} the continuous categorical, which arises as a nontrivial multivariate generalization of the recently discovered continuous Bernoulli. Unlike the Dirichlet and other typical choices, the continuous categorical results in a well-behaved probabilistic loss function that produces unbiased estimators, while preserving the mathematical simplicity of the Dirichlet. As well as exploring its theoretical properties, we introduce sampling methods for this distribution that are

amenable to the reparameterization trick, and evaluate their performance. Lastly, we demonstrate that the continuous categorical outperforms standard choices empirically, across a simulation study, an applied example on multi-party elections, and a neural network compression task.

Automatic Reparameterisation of Probabilistic Programs

- Maria Gorinova, Dave Moore, Matthew Hoffman
- abstract: Probabilistic programming has emerged as a powerful paradigm in statistics, applied science, and machine learning: by decoupling modelling from inference, it promises to allow modellers to directly reason about the processes generating data. However, the performance of inference algorithms can be dramatically affected by the parameterisation used to express a model, requiring users to transform their programs in non-intuitive ways. We argue for automating these transformations, and demonstrate that mechanisms available in recent modelling frameworks can implement non-centring and related reparameterisations. This enables new inference algorithms, and we propose two: a simple approach using interleaved sampling and a novel variational formulation that searches over a continuous space of parameterisations. We show that these approaches enable robust inference across a range of models, and can yield more efficient samplers than the best fixed parameterisation.

Interpretable Off-Policy Evaluation in Reinforcement Learning by Highlighting Influential Transitions

- Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Celi, Emma Brunskill, Finale Doshi-Velez
- abstract: Off-policy evaluation in reinforcement learning offers the chance of using observational data to improve future outcomes in domains such as healthcare and education, but safe deployment in high stakes settings requires ways of assessing its validity. Traditional measures such as confidence intervals may be insufficient due to noise, limited data and confounding. In this paper we develop a method that could serve as a hybrid human-AI system, to enable human experts to analyze the validity of policy evaluation estimates. This is accomplished by highlighting observations in the data whose removal will have a large effect on the OPE estimate, and formulating a set of rules for choosing which ones to present to domain experts for validation. We develop methods to compute exactly the influence functions for fitted Q-evaluation with two different function classes: kernel-based and linear least squares, as well as importance sampling methods. Experiments on medical simulations and real-world intensive care unit data demonstrate that our method can be used to identify limitations in the evaluation process and make evaluation more robust.

Learning to Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning

- Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Shengchao Liu, Simon Blackburn, Karam Thomas, Connor Coley, Jian Tang, Sarath Chandar, Yoshua Bengio
- abstract: Over the last decade, there has been significant progress in the field of machine learning for de novo drug design, particularly in generative modeling of novel chemical structures. However, current generative approaches exhibit a significant challenge: they do not ensure that the proposed molecular structures can be feasibly synthesized nor do they provide the synthesis routes of the proposed small molecules, thereby seriously limiting their practical applicability. In this work, we propose a novel reinforcement learning (RL) setup for de novo drug design: Policy Gradient for Forward Synthesis (PGFS), that addresses this challenge by embedding the concept of synthetic accessibility directly into the de novo drug design system. In this setup, the agent learns to navigate through the immense synthetically accessible chemical space by subjecting initial commercially available molecules to valid chemical reactions at every time step of the iterative virtual synthesis process. The proposed environment for drug discovery provides a highly challenging test-bed for RL algorithms owing to the large state space and high-dimensional continuous action space with hierarchical actions. PGFS achieves state-of-the-art performance in generating structures with high QED and clogP. Moreover, we validate PGFS in an in-silico proof-of-concept associated with three HIV targets. Finally, we describe how the end-to-end training conceptualized in this study represents an important paradigm in radically expanding the synthesizable chemical space and automating the drug discovery process.

Ordinal Non-negative Matrix Factorization for Recommendation

- Olivier Gouvert, Thomas Oberlin, Cédric Févotte
- abstract: We introduce a new non-negative matrix factorization (NMF) method for ordinal data, called OrdNMF. Ordinal data are categorical data which exhibit a natural ordering between the categories. In particular, they can be found in recommender systems, either with explicit data (such as ratings) or implicit data (such as quantized play counts). OrdNMF is a probabilistic latent factor model that generalizes Bernoulli-Poisson factorization (BePoF) and Poisson factorization (PF) applied to binarized data. Contrary to these methods, OrdNMF circumvents binarization and can exploit a more informative representation of the data. We design an efficient variational algorithm based on a suitable model augmentation and related to variational PF. In particular, our algorithm preserves the scalability of PF and can be applied to huge sparse datasets. We report recommendation experiments on explicit and implicit datasets, and show that OrdNMF outperforms BePoF and PF applied to binarized data.

PoWER-BERT: Accelerating BERT Inference via Progressive Word-vector Elimination

- Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, Ashish Verma
- abstract: We develop a novel method, called PoWER-BERT, for improving the inference time of the popular BERT model, while maintaining the accuracy. It works by: a) exploiting redundancy pertaining to word-vectors (intermediate transformer block outputs) and eliminating the redundant vectors. b) determining which word-vectors to eliminate by developing a strategy for measuring their significance, based on the self-attention mechanism. c) learning how many word-vectors to eliminate by augmenting the BERT model and the loss function. Experiments on the standard GLUE benchmark shows that PoWER-BERT achieves up to 4.5x reduction in inference time over BERT with < 1% loss in accuracy. We show that PoWER-BERT offers significantly better trade-off between accuracy and inference time compared to prior methods. We demonstrate that our method attains up to 6.8x reduction in inference time with < 1% loss in accuracy when applied over ALBERT, a highly compressed version of BERT. The code for PoWER-BERT is publicly available at https://github.com/IBM/PoWER-BERT.

PackIt: A Virtual Environment for Geometric Planning

- Ankit Goyal, Jia Deng
- abstract: The ability to jointly understand the geometry of objects and plan actions for manipulating them is crucial for intelligent agents. We refer to this ability as geometric planning. Recently, many interactive environments have been proposed to evaluate intelligent agents on various skills, however, none of them cater to the needs of geometric planning. We present PackIt, a virtual environment to evaluate and potentially learn the ability to do geometric planning, where an agent needs to take a sequence of actions to pack a set of objects into a box with limited space. We also construct a set of challenging packing tasks using an evolutionary algorithm. Further, we study various baselines for the task that include model-free learning-based and heuristic-based methods, as well as search-based optimization methods that assume access to the model of the environment.

DROCC: Deep Robust One-Class Classification

- Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, Prateek Jain
- abstract: Classical approaches for one-class problems such as one-class SVM and isolation forest require careful feature engineering when applied to structured domains like images. State-of-the-art methods aim to leverage deep learning to learn appropriate features via two main approaches. The first approach based on predicting transformations (Golan & El-Yaniv, 2018; Hendrycks et al., 2019a) while successful in some domains, crucially depends on

an appropriate domain-specific set of transformations that are hard to obtain in general. The second approach of minimizing a classical one-class loss on the learned final layer representations, e.g., DeepSVDD (Ruff et al., 2018) suffers from the fundamental drawback of representation collapse. In this work, we propose Deep Robust One Class Classification (DROCC) that is both applicable to most standard domains without requiring any side-information and robust to representation collapse. DROCC is based on the assumption that the points from the class of interest lie on a well-sampled, locally linear low dimensional manifold. Empirical evaluation demonstrates that DROCC is highly effective in two different one-class problem settings and on a range of real-world datasets across different domains: tabular data, images (CIFAR and ImageNet), audio, and time-series, offering up to 20% increase in accuracy over the state-of-the-art in anomaly detection. Code is available at https://github.com/microsoft/EdgeML

Scalable Gaussian Process Separation for Kernels with a Non-Stationary Phase

- Jan Graßhoff, Alexandra Jankowski, Philipp Rostalski
- abstract: The application of Gaussian processes (GPs) to large data sets is limited due to heavy memory and computational requirements. A variety of methods has been proposed to enable scalability, one of which is to exploit structure in the kernel matrix. Previous methods, however, cannot easily deal with mixtures of non-stationary processes. This paper investigates an efficient GP framework, that extends structured kernel interpolation methods to GPs with a non-stationary phase. We particularly treat the separation of nonstationary sources, which is a problem that commonly arises e.g. in spatio-temporal biomedical datasets. Our approach employs multiple sets of non-equidistant inducing points to account for the non-stationarity and retrieve Toeplitz and Kronecker structure in the kernel matrix allowing for efficient inference and kernel learning. Our approach is demonstrated on numerical examples and large spatio-temporal biomedical problems.

Learning the Stein Discrepancy for Training and Evaluating Energy-Based Models without Sampling

- Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Richard Zemel
- abstract: We present a new method for evaluating and training unnormalized density models. Our approach only requires access to the gradient of the unnormalized model's log-density. We estimate the Stein discrepancy between the data density p(x) and the model density q(x) based on a vector function of the data. We parameterize this function with a neural network and fit its parameters to maximize this discrepancy. This yields a novel goodness-of-fit test which outperforms existing methods on high dimensional data. Furthermore, optimizing q(x) to minimize this discrepancy produces a novel method for training unnormalized models. This training method can fit large unnormalized models faster than existing approaches. The ability to both learn and compare models is a unique feature of the proposed method.

On the Iteration Complexity of Hypergradient Computation

- Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, Saverio Salzo
- abstract: We study a general class of bilevel problems, consisting in the minimization of an upper-level objective which depends on the solution to a parametric fixed-point equation. Important instances arising in machine learning include hyperparameter optimization, meta-learning, and certain graph and recurrent neural networks. Typically the gradient of the upper-level objective (hypergradient) is hard or even impossible to compute exactly, which has raised the interest in approximation methods. We investigate some popular approaches to compute the hypergradient, based on reverse mode iterative differentiation and approximate implicit differentiation. Under the hypothesis that the fixed point equation is defined by a contraction mapping, we present a unified analysis which allows for the first time to quantitatively compare these methods, providing explicit bounds for their iteration complexity. This analysis suggests a hierarchy in terms of computational efficiency among the above methods, with approximate implicit differentiation based on conjugate gradient performing best. We present an extensive experimental comparison among the methods which confirm the theoretical findings.

Robust Learning with the Hilbert-Schmidt Independence Criterion

- Daniel Greenfeld, Uri Shalit
- abstract: We investigate the use of a non-parametric independence measure, the Hilbert-Schmidt Independence Criterion (HSIC), as a loss-function for learning robust regression and classification models. This loss-function encourages learning models where the distribution of the residuals between the label and the model prediction is statistically independent of the distribution of the instances themselves. This loss-function was first proposed by \citet{mooij2009regression} in the context of learning causal graphs. We adapt it to the task of learning for unsupervised covariate shift: learning on a source domain without access to any instances or labels from the unknown target domain, but with the assumption that $p(y|x)$ (the conditional probability of labels given instances) remains the same in the target domain. We show that the proposed loss is expected to give rise to models that generalize well on a class of target domains characterised by the complexity of their description within a reproducing kernel Hilbert space. Experiments on unsupervised covariate shift tasks demonstrate that models learned with the proposed loss-function outperform models learned with standard loss functions, achieving state-of-the-art results on a challenging cell-microscopy unsupervised covariate shift task.

Monte-Carlo Tree Search as Regularized Policy Optimization

- Jean-Bastien Grill, Florent Altché, Yunhao Tang, Thomas Hubert, Michal Valko, Ioannis Antonoglou, Remi Munos
- abstract: The combination of Monte-Carlo tree search (MCTS) with deep reinforcement learning has led to groundbreaking results in artificial intelligence. However, AlphaZero, the current state-of-the-art MCTS algorithm still relies on handcrafted heuristics that are only partially understood. In this paper, we show that AlphaZero's search heuristic, along with other common ones, can be interpreted as an approximation to the solution of a specific regularized policy optimization problem. With this insight, we propose a variant of AlphaZero which uses the exact solution to this policy optimization problem, and show experimentally that it reliably outperforms the original algorithm in multiple domains.

Near-Tight Margin-Based Generalization Bounds for Support Vector Machines

- Allan Grønlund, Lior Kamma, Kasper Green Larsen
- abstract: Support Vector Machines (SVMs) are among the most fundamental tools for binary classification. In its simplest formulation, an SVM produces a hyperplane separating two classes of data using the largest possible margin to the data. The focus on maximizing the margin has been well motivated through numerous generalization bounds. In this paper, we revisit and improve the classic generalization bounds in terms of margins. Furthermore, we complement our new generalization bound by a nearly matching lower bound, thus almost settling the generalization performance of SVMs in terms of margins.

Implicit Geometric Regularization for Learning Shapes

- Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, Yaron Lipman
- abstract: Representing shapes as level-sets of neural networks has been recently proved to be useful for different shape analysis and reconstruction tasks. So far, such representations were computed using either: (i) pre-computed implicit shape representations; or (ii) loss functions explicitly defined over the neural level-sets. In this paper we offer a new paradigm for computing high fidelity implicit neural representations directly from raw data (i.e., point clouds, with or without normal information). We observe that a rather simple loss function, encouraging the neural network to vanish on the input point cloud and to have a unit norm gradient, possesses an implicit geometric regularization property that favors smooth and natural zero level-set surfaces,

avoiding bad zero-loss solutions. We provide a theoretical analysis of this property for the linear case, and show that, in practice, our method leads to state-of-the-art implicit neural representations with higher level-of-details and fidelity compared to previous methods.

Improving the Gating Mechanism of Recurrent Neural Networks

- Albert Gu, Caglar Gulcehre, Thomas Paine, Matt Hoffman, Razvan Pascanu
- abstract: Gating mechanisms are widely used in neural network models, where they allow gradients to backpropagate easily through depth or time. However, their saturation property introduces problems of its own. For example, in recurrent models these gates need to have outputs near 1 to propagate information over long time-delays, which requires them to operate in their saturation regime and hinders gradient-based learning of the gate mechanism. We address this problem by deriving two synergistic modifications to the standard gating mechanism that are easy to implement, introduce no additional hyperparameters, and improve learnability of the gates when they are close to saturation. We show how these changes are related to and improve on alternative recently proposed gating mechanisms such as chrono-initialization and Ordered Neurons. Empirically, our simple gating mechanisms robustly improve the performance of recurrent models on a range of applications, including synthetic memorization tasks, sequential image classification, language modeling, and reinforcement learning, particularly when long-term dependencies are involved.

Recurrent Hierarchical Topic-Guided RNN for Language Generation

- Dandan Guo, Bo Chen, Ruiying Lu, Mingyuan Zhou
- abstract: To simultaneously capture syntax and global semantics from a text corpus, we propose a new larger-context recurrent neural network (RNN) based language model, which extracts recurrent hierarchical semantic structure via a dynamic deep topic model to guide natural language generation. Moving beyond a conventional RNN-based language model that ignores long-range word dependencies and sentence order, the proposed model captures not only intra-sentence word dependencies, but also temporal transitions between sentences and inter-sentence topic dependencies. For inference, we develop a hybrid of stochastic-gradient Markov chain Monte Carlo and recurrent autoencoding variational Bayes. Experimental results on a variety of real-world text corpora demonstrate that the proposed model not only outperforms larger-context RNN-based language models, but also learns interpretable recurrent multilayer topics and generates diverse sentences and paragraphs that are syntactically correct and semantically coherent.

Breaking the Curse of Space Explosion: Towards Efficient NAS with Curriculum Search

- Yong Guo, Yaofo Chen, Yin Zheng, Peilin Zhao, Jian Chen, Junzhou Huang, Mingkui Tan
- abstract: Neural architecture search (NAS) has become an important approach to automatically find effective architectures. To cover all possible good architectures, we need to search in an extremely large search space with billions of candidate architectures. More critically, given a large search space, we may face a very challenging issue of space explosion. However, due to the limitation of computational resources, we can only sample a very small proportion of the architectures, which provides insufficient information for the training. As a result, existing methods may often produce sub-optimal architectures. To alleviate this issue, we propose a curriculum search method that starts from a small search space and gradually incorporates the learned knowledge to guide the search in a large space. With the proposed search strategy, our Curriculum Neural Architecture Search (CNAS) method significantly improves the search efficiency and finds better architectures than existing NAS methods. Extensive experiments on CIFAR-10 and ImageNet demonstrate the effectiveness of the proposed method.

Certified Data Removal from Machine Learning Models

- Chuan Guo, Tom Goldstein, Awni Hannun, Laurens Van Der Maaten
- abstract: Good data stewardship requires removal of data at the request of the data’s owner. This raises the question if and how a trained machine-learning model, which implicitly stores information about its training data, should be affected by such a removal request. Is it possible to “remove” data from a machine-learning model? We study this problem by defining certified removal: a very strong theoretical guarantee that a model from which data is removed cannot be distinguished from a model that never observed the data to begin with. We develop a certified-removal mechanism for linear classifiers and empirically study learning settings in which this mechanism is practical.

LTF: A Label Transformation Framework for Correcting Label Shift

- Jiaxian Guo, Mingming Gong, Tongliang Liu, Kun Zhang, Dacheng Tao
- abstract: Distribution shift is a major obstacle to the deployment of current deep learning models on real-world problems. Let Y be the class label and X the features. We focus on one type of distribution shift, \emph{ label shift}, where the label marginal distribution P_Y changes but the conditional distribution $P_{X|Y}$ does not. Most existing methods estimate the density ratio between the source- and target-domain label distributions by density matching. However, these methods are either computationally infeasible for large-scale data or restricted to shift correction for discrete labels. In this paper, we propose an end-to-end Label Transformation Framework (LTF) for correcting label shift, which implicitly models the shift of P_Y and the conditional distribution $P_{X|Y}$ using neural networks. Thanks to the flexibility of deep networks, our framework can handle continuous, discrete, and even multi-dimensional labels in a unified way and is scalable to large data. Moreover, for high dimensional X, such as images, we find that the redundant information in X severely degrades the estimation accuracy. To remedy this issue, we propose to match the distribution implied by our generative model and the target-domain distribution in a low-dimensional feature space that discards information irrelevant to Y. Both theoretical and empirical studies demonstrate the superiority of our method over previous approaches.

Learning to Branch for Multi-Task Learning

- Pengsheng Guo, Chen-Yu Lee, Daniel Ulbricht
- abstract: Training multiple tasks jointly in one deep network yields reduced latency during inference and better performance over the single-task counterpart by sharing certain layers of a network. However, over-sharing a network could erroneously enforce over-generalization, causing negative knowledge transfer across tasks. Prior works rely on human intuition or pre-computed task relatedness scores for ad hoc branching structures. They provide sub-optimal end results and often require huge efforts for the trial-and-error process. In this work, we present an automated multi-task learning algorithm that learns where to share or branch within a network, designing an effective network topology that is directly optimized for multiple objectives across tasks. Specifically, we propose a novel tree-structured design space that casts a tree branching operation as a gumbel-softmax sampling procedure. This enables differentiable network splitting that is end-to-end trainable. We validate the proposed method on controlled synthetic data, CelebA, and Taskonomy.

Communication-Efficient Distributed Stochastic AUC Maximization with Deep Neural Networks

- Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, Tianbao Yang
- abstract: In this paper, we study distributed algorithms for large-scale AUC maximization with a deep neural network as a predictive model. Although distributed learning techniques have been investigated extensively in deep learning, they are not directly applicable to stochastic AUC maximization with deep neural networks due to its striking differences from standard loss minimization problems (e.g., cross-entropy). Towards addressing this challenge, we propose and analyze a communication-efficient distributed optimization algorithm based on a \emph{non-convex concave} reformulation of the AUC maximization, in which the communication of both the primal variable and the dual variable between each worker and the parameter server only occurs

after multiple steps of gradient-based updates in each worker. Compared with the naive parallel version of an existing algorithm that computes stochastic gradients at individual machines and averages them for updating the model parameter, our algorithm requires a much less number of communication rounds and still achieves linear speedup in theory. To the best of our knowledge, this is the \textbf{first} work that solves the \emph{non-convex concave min-max} problem for AUC maximization with deep neural networks in a communication-efficient distributed manner while still maintaining the linear speedup property in theory. Our experiments on several benchmark datasets show the effectiveness of our algorithm and also confirm our theory.

Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning

- Zhaohan Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-Bastien Grill, Florent Altché, Remi Munos, Mohammad Gheshlaghi Azar
- abstract: Learning a good representation is an essential component for deep reinforcement learning (RL). Representation learning is especially important in multitask and partially observable settings where building a representation of the unknown environment is crucial to solve the tasks. Here we introduce Predictions of Bootstrapped Latents (PBL), a simple and flexible self-supervised representation learning algorithm for multitask deep RL. PBL builds on multistep predictive representations of future observations, and focuses on capturing structured information about environment dynamics. Specifically, PBL trains its representation by predicting latent embeddings of future observations. These latent embeddings are themselves trained to be predictive of the aforementioned representations. These predictions form a bootstrapping effect, allowing the agent to learn more about the key aspects of the environment dynamics. In addition, by defining prediction tasks completely in latent space, PBL provides the flexibility of using multimodal observations involving pixel images, language instructions, rewards and more. We show in our experiments that PBL delivers across-the-board improved performance over state of the art deep RL agents in the DMLab-30 multitask setting.

Accelerating Large-Scale Inference with Anisotropic Vector Quantization

- Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, Sanjiv Kumar
- abstract: Quantization based techniques are the current state-of-the-art for scaling maximum inner product search to massive databases. Traditional approaches to quantization aim to minimize the reconstruction error of the database points. Based on the observation that for a given query, the database points that have the largest inner products are more relevant, we develop a family of anisotropic quantization loss functions. Under natural statistical assumptions, we show that quantization with these loss functions leads to a new variant of vector quantization that more greatly penalizes the parallel component of a datapoint's residual relative to its orthogonal component. The proposed approach, whose implementation is open-source, achieves state-of-the-art results on the public benchmarks available at ann-benchmarks.com.

Safe Deep Semi-Supervised Learning for Unseen-Class Unlabeled Data

- Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, Zhi-Hua Zhou
- abstract: Deep semi-supervised learning (SSL) has been recently shown very effectively. However, its performance is seriously decreased when the class distribution is mismatched, among which a common situation is that unlabeled data contains some classes not seen in the labeled data. Efforts on this issue remain to be limited. This paper proposes a simple and effective safe deep SSL method to alleviate the harm caused by it. In theory, the result learned from the new method is never worse than learning from merely labeled data, and it is theoretically guaranteed that its generalization approaches the optimal in the order $O(\sqrt{d\ln(n)/n})$, even faster than the convergence rate in supervised learning associated with massive parameters. In the experiment of benchmark data, unlike the existing deep SSL methods which are no longer as good as supervised learning in 40% of unseen-class unlabeled data, the new method can still achieve performance gain in more than 60% of unseen-class unlabeled data. Moreover, the proposal is suitable for many deep SSL algorithms and can be easily extended to handle other cases of class distribution mismatch.

Neural Topic Modeling with Continual Lifelong Learning

- Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, Hinrich Schuetze
- abstract: Lifelong learning has recently attracted attention in building machine learning systems that continually accumulate and transfer knowledge to help future learning. Unsupervised topic modeling has been popularly used to discover topics from document collections. However, the application of topic modeling is challenging due to data sparsity, e.g., in a small collection of (short) documents and thus, generate incoherent topics and sub-optimal document representations. To address the problem, we propose a lifelong learning framework for neural topic modeling that can continuously process streams of document collections, accumulate topics and guide future topic modeling tasks by knowledge transfer from several sources to better deal with the sparse data. In the lifelong process, we particularly investigate jointly: (1) sharing generative homologies (latent topics) over lifetime to transfer prior knowledge, and (2) minimizing catastrophic forgetting to retain the past learning via novel selective data augmentation, co-training and topic regularization approaches. Given a stream of document collections, we apply the proposed Lifelong Neural Topic Modeling (LNTM) framework in modeling three sparse document collections as future tasks and demonstrate improved performance quantified by perplexity, topic coherence and information retrieval task. Code: https://github.com/pgcool/Lifelong-Neural-Topic-Modeling

Multidimensional Shape Constraints

- Maya Gupta, Erez Louidor, Oleksandr Mangylov, Nobu Morioka, Taman Narayan, Sen Zhao
- abstract: We propose new multi-input shape constraints across four intuitive categories: complements, diminishers, dominance, and unimodality constraints. We show these shape constraints can be checked and even enforced when training machine-learned models for linear models, generalized additive models, and the nonlinear function class of multi-layer lattice models. Real-world experiments illustrate how the different shape constraints can be used to increase explainability and improve regularization, especially for non-IID train-test distribution shift.

Retrieval Augmented Language Model Pre-Training

- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Mingwei Chang
- abstract: Language model pre-training has been shown to capture a surprising amount of world knowledge, crucial for NLP tasks such as question answering. However, this knowledge is stored implicitly in the parameters of a neural network, requiring ever-larger networks to cover more facts. To capture knowledge in a more modular and interpretable way, we augment language model pre-training with a latent knowledge retriever, which allows the model to retrieve and attend over documents from a large corpus such as Wikipedia, used during pre-training, fine-tuning and inference. For the first time, we show how to pre-train such a knowledge retriever in an unsupervised manner, using masked language modeling as the learning signal and backpropagating through a retrieval step that considers millions of documents. We demonstrate the effectiveness of Retrieval-Augmented Language Model pre-training (REALM) by fine-tuning on the challenging task of Open-domain Question Answering (Open-QA). We compare against state-of-the-art models for both explicit and implicit knowledge storage on three popular Open-QA benchmarks, and find that we outperform all previous methods by a significant margin (4-16% absolute accuracy), while also providing qualitative benefits such as interpretability and modularity.

Streaming Submodular Maximization under a k-Set System Constraint

- Ran Haba, Ehsan Kazemi, Moran Feldman, Amin Karbasi
- abstract: In this paper, we propose a novel framework that converts streaming algorithms for monotone submodular maximization into streaming algorithms for non-monotone submodular maximization. This reduction readily leads to the currently tightest deterministic approximation ratio for

submodular maximization subject to a k-matchoid constraint. Moreover, we propose the first streaming algorithm for monotone submodular maximization subject to k-extendible and k-set system constraints. Together with our proposed reduction, we obtain $O(k\log k)$ and $O(k^2\log k)$ approximation ratio for submodular maximization subject to the above constraints, respectively. We extensively evaluate the empirical performance of our algorithm against the existing work in a series of experiments including finding the maximum independent set in randomly generated graphs, maximizing linear functions over social networks, movie recommendation, Yelp location summarization, and Twitter data summarization.

Let's Agree to Agree: Neural Networks Share Classification Order on Real Datasets

- Guy Hacohen, Leshem Choshen, Daphna Weinshall
- abstract: We report a series of robust empirical observations, demonstrating that deep Neural Networks learn the examples in both the training and test sets in a similar order. This phenomenon is observed in all the commonly used benchmarks we evaluated, including many image classification benchmarks, and one text classification benchmark. While this phenomenon is strongest for models of the same architecture, it also crosses architectural boundaries – models of different architectures start by learning the same examples, after which the more powerful model may continue to learn additional examples. We further show that this pattern of results reflects the interplay between the way neural networks learn benchmark datasets. Specifically, when fixing the architecture, we describe synthetic datasets for which this pattern is no longer observed. When fixing the dataset, we show that other learning paradigms may learn the data in a different order. We hypothesize that our results reflect how neural networks discover structure in natural datasets.

Optimal approximation for unconstrained non-submodular minimization

- Marwa El Halabi, Stefanie Jegelka
- abstract: Submodular function minimization is well studied, and existing algorithms solve it exactly or up to arbitrary accuracy. However, in many applications, such as structured sparse learning or batch Bayesian optimization, the objective function is not exactly submodular, but close. In this case, no theoretical guarantees exist. Indeed, submodular minimization algorithms rely on intricate connections between submodularity and convexity. We show how these relations can be extended to obtain approximation guarantees for minimizing non-submodular functions, characterized by how close the function is to submodular. We also extend this result to noisy function evaluations. Our approximation results are the first for minimizing non-submodular functions, and are optimal, as established by our matching lower bound.

FedBoost: A Communication-Efficient Algorithm for Federated Learning

- Jenny Hamer, Mehryar Mohri, Ananda Theertha Suresh
- abstract: Communication cost is often a bottleneck in federated learning and other client-based distributed learning scenarios. To overcome this, several gradient compression and model compression algorithms have been proposed. In this work, we propose an alternative approach whereby an ensemble of pre-trained base predictors is trained via federated learning. This method allows for training a model which may otherwise surpass the communication bandwidth and storage capacity of the clients to be learned with on-device data through federated learning. Motivated by language modeling, we prove the optimality of ensemble methods for density estimation for standard empirical risk minimization and agnostic risk minimization. We provide communication-efficient ensemble algorithms for federated learning, where per-round communication cost is independent of the size of the ensemble. Furthermore, unlike works on gradient compression, our proposed approach reduces the communication cost of both server-to-client and client-to-server communication.

Polynomial Tensor Sketch for Element-wise Function of Low-Rank Matrix

- Insu Han, Haim Avron, Jinwoo Shin
- abstract: This paper studies how to sketch element-wise functions of low-rank matrices. Formally, given low-rank matrix A = [Aij] and scalar non-linear function f, we aim for finding an approximated low-rank representation of the (possibly high-rank) matrix [f(Aij)]. To this end, we propose an efficient sketching-based algorithm whose complexity is significantly lower than the number of entries of A, i.e., it runs without accessing all entries of [f(Aij)] explicitly. The main idea underlying our method is to combine a polynomial approximation of f with the existing tensor sketch scheme for approximating monomials of entries of A. To balance the errors of the two approximation components in an optimal manner, we propose a novel regression formula to find polynomial coefficients given A and f. In particular, we utilize a coreset-based regression with a rigorous approximation guarantee. Finally, we demonstrate the applicability and superiority of the proposed scheme under various machine learning tasks.

DRWR: A Differentiable Renderer without Rendering for Unsupervised 3D Structure Learning from Silhouette Images

- Zhizhong Han, Chao Chen, Yu-Shen Liu, Matthias Zwicker
- abstract: Differentiable renderers have been used successfully for unsupervised 3D structure learning from 2D images because they can bridge the gap between 3D and 2D. To optimize 3D shape parameters, current renderers rely on pixel-wise losses between rendered images of 3D reconstructions and ground truth images from corresponding viewpoints. Hence they require interpolation of the recovered 3D structure at each pixel, visibility handling, and optionally evaluating a shading model. In contrast, here we propose a Differentiable Renderer Without Rendering (DRWR) that omits these steps. DRWR only relies on a simple but effective loss that evaluates how well the projections of reconstructed 3D point clouds cover the ground truth object silhouette. Specifically, DRWR employs a smooth silhouette loss to pull the projection of each individual 3D point inside the object silhouette, and a structure-aware repulsion loss to push each pair of projections that fall inside the silhouette far away from each other. Although we omit surface interpolation, visibility handling, and shading, our results demonstrate that DRWR achieves state-of-the-art accuracies under widely used benchmarks, outperforming previous methods both qualitatively and quantitatively. In addition, our training times are significantly lower due to the simplicity of DRWR.

SIGUA: Forgetting May Make Learning with Noisy Labels More Robust

- Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, Masashi Sugiyama
- abstract: Given data with noisy labels, over-parameterized deep networks can gradually memorize the data, and fit everything in the end. Although equipped with corrections for noisy labels, many learning methods in this area still suffer overfitting due to undesired memorization. In this paper, to relieve this issue, we propose stochastic integrated gradient underweighted ascent (SIGUA): in a mini-batch, we adopt gradient descent on good data as usual, and learning-rate-reduced gradient ascent on bad data; the proposal is a versatile approach where data goodness or badness is w.r.t. desired or undesired memorization given a base learning method. Technically, SIGUA pulls optimization back for generalization when their goals conflict with each other; philosophically, SIGUA shows forgetting undesired memorization can reinforce desired memorization. Experiments demonstrate that SIGUA successfully robustifies two typical base learning methods, so that their performance is often significantly improved.

Training Binary Neural Networks through Learning with Noisy Supervision

- Kai Han, Yunhe Wang, Yixing Xu, Chunjing Xu, Enhua Wu, Chang Xu
- abstract: This paper formalizes the binarization operations over neural networks from a learning perspective. In contrast to classical hand crafted rules (\eg hard thresholding) to binarize full-precision neurons, we propose to learn a mapping from full-precision neurons to the target binary ones. Each individual weight entry will not be binarized independently. Instead, they are taken as a whole to accomplish the binarization, just as they work together in generating convolution features. To help the training of the binarization mapping, the full-precision neurons after taking sign operations is regarded as some auxiliary

supervision signal, which is noisy but still has valuable guidance. An unbiased estimator is therefore introduced to mitigate the influence of the supervision noise. Experimental results on benchmark datasets indicate that the proposed binarization technique attains consistent improvements over baselines.

Stochastic Subspace Cubic Newton Method

- Filip Hanzely, Nikita Doikov, Yurii Nesterov, Peter Richtarik
- abstract: In this paper, we propose a new randomized second-order optimization algorithm—Stochastic Subspace Cubic Newton (SSCN)—for minimizing a high dimensional convex function f. Our method can be seen both as a \emph{stochastic} extension of the cubically-regularized Newton method of Nesterov and Polyak (2006), and a \emph{second-order} enhancement of stochastic subspace descent of Kozak et al. (2019). We prove that as we vary the minibatch size, the global convergence rate of SSCN interpolates between the rate of stochastic coordinate descent (CD) and the rate of cubic regularized Newton, thus giving new insights into the connection between first and second-order methods. Remarkably, the local convergence rate of SSCN matches the rate of stochastic subspace descent applied to the problem of minimizing the quadratic function $\frac12 (x-x^)^\top \nabla^2f(x^)(x-x^)$, where $x^$ is the minimizer of f, and hence depends on the properties of f at the optimum only. Our numerical experiments show that SSCN outperforms non-accelerated first-order CD algorithms while being competitive to their accelerated variants.

Variance Reduced Coordinate Descent with Acceleration: New Method With a Surprising Application to Finite-Sum Problems

- Filip Hanzely, Dmitry Kovalev, Peter Richtarik
- abstract: We propose an accelerated version of stochastic variance reduced coordinate descent – ASVRCD. As other variance reduced coordinate descent methods such as SEGA or SVRCD, our method can deal with problems that include a non-separable and non-smooth regularizer, while accessing a random block of partial derivatives in each iteration only. However, ASVRCD incorporates Nesterov's momentum, which offers favorable iteration complexity guarantees over both SEGA and SVRCD. As a by-product of our theory, we show that a variant of Katyusha (Allen-Zhu, 2017) is a specific case of ASVRCD, recovering the optimal oracle complexity for the finite sum objective.

Data Amplification: Instance-Optimal Property Estimation

- Yi Hao, Alon Orlitsky
- abstract: The best-known and most commonly used technique for distribution-property estimation uses a plug-in estimator, with empirical frequency replacing the underlying distribution. We present novel linear-time-computable estimators that significantly "amplify" the effective amount of data available. For a large variety of distribution properties including four of the most popular ones and for every underlying distribution, they achieve the accuracy that the empirical-frequency plug-in estimators would attain using a logarithmic-factor more samples. Specifically, for Shannon entropy and a broad class of Lipschitz properties including the L_1 distance to a fixed distribution, the new estimators use n samples to achieve the accuracy attained by the empirical estimators with $n\log n$ samples. For support-size and coverage, the new estimators use n samples to achieve the performance of empirical frequency with sample size n times the logarithm of the property value. Significantly strengthening the traditional min-max formulation, these results hold not only for the worst distributions, but for each and every underlying distribution. Furthermore, the logarithmic amplification factors are optimal. Experiments on a wide variety of distributions show that the new estimators outperform the previous state-of-the-art estimators designed for each specific property.

Dynamic Knapsack Optimization Towards Efficient Multi-Channel Sequential Advertising

- Xiaotian Hao, Zhaoqing Peng, Yi Ma, Guan Wang, Junqi Jin, Jianye Hao, Shan Chen, Rongquan Bai, Mingzhou Xie, Miao Xu, Zhenzhe Zheng, Chuan Yu, Han Li, Jian Xu, Kun Gai
- abstract: In E-commerce, advertising is essential for merchants to reach their target users. The typical objective is to maximize the advertiser's cumulative revenue over a period of time under a budget constraint. In real applications, an advertisement (ad) usually needs to be exposed to the same user multiple times until the user finally contributes revenue (e.g., places an order). However, existing advertising systems mainly focus on the immediate revenue with single ad exposures, ignoring the contribution of each exposure to the final conversion, thus usually falls into suboptimal solutions. In this paper, we formulate the sequential advertising strategy optimization as a dynamic knapsack problem. We propose a theoretically guaranteed bilevel optimization framework, which significantly reduces the solution space of the original optimization space while ensuring the solution quality. To improve the exploration efficiency of reinforcement learning, we also devise an effective action space reduction approach. Extensive offline and online experiments show the superior performance of our approaches over state-of-the-art baselines in terms of cumulative revenue.

Improving generalization by controlling label-noise information in neural network weights

- Hrayr Harutyunyan, Kyle Reing, Greg Ver Steeg, Aram Galstyan
- abstract: In the presence of noisy or incorrect labels, neural networks have the undesirable tendency to memorize information about the noise. Standard regularization techniques such as dropout, weight decay or data augmentation sometimes help, but do not prevent this behavior. If one considers neural network weights as random variables that depend on the data and stochasticity of training, the amount of memorized information can be quantified with the Shannon mutual information between weights and the vector of all training labels given inputs, $I(w; \mathbf{y} \mid \mathbf{x})$. We show that for any training algorithm, low values of this term correspond to reduction in memorization of label-noise and better generalization bounds. To obtain these low values, we propose training algorithms that employ an auxiliary network that predicts gradients in the final layers of a classifier without accessing labels. We illustrate the effectiveness of our approach on versions of MNIST, CIFAR-10, and CIFAR-100 corrupted with various noise models, and on a large-scale dataset Clothing1M that has noisy labels.

A Natural Lottery Ticket Winner: Reinforcement Learning with Ordinary Neural Circuits

- Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, Radu Grosu
- abstract: We propose a neural information processing system obtained by re-purposing the function of a biological neural circuit model to govern simulated and real-world control tasks. Inspired by the structure of the nervous system of the soil-worm, C. elegans, we introduce ordinary neural circuits (ONCs), defined as the model of biological neural circuits reparameterized for the control of alternative tasks. We first demonstrate that ONCs realize networks with higher maximum flow compared to arbitrary wired networks. We then learn instances of ONCs to control a series of robotic tasks, including the autonomous parking of a real-world rover robot. For reconfiguration of the purpose of the neural circuit, we adopt a search-based optimization algorithm. Ordinary neural circuits perform on par and, in some cases, significantly surpass the performance of contemporary deep learning models. ONC networks are compact, 77% sparser than their counterpart neural controllers, and their neural dynamics are fully interpretable at the cell-level.

Bayesian Graph Neural Networks with Adaptive Connection Sampling

- Arman Hasanzadeh, Ehsan Hajiramezanali, Shahin Boluki, Mingyuan Zhou, Nick Duffield, Krishna Narayanan, Xiaoning Qian
- abstract: We propose a unified framework for adaptive connection sampling in graph neural networks (GNNs) that generalizes existing stochastic regularization methods for training GNNs. The proposed framework not only alleviates over-smoothing and over-fitting tendencies of deep GNNs, but also enables learning with uncertainty in graph analytic tasks with GNNs. Instead of using fixed sampling rates or hand-tuning themas model

hyperparameters in existing stochastic regularization methods, our adaptive connection sampling can be trained jointly with GNN model parameters in both global and local fashions. GNN training with adaptive connection sampling is shown to be mathematically equivalent to an efficient approximation of training BayesianGNNs. Experimental results with ablation studies on benchmark datasets validate that adaptively learning the sampling rate given graph training data is the key to boost the performance of GNNs in semi-supervised node classification, less prone to over-smoothing and over-fitting with more robust prediction.

CoMic: Complementary Task Learning & Mimicry for Reusable Skills

- Leonard Hasenclever, Fabio Pardo, Raia Hadsell, Nicolas Heess, Josh Merel
- abstract: Learning to control complex bodies and reuse learned behaviors is a longstanding challenge in continuous control. We study the problem of learning reusable humanoid skills by imitating motion capture data and joint training with complementary tasks. We show that it is possible to learn reusable skills through reinforcement learning on 50 times more motion capture data than prior work. We systematically compare a variety of different network architectures across different data regimes both in terms of imitation performance as well as transfer to challenging locomotion tasks. Finally we show that it is possible to interleave the motion capture tracking with training on complementary tasks, enriching the resulting skill space, and enabling the reuse of skills not well covered by the motion capture data such as getting up from the ground or catching a ball.

Contrastive Multi-View Representation Learning on Graphs

- Kaveh Hassani, Amir Hosein Khasahmadi
- abstract: We introduce a self-supervised approach for learning node and graph level representations by contrasting structural views of graphs. We show that unlike visual representation learning, increasing the number of views to more than two or contrasting multi-scale encodings do not improve performance, and the best performance is achieved by contrasting encodings from first-order neighbors and a graph diffusion. We achieve new state-of-the-art results in self-supervised learning on 8 out of 8 node and graph classification benchmarks under the linear evaluation protocol. For example, on Cora (node) and Reddit-Binary (graph) classification benchmarks, we achieve 86.8% and 84.5% accuracy, which are 5.5% and 2.4% relative improvements over previous state-of-the-art. When compared to supervised baselines, our approach outperforms them in 4 out of 8 benchmarks.

Nested Subspace Arrangement for Representation of Relational Data

- Nozomi Hata, Shizuo Kaji, Akihiro Yoshida, Katsuki Fujisawa
- abstract: Studies of acquiring appropriate continuous representations of a discrete objects such as graph and knowledge based data have been conducted by many researches in the field of machine learning. In this paper, we introduce Nested SubSpace arrangement (NSS arrangement), a comprehensive framework for representation learning. We show that existing embedding techniques can be regarded as a member of NSS arrangement. Based on the concept of the NSS arrangement, we implemented Disk-ANChor ARrangement (DANCAR), a representation learning method specializing to reproduce general graphs. Numerical experiments have shown that DANCAR has successfully embedded WordNet in ${\mathbb R}^{20}$ with the F1 score of 0.993 in the reconstruction task. DANCAR is also suitable for visualization to understand the characteristics of graph.

The Tree Ensemble Layer: Differentiability meets Conditional Computation

- Hussein Hazimeh, Natalia Ponomareva, Petros Mol, Zhenyu Tan, Rahul Mazumder
- abstract: Neural networks and tree ensembles are state-of-the-art learners, each with its unique statistical and computational advantages. We aim to combine these advantages by introducing a new layer for neural networks, composed of an ensemble of differentiable decision trees (a.k.a. soft trees). While differentiable trees demonstrate promising results in the literature, they are typically slow in training and inference as they do not support conditional computation. We mitigate this issue by introducing a new sparse activation function for sample routing, and implement true conditional computation by developing specialized forward and backward propagation algorithms that exploit sparsity. Our efficient algorithms pave the way for jointly training over deep and wide tree ensembles using first-order methods (e.g., SGD). Experiments on 23 classification datasets indicate over 10x speed-ups compared to the differentiable trees used in the literature and over 20x reduction in the number of parameters compared to gradient boosted trees, while maintaining competitive performance. Moreover, experiments on CIFAR, MNIST, and Fashion MNIST indicate that replacing dense layers in CNNs with our tree layer reduces the test loss by 7-53% and the number of parameters by 8x. We provide an open-source TensorFlow implementation with a Keras API.

Compressive sensing with un-trained neural networks: Gradient descent finds a smooth approximation

- Reinhard Heckel, Mahdi Soltanolkotabi
- abstract: Un-trained convolutional neural networks have emerged as highly successful tools for image recovery and restoration. They are capable of solving standard inverse problems such as denoising and compressive sensing with excellent results by simply fitting a neural network model to measurements from a single image or signal without the need for any additional training data. For some applications, this critically requires additional regularization in the form of early stopping the optimization. For signal recovery from a few measurements, however, un-trained convolutional networks have an intriguing self-regularizing property: Even though the network can perfectly fit any image, the network recovers a natural image from few measurements when trained with gradient descent until convergence. In this paper, we provide numerical evidence for this property and study it theoretically. We show that—without any further regularization—an un-trained convolutional neural network can approximately reconstruct signals and images that are sufficiently structured, from a near minimal number of random measurements.

Hierarchically Decoupled Imitation For Morphological Transfer

- Donald Hejna, Lerrel Pinto, Pieter Abbeel
- abstract: Learning long-range behaviors on complex high-dimensional agents is a fundamental problem in robot learning. For such tasks, we argue that transferring learned information from a morphologically simpler agent can massively improve the sample efficiency of a more complex one. To this end, we propose a hierarchical decoupling of policies into two parts: an independently learned low-level policy and a transferable high-level policy. To remedy poor transfer performance due to mismatch in morphologies, we contribute two key ideas. First, we show that incentivizing a complex agent's low-level to imitate a simpler agent's low-level significantly improves zero-shot high-level transfer. Second, we show that KL-regularized training of the high level stabilizes learning and prevents mode-collapse. Finally, on a suite of publicly released navigation and manipulation environments, we demonstrate the applicability of hierarchical transfer on long-range tasks across morphologies. Our code and videos can be found at https://sites.google.com/berkeley.edu/morphology-transfer.

Gradient-free Online Learning in Continuous Games with Delayed Rewards

- Amélie Héliou, Panayotis Mertikopoulos, Zhengyuan Zhou
- abstract: Motivated by applications to online advertising and recommender systems, we consider a game-theoretic model with delayed rewards and asynchronous, payoff-based feedback. In contrast to previous work on delayed multi-armed bandits, we focus on games with continuous action spaces, and we examine the long-run behavior of strategic agents that follow a no-regret learning policy (but are otherwise oblivious to the game being played, the objectives of their opponents, etc.). To account for the lack of a consistent stream of information (for instance, rewards can arrive out of order and with an

a priori unbounded delay), we introduce a gradient-free learning policy where payoff information is placed in a priority queue as it arrives. Somewhat surprisingly, we find that under a standard diagonal concavity assumption, the induced sequence of play converges to Nash Equilibrium (NE) with probability 1, even if the delay between choosing an action and receiving the corresponding reward is unbounded.

Data-Efficient Image Recognition with Contrastive Predictive Coding

- Olivier Henaff
- abstract: Human observers can learn to recognize new categories of images from a handful of examples, yet doing so with artificial ones remains an open challenge. We hypothesize that data-efficient recognition is enabled by representations which make the variability in natural signals more predictable. We therefore revisit and improve Contrastive Predictive Coding, an unsupervised objective for learning such representations. This new implementation produces features which support state-of-the-art linear classification accuracy on the ImageNet dataset. When used as input for non-linear classification with deep neural networks, this representation allows us to use 2-5x less labels than classifiers trained directly on image pixels. Finally, this unsupervised representation substantially improves transfer learning to object detection on the PASCAL VOC dataset, surpassing fully supervised pre-trained ImageNet classifiers.

Minimax Rate for Learning From Pairwise Comparisons in the BTL Model

- Julien Hendrickx, Alex Olshevsky, Venkatesh Saligrama
- abstract: We consider the problem of learning the qualities w_1, ... , w_n of a collection of items by performing noisy comparisons among them. We assume there is a fixed “comparison graph” and every neighboring pair of items is compared k times. We will study the popular Bradley-Terry-Luce model, where the probability that item i wins a comparison against j equals w_i/(w_i + w_j). We are interested in how the expected error in estimating the vector w = (w_1, ... , w_n) behaves in the regime when the number of comparisons k is large. Our contribution is the determination of the minimax rate up to a constant factor. We show that this rate is achieved by a simple algorithm based on weighted least squares, with weights determined from the empirical outcomes of the comparisons. This algorithm can be implemented in nearly linear time in the total number of comparisons.

Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization

- Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, Laurent Massoulie
- abstract: We consider the setting of distributed empirical risk minimization where multiple machines compute the gradients in parallel and a centralized server updates the model parameters. In order to reduce the number of communications required to reach a given accuracy, we propose a preconditioned accelerated gradient method where the preconditioning is done by solving a local optimization problem over a subsampled dataset at the server. The convergence rate of the method depends on the square root of the relative condition number between the global and local loss functions. We estimate the relative condition number for linear prediction models by studying uniform concentration of the Hessians over a bounded domain, which allows us to derive improved convergence rates for existing preconditioned gradient methods and our accelerated method. Experiments on real-world datasets illustrate the benefits of acceleration in the ill-conditioned regime.

Cost-Effective Interactive Attention Learning with Neural Attention Processes

- Jay Heo, Junhyeon Park, Hyewon Jeong, Kwang Joon Kim, Juho Lee, Eunho Yang, Sung Ju Hwang
- abstract: We propose a novel interactive learning framework which we refer to as Interactive Attention Learning (IAL), in which the human supervisors interactively manipulate the allocated attentions, to correct the model’s behaviour by updating the attention-generating network. However, such a model is prone to overfitting due to scarcity of human annotations, and requires costly retraining. Moreover, it is almost infeasible for the human annotators to examine attentions on tons of instances and features. We tackle these challenges by proposing a sample-efficient attention mechanism and a cost-effective reranking algorithm for instances and features. First, we propose Neural Attention Processes (NAP), which is an attention generator that can update its behaviour by incorporating new attention-level supervisions without any retraining. Secondly, we propose an algorithm which prioritizes the instances and the features by their negative impacts, such that the model can yield large improvements with minimal human feedback. We validate IAL on various time-series datasets from multiple domains (healthcare, real-estate, and computer vision) on which it significantly outperforms baselines with conventional attention mechanisms, or without cost-effective reranking, with substantially less retraining and human-model interaction cost.

Likelihood-free MCMC with Amortized Approximate Ratio Estimators

- Joeri Hermans, Volodimir Begy, Gilles Louppe
- abstract: Posterior inference with an intractable likelihood is becoming an increasingly common task in scientific domains which rely on sophisticated computer simulations. Typically, these forward models do not admit tractable densities forcing practitioners to rely on approximations. This work introduces a novel approach to address the intractability of the likelihood and the marginal model. We achieve this by learning a flexible amortized estimator which approximates the likelihood-to-evidence ratio. We demonstrate that the learned ratio estimator can be embedded in \textsc{mcmc} samplers to approximate likelihood-ratios between consecutive states in the Markov chain, allowing us to draw samples from the intractable posterior. Techniques are presented to improve the numerical stability and to measure the quality of an approximation. The accuracy of our approach is demonstrated on a variety of benchmarks against well-established techniques. Scientific applications in physics show its applicability.

Towards Non-Parametric Drift Detection via Dynamic Adapting Window Independence Drift Detection (DAWIDD)

- Fabian Hinder, André Artelt, Barbara Hammer
- abstract: The notion of concept drift refers to the phenomenon that the distribution, which is underlying the observed data, changes over time; as a consequence machine learning models may become inaccurate and need adjustment. Many online learning schemes include drift detection to actively detect and react to observed changes. Yet, reliable drift detection constitutes a challenging problem in particular in the context of high dimensional data, varying drift characteristics, and the absence of a parametric model such as a classification scheme which reflects the drift. In this paper we present a novel concept drift detection method, Dynamic Adapting Window Independence Drift Detection (DAWIDD), which aims for non-parametric drift detection of diverse drift characteristics. For this purpose, we establish a mathematical equivalence of the presence of drift to the dependency of specific random variables in an according drift process. This allows us to rely on independence tests rather than parametric models or the classification loss, resulting in a fairly robust scheme to universally detect different types of drift, as it is also confirmed in experiments.

Optimization and Analysis of the pAp@k Metric for Recommender Systems

- Gaurush Hiranandani, Warut Vijitbenjaronk, Sanmi Koyejo, Prateek Jain
- abstract: Modern recommendation and notification systems must be robust to data imbalance, limitations on the number of recommendations/notifications, and heterogeneous engagement profiles across users. The pAp@k metric, which combines the partial-AUC and the precision@k metrics, was recently proposed to evaluate such recommendation systems and has been used in real-world deployments. Conceptually, pAp@k measures the probability of correctly ranking a top-ranked positive instance over top-ranked negative instances. Due to the combinatorial aspect surfaced by top-ranked points, little is known about the characteristics and optimization methods of pAp@k. In this paper, we analyze the learning-theoretic properties of pAp@k, particularly its benefits in evaluating modern recommender systems, and propose novel surrogates that are consistent under certain data regularity conditions. We then

provide gradient descent based algorithms to optimize the surrogates directly. Our analysis and experimental evaluation suggest that pAp@k indeed exhibits a certain dual behavior with respect to partial-AUC and precision@k. Moreover, the proposed methods outperform all the baselines in various applications. Taken together, our results motivate the use of pAp@k for large-scale recommender systems with heterogeneous user-engagement.

Optimizing Dynamic Structures with Bayesian Generative Search

- Minh Hoang, Carleton Kingsford
- abstract: Kernel selection for kernel-based methods is prohibitively expensive due to the NP-hard nature of discrete optimization. Since gradient-based optimizers are not applicable due to the lack of a differentiable objective function, many state-of-the-art solutions resort to heuristic search or gradient-free optimization. These approaches, however, require imposing restrictive assumptions on the explorable space of structures such as limiting the active candidate pool, thus depending heavily on the intuition of domain experts. This paper instead proposes \textbf{DTERGENS}, a novel generative search framework that constructs and optimizes a high-performance composite kernel expressions generator. \textbf{DTERGENS} does not restrict the space of candidate kernels and is capable of obtaining flexible length expressions by jointly optimizing a generative termination criterion. We demonstrate that our framework explores more diverse kernels and obtains better performance than state-of-the-art approaches on many real-world predictive tasks.

Learning Task-Agnostic Embedding of Multiple Black-Box Experts for Multi-Task Model Fusion

- Nghia Hoang, Thanh Lam, Bryan Kian Hsiang Low, Patrick Jaillet
- abstract: Model fusion is an emerging study in collective learning where heterogeneous experts with private data and learning architectures need to combine their black-box knowledge for better performance. Existing literature achieves this via a local knowledge distillation scheme that transfuses the predictive patterns of each pre-trained expert onto a white-box imitator model, which can be incorporated efficiently into a global model. This scheme however does not extend to multi-task scenarios where different experts were trained to solve different tasks and only part of their distilled knowledge is relevant to a new task. To address this multi-task challenge, we develop a new fusion paradigm that represents each expert as a distribution over a spectrum of predictive prototypes, which are isolated from task-specific information encoded within the prototype distribution. The task-agnostic prototypes can then be reintegrated to generate a new model that solves a new task encoded with a different prototype distribution. The fusion and adaptation performance of the proposed framework is demonstrated empirically on several real-world benchmark datasets.

Parameterized Rate-Distortion Stochastic Encoder

- Quan Hoang, Trung Le, Dinh Phung
- abstract: We propose a novel gradient-based tractable approach for the Blahut-Arimoto (BA) algorithm to compute the rate-distortion function where the BA algorithm is fully parameterized. This results in a rich and flexible framework to learn a new class of stochastic encoders, termed PArameterized RAte-DIstortion Stochastic Encoder (PARADISE). The framework can be applied to a wide range of settings from semi-supervised, multi-task to supervised and robust learning. We show that the training objective of PARADISE can be seen as a form of regularization that helps improve generalization. With an emphasis on robust learning we further develop a novel posterior matching objective to encourage smoothness on the loss function and show that PARADISE can significantly improve interpretability as well as robustness to adversarial attacks on the CIFAR-10 and ImageNet datasets. In particular, on the CIFAR-10 dataset, our model reduces standard and adversarial error rates in comparison to the state-of-the-art by 50% and 41%, respectively without the expensive computational cost of adversarial training.

Topologically Densified Distributions

- Christoph Hofer, Florian Graf, Marc Niethammer, Roland Kwitt
- abstract: We study regularization in the context of small sample-size learning with over-parametrized neural networks. Specifically, we shift focus from architectural properties, such as norms on the network weights, to properties of the internal representations before a linear classifier. Specifically, we impose a topological constraint on samples drawn from the probability measure induced in that space. This provably leads to mass concentration effects around the representations of training instances, i.e., a property beneficial for generalization. By leveraging previous work to impose topological constrains in a neural network setting, we provide empirical evidence (across various vision benchmarks) to support our claim for better generalization.

Graph Filtration Learning

- Christoph Hofer, Florian Graf, Bastian Rieck, Marc Niethammer, Roland Kwitt
- abstract: We propose an approach to learning with graph-structured data in the problem domain of graph classification. In particular, we present a novel type of readout operation to aggregate node features into a graph-level representation. To this end, we leverage persistent homology computed via a real-valued, learnable, filter function. We establish the theoretical foundation for differentiating through the persistent homology computation. Empirically, we show that this type of readout operation compares favorably to previous techniques, especially when the graph connectivity structure is informative for the learning problem.

Black-Box Variational Inference as a Parametric Approximation to Langevin Dynamics

- Matthew Hoffman, Yian Ma
- abstract: Variational inference (VI) and Markov chain Monte Carlo (MCMC) are approximate posterior inference algorithms that are often said to have complementary strengths, with VI being fast but biased and MCMC being slower but asymptotically unbiased. In this paper, we analyze gradient-based MCMC and VI procedures and find theoretical and empirical evidence that these procedures are not as different as one might think. In particular, a close examination of the Fokker-Planck equation that governs the Langevin dynamics (LD) MCMC procedure reveals that LD implicitly follows a gradient flow that corresponds to a variational inference procedure based on optimizing a nonparametric normalizing flow. This result suggests that the transient bias of LD (due to the Markov chain not having burned in) may track that of VI (due to the optimizer not having converged), up to differences due to VI's asymptotic bias and parameterization. Empirically, we find that the transient biases of these algorithms (and their momentum-accelerated counterparts) do evolve similarly. This suggests that practitioners with a limited time budget may get more accurate results by running an MCMC procedure (even if it's far from burned in) than a VI procedure, as long as the variance of the MCMC estimator can be dealt with (e.g., by running many parallel chains).

Learning Mixtures of Graphs from Epidemic Cascades

- Jessica Hoffmann, Soumya Basu, Surbhi Goel, Constantine Caramanis
- abstract: We consider the problem of learning the weighted edges of a balanced mixture of two undirected graphs from epidemic cascades. While mixture models are popular modeling tools, algorithmic development with rigorous guarantees has lagged. Graph mixtures are apparently no exception: until now, very little is known about whether this problem is solvable. To the best of our knowledge, we establish the first necessary and sufficient conditions for this problem to be solvable in polynomial time on edge-separated graphs. When the conditions are met, i.e., when the graphs are connected with at least three edges, we give an efficient algorithm for learning the weights of both graphs with optimal sample complexity (up to log factors). We give complementary results and provide sample-optimal (up to log factors) algorithms for mixtures of directed graphs of out-degree at least three, and for mixture of undirected graphs of unbalanced and/or unknown priors.

Set Functions for Time Series

- Max Horn, Michael Moor, Christian Bock, Bastian Rieck, Karsten Borgwardt
- abstract: Despite the eminent successes of deep neural networks, many architectures are often hard to transfer to irregularly-sampled and asynchronous time series that commonly occur in real-world datasets, especially in healthcare applications. This paper proposes a novel approach for classifying irregularly-sampled time series with unaligned measurements, focusing on high scalability and data efficiency. Our method SeFT (Set Functions for Time Series) is based on recent advances in differentiable set function learning, extremely parallelizable with a beneficial memory footprint, thus scaling well to large datasets of long time series and online monitoring scenarios. Furthermore, our approach permits quantifying per-observation contributions to the classification outcome. We extensively compare our method with existing algorithms on multiple healthcare time series datasets and demonstrate that it performs competitively whilst significantly reducing runtime.

Lifted Disjoint Paths with Application in Multiple Object Tracking

- Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, Paul Swoboda
- abstract: We present an extension to the disjoint paths problem in which additional lifted edges are introduced to provide path connectivity priors. We call the resulting optimization problem the lifted disjoint paths problem. We show that this problem is NP-hard by reduction from integer multicommodity flow and 3-SAT. To enable practical global optimization, we propose several classes of linear inequalities that produce a high-quality LP-relaxation. Additionally, we propose efficient cutting plane algorithms for separating the proposed linear inequalities. The lifted disjoint path problem is a natural model for multiple object tracking and allows an elegant mathematical formulation for long range temporal interactions. Lifted edges help to prevent id switches and to re-identify persons. Our lifted disjoint paths tracker achieves nearly optimal assignments with respect to input detections. As a consequence, it leads on all three main benchmarks of the MOT challenge, improving significantly over state-of-the-art.

Infinite attention: NNGP and NTK for deep attention networks

- Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, Roman Novak
- abstract: There is a growing amount of literature on the relationship between wide neural networks (NNs) and Gaussian processes (GPs), identifying an equivalence between the two for a variety of NN architectures. This equivalence enables, for instance, accurate approximation of the behaviour of wide Bayesian NNs without MCMC or variational approximations, or characterisation of the distribution of randomly initialised wide NNs optimised by gradient descent without ever running an optimiser. We provide a rigorous extension of these results to NNs involving attention layers, showing that unlike single-head attention, which induces non-Gaussian behaviour, multi-head attention architectures behave as GPs as the number of heads tends to infinity. We further discuss the effects of positional encodings and layer normalisation, and propose modifications of the attention mechanism which lead to improved results for both finite and infinitely wide NNs. We evaluate attention kernels empirically, leading to a moderate improvement upon the previous state-of-the-art on CIFAR-10 for GPs without trainable kernels and advanced data preprocessing. Finally, we introduce new features to the Neural Tangents library (Novak et al.,2020) allowing applications of NNGP/NTK models, with and without attention, to variable-length sequences, with an example on the IMDb reviews dataset.

The Non-IID Data Quagmire of Decentralized Machine Learning

- Kevin Hsieh, Amar Phanishayee, Onur Mutlu, Phillip Gibbons
- abstract: Many large-scale machine learning (ML) applications need to perform decentralized learning over datasets generated at different devices and locations. Such datasets pose a significant challenge to decentralized learning because their different contexts result in significant data distribution skew across devices/locations. In this paper, we take a step toward better understanding this challenge by presenting a detailed experimental study of decentralized DNN training on a common type of data skew: skewed distribution of data labels across devices/locations. Our study shows that: (i) skewed data labels are a fundamental and pervasive problem for decentralized learning, causing significant accuracy loss across many ML applications, DNN models, training datasets, and decentralized learning algorithms; (ii) the problem is particularly challenging for DNN models with batch normalization; and (iii) the degree of data skew is a key determinant of the difficulty of the problem. Based on these findings, we present SkewScout, a system-level approach that adapts the communication frequency of decentralized learning algorithms to the (skew-induced) accuracy loss between data partitions. We also show that group normalization can recover much of the accuracy loss of batch normalization.

“Other-Play” for Zero-Shot Coordination

- Hengyuan Hu, Adam Lerer, Alex Peysakhovich, Jakob Foerster
- abstract: We consider the problem of zero-shot coordination - constructing AI agents that can coordinate with novel partners they have not seen before (e.g.humans). Standard Multi-Agent Reinforcement Learning (MARL) methods typically focus on the self-play (SP) setting where agents construct strategies by playing the game with themselves repeatedly. Unfortunately, applying SP naively to the zero-shot coordination problem can produce agents that establish highly specialized conventions that do not carry over to novel partners they have not been trained with. We introduce a novel learning algorithm called other-play (OP), that enhances self-play by looking for more robust strategies. We characterize OP theoretically as well as experimentally. We study the cooperative card game Hanabi and show that OP agents achieve higher scores when paired with independently trained agents as well as with human players than SP agents.

XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation

- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, Melvin Johnson
- abstract: Much recent progress in applications of machine learning models to NLP has been driven by benchmarks that evaluate models across a wide variety of tasks. However, these broad-coverage benchmarks have been mostly limited to English, and despite an increasing interest in multilingual models, a benchmark that enables the comprehensive evaluation of such methods on a diverse range of languages and tasks is still missing. To this end, we introduce the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark, a multi-task benchmark for evaluating the cross-lingual generalization capabilities of multilingual representations across 40 languages and 9 tasks. We demonstrate that while models tested on English reach human performance on many tasks, there is still a sizable gap in the performance of cross-lingually transferred models, particularly on syntactic and sentence retrieval tasks. There is also a wide spread of results across languages. We will release the benchmark to encourage research on cross-lingual learning methods that transfer linguistic knowledge across a diverse and representative set of languages and tasks.

Momentum-Based Policy Gradient Methods

- Feihu Huang, Shangqian Gao, Jian Pei, Heng Huang
- abstract: In the paper, we propose a class of efficient momentum-based policy gradient methods for the model-free reinforcement learning, which use adaptive learning rates and do not require any large batches. Specifically, we propose a fast important-sampling momentum-based policy gradient (IS-MBPG) method based on a new momentum-based variance reduced technique and the importance sampling technique. We also propose a fast Hessian-aided momentum-based policy gradient (HA-MBPG) method based on the momentum-based variance reduced technique and the Hessian-aided technique. Moreover, we prove that both the IS-MBPG and HA-MBPG methods reach the best known sample complexity of $O(\epsilon^{-3})$ for finding an ϵ-stationary point of the nonconcave performance function, which only require one trajectory at each iteration. In particular, we present a non-

adaptive version of IS-MBPG method, i.e., IS-MBPG*, which also reaches the best known sample complexity of $O(\epsilon^{-3})$ without any large batches. In the experiments, we apply four benchmark tasks to demonstrate the effectiveness of our algorithms.

From Importance Sampling to Doubly Robust Policy Gradient

- Jiawei Huang, Nan Jiang
- abstract: We show that on-policy policy gradient (PG) and its variance reduction variants can be derived by taking finite-difference of function evaluations supplied by estimators from the importance sampling (IS) family for off-policy evaluation (OPE). Starting from the doubly robust (DR) estimator (Jiang & Li, 2016), we provide a simple derivation of a very general and flexible form of PG, which subsumes the state-of-the-art variance reduction technique (Cheng et al., 2019) as its special case and immediately hints at further variance reduction opportunities overlooked by existing literature. We analyze the variance of the new DR-PG estimator, compare it to existing methods as well as the Cramer-Rao lower bound of policy gradient, and empirically show its effectiveness.

Evaluating Lossy Compression Rates of Deep Generative Models

- Sicong Huang, Alireza Makhzani, Yanshuai Cao, Roger Grosse
- abstract: The field of deep generative modeling has succeeded in producing astonishingly realistic-seeming images and audio, but quantitative evaluation remains a challenge. Log-likelihood is an appealing metric due to its grounding in statistics and information theory, but it can be challenging to estimate for implicit generative models, and scalar-valued metrics give an incomplete picture of a model's quality. In this work, we propose to use rate distortion (RD) curves to evaluate and compare deep generative models. While estimating RD curves is seemingly even more computationally demanding than log-likelihood estimation, we show that we can approximate the entire RD curve using nearly the same computations as were previously used to achieve a single log-likelihood estimate. We evaluate lossy compression rates of VAEs, GANs, and adversarial autoencoders (AAEs) on the MNIST and CIFAR10 datasets. Measuring the entire RD curve gives a more complete picture than scalar-valued metrics, and we arrive at a number of insights not obtainable from log-likelihoods alone.

One Policy to Control Them All: Shared Modular Policies for Agent-Agnostic Control

- Wenlong Huang, Igor Mordatch, Deepak Pathak
- abstract: Reinforcement learning is typically concerned with learning control policies tailored to a particular agent. We investigate whether there exists a single global policy that can generalize to control a wide variety of agent morphologies – ones in which even dimensionality of state and action spaces changes. We propose to express this global policy as a collection of identical modular neural networks, dubbed as Shared Modular Policies (SMP), that correspond to each of the agent's actuators. Every module is only responsible for controlling its corresponding actuator and receives information from only its local sensors. In addition, messages are passed between modules, propagating information between distant modules. We show that a single modular policy can successfully generate locomotion behaviors for several planar agents with different skeletal structures such as monopod hoppers, quadrupeds, bipeds, and generalize to variants not seen during training – a process that would normally require training and manual hyperparameter tuning for each morphology. We observe that a wide variety of drastically diverse locomotion styles across morphologies as well as centralized coordination emerges via message passing between decentralized modules purely from the reinforcement learning objective. Videos and code at https://huangwl18.github.io/modular-rl/

Communication-Efficient Distributed PCA by Riemannian Optimization

- Long-Kai Huang, Sinno Pan
- abstract: In this paper, we study the leading eigenvector problem in a statistically distributed setting and propose a communication-efficient algorithm based on Riemannian optimization, which trades local computation for global communication. Theoretical analysis shows that the proposed algorithm linearly converges to the centralized empirical risk minimization solution regarding the number of communication rounds. When the number of data points in local machines is sufficiently large, the proposed algorithm achieves a significant reduction of communication cost over existing distributed PCA algorithms. Superior performance in terms of communication cost of the proposed algorithm is verified on real-world and synthetic datasets.

Improving Transformer Optimization Through Better Initialization

- Xiao Shi Huang, Felipe Perez, Jimmy Ba, Maksims Volkovs
- abstract: The Transformer architecture has achieved considerable success recently; the key component of the Transformer is the attention layer that enables the model to focus on important regions within an input sequence. Gradient optimization with attention layers can be notoriously difficult requiring tricks such as learning rate warmup to prevent divergence. As Transformer models are becoming larger and more expensive to train, recent research has focused on understanding and improving optimization in these architectures. In this work our contributions are two-fold: we first investigate and empirically validate the source of optimization problems in the encoder-decoder Transformer architecture; we then propose a new weight initialization scheme with theoretical justification, that enables training without warmup or layer normalization. Empirical results on public machine translation benchmarks show that our approach achieves leading accuracy, allowing to train deep Transformer models with 200 layers in both encoder and decoder (over 1000 attention/MLP blocks) without difficulty. Code for this work is available here: \url{https://github.com/layer6ai-labs/T-Fixup}.

More Information Supervised Probabilistic Deep Face Embedding Learning

- Ying Huang, Shangfeng Qiu, Wenwei Zhang, Xianghui Luo, Jinzhuo Wang
- abstract: Researches using margin based comparison loss demonstrate the effectiveness of penalizing the distance between face feature and their corresponding class centers. Despite their popularity and excellent performance, they do not explicitly encourage the generic embedding learning for an open set recognition problem. In this paper, we analyse margin based softmax loss in probability view. With this perspective, we propose two general principles: 1) monotonically decreasing and 2) margin probability penalty, for designing new margin loss functions. Unlike methods optimized with single comparison metric, we provide a new perspective to treat open set face recognition as a problem of information transmission. And the generalization capability for face embedding is gained with more clean information. An auto-encoder architecture called Linear-Auto-TS-Encoder(LATSE) is proposed to corroborate this finding. Extensive experiments on several benchmarks demonstrate that LATSE help face embedding to gain more generalization capability and it boost the single model performance with open training dataset to more than 99% on MegaFace test.

Generating Programmatic Referring Expressions via Program Synthesis

- Jiani Huang, Calvin Smith, Osbert Bastani, Rishabh Singh, Aws Albarghouthi, Mayur Naik
- abstract: Incorporating symbolic reasoning into machine learning algorithms is a promising approach to improve performance on learning tasks that require logical reasoning. We study the problem of generating a programmatic variant of referring expressions that we call referring relational programs. In particular, given a symbolic representation of an image and a target object in that image, the goal is to generate a relational program that uniquely identifies the target object in terms of its attributes and its relations to other objects in the image. We propose a neurosymbolic program synthesis algorithm that combines a policy neural network with enumerative search to generate such relational programs. The policy neural network employs a program interpreter that provides immediate feedback on the consequences of the decisions made by the policy, and also takes into account the uncertainty

in the symbolic representation of the image. We evaluate our algorithm on challenging benchmarks based on the CLEVR dataset, and demonstrate that our approach significantly outperforms several baselines.

InstaHide: Instance-hiding Schemes for Private Distributed Learning

- Yangsibo Huang, Zhao Song, Kai Li, Sanjeev Arora
- abstract: How can multiple distributed entities train a shared deep net on their private data while protecting data privacy? This paper introduces InstaHide, a simple encryption of training images. Encrypted images can be used in standard deep learning pipelines (PyTorch, Federated Learning etc.) with no additional setup or infrastructure. The encryption has a minor effect on test accuracy (unlike differential privacy). Encryption consists of mixing the image with a set of other images (in the sense of Mixup data augmentation technique (Zhang et al., 2018)) followed by applying a random pixel-wise mask on the mixed image. Other contributions of this paper are: (a) Use of large public dataset of images (e.g. ImageNet) for mixing during encryption; this improves security. (b) Experiments demonstrating effectiveness in protecting privacy against known attacks while preserving model accuracy. (c) Theoretical analysis showing that successfully attacking privacy requires attackers to solve a difficult computational problem. (d) Demonstration that Mixup alone is insecure as (contrary to recent proposals), by showing some efficient attacks. (e) Release of a challenge dataset to allow design of new attacks.

Accelerated Stochastic Gradient-free and Projection-free Methods

- Feihu Huang, Lue Tao, Songcan Chen
- abstract: In the paper, we propose a class of accelerated stochastic gradient-free and projection-free (a.k.a., zeroth-order Frank-Wolfe) methods to solve the constrained stochastic and finite-sum nonconvex optimization. Specifically, we propose an accelerated stochastic zeroth-order Frank-Wolfe (Acc-SZOFW) method based on the variance reduced technique of SPIDER/SpiderBoost and a novel momentum accelerated technique. Moreover, under some mild conditions, we prove that the Acc-SZOFW has the function query complexity of $O(d\sqrt{n}\epsilon^{-2})$ for finding an ϵ-stationary point in the finite-sum problem, which improves the exiting best result by a factor of $O(\sqrt{n}\epsilon^{-2})$, and has the function query complexity of $O(d\epsilon^{-3})$ in the stochastic problem, which improves the exiting best result by a factor of $O(\epsilon^{-1})$. To relax the large batches required in the Acc-SZOFW, we further propose a novel accelerated stochastic zeroth-order Frank-Wolfe (Acc-SZOFW*) based on a new variance reduced technique of STORM, which still reaches the function query complexity of $O(d\epsilon^{-3})$ in the stochastic problem without relying on any large batches. In particular, we present an accelerated framework of the Frank-Wolfe methods based on the proposed momentum accelerated technique. The extensive experimental results on black-box adversarial attack and robust black-box classification demonstrate the efficiency of our algorithms.

Deep Graph Random Process for Relational-Thinking-Based Speech Recognition

- Hengguan Huang, Fuzhao Xue, Hao Wang, Ye Wang
- abstract: Lying at the core of human intelligence, relational thinking is characterized by initially relying on innumerable unconscious percepts pertaining to relations between new sensory signals and prior knowledge, consequently becoming a recognizable concept or object through coupling and transformation of these percepts. Such mental processes are difficult to model in real-world problems such as in conversational automatic speech recognition (ASR), as the percepts (if they are modelled as graphs indicating relationships among utterances) are supposed to be innumerable and not directly observable. In this paper, we present a Bayesian nonparametric deep learning method called deep graph random process (DGP) that can generate an infinite number of probabilistic graphs representing percepts. We further provide a closed-form solution for coupling and transformation of these percept graphs for acoustic modeling. Our approach is able to successfully infer relations among utterances without using any relational data during training. Experimental evaluations on ASR tasks including CHiME-2 and CHiME-5 demonstrate the effectiveness and benefits of our method.

Dynamics of Deep Neural Networks and Neural Tangent Hierarchy

- Jiaoyang Huang, Horng-Tzer Yau
- abstract: The evolution of a deep neural network trained by the gradient descent in the overparametrization regime can be described by its neural tangent kernel (NTK) \cite{jacot2018neural, du2018gradient1,du2018gradient2,arora2019fine}. It was observed \cite{arora2019exact} that there is a performance gap between the kernel regression using the limiting NTK and the deep neural networks. We study the dynamic of neural networks of finite width and derive an infinite hierarchy of differential equations, the neural tangent hierarchy (NTH). We prove that the NTH hierarchy truncated at the level $p\geq 2$ approximates the dynamic of the NTK up to arbitrary precision under certain conditions on the neural network width and the data set dimension. The assumptions needed for these approximations become weaker as p increases. Finally, NTH can be viewed as higher order extensions of NTK. In particular, the NTH truncated at $p=2$ recovers the NTK dynamics.

Curvature-corrected learning dynamics in deep neural networks

- Dongsung Huh
- abstract: Deep neural networks exhibit complex learning dynamics due to their non-convex loss landscapes. Second-order optimization methods facilitate learning dynamics by compensating for ill-conditioned curvature. In this work, we investigate how curvature correction modifies the learning dynamics in deep linear neural networks and provide analytical solutions. We derive a generalized conservation law that preserves the path of parameter dynamics from curvature correction, which shows that curvature correction only modifies the temporal profiles of dynamics along the path. We show that while curvature correction accelerates the convergence dynamics of the input-output map, it can also negatively affect the generalization performance. Our analysis also reveals an undesirable effect of curvature correction that compromises stability of parameters dynamics during learning, especially with block-diagonal approximation of natural gradient descent. We introduce fractional curvature correction that resolves this problem while retaining most of the acceleration benefits of full curvature correction.

Multigrid Neural Memory

- Tri Huynh, Michael Maire, Matthew Walter
- abstract: We introduce a novel approach to endowing neural networks with emergent, long-term, large-scale memory. Distinct from strategies that connect neural networks to external memory banks via intricately crafted controllers and hand-designed attentional mechanisms, our memory is internal, distributed, co-located alongside computation, and implicitly addressed, while being drastically simpler than prior efforts. Architecting networks with multigrid structure and connectivity, while distributing memory cells alongside computation throughout this topology, we observe the emergence of coherent memory subsystems. Our hierarchical spatial organization, parameterized convolutionally, permits efficient instantiation of large-capacity memories, while multigrid topology provides short internal routing pathways, allowing convolutional networks to efficiently approximate the behavior of fully connected networks. Such networks have an implicit capacity for internal attention; augmented with memory, they learn to read and write specific memory locations in a dynamic data-dependent manner. We demonstrate these capabilities on exploration and mapping tasks, where our network is able to self-organize and retain long-term memory for trajectories of thousands of time steps. On tasks decoupled from any notion of spatial geometry: sorting, associative recall, and question answering, our design functions as a truly generic memory and yields excellent results.

Meta-Learning with Shared Amortized Variational Inference

- Ekaterina Iakovleva, Jakob Verbeek, Karteek Alahari
- abstract: We propose a novel amortized variational inference scheme for an empirical Bayes meta-learning model, where model parameters are treated as latent variables. We learn the prior distribution over model parameters conditioned on limited training data using a variational autoencoder approach. Our framework proposes sharing the same amortized inference network between the conditional prior and variational posterior distributions over the model parameters. While the posterior leverages both the labeled support and query data, the conditional prior is based only on the labeled support data. We show that in earlier work, relying on Monte-Carlo approximation, the conditional prior collapses to a Dirac delta function. In contrast, our variational approach prevents this collapse and preserves uncertainty over the model parameters. We evaluate our approach on the miniImageNet, CIFAR-FS and FC100 datasets, and present results demonstrating its advantages over previous work.

Linear Lower Bounds and Conditioning of Differentiable Games

- Adam Ibrahim, Waïss Azizian, Gauthier Gidel, Ioannis Mitliagkas
- abstract: Recent successes of game-theoretic formulations in ML have caused a resurgence of research interest in differentiable games. Overwhelmingly, that research focuses on methods and upper bounds on their speed of convergence. In this work, we approach the question of fundamental iteration complexity by providing lower bounds to complement the linear (i.e. geometric) upper bounds observed in the literature on a wide class of problems. We cast saddle-point and min-max problems as 2-player games. We leverage tools from single-objective convex optimisation to propose new linear lower bounds for convex-concave games. Notably, we give a linear lower bound for n-player differentiable games, by using the spectral properties of the update operator. We then propose a new definition of the condition number arising from our lower bound analysis. Unlike past definitions, our condition number captures the fact that linear rates are possible in games, even in the absence of strong convexity or strong concavity in the variables.

Fast Deterministic CUR Matrix Decomposition with Accuracy Assurance

- Yasutoshi Ida, Sekitoshi Kanai, Yasuhiro Fujiwara, Tomoharu Iwata, Koh Takeuchi, Hisashi Kashima
- abstract: The deterministic CUR matrix decomposition is a low-rank approximation method to analyze a data matrix. It has attracted considerable attention due to its high interpretability, which results from the fact that the decomposed matrices consist of subsets of the original columns and rows of the data matrix. The subset is obtained by optimizing an objective function with sparsity-inducing norms via coordinate descent. However, the existing algorithms for optimization incur high computation costs. This is because coordinate descent iteratively updates all the parameters in the objective until convergence. This paper proposes a fast deterministic CUR matrix decomposition. Our algorithm safely skips unnecessary updates by efficiently evaluating the optimality conditions for the parameters to be zeros. In addition, we preferentially update the parameters that must be nonzeros. Theoretically, our approach guarantees the same result as the original approach. Experiments demonstrate that our algorithm speeds up the deterministic CUR while achieving the same accuracy.

Do We Need Zero Training Loss After Achieving Zero Training Error?

- Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, Masashi Sugiyama
- abstract: Overparameterized deep networks have the capacity to memorize training data with zero \emph{training error}. Even after memorization, the \emph{training loss} continues to approach zero, making the model overconfident and the test performance degraded. Since existing regularizers do not directly aim to avoid zero training loss, it is hard to tune their hyperparameters in order to maintain a fixed/preset level of training loss. We propose a direct solution called \emph{flooding} that intentionally prevents further reduction of the training loss when it reaches a reasonably small value, which we call the \emph{flood level}. Our approach makes the loss float around the flood level by doing mini-batched gradient descent as usual but gradient ascent if the training loss is below the flood level. This can be implemented with one line of code and is compatible with any stochastic optimizer and other regularizers. With flooding, the model will continue to “random walk” with the same non-zero training loss, and we expect it to drift into an area with a flat loss landscape that leads to better generalization. We experimentally show that flooding improves performance and, as a byproduct, induces a double descent curve of the test loss.

Semi-Supervised Learning with Normalizing Flows

- Pavel Izmailov, Polina Kirichenko, Marc Finzi, Andrew Gordon Wilson
- abstract: Normalizing flows transform a latent distribution through an invertible neural network for a flexible and pleasingly simple approach to generative modelling, while preserving an exact likelihood. We propose FlowGMM, an end-to-end approach to generative semi supervised learning with normalizing flows, using a latent Gaussian mixture model. FlowGMM is distinct in its simplicity, unified treatment of labelled and unlabelled data with an exact likelihood, interpretability, and broad applicability beyond image data. We show promising results on a wide range of applications, including AG-News and Yahoo Answers text data, tabular data, and semi-supervised image classification. We also show that FlowGMM can discover interpretable structure, provide real-time optimization-free feature visualizations, and specify well calibrated predictive distributions.

Implicit Regularization of Random Feature Models

- Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, Franck Gabriel
- abstract: Random Features (RF) models are used as efficient parametric approximations of kernel methods. We investigate, by means of random matrix theory, the connection between Gaussian RF models and Kernel Ridge Regression (KRR). For a Gaussian RF model with P features, N data points, and a ridge λ, we show that the average (i.e. expected) RF predictor is close to a KRR predictor with an \emph{effective ridge} $\tilde{\lambda}$. We show that $\tilde{\lambda} > \lambda$ and $\tilde{\lambda} \searrow \lambda$ monotonically as P grows, thus revealing the \emph{implicit regularization effect} of finite RF sampling. We then compare the risk (i.e. test error) of the $\tilde{\lambda}$-KRR predictor with the average risk of the λ-RF predictor and obtain a precise and explicit bound on their difference. Finally, we empirically find an extremely good agreement between the test errors of the average λ-RF predictor and $\tilde{\lambda}$-KRR predictor.

Correlation Clustering with Asymmetric Classification Errors

- Jafar Jafarov, Sanchit Kalhan, Konstantin Makarychev, Yury Makarychev
- abstract: In the Correlation Clustering problem, we are given a weighted graph G with its edges labelled as "similar" or "dissimilar" by a binary classifier. The goal is to produce a clustering that minimizes the weight of "disagreements": the sum of the weights of "similar" edges across clusters and "dissimilar" edges within clusters. We study the correlation clustering problem under the following assumption: Every "similar" edge e has weight $w_e \in [\alpha w, w]$ and every "dissimilar" edge e has weight $w_e \geq \alpha w$ (where $\alpha \leq 1$ and $w > 0$ is a scaling parameter). We give a $(3 + 2 \log_e (1/\alpha))$ approximation algorithm for this problem. This assumption captures well the scenario when classification errors are asymmetric. Additionally, we show an asymptotically matching Linear Programming integrality gap of $\Omega(\log 1/\alpha)$.

Optimal Robust Learning of Discrete Distributions from Batches

- Ayush Jain, Alon Orlitsky
- abstract: Many applications, including natural language processing, sensor networks, collaborative filtering, and federated learning, call for estimating discrete distributions from data collected in batches, some of which may be untrustworthy, erroneous, faulty, or even adversarial. Previous estimators for

this setting ran in exponential time, and for some regimes required a suboptimal number of batches. We provide the first polynomial-time estimator that is optimal in the number of batches and achieves essentially the best possible estimation accuracy.

Generalization to New Actions in Reinforcement Learning

- Ayush Jain, Andrew Szot, Joseph Lim
- abstract: A fundamental trait of intelligence is the ability to achieve goals in the face of novel circumstances, such as making decisions from new action choices. However, standard reinforcement learning assumes a fixed set of actions and requires expensive retraining when given a new action set. To make learning agents more adaptable, we introduce the problem of zero-shot generalization to new actions. We propose a two-stage framework where the agent first infers action representations from action information acquired separately from the task. A policy flexible to varying action sets is then trained with generalization objectives. We benchmark generalization on sequential tasks, such as selecting from an unseen tool-set to solve physical reasoning puzzles and stacking towers with novel 3D shapes. Videos and code are available at https://sites.google.com/view/action-generalization.

Tails of Lipschitz Triangular Flows

- Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, Marcus Brubaker
- abstract: We investigate the ability of popular flow models to capture tail-properties of a target density by studying the increasing triangular maps used in these flow methods acting on a tractable source density. We show that the density quantile functions of the source and target density provide a precise characterization of the slope of transformation required to capture tails in a target density. We further show that any Lipschitz-continuous transport map acting on a source density will result in a density with similar tail properties as the source, highlighting the trade-off between the importance of choosing a complex source density and a sufficiently expressive transformation to capture desirable properties of a target density. Subsequently, we illustrate that flow models like Real-NVP, MAF, and Glow as implemented lack the ability to capture a distribution with non-Gaussian tails. We circumvent this problem by proposing tail-adaptive flows consisting of a source distribution that can be learned simultaneously with the triangular map to capture tail-properties of a target density. We perform several synthetic and real-world experiments to complement our theoretical findings.

Learning Portable Representations for High-Level Planning

- Steven James, Benjamin Rosman, George Konidaris
- abstract: We present a framework for autonomously learning a portable representation that describes a collection of low-level continuous environments. We show that these abstract representations can be learned in a task-independent egocentric space specific to the agent that, when grounded with problem-specific information, are provably sufficient for planning. We demonstrate transfer in two different domains, where an agent learns a portable, task-independent symbolic vocabulary, as well as operators expressed in that vocabulary, and then learns to instantiate those operators on a per-task basis. This reduces the number of samples required to learn a representation of a new task.

Debiased Sinkhorn barycenters

- Hicham Janati, Marco Cuturi, Alexandre Gramfort
- abstract: Entropy regularization in optimal transport (OT) has been the driver of many recent interests for Wasserstein metrics and barycenters in machine learning. It allows to keep the appealing geometrical properties of the unregularized Wasserstein distance while having a significantly lower complexity thanks to Sinkhorn's algorithm. However, entropy brings some inherent smoothing bias, resulting for example in blurred barycenters. This side effect has prompted an increasing temptation in the community to settle for a slower algorithm such as log-domain stabilized Sinkhorn which breaks the parallel structure that can be leveraged on GPUs, or even go back to unregularized OT. Here we show how this bias is tightly linked to the reference measure that defines the entropy regularizer and propose debiased Sinkhorn barycenters that preserve the best of worlds: fast Sinkhorn-like iterations without entropy smoothing. Theoretically, we prove that this debiasing is perfect for Gaussian distributions with equal variance. Empirically, we illustrate the reduced blurring and the computational advantage.

Parametric Gaussian Process Regressors

- Martin Jankowiak, Geoff Pleiss, Jacob Gardner
- abstract: The combination of inducing point methods with stochastic variational inference has enabled approximate Gaussian Process (GP) inference on large datasets. Unfortunately, the resulting predictive distributions often exhibit substantially underestimated uncertainties. Notably, in the regression case the predictive variance is typically dominated by observation noise, yielding uncertainty estimates that make little use of the input-dependent function uncertainty that makes GP priors attractive. In this work we propose two simple methods for scalable GP regression that address this issue and thus yield substantially improved predictive uncertainties. The first applies variational inference to FITC (Fully Independent Training Conditional; Snelson et. al. 2006). The second bypasses posterior approximations and instead directly targets the posterior predictive distribution. In an extensive empirical comparison with a number of alternative methods for scalable GP regression, we find that the resulting predictive distributions exhibit significantly better calibrated uncertainties and higher log likelihoods–often by as much as half a nat per datapoint.

Inverse Active Sensing: Modeling and Understanding Timely Decision-Making

- Daniel Jarrett, Mihaela Van Der Schaar
- abstract: Evidence-based decision-making entails collecting (costly) observations about an underlying phenomenon of interest, and subsequently committing to an (informed) decision on the basis of accumulated evidence. In this setting, *active sensing* is the goal-oriented problem of efficiently selecting which acquisitions to make, and when and what decision to settle on. As its complement, *inverse active sensing* seeks to uncover an agent's preferences and strategy given their observable decision-making behavior. In this paper, we develop an expressive, unified framework for the general setting of evidence-based decision-making under endogenous, context-dependent time pressure—which requires negotiating (subjective) tradeoffs between accuracy, speediness, and cost of information. Using this language, we demonstrate how it enables *modeling* intuitive notions of surprise, suspense, and optimality in decision strategies (the forward problem). Finally, we illustrate how this formulation enables *understanding* decision-making behavior by quantifying preferences implicit in observed decision strategies (the inverse problem).

Source Separation with Deep Generative Priors

- Vivek Jayaram, John Thickstun
- abstract: Despite substantial progress in signal source separation, results for richly structured data continue to contain perceptible artifacts. In contrast, recent deep generative models can produce authentic samples in a variety of domains that are indistinguishable from samples of the data distribution. This paper introduces a Bayesian approach to source separation that uses deep generative models as priors over the components of a mixture of sources, and noise-annealed Langevin dynamics to sample from the posterior distribution of sources given a mixture. This decouples the source separation problem from generative modeling, enabling us to directly use cutting-edge generative models as priors. The method achieves state-of-the-art performance for MNIST digit separation. We introduce new methodology for evaluating separation quality on richer datasets, providing quantitative evaluation and qualitative discussion of results for CIFAR-10 image separation. We also provide qualitative results on LSUN.

Extra-gradient with player sampling for faster convergence in n-player games

- Samy Jelassi, Carles Domingo-Enrich, Damien Scieur, Arthur Mensch, Joan Bruna
- abstract: Data-driven modeling increasingly requires to find a Nash equilibrium in multi-player games, e.g. when training GANs. In this paper, we analyse a new extra-gradient method for Nash equilibrium finding, that performs gradient extrapolations and updates on a random subset of players at each iteration. This approach provably exhibits a better rate of convergence than full extra-gradient for non-smooth convex games with noisy gradient oracle. We propose an additional variance reduction mechanism to obtain speed-ups in smooth convex games. Our approach makes extrapolation amenable to massive multiplayer settings, and brings empirical speed-ups, in particular when using a heuristic cyclic sampling scheme. Most importantly, it allows to train faster and better GANs and mixtures of GANs.

T-GD: Transferable GAN-generated Images Detection Framework

- Hyeonseong Jeon, Young Oh Bang, Junyaup Kim, Simon Woo
- abstract: Recent advancements in Generative Adversarial Networks (GANs) enable the generation of highly realistic images, raising concerns about their misuse for malicious purposes. Detecting these GAN-generated images (GAN-images) becomes increasingly challenging due to the significant reduction of underlying artifacts and specific patterns. The absence of such traces can hinder detection algorithms from identifying GAN-images and transferring knowledge to identify other types of GAN-images as well. In this work, we present the Transferable GAN-images Detection framework T-GD, a robust transferable framework for an effective detection of GAN-images. T-GD is composed of a teacher and a student model that can iteratively teach and evaluate each other to improve the detection performance. First, we train the teacher model on the source dataset and use it as a starting point for learning the target dataset. To train the student model, we inject noise by mixing up the source and target datasets, while constraining the weight variation to preserve the starting point. Our approach is a self-training method, but distinguishes itself from prior approaches by focusing on improving the transferability of GAN-image detection. T-GD achieves high performance on the source dataset by overcoming catastrophic forgetting and effectively detecting state-of-the-art GAN-images with only a small volume of data without any metadata information.

History-Gradient Aided Batch Size Adaptation for Variance Reduced Algorithms

- Kaiyi Ji, Zhe Wang, Bowen Weng, Yi Zhou, Wei Zhang, Yingbin Liang
- abstract: Variance-reduced algorithms, although achieve great theoretical performance, can run slowly in practice due to the periodic gradient estimation with a large batch of data. Batch-size adaptation thus arises as a promising approach to accelerate such algorithms. However, existing schemes either apply prescribed batch-size adaption rule or exploit the information along optimization path via additional backtracking and condition verification steps. In this paper, we propose a novel scheme, which eliminates backtracking line search but still exploits the information along optimization path by adapting the batch size via history stochastic gradients. We further theoretically show that such a scheme substantially reduces the overall complexity for popular variance-reduced algorithms SVRG and SARAH/SPIDER for both conventional nonconvex optimization and reinforcement learning problems. To this end, we develop a new convergence analysis framework to handle the dependence of the batch size on history stochastic gradients. Extensive experiments validate the effectiveness of the proposed batch-size adaptation scheme.

Information-Theoretic Local Minima Characterization and Regularization

- Zhiwei Jia, Hao Su
- abstract: Recent advances in deep learning theory have evoked the study of generalizability across different local minima of deep neural networks (DNNs). While current work focused on either discovering properties of good local minima or developing regularization techniques to induce good local minima, no approach exists that can tackle both problems. We achieve these two goals successfully in a unified manner. Specifically, based on the observed Fisher information we propose a metric both strongly indicative of generalizability of local minima and effectively applied as a practical regularizer. We provide theoretical analysis including a generalization bound and empirically demonstrate the success of our approach in both capturing and improving the generalizability of DNNs. Experiments are performed on CIFAR-10, CIFAR-100 and ImageNet for various network architectures.

Optimizing Black-box Metrics with Adaptive Surrogates

- Qijia Jiang, Olaoluwa Adigun, Harikrishna Narasimhan, Mahdi Milani Fard, Maya Gupta
- abstract: We address the problem of training models with black-box and hard-to-optimize metrics by expressing the metric as a monotonic function of a small number of easy-to-optimize surrogates. We pose the training problem as an optimization over a relaxed surrogate space, which we solve by estimating local gradients for the metric and performing inexact convex projections. We analyze gradient estimates based on finite differences and local linear interpolations, and show convergence of our approach under smoothness assumptions with respect to the surrogates. Experimental results on classification and ranking problems verify the proposal performs on par with methods that know the mathematical formulation, and adds notable value when the form of the metric is unknown.

BINOCULARS for efficient, nonmyopic sequential experimental design

- Shali Jiang, Henry Chai, Javier Gonzalez, Roman Garnett
- abstract: Finite-horizon sequential experimental design (SED) arises naturally in many contexts, including hyperparameter tuning in machine learning among more traditional settings. Computing the optimal policy for such problems requires solving Bellman equations, which are generally intractable. Most existing work resorts to severely myopic approximations by limiting the decision horizon to only a single time-step, which can underweight exploration in favor of exploitation. We present BINOCULARS: Batch-Informed NOnmyopic Choices, Using Long-horizons for Adaptive, Rapid SED, a general framework for deriving efficient, nonmyopic approximations to the optimal experimental policy. Our key idea is simple and surprisingly effective: we first compute a one-step optimal batch of experiments, then select a single point from this batch to evaluate. We realize BINOCULARS for Bayesian optimization and Bayesian quadrature – two notable example problems with radically different objectives – and demonstrate that BINOCULARS significantly outperforms significantly outperforms myopic alternatives in real-world scenarios.

Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels

- Lu Jiang, Di Huang, Mason Liu, Weilong Yang
- abstract: Performing controlled experiments on noisy data is essential in understanding deep learning across noise levels. Due to the lack of suitable datasets, previous research has only examined deep learning on controlled synthetic label noise, and real-world label noise has never been studied in a controlled setting. This paper makes three contributions. First, we establish the first benchmark of controlled real-world label noise from the web. This new benchmark enables us to study the web label noise in a controlled setting for the first time. The second contribution is a simple but effective method to overcome both synthetic and real noisy labels. We show that our method achieves the best result on our dataset as well as on two public benchmarks (CIFAR and WebVision). Third, we conduct the largest study by far into understanding deep neural networks trained on noisy labels across different noise levels, noise types, network architectures, and training settings.

Implicit Class-Conditioned Domain Alignment for Unsupervised Domain Adaptation

- Xiang Jiang, Qicheng Lao, Stan Matwin, Mohammad Havaei
- abstract: We present an approach for unsupervised domain adaptation{—}with a strong focus on practical considerations of within-domain class imbalance and between-domain class distribution shift{—}from a class-conditioned domain alignment perspective. Current methods for class-conditioned domain alignment aim to explicitly minimize a loss function based on pseudo-label estimations of the target domain. However, these methods suffer from pseudo-label bias in the form of error accumulation. We propose a method that removes the need for explicit optimization of model parameters from pseudo-labels. Instead, we present a sampling-based implicit alignment approach, where the sample selection is implicitly guided by the pseudo-labels. Theoretical analysis reveals the existence of a domain-discriminator shortcut in misaligned classes, which is addressed by the proposed approach to facilitate domain-adversarial learning. Empirical results and ablation studies confirm the effectiveness of the proposed approach, especially in the presence of within-domain class imbalance and between-domain class distribution shift.

Associative Memory in Iterated Overparameterized Sigmoid Autoencoders

- Yibo Jiang, Cengiz Pehlevan
- abstract: Recent work showed that overparameterized autoencoders can be trained to implement associative memory via iterative maps, when the trained input-output Jacobian of the network has all of its eigenvalue norms strictly below one. Here, we theoretically analyze this phenomenon for sigmoid networks by leveraging recent developments in deep learning theory, especially the correspondence between training neural networks in the infinite-width limit and performing kernel regression with the Neural Tangent Kernel (NTK). We find that overparameterized sigmoid autoencoders can have attractors in the NTK limit for both training with a single example and multiple examples under certain conditions. In particular, for multiple training examples, we find that the norm of the largest Jacobian eigenvalue drops below one with increasing input norm, leading to associative memory.

Hierarchical Generation of Molecular Graphs using Structural Motifs

- Wengong Jin, Dr.Regina Barzilay, Tommi Jaakkola
- abstract: Graph generation techniques are increasingly being adopted for drug discovery. Previous graph generation approaches have utilized relatively small molecular building blocks such as atoms or simple cycles, limiting their effectiveness to smaller molecules. Indeed, as we demonstrate, their performance degrades significantly for larger molecules. In this paper, we propose a new hierarchical graph encoder-decoder that employs significantly larger and more flexible graph motifs as basic building blocks. Our encoder produces a multi-resolution representation for each molecule in a fine-to-coarse fashion, from atoms to connected motifs. Each level integrates the encoding of constituents below with the graph at that level. Our autoregressive coarse-to-fine decoder adds one motif at a time, interleaving the decision of selecting a new motif with the process of resolving its attachments to the emerging molecule. We evaluate our model on multiple molecule generation tasks, including polymers, and show that our model significantly outperforms previous state-of-the-art baselines.

Multi-Objective Molecule Generation using Interpretable Substructures

- Wengong Jin, Dr.Regina Barzilay, Tommi Jaakkola
- abstract: Drug discovery aims to find novel compounds with specified chemical property profiles. In terms of generative modeling, the goal is to learn to sample molecules in the intersection of multiple property constraints. This task becomes increasingly challenging when there are many property constraints. We propose to offset this complexity by composing molecules from a vocabulary of substructures that we call molecular rationales. These rationales are identified from molecules as substructures that are likely responsible for each property of interest. We then learn to expand rationales into a full molecule using graph generative models. Our final generative model composes molecules as mixtures of multiple rationale completions, and this mixture is fine-tuned to preserve the properties of interest. We evaluate our model on various drug design tasks and demonstrate significant improvements over state-of-the-art baselines in terms of accuracy, diversity, and novelty of generated compounds.

Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition

- Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, Tiancheng Yu
- abstract: We consider the task of learning in episodic finite-horizon Markov decision processes with an unknown transition function, bandit feedback, and adversarial losses. We propose an efficient algorithm that achieves $\mathcal{\tilde{O}}(L|X|\sqrt{|A|T})$ regret with high probability, where L is the horizon, $|X|$ the number of states, $|A|$ the number of actions, and T the number of episodes. To our knowledge, our algorithm is the first to ensure $\mathcal{\tilde{O}}(\sqrt{T})$ regret in this challenging setting; in fact, it achieves the same regret as (Rosenberg & Mansour, 2019a) who consider the easier setting with full-information. Our key contributions are two-fold: a tighter confidence set for the transition function; and an optimistic loss estimator that is inversely weighted by an "upper occupancy bound".

Reward-Free Exploration for Reinforcement Learning

- Chi Jin, Akshay Krishnamurthy, Max Simchowitz, Tiancheng Yu
- abstract: Exploration is widely regarded as one of the most challenging aspects of reinforcement learning (RL), with many naive approaches succumbing to exponential sample complexity. To isolate the challenges of exploration, we propose the following “reward-free RL” framework. In the exploration phase, the agent first collects trajectories from an MDP M without a pre-specified reward function. After exploration, it is tasked with computing a near-policies under the transitions of $\mathcal{M}$ for a collection of given reward functions. This framework is particularly suitable where there are many reward functions of interest, or where the reward function is shaped by an external agent to elicit desired behavior. We give an efficient algorithm that conducts $\widetilde{O}(S^2A\mathrm{poly}(H)/\epsilon^2)$ episodes of exploration, and returns ϵ-suboptimal policies for an arbitrary number of reward functions. We achieve this by finding exploratory policies that jointly visit each “significant” state with probability proportional to its maximum visitation probability under any possible policy. Moreover, our planning procedure can be instantiated by any black-box approximate planner, such as value iteration or natural policy gradient. Finally, we give a nearly-matching $\Omega(S^2AH^2/\epsilon^2)$ lower bound, demonstrating the near-optimality of our algorithm in this setting.

What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?

- Chi Jin, Praneeth Netrapalli, Michael Jordan
- abstract: Minimax optimization has found extensive applications in modern machine learning, in settings such as generative adversarial networks (GANs), adversarial training and multi-agent reinforcement learning. As most of these applications involve continuous nonconvex-nonconcave formulations, a very basic question arises—“what is a proper definition of local optima?” Most previous work answers this question using classical notions of equilibria from simultaneous games, where the min-player and the max-player act simultaneously. In contrast, most applications in machine learning, including GANs and adversarial training, correspond to sequential games, where the order of which player acts first is crucial (since minimax is in general not equal to maximin due to the nonconvex-nonconcave nature of the problems). The main contribution of this paper is to propose a proper mathematical definition of local optimality for this sequential setting—local minimax, as well as to present its properties and existence results. Finally, we establish a strong connection to a basic local search algorithm—gradient descent ascent (GDA): under mild conditions, all stable limit points of GDA are exactly local minimax points up to some degenerate points.

Efficiently Solving MDPs with Stochastic Mirror Descent

- Yujia Jin, Aaron Sidford
- abstract: We present a unified framework based on primal-dual stochastic mirror descent for approximately solving infinite-horizon Markov decision processes (MDPs) given a generative model. When applied to an average-reward MDP with A_{tot} total actions and mixing time bound t_{mix} our method computes an ϵ-optimal policy with an expected $\widetilde{O}(t_{mix}^2 A_{tot} \epsilon^{-2})$ samples from the state-transition matrix, removing the ergodicity dependence of prior art. When applied to a γ-discounted MDP with A_{tot} total actions our method computes an ϵ-optimal policy with an expected $\widetilde{O}((1-\gamma)^{-4} A_{tot} \epsilon^{-2})$ samples, improving over the best-known primal-dual methods while matching the state-of-the-art up to a $(1-\gamma)^{-1}$ factor. Both methods are model-free, update state values and policies simultaneously, and run in time linear in the number of samples taken. We achieve these results through a more general stochastic mirror descent framework for solving bilinear saddle-point problems with simplex and box domains and we demonstrate the flexibility of this framework by providing further applications to constrained MDPs.

Computational and Statistical Tradeoffs in Inferring Combinatorial Structures of Ising Model

- Ying Jin, Zhaoran Wang, Junwei Lu
- abstract: We study the computational and statistical tradeoffs in inferring combinatorial structures of high dimensional simple zero-field ferromagnetic Ising model. Under the framework of oracle computational model where an algorithm interacts with an oracle that discourses a randomized version of truth, we characterize the computational lower bounds of learning combinatorial structures in polynomial time, under which no algorithms within polynomial-time can distinguish between graphs with and without certain structures. This hardness of learning with limited computational budget is shown to be characterized by a novel quantity called vertex overlap ratio. Such quantity is universally valid for many specific graph structures including cliques and nearest neighbors. On the other side, we attain the optimal rates for testing these structures against empty graph by proposing the quadratic testing statistics to match the lower bounds. We also investigate the relationship between computational bounds and information-theoretic bounds for such problems, and found gaps between the two boundaries in inferring some particular structures, especially for those with dense edges.

AdaScale SGD: A User-Friendly Algorithm for Distributed Training

- Tyler Johnson, Pulkit Agrawal, Haijie Gu, Carlos Guestrin
- abstract: When using large-batch training to speed up stochastic gradient descent, learning rates must adapt to new batch sizes in order to maximize speed-ups and preserve model quality. Re-tuning learning rates is resource intensive, while fixed scaling rules often degrade model quality. We propose AdaScale SGD, an algorithm that reliably adapts learning rates to large-batch training. By continually adapting to the gradient's variance, AdaScale automatically achieves speed-ups for a wide range of batch sizes. We formally describe this quality with AdaScale's convergence bound, which maintains final objective values, even as batch sizes grow large and the number of iterations decreases. In empirical comparisons, AdaScale trains well beyond the batch size limits of popular "linear learning rate scaling" rules. This includes large-batch training with no model degradation for machine translation, image classification, object detection, and speech recognition tasks. AdaScale's qualitative behavior is similar to that of "warm-up" heuristics, but unlike warm-up, this behavior emerges naturally from a principled mechanism. The algorithm introduces negligible computational overhead and no new hyperparameters, making AdaScale an attractive choice for large-scale training in practice.

Guided Learning of Nonconvex Models through Successive Functional Gradient Optimization

- Rie Johnson, Tong Zhang
- abstract: This paper presents a framework of successive functional gradient optimization for training nonconvex models such as neural networks, where training is driven by mirror descent in a function space. We provide a theoretical analysis and empirical study of the training method derived from this framework. It is shown that the method leads to better performance than that of standard training techniques.

On Relativistic f-Divergences

- Alexia Jolicoeur-Martineau
- abstract: We take a more rigorous look at Relativistic Generative Adversarial Networks (RGANs) and prove that the objective function of the discriminator is a statistical divergence for any concave function f with minimal properties. We devise additional variants of relativistic f-divergences. We show that the Wasserstein distance is weaker than f-divergences which are weaker than relativistic f-divergences. Given the good performance of RGANs, this suggests that Wasserstein GAN does not performs well primarily because of the weak metric, but rather because of regularization and the use of a relativistic discriminator. We introduce the minimum-variance unbiased estimator (MVUE) for Relativistic paired GANs (RpGANs; originally called RGANs which could bring confusion) and show that it does not perform better. We show that the estimator of Relativistic average GANs (RaGANs) is asymptotically unbiased and that the finite-sample bias is small; removing this bias does not improve performance.

Fair k-Centers via Maximum Matching

- Matthew Jones, Huy Nguyen, Thy Nguyen
- abstract: The field of algorithms has seen a push for fairness, or the removal of inherent bias, in recent history. In data summarization, where a much smaller subset of a data set is chosen to represent the whole of the data, fairness can be introduced by guaranteeing each "demographic group" a specific portion of the representative subset. Specifically, this paper examines this fair variant of the k-centers problem, where a subset of the data with cardinality k is chosen to minimize distance to the rest of the data. Previous papers working on this problem presented both a 3-approximation algorithm with a super-linear runtime and a linear-time algorithm whose approximation factor is exponential in the number of demographic groups. This paper combines the best of each algorithm by presenting a linear-time algorithm with a guaranteed 3-approximation factor and provides empirical evidence of both the algorithm's runtime and effectiveness.

Being Bayesian about Categorical Probability

- Taejong Joo, Uijung Chung, Min-Gwan Seo
- abstract: Neural networks utilize the softmax as a building block in classification tasks, which contains an overconfidence problem and lacks an uncertainty representation ability. As a Bayesian alternative to the softmax, we consider a random variable of a categorical probability over class labels. In this framework, the prior distribution explicitly models the presumed noise inherent in the observed label, which provides consistent gains in generalization performance in multiple challenging tasks. The proposed method inherits advantages of Bayesian approaches that achieve better uncertainty estimation and model calibration. Our method can be implemented as a plug-and-play loss function with negligible computational overhead compared to the softmax with the cross-entropy loss function.

Evaluating the Performance of Reinforcement Learning Algorithms

- Scott Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, Philip Thomas
- abstract: Performance evaluations are critical for quantifying algorithmic advances in reinforcement learning. Recent reproducibility analyses have shown that reported performance results are often inconsistent and difficult to replicate. In this work, we argue that the inconsistency of performance stems from the use of flawed evaluation metrics. Taking a step towards ensuring that reported results are consistent, we propose a new comprehensive evaluation

methodology for reinforcement learning algorithms that produces reliable measurements of performance both on a single environment and when aggregated across environments. We demonstrate this method by evaluating a broad class of reinforcement learning algorithms on standard benchmark tasks.

Stochastic Differential Equations with Variational Wishart Diffusions

- Martin Jørgensen, Marc Deisenroth, Hugh Salimbeni
- abstract: We present a Bayesian non-parametric way of inferring stochastic differential equations for both regression tasks and continuous-time dynamical modelling. The work has high emphasis on the stochastic part of the differential equation, also known as the diffusion, and modelling it by means of Wishart processes. Further, we present a semiparametric approach that allows the framework to scale to high dimensions. This successfully leads us onto how to model both latent and autoregressive temporal systems with conditional heteroskedastic noise. We provide experimental evidence that modelling diffusion often improves performance and that this randomness in the differential equation can be essential to avoid overfitting.

A simpler approach to accelerated optimization: iterative averaging meets optimism

- Pooria Joulani, Anant Raj, Andras Gyorgy, Csaba Szepesvari
- abstract: Recently there have been several attempts to extend Nesterov's accelerated algorithm to smooth stochastic and variance-reduced optimization. In this paper, we show that there is a simpler approach to acceleration: applying optimistic online learning algorithms and querying the gradient oracle at the online average of the intermediate optimization iterates. In particular, we tighten a recent result of Cutkosky (2019) to demonstrate theoretically that online iterate averaging results in a reduced optimization gap, independently of the algorithm involved. We show that carefully combining this technique with existing generic optimistic online learning algorithms yields the optimal accelerated rates for optimizing strongly-convex and non-strongly-convex, possibly composite objectives, with deterministic as well as stochastic first-order oracles. We further extend this idea to variance-reduced optimization. Finally, we also provide "universal" algorithms that achieve the optimal rate for smooth and non-smooth composite objectives simultaneously without further tuning, generalizing the results of Kavis et al. (2019) and solving a number of their open problems.

Sets Clustering

- Ibrahim Jubran, Murad Tukan, Alaa Maalouf, Dan Feldman
- abstract: The input to the \emph{sets-k-means} problem is an integer $k\geq 1$ and a set $\mathcal{P}={P_1,\cdots,P_n}$ of fixed sized sets in $\mathbb{R}^d$. The goal is to compute a set C of k centers (points) in $\mathbb{R}^d$ that minimizes the sum $\sum_{P\in \mathcal{P}} \min_{p\in P, c\in C}\left\|{p}-c \right\|^2$ of squared distances to these sets. An \emph{ε-core-set} for this problem is a weighted subset of $\mathcal{P}$ that approximates this sum up to $1\pm\varepsilon$ factor, for \emph{every} set C of k centers in $\mathbb{R}^d$. We prove that such a core-set of $O(\log^2{n})$ sets always exists, and can be computed in $O(n\log{n})$ time, for every input $\mathcal{P}$ and every fixed $d,k\geq 1$ and $\varepsilon \in (0,1)$. The result easily generalized for any metric space, distances to the power of $z>0$, and M-estimators that handle outliers. Applying an inefficient but optimal algorithm on this coreset allows us to obtain the first PTAS ($1+\varepsilon$ approximation) for the sets-k-means problem that takes time near linear in n. This is the first result even for sets-mean on the plane ($k=1$, $d=2$). Open source code and experimental results for document classification and facility locations are also provided.

Distribution Augmentation for Generative Modeling

- Heewoo Jun, Rewon Child, Mark Chen, John Schulman, Aditya Ramesh, Alec Radford, Ilya Sutskever
- abstract: We present distribution augmentation (DistAug), a simple and powerful method of regularizing generative models. Our approach applies augmentation functions to data and, importantly, conditions the generative model on the specific function used. Unlike typical data augmentation, DistAug allows usage of functions which modify the target density, enabling aggressive augmentations more commonly seen in supervised and self-supervised learning. We demonstrate this is a more effective regularizer than standard methods, and use it to train a 152M parameter autoregressive model on CIFAR-10 to 2.56 bits per dim (relative to the state-of-the-art 2.80). Samples from this model attain FID 12.75 and IS 8.40, outperforming the majority of GANs. We further demonstrate the technique is broadly applicable across model architectures and problem domains.

Sub-Goal Trees a Framework for Goal-Based Reinforcement Learning

- Tom Jurgenson, Or Avner, Edward Groshev, Aviv Tamar
- abstract: Many AI problems, in robotics and other domains, are goal-directed, essentially seeking a trajectory leading to some goal state. Reinforcement learning (RL), building on Bellman's optimality equation, naturally optimizes for a single goal, yet can be made goal-directed by augmenting the state with the goal. Instead, we propose a new RL framework, derived from a dynamic programming equation for the all pairs shortest path (APSP) problem, which naturally solves goal-directed queries. We show that this approach has computational benefits for both standard and approximate dynamic programming. Interestingly, our formulation prescribes a novel protocol for computing a trajectory: instead of predicting the next state given its predecessor, as in standard RL, a goal-conditioned trajectory is constructed by first predicting an intermediate state between start and goal, partitioning the trajectory into two. Then, recursively, predicting intermediate points on each sub-segment, until a complete trajectory is obtained. We call this trajectory structure a sub-goal tree. Building on it, we additionally extend the policy gradient methodology to recursively predict sub-goals, resulting in novel goal-based algorithms. Finally, we apply our method to neural motion planning, where we demonstrate significant improvements compared to standard RL on navigating a 7-DoF robot arm between obstacles.

Partial Trace Regression and Low-Rank Kraus Decomposition

- Hachem Kadri, Stephane Ayache, Riikka Huusari, Alain Rakotomamonjy, Ralaivola Liva
- abstract: The trace regression model, a direct extension of the well-studied linear regression model, allows one to map matrices to real-valued outputs. We here introduce an even more general model, namely the partial-trace regression model, a family of linear mappings from matrix-valued inputs to matrix-valued outputs; this model subsumes the trace regression model and thus the linear regression model. Borrowing tools from quantum information theory, where partial trace operators have been extensively studied, we propose a framework for learning partial trace regression models from data by taking advantage of the so-called low-rank Kraus representation of completely positive maps. We show the relevance of our framework with synthetic and real-world experiments conducted for both i) matrix-to-matrix regression and ii) positive semidefinite matrix completion, two tasks which can be formulated as partial trace regression problems.

Strategyproof Mean Estimation from Multiple-Choice Questions

- Anson Kahng, Gregory Kehne, Ariel Procaccia
- abstract: Given n values possessed by n agents, we study the problem of estimating the mean by truthfully eliciting agents' answers to multiple-choice questions about their values. We consider two natural candidates for estimation error: mean squared error (MSE) and mean absolute error (MAE). We design a randomized estimator which is asymptotically optimal for both measures in the worst case. In the case where prior distributions over the agents' values are known, we give an optimal, polynomial-time algorithm for MSE, and show that the task of computing an optimal estimate for MAE is #P-hard. Finally, we demonstrate empirically that knowledge of prior distributions gives a significant edge.

Variational Autoencoders with Riemannian Brownian Motion Priors

- Dimitrios Kalatzis, David Eklund, Georgios Arvanitidis, Soren Hauberg
- abstract: Variational Autoencoders (VAEs) represent the given data in a low-dimensional latent space, which is generally assumed to be Euclidean. This assumption naturally leads to the common choice of a standard Gaussian prior over continuous latent variables. Recent work has, however, shown that this prior has a detrimental effect on model capacity, leading to subpar performance. We propose that the Euclidean assumption lies at the heart of this failure mode. To counter this, we assume a Riemannian structure over the latent space, which constitutes a more principled geometric view of the latent codes, and replace the standard Gaussian prior with a Riemannian Brownian motion prior. We propose an efficient inference scheme that does not rely on the unknown normalizing factor of this prior. Finally, we demonstrate that this prior significantly increases model capacity using only one additional scalar parameter.

DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training

- Nathan Kallus
- abstract: We study optimal covariate balance for causal inferences from observational data when rich covariates and complex relationships necessitate flexible modeling with neural networks. Standard approaches such as propensity weighting and matching/balancing fail in such settings due to miscalibrated propensity nets and inappropriate covariate representations, respectively. We propose a new method based on adversarial training of a weighting and a discriminator network that effectively addresses this methodological gap. This is demonstrated through new theoretical characterizations and empirical results on both synthetic and clinical data showing how causal analyses can be salvaged in such challenging settings.

Double Reinforcement Learning for Efficient and Robust Off-Policy Evaluation

- Nathan Kallus, Masatoshi Uehara
- abstract: Off-policy evaluation (OPE) in reinforcement learning allows one to evaluate novel decision policies without needing to conduct exploration, which is often costly or otherwise infeasible. We consider for the first time the semiparametric efficiency limits of OPE in Markov decision processes (MDPs), where actions, rewards, and states are memoryless. We show existing OPE estimators may fail to be efficient in this setting. We develop a new estimator based on cross-fold estimation of q-functions and marginalized density ratios, which we term double reinforcement learning (DRL). We show that DRL is efficient when both components are estimated at fourth-root rates and is also doubly robust when only one component is consistent. We investigate these properties empirically and demonstrate the performance benefits due to harnessing memorylessness.

Statistically Efficient Off-Policy Policy Gradients

- Nathan Kallus, Masatoshi Uehara
- abstract: Policy gradient methods in reinforcement learning update policy parameters by taking steps in the direction of an estimated gradient of policy value. In this paper, we consider the efficient estimation of policy gradients from off-policy data, where the estimation is particularly non-trivial. We derive the asymptotic lower bound on the feasible mean-squared error in both Markov and non-Markov decision processes and show that existing estimators fail to achieve it in general settings. We propose a meta-algorithm that achieves the lower bound without any parametric assumptions and exhibits a unique 4-way double robustness property. We discuss how to estimate nuisances that the algorithm relies on. Finally, we establish guarantees at the rate at which we approach a stationary point when we take steps in the direction of our new estimated policy gradient.

On the Power of Compressed Sensing with Generative Models

- Akshay Kamath, Eric Price, Sushrut Karmalkar
- abstract: The goal of compressed sensing is to learn a structured signal x from a limited number of noisy linear measurements $y \approx Ax$. In traditional compressed sensing, “structure” is represented by sparsity in some known basis. Inspired by the success of deep learning in modeling images, recent work starting with Bora-Jalal-Price-Dimakis’17 has instead considered structure to come from a generative model $G: \mathbb{R}^k \to \mathbb{R}^n$. We present two results establishing the difficulty and strength of this latter task, showing that existing bounds are tight: First, we provide a lower bound matching the Bora et.al upper bound for compressed sensing with L-Lipschitz generative models G which holds even for the more relaxed goal of \emph{non-uniform} recovery. Second, we show that generative models generalize sparsity as a representation of structure by constructing a ReLU-based neural network with 2 hidden layers and $O(n)$ activations per layer whose range is precisely the set of all k-sparse vectors.

Learning and Evaluating Contextual Embedding of Source Code

- Aditya Kanade, Petros Maniatis, Gogul Balakrishnan, Kensen Shi
- abstract: Recent research has achieved impressive results on understanding and improving source code by building up on machine-learning techniques developed for natural languages. A significant advancement in natural-language understanding has come with the development of pre-trained contextual embeddings, such as BERT, which can be fine-tuned for downstream tasks with less labeled data and training budget, while achieving better accuracies. However, there is no attempt yet to obtain a high-quality contextual embedding of source code, and to evaluate it on multiple program-understanding tasks simultaneously; that is the gap that this paper aims to mitigate. Specifically, first, we curate a massive, deduplicated corpus of 7.4M Python files from GitHub, which we use to pre-train CuBERT, an open-sourced code-understanding BERT model; and, second, we create an open-sourced benchmark that comprises five classification tasks and one program-repair task, akin to code-understanding tasks proposed in the literature before. We fine-tune CuBERT on our benchmark tasks, and compare the resulting models to different variants of Word2Vec token embeddings, BiLSTM and Transformer models, as well as published state-of-the-art models, showing that CuBERT outperforms them all, even with shorter training, and with fewer labeled examples. Future work on source-code embedding can benefit from reusing our benchmark, and from comparing against CuBERT models as a strong baseline.

Operation-Aware Soft Channel Pruning using Differentiable Masks

- Minsoo Kang, Bohyung Han
- abstract: We propose a simple but effective data-driven channel pruning algorithm, which compresses deep neural networks in a differentiable way by exploiting the characteristics of operations. The proposed approach makes a joint consideration of batch normalization (BN) and rectified linear unit (ReLU) for channel pruning; it estimates how likely the two successive operations deactivate each feature map and prunes the channels with high probabilities. To this end, we learn differentiable masks for individual channels and make soft decisions throughout the optimization procedure, which facilitates to explore larger search space and train more stable networks. The proposed framework enables us to identify compressed models via a joint learning of model parameters and channel pruning without an extra procedure of fine-tuning. We perform extensive experiments and achieve outstanding performance in terms of the accuracy of output networks given the same amount of resources when compared with the state-of-the-art methods.

SCAFFOLD: Stochastic Controlled Averaging for Federated Learning

- Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, Ananda Theertha Suresh
- abstract: Federated learning is a key scenario in modern large-scale machine learning where the data remains distributed over a large number of clients and the task is to learn a centralized model without transmitting the client data. The standard optimization algorithm used in this setting is Federated Averaging

(FedAvg) due to its low communication cost. We obtain a tight characterization of the convergence of FedAvg and prove that heterogeneity (non-iid-ness) in the client's data results in a 'drift' in the local updates resulting in poor performance. As a solution, we propose a new algorithm (SCAFFOLD) which uses control variates (variance reduction) to correct for the 'client drift'. We prove that SCAFFOLD requires significantly fewer communication rounds and is not affected by data heterogeneity or client sampling. Further, we show that (for quadratics) SCAFFOLD can take advantage of similarity in the client's data yielding even faster convergence. The latter is the first result to quantify the usefulness of local-steps in distributed optimization.

Non-autoregressive Machine Translation with Disentangled Context Transformer

- Jungo Kasai, James Cross, Marjan Ghazvininejad, Jiatao Gu
- abstract: State-of-the-art neural machine translation models generate a translation from left to right and every step is conditioned on the previously generated tokens. The sequential nature of this generation process causes fundamental latency in inference since we cannot generate multiple tokens in each sentence in parallel. We propose an attention-masking based model, called Disentangled Context (DisCo) transformer, that simultaneously generates all tokens given different contexts. The DisCo transformer is trained to predict every output token given an arbitrary subset of the other reference tokens. We also develop the parallel easy-first inference algorithm, which iteratively refines every token in parallel and reduces the number of required iterations. Our extensive experiments on 7 translation directions with varying data sizes demonstrate that our model achieves competitive, if not better, performance compared to the state of the art in non-autoregressive machine translation while significantly reducing decoding time on average.

Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

- Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, François Fleuret
- abstract: Transformers achieve remarkable performance in several tasks but due to their quadratic complexity, with respect to the input's length, they are prohibitively slow for very long sequences. To address this limitation, we express the self-attention as a linear dot-product of kernel feature maps and make use of the associativity property of matrix products to reduce the complexity from $\bigO{N^2}$ to $\bigO{N}$, where N is the sequence length. We show that this formulation permits an iterative implementation that dramatically accelerates autoregressive transformers and reveals their relationship to recurrent neural networks. Our \emph{Linear Transformers} achieve similar performance to vanilla Transformers and they are up to 4000x faster on autoregressive prediction of very long sequences.

Rate-distortion optimization guided autoencoder for isometric embedding in Euclidean latent space

- Keizo Kato, Jing Zhou, Tomotake Sasaki, Akira Nakagawa
- abstract: To analyze high-dimensional and complex data in the real world, deep generative models, such as variational autoencoder (VAE) embed data in a low-dimensional space (latent space) and learn a probabilistic model in the latent space. However, they struggle to accurately reproduce the probability distribution function (PDF) in the input space from that in the latent space. If the embedding were isometric, this issue can be solved, because the relation of PDFs can become tractable. To achieve isometric property, we propose Rate-Distortion Optimization guided autoencoder inspired by orthonormal transform coding. We show our method has the following properties: (i) the Jacobian matrix between the input space and a Euclidean latent space forms a constantly-scaled orthonormal system and enables isometric data embedding; (ii) the relation of PDFs in both spaces can become tractable one such as proportional relation. Furthermore, our method outperforms state-of-the-art methods in unsupervised anomaly detection with four public datasets.

Efficient Non-conjugate Gaussian Process Factor Models for Spike Count Data using Polynomial Approximations

- Stephen Keeley, David Zoltowski, Yiyi Yu, Spencer Smith, Jonathan Pillow
- abstract: Gaussian Process Factor Analysis (GPFA) has been broadly applied to the problem of identifying smooth, low-dimensional temporal structure underlying large-scale neural recordings. However, spike trains are non-Gaussian, which motivates combining GPFA with discrete observation models for binned spike count data. The drawback to this approach is that GPFA priors are not conjugate to count model likelihoods, which makes inference challenging. Here we address this obstacle by introducing a fast, approximate inference method for non-conjugate GPFA models. Our approach uses orthogonal second-order polynomials to approximate the nonlinear terms in the non-conjugate log-likelihood, resulting in a method we refer to as polynomial approximate log-likelihood (PAL) estimators. This approximation allows for accurate closed-form evaluation of marginal likelihoods and fast numerical optimization for parameters and hyperparameters. We derive PAL estimators for GPFA models with binomial, Poisson, and negative binomial observations and find the PAL estimation is highly accurate, and achieves faster convergence times compared to existing state-of-the-art inference methods. We also find that PAL hyperparameters can provide sensible initialization for black box variational inference (BBVI), which improves BBVI accuracy. We demonstrate that PAL estimators achieve fast and accurate extraction of latent structure from multi-neuron spike train data.

Quantum Expectation-Maximization for Gaussian mixture models

- Iordanis Kerenidis, Alessandro Luongo, Anupam Prakash
- abstract: We define a quantum version of Expectation-Maximization (QEM), a fundamental tool in unsupervised machine learning, often used to solve Maximum Likelihood (ML) and Maximum A Posteriori (MAP) estimation problems. We use QEM to fit a Gaussian Mixture Model, and show how to generalize it to fit mixture models with base distributions in the exponential family. Given quantum access to a dataset, our algorithm has convergence and precision guarantees similar to the classical algorithm, while the runtime is polylogarithmic in the number of elements in the training set and polynomial in other parameters, such as the dimension of the feature space and the number of components in the mixture. We discuss the performance of the algorithm on a dataset that is expected to be classified successfully by classical EM and provide guarantees for its runtime.

Differentiable Likelihoods for Fast Inversion of 'Likelihood-Free' Dynamical Systems

- Hans Kersting, Nicholas Krämer, Martin Schiegg, Christian Daniel, Michael Tiemann, Philipp Hennig
- abstract: Likelihood-free (a.k.a. simulation-based) inference problems are inverse problems with expensive, or intractable, forward models. ODE inverse problems are commonly treated as likelihood-free, as their forward map has to be numerically approximated by an ODE solver. This, however, is not a fundamental constraint but just a lack of functionality in classic ODE solvers, which do not return a likelihood but a point estimate. To address this shortcoming, we employ Gaussian ODE filtering (a probabilistic numerical method for ODEs) to construct a local Gaussian approximation to the likelihood. This approximation yields tractable estimators for the gradient and Hessian of the (log-)likelihood. Insertion of these estimators into existing gradient-based optimization and sampling methods engenders new solvers for ODE inverse problems. We demonstrate that these methods outperform standard likelihood-free approaches on three benchmark-systems.

Feature Noise Induces Loss Discrepancy Across Groups

- Fereshte Khani, Percy Liang
- abstract: The performance of standard learning procedures has been observed to differ widely across groups. Recent studies usually attribute this loss discrepancy to an information deficiency for one group (e.g., one group has less data). In this work, we point to a more subtle source of loss discrepancy—feature noise. Our main result is that even when there is no information deficiency specific to one group (e.g., both groups have infinite data), adding the same amount of feature noise to all individuals leads to loss discrepancy. For linear regression, we thoroughly characterize the effect of feature noise on loss discrepancy in terms of the amount of noise, the difference between moments of the two groups, and whether group information is used or not. We

then show this loss discrepancy does not vanish immediately if a shift in distribution causes the groups to have similar moments. On three real-world datasets, we show feature noise increases the loss discrepancy if groups have different distributions, while it does not affect the loss discrepancy on datasets where groups have similar distributions.

Entropy Minimization In Emergent Languages

- Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, Marco Baroni
- abstract: There is growing interest in studying the languages that emerge when neural agents are jointly trained to solve tasks requiring communication through a discrete channel. We investigate here the information-theoretic complexity of such languages, focusing on the basic two-agent, one-exchange setup. We find that, under common training procedures, the emergent languages are subject to an entropy minimization pressure that has also been detected in human language, whereby the mutual information between the communicating agent's inputs and the messages is minimized, within the range afforded by the need for successful communication. That is, emergent languages are (nearly) as simple as the task they are developed for allow them to be. This pressure is amplified as we increase communication channel discreteness. Further, we observe that stronger discrete-channel-driven entropy minimization leads to representations with increased robustness to overfitting and adversarial attacks. We conclude by discussing the implications of our findings for the study of natural and artificial communication systems.

Private Outsourced Bayesian Optimization

- Dmitrii Kharkovskii, Zhongxiang Dai, Bryan Kian Hsiang Low
- abstract: This paper presents the private-outsourced-Gaussian process-upper confidence bound (PO-GP-UCB) algorithm, which is the first algorithm for privacy-preserving Bayesian optimization (BO) in the outsourced setting with a provable performance guarantee. We consider the outsourced setting where the entity holding the dataset and the entity performing BO are represented by different parties, and the dataset cannot be released non-privately. For example, a hospital holds a dataset of sensitive medical records and outsources the BO task on this dataset to an industrial AI company. The key idea of our approach is to make the BO performance of our algorithm similar to that of non-private GP-UCB run using the original dataset, which is achieved by using a random projection-based transformation that preserves both privacy and the pairwise distances between inputs. Our main theoretical contribution is to show that a regret bound similar to that of the standard GP-UCB algorithm can be established for our PO-GP-UCB algorithm. We empirically evaluate the performance of our PO-GP-UCB algorithm with synthetic and real-world datasets.

What can I do here? A Theory of Affordances in Reinforcement Learning

- Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, David Abel, Doina Precup
- abstract: Reinforcement learning algorithms usually assume that all actions are always available to an agent. However, both people and animals understand the general link between the features of their environment and the actions that are feasible. Gibson (1977) coined the term "affordances" to describe the fact that certain states enable an agent to do certain actions, in the context of embodied agents. In this paper, we develop a theory of affordances for agents who learn and plan in Markov Decision Processes. Affordances play a dual role in this case. On one hand, they allow faster planning, by reducing the number of actions available in any given situation. On the other hand, they facilitate more efficient and precise learning of transition models from data, especially when such models require function approximation. We establish these properties through theoretical results as well as illustrative examples. We also propose an approach to learn affordances and use it to estimate transition models that are simpler and generalize better.

Uniform Convergence of Rank-weighted Learning

- Justin Khim, Liu Leqi, Adarsh Prasad, Pradeep Ravikumar
- abstract: The decision-theoretic foundations of classical machine learning models have largely focused on estimating model parameters that minimize the expectation of a given loss function. However, as machine learning models are deployed in varied contexts, such as in high-stakes decision-making and societal settings, it is clear that these models are not just evaluated by their average performances. In this work, we study a novel notion of L-Risk based on the classical idea of rank-weighted learning. These L-Risks, induced by rank-dependent weighting functions with bounded variation, is a unification of popular risk measures such as conditional value-at-risk and those defined by cumulative prospect theory. We give uniform convergence bounds of this broad class of risk measures and study their consequences on a logistic regression example.

FACT: A Diagnostic for Group Fairness Trade-offs

- Joon Sik Kim, Jiahao Chen, Ameet Talwalkar
- abstract: Group fairness, a class of fairness notions that measure how different groups of individuals are treated differently according to their protected attributes, has been shown to conflict with one another, often with a necessary cost in loss of model's predictive performance. We propose a general diagnostic that enables systematic characterization of these trade-offs in group fairness. We observe that the majority of group fairness notions can be expressed via the fairness-confusion tensor, which is the confusion matrix split according to the protected attribute values. We frame several optimization problems that directly optimize both accuracy and fairness objectives over the elements of this tensor, which yield a general perspective for understanding multiple trade-offs including group fairness incompatibilities. It also suggests an alternate post-processing method for designing fair classifiers. On synthetic and real datasets, we demonstrate the use cases of our diagnostic, particularly on understanding the trade-off landscape between accuracy and fairness.

Puzzle Mix: Exploiting Saliency and Local Statistics for Optimal Mixup

- Jang-Hyun Kim, Wonho Choo, Hyun Oh Song
- abstract: While deep neural networks achieve great performance on fitting the training distribution, the learned networks are prone to overfitting and are susceptible to adversarial attacks. In this regard, a number of mixup based augmentation methods have been recently proposed. However, these approaches mainly focus on creating previously unseen virtual examples and can sometimes provide misleading supervisory signal to the network. To this end, we propose Puzzle Mix, a mixup method for explicitly utilizing the saliency information and the underlying statistics of the natural examples. This leads to an interesting optimization problem alternating between the multi-label objective for optimal mixing mask and saliency discounted optimal transport objective. Our experiments show Puzzle Mix achieves the state of the art generalization and the adversarial robustness results compared to other mixup methods on CIFAR-100, Tiny-ImageNet, and ImageNet datasets, and the source code is available at https://github.com/snu-mllab/PuzzleMix.

Domain Adaptive Imitation Learning

- Kuno Kim, Yihong Gu, Jiaming Song, Shengjia Zhao, Stefano Ermon
- abstract: We study the question of how to imitate tasks across domains with discrepancies such as embodiment, viewpoint, and dynamics mismatch. Many prior works require paired, aligned demonstrations and an additional RL step that requires environment interactions. However, paired, aligned demonstrations are seldom obtainable and RL procedures are expensive. In this work, we formalize the Domain Adaptive Imitation Learning (DAIL) problem - a unified framework for imitation learning in the presence of viewpoint, embodiment, and/or dynamics mismatch. Informally, DAIL is the process of learning how to perform a task optimally, given demonstrations of the task in a distinct domain. We propose a two step approach to DAIL: alignment followed by adaptation. In the alignment step we execute a novel unsupervised MDP alignment algorithm, Generative Adversarial MDP

Alignment (GAMA), to learn state and action correspondences from \emph{unpaired, unaligned} demonstrations. In the adaptation step we leverage the correspondences to zero-shot imitate tasks across domains. To describe when DAIL is feasible via alignment and adaptation, we introduce a theory of MDP alignability. We experimentally evaluate GAMA against baselines in embodiment, viewpoint, and dynamics mismatch scenarios where aligned demonstrations don't exist and show the effectiveness of our approach

Variational Inference for Sequential Data with Future Likelihood Estimates

- Geon-Hyeong Kim, Youngsoo Jang, Hongseok Yang, Kee-Eung Kim
- abstract: The recent development of flexible and scalable variational inference algorithms has popularized the use of deep probabilistic models in a wide range of applications. However, learning and reasoning about high-dimensional models with nondifferentiable densities are still a challenge. For such a model, inference algorithms struggle to estimate the gradients of variational objectives accurately, due to high variance in their estimates. To tackle this challenge, we present a novel variational inference algorithm for sequential data, which performs well even when the density from the model is not differentiable, for instance, due to the use of discrete random variables. The key feature of our algorithm is that it estimates future likelihoods at all time steps. The estimated future likelihoods form the core of our new low-variance gradient estimator. We formally analyze our gradient estimator from the perspective of variational objective, and show the effectiveness of our algorithm with synthetic and real datasets.

Active World Model Learning with Progress Curiosity

- Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, Daniel Yamins
- abstract: World models are self-supervised predictive models of how the world evolves. Humans learn world models by curiously exploring their environment, in the process acquiring compact abstractions of high bandwidth sensory inputs, the ability to plan across long temporal horizons, and an understanding of the behavioral patterns of other agents. In this work, we study how to design such a curiosity-driven Active World Model Learning (AWML) system. To do so, we construct a curious agent building world models while visually exploring a 3D physical environment rich with distillations of representative real-world agents. We propose an AWML system driven by γ-Progress: a scalable and effective learning progress-based curiosity signal and show that γ-Progress naturally gives rise to an exploration policy that directs attention to complex but learnable dynamics in a balanced manner, as a result overcoming the "white noise problem". As a result, our γ-Progress-driven controller achieves significantly higher AWML performance than baseline controllers equipped with state-of-the-art exploration strategies such as Random Network Distillation and Model Disagreement.

Bayesian Experimental Design for Implicit Models by Mutual Information Neural Estimation

- Steven Kleinegesse, Michael U. Gutmann
- abstract: Implicit stochastic models, where the data-generation distribution is intractable but sampling is possible, are ubiquitous in the natural sciences. The models typically have free parameters that need to be inferred from data collected in scientific experiments. A fundamental question is how to design the experiments so that the collected data are most useful. The field of Bayesian experimental design advocates that, ideally, we should choose designs that maximise the mutual information (MI) between the data and the parameters. For implicit models, however, this approach is severely hampered by the high computational cost of computing posteriors and maximising MI, in particular when we have more than a handful of design variables to optimise. In this paper, we propose a new approach to Bayesian experimental design for implicit models that leverages recent advances in neural MI estimation to deal with these issues. We show that training a neural network to maximise a lower bound on MI allows us to jointly determine the optimal design and the posterior. Simulation studies illustrate that this gracefully extends Bayesian experimental design for implicit models to higher design dimensions.

Optimal Continual Learning has Perfect Memory and is NP-hard

- Jeremias Knoblauch, Hisham Husain, Tom Diethe
- abstract: Continual Learning (CL) algorithms incrementally learn a predictor or representation across multiple sequentially observed tasks. Designing CL algorithms that perform reliably and avoid so-called catastrophic forgetting has proven a persistent challenge. The current paper develops a theoretical approach that explains why. In particular, we derive the computational properties which CL algorithms would have to possess in order to avoid catastrophic forgetting. Our main finding is that such optimal CL algorithms generally solve an NP-hard problem and will require perfect memory to do so. The findings are of theoretical interest, but also explain the excellent performance of CL algorithms using experience replay, episodic memory and core sets relative to regularization-based approaches.

Concept Bottleneck Models

- Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, Percy Liang
- abstract: We seek to learn models that we can interact with using high-level concepts: if the model did not think there was a bone spur in the x-ray, would it still predict severe arthritis? State-of-the-art models today do not typically support the manipulation of concepts like "the existence of bone spurs", as they are trained end-to-end to go directly from raw input (e.g., pixels) to output (e.g., arthritis severity). We revisit the classic idea of first predicting concepts that are provided at training time, and then using these concepts to predict the label. By construction, we can intervene on these concept bottleneck models by editing their predicted concept values and propagating these changes to the final prediction. On x-ray grading and bird identification, concept bottleneck models achieve competitive accuracy with standard end-to-end models, while enabling interpretation in terms of high-level clinical concepts ("bone spurs") or bird attributes ("wing color"). These models also allow for richer human-model interaction: accuracy improves significantly if we can correct model mistakes on concepts at test time.

Learning Similarity Metrics for Numerical Simulations

- Georg Kohl, Kiwon Um, Nils Thuerey
- abstract: We propose a neural network-based approach that computes a stable and generalizing metric (LSiM) to compare data from a variety of numerical simulation sources. We focus on scalar time-dependent 2D data that commonly arises from motion and transport-based partial differential equations (PDEs). Our method employs a Siamese network architecture that is motivated by the mathematical properties of a metric. We leverage a controllable data generation setup with PDE solvers to create increasingly different outputs from a reference simulation in a controlled environment. A central component of our learned metric is a specialized loss function that introduces knowledge about the correlation between single data samples into the training process. To demonstrate that the proposed approach outperforms existing metrics for vector spaces and other learned, image-based metrics, we evaluate the different methods on a large range of test data. Additionally, we analyze generalization benefits of an adjustable training data difficulty and demonstrate the robustness of LSiM via an evaluation on three real-world data sets.

Equivariant Flows: Exact Likelihood Generative Learning for Symmetric Densities

- Jonas Köhler, Leon Klein, Frank Noe
- abstract: Normalizing flows are exact-likelihood generative neural networks which approximately transform samples from a simple prior distribution to samples of the probability distribution of interest. Recent work showed that such generative models can be utilized in statistical mechanics to sample equilibrium states of many-body systems in physics and chemistry. To scale and generalize these results, it is essential that the natural symmetries in the

probability density – in physics defined by the invariances of the target potential – are built into the flow. We provide a theoretical sufficient criterion showing that the distribution generated by equivariant normalizing flows is invariant with respect to these symmetries by design. Furthermore, we propose building blocks for flows which preserve symmetries which are usually found in physical/chemical many-body particle systems. Using benchmark systems motivated from molecular physics, we demonstrate that those symmetry preserving flows can provide better generalization capabilities and sampling efficiency.

Online Learning for Active Cache Synchronization

- Andrey Kolobov, Sebastien Bubeck, Julian Zimmert
- abstract: Existing multi-armed bandit (MAB) models make two implicit assumptions: an arm generates a payoff only when it is played, and the agent observes every payoff that is generated. This paper introduces synchronization bandits, a MAB variant where all arms generate costs at all times, but the agent observes an arm's instantaneous cost only when the arm is played. Synchronization MABs are inspired by online caching scenarios such as Web crawling, where an arm corresponds to a cached item and playing the arm means downloading its fresh copy from a server. We present MirrorSync, an online learning algorithm for synchronization bandits, establish an adversarial regret of $O(T^{2/3})$ for it, and show how to make it practical.

A Unified Theory of Decentralized SGD with Changing Topology and Local Updates

- Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, Sebastian Stich
- abstract: Decentralized stochastic optimization methods have gained a lot of attention recently, mainly because of their cheap per iteration cost, data locality, and their communication-efficiency. In this paper we introduce a unified convergence analysis that covers a large variety of decentralized SGD methods which so far have required different intuitions, have different applications, and which have been developed separately in various communities. Our algorithmic framework covers local SGD updates and synchronous and pairwise gossip updates on adaptive network topology. We derive universal convergence rates for smooth (convex and non-convex) problems and the rates interpolate between the heterogeneous (non-identically distributed data) and iid-data settings, recovering linear convergence rates in many special cases, for instance for over-parametrized models. Our proofs rely on weak assumptions (typically improving over prior work in several aspects) and recover (and improve) the best known complexity results for a host of important scenarios, such as for instance coorperative SGD and federated averaging (local SGD).

Meta-learning for Mixed Linear Regression

- Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, Sewoong Oh
- abstract: In modern supervised learning, there are a large number of tasks, but many of them are associated with only a small amount of labelled data. These include data from medical image processing and robotic interaction. Even though each individual task cannot be meaningfully trained in isolation, one seeks to meta-learn across the tasks from past experiences by exploiting some similarities. We study a fundamental question of interest: When can abundant tasks with small data compensate for lack of tasks with big data? We focus on a canonical scenario where each task is drawn from a mixture of k linear regressions, and identify sufficient conditions for such a graceful exchange to hold; there is little loss in sample complexity even when we only have access to small data tasks. To this end, we introduce a novel spectral approach and show that we can efficiently utilize small data tasks with the help of $\tilde\Omega(k^{3/2})$ medium data tasks each with $\tilde\Omega(k^{1/2})$ examples.

SDE-Net: Equipping Deep Neural Networks with Uncertainty Estimates

- Lingkai Kong, Jimeng Sun, Chao Zhang
- abstract: Uncertainty quantification is a fundamental yet unsolved problem for deep learning. The Bayesian framework provides a principled way of uncertainty estimation but is often not scalable to modern deep neural nets (DNNs) that have a large number of parameters. Non-Bayesian methods are simple to implement but often conflate different sources of uncertainties and require huge computing resources. We propose a new method for quantifying uncertainties of DNNs from a dynamical system perspective. The core of our method is to view DNN transformations as state evolution of a stochastic dynamical system and introduce a Brownian motion term for capturing epistemic uncertainty. Based on this perspective, we propose a neural stochastic differential equation model (SDE-Net) which consists of (1) a drift net that controls the system to fit the predictive function; and (2) a diffusion net that captures epistemic uncertainty. We theoretically analyze the existence and uniqueness of the solution to SDE-Net. Our experiments demonstrate that the SDE-Net model can outperform existing uncertainty estimation methods across a series of tasks where uncertainty plays a fundamental role.

On the Sample Complexity of Adversarial Multi-Source PAC Learning

- Nikola Konstantinov, Elias Frantar, Dan Alistarh, Christoph Lampert
- abstract: We study the problem of learning from multiple untrusted data sources, a scenario of increasing practical relevance given the recent emergence of crowdsourcing and collaborative learning paradigms. Specifically, we analyze the situation in which a learning system obtains datasets from multiple sources, some of which might be biased or even adversarially perturbed. It is known that in the single-source case, an adversary with the power to corrupt a fixed fraction of the training data can prevent PAC-learnability, that is, even in the limit of infinitely much training data, no learning system can approach the optimal test error. In this work we show that, surprisingly, the same is not true in the multi-source setting, where the adversary can arbitrarily corrupt a fixed fraction of the data sources. Our main results are a generalization bound that provides finite-sample guarantees for this learning setting, as well as corresponding lower bounds. Besides establishing PAC-learnability our results also show that in a cooperative learning setting sharing data with other parties has provable benefits, even if some participants are malicious.

Asynchronous Coagent Networks

- James Kostas, Chris Nota, Philip Thomas
- abstract: Coagent policy gradient algorithms (CPGAs) are reinforcement learning algorithms for training a class of stochastic neural networks called coagent networks. In this work, we prove that CPGAs converge to locally optimal policies. Additionally, we extend prior theory to encompass asynchronous and recurrent coagent networks. These extensions facilitate the straightforward design and analysis of hierarchical reinforcement learning algorithms like the option-critic, and eliminate the need for complex derivations of customized learning rules for these algorithms.

Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks

- Agustinus Kristiadi, Matthias Hein, Philipp Hennig
- abstract: The point estimates of ReLU classification networks—arguably the most widely used neural network architecture—have been shown to yield arbitrarily high confidence far away from the training data. This architecture, in conjunction with a maximum a posteriori estimation scheme, is thus not calibrated nor robust. Approximate Bayesian inference has been empirically demonstrated to improve predictive uncertainty in neural networks, although the theoretical analysis of such Bayesian approximations is limited. We theoretically analyze approximate Gaussian distributions on the weights of ReLU networks and show that they fix the overconfidence problem. Furthermore, we show that even a simplistic, thus cheap, Bayesian approximation, also fixes these issues. This indicates that a sufficient condition for a calibrated uncertainty on a ReLU network is "to be a bit Bayesian". These theoretical results validate the usage of last-layer Bayesian approximation and motivate a range of a fidelity-cost trade-off. We further validate these findings empirically via various standard experiments using common deep ReLU networks and Laplace approximations.

A Sequential Self Teaching Approach for Improving Generalization in Sound Event Recognition

- Anurag Kumar, Vamsi Ithapu
- abstract: An important problem in machine auditory perception is to recognize and detect sound events. In this paper, we propose a sequential self-teaching approach to learning sounds. Our main proposition is that it is harder to learn sounds in adverse situations such as from weakly labeled and/or noisy labeled data, and in these situations a single stage of learning is not sufficient. Our proposal is a sequential stage-wise learning process that improves generalization capabilities of a given modeling system. We justify this method via technical results and on Audioset, the largest sound events dataset, our sequential learning approach can lead to up to 9% improvement in performance. A comprehensive evaluation also shows that the method leads to improved transferability of knowledge from previously trained models, thereby leading to improved generalization capabilities on transfer learning tasks.

Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness

- Aounon Kumar, Alexander Levine, Tom Goldstein, Soheil Feizi
- abstract: Randomized smoothing, using just a simple isotropic Gaussian distribution, has been shown to produce good robustness guarantees against ℓ_2-norm bounded adversaries. In this work, we show that extending the smoothing technique to defend against other attack models can be challenging, especially in the high-dimensional regime. In particular, for a vast class of i.i.d. smoothing distributions, we prove that the largest ℓ_p-radius that can be certified decreases as $O(1/d^{\frac{1}{2} - \frac{1}{p}})$ with dimension d for $p > 2$. Notably, for $p \geq 2$, this dependence on d is no better than that of the ℓ_p-radius that can be certified using isotropic Gaussian smoothing, essentially putting a matching lower bound on the robustness radius. When restricted to \emph{generalized} Gaussian smoothing, these two bounds can be shown to be within a constant factor of each other in an asymptotic sense, establishing that Gaussian smoothing provides the best possible results, up to a constant factor, when $p \geq 2$. We present experimental results on CIFAR to validate our theory. For other smoothing distributions, such as, a uniform distribution within an ℓ_1 or an ℓ_∞-norm ball, we show upper bounds of the form $O(1 / d)$ and $O(1 / d^{1 - \frac{1}{p}})$ respectively, which have an even worse dependence on d.

Understanding Self-Training for Gradual Domain Adaptation

- Ananya Kumar, Tengyu Ma, Percy Liang
- abstract: Machine learning systems must adapt to data distributions that evolve over time, in applications ranging from sensor networks and self-driving car perception modules to brain-machine interfaces. Traditional domain adaptation is only guaranteed to work when the distribution shift is small; empirical methods combine several heuristics for larger shifts but can be dataset specific. To adapt to larger shifts we consider gradual domain adaptation, where the goal is to adapt an initial classifier trained on a source domain given only unlabeled data that shifts gradually in distribution towards a target domain. We prove the first non-vacuous upper bound on the error of self-training with gradual shifts, under settings where directly adapting to the target domain can result in unbounded error. The theoretical analysis leads to algorithmic insights, highlighting that regularization and label sharpening are essential even when we have infinite data. Leveraging the gradual shift structure leads to higher accuracies on a rotating MNIST dataset, a forest Cover Type dataset, and a realistic Portraits dataset.

On Implicit Regularization in $β$-VAEs

- Abhishek Kumar, Ben Poole
- abstract: While the impact of variational inference (VI) on posterior inference in a fixed generative model is well-characterized, its role in regularizing a learned generative model when used in variational autoencoders (VAEs) is poorly understood. We study the regularizing effects of variational distributions on learning in generative models from two perspectives. First, we analyze the role that the choice of variational family plays in imparting uniqueness to the learned model by restricting the set of optimal generative models. Second, we study the regularization effect of the variational family on the local geometry of the decoding model. This analysis uncovers the regularizer implicit in the β-VAE objective, and leads to an approximation consisting of a deterministic autoencoding objective plus analytic regularizers that depend on the Hessian or Jacobian of the decoding model, unifying VAEs with recent heuristics proposed for training regularized autoencoders. We empirically verify these findings, observing that the proposed deterministic objective exhibits similar behavior to the β-VAE in terms of objective value and sample quality.

Problems with Shapley-value-based explanations as feature importance measures

- I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, Sorelle Friedler
- abstract: Game-theoretic formulations of feature importance have become popular as a way to "explain" machine learning models. These methods define a cooperative game between the features of a model and distribute influence among these input elements using some form of the game's unique Shapley values. Justification for these methods rests on two pillars: their desirable mathematical properties, and their applicability to specific motivations for explanations. We show that mathematical problems arise when Shapley values are used for feature importance and that the solutions to mitigate these necessarily induce further complexity, such as the need for causal reasoning. We also draw on additional literature to argue that Shapley values do not provide explanations which suit human-centric goals of explainability.

Efficient Identification in Linear Structural Causal Models with Auxiliary Cutsets

- Daniel Kumor, Carlos Cinelli, Elias Bareinboim
- abstract: We develop a polynomial-time algorithm for identification of structural coefficients in linear causal models that subsumes previous efficient state-of-the-art methods, unifying several disparate approaches to identification in this setting. Building on these results, we develop a procedure for identifying total causal effects in linear systems.

Two Routes to Scalable Credit Assignment without Weight Symmetry

- Daniel Kunin, Aran Nayebi, Javier Sagastuy-Brena, Surya Ganguli, Jonathan Bloom, Daniel Yamins
- abstract: The neural plausibility of backpropagation has long been disputed, primarily for its use of non-local weight transport — the biologically dubious requirement that one neuron instantaneously measure the synaptic weights of another. Until recently, attempts to create local learning rules that avoid weight transport have typically failed in the large-scale learning scenarios where backpropagation shines, e.g. ImageNet categorization with deep convolutional networks. Here, we investigate a recently proposed local learning rule that yields competitive performance with backpropagation and find that it is highly sensitive to metaparameter choices, requiring laborious tuning that does not transfer across network architecture. Our analysis indicates the underlying mathematical reason for this instability, allowing us to identify a more robust local learning rule that better transfers without metaparameter tuning. Nonetheless, we find a performance and stability gap between this local rule and backpropagation that widens with increasing model depth. We then investigate several non-local learning rules that relax the need for instantaneous weight transport into a more biologically-plausible "weight estimation" process, showing that these rules match state-of-the-art performance on deep networks and operate effectively in the presence of noisy updates. Taken together, our results suggest two routes towards the discovery of neural implementations for credit assignment without weight symmetry: further improvement of local rules so that they perform consistently across architectures and the identification of biological implementations for non-local learning mechanisms.

Online Dense Subgraph Discovery via Blurred-Graph Feedback

- Yuko Kuroki, Atsushi Miyauchi, Junya Honda, Masashi Sugiyama
- abstract: Dense subgraph discovery aims to find a dense component in edge-weighted graphs. This is a fundamental graph-mining task with a variety of applications and thus has received much attention recently. Although most existing methods assume that each individual edge weight is easily obtained, such an assumption is not necessarily valid in practice. In this paper, we introduce a novel learning problem for dense subgraph discovery in which a learner queries edge subsets rather than only single edges and observes a noisy sum of edge weights in a queried subset. For this problem, we first propose a polynomial-time algorithm that obtains a nearly-optimal solution with high probability. Moreover, to deal with large-sized graphs, we design a more scalable algorithm with a theoretical guarantee. Computational experiments using real-world graphs demonstrate the effectiveness of our algorithms.

Inducing and Exploiting Activation Sparsity for Fast Inference on Deep Neural Networks

- Mark Kurtz, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Nir Shavit, Dan Alistarh
- abstract: Optimizing convolutional neural networks for fast inference has recently become an extremely active area of research. One of the go-to solutions in this context is weight pruning, which aims to reduce computational and memory footprint by removing large subsets of the connections in a neural network. Surprisingly, much less attention has been given to exploiting sparsity in the activation maps, which tend to be naturally sparse in many settings thanks to the structure of rectified linear (ReLU) activation functions. In this paper, we present an in-depth analysis of methods for maximizing the sparsity of the activations in a trained neural network, and show that, when coupled with an efficient sparse-input convolution algorithm, we can leverage this sparsity for significant performance gains. To induce highly sparse activation maps without accuracy loss, we introduce a new regularization technique, coupled with a new threshold-based sparsification method based on a parameterized activation function called Forced-Activation-Threshold Rectified Linear Unit (FATReLU). We examine the impact of our methods on popular image classification models, showing that most architectures can adapt to significantly sparser activation maps without any accuracy loss. Our second contribution is showing that these these compression gains can be translated into inference speedups: we provide a new algorithm to enable fast convolution operations over networks with sparse activations, and show that it can enable significant speedups for end-to-end inference on a range of popular models on the large-scale ImageNet image classification task on modern Intel CPUs, with little or no retraining cost.

Soft Threshold Weight Reparameterization for Learnable Sparsity

- Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, Ali Farhadi
- abstract: Sparsity in Deep Neural Networks (DNNs) is studied extensively with the focus of maximizing prediction accuracy given an overall parameter budget. Existing methods rely on uniform or heuristic non-uniform sparsity budgets which have sub-optimal layer-wise parameter allocation resulting in a) lower prediction accuracy or b) higher inference cost (FLOPs). This work proposes Soft Threshold Reparameterization (STR), a novel use of the soft-threshold operator on DNN weights. STR smoothly induces sparsity while learning pruning thresholds thereby obtaining a non-uniform sparsity budget. Our method achieves state-of-the-art accuracy for unstructured sparsity in CNNs (ResNet50 and MobileNetV1 on ImageNet-1K), and, additionally, learns non-uniform budgets that empirically reduce the FLOPs by up to 50%. Notably, STR boosts the accuracy over existing results by up to 10% in the ultra sparse (99%) regime and can also be used to induce low-rank (structured sparsity) in RNNs. In short, STR is a simple mechanism which learns effective sparsity budgets that contrast with popular heuristics. Code, pretrained models and sparsity budgets are at https://github.com/RAIVNLab/STR.

Controlling Overestimation Bias with Truncated Mixture of Continuous Distributional Quantile Critics

- Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, Dmitry Vetrov
- abstract: The overestimation bias is one of the major impediments to accurate off-policy learning. This paper investigates a novel way to alleviate the overestimation bias in a continuous control setting. Our method—Truncated Quantile Critics, TQC,—blends three ideas: distributional representation of a critic, truncation of critics prediction, and ensembling of multiple critics. Distributional representation and truncation allow for arbitrary granular overestimation control, while ensembling provides additional score improvements. TQC outperforms the current state of the art on all environments from the continuous control benchmark suite, demonstrating 25% improvement on the most challenging Humanoid environment.

Principled learning method for Wasserstein distributionally robust optimization with local perturbations

- Yongchan Kwon, Wonyoung Kim, Joong-Ho Won, Myunghee Cho Paik
- abstract: Wasserstein distributionally robust optimization (WDRO) attempts to learn a model that minimizes the local worst-case risk in the vicinity of the empirical data distribution defined by Wasserstein ball. While WDRO has received attention as a promising tool for inference since its introduction, its theoretical understanding has not been fully matured. Gao et al. (2017) proposed a minimizer based on a tractable approximation of the local worst-case risk, but without showing risk consistency. In this paper, we propose a minimizer based on a novel approximation theorem and provide the corresponding risk consistency results. Furthermore, we develop WDRO inference for locally perturbed data that include the Mixup (Zhang et al., 2017) as a special case. We show that our approximation and risk consistency results naturally extend to the cases when data are locally perturbed. Numerical experiments demonstrate robustness of the proposed method using image classification datasets. Our results show that the proposed method achieves significantly higher accuracy than baseline models on noisy datasets.

Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions

- Prashanth L.A., Krishna Jagannathan, Ravi Kolla
- abstract: Conditional Value-at-Risk (CVaR) is a widely used risk metric in applications such as finance. We derive concentration bounds for CVaR estimates, considering separately the cases of sub-Gaussian, light-tailed and heavy-tailed distributions. For the sub-Gaussian and light-tailed cases, we use a classical CVaR estimator based on the empirical distribution constructed from the samples. For heavy-tailed random variables, we assume a mild ‘bounded moment’ condition, and derive a concentration bound for a truncation-based estimator. Our concentration bounds exhibit exponential decay in the sample size, and are tighter than those available in the literature for the above distribution classes. To demonstrate the applicability of our concentration results, we consider the CVaR optimization problem in a multi-armed bandit setting. Specifically, we address the best CVaR-arm identification problem under a fixed budget. Using our CVaR concentration results, we derive an upper-bound on the probability of incorrect arm identification.

Optimal Randomized First-Order Methods for Least-Squares Problems

- Jonathan Lacotte, Mert Pilanci
- abstract: We provide an exact analysis of a class of randomized algorithms for solving overdetermined least-squares problems. We consider first-order methods, where the gradients are pre-conditioned by an approximation of the Hessian, based on a subspace embedding of the data matrix. This class of algorithms encompasses several randomized methods among the fastest solvers for least-squares problems. We focus on two classical embeddings, namely, Gaussian projections and subsampled randomized Hadamard transforms (SRHT). Our key technical innovation is the derivation of the limiting spectral density of SRHT embeddings. Leveraging this novel result, we derive the family of normalized orthogonal polynomials of the SRHT density and we find the optimal pre-conditioned first-order method along with its rate of convergence. Our analysis of Gaussian embeddings proceeds similarly, and leverages classical random matrix theory results. In particular, we show that for a given sketch size, SRHT embeddings exhibits a faster rate of

convergence than Gaussian embeddings. Then, we propose a new algorithm by optimizing the computational complexity over the choice of the sketching dimension. To our knowledge, our resulting algorithm yields the best known complexity for solving least-squares problems with no condition number dependence.

Duality in RKHSs with Infinite Dimensional Outputs: Application to Robust Losses

- Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, Florence D'Alché-Buc
- abstract: Operator-Valued Kernels (OVKs) and associated vector-valued Reproducing Kernel Hilbert Spaces provide an elegant way to extend scalar kernel methods when the output space is a Hilbert space. Although primarily used in finite dimension for problems like multi-task regression, the ability of this framework to deal with infinite dimensional output spaces unlocks many more applications, such as functional regression, structured output prediction, and structured data representation. However, these sophisticated schemes crucially rely on the kernel trick in the output space, so that most of previous works have focused on the square norm loss function, completely neglecting robustness issues that may arise in such surrogate problems. To overcome this limitation, this paper develops a duality approach that allows to solve OVK machines for a wide range of loss functions. The infinite dimensional Lagrange multipliers are handled through a Double Representer Theorem, and algorithms for \epsilon-insensitive losses and the Huber loss are thoroughly detailed. Robustness benefits are emphasized by a theoretical stability analysis, as well as empirical improvements on structured data applications.

Recht-Re Noncommutative Arithmetic-Geometric Mean Conjecture is False

- Zehua Lai, Lek-Heng Lim
- abstract: Stochastic optimization algorithms have become indispensable in modern machine learning. An unresolved foundational question in this area is the difference between with-replacement sampling and without-replacement sampling — does the latter have superior convergence rate compared to the former? A groundbreaking result of Recht and Ré reduces the problem to a noncommutative analogue of the arithmetic-geometric mean inequality where n positive numbers are replaced by n positive definite matrices. If this inequality holds for all n, then without-replacement sampling (also known as random reshuffling) indeed outperforms with-replacement sampling in some important optimization problems. The conjectured Recht–Ré inequality has so far only been established for $n = 2$ and a special case of $n = 3$. We will show that the Recht–Ré conjecture is false for general n. Our approach relies on the noncommutative Positivstellensatz, which allows us to reduce the conjectured inequality to a semidefinite program and the validity of the conjecture to certain bounds for the optimum values, which we show are false as soon as $n = 5$.

Bidirectional Model-based Policy Optimization

- Hang Lai, Jian Shen, Weinan Zhang, Yong Yu
- abstract: Model-based reinforcement learning approaches leverage a forward dynamics model to support planning and decision making, which, however, may fail catastrophically if the model is inaccurate. Although there are several existing methods dedicated to combating the model error, the potential of the single forward model is still limited. In this paper, we propose to additionally construct a backward dynamics model to reduce the reliance on accuracy in forward model predictions. We develop a novel method, called Bidirectional Model-based Policy Optimization (BMPO) to utilize both the forward model and backward model to generate short branched rollouts for policy optimization. Furthermore, we theoretically derive a tighter bound of return discrepancy, which shows the superiority of BMPO against the one using merely the forward model. Extensive experiments demonstrate that BMPO outperforms state-of-the-art model-based methods in terms of sample efficiency and asymptotic performance.

Robust and Stable Black Box Explanations

- Himabindu Lakkaraju, Nino Arsov, Osbert Bastani
- abstract: As machine learning black boxes are increasingly being deployed in real-world applications, there has been a growing interest in developing post hoc explanations that summarize the behaviors of these black boxes. However, existing algorithms for generating such explanations have been shown to lack stability and robustness to distribution shifts. We propose a novel framework for generating robust and stable explanations of black box models based on adversarial training. Our framework optimizes a minimax objective that aims to construct the highest fidelity explanation with respect to the worst-case over a set of adversarial perturbations. We instantiate this algorithm for explanations in the form of linear models and decision sets by devising the required optimization procedures. To the best of our knowledge, this work makes the first attempt at generating post hoc explanations that are robust to a general class of adversarial perturbations that are of practical interest. Experimental evaluation with real-world and synthetic datasets demonstrates that our approach substantially improves robustness of explanations without sacrificing their fidelity on the original data distribution.

CURL: Contrastive Unsupervised Representations for Reinforcement Learning

- Michael Laskin, Aravind Srinivas, Pieter Abbeel
- abstract: We present CURL: Contrastive Unsupervised Representations for Reinforcement Learning. CURL extracts high-level features from raw pixels using contrastive learning and performs off-policy control on top of the extracted features. CURL outperforms prior pixel-based methods, both model-based and model-free, on complex tasks in the DeepMind Control Suite and Atari Games showing 1.9x and 1.2x performance gains at the 100K environment and interaction steps benchmarks respectively. On the DeepMind Control Suite, CURL is the first image-based algorithm to nearly match the sample-efficiency of methods that use state-based features. Our code is open-sourced and available at https://www.github.com/MishaLaskin/curl.

Efficient Proximal Mapping of the 1-path-norm of Shallow Networks

- Fabian Latorre, Paul Rolland, Nadav Hallak, Volkan Cevher
- abstract: We demonstrate two new important properties of the 1-path-norm of shallow neural networks. First, despite its non-smoothness and non-convexity it allows a closed form proximal operator which can be efficiently computed, allowing the use of stochastic proximal-gradient-type methods for regularized empirical risk minimization. Second, when the activation functions is differentiable, it provides an upper bound on the Lipschitz constant of the network. Such bound is tighter than the trivial layer-wise product of Lipschitz constants, motivating its use for training networks robust to adversarial perturbations. In practical experiments we illustrate the advantages of using the proximal mapping and we compare the robustness-accuracy trade-off induced by the 1-path-norm, L1-norm and layer-wise constraints on the Lipschitz constant (Parseval networks).

Learning with Good Feature Representations in Bandits and in RL with a Generative Model

- Tor Lattimore, Csaba Szepesvari, Gellert Weisz
- abstract: The construction in the recent paper by Du et al. [2019] implies that searching for a near-optimal action in a bandit sometimes requires examining essentially all the actions, even if the learner is given linear features in R^d that approximate the rewards with a small uniform error. We use the Kiefer-Wolfowitz theorem to prove a positive result that by checking only a few actions, a learner can always find an action that is suboptimal with an error of at most O($\epsilon$$\sqrt{}$d) where ϵ is the approximation error of the features. Thus, features are useful when the approximation error is small relative to the dimensionality of the features. The idea is applied to stochastic bandits and reinforcement learning with a generative model where the learner has access to d-dimensional linear features that approximate the action-value functions for all policies to an accuracy of ϵ. For linear bandits, we prove a bound on the regret of order d$\sqrt{}$(n log(k)) + ϵn$\sqrt{}$d log(n) with k the number of actions and n the horizon. For

RL we show that approximate policy iteration can learn a policy that is optimal up to an additive error of order $\epsilon$$\sqrt{}$d/(1 – γ)^2 and using about d/(ϵ^2(1 – γ)^4) samples from the generative model. These bounds are independent of the finer details of the features. We also investigate how the structure of the feature set impacts the tradeoff between sample complexity and estimation error.

Inertial Block Proximal Methods for Non-Convex Non-Smooth Optimization

- Hien Le, Nicolas Gillis, Panagiotis Patrinos
- abstract: We propose inertial versions of block coordinate descent methods for solving non-convex non-smooth composite optimization problems. Our methods possess three main advantages compared to current state-of-the-art accelerated first-order methods: (1) they allow using two different extrapolation points to evaluate the gradients and to add the inertial force (we will empirically show that it is more efficient than using a single extrapolation point), (2) they allow to randomly select the block of variables to update, and (3) they do not require a restarting step. We prove the subsequential convergence of the generated sequence under mild assumptions, prove the global convergence under some additional assumptions, and provide convergence rates. We deploy the proposed methods to solve non-negative matrix factorization (NMF) and show that they compete favorably with the state-of-the-art NMF algorithms. Additional experiments on non-negative approximate canonical polyadic decomposition, also known as nonnegative tensor factorization, are also provided.

Self-Attentive Associative Memory

- Hung Le, Truyen Tran, Svetha Venkatesh
- abstract: Heretofore, neural networks with external memory are restricted to single memory with lossy representations of memory interactions. A rich representation of relationships between memory pieces urges a high-order and segregated relational memory. In this paper, we propose to separate the storage of individual experiences (item memory) and their occurring relationships (relational memory). The idea is implemented through a novel Self-attentive Associative Memory (SAM) operator. Found upon outer product, SAM forms a set of associative memories that represent the hypothetical high-order relationships between arbitrary pairs of memory elements, through which a relational memory is constructed from an item memory. The two memories are wired into a single sequential model capable of both memorization and relational reasoning. We achieve competitive results with our proposed two-memory model in a diversity of machine learning tasks, from challenging synthetic problems to practical testbeds such as geometry, graph, reinforcement learning, and question answering.

Causal Effect Identifiability under Partial-Observability

- Sanghack Lee, Elias Bareinboim
- abstract: Causal effect identifiability is concerned with establishing the effect of intervening on a set of variables on another set of variables from observational or interventional distributions under causal assumptions that are usually encoded in the form of a causal graph. Most of the results of this literature implicitly assume that every variable modeled in the graph is measured in the available distributions. In practice, however, the data collections of the different studies considered do not measure the same variables, consistently. In this paper, we study the causal effect identifiability problem when the available distributions encompass different sets of variables, which we refer to as identification under partial-observability. We study a number of properties of the factors that comprise a causal effect under various levels of abstraction, and then characterize the relationship between them with respect to their status relative to the identification of a targeted intervention. We establish a sufficient graphical criterion for determining whether the effects are identifiable from partially-observed distributions. Finally, building on these graphical properties, we develop an algorithm that returns a formula for a causal effect in terms of the available distributions.

Estimating Model Uncertainty of Neural Networks in Sparse Information Form

- Jongseok Lee, Matthias Humt, Jianxiang Feng, Rudolph Triebel
- abstract: We present a sparse representation of model uncertainty for Deep Neural Networks (DNNs) where the parameter posterior is approximated with an inverse formulation of the Multivariate Normal Distribution (MND), also known as the information form. The key insight of our work is that the information matrix, i.e. the inverse of the covariance matrix tends to be sparse in its spectrum. Therefore, dimensionality reduction techniques such as low rank approximations (LRA) can be effectively exploited. To achieve this, we develop a novel sparsification algorithm and derive a cost-effective analytical sampler. As a result, we show that the information form can be scalably applied to represent model uncertainty in DNNs. Our exhaustive theoretical analysis and empirical evaluations on various benchmarks show the competitiveness of our approach over the current methods.

Self-supervised Label Augmentation via Input Transformations

- Hankook Lee, Sung Ju Hwang, Jinwoo Shin
- abstract: Self-supervised learning, which learns by constructing artificial labels given only the input signals, has recently gained considerable attention for learning representations with unlabeled datasets, i.e., learning without any human-annotated supervision. In this paper, we show that such a technique can be used to significantly improve the model accuracy even under fully-labeled datasets. Our scheme trains the model to learn both original and self-supervised tasks, but is different from conventional multi-task learning frameworks that optimize the summation of their corresponding losses. Our main idea is to learn a single unified task with respect to the joint distribution of the original and self-supervised labels, i.e., we augment original labels via self-supervision. This simple, yet effective approach allows to train models easier by relaxing a certain invariant constraint during learning the original and self-supervised tasks simultaneously. It also enables an aggregated inference which combines the predictions from different augmentations to improve the prediction accuracy. Furthermore, we propose a novel knowledge transfer technique, which we refer to as self-distillation, that has the effect of the aggregated inference in a single (faster) inference. We demonstrate the large accuracy improvement and wide applicability of our framework on various fully-supervised settings, e.g., the few-shot and imbalanced classification scenarios.

Batch Reinforcement Learning with Hyperparameter Gradients

- Byungjun Lee, Jongmin Lee, Peter Vrancx, Dongho Kim, Kee-Eung Kim
- abstract: We consider the batch reinforcement learning problem where the agent needs to learn only from a fixed batch of data, without further interaction with the environment. In such a scenario, we want to prevent the optimized policy from deviating too much from the data collection policy since the estimation becomes highly unstable otherwise due to the off-policy nature of the problem. However, imposing this requirement too strongly will result in a policy that merely follows the data collection policy. Unlike prior work where this trade-off is controlled by hand-tuned hyperparameters, we propose a novel batch reinforcement learning approach, batch optimization of policy and hyperparameter (BOPAH), that uses a gradient-based optimization of the hyperparameter using held-out data. We show that BOPAH outperforms other batch reinforcement learning algorithms in tabular and continuous control tasks, by finding a good balance to the trade-off between adhering to the data collection policy and pursuing the possible policy improvement.

Accelerated Message Passing for Entropy-Regularized MAP Inference

- Jonathan Lee, Aldo Pacchiano, Peter Bartlett, Michael Jordan
- abstract: Maximum a posteriori (MAP) inference in discrete-valued Markov random fields is a fundamental problem in machine learning that involves identifying the most likely configuration of random variables given a distribution. Due to the difficulty of this combinatorial problem, linear programming

(LP) relaxations are commonly used to derive specialized message passing algorithms that are often interpreted as coordinate descent on the dual LP. To achieve more desirable computational properties, a number of methods regularize the LP with an entropy term, leading to a class of smooth message passing algorithms with convergence guarantees. In this paper, we present randomized methods for accelerating these algorithms by leveraging techniques that underlie classical accelerated gradient methods. The proposed algorithms incorporate the familiar steps of standard smooth message passing algorithms, which can be viewed as coordinate minimization steps. We show that these accelerated variants achieve faster rates for finding ϵ-optimal points of the unregularized problem, and, when the LP is tight, we prove that the proposed algorithms recover the true MAP solution in fewer iterations than standard message passing algorithms.

Learning Compound Tasks without Task-specific Knowledge via Imitation and Self-supervised Learning

- Sang-Hyun Lee, Seung-Woo Seo
- abstract: Most real-world tasks are compound tasks that consist of multiple simpler sub-tasks. The main challenge of learning compound tasks is that we have no explicit supervision to learn the hierarchical structure of compound tasks. To address this challenge, previous imitation learning methods exploit task-specific knowledge, e.g., labeling demonstrations manually or specifying termination conditions for each sub-task. However, the need for task-specific knowledge makes it difficult to scale imitation learning to real-world tasks. In this paper, we propose an imitation learning method that can learn compound tasks without task-specific knowledge. The key idea behind our method is to leverage a self-supervised learning framework to learn the hierarchical structure of compound tasks. Our work also proposes a task-agnostic regularization technique to prevent unstable switching between sub-tasks, which has been a common degenerate case in previous works. We evaluate our method against several baselines on compound tasks. The results show that our method achieves state-of-the-art performance on compound tasks, outperforming prior imitation learning methods.

Context-aware Dynamics Model for Generalization in Model-Based Reinforcement Learning

- Kimin Lee, Younggyo Seo, Seunghyun Lee, Honglak Lee, Jinwoo Shin
- abstract: Model-based reinforcement learning (RL) enjoys several benefits, such as data-efficiency and planning, by learning a model of the environment's dynamics. However, learning a global model that can generalize across different dynamics remains a challenge. To tackle this problem, we decompose the task of learning a global dynamics model into two stages: (a) learning a context latent vector that captures the local dynamics, then (b) predicting the next state conditioned on it. In order to encode dynamics-specific information into the context latent vector, we introduce a novel loss function that encourages the context latent vector to be useful for predicting both forward and backward dynamics. The proposed method achieves superior generalization ability across various simulated robotics and control tasks, compared to existing RL schemes.

Temporal Phenotyping using Deep Predictive Clustering of Disease Progression

- Changhee Lee, Mihaela Van Der Schaar
- abstract: Due to the wider availability of modern electronic health records, patient care data is often being stored in the form of time-series. Clustering such time-series data is crucial for patient phenotyping, anticipating patients' prognoses by identifying "similar" patients, and designing treatment guidelines that are tailored to homogeneous patient subgroups. In this paper, we develop a deep learning approach for clustering time-series data, where each cluster comprises patients who share similar future outcomes of interest (e.g., adverse events, the onset of comorbidities). To encourage each cluster to have homogeneous future outcomes, the clustering is carried out by learning discrete representations that best describe the future outcome distribution based on novel loss functions. Experiments on two real-world datasets show that our model achieves superior clustering performance over state-of-the-art benchmarks and identifies meaningful clusters that can be translated into actionable information for clinical decision-making.

Tensor denoising and completion based on ordinal observations

- Chanwoo Lee, Miaoyan Wang
- abstract: Higher-order tensors arise frequently in applications such as neuroimaging, recommendation system, and social network analysis. We consider the problem of low-rank tensor estimation from possibly incomplete, ordinal-valued observations. Two related problems are studied, one on tensor denoising and another on tensor completion. We propose a multi-linear cumulative link model, develop a rank-constrained M-estimator, and obtain theoretical accuracy guarantees. Our mean squared error bound enjoys a faster convergence rate than previous results, and we show that the proposed estimator is minimax optimal under the class of low-rank models. Furthermore, the procedure developed serves as an efficient completion method which guarantees consistent recovery of an order-K (d,...,d)-dimensional low-rank tensor using only O(Kd) noisy, quantized observations. We demonstrate the outperformance of our approach over previous methods on the tasks of clustering and collaborative filtering.

Analytic Marching: An Analytic Meshing Solution from Deep Implicit Surface Networks

- Jiabao Lei, Kui Jia
- abstract: This paper studies a problem of learning surface mesh via implicit functions in an emerging field of deep learning surface reconstruction, where implicit functions are popularly implemented as multi-layer perceptrons (MLPs) with rectified linear units (ReLU). To achieve meshing from the learned implicit functions, existing methods adopt the de-facto standard algorithm of marching cubes; while promising, they suffer from loss of precision learned in the MLPs, due to the discretization nature of marching cubes. Motivated by the knowledge that a ReLU based MLP partitions its input space into a number of linear regions, we identify from these regions analytic cells and faces that are associated with zero-level isosurface of the implicit function, and characterize the conditions under which the identified faces are guaranteed to connect and form a closed, piecewise planar surface. We propose a naturally parallelizable algorithm of analytic marching to exactly recover the mesh captured by a learned MLP. Experiments on deep learning mesh reconstruction verify the advantages of our algorithm over existing ones.

SGD Learns One-Layer Networks in WGANs

- Qi Lei, Jason Lee, Alex Dimakis, Constantinos Daskalakis
- abstract: Generative adversarial networks (GANs) are a widely used framework for learning generative models. Wasserstein GANs (WGANs), one of the most successful variants of GANs, require solving a minmax optimization problem to global optimality, but are in practice successfully trained using stochastic gradient descent-ascent. In this paper, we show that, when the generator is a one-layer network, stochastic gradient descent-ascent converges to a global solution with polynomial time and sample complexity.

Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent

- Yunwen Lei, Yiming Ying
- abstract: Recently there are a considerable amount of work devoted to the study of the algorithmic stability and generalization for stochastic gradient descent (SGD). However, the existing stability analysis requires to impose restrictive assumptions on the boundedness of gradients, smoothness and convexity of loss functions. In this paper, we provide a fine-grained analysis of stability and generalization for SGD by substantially relaxing these assumptions. Firstly, we establish stability and generalization for SGD by removing the existing bounded gradient assumptions. The key idea is the introduction of a new stability measure called on-average model stability, for which we develop novel bounds controlled by the risks of SGD iterates. This yields generalization bounds depending on the behavior of the best model, and leads to the first-ever-known fast bounds in the low-noise setting using

stability approach. Secondly, the smoothness assumption is relaxed by considering loss functions with Holder continuous (sub)gradients for which we show that optimal bounds are still achieved by balancing computation and stability. To our best knowledge, this gives the first-ever-known stability and generalization bounds for SGD with non-smooth loss functions (e.g., hinge loss). Finally, we study learning problems with (strongly) convex objectives but non-convex loss functions.

Learning Quadratic Games on Networks

- Yan Leng, Xiaowen Dong, Junfeng Wu, Alex Pentland
- abstract: Individuals, or organizations, cooperate with or compete against one another in a wide range of practical situations. Such strategic interactions are often modeled as games played on networks, where an individual's payoff depends not only on her action but also on that of her neighbors. The current literature has largely focused on analyzing the characteristics of network games in the scenario where the structure of the network, which is represented by a graph, is known beforehand. It is often the case, however, that the actions of the players are readily observable while the underlying interaction network remains hidden. In this paper, we propose two novel frameworks for learning, from the observations on individual actions, network games with linear-quadratic payoffs, and in particular, the structure of the interaction network. Our frameworks are based on the Nash equilibrium of such games and involve solving a joint optimization problem for the graph structure and the individual marginal benefits. Both synthetic and real-world experiments demonstrate the effectiveness of the proposed frameworks, which have theoretical as well as practical implications for understanding strategic interactions in a network environment.

ACFlow: Flow Models for Arbitrary Conditional Likelihoods

- Yang Li, Shoaib Akbar, Junier Oliva
- abstract: Understanding the dependencies among features of a dataset is at the core of most unsupervised learning tasks. However, a majority of generative modeling approaches are focused solely on the joint distribution $p(x)$ and utilize models where it is intractable to obtain the conditional distribution of some arbitrary subset of features x_u given the rest of the observed covariates x_o: $p(x_u \mid x_o)$. Traditional conditional approaches provide a model for a \emph{fixed} set of covariates conditioned on another \emph{fixed} set of observed covariates. Instead, in this work we develop a model that is capable of yielding \emph{all} conditional distributions $p(x_u \mid x_o)$ (for arbitrary x_u) via tractable conditional likelihoods. We propose a novel extension of (change of variables based) flow generative models, arbitrary conditioning flow models (ACFlow). ACFlow can be conditioned on arbitrary subsets of observed covariates, which was previously infeasible. We further extend ACFlow to model the joint distributions $p(x)$ and arbitrary marginal distributions $p(x_u)$. We also apply ACFlow to the imputation of features, and develop a unified platform for both multiple and single imputation by introducing an auxiliary objective that provides a principled single "best guess" for flow models. Extensive empirical evaluations show that our model achieves state-of-the-art performance in modeling arbitrary conditional likelihoods in addition to both single and multiple imputation in synthetic and real-world datasets.

Manifold Identification for Ultimately Communication-Efficient Distributed Optimization

- Yu-Sheng Li, Wei-Lin Chiang, Ching-Pei Lee
- abstract: This work proposes a progressive manifold identification approach for distributed optimization with sound theoretical justifications to greatly reduce both the rounds of communication and the bytes communicated per round for partly-smooth regularized problems such as the ℓ_1- and group-LASSO-regularized ones. Our two-stage method first uses an inexact proximal quasi-Newton method to iteratively identify a sequence of low-dimensional manifolds in which the final solution would lie, and restricts the model update within the current manifold to gradually lower the order of the per-round communication cost from the problem dimension to the dimension of the manifold that contains a solution and makes the problem within it smooth. After identifying this manifold, we take superlinear-convergent truncated semismooth Newton steps computed by preconditioned conjugate gradient to largely reduce the communication rounds by improving the convergence rate from the existing linear or sublinear ones to a superlinear rate. Experiments show that our method can be orders of magnitudes lower in the communication cost and an order of magnitude faster in the running time than the state of the art.

Neural Architecture Search in A Proxy Validation Loss Landscape

- Yanxi Li, Minjing Dong, Yunhe Wang, Chang Xu
- abstract: This paper searches for the optimal neural architecture by minimizing a proxy of validation loss. Existing neural architecture search (NAS) methods used to discover the optimal neural architecture that best fits the validation examples given the up-to-date network weights. However, back propagation with a number of validation examples could be time consuming, especially when it needs to be repeated many times in NAS. Though these intermediate validation results are invaluable, they would be wasted if we cannot use them to predict the future from the past. In this paper, we propose to approximate the validation loss landscape by learning a mapping from neural architectures to their corresponding validate losses. The optimal neural architecture thus can be easily identified as the minimum of this proxy validation loss landscape. A novel sampling strategy is further developed for an efficient approximation of the loss landscape. Theoretical analysis indicates that the validation loss estimator learnt with our sampling strategy can reach a lower error rate and a lower label complexity compared with a uniform sampling. Experimental results on benchmarks demonstrate that the architecture searched by the proposed algorithm can achieve a satisfactory accuracy with less time cost.

PENNI: Pruned Kernel Sharing for Efficient CNN Inference

- Shiyu Li, Edward Hanson, Hai Li, Yiran Chen
- abstract: Although state-of-the-art (SOTA) CNNs achieve outstanding performance on various tasks, their high computation demand and massive number of parameters make it difficult to deploy these SOTA CNNs onto resource-constrained devices. Previous works on CNN acceleration utilize low-rank approximation of the original convolution layers to reduce computation cost. However, these methods are very difficult to conduct upon sparse models, which limits execution speedup since redundancies within the CNN model are not fully exploited. We argue that kernel granularity decomposition can be conducted with low-rank assumption while exploiting the redundancy within the remaining compact coefficients. Based on this observation, we propose PENNI, a CNN model compression framework that is able to achieve model compactness and hardware efficiency simultaneously by (1) implementing kernel sharing in convolution layers via a small number of basis kernels and (2) alternately adjusting bases and coefficients with sparse constraints. Experiments show that we can prune 97% parameters and 92% FLOPs on ResNet18 CIFAR10 with no accuracy loss, and achieve a 44% reduction in run-time memory consumption and a 53% reduction in inference latency.

Implicit Euler Skip Connections: Enhancing Adversarial Robustness via Numerical Stability

- Mingjie Li, Lingshen He, Zhouchen Lin
- abstract: Deep neural networks have achieved great success in various areas, but recent works have found that neural networks are vulnerable to adversarial attacks, which leads to a hot topic nowadays. Although many approaches have been proposed to enhance the robustness of neural networks, few of them explored robust architectures for neural networks. On this account, we try to address such an issue from the perspective of dynamic system in this work. By viewing ResNet as an explicit Euler discretization of an ordinary differential equation (ODE), for the first time, we find that the adversarial robustness of ResNet is connected to the numerical stability of the corresponding dynamic system, i.e., more stable numerical schemes may correspond to more robust deep networks. Furthermore, inspired by the implicit Euler method for solving numerical ODE problems, we propose Implicit Euler skip connections (IE-Skips) by modifying the original skip connection in ResNet or its variants. Then we theoretically prove its advantages under the

adversarial attack and the experimental results show that our ResNet with IE-Skips can largely improve the robustness and the generalization ability under adversarial attacks when compared with the vanilla ResNet of the same parameter size.

Closed Loop Neural-Symbolic Learning via Integrating Neural Perception, Grammar Parsing, and Symbolic Reasoning

- Qing Li, Siyuan Huang, Yining Hong, Yixin Chen, Ying Nian Wu, Song-Chun Zhu
- abstract: The goal of neural-symbolic computation is to integrate the connectionist and symbolist paradigms. Prior methods learn the neural-symbolic models using reinforcement learning (RL) approaches, which ignore the error propagation in the symbolic reasoning module and thus converge slowly with sparse rewards. In this paper, we address these issues and close the loop of neural-symbolic learning by (1) introducing the grammar model as a symbolic prior to bridge neural perception and symbolic reasoning, and (2) proposing a novel back-search algorithm which mimics the top-down human-like learning procedure to propagate the error through the symbolic reasoning module efficiently. We further interpret the proposed learning framework as maximum likelihood estimation using Markov chain Monte Carlo sampling and the back-search algorithm as a Metropolis-Hastings sampler. The experiments are conducted on two weakly-supervised neural-symbolic tasks: (1) handwritten formula recognition on the newly introduced HWF dataset; (2) visual question answering on the CLEVR dataset. The results show that our approach significantly outperforms the RL methods in terms of performance, converging speed, and data efficiency. Our code and data are released at https://liqing-ustc.github.io/NGS.

Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization

- Zhize Li, Dmitry Kovalev, Xun Qian, Peter Richtarik
- abstract: Due to the high communication cost in distributed and federated learning problems, methods relying on compression of communicated messages are becoming increasingly popular. While in other contexts the best performing gradient-type methods invariably rely on some form of acceleration/momentum to reduce the number of iterations, there are no methods which combine the benefits of both gradient compression and acceleration. In this paper, we remedy this situation and propose the first \emph{accelerated compressed gradient descent (ACGD)} methods. In the single machine regime, we prove that ACGD enjoys the rate $O\Big((1+\omega)\sqrt{\frac{L}{\mu}}\log \frac{1}{\epsilon}\Big)$ for μ-strongly convex problems and $O\Big((1+\omega)\sqrt{\frac{L}{\epsilon}}\Big)$ for convex problems, respectively, where ω is the compression parameter. Our results improve upon the existing non-accelerated rates $O\Big((1+\omega)\frac{L}{\mu}\log \frac{1}{\epsilon}\Big)$ and $O\Big((1+\omega)\frac{L}{\epsilon}\Big)$, respectively, and recover the optimal rates of accelerated gradient descent as a special case when no compression ($\omega=0$) is applied. We further propose a distributed variant of ACGD (called ADIANA) and prove the convergence rate $\widetilde{O}\Big(\omega+\sqrt{\frac{L}{\mu}}+\sqrt{\big(\frac{\omega}{n}+\sqrt{\frac{\omega}{n}}\big)\frac{\omega L}{\mu}}\Big)$, where n is the number of devices/workers and $\widetilde{O}$ hides the logarithmic factor $\log \frac{1}{\epsilon}$. This improves upon the previous best result $\widetilde{O}\Big(\omega + \frac{L}{\mu}+\frac{\omega L}{n\mu} \Big)$ achieved by the DIANA method of Mishchenko et al. (2019). Finally, we conduct several experiments on real-world datasets which corroborate our theoretical results and confirm the practical superiority of our accelerated methods.

On the Relation between Quality-Diversity Evaluation and Distribution-Fitting Goal in Text Generation

- Jianing Li, Yanyan Lan, Jiafeng Guo, Xueqi Cheng
- abstract: The goal of text generation models is to fit the underlying real probability distribution of text. For performance evaluation, quality and diversity metrics are usually applied. However, it is still not clear to what extend can the quality-diversity evaluation reflect the distribution-fitting goal. In this paper, we try to reveal such relation in a theoretical approach. We prove that under certain conditions, a linear combination of quality and diversity constitutes a divergence metric between the generated distribution and the real distribution. We also show that the commonly used BLEU/Self-BLEU metric pair fails to match any divergence metric, thus propose CR/NRR as a substitute for quality/diversity metric pair.

Latent Space Factorisation and Manipulation via Matrix Subspace Projection

- Xiao Li, Chenghua Lin, Ruizhe Li, Chaozheng Wang, Frank Guerin
- abstract: We tackle the problem disentangling the latent space of an autoencoder in order to separate labelled attribute information from other characteristic information. This then allows us to change selected attributes while preserving other information. Our method, matrix subspace projection, is much simpler than previous approaches to latent space factorisation, for example not requiring multiple discriminators or a careful weighting among their loss functions. Furthermore our new model can be applied to autoencoders as a plugin, and works across diverse domains such as images or text. We demonstrate the utility of our method for attribute manipulation in autoencoders trained across varied domains, using both human evaluation and automated methods. The quality of generation of our new model (e.g. reconstruction, conditional generation) is highly competitive to a number of strong baselines.

Visual Grounding of Learned Physical Models

- Yunzhu Li, Toru Lin, Kexin Yi, Daniel Bear, Daniel Yamins, Jiajun Wu, Joshua Tenenbaum, Antonio Torralba
- abstract: Humans intuitively recognize objects' physical properties and predict their motion, even when the objects are engaged in complicated interactions. The abilities to perform physical reasoning and to adapt to new environments, while intrinsic to humans, remain challenging to state-of-the-art computational models. In this work, we present a neural model that simultaneously reasons about physics and makes future predictions based on visual and dynamics priors. The visual prior predicts a particle-based representation of the system from visual observations. An inference module operates on those particles, predicting and refining estimates of particle locations, object states, and physical parameters, subject to the constraints imposed by the dynamics prior, which we refer to as visual grounding. We demonstrate the effectiveness of our method in environments involving rigid objects, deformable materials, and fluids. Experiments show that our model can infer the physical properties within a few observations, which allows the model to quickly adapt to unseen scenarios and make accurate predictions into the future.

Learning from Irregularly-Sampled Time Series: A Missing Data Perspective

- Steven Cheng-Xian Li, Benjamin Marlin
- abstract: Irregularly-sampled time series occur in many domains including healthcare. They can be challenging to model because they do not naturally yield a fixed-dimensional representation as required by many standard machine learning models. In this paper, we consider irregular sampling from the perspective of missing data. We model observed irregularly-sampled time series data as a sequence of index-value pairs sampled from a continuous but unobserved function. We introduce an encoder-decoder framework for learning from such generic indexed sequences. We propose learning methods for this framework based on variational autoencoders and generative adversarial networks. For continuous irregularly-sampled time series, we introduce continuous convolutional layers that can efficiently interface with existing neural network architectures. Experiments show that our models are able to achieve competitive or better classification results on irregularly-sampled multivariate time series compared to recent RNN models while offering significantly faster training times.

Evolutionary Topology Search for Tensor Network Decomposition

- Chao Li, Zhun Sun

- abstract: Tensor network (TN) decomposition is a promising framework to represent extremely high-dimensional problems with few parameters. However, it is challenging to search the (near-)optimal topological structures for TN decomposition, since the number of candidate solutions exponentially grows with increasing the order of a tensor. In this paper, we claim that the issue can be practically tackled by evolutionary algorithms in an affordable manner. We encode the complex topological structures into binary strings, and develop a simple genetic meta-algorithm to search the optimal topology on Hamming space. The experimental results by both synthetic and real-world data demonstrate that our method can effectively discover the ground-truth topology or even better structures with a small number of generations, and significantly boost the representational power of TN decomposition compared with well-known tensor-train (TT) or tensor-ring (TR) models.

Train Big, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers

- Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, Joey Gonzalez
- abstract: Since hardware resources are limited, the objective of training deep learning models is typically to maximize accuracy subject to the time and memory constraints of training and inference. We study the impact of model size in this setting, focusing on Transformer models for NLP tasks that are limited by compute: self-supervised pretraining and high-resource machine translation. We first show that even though smaller Transformer models execute faster per iteration, wider and deeper models converge in significantly fewer steps. Moreover, this acceleration in convergence typically outpaces the additional computational overhead of using larger models. Therefore, the most compute-efficient training strategy is to counterintuitively train extremely large models but stop after a small number of iterations. This leads to an apparent trade-off between the training efficiency of large Transformer models and the inference efficiency of small Transformer models. However, we show that large models are more robust to compression techniques such as quantization and pruning than small models. Consequently, one can get the best of both worlds: heavily compressed, large models achieve higher accuracy than lightly compressed, small models.

Almost Tune-Free Variance Reduction

- Bingcong Li, Lingda Wang, Georgios B. Giannakis
- abstract: The variance reduction class of algorithms including the representative ones, SVRG and SARAH, have well documented merits for empirical risk minimization problems. However, they require grid search to tune parameters (step size and the number of iterations per inner loop) for optimal performance. This work introduces ‘almost tune-free’ SVRG and SARAH schemes equipped with i) Barzilai-Borwein (BB) step sizes; ii) averaging; and, iii) the inner loop length adjusted to the BB step sizes. In particular, SVRG, SARAH, and their BB variants are first reexamined through an ‘estimate sequence’ lens to enable new averaging methods that tighten their convergence rates theoretically, and improve their performance empirically when the step size or the inner loop length is chosen large. Then a simple yet effective means to adjust the number of iterations per inner loop is developed to enhance the merits of the proposed averaging schemes and BB step sizes. Numerical tests corroborate the proposed methods.

Nearly Linear Row Sampling Algorithm for Quantile Regression

- Yi Li, Ruosong Wang, Lin Yang, Hanrui Zhang
- abstract: We give a row sampling algorithm for the quantile loss function with sample complexity nearly linear in the dimensionality of the data, improving upon the previous best algorithm whose sampling complexity has at least cubic dependence on the dimensionality. Based upon our row sampling algorithm, we give the fastest known algorithm for quantile regression and a graph sparsification algorithm for balanced directed graphs. Our main technical contribution is to show that Lewis weights sampling, which has been used in row sampling algorithms for ℓ_p norms, can also be applied in row sampling algorithms for a variety of loss functions. We complement our theoretical results by experiments to demonstrate the practicality of our approach.

Temporal Logic Point Processes

- Shuang Li, Lu Wang, Ruizhi Zhang, Xiaofu Chang, Xuqin Liu, Yao Xie, Yuan Qi, Le Song
- abstract: We propose a modeling framework for event data and aim to answer questions such as \emph{when} and \emph{why} the next event would happen. Our proposed model excels in small data regime with the ability to incorporate domain knowledge in terms of logic rules. We model the dynamics of the event starts and ends via intensity function with the structures informed by a set of first-order temporal logic rules. Using the softened representation of temporal relations, and a weighted combination of logic rules, our probabilistic model can deal with uncertainty in events. Furthermore, many well-known point processes (e.g., Hawkes process, self-correcting point process) can be interpreted as special cases of our model given simple temporal logic rules. Our model, therefore, riches the family of point processes. We derive a maximum likelihood estimation procedure for our model and show that it can lead to accurate predictions when data are sparse and domain knowledge is critical.

Input-Sparsity Low Rank Approximation in Schatten Norm

- Yi Li, David Woodruff
- abstract: We give the first input-sparsity time algorithms for the rank-k low rank approximation problem in every Schatten norm. Specifically, for a given $n\times n$ matrix A, our algorithm computes $Y,Z\in \R^{n\times k}$, which, with high probability, satisfy $\|A-YZ^T\|_p \leq (1+\eps)\|A-A_k\|_p$, *where* $\|M\|_p = \left (\sum{i=1}^n \sigma_i(M)^p \right)^{1/p}$ is the Schatten p-norm of a matrix M with singular values $\sigma_1(M), \ldots, \sigma_n(M)$, and where A_k is the best rank-k approximation to A. Our algorithm runs in time $\tilde{O}(\nnz(A) + n^{\alpha_p}\poly(k/\eps))$, where $\alpha_p = 1$ for $p\in [1,2)$ and $\alpha_p = 1 + (\omega-1)(1-2/p)$ for $p>2$ and $\omega \approx 2.374$ is the exponent of matrix multiplication. For the important case of $p = 1$, which corresponds to the more “robust” nuclear norm, we obtain $\tilde{O}(\nnz(A) + n \cdot \poly(k/\epsilon))$ time, which was previously only known for the Frobenius norm $(p = 2)$. Moreover, since $\alpha_p < \omega$ for every p, our algorithm has a better dependence on n than that in the singular value decomposition for every p. Crucial to our analysis is the use of dimensionality reduction for Ky-Fan p-norms.

RIFLE: Backpropagation in Depth for Deep Transfer Learning through Re-Initializing the Fully-connected LayEr

- Xingjian Li, Haoyi Xiong, Haozhe An, Cheng-Zhong Xu, Dejing Dou
- abstract: Fine-tuning the deep convolution neural network (CNN) using a pre-trained model helps transfer knowledge learned from larger datasets to the target task. While the accuracy could be largely improved even when the training dataset is small, the transfer learning outcome is similar with the pre-trained one with closed CNN weights[17], as the backpropagation here brings less updates to deeper CNN layers. In this work, we propose RIFLE - a simple yet effective strategy that deepens backpropagation in transfer learning settings, through periodically ReInitializing the Fully-connected LayEr with random scratch during the fine-tuning procedure. RIFLE brings significant perturbation to the backpropagation process and leads to deep CNN weights update, while the affects of perturbation can be easily converged throughout the overall learning procedure. The experiments show that the use of RIFLE significantly improves deep transfer learning accuracy on a wide range of datasets, outperforming known tricks for the similar purpose, such as dropout, dropconnect, stochastic depth, and cyclic learning rate, under the same settings with 0.5%-2% higher testing accuracy. Empirical cases and ablation studies further indicate RIFLE brings meaningful updates to deep CNN layers with accuracy improved.

On a projective ensemble approach to two sample test for equality of distributions

- Zhimei Li, Yaowu Zhang
- abstract: In this work, we propose a robust test for the multivariate two-sample problem through projective ensemble, which is a generalization of the Cramer-von Mises statistic. The proposed test statistic has a simple closed-form expression without any tuning parameters involved, it is easy to implement can be computed in quadratic time. Moreover, our test is insensitive to the dimension and consistent against all fixed alternatives, it does not require the moment assumption and is robust to the presence of outliers. We study the asymptotic behaviors of the test statistic under the null and two kinds of alternative hypotheses. We also suggest a permutation procedure to approximate critical values and employ its consistency. We demonstrate the effectiveness of our test through extensive simulation studies and a real data application.

Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation

- Jian Liang, Dapeng Hu, Jiashi Feng
- abstract: Unsupervised domain adaptation (UDA) aims to leverage the knowledge learned from a labeled source dataset to solve similar tasks in a new unlabeled domain. Prior UDA methods typically require to access the source data when learning to adapt the model, making them risky and inefficient for decentralized private data. This work tackles a practical setting where only a trained source model is available and investigates how we can effectively utilize such a model without source data to solve UDA problems. We propose a simple yet generic representation learning framework, named \emph{Source HypOthesis Transfer} (SHOT). SHOT freezes the classifier module (hypothesis) of the source model and learns the target-specific feature extraction module by exploiting both information maximization and self-supervised pseudo-labeling to implicitly align representations from the target domains to the source hypothesis. To verify its versatility, we evaluate SHOT in a variety of adaptation cases including closed-set, partial-set, and open-set domain adaptation. Experiments indicate that SHOT yields state-of-the-art results among multiple domain adaptation benchmarks.

Variable Skipping for Autoregressive Range Density Estimation

- Eric Liang, Zongheng Yang, Ion Stoica, Pieter Abbeel, Yan Duan, Peter Chen
- abstract: Deep autoregressive models compute point likelihood estimates of individual data points. However, many applications (i.e., database cardinality estimation), require estimating range densities, a capability that is under-explored by current neural density estimation literature. In these applications, fast and accurate range density estimates over high-dimensional data directly impact user-perceived performance. In this paper, we explore a technique for accelerating range density estimation over deep autoregressive models. This technique, called variable skipping, exploits the sparse structure of range density queries to avoid sampling unnecessary variables during approximate inference. We show that variable skipping provides 10-100x efficiency improvements when targeting challenging high-quantile error metrics, enables complex applications such as text pattern matching, and can be realized via a simple data augmentation procedure without changing the usual maximum likelihood objective.

Adaptive Droplet Routing in Digital Microfluidic Biochips Using Deep Reinforcement Learning

- Tung-Che Liang, Zhanwei Zhong, Yaas Bigdeli, Tsung-Yi Ho, Krishnendu Chakrabarty, Richard Fair
- abstract: We present and investigate a novel application domain for deep reinforcement learning (RL): droplet routing on digital microfluidic biochips (DMFBs). A DMFB, composed of a two-dimensional electrode array, manipulates discrete fluid droplets to automatically execute biochemical protocols such as point-of-care clinical diagnosis. However, a major concern associated with the use of DMFBs is that electrodes in a biochip can degrade over time. Droplet-transportation operations associated with the degraded electrodes can fail, thereby compromising the integrity of the bioassay outcome. We show that casting droplet transportation as an RL problem enables the training of deep network policies to capture the underlying health conditions of electrodes and to provide reliable fluidic operations. We propose a new RL-based droplet-routing flow that can be used for various sizes of DMFBs, and demonstrate reliable execution of an epigenetic bioassay with the RL droplet router on a fabricated DMFB. To facilitate further research, we also present a simulation environment based on the OpenAI Gym Interface for RL-guided droplet-routing problems on DMFBs.

AR-DAE: Towards Unbiased Neural Entropy Gradient Estimation

- Jae Hyun Lim, Aaron Courville, Christopher Pal, Chin-Wei Huang
- abstract: Entropy is ubiquitous in machine learning, but it is in general intractable to compute the entropy of the distribution of an arbitrary continuous random variable. In this paper, we propose the amortized residual denoising autoencoder (AR-DAE) to approximate the gradient of the log density function, which can be used to estimate the gradient of entropy. Amortization allows us to significantly reduce the error of the gradient approximator by approaching asymptotic optimality of a regular DAE, in which case the estimation is in theory unbiased. We conduct theoretical and experimental analyses on the approximation error of the proposed method, as well as extensive studies on heuristics to ensure its robustness. Finally, using the proposed gradient approximator to estimate the gradient of entropy, we demonstrate state-of-the-art performance on density estimation with variational autoencoders and continuous control with soft actor-critic.

Hierarchical Verification for Adversarial Robustness

- Cong Han Lim, Raquel Urtasun, Ersin Yumer
- abstract: We introduce a new framework for the exact point-wise ℓp robustness verification problem that exploits the layer-wise geometric structure of deep feed-forward networks with rectified linear activations (ReLU networks). The activation regions of the network partition the input space, and one can verify the ℓp robustness around a point by checking all the activation regions within the desired radius. The GeoCert algorithm (Jordan et al., NeurIPS 2019) treats this partition as a generic polyhedral complex in order to detect which region to check next. In contrast, our LayerCert framework considers the nested hyperplane arrangement structure induced by the layers of the ReLU network and explores regions in a hierarchical manner. We show that, under certain conditions on the algorithm parameters, LayerCert provably reduces the number and size of the convex programs that one needs to solve compared to GeoCert. Furthermore, our LayerCert framework allows the incorporation of lower bounding routines based on convex relaxations to further improve performance. Experimental results demonstrate that LayerCert can significantly reduce both the number of convex programs solved and the running time over the state-of-the-art.

On Gradient Descent Ascent for Nonconvex-Concave Minimax Problems

- Tianyi Lin, Chi Jin, Michael Jordan
- abstract: We consider nonconvex-concave minimax problems, $\min_{\mathbf{x}} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$, where f is nonconvex in $\mathbf{x}$ but concave in $\mathbf{y}$ and $\mathcal{Y}$ is a convex and bounded set. One of the most popular algorithms for solving this problem is the celebrated gradient descent ascent (GDA) algorithm, which has been widely used in machine learning, control theory and economics. Despite the extensive convergence results for the convex-concave setting, GDA with equal stepsize can converge to limit cycles or even diverge in a general setting. In this paper, we present the complexity results on two-time-scale GDA for solving nonconvex-concave minimax problems, showing that the algorithm can find a stationary point of the function $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ efficiently. To the best our knowledge, this is the first nonasymptotic analysis for two-time-scale GDA in this setting, shedding light on its superior practical performance in training generative adversarial networks (GANs) and other real applications.

Extrapolation for Large-batch Training in Deep Learning

- Tao Lin, Lingjing Kong, Sebastian Stich, Martin Jaggi
- abstract: Deep learning networks are typically trained by Stochastic Gradient Descent (SGD) methods that iteratively improve the model parameters by estimating a gradient on a very small fraction of the training data. A major roadblock faced when increasing the batch size to a substantial fraction of the training data for reducing training time is the persistent degradation in performance (generalization gap). To address this issue, recent work propose to add small perturbations to the model parameters when computing the stochastic gradients and report improved generalization performance due to smoothing effects. However, this approach is poorly understood; it requires often model-specific noise and fine-tuning. To alleviate these drawbacks, we propose to use instead computationally efficient extrapolation (extragradient) to stabilize the optimization trajectory while still benefiting from smoothing to avoid sharp minima. This principled approach is well grounded from an optimization perspective and we show that a host of variations can be covered in a unified framework that we propose. We prove the convergence of this novel scheme and rigorously evaluate its empirical performance on ResNet, LSTM, and Transformer. We demonstrate that in a variety of experiments the scheme allows scaling to much larger batch sizes than before whilst reaching or surpassing SOTA accuracy.

On the Theoretical Properties of the Network Jackknife

- Qiaohui Lin, Robert Lunde, Purnamrita Sarkar
- abstract: We study the properties of a leave-node-out jackknife procedure for network data. Under the sparse graphon model, we prove an Efron-Stein-type inequality, showing that the network jackknife leads to conservative estimates of the variance (in expectation) for any network functional that is invariant to node permutation. For a general class of count functionals, we also establish consistency of the network jackknife. We complement our theoretical analysis with a range of simulated and real-data examples and show that the network jackknife offers competitive performance in cases where other resampling methods are known to be valid. In fact, for several network statistics, we see that the jackknife provides more accurate inferences compared to related methods such as subsampling.

Handling the Positive-Definite Constraint in the Bayesian Learning Rule

- Wu Lin, Mark Schmidt, Mohammad Emtiyaz Khan
- abstract: The Bayesian learning rule is a natural-gradient variational inference method, which not only contains many existing learning algorithms as special cases but also enables the design of new algorithms. Unfortunately, when variational parameters lie in an open constraint set, the rule may not satisfy the constraint and requires line-searches which could slow down the algorithm. In this work, we address this issue for positive-definite constraints by proposing an improved rule that naturally handles the constraints. Our modification is obtained by using Riemannian gradient methods, and is valid when the approximation attains a block-coordinate natural parameterization (e.g., Gaussian distributions and their mixtures). Our method outperforms existing methods without any significant increase in computation. Our work makes it easier to apply the rule in the presence of positive-definite constraints in parameter spaces.

InfoGAN-CR and ModelCentrality: Self-supervised Model Training and Selection for Disentangling GANs

- Zinan Lin, Kiran Thekumparampil, Giulia Fanti, Sewoong Oh
- abstract: Disentangled generative models map a latent code vector to a target space, while enforcing that a subset of the learned latent codes are interpretable and associated with distinct properties of the target distribution. Recent advances have been dominated by Variational AutoEncoder (VAE)-based methods, while training disentangled generative adversarial networks (GANs) remains challenging. In this work, we show that the dominant challenges facing disentangled GANs can be mitigated through the use of self-supervision. We make two main contributions: first, we design a novel approach for training disentangled GANs with self-supervision. We propose contrastive regularizer, which is inspired by a natural notion of disentanglement: latent traversal. This achieves higher disentanglement scores than state-of-the-art VAE- and GAN-based approaches. Second, we propose an unsupervised model selection scheme called ModelCentrality, which uses generated synthetic samples to compute the medoid (multi-dimensional generalization of median) of a collection of models. The current common practice of hyper-parameter tuning requires using ground-truths samples, each labelled with known perfect disentangled latent codes. As real datasets are not equipped with such labels, we propose an unsupervised model selection scheme and show that it finds a model close to the best one, for both VAEs and GANs. Combining contrastive regularization with ModelCentrality, we improve upon the state-of-the-art disentanglement scores significantly, without accessing the supervised data.

Improving Generative Imagination in Object-Centric World Models

- Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, Sungjin Ahn
- abstract: The remarkable recent advances in object-centric generative world models raise a few questions. First, while many of the recent achievements are indispensable for making a general and versatile world model, it is quite unclear how these ingredients can be integrated into a unified framework. Second, despite using generative objectives, abilities for object detection and tracking are mainly investigated, leaving the crucial ability of temporal imagination largely under question. Third, a few key abilities for more faithful temporal imagination such as multimodal uncertainty and situation-awareness are missing. In this paper, we introduce Generative Structured World Models (G-SWM). The G-SWM achieves the versatile world modeling not only by unifying the key properties of previous models in a principled framework but also by achieving two crucial new abilities, multimodal uncertainty and situation-awareness. Our thorough investigation on the temporal generation ability in comparison to the previous models demonstrates that G-SWM achieves the versatility with the best or comparable performance for all experiment settings including a few complex settings that have not been tested before. https://sites.google.com/view/gswm

Generalized and Scalable Optimal Sparse Decision Trees

- Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, Margo Seltzer
- abstract: Decision tree optimization is notoriously difficult from a computational perspective but essential for the field of interpretable machine learning. Despite efforts over the past 40 years, only recently have optimization breakthroughs been made that have allowed practical algorithms to find optimal decision trees. These new techniques have the potential to trigger a paradigm shift, where, it is possible to construct sparse decision trees to efficiently optimize a variety of objective functions, without relying on greedy splitting and pruning heuristics that often lead to suboptimal solutions. The contribution in this work is to provide a general framework for decision tree optimization that addresses the two significant open problems in the area: treatment of imbalanced data and fully optimizing over continuous variables. We present techniques that produce optimal decision trees over variety of objectives including F-score, AUC, and partial area under the ROC convex hull. We also introduce a scalable algorithm that produces provably optimal results in the presence of continuous variables and speeds up decision tree construction by several order of magnitude relative to the state-of-the art.

Finite-Time Last-Iterate Convergence for Multi-Agent Learning in Games

- Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, Michael Jordan
- abstract: In this paper, we consider multi-agent learning via online gradient descent in a class of games called λ-cocoercive games, a fairly broad class of games that admits many Nash equilibria and that properly includes unconstrained strongly monotone games. We characterize the finite-time last-iterate convergence rate for joint OGD learning on λ-cocoercive games; further, building on this result, we develop a fully adaptive OGD learning algorithm that does not require any knowledge of problem parameter (e.g. cocoercive constant λ) and show, via a novel double-stopping time technique, that this adaptive algorithm achieves same finite-time last-iterate convergence rate as non-adaptive counterpart. Subsequently, we extend OGD learning to the noisy gradient feedback case and establish last-iterate convergence results–first qualitative almost sure convergence, then quantitative

finite-time convergence rates– all under non-decreasing step-sizes. To our knowledge, we provide the first set of results that fill in several gaps of the existing multi-agent online learning literature, where three aspects–finite-time convergence rates, non-decreasing step-sizes, and fully adaptive algorithms have been unexplored before.

Time-aware Large Kernel Convolutions

- Vasileios Lioutas, Yuhong Guo
- abstract: To date, most state-of-the-art sequence modeling architectures use attention to build generative models for language based tasks. Some of these models use all the available sequence tokens to generate an attention distribution which results in time complexity of $O(n^2)$. Alternatively, they utilize depthwise convolutions with softmax normalized kernels of size k acting as a limited-window self-attention, resulting in time complexity of $O(k{\cdot}n)$. In this paper, we introduce Time-aware Large Kernel (TaLK) Convolutions, a novel adaptive convolution operation that learns to predict the size of a summation kernel instead of using a fixed-sized kernel matrix. This method yields a time complexity of $O(n)$, effectively making the sequence encoding process linear to the number of tokens. We evaluate the proposed method on large-scale standard machine translation, abstractive summarization and language modeling datasets and show that TaLK Convolutions constitute an efficient improvement over other attention/convolution based approaches.

Understanding the Curse of Horizon in Off-Policy Evaluation via Conditional Importance Sampling

- Yao Liu, Pierre-Luc Bacon, Emma Brunskill
- abstract: Off-policy policy estimators that use importance sampling (IS) can suffer from high variance in long-horizon domains, and there has been particular excitement over new IS methods that leverage the structure of Markov decision processes. We analyze the variance of the most popular approaches through the viewpoint of conditional Monte Carlo. Surprisingly, we find that in finite horizon MDPs there is no strict variance reduction of per-decision importance sampling or marginalized importance sampling, comparing with vanilla importance sampling. We then provide sufficient conditions under which the per-decision or marginalized estimators will provably reduce the variance over importance sampling with finite horizons. For the asymptotic (in terms of horizon T) case, we develop upper and lower bounds on the variance of those estimators which yields sufficient conditions under which there exists an exponential v.s. polynomial gap between the variance of importance sampling and that of the per-decision or stationary/marginalized estimators. These results help advance our understanding of if and when new types of IS estimators will improve the accuracy of off-policy estimation.

Sparse Shrunk Additive Models

- Guodong Liu, Hong Chen, Heng Huang
- abstract: Most existing feature selection methods in literature are linear models, so that the nonlinear relations between features and response variables are not considered. Meanwhile, in these feature selection models, the interactions between features are often ignored or just discussed under prior structure information. To address these challenging issues, we consider the problem of sparse additive models for high-dimensional nonparametric regression with the allowance of the flexible interactions between features. A new method, called as sparse shrunk additive models (SSAM), is proposed to explore the structure information among features. This method bridges sparse kernel regression and sparse feature selection. Theoretical results on the convergence rate and sparsity characteristics of SSAM are established by the novel analysis techniques with integral operator and concentration estimate. In particular, our algorithm and theoretical analysis only require the component functions to be continuous and bounded, which are not necessary to be in reproducing kernel Hilbert spaces. Experiments on both synthetic and real-world data demonstrate the effectiveness of the proposed approach.

Boosting Deep Neural Network Efficiency with Dual-Module Inference

- Liu Liu, Lei Deng, Zhaodong Chen, Yuke Wang, Shuangchen Li, Jingwei Zhang, Yihua Yang, Zhenyu Gu, Yufei Ding, Yuan Xie
- abstract: Using deep neural networks (DNNs) in machine learning tasks is promising in delivering high-quality results but challenging to meet stringent latency requirements and energy constraints because of the memory-bound and the compute-bound execution pattern of DNNs. We propose a big-little dual-module inference to dynamically skip unnecessary memory accesses and computations to accelerate DNN inference. Leveraging the noise-resilient feature of nonlinear activation functions, we propose to use a lightweight little module that approximates the original DNN layer, termed as the big module, to compute activations of the insensitive region that are more noise-resilient. Hence, the expensive memory accesses and computations of the big module can be reduced as the results are only calculated in the sensitive region. For memory-bound models such as recurrent neural networks (RNNs), our method can reduce the overall memory accesses by 40% on average and achieve 1.54x to 1.75x speedup on a commodity CPU-based server platform with a negligible impact on model quality. In addition, our method can reduce the operations of the compute-bound models such as convolutional neural networks (CNNs) by 3.02x, with only a 0.5% accuracy drop.

Sample Complexity Bounds for 1-bit Compressive Sensing and Binary Stable Embeddings with Generative Priors

- Zhaoqiang Liu, Selwyn Gomes, Avtansh Tiwari, Jonathan Scarlett
- abstract: The goal of standard 1-bit compressive sensing is to accurately recover an unknown sparse vector from binary-valued measurements, each indicating the sign of a linear function of the vector. Motivated by recent advances in compressive sensing with generative models, where a generative modeling assumption replaces the usual sparsity assumption, we study the problem of 1-bit compressive sensing with generative models. We first consider noiseless 1-bit measurements, and provide sample complexity bounds for approximate recovery under i.i.d. Gaussian measurements and a Lipschitz continuous generative prior, as well as a near-matching algorithm-independent lower bound. Moreover, we demonstrate that the Binary ϵ-Stable Embedding property, which characterizes the robustness of the reconstruction to measurement errors and noise, also holds for 1-bit compressive sensing with Lipschitz continuous generative models with sufficiently many Gaussian measurements. In addition, we apply our results to neural network generative models, and provide a proof-of-concept numerical experiment demonstrating significant improvements over sparsity-based approaches.

Peer Loss Functions: Learning from Noisy Labels without Knowing Noise Rates

- Yang Liu, Hongyi Guo
- abstract: Learning with noisy labels is a common challenge in supervised learning. Existing approaches often require practitioners to specify noise rates, i.e., a set of parameters controlling the severity of label noises in the problem, and the specifications are either assumed to be given or estimated using additional steps. In this work, we introduce a new family of loss functions that we name as peer loss functions, which enables learning from noisy labels and does not require a priori specification of the noise rates. Peer loss functions work within the standard empirical risk minimization (ERM) framework. We show that, under mild conditions, performing ERM with peer loss functions on the noisy data leads to the optimal or a near-optimal classifier as if performing ERM over the clean training data, which we do not have access to. We pair our results with an extensive set of experiments. Peer loss provides a way to simplify model development when facing potentially noisy training labels, and can be promoted as a robust candidate loss function in such situations.

An Imitation Learning Approach for Cache Replacement

- Evan Liu, Milad Hashemi, Kevin Swersky, Parthasarathy Ranganathan, Junwhan Ahn

- abstract: Program execution speed critically depends on increasing cache hits, as cache hits are orders of magnitude faster than misses. To increase cache hits, we focus on the problem of cache replacement: choosing which cache line to evict upon inserting a new line. This is challenging because it requires planning far ahead and currently there is no known practical solution. As a result, current replacement policies typically resort to heuristics designed for specific common access patterns, which fail on more diverse and complex access patterns. In contrast, we propose an imitation learning approach to automatically learn cache access patterns by leveraging Belady's, an oracle policy that computes the optimal eviction decision given the future cache accesses. While directly applying Belady's is infeasible since the future is unknown, we train a policy conditioned only on past accesses that accurately approximates Belady's even on diverse and complex access patterns, and call this approach Parrot. When evaluated on 13 of the most memory-intensive SPEC applications, Parrot increases cache miss rates by 20% over the current state of the art. In addition, on a large-scale web search benchmark, Parrot increases cache hit rates by 61% over a conventional LRU policy. We release a Gym environment to facilitate research in this area, as data is plentiful, and further advancements can have significant real-world impact.

Exploration Through Reward Biasing: Reward-Biased Maximum Likelihood Estimation for Stochastic Multi-Armed Bandits

- Xi Liu, Ping-Chun Hsieh, Yu Heng Hung, Anirban Bhattacharya, P. Kumar
- abstract: Inspired by the Reward-Biased Maximum Likelihood Estimate method of adaptive control, we propose RBMLE – a novel family of learning algorithms for stochastic multi-armed bandits (SMABs). For a broad range of SMABs including both the parametric Exponential Family as well as the non-parametric sub-Gaussian/Exponential family, we show that RBMLE yields an index policy. To choose the bias-growth rate $\alpha(t)$ in RBMLE, we reveal the nontrivial interplay between $\alpha(t)$ and the regret bound that generally applies in both the Exponential Family as well as the sub-Gaussian/Exponential family bandits. To quantify the finite-time performance, we prove that RBMLE attains order-optimality by adaptively estimating the unknown constants in the expression of $\alpha(t)$ for Gaussian and sub-Gaussian bandits. Extensive experiments demonstrate that the proposed RBMLE achieves empirical regret performance competitive with the state-of-the-art methods, while being more computationally efficient and scalable in comparison to the best-performing ones among them.

Hallucinative Topological Memory for Zero-Shot Visual Planning

- Kara Liu, Thanard Kurutach, Christine Tung, Pieter Abbeel, Aviv Tamar
- abstract: In visual planning (VP), an agent learns to plan goal-directed behavior from observations of a dynamical system obtained offline, e.g., images obtained from self-supervised robot interaction. Most previous works on VP approached the problem by planning in a learned latent space, resulting in low-quality visual plans, and difficult training algorithms. Here, instead, we propose a simple VP method that plans directly in image space and displays competitive performance. We build on the semi-parametric topological memory (SPTM) method: image samples are treated as nodes in a graph, the graph connectivity is learned from image sequence data, and planning can be performed using conventional graph search methods. We propose two modifications on SPTM. First, we train an energy-based graph connectivity function using contrastive predictive coding that admits stable training. Second, to allow zero-shot planning in new domains, we learn a conditional VAE model that generates images given a context describing the domain, and use these hallucinated samples for building the connectivity graph and planning. We show that this simple approach significantly outperform the SOTA VP methods, in terms of both plan interpretability and success rate when using the plan to guide a trajectory-following controller. Interestingly, our method can pick up non-trivial visual properties of objects, such as their geometry, and account for it in the plans.

A Chance-Constrained Generative Framework for Sequence Optimization

- Xianggen Liu, Qiang Liu, Sen Song, Jian Peng
- abstract: Deep generative modeling has achieved many successes for continuous data generation, such as producing realistic images and controlling their properties (e.g., styles). However, the development of generative modeling techniques for optimizing discrete data, such as sequences or strings, still lags behind largely due to the challenges in modeling complex and long-range constraints, including both syntax and semantics, in discrete structures. In this paper, we formulate the sequence optimization task as a chance-constrained optimization problem. The key idea is to enforce a high probability of generating valid sequences and also optimize the property of interest. We propose a novel minimax algorithm to simultaneously tighten a bound of the valid chance and optimize the expected property. Extensive experimental results in three domains demonstrate the superiority of our approach over the existing sequence optimization methods.

Min-Max Optimization without Gradients: Convergence and Applications to Black-Box Evasion and Poisoning Attacks

- Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Mingyi Hong, Una-May O'Reilly
- abstract: In this paper, we study the problem of constrained min-max optimization in a black-box setting, where the desired optimizer cannot access the gradients of the objective function but may query its values. We present a principled optimization framework, integrating a zeroth-order (ZO) gradient estimator with an alternating projected stochastic gradient descent-ascent method, where the former only requires a small number of function queries and the later needs just one-step descent/ascent update. We show that the proposed framework, referred to as ZO-Min-Max, has a sublinear convergence rate under mild conditions and scales gracefully with problem size. We also explore a promising connection between black-box min-max optimization and black-box evasion and poisoning attacks in adversarial machine learning (ML). Our empirical evaluations on these use cases demonstrate the effectiveness of our approach and its scalability to dimensions that prohibit using recent black-box solvers.

Median Matrix Completion: from Embarrassment to Optimality

- Weidong Liu, Xiaojun Mao, Raymond K. W. Wong
- abstract: In this paper, we consider matrix completion with absolute deviation loss and obtain an estimator of the median matrix. Despite several appealing properties of median, the non-smooth absolute deviation loss leads to computational challenge for large-scale data sets which are increasingly common among matrix completion problems. A simple solution to large-scale problems is parallel computing. However, embarrassingly parallel fashion often leads to inefficient estimators. Based on the idea of pseudo data, we propose a novel refinement step, which turns such inefficient estimators into a rate (near-)optimal matrix completion procedure. The refined estimator is an approximation of a regularized least median estimator, and therefore not an ordinary regularized empirical risk estimator. This leads to a non-standard analysis of asymptotic behaviors. Empirical results are also provided to confirm the effectiveness of the proposed method.

A Generic First-Order Algorithmic Framework for Bi-Level Programming Beyond Lower-Level Singleton

- Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, Jin Zhang
- abstract: In recent years, a variety of gradient-based bi-level optimization methods have been developed for learning tasks. However, theoretical guarantees of these existing approaches often heavily rely on the simplification that for each fixed upper-level variable, the lower-level solution must be a singleton (a.k.a., Lower-Level Singleton, LLS). In this work, by formulating bi-level models from the optimistic viewpoint and aggregating hierarchical objective information, we establish Bi-level Descent Aggregation (BDA), a flexible and modularized algorithmic framework for bi-level programming. Theoretically, we derive a new methodology to prove the convergence of BDA without the LLS condition. Furthermore, we improve the convergence properties of conventional first-order bi-level schemes (under the LLS simplification) based on our proof recipe. Extensive experiments justify our theoretical results and demonstrate the superiority of the proposed BDA for different tasks, including hyper-parameter optimization and meta learning.

Learning Deep Kernels for Non-Parametric Two-Sample Tests

- Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, Danica J. Sutherland
- abstract: We propose a class of kernel-based two-sample tests, which aim to determine whether two sets of samples are drawn from the same distribution. Our tests are constructed from kernels parameterized by deep neural nets, trained to maximize test power. These tests adapt to variations in distribution smoothness and shape over space, and are especially suited to high dimensions and complex data. By contrast, the simpler kernels used in prior kernel testing work are spatially homogeneous, and adaptive only in lengthscale. We explain how this scheme includes popular classifier-based two-sample tests as a special case, but improves on them in general. We provide the first proof of consistency for the proposed adaptation method, which applies both to kernels on deep features and to simpler radial basis kernels or multiple kernel learning. In experiments, we establish the superior performance of our deep kernels in hypothesis testing on benchmark and real-world data. The code of our deep-kernel-based two-sample tests is available at github.com/fengliu90/DK-for-TST.

Learning to Encode Position for Transformer with Continuous Dynamical Model

- Xuanqing Liu, Hsiang-Fu Yu, Inderjit Dhillon, Cho-Jui Hsieh
- abstract: We introduce a new way of learning to encode position information for non-recurrent models, such as Transformer models. Unlike RNN and LSTM, which contain inductive bias by loading the input tokens sequentially, non-recurrent models are less sensitive to position. The main reason is that position information among input units is not encoded inherently, i.e., they are permutation equivalent, this problem justifies why all of the existing models are accompanied by position encoding/embedding layer at the input. However, this solution has clear limitations: the sinusoidal position encoding is not flexible enough as it is manually designed and does not contain any learnable parameters, whereas the position embedding restricts the maximum length of input sequences. It is thus desirable to design a new position layer that contains learnable parameters to adjust to different datasets and different architectures. At the same time, we would also like it to extrapolate in accordance with the variable length of inputs. In our proposed solution, we borrow from the recent Neural ODE approach, which may be viewed as a versatile continuous version of a ResNet. This model is capable of modeling many kinds of dynamical systems. We model the evolution of encoded results along position index by such a dynamical system, thereby overcoming the above limitations of existing methods. We evaluate our new position layers on a variety of neural machine translation and language understanding tasks, the experimental results show consistent improvements over the baselines.

Finding trainable sparse networks through Neural Tangent Transfer

- Tianlin Liu, Friedemann Zenke
- abstract: Deep neural networks have dramatically transformed machine learning, but their memory and energy demands are substantial. The requirements of real biological neural networks are rather modest in comparison, and one feature that might underlie this austerity is their sparse connectivity. In deep learning, trainable sparse networks that perform well on a specific task are usually constructed using label-dependent pruning criteria. In this article, we introduce Neural Tangent Transfer, a method that instead finds trainable sparse networks in a label-free manner. Specifically, we find sparse networks whose training dynamics, as characterized by the neural tangent kernel, mimic those of dense networks in function space. Finally, we evaluate our label-agnostic approach on several standard classification tasks and show that the resulting sparse networks achieve higher classification performance while converging faster.

Weakly-Supervised Disentanglement Without Compromises

- Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, Michael Tschannen
- abstract: Intelligent agents should be able to learn useful representations by observing changes in their environment. We model such observations as pairs of non-i.i.d. images sharing at least one of the underlying factors of variation. First, we theoretically show that only knowing how many factors have changed, but not which ones, is sufficient to learn disentangled representations. Second, we provide practical algorithms that learn disentangled representations from pairs of images without requiring annotation of groups, individual factors, or the number of factors that have changed. Third, we perform a large-scale empirical study and show that such pairs of observations are sufficient to reliably learn disentangled representations on several benchmark data sets. Finally, we evaluate our learned representations and find that they are simultaneously useful on a diverse suite of tasks, including generalization under covariate shifts, fairness, and abstract reasoning. Overall, our results demonstrate that weak supervision enables learning of useful disentangled representations in realistic scenarios.

Too Relaxed to Be Fair

- Michael Lohaus, Michael Perrot, Ulrike Von Luxburg
- abstract: We address the problem of classification under fairness constraints. Given a notion of fairness, the goal is to learn a classifier that is not discriminatory against a group of individuals. In the literature, this problem is often formulated as a constrained optimization problem and solved using relaxations of the fairness constraints. We show that many existing relaxations are unsatisfactory: even if a model satisfies the relaxed constraint, it can be surprisingly unfair. We propose a principled framework to solve this problem. This new approach uses a strongly convex formulation and comes with theoretical guarantees on the fairness of its solution. In practice, we show that this method gives promising results on real data.

Stochastic Hamiltonian Gradient Methods for Smooth Games

- Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, Ioannis Mitliagkas
- abstract: The success of adversarial formulations in machine learning has brought renewed motivation for smooth games. In this work, we focus on the class of stochastic Hamiltonian methods and provide the first convergence guarantees for certain classes of stochastic smooth games. We propose a novel unbiased estimator for the stochastic Hamiltonian gradient descent (SHGD) and highlight its benefits. Using tools from the optimization literature we show that SHGD converges linearly to the neighbourhood of a stationary point. To guarantee convergence to the exact solution, we analyze SHGD with a decreasing step-size and we also present the first stochastic variance reduced Hamiltonian method. Our results provide the first global non-asymptotic last-iterate convergence guarantees for the class of stochastic unconstrained bilinear games and for the more general class of stochastic games that satisfy a “sufficiently bilinear" condition, notably including some non-convex non-concave problems. We supplement our analysis with experiments on stochastic bilinear and sufficiently bilinear games, where our theory is shown to be tight, and on simple adversarial machine learning formulations.

Error Estimation for Sketched SVD via the Bootstrap

- Miles Lopes, N. Benjamin Erichson, Michael Mahoney
- abstract: In order to compute fast approximations to the singular value decompositions (SVD) of very large matrices, randomized sketching algorithms have become a leading approach. However, a key practical difficulty of sketching an SVD is that the user does not know how far the sketched singular vectors/values are from the exact ones. Indeed, the user may be forced to rely on analytical worst-case error bounds, which may not account for the unique structure of a given problem. As a result, the lack of tools for error estimation often leads to much more computation than is really necessary. To overcome these challenges, this paper develops a fully data-driven bootstrap method that numerically estimates the actual error of sketched singular vectors/values. Furthermore, the method is computationally inexpensive, because it operates only on sketched objects, and hence it requires no extra passes over the full matrix being factored.

Differentiating through the Fréchet Mean

- Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, Christopher De Sa
- abstract: Recent advances in deep representation learning on Riemannian manifolds extend classical deep learning operations to better capture the geometry of the manifold. One possible extension is the Fr{é}chet mean, the generalization of the Euclidean mean; however, it has been difficult to apply because it lacks a closed form with an easily computable derivative. In this paper, we show how to differentiate through the Fr{é}chet mean for arbitrary Riemannian manifolds. Then, focusing on hyperbolic space, we derive explicit gradient expressions and a fast, accurate, and hyperparameter-free Fr{é}chet mean solver. This fully integrates the Fr{é}chet mean into the hyperbolic neural network pipeline. To demonstrate this integration, we present two case studies. First, we apply our Fr{é}chet mean to the existing Hyperbolic Graph Convolutional Network, replacing its projected aggregation to obtain state-of-the-art results on datasets with high hyperbolicity. Second, to demonstrate the Fr{é}chet mean's capacity to generalize Euclidean neural network operations, we develop a hyperbolic batch normalization method that gives an improvement parallel to the one observed in the Euclidean setting.

Working Memory Graphs

- Ricky Loynd, Roland Fernandez, Asli Celikyilmaz, Adith Swaminathan, Matthew Hausknecht
- abstract: Transformers have increasingly outperformed gated RNNs in obtaining new state-of-the-art results on supervised tasks involving text sequences. Inspired by this trend, we study the question of how Transformer-based models can improve the performance of sequential decision-making agents. We present the Working Memory Graph (WMG), an agent that employs multi-head self-attention to reason over a dynamic set of vectors representing observed and recurrent state. We evaluate WMG in three environments featuring factored observation spaces: a Pathfinding environment that requires complex reasoning over past observations, BabyAI gridworld levels that involve variable goals, and Sokoban which emphasizes future planning. We find that the combination of WMG's Transformer-based architecture with factored observation spaces leads to significant gains in learning efficiency compared to baseline architectures across all tasks. WMG demonstrates how Transformer-based models can dramatically boost sample efficiency in RL environments for which observations can be factored.

Moniqua: Modulo Quantized Communication in Decentralized SGD

- Yucheng Lu, Christopher De Sa
- abstract: Running Stochastic Gradient Descent (SGD) in a decentralized fashion has shown promising results. In this paper we propose Moniqua, a technique that allows decentralized SGD to use quantized communication. We prove in theory that Moniqua communicates a provably bounded number of bits per iteration, while converging at the same asymptotic rate as the original algorithm does with full-precision communication. Moniqua improves upon prior works in that it (1) requires zero additional memory, (2) works with 1-bit quantization, and (3) is applicable to a variety of decentralized algorithms. We demonstrate empirically that Moniqua converges faster with respect to wall clock time than other quantized decentralized algorithms. We also show that Moniqua is robust to very low bit-budgets, allowing 1-bit-per-parameter communication without compromising validation accuracy when training ResNet20 and ResNet110 on CIFAR10.

A Mean Field Analysis Of Deep ResNet And Beyond: Towards Provably Optimization Via Overparameterization From Depth

- Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, Lexing Ying
- abstract: Training deep neural networks with stochastic gradient descent (SGD) can often achieve zero training loss on real-world tasks although the optimization landscape is known to be highly non-convex. To understand the success of SGD for training deep neural networks, this work presents a mean-field analysis of deep residual networks, based on a line of works which interpret the continuum limit of the deep residual network as an ordinary differential equation as the the network capacity tends to infinity. Specifically, we propose a \textbf{new continuum limit} of deep residual networks, which enjoys a good landscape in the sense that \textbf{every local minimizer is global}. This characterization enables us to derive the first global convergence result for multilayer neural networks in the mean-field regime. Furthermore, our proof does not rely on the convexity of the loss landscape, but instead, an assumption on the global minimizer should achieve zero loss which can be achieved when the model shares a universal approximation property. Key to our result is the observation that a deep residual network resembles a shallow network ensemble \cite{veit2016residual}, \emph{i.e.} a two-layer network. We bound the difference between the shallow network and our ResNet model via the adjoint sensitivity method, which enables us to transfer previous mean-field analysis of two-layer networks to deep networks. Furthermore, we propose several novel training schemes based on our new continuous model, among which one new training procedure introduces the operation of switching the order of the residual blocks and results in strong empirical performance on benchmark datasets.

Countering Language Drift with Seeded Iterated Learning

- Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, Olivier Pietquin
- abstract: Pretraining on human corpus and then finetuning in a simulator has become a standard pipeline for training a goal-oriented dialogue agent. Nevertheless, as soon as the agents are finetuned to maximize task completion, they suffer from the so-called language drift phenomenon: they slowly lose syntactic and semantic properties of language as they only focus on solving the task. In this paper, we propose a generic approach to counter language drift called Seeded iterated learning (SIL). We periodically refine a pretrained student agent by imitating data sampled from a newly generated teacher agent. At each time step, the teacher is created by copying the student agent, before being finetuned to maximize task completion. SIL does not require external syntactic constraint nor semantic knowledge, making it a valuable task-agnostic finetuning protocol. We evaluate SIL in a toy-setting Lewis Game, and then scale it up to the translation game with natural language. In both settings, SIL helps counter language drift as well as it improves the task completion compared to baselines.

Does label smoothing mitigate label noise?

- Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, Sanjiv Kumar
- abstract: Label smoothing is commonly used in training deep learning models, wherein one-hot training labels are mixed with uniform label vectors. Empirically, smoothing has been shown to improve both predictive performance and model calibration. In this paper, we study whether label smoothing is also effective as a means of coping with label noise. While label smoothing apparently amplifies this problem — being equivalent to injecting symmetric noise to the labels — we show how it relates to a general family of loss-correction techniques from the label noise literature. Building on this connection, we show that label smoothing is competitive with loss-correction under label noise. Further, we show that when distilling models from noisy data, label smoothing of the teacher is beneficial; this is in contrast to recent findings for noise-free problems, and sheds further light on settings where label smoothing is beneficial.

Improved Communication Cost in Distributed PageRank Computation – A Theoretical Study

- Siqiang Luo
- abstract: PageRank is a widely used approach for measuring the importance of a node in a graph. Due to the rapid growth of the graph size in the real world, the importance of computing PageRanks in a distributed environment has been increasingly recognized. However, only a few previous works can provide a provable complexity and accuracy for distributed PageRank computation. Given a constant $d\ge 1$ and a graph of n nodes, the state-of-the-art approach, Radar-Push, uses $O(\log\log{n}+\log{d})$ communication rounds to approximate the PageRanks within a relative error $\Theta(\frac{1}

{\log^d{n}})$ under a generalized congested clique distributed computation model. However, Radar-Push entails as large as $O(\log^{2d+3}{n})$ bits of bandwidth (e.g., the communication cost between a pair of nodes per round). In this paper, we provide a new algorithm that uses asymptotically the same communication round complexity while using only $O(d\log^3{n})$ bits of bandwidth.

Progressive Graph Learning for Open-Set Domain Adaptation

- Yadan Luo, Zijian Wang, Zi Huang, Mahsa Baktashmotlagh
- abstract: Domain shift is a fundamental problem in visual recognition which typically arises when the source and target data follow different distributions. The existing domain adaptation approaches which tackle this problem work in the "closed-set" setting with the assumption that the source and the target data share exactly the same classes of objects. In this paper, we tackle a more realistic problem of the "open-set" domain shift where the target data contains additional classes that were not present in the source data. More specifically, we introduce an end-to-end Progressive Graph Learning (PGL) framework where a graph neural network with episodic training is integrated to suppress underlying conditional shift and adversarial learning is adopted to close the gap between the source and target distributions. Compared to the existing open-set adaptation approaches, our approach guarantees to achieve a tighter upper bound of the target error. Extensive experiments on three standard open-set benchmarks evidence that our approach significantly outperforms the state-of-the-arts in open-set domain adaptation.

Adversarial Nonnegative Matrix Factorization

- Lei Luo, Yanfu Zhang, Heng Huang
- abstract: Nonnegative Matrix Factorization (NMF) has become an increasingly important research topic in machine learning. Despite all the practical success, most of existing NMF models are still vulnerable to adversarial attacks. To overcome this limitation, we propose a novel Adversarial NMF (ANMF) approach in which an adversary can exercise some control over the perturbed data generation process. Different from the traditional NMF models which focus on either the regular input or certain types of noise, our model considers potential test adversaries that are beyond the pre-defined constraints, which can cope with various noises (or perturbations). We formulate the proposed model as a bilevel optimization problem and use Alternating Direction Method of Multipliers (ADMM) to solve it with convergence analysis. Theoretically, the robustness analysis of ANMF is established under mild conditions dedicating asymptotically unbiased prediction. Extensive experiments verify that ANMF is robust to a broad categories of perturbations, and achieves state-of-the-art performances on distinct real-world benchmark datasets.

Learning Algebraic Multigrid Using Graph Neural Networks

- Ilay Luz, Meirav Galun, Haggai Maron, Ronen Basri, Irad Yavneh
- abstract: Efficient numerical solvers for sparse linear systems are crucial in science and engineering. One of the fastest methods for solving large-scale sparse linear systems is algebraic multigrid (AMG). The main challenge in the construction of AMG algorithms is the selection of the prolongation operator—a problem-dependent sparse matrix which governs the multiscale hierarchy of the solver and is critical to its efficiency. Over many years, numerous methods have been developed for this task, and yet there is no known single right answer except in very special cases. Here we propose a framework for learning AMG prolongation operators for linear systems with sparse symmetric positive (semi-) definite matrices. We train a single graph neural network to learn a mapping from an entire class of such matrices to prolongation operators, using an efficient unsupervised loss function. Experiments on a broad class of problems demonstrate improved convergence rates compared to classical AMG, demonstrating the potential utility of neural networks for developing sparse system solvers.

Progressive Identification of True Labels for Partial-Label Learning

- Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, Masashi Sugiyama
- abstract: Partial-label learning (PLL) is a typical weakly supervised learning problem, where each training instance is equipped with a set of candidate labels among which only one is the true label. Most existing methods elaborately designed learning objectives as constrained optimizations that must be solved in specific manners, making their computational complexity a bottleneck for scaling up to big data. The goal of this paper is to propose a novel framework of PLL with flexibility on the model and optimization algorithm. More specifically, we propose a novel estimator of the classification risk, theoretically analyze the classifier-consistency, and establish an estimation error bound. Then we propose a progressive identification algorithm for approximately minimizing the proposed risk estimator, where the update of the model and identification of true labels are conducted in a seamless manner. The resulting algorithm is model-independent and loss-independent, and compatible with stochastic optimization. Thorough experiments demonstrate it sets the new state of the art.

Bandits with Adversarial Scaling

- Thodoris Lykouris, Vahab Mirrokni, Renato Paes Leme
- abstract: We study "adversarial scaling", a multi-armed bandit model where rewards have a stochastic and an adversarial component. Our model captures display advertising where the "click-through-rate" can be decomposed to a (fixed across time) arm-quality component and a non-stochastic user-relevance component (fixed across arms). Despite the relative stochasticity of our model, we demonstrate two settings where most bandit algorithms suffer. On the positive side, we show that two algorithms, one from the action elimination and one from the mirror descent family are adaptive enough to be robust to adversarial scaling. Our results shed light on the robustness of adaptive parameter selection in stochastic bandits, which may be of independent interest.

Efficient Continuous Pareto Exploration in Multi-Task Learning

- Pingchuan Ma, Tao Du, Wojciech Matusik
- abstract: Tasks in multi-task learning often correlate, conflict, or even compete with each other. As a result, a single solution that is optimal for all tasks rarely exists. Recent papers introduced the concept of Pareto optimality to this field and directly cast multi-task learning as multi-objective optimization problems, but solutions returned by existing methods are typically finite, sparse, and discrete. We present a novel, efficient method that generates locally continuous Pareto sets and Pareto fronts, which opens up the possibility of continuous analysis of Pareto optimal solutions in machine learning problems. We scale up theoretical results in multi-objective optimization to modern machine learning problems by proposing a sample-based sparse linear system, for which standard Hessian-free solvers in machine learning can be applied. We compare our method to the state-of-the-art algorithms and demonstrate its usage of analyzing local Pareto sets on various multi-task classification and regression problems. The experimental results confirm that our algorithm reveals the primary directions in local Pareto sets for trade-off balancing, finds more solutions with different trade-offs efficiently, and scales well to tasks with millions of parameters.

Convex Representation Learning for Generalized Invariance in Semi-Inner-Product Space

- Yingyi Ma, Vignesh Ganapathiraman, Yaoliang Yu, Xinhua Zhang
- abstract: Invariance (defined in a general sense) has been one of the most effective priors for representation learning. Direct factorization of parametric models is feasible only for a small range of invariances, while regularization approaches, despite improved generality, lead to nonconvex optimization. In this work, we develop a \emph{convex} representation learning algorithm for a variety of generalized invariances that can be modeled as semi-norms.

Novel Euclidean embeddings are introduced for kernel representers in a semi-inner-product space, and approximation bounds are established. This allows invariant representations to be learned efficiently and effectively as confirmed in our experiments, along with accurate predictions.

Normalized Loss Functions for Deep Learning with Noisy Labels

- Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, James Bailey
- abstract: Robust loss functions are essential for training accurate deep neural networks (DNNs) in the presence of noisy (incorrect) labels. It has been shown that the commonly used Cross Entropy (CE) loss is not robust to noisy labels. Whilst new loss functions have been designed, they are only partially robust. In this paper, we theoretically show by applying a simple normalization that: \emph{any loss can be made robust to noisy labels}. However, in practice, simply being robust is not sufficient for a loss function to train accurate DNNs. By investigating several robust loss functions, we find that they suffer from a problem of \emph{underfitting}. To address this, we propose a framework to build robust loss functions called \emph{Active Passive Loss} (APL). APL combines two robust loss functions that mutually boost each other. Experiments on benchmark datasets demonstrate that the family of new loss functions created by our APL framework can consistently outperform state-of-the-art methods by large margins, especially under large noise rates such as 60% or 80% incorrect labels.

Quadratically Regularized Subgradient Methods for Weakly Convex Optimization with Weakly Convex Constraints

- Runchao Ma, Qihang Lin, Tianbao Yang
- abstract: Optimization models with non-convex constraints arise in many tasks in machine learning, e.g., learning with fairness constraints or Neyman-Pearson classification with non-convex loss. Although many efficient methods have been developed with theoretical convergence guarantees for non-convex unconstrained problems, it remains a challenge to design provably efficient algorithms for problems with non-convex functional constraints. This paper proposes a class of subgradient methods for constrained optimization where the objective function and the constraint functions are weakly convex and nonsmooth. Our methods solve a sequence of strongly convex subproblems, where a quadratic regularization term is added to both the objective function and each constraint function. Each subproblem can be solved by various algorithms for strongly convex optimization. Under a uniform Slater's condition, we establish the computation complexities of our methods for finding a nearly stationary point.

Understanding the Impact of Model Incoherence on Convergence of Incremental SGD with Random Reshuffle

- Shaocong Ma, Yi Zhou
- abstract: Although SGD with random reshuffle has been widely-used in machine learning applications, there is a limited understanding of how model characteristics affect the convergence of the algorithm. In this work, we introduce model incoherence to characterize the diversity of model characteristics and study its impact on convergence of SGD with random reshuffle under weak strong convexity. Specifically, minimizer incoherence measures the discrepancy between the global minimizers of a sample loss and those of the total loss and affects the convergence error of SGD with random reshuffle. In particular, we show that the variable sequence generated by SGD with random reshuffle converges to a certain global minimizer of the total loss under full minimizer coherence. The other curvature incoherence measures the quality of condition numbers of the sample losses and determines the convergence rate of SGD. With model incoherence, our results show that SGD has a faster convergence rate and smaller convergence error under random reshuffle than those under random sampling, and hence provide justifications to the superior practical performance of SGD with random reshuffle.

Adversarial Neural Pruning with Latent Vulnerability Suppression

- Divyam Madaan, Jinwoo Shin, Sung Ju Hwang
- abstract: Despite the remarkable performance of deep neural networks on various computer vision tasks, they are known to be susceptible to adversarial perturbations, which makes it challenging to deploy them in real-world safety-critical applications. In this paper, we conjecture that the leading cause of adversarial vulnerability is the distortion in the latent feature space, and provide methods to suppress them effectively. Explicitly, we define \emph{vulnerability} for each latent feature and then propose a new loss for adversarial learning, \emph{Vulnerability Suppression (VS)} loss, that aims to minimize the feature-level vulnerability during training. We further propose a Bayesian framework to prune features with high vulnerability to reduce both vulnerability and loss on adversarial samples. We validate our \emph{Adversarial Neural Pruning with Vulnerability Suppression (ANP-VS)} method on multiple benchmark datasets, on which it not only obtains state-of-the-art adversarial robustness but also improves the performance on clean examples, using only a fraction of the parameters used by the full network. Further qualitative analysis suggests that the improvements come from the suppression of feature-level vulnerability.

Individual Fairness for k-Clustering

- Sepideh Mahabadi, Ali Vakilian
- abstract: We give a local search based algorithm for k-median and k-means (and more generally for any k-clustering with ℓ_p norm cost function) from the perspective of individual fairness. More precisely, for a point x in a point set P of size n, let $r(x)$ be the minimum radius such that the ball of radius $r(x)$ centered at x has at least n/k points from P. Intuitively, if a set of k random points are chosen from P as centers, every point $x\in P$ expects to have a center within radius $r(x)$. In this work, we show how to get an approximately optimal such fair k-clustering: The k-median (k-means) cost of our solution is within a constant factor of the cost of an optimal fair k-clustering, and our solution approximately satisfies the fairness condition (also within a constant factor).

Multi-Task Learning with User Preferences: Gradient Descent with Controlled Ascent in Pareto Optimization

- Debabrata Mahapatra, Vaibhav Rajan
- abstract: Multi-Task Learning (MTL) is a well established paradigm for jointly learning models for multiple correlated tasks. Often the tasks conflict, requiring trade-offs between them during optimization. In such cases, multi-objective optimization based MTL methods can be used to find one or more Pareto optimal solutions. A common requirement in MTL applications, that cannot be addressed by these methods, is to find a solution satisfying userspecified preferences with respect to task-specific losses. We advance the state-of-the-art by developing the first gradient-based multi-objective MTL algorithm to solve this problem. Our unique approach combines multiple gradient descent with carefully controlled ascent to traverse the Pareto front in a principled manner, which also makes it robust to initialization. The scalability of our algorithm enables its use in large-scale deep networks for MTL. Assuming only differentiability of the task-specific loss functions, we provide theoretical guarantees for convergence. Our experiments show that our algorithm outperforms the best competing methods on benchmark datasets.

How recurrent networks implement contextual processing in sentiment analysis

- Niru Maheswaranathan, David Sussillo
- abstract: Neural networks have a remarkable capacity for contextual processing{—}using recent or nearby inputs to modify processing of current input. For example, in natural language, contextual processing is necessary to correctly interpret negation (e.g. phrases such as "not bad"). However, our ability to understand how networks process context is limited. Here, we propose general methods for reverse engineering recurrent neural networks (RNNs) to identify and elucidate contextual processing. We apply these methods to understand RNNs trained on sentiment classification. This analysis reveals inputs that induce contextual effects, quantifies the strength and timescale of these effects, and identifies sets of these inputs with similar properties. Additionally,

we analyze contextual effects related to differential processing of the beginning and end of documents. Using the insights learned from the RNNs we improve baseline Bag-of-Words models with simple extensions that incorporate contextual modification, recovering greater than 90% of the RNN's performance increase over the baseline. This work yields a new understanding of how RNNs process contextual information, and provides tools that should provide similar insight more broadly.

Anderson Acceleration of Proximal Gradient Methods

- Vien Mai, Mikael Johansson
- abstract: Anderson acceleration is a well-established and simple technique for speeding up fixed-point computations with countless applications. This work introduces novel methods for adapting Anderson acceleration to proximal gradient algorithms. Under some technical conditions, we extend existing local convergence results of Anderson acceleration for smooth fixed-point mappings to the proposed non-smooth setting. We also prove analytically that it is in general, impossible to guarantee global convergence of native Anderson acceleration. We therefore propose a simple scheme for stabilization that combines the global worst-case guarantees of proximal gradient methods with the local adaptation and practical speed-up of Anderson acceleration. Finally, we provide the first applications of Anderson acceleration to non-Euclidean geometry.

Convergence of a Stochastic Gradient Method with Momentum for Non-Smooth Non-Convex Optimization

- Vien Mai, Mikael Johansson
- abstract: Stochastic gradient methods with momentum are widely used in applications and at the core of optimization subroutines in many popular machine learning libraries. However, their sample complexities have not been obtained for problems beyond those that are convex or smooth. This paper establishes the convergence rate of a stochastic subgradient method with a momentum term of Polyak type for a broad class of non-smooth, non-convex, and constrained optimization problems. Our key innovation is the construction of a special Lyapunov function for which the proven complexity can be achieved without any tuning of the momentum parameter. For smooth problems, we extend the known complexity bound to the constrained case and demonstrate how the unconstrained case can be analyzed under weaker assumptions than the state-of-the-art. Numerical results confirm our theoretical developments.

Adversarial Robustness Against the Union of Multiple Perturbation Models

- Pratyush Maini, Eric Wong, Zico Kolter
- abstract: Owing to the susceptibility of deep learning systems to adversarial attacks, there has been a great deal of work in developing (both empirically and certifiably) robust classifiers. While most work has defended against a single type of attack, recent work has looked at defending against multiple perturbation models using simple aggregations of multiple attacks. However, these methods can be difficult to tune, and can easily result in imbalanced degrees of robustness to individual perturbation models, resulting in a sub-optimal worst-case loss over the union. In this work, we develop a natural generalization of the standard PGD-based procedure to incorporate multiple perturbation models into a single attack, by taking the worst-case over all steepest descent directions. This approach has the advantage of directly converging upon a trade-off between different perturbation models which minimizes the worst-case performance over the union. With this approach, we are able to train standard architectures which are simultaneously robust against ℓ_∞, ℓ_2, and ℓ_1 attacks, outperforming past approaches on the MNIST and CIFAR10 datasets and achieving adversarial accuracy of 47.0% against the union of (ℓ_∞, ℓ_2, ℓ_1) perturbations with radius = (0.03, 0.5, 12) on the latter, improving upon previous approaches which achieve 40.6% accuracy.

Evolutionary Reinforcement Learning for Sample-Efficient Multiagent Coordination

- Somdeb Majumdar, Shauharda Khadka, Santiago Miret, Stephen Mcaleer, Kagan Tumer
- abstract: Many cooperative multiagent reinforcement learning environments provide agents with a sparse team-based reward, as well as a dense agent-specific reward that incentivizes learning basic skills. Training policies solely on the team-based reward is often difficult due to its sparsity. Also, relying solely on the agent-specific reward is sub-optimal because it usually does not capture the team coordination objective. A common approach is to use reward shaping to construct a proxy reward by combining the individual rewards. However, this requires manual tuning for each environment. We introduce Multiagent Evolutionary Reinforcement Learning (MERL), a split-level training platform that handles the two objectives separately through two optimization processes. An evolutionary algorithm maximizes the sparse team-based objective through neuroevolution on a population of teams. Concurrently, a gradient-based optimizer trains policies to only maximize the dense agent-specific rewards. The gradient-based policies are periodically added to the evolutionary population as a way of information transfer between the two optimization processes. This enables the evolutionary algorithm to use skills learned via the agent-specific rewards toward optimizing the global objective. Results demonstrate that MERL significantly outperforms state-of-the-art methods, such as MADDPG, on a number of difficult coordination benchmarks.

Estimation of Bounds on Potential Outcomes For Decision Making

- Maggie Makar, Fredrik Johansson, John Guttag, David Sontag
- abstract: Estimation of individual treatment effects is commonly used as the basis for contextual decision making in fields such as healthcare, education, and economics. However, it is often sufficient for the decision maker to have estimates of upper and lower bounds on the potential outcomes of decision alternatives to assess risks and benefits. We show that, in such cases, we can improve sample efficiency by estimating simple functions that bound these outcomes instead of estimating their conditional expectations, which may be complex and hard to estimate. Our analysis highlights a trade-off between the complexity of the learning task and the confidence with which the learned bounds hold. Guided by these findings, we develop an algorithm for learning upper and lower bounds on potential outcomes which optimize an objective function defined by the decision maker, subject to the probability that bounds are violated being small. Using a clinical dataset and a well-known causality benchmark, we demonstrate that our algorithm outperforms baselines, providing tighter, more reliable bounds.

Optimal transport mapping via input convex neural networks

- Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, Jason Lee
- abstract: In this paper, we present a novel and principled approach to learn the optimal transport between two distributions, from samples. Guided by the optimal transport theory, we learn the optimal Kantorovich potential which induces the optimal transport map. This involves learning two convex functions, by solving a novel minimax optimization. Building upon recent advances in the field of input convex neural networks, we propose a new framework to estimate the optimal transport mapping as the gradient of a convex function that is trained via minimax optimization. Numerical experiments confirm the accuracy of the learned transport map. Our approach can be readily used to train a deep generative model. When trained between a simple distribution in the latent space and a target distribution, the learned optimal transport map acts as a deep generative model. Although scaling this to a large dataset is challenging, we demonstrate two important strengths over standard adversarial training: robustness and discontinuity. As we seek the optimal transport, the learned generative model provides the same mapping regardless of how we initialize the neural networks. Further, a gradient of a neural network can easily represent discontinuous mappings, unlike standard neural networks that are constrained to be continuous. This allows the learned transport map to match any target distribution with many discontinuous supports and achieve sharp boundaries.

Proving the Lottery Ticket Hypothesis: Pruning is All You Need

- Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, Ohad Shamir
- abstract: The lottery ticket hypothesis (Frankle and Carbin, 2018), states that a randomly-initialized network contains a small subnetwork such that, when trained in isolation, can compete with the performance of the original network. We prove an even stronger hypothesis (as was also conjectured in Ramanujan et al., 2019), showing that for every bounded distribution and every target network with bounded weights, a sufficiently over-parameterized neural network with random weights contains a subnetwork with roughly the same accuracy as the target network, without any further training.

From Local SGD to Local Fixed-Point Methods for Federated Learning

- Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, Peter Richtarik
- abstract: Most algorithms for solving optimization problems or finding saddle points of convex-concave functions are fixed-point algorithms. In this work we consider the generic problem of finding a fixed point of an average of operators, or an approximation thereof, in a distributed setting. Our work is motivated by the needs of federated learning. In this context, each local operator models the computations done locally on a mobile device. We investigate two strategies to achieve such a consensus: one based on a fixed number of local steps, and the other based on randomized computations. In both cases, the goal is to limit communication of the locally-computed variables, which is often the bottleneck in distributed frameworks. We perform convergence analysis of both methods and conduct a number of experiments highlighting the benefits of our approach.

Adaptive Gradient Descent without Descent

- Yura Malitsky, Konstantin Mishchenko
- abstract: We present a strikingly simple proof that two rules are sufficient to automate gradient descent: 1) don’t increase the stepsize too fast and 2) don’t overstep the local curvature. No need for functional values, no line search, no information about the function except for the gradients. By following these rules, you get a method adaptive to the local geometry, with convergence guarantees depending only on the smoothness in a neighborhood of a solution. Given that the problem is convex, our method converges even if the global smoothness constant is infinity. As an illustration, it can minimize arbitrary continuously twice-differentiable convex function. We examine its performance on a range of convex and nonconvex problems, including logistic regression and matrix factorization.

Emergence of Separable Manifolds in Deep Language Representations

- Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, Sueyeon Chung
- abstract: Deep neural networks (DNNs) have shown much empirical success in solving perceptual tasks across various cognitive modalities. While they are only loosely inspired by the biological brain, recent studies report considerable similarities between representations extracted from task-optimized DNNs and neural populations in the brain. DNNs have subsequently become a popular model class to infer computational principles underlying complex cognitive functions, and in turn, they have also emerged as a natural testbed for applying methods originally developed to probe information in neural populations. In this work, we utilize mean-field theoretic manifold analysis, a recent technique from computational neuroscience that connects geometry of feature representations with linear separability of classes, to analyze language representations from large-scale contextual embedding models. We explore representations from different model families (BERT, RoBERTa, GPT, etc.) and find evidence for emergence of linguistic manifolds across layer depth (e.g., manifolds for part-of-speech tags), especially in ambiguous data (i.e, words with multiple part-of-speech tags, or part-of-speech classes including many words). In addition, we find that the emergence of linear separability in these manifolds is driven by a combined reduction of manifolds’ radius, dimensionality and inter-manifold correlations.

Adaptive Adversarial Multi-task Representation Learning

- Yuren Mao, Weiwei Liu, Xuemin Lin
- abstract: Adversarial Multi-task Representation Learning (AMTRL) methods are able to boost the performance of Multi-task Representation Learning (MTRL) models. However, the theoretical mechanism behind AMTRL is less investigated. To fill this gap, we study the generalization error bound of AMTRL through the lens of Lagrangian duality . Based on the duality, we proposed an novel adaptive AMTRL algorithm which improves the performance of original AMTRL methods. The extensive experiments back up our theoretical analysis and validate the superiority of our proposed algorithm.

On Learning Sets of Symmetric Elements

- Haggai Maron, Or Litany, Gal Chechik, Ethan Fetaya
- abstract: Learning from unordered sets is a fundamental learning setup, recently attracting increasing attention. Research in this area has focused on the case where elements of the set are represented by feature vectors, and far less emphasis has been given to the common case where set elements themselves adhere to their own symmetries. That case is relevant to numerous applications, from deblurring image bursts to multi-view 3D shape recognition and reconstruction. In this paper, we present a principled approach to learning sets of general symmetric elements. We first characterize the space of linear layers that are equivariant both to element reordering and to the inherent symmetries of elements, like translation in the case of images. We further show that networks that are composed of these layers, called Deep Sets for Symmetric Elements layers (DSS), are universal approximators of both invariant and equivariant functions. DSS layers are also straightforward to implement. Finally, we show that they improve over existing set-learning architectures in a series of experiments with images, graphs, and point-clouds.

Stochastically Dominant Distributional Reinforcement Learning

- John Martin, Michal Lyskawinski, Xiaohu Li, Brendan Englot
- abstract: We describe a new approach for managing aleatoric uncertainty in the Reinforcement Learning (RL) paradigm. Instead of selecting actions according to a single statistic, we propose a distributional method based on the second-order stochastic dominance (SSD) relation. This compares the inherent dispersion of random returns induced by actions, producing a comprehensive evaluation of the environment’s uncertainty. The necessary conditions for SSD require estimators to predict accurate second moments. To accommodate this, we map the distributional RL problem to a Wasserstein gradient flow, treating the distributional Bellman residual as a potential energy functional. We propose a particle-based algorithm for which we prove optimality and convergence. Our experiments characterize the algorithm’s performance and demonstrate how uncertainty and performance are better balanced using an SSD policy than with other risk measures.

Minimax Pareto Fairness: A Multi Objective Perspective

- Natalia Martinez, Martin Bertran, Guillermo Sapiro
- abstract: In this work we formulate and formally characterize group fairness as a multi-objective optimization problem, where each sensitive group risk is a separate objective. We propose a fairness criterion where a classifier achieves minimax risk and is Pareto-efficient w.r.t. all groups, avoiding unnecessary harm, and can lead to the best zero-gap model if policy dictates so. We provide a simple optimization algorithm compatible with deep neural networks to satisfy these constraints. Since our method does not require test-time access to sensitive attributes, it can be applied to reduce worst-case classification errors between outcomes in unbalanced classification problems. We test the proposed methodology on real case-studies of predicting income, ICU patient mortality, skin lesions classification, and assessing credit risk, demonstrating how our framework compares favorably to other approaches.

Predictive Multiplicity in Classification

- Charles Marx, Flavio Calmon, Berk Ustun
- abstract: Prediction problems often admit competing models that perform almost equally well. This effect challenges key assumptions in machine learning when competing models assign conflicting predictions. In this paper, we define predictive multiplicity as the ability of a prediction problem to admit competing models with conflicting predictions. We introduce measures to evaluate the severity of predictive multiplicity, and develop integer programming tools to compute these measures exactly for linear classification problems. We apply our tools to measure predictive multiplicity in recidivism prediction problems. Our results show that real-world datasets may admit competing models that assign wildly conflicting predictions, and motivate the need to report predictive multiplicity in model development.

Adding seemingly uninformative labels helps in low data regimes

- Christos Matsoukas, Albert Bou Hernandez, Yue Liu, Karin Dembrower, Gisele Miranda, Emir Konuk, Johan Fredin Haslum, Athanasios Zouzos, Peter Lindholm, Fredrik Strand, Kevin Smith
- abstract: Evidence suggests that networks trained on large datasets generalize well not solely because of the numerous training examples, but also class diversity which encourages learning of enriched features. This raises the question of whether this remains true when data is scarce - is there an advantage to learning with additional labels in low-data regimes? In this work, we consider a task that requires difficult-to-obtain expert annotations: tumor segmentation in mammography images. We show that, in low-data settings, performance can be improved by complementing the expert annotations with seemingly uninformative labels from non-expert annotators, turning the task into a multi-class problem. We reveal that these gains increase when less expert data is available, and uncover several interesting properties through further studies. We demonstrate our findings on CSAW-S, a new dataset that we introduce here, and confirm them on two public datasets.

Fast and Consistent Learning of Hidden Markov Models by Incorporating Non-Consecutive Correlations

- Robert Mattila, Cristian Rojas, Eric Moulines, Vikram Krishnamurthy, Bo Wahlberg
- abstract: Can the parameters of a hidden Markov model (HMM) be estimated from a single sweep through the observations – and additionally, without being trapped at a local optimum in the likelihood surface? That is the premise of recent method of moments algorithms devised for HMMs. In these, correlations between consecutive pair- or triplet-wise observations are empirically estimated and used to compute estimates of the HMM parameters. Albeit computationally very attractive, the main drawback is that by restricting to only low-order correlations in the data, information is being neglected which results in a loss of accuracy (compared to standard maximum likelihood schemes). In this paper, we propose extending these methods (both pair- and triplet-based) by also including non-consecutive correlations in a way which does not significantly increase the computational cost (which scales linearly with the number of additional lags included). We prove strong consistency of the new methods, and demonstrate an improved performance in numerical experiments on both synthetic and real-world financial time-series datasets.

On Approximate Thompson Sampling with Langevin Algorithms

- Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, Peter Bartlett
- abstract: Thompson sampling for multi-armed bandit problems is known to enjoy favorable performance in both theory and practice. However, its wider deployment is restricted due to a significant computational limitation: the need for samples from posterior distributions at every iteration. In practice, this limitation is alleviated by making use of approximate sampling methods, yet provably incorporating approximate samples into Thompson Sampling algorithms remains an open problem. In this work we address this by proposing two efficient Langevin MCMC algorithms tailored to Thompson sampling. The resulting approximate Thompson Sampling algorithms are efficiently implementable and provably achieve optimal instance-dependent regret for the Multi-Armed Bandit (MAB) problem. To prove these results we derive novel posterior concentration bounds and MCMC convergence rates for log-concave distributions which may be of independent interest.

Neural Datalog Through Time: Informed Temporal Modeling via Logical Specification

- Hongyuan Mei, Guanghui Qin, Minjie Xu, Jason Eisner
- abstract: Learning how to predict future events from patterns of past events is difficult when the set of possible event types is large. Training an unrestricted neural model might overfit to spurious patterns. To exploit domain-specific knowledge of how past events might affect an event's present probability, we propose using a temporal deductive database to track structured facts over time. Rules serve to prove facts from other facts and from past events. Each fact has a time-varying state—a vector computed by a neural net whose topology is determined by the fact's provenance, including its experience of past events. The possible event types at any time are given by special facts, whose probabilities are neurally modeled alongside their states. In both synthetic and real-world domains, we show that neural probabilistic models derived from concise Datalog programs improve prediction by encoding appropriate domain knowledge in their architecture.

On the Global Convergence Rates of Softmax Policy Gradient Methods

- Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, Dale Schuurmans
- abstract: We make three contributions toward better understanding policy gradient methods in the tabular setting. First, we show that with the true gradient, policy gradient with a softmax parametrization converges at a $O(1/t)$ rate, with constants depending on the problem and initialization. This result significantly expands the recent asymptotic convergence results. The analysis relies on two findings: that the softmax policy gradient satisfies a Ł{}ojasiewicz inequality, and the minimum probability of an optimal action during optimization can be bounded in terms of its initial value. Second, we analyze entropy regularized policy gradient and show that it enjoys a significantly faster linear convergence rate $O(e^{-t})$ toward softmax optimal policy. This result resolves an open question in the recent literature. Finally, combining the above two results and additional new $\Omega(1/t)$ lower bound results, we explain how entropy regularization improves policy optimization, even with the true gradient, from the perspective of convergence rate. The separation of rates is further explained using the notion of non-uniform Ł{}ojasiewicz degree. These results provide a theoretical understanding of the impact of entropy and corroborate existing empirical studies.

Scalable Identification of Partially Observed Systems with Certainty-Equivalent EM

- Kunal Menda, Jean De Becdelievre, Jayesh Gupta, Ilan Kroo, Mykel Kochenderfer, Zachary Manchester
- abstract: System identification is a key step for model-based control, estimator design, and output prediction. This work considers the offline identification of partially observed nonlinear systems. We empirically show that the certainty-equivalent approximation to expectation-maximization can be a reliable and scalable approach for high-dimensional deterministic systems, which are common in robotics. We formulate certainty-equivalent expectation-maximization as block coordinate-ascent, and provide an efficient implementation. The algorithm is tested on a simulated system of coupled Lorenz attractors, demonstrating its ability to identify high-dimensional systems that can be intractable for particle-based approaches. Our approach is also used to identify the dynamics of an aerobatic helicopter. By augmenting the state with unobserved fluid states, a model is learned that predicts the acceleration of the helicopter better than state-of-the-art approaches. The codebase for this work is available at https://github.com/sisl/CEEM.

Randomized Block-Diagonal Preconditioning for Parallel Learning

- Celestine Mendler-Dünner, Aurelien Lucchi
- abstract: We study preconditioned gradient-based optimization methods where the preconditioning matrix has block-diagonal form. Such a structural constraint comes with the advantage that the update computation can be parallelized across multiple independent tasks. Our main contribution is to demonstrate that the convergence of these methods can significantly be improved by a randomization technique which corresponds to repartitioning coordinates across tasks during the optimization procedure. We provide a theoretical analysis that accurately characterizes the expected convergence gains of repartitioning and validate our findings empirically on various traditional machine learning tasks. From an implementation perspective, block-separable models are well suited for parallelization and, when shared memory is available, randomization can be implemented on top of existing methods very efficiently to improve convergence.

Training Binary Neural Networks using the Bayesian Learning Rule

- Xiangming Meng, Roman Bachmann, Mohammad Emtiyaz Khan
- abstract: Neural networks with binary weights are computation-efficient and hardware-friendly, but their training is challenging because it involves a discrete optimization problem. Surprisingly, ignoring the discrete nature of the problem and using gradient-based methods, such as the Straight-Through Estimator, still works well in practice. This raises the question: are there principled approaches which justify such methods? In this paper, we propose such an approach using the Bayesian learning rule. The rule, when applied to estimate a Bernoulli distribution over the binary weights, results in an algorithm which justifies some of the algorithmic choices made by the previous approaches. The algorithm not only obtains state-of-the-art performance, but also enables uncertainty estimation and continual learning to avoid catastrophic forgetting. Our work provides a principled approach for training binary neural networks which also justifies and extends existing approaches.

Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning

- Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, Marcello Restelli
- abstract: The choice of the control frequency of a system has a relevant impact on the ability of reinforcement learning algorithms to learn a highly performing policy. In this paper, we introduce the notion of action persistence that consists in the repetition of an action for a fixed number of decision steps, having the effect of modifying the control frequency. We start analyzing how action persistence affects the performance of the optimal policy, and then we present a novel algorithm, Persistent Fitted Q-Iteration (PFQI), that extends FQI, with the goal of learning the optimal value function at a given persistence. After having provided a theoretical study of PFQI and a heuristic approach to identify the optimal persistence, we present an experimental campaign on benchmark domains to show the advantages of action persistence and proving the effectiveness of our persistence selection method.

The Role of Regularization in Classification of High-dimensional Noisy Gaussian Mixture

- Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, Lenka Zdeborova
- abstract: We consider a high-dimensional mixture of two Gaussians in the noisy regime where even an oracle knowing the centers of the clusters misclassifies a small but finite fraction of the points. We provide a rigorous analysis of the generalization error of regularized convex classifiers, including ridge, hinge and logistic regression, in the high-dimensional limit where the number n of samples and their dimension d go to infinity while their ratio is fixed to $\alpha=n/d$. We discuss surprising effects of the regularization that in some cases allows to reach the Bayes-optimal performances. We also illustrate the interpolation peak at low regularization, and analyze the role of the respective sizes of the two clusters.

Projective Preferential Bayesian Optimization

- Petrus Mikkola, Milica Todorović, Jari Järvi, Patrick Rinke, Samuel Kaski
- abstract: Bayesian optimization is an effective method for finding extrema of a black-box function. We propose a new type of Bayesian optimization for learning user preferences in high-dimensional spaces. The central assumption is that the underlying objective function cannot be evaluated directly, but instead a minimizer along a projection can be queried, which we call a projective preferential query. The form of the query allows for feedback that is natural for a human to give, and which enables interaction. This is demonstrated in a user experiment in which the user feedback comes in the form of optimal position and orientation of a molecule adsorbing to a surface. We demonstrate that our framework is able to find a global minimum of a high-dimensional black-box function, which is an infeasible task for existing preferential Bayesian optimization frameworks that are based on pairwise comparisons.

VideoOneNet: Bidirectional Convolutional Recurrent OneNet with Trainable Data Steps for Video Processing

- Zoltán Milacski, Barnabas Poczos, Andras Lorincz
- abstract: Deep Neural Networks (DNNs) achieve the state-of-the-art results on a wide range of image processing tasks, however, the majority of such solutions are problem-specific, like most AI algorithms. The One Network to Solve Them All (OneNet) procedure has been suggested to resolve this issue by exploiting a DNN as the proximal operator in Alternating Direction Method of Multipliers (ADMM) solvers for various imaging problems. In this work, we make two contributions, both facilitating end-to-end learning using backpropagation. First, we generalize OneNet to videos by augmenting its convolutional prior network with bidirectional recurrent connections; second, we extend the fixed fully connected linear ADMM data step with another trainable bidirectional convolutional recurrent network. In our computational experiments on the Rotated MNIST, Scanned CIFAR-10 and UCF-101 data sets, the proposed modifications improve performance by a large margin compared to end-to-end convolutional OneNet and 3D Wavelet sparsity on several video processing problems: pixelwise inpainting-denoising, blockwise inpainting, scattered inpainting, super resolution, compressive sensing, deblurring, frame interpolation, frame prediction and colorization. Our two contributions are complementary, and using them together yields the best results.

The Effect of Natural Distribution Shift on Question Answering Models

- John Miller, Karl Krauth, Benjamin Recht, Ludwig Schmidt
- abstract: We build four new test sets for the Stanford Question Answering Dataset (SQuAD) and evaluate the ability of question-answering systems to generalize to new data. Our first test set is from the original Wikipedia domain and measures the extent to which existing systems overfit the original test set. Despite several years of heavy test set re-use, we find no evidence of adaptive overfitting. The remaining three test sets are constructed from New York Times articles, Reddit posts, and Amazon product reviews and measure robustness to natural distribution shifts. Across a broad range of models, we observe average performance drops of 3.8, 14.0, and 17.4 F1 points, respectively. In contrast, a strong human baseline matches or exceeds the performance of SQuAD models on the original domain and exhibits little to no drop in new domains. Taken together, our results confirm the surprising resilience of the holdout method and emphasize the need to move towards evaluation metrics that incorporate robustness to natural distribution shifts.

Strategic Classification is Causal Modeling in Disguise

- John Miller, Smitha Milli, Moritz Hardt
- abstract: Consequential decision-making incentivizes individuals to strategically adapt their behavior to the specifics of the decision rule. While a long line of work has viewed strategic adaptation as gaming and attempted to mitigate its effects, recent work has instead sought to design classifiers that incentivize individuals to improve a desired quality. Key to both accounts is a cost function that dictates which adaptations are rational to undertake. In

this work, we develop a causal framework for strategic adaptation. Our causal perspective clearly distinguishes between gaming and improvement and reveals an important obstacle to incentive design. We prove any procedure for designing classifiers that incentivize improvement must inevitably solve a non-trivial causal inference problem. We show a similar result holds for designing cost functions that satisfy the requirements of previous work. With the benefit of hindsight, our results show much of the prior work on strategic classification is causal modeling in disguise.

Automatic Shortcut Removal for Self-Supervised Representation Learning

- Matthias Minderer, Olivier Bachem, Neil Houlsby, Michael Tschannen
- abstract: In self-supervised visual representation learning, a feature extractor is trained on a "pretext task" for which labels can be generated cheaply, without human annotation. A central challenge in this approach is that the feature extractor quickly learns to exploit low-level visual features such as color aberrations or watermarks and then fails to learn useful semantic representations. Much work has gone into identifying such "shortcut" features and hand-designing schemes to reduce their effect. Here, we propose a general framework for mitigating the effect shortcut features. Our key assumption is that those features which are the first to be exploited for solving the pretext task may also be the most vulnerable to an adversary trained to make the task harder. We show that this assumption holds across common pretext tasks and datasets by training a "lens" network to make small image changes that maximally reduce performance in the pretext task. Representations learned with the modified images outperform those learned without in all tested cases. Additionally, the modifications made by the lens reveal how the choice of pretext task and dataset affects the features learned by self-supervision.

Learning Reasoning Strategies in End-to-End Differentiable Proving

- Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, Tim Rocktäschel
- abstract: Attempts to render deep learning models interpretable, data-efficient, and robust have seen some success through hybridisation with rule-based systems, for example, in Neural Theorem Provers (NTPs). These neuro-symbolic models can induce interpretable rules and learn representations from data via back-propagation, while providing logical explanations for their predictions. However, they are restricted by their computational complexity, as they need to consider all possible proof paths for explaining a goal, thus rendering them unfit for large-scale applications. We present Conditional Theorem Provers (CTPs), an extension to NTPs that learns an optimal rule selection strategy via gradient-based optimisation. We show that CTPs are scalable and yield state-of-the-art results on the CLUTRR dataset, which tests systematic generalisation of neural models by learning to reason over smaller graphs and evaluating on larger ones. Finally, CTPs show better link prediction results on standard benchmarks in comparison with other neural-symbolic models, while being explainable.

Coresets for Data-efficient Training of Machine Learning Models

- Baharan Mirzasoleiman, Jeff Bilmes, Jure Leskovec
- abstract: Incremental gradient (IG) methods, such as stochastic gradient descent and its variants are commonly used for large scale optimization in machine learning. Despite the sustained effort to make IG methods more data-efficient, it remains an open question how to select a training data subset that can theoretically and practically perform on par with the full dataset. Here we develop CRAIG, a method to select a weighted subset (or coreset) of training data that closely estimates the full gradient by maximizing a submodular function. We prove that applying IG to this subset is guaranteed to converge to the (near)optimal solution with the same convergence rate as that of IG for convex optimization. As a result, CRAIG achieves a speedup that is inversely proportional to the size of the subset. To our knowledge, this is the first rigorous method for data-efficient training of general machine learning models. Our extensive set of experiments show that CRAIG, while achieving practically the same solution, speeds up various IG methods by up to 6x for logistic regression and 3x for training deep neural networks.

Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning

- Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, John Langford
- abstract: We present an algorithm, HOMER, for exploration and reinforcement learning in rich observation environments that are summarizable by an unknown latent state space. The algorithm interleaves representation learning to identify a new notion of kinematic state abstraction with strategic exploration to reach new states using the learned abstraction. The algorithm provably explores the environment with sample complexity scaling polynomially in the number of latent states and the time horizon, and, crucially, with no dependence on the size of the observation space, which could be infinitely large. This exploration guarantee further enables sample-efficient global policy optimization for any reward function. On the computational side, we show that the algorithm can be implemented efficiently whenever certain supervised learning problems are tractable. Empirically, we evaluate HOMER on a challenging exploration problem, where we show that the algorithm is more sample efficient than standard reinforcement learning baselines.

Learning to Combine Top-Down and Bottom-Up Signals in Recurrent Neural Networks with Attention over Modules

- Sarthak Mittal, Alex Lamb, Anirudh Goyal, Vikram Voleti, Murray Shanahan, Guillaume Lajoie, Michael Mozer, Yoshua Bengio
- abstract: Robust perception relies on both bottom-up and top-down signals. Bottom-up signals consist of what’s directly observed through sensation. Top-down signals consist of beliefs and expectations based on past experience and the current reportable short-term memory, such as how the phrase ‘peanut butter and ...’ will be completed. The optimal combination of bottom-up and top-down information remains an open question, but the manner of combination must be dynamic and both context and task dependent. To effectively utilize the wealth of potential top-down information available, and to prevent the cacophony of intermixed signals in a bidirectional architecture, mechanisms are needed to restrict information flow. We explore deep recurrent neural net architectures in which bottom-up and top-down signals are dynamically combined using attention. Modularity of the architecture further restricts the sharing and communication of information. Together, attention and modularity direct information flow, which leads to reliable performance improvements in perceptual and language tasks, and in particular improves robustness to distractions and noisy data. We demonstrate on a variety of benchmarks in language modeling, sequential image classification, video prediction and reinforcement learning that the \emph{bidirectional} information flow can improve results over strong baselines.

Optimizing Long-term Social Welfare in Recommender Systems: A Constrained Matching Approach

- Martin Mladenov, Elliot Creager, Omer Ben-Porat, Kevin Swersky, Richard Zemel, Craig Boutilier
- abstract: Most recommender systems (RS) research assumes that a user’s utility can be maximized independently of the utility of the other agents (e.g., other users, content providers). In realistic settings, this is often not true – the dynamics of an RS ecosystem couple the long-term utility of all agents. In this work, we explore settings in which content providers cannot remain viable unless they receive a certain level of user engagement. We formulate this problem as one of equilibrium selection in the induced dynamical system, and show that it can be solved as an optimal constrained matching problem. Our model ensures the system reaches an equilibrium with maximal social welfare supported by a sufficiently diverse set of viable providers. We demonstrate that even in a simple, stylized dynamical RS model, the standard myopic approach to recommendation - always matching a user to the best provider - performs poorly. We develop several scalable techniques to solve the matching problem, and also draw connections to various notions of user regret and fairness, arguing that these outcomes are fairer in a utilitarian sense.

Transformation of ReLU-based recurrent neural networks from discrete-time to continuous-time

- Zahra Monfared, Daniel Durstewitz

- abstract: Recurrent neural networks (RNN) as used in machine learning are commonly formulated in discrete time, i.e. as recursive maps. This brings a lot of advantages for training models on data, e.g. for the purpose of time series prediction or dynamical systems identification, as powerful and efficient inference algorithms exist for discrete time systems and numerical integration of differential equations is not necessary. On the other hand, mathematical analysis of dynamical systems inferred from data is often more convenient and enables additional insights if these are formulated in continuous time, i.e. as systems of ordinary or partial differential equations (ODE/ PDE). Here we show how to perform such a translation from discrete to continuous time for a particular class of ReLU-based RNN. We prove three theorems on the mathematical equivalence between the discrete and continuous time formulations under a variety of conditions, and illustrate how to use our mathematical results on different machine learning and nonlinear dynamical systems examples.

Efficiently Learning Adversarially Robust Halfspaces with Noise

- Omar Montasser, Surbhi Goel, Ilias Diakonikolas, Nathan Srebro
- abstract: We study the problem of learning adversarially robust halfspaces in the distribution-independent setting. In the realizable setting, we provide necessary and sufficient conditions on the adversarial perturbation sets under which halfspaces are efficiently robustly learnable. In the presence of random label noise, we give a simple computationally efficient algorithm for this problem with respect to any ℓ_p-perturbation.

An end-to-end approach for the verification problem: learning the right distance

- Joao Monteiro, Isabela Albuquerque, Jahangir Alam, R Devon Hjelm, Tiago Falk
- abstract: In this contribution, we augment the metric learning setting by introducing a parametric pseudo-distance, trained jointly with the encoder. Several interpretations are thus drawn for the learned distance-like model's output. We first show it approximates a likelihood ratio which can be used for hypothesis tests, and that it further induces a large divergence across the joint distributions of pairs of examples from the same and from different classes. Evaluation is performed under the verification setting consisting of determining whether sets of examples belong to the same class, even if such classes are novel and were never presented to the model during training. Empirical evaluation shows such method defines an end-to-end approach for the verification problem, able to attain better performance than simple scorers such as those based on cosine similarity and further outperforming widely used downstream classifiers. We further observe training is much simplified under the proposed approach compared to metric learning with actual distances, requiring no complex scheme to harvest pairs of examples.

Confidence-Aware Learning for Deep Neural Networks

- Jooyoung Moon, Jihyo Kim, Younghak Shin, Sangheum Hwang
- abstract: Despite the power of deep neural networks for a wide range of tasks, an overconfident prediction issue has limited their practical use in many safety-critical applications. Many recent works have been proposed to mitigate this issue, but most of them require either additional computational costs in training and/or inference phases or customized architectures to output confidence estimates separately. In this paper, we propose a method of training deep neural networks with a novel loss function, named Correctness Ranking Loss, which regularizes class probabilities explicitly to be better confidence estimates in terms of ordinal ranking according to confidence. The proposed method is easy to implement and can be applied to the existing architectures without any modification. Also, it has almost the same computational costs for training as conventional deep classifiers and outputs reliable predictions by a single inference. Extensive experimental results on classification benchmark datasets indicate that the proposed method helps networks to produce well-ranked confidence estimates. We also demonstrate that it is effective for the tasks closely related to confidence estimation, out-of-distribution detection and active learning.

Topological Autoencoders

- Michael Moor, Max Horn, Bastian Rieck, Karsten Borgwardt
- abstract: We propose a novel approach for preserving topological structures of the input space in latent representations of autoencoders. Using persistent homology, a technique from topological data analysis, we calculate topological signatures of both the input and latent space to derive a topological loss term. Under weak theoretical assumptions, we construct this loss in a differentiable manner, such that the encoding learns to retain multi-scale connectivity information. We show that our approach is theoretically well-founded and that it exhibits favourable latent representations on a synthetic manifold as well as on real-world image data sets, while preserving low reconstruction errors.

Explainable k-Means and k-Medians Clustering

- Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, Nave Frost
- abstract: Many clustering algorithms lead to cluster assignments that are hard to explain, partially because they depend on all the features of the data in a complicated way. To improve interpretability, we consider using a small decision tree to partition a data set into clusters, so that clusters can be characterized in a straightforward manner. We study this problem from a theoretical viewpoint, measuring cluster quality by the k-means and k-medians objectives. In terms of negative results, we show that popular top-down decision tree algorithms may lead to clusterings with arbitrarily large cost, and any clustering based on a tree with k leaves must incur an Omega(log k) approximation factor compared to the optimal clustering. On the positive side, for two means/medians, we show that a single threshold cut can achieve a constant factor approximation, and we give nearly-matching lower bounds; for general k > 2, we design an efficient algorithm that leads to an O(k) approximation to the optimal k-medians and an O(k^2) approximation to the optimal k-means. Prior to our work, no algorithms were known with provable guarantees independent of dimension and input size.

Fair Learning with Private Demographic Data

- Hussein Mozannar, Mesrob Ohannessian, Nathan Srebro
- abstract: Sensitive attributes such as race are rarely available to learners in real world settings as their collection is often restricted by laws and regulations. We give a scheme that allows individuals to release their sensitive information privately while still allowing any downstream entity to learn non-discriminatory predictors. We show how to adapt non-discriminatory learners to work with privatized protected attributes giving theoretical guarantees on performance. Finally, we highlight how the methodology could apply to learning fair predictors in settings where protected attributes are only available for a subset of the data.

Consistent Estimators for Learning to Defer to an Expert

- Hussein Mozannar, David Sontag
- abstract: Learning algorithms are often used in conjunction with expert decision makers in practical scenarios, however this fact is largely ignored when designing these algorithms. In this paper we explore how to learn predictors that can either predict or choose to defer the decision to a downstream expert. Given only samples of the expert's decisions, we give a procedure based on learning a classifier and a rejector and analyze it theoretically. Our approach is based on a novel reduction to cost sensitive learning where we give a consistent surrogate loss for cost sensitive learning that generalizes the cross entropy loss. We show the effectiveness of our approach on a variety of experimental tasks.

Continuous-time Lower Bounds for Gradient-based Algorithms

- Michael Muehlebach, Michael Jordan
- abstract: This article derives lower bounds on the convergence rate of continuous-time gradient-based optimization algorithms. The algorithms are subjected to a time-normalization constraint that avoids a reparametrization of time in order to make the discussion of continuous-time convergence rates meaningful. We reduce the multi-dimensional problem to a single dimension, recover well-known lower bounds from the discrete-time setting, and provide insight into why these lower bounds occur. We present algorithms that achieve the proposed lower bounds, even when the function class under consideration includes certain nonconvex functions.

Two Simple Ways to Learn Individual Fairness Metrics from Data

- Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, Yuekai Sun
- abstract: Individual fairness is an intuitive definition of algorithmic fairness that addresses some of the drawbacks of group fairness. Despite its benefits, it depends on a task specific fair metric that encodes our intuition of what is fair and unfair for the ML task at hand, and the lack of a widely accepted fair metric for many ML tasks is the main barrier to broader adoption of individual fairness. In this paper, we present two simple ways to learn fair metrics from a variety of data types. We show empirically that fair training with the learned metrics leads to improved fairness on three machine learning tasks susceptible to gender and racial biases. We also provide theoretical guarantees on the statistical performance of both approaches.

Unique Properties of Flat Minima in Deep Networks

- Rotem Mulayoff, Tomer Michaeli
- abstract: It is well known that (stochastic) gradient descent has an implicit bias towards flat minima. In deep neural network training, this mechanism serves to screen out minima. However, the precise effect that this has on the trained network is not yet fully understood. In this paper, we characterize the flat minima in linear neural networks trained with a quadratic loss. First, we show that linear ResNets with zero initialization necessarily converge to the flattest of all minima. We then prove that these minima correspond to nearly balanced networks whereby the gain from the input to any intermediate representation does not change drastically from one layer to the next. Finally, we show that consecutive layers in flat minima solutions are coupled. That is, one of the left singular vectors of each weight matrix, equals one of the right singular vectors of the next matrix. This forms a distinct path from input to output, that, as we show, is dedicated to the signal that experiences the largest gain end-to-end. Experiments indicate that these properties are characteristic of both linear and nonlinear models trained in practice.

Fast computation of Nash Equilibria in Imperfect Information Games

- Remi Munos, Julien Perolat, Jean-Baptiste Lespiau, Mark Rowland, Bart De Vylder, Marc Lanctot, Finbarr Timbers, Daniel Hennes, Shayegan Omidshafiei, Audrunas Gruslys, Mohammad Gheshlaghi Azar, Edward Lockhart, Karl Tuyls
- abstract: We introduce and analyze a class of algorithms, called Mirror Ascent against an Improved Opponent (MAIO), for computing Nash equilibria in two-player zero-sum games, both in normal form and in sequential form with imperfect information. These algorithms update the policy of each player with a mirror-ascent step to maximize the value of playing against an improved opponent. An improved opponent can be a best response, a greedy policy, a policy improved by policy gradient, or by any other reinforcement learning or search techniques. We establish a convergence result of the last iterate to the set of Nash equilibria and show that the speed of convergence depends on the amount of improvement offered by these improved policies. In addition, we show that under some condition, if we use a best response as improved policy, then an exponential convergence rate is achieved.

Missing Data Imputation using Optimal Transport

- Boris Muzellec, Julie Josse, Claire Boyer, Marco Cuturi
- abstract: Missing data is a crucial issue when applying machine learning algorithms to real-world datasets. Starting from the simple assumption that two batches extracted randomly from the same dataset should share the same distribution, we leverage optimal transport distances to quantify that criterion and turn it into a loss function to impute missing data values. We propose practical methods to minimize these losses using end-to-end learning, that can exploit or not parametric assumptions on the underlying distributions of values. We evaluate our methods on datasets from the UCI repository, in MCAR, MAR and MNAR settings. These experiments show that OT-based methods match or out-perform state-of-the-art imputation methods, even for high percentages of missing values.

Semiparametric Nonlinear Bipartite Graph Representation Learning with Provable Guarantees

- Sen Na, Yuwei Luo, Zhuoran Yang, Zhaoran Wang, Mladen Kolar
- abstract: Graph representation learning is a ubiquitous task in machine learning where the goal is to embed each vertex into a low-dimensional vector space. We consider the bipartite graph and formalize its representation learning problem as a statistical estimation problem of parameters in a semiparametric exponential family distribution: the bipartite graph is assumed to be generated by a semiparametric exponential family distribution, whose parametric component is given by the proximity of outputs of two one-layer neural networks that take high-dimensional features as inputs, while nonparametric (nuisance) component is the base measure. In this setting, the representation learning problem is equivalent to recovering the weight matrices, and the main challenges of estimation arise from the nonlinearity of activation functions and the nonparametric nuisance component of the distribution. To overcome these challenges, we propose a pseudo-likelihood objective based on the rank-order decomposition technique and show that the proposed objective is strongly convex in a neighborhood around the ground truth, so that a gradient descent-based method achieves linear convergence rate. Moreover, we prove that the sample complexity of the problem is linear in dimensions (up to logarithmic factors), which is consistent with parametric Gaussian models. However, our estimator is robust to any model misspecification within the exponential family, which is validated in extensive experiments.

Full Law Identification in Graphical Models of Missing Data: Completeness Results

- Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser
- abstract: Missing data has the potential to affect analyses conducted in all fields of scientific study including healthcare, economics, and the social sciences. Several approaches to unbiased inference in the presence of non-ignorable missingness rely on the specification of the target distribution and its missingness process as a probability distribution that factorizes with respect to a directed acyclic graph. In this paper, we address the longstanding question of the characterization of models that are identifiable within this class of missing data distributions. We provide the first completeness result in this field of study – necessary and sufficient graphical conditions under which, the full data distribution can be recovered from the observed data distribution. We then simultaneously address issues that may arise due to the presence of both missing data and unmeasured confounding, by extending these graphical conditions and proofs of completeness, to settings where some variables are not just missing, but completely unobserved.

Voice Separation with an Unknown Number of Multiple Speakers

- Eliya Nachmani, Yossi Adi, Lior Wolf
- abstract: We present a new method for separating a mixed audio sequence, in which multiple voices speak simultaneously. The new method employs gated neural networks that are trained to separate the voices at multiple processing steps, while maintaining the speaker in each output channel fixed. A different

model is trained for every number of possible speakers, and the model with the largest number of speakers is employed to select the actual number of speakers in a given sample. Our method greatly outperforms the current state of the art, which, as we show, is not competitive for more than two speakers.

Reliable Fidelity and Diversity Metrics for Generative Models

- Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, Jaejun Yoo
- abstract: Devising indicative evaluation metrics for the image generation task remains an open problem. The most widely used metric for measuring the similarity between real and generated images has been the Frechet Inception Distance (FID) score. Since it does not differentiate the fidelity and diversity aspects of the generated images, recent papers have introduced variants of precision and recall metrics to diagnose those properties separately. In this paper, we show that even the latest version of the precision and recall metrics are not reliable yet. For example, they fail to detect the match between two identical distributions, they are not robust against outliers, and the evaluation hyperparameters are selected arbitrarily. We propose density and coverage metrics that solve the above issues. We analytically and experimentally show that density and coverage provide more interpretable and reliable signals for practitioners than the existing metrics.

From Chaos to Order: Symmetry and Conservation Laws in Game Dynamics

- Sai Ganesh Nagarajan, David Balduzzi, Georgios Piliouras
- abstract: Games are an increasingly useful tool for training and testing learning algorithms. Recent examples include GANs, AlphaZero and the AlphaStar league. However, multi-agent learning can be extremely difficult to predict and control. Learning dynamics even in simple games can yield chaotic behavior. In this paper, we present basic \emph{mechanism design} tools for constructing games with predictable and controllable dynamics. We show that arbitrarily large and complex network games, encoding both cooperation (team play) and competition (zero-sum interaction), exhibit conservation laws when agents use the standard regret-minimizing dynamics known as Follow-the-Regularized-Leader. These laws persist when different agents use different dynamics and encode long-range correlations between agents' behavior, even though the agents may not interact directly. Moreover, we provide sufficient conditions under which the dynamics have multiple, linearly independent, conservation laws. Increasing the number of conservation laws results in more predictable dynamics, eventually making chaotic behavior formally impossible in some cases.

Up or Down? Adaptive Rounding for Post-Training Quantization

- Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, Tijmen Blankevoort
- abstract: When quantizing neural networks, assigning each floating-point weight to its nearest fixed-point value is the predominant approach. We find that, perhaps surprisingly, this is not the best we can do. In this paper, we propose AdaRound, a better weight-rounding mechanism for post-training quantization that adapts to the data and the task loss. AdaRound is fast, does not require fine-tuning of the network, and only uses a small amount of unlabelled data. We start by theoretically analyzing the rounding problem for a pre-trained neural network. By approximating the task loss with a Taylor series expansion, the rounding task is posed as a quadratic unconstrained binary optimization problem. We simplify this to a layer-wise local loss and propose to optimize this loss with a soft relaxation. AdaRound not only outperforms rounding-to-nearest by a significant margin but also establishes a new state-of-the-art for post-training quantization on several networks and tasks. Without fine-tuning, we can quantize the weights of Resnet18 and Resnet50 to 4 bits while staying within an accuracy loss of 1%.

Goal-Aware Prediction: Learning to Model What Matters

- Suraj Nair, Silvio Savarese, Chelsea Finn
- abstract: Learned dynamics models combined with both planning and policy learning algorithms have shown promise in enabling artificial agents to learn to perform many diverse tasks with limited supervision. However, one of the fundamental challenges in using a learned forward dynamics model is the mismatch between the objective of the learned model (future state reconstruction), and that of the downstream planner or policy (completing a specified task). This issue is exacerbated by vision-based control tasks in diverse real-world environments, where the complexity of the real world dwarfs model capacity. In this paper, we propose to direct prediction towards task relevant information, enabling the model to be aware of the current task and encouraging it to only model relevant quantities of the state space, resulting in a learning objective that more closely matches the downstream task. Further, we do so in an entirely self-supervised manner, without the need for a reward function or image labels. We find that our method more effectively models the relevant parts of the scene conditioned on the goal, and as a result outperforms standard task-agnostic dynamics models and model-free reinforcement learning.

PolyGen: An Autoregressive Generative Model of 3D Meshes

- Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, Peter Battaglia
- abstract: Polygon meshes are an efficient representation of 3D geometry, and are of central importance in computer graphics, robotics and games development. Existing learning-based approaches for object synthesis have avoided the challenges of working with 3D meshes, instead using alternative object representations that are more compatible with neural architectures and training approaches. We present PolyGen, a generative model of 3D objects which models the mesh directly, predicting vertices and faces sequentially using a Transformer-based architecture. Our model can condition on a range of inputs, including object classes, voxels, and images, and because the model is probabilistic it can produce samples that capture uncertainty in ambiguous scenarios. We show that the model is capable of producing high-quality, usable meshes, and establish log-likelihood benchmarks for the mesh-modelling task. We also evaluate the conditional models on surface reconstruction metrics against alternative methods, and demonstrate competitive performance despite not training directly on this task.

Bayesian Sparsification of Deep C-valued Networks

- Ivan Nazarov, Evgeny Burnaev
- abstract: With continual miniaturization ever more applications of deep learning can be found in embedded systems, where it is common to encounter data with natural representation in the complex domain. To this end we extend Sparse Variational Dropout to complex-valued neural networks and verify the proposed Bayesian technique by conducting a large numerical study of the performance-compression trade-off of C-valued networks on two tasks: image recognition on MNIST-like and CIFAR10 datasets and music transcription on MusicNet. We replicate the state-of-the-art result by Trabelsi et al. (2018) on MusicNet with a complex-valued network compressed by 50-100x at a small performance penalty.

Oracle Efficient Private Non-Convex Optimization

- Seth Neel, Aaron Roth, Giuseppe Vietri, Steven Wu
- abstract: One of the most effective algorithms for differentially private learning and optimization is \emph{objective perturbation}. This technique augments a given optimization problem (e.g. deriving from an ERM problem) with a random linear term, and then exactly solves it. However, to date, analyses of this approach crucially rely on the convexity and smoothness of the objective function. We give two algorithms that extend this approach substantially. The first algorithm requires nothing except boundedness of the loss function, and operates over a discrete domain. Its privacy and accuracy guarantees hold even without assuming convexity. We are able to extend traditional analyses of objective perturbation by introducing a novel "normalization" step into the algorithm, which provides enough stability to be differentially private even without second-order conditions. The second

algorithm operates over a continuous domain and requires only that the loss function be bounded and Lipschitz in its continuous parameter. Its privacy analysis does not even require convexity. Its accuracy analysis does require convexity, but does not require second order conditions like smoothness. We complement our theoretical results with an empirical evaluation of the non-convex case, in which we use an integer program solver as our optimization oracle. We find that for the problem of learning linear classifiers, directly optimizing for 0/1 loss using our approach can out-perform the more standard approach of privately optimizing a convex-surrogate loss function on the Adult dataset.

Stochastic Frank-Wolfe for Constrained Finite-Sum Minimization

- Geoffrey Negiar, Gideon Dresdner, Alicia Tsai, Laurent El Ghaoui, Francesco Locatello, Robert Freund, Fabian Pedregosa
- abstract: We propose a novel Stochastic Frank-Wolfe (a. k. a. conditional gradient) algorithm for constrained smooth finite-sum minimization with a generalized linear prediction/structure. This class of problems includes empirical risk minimization with sparse, low-rank, or other structured constraints. The proposed method is simple to implement, does not require step-size tuning, and has a constant per-iteration cost that is independent of the dataset size. Furthermore, as a byproduct of the method we obtain a stochastic estimator of the Frank-Wolfe gap that can be used as a stopping criterion. Depending on the setting, the proposed method matches or improves on the best computational guarantees for Stochastic Frank-Wolfe algorithms. Benchmarks on several datasets highlight different regimes in which the proposed method exhibits a faster empirical convergence than related methods. Finally, we provide an implementation of all considered methods in an open-source package.

In Defense of Uniform Convergence: Generalization via Derandomization with an Application to Interpolating Predictors

- Jeffrey Negrea, Gintare Karolina Dziugaite, Daniel Roy
- abstract: We propose to study the generalization error of a learned predictor in terms of that of a surrogate (potentially randomized) predictor that is coupled to $\hh$ and designed to trade empirical risk for control of generalization error. In the case where the learned predictor interpolates the data, it is interesting to consider theoretical surrogate classifiers that are partially derandomized or rerandomized, e.g., fit to the training data but with modified label noise. We also show that replacing the learned predictor by its conditional distribution with respect to an arbitrary σ-field is a convenient way to derandomize. We study two examples, inspired by the work of Nagarajan and Kolter (2019) and Bartlett et al. (2020), where the learned predictor interpolates the training data with high probability, has small risk, and, yet, does not belong to a nonrandom class with a tight uniform bound on two-sided generalization error. At the same time, we bound the risk of the learned predictor in terms of surrogates constructed by conditioning and denoising, respectively, and shown to belong to nonrandom classes with uniformly small generalization error.

Involutive MCMC: a Unifying Framework

- Kirill Neklyudov, Max Welling, Evgenii Egorov, Dmitry Vetrov
- abstract: Markov Chain Monte Carlo (MCMC) is a computational approach to fundamental problems such as inference, integration, optimization, and simulation. The field has developed a broad spectrum of algorithms, varying in the way they are motivated, the way they are applied and how efficiently they sample. Despite all the differences, many of them share the same core principle, which we unify as the Involutive MCMC (iMCMC) framework. Building upon this, we describe a wide range of MCMC algorithms in terms of iMCMC, and formulate a number of “tricks” which one can use as design principles for developing new MCMC algorithms. Thus, iMCMC provides a unified view of many known MCMC algorithms, which facilitates the derivation of powerful extensions. We demonstrate the latter with two examples where we transform known reversible MCMC algorithms into more efficient irreversible ones.

Aggregation of Multiple Knockoffs

- Tuan-Binh Nguyen, Jerome-Alexis Chevalier, Bertrand Thirion, Sylvain Arlot
- abstract: We develop an extension of the knockoff inference procedure, introduced by Barber & Candes (2015). This new method, called Aggregation of Multiple Knockoffs (AKO), addresses the instability inherent to the random nature of knockoff-based inference. Specifically, AKO improves both the stability and power compared with the original knockoff algorithm while still maintaining guarantees for false discovery rate control. We provide a new inference procedure, prove its core properties, and demonstrate its benefits in a set of experiments on synthetic and real datasets.

LEEP: A New Measure to Evaluate Transferability of Learned Representations

- Cuong Nguyen, Tal Hassner, Matthias Seeger, Cedric Archambeau
- abstract: We introduce a new measure to evaluate the transferability of representations learned by classifiers. Our measure, the Log Expected Empirical Prediction (LEEP), is simple and easy to compute: when given a classifier trained on a source data set, it only requires running the target data set through this classifier once. We analyze the properties of LEEP theoretically and demonstrate its effectiveness empirically. Our analysis shows that LEEP can predict the performance and convergence speed of both transfer and meta-transfer learning methods, even for small or imbalanced data. Moreover, LEEP outperforms recently proposed transferability measures such as negative conditional entropy and H scores. Notably, when transferring from ImageNet to CIFAR100, LEEP can achieve up to 30% improvement compared to the best competing method in terms of the correlations with actual transfer accuracy.

Graph Homomorphism Convolution

- Hoang Nguyen, Takanori Maehara
- abstract: In this paper, we study the graph classification problem from the graph homomorphism perspective. We consider the homomorphisms from F to G, where G is a graph of interest (e.g. molecules or social networks) and F belongs to some family of graphs (e.g. paths or non-isomorphic trees). We show that graph homomorphism numbers provide a natural invariant (isomorphism invariant and $\mathcal{F}$-invariant) embedding maps which can be used for graph classification. Viewing the expressive power of a graph classifier by the $\mathcal{F}$-indistinguishable concept, we prove the universality property of graph homomorphism vectors in approximating $\mathcal{F}$-invariant functions. In practice, by choosing $\mathcal{F}$ whose elements have bounded tree-width, we show that the homomorphism method is efficient compared with other methods.

Knowing The What But Not The Where in Bayesian Optimization

- Vu Nguyen, Michael A. Osborne
- abstract: Bayesian optimization has demonstrated impressive success in finding the optimum input x$\ast$ and output f$\ast$ = f(x$\ast$) = max f(x) of a black-box function f. In some applications, however, the optimum output is known in advance and the goal is to find the corresponding optimum input. Existing work in Bayesian optimization (BO) has not effectively exploited the knowledge of f$\ast$ for optimization. In this paper, we consider a new setting in BO in which the knowledge of the optimum output is available. Our goal is to exploit the knowledge about f$\ast$ to search for the input x$\ast$ efficiently. To achieve this goal, we first transform the Gaussian process surrogate using the information about the optimum output. Then, we propose two acquisition functions, called confidence bound minimization and expected regret minimization, which exploit the knowledge about the optimum output to identify the optimum input more efficient. We show that our approaches work intuitively and quantitatively better performance against standard BO methods. We demonstrate real applications in tuning a deep reinforcement learning algorithm on the CartPole problem and XGBoost on Skin Segmentation dataset in which the optimum values are publicly available.

Robust Bayesian Classification Using An Optimistic Score Ratio

- Viet Anh Nguyen, Nian Si, Jose Blanchet
- abstract: We build a Bayesian contextual classification model using an optimistic score ratio for robust binary classification when there is limited information on the class-conditional, or contextual, distribution. The optimistic score searches for the distribution that is most plausible to explain the observed outcomes in the testing sample among all distributions belonging to the contextual ambiguity set which is prescribed using a limited structural constraint on the mean vector and the covariance matrix of the underlying contextual distribution. We show that the Bayesian classifier using the optimistic score ratio is conceptually attractive, delivers solid statistical guarantees and is computationally tractable. We showcase the power of the proposed optimistic score ratio classifier on both synthetic and empirical data.

Streaming k-Submodular Maximization under Noise subject to Size Constraint

- Lan Nguyen, My T. Thai
- abstract: Maximizing on k-submodular functions subject to size constraint has received extensive attention recently. In this paper, we investigate a more realistic scenario of this problem that (1) obtaining exact evaluation of an objective function is impractical, instead, its noisy version is acquired; and (2) algorithms are required to take only one single pass over dataset, producing solutions in a timely manner. We propose two novel streaming algorithms, namely DStream and RStream, with their theoretical performance guarantees. We further demonstrate the efficiency of our algorithms in two application, showing that our algorithms can return comparative results to state-of-the-art non-streaming methods while using a much fewer number of queries.

LP-SparseMAP: Differentiable Relaxed Optimization for Sparse Structured Prediction

- Vlad Niculae, Andre Martins
- abstract: Structured predictors require solving a combinatorial optimization problem over a large number of structures, such as dependency trees or alignments. When embedded as structured hidden layers in a neural net, argmin differentiation and efficient gradient computation are further required. Recently, SparseMAP has been proposed as a differentiable, sparse alternative to maximum a posteriori (MAP) and marginal inference. SparseMAP returns an interpretable combination of a small number of structures; its sparsity being the key to efficient optimization. However, SparseMAP requires access to an exact MAP oracle in the structured model, excluding, e.g., loopy graphical models or logic constraints, which generally require approximate inference. In this paper, we introduce LP-SparseMAP, an extension of SparseMAP addressing this limitation via a local polytope relaxation. LP-SparseMAP uses the flexible and powerful language of factor graphs to define expressive hidden structures, supporting coarse decompositions, hard logic constraints, and higher-order correlations. We derive the forward and backward algorithms needed for using LP-SparseMAP as a structured hidden or output layer. Experiments in three structured tasks show benefits versus SparseMAP and Structured SVM.

Semi-Supervised StyleGAN for Disentanglement Learning

- Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, Animashree Anandkumar
- abstract: Disentanglement learning is crucial for obtaining disentangled representations and controllable generation. Current disentanglement methods face several inherent limitations: difficulty with high-resolution images, primarily focusing on learning disentangled representations, and non-identifiability due to the unsupervised setting. To alleviate these limitations, we design new architectures and loss functions based on StyleGAN (Karras et al., 2019), for semi-supervised high-resolution disentanglement learning. We create two complex high-resolution synthetic datasets for systematic testing. We investigate the impact of limited supervision and find that using only 0.25% 2.5% of labeled data is sufficient for good disentanglement on both synthetic and real datasets. We propose new metrics to quantify generator controllability, and observe there may exist a crucial trade-off between disentangled representation learning and controllable generation. We also consider semantic fine-grained image editing to achieve better generalization to unseen images.

Supervised learning: no loss no cry

- Richard Nock, Aditya Menon
- abstract: Supervised learning requires the specification of a loss function to minimise. While the theory of admissible losses from both a computational and statistical perspective is well-developed, these offer a panoply of different choices. In practice, this choice is typically made in an \emph{ad hoc} manner. In hopes of making this procedure more principled, the problem of \emph{learning the loss function} for a downstream task (e.g., classification) has garnered recent interest. However, works in this area have been generally empirical in nature. In this paper, we revisit the {\sc SLIsotron} algorithm of Kakade et al. (2011) through a novel lens, derive a generalisation based on Bregman divergences, and show how it provides a principled procedure for learning the loss. In detail, we cast {\sc SLIsotron} as learning a loss from a family of composite square losses. By interpreting this through the lens of \emph{proper losses}, we derive a generalisation of {\sc SLIsotron} based on Bregman divergences. The resulting {\sc BregmanTron} algorithm jointly learns the loss along with the classifier. It comes equipped with a simple guarantee of convergence for the loss it learns, and its set of possible outputs comes with a guarantee of agnostic approximability of Bayes rule. Experiments indicate that the {\sc BregmanTron} significantly outperforms the {\sc SLIsotron}, and that the loss it learns can be minimized by other algorithms for different tasks, thereby opening the interesting problem of \emph{loss transfer} between domains.

Consistent Structured Prediction with Max-Min Margin Markov Networks

- Alex Nowak, Francis Bach, Alessandro Rudi
- abstract: Max-margin methods for binary classification such as the support vector machine (SVM) have been extended to the structured prediction setting under the name of max-margin Markov networks (M^3N), or more generally structural SVMs. Unfortunately, these methods are statistically inconsistent when the relationship between inputs and labels is far from deterministic. We overcome such limitations by defining the learning problem in terms of a “max-min” margin formulation, naming the resulting method max-min margin Markov networks (M^4N). We prove consistency and finite sample generalization bounds for M^4N and provide an explicit algorithm to compute the estimator. The algorithm achieves a generalization error of $O(1/\sqrt{n})$ for a total cost of $O(n)$ projection-oracle calls (which have at most the same cost as the max-oracle from M^3N). Experiments on multi-class classification, ordinal regression, sequence prediction and ranking demonstrate the effectiveness of the proposed method.

T-Basis: a Compact Representation for Neural Networks

- Anton Obukhov, Maxim Rakhuba, Stamatios Georgoulis, Menelaos Kanakis, Dengxin Dai, Luc Van Gool
- abstract: We introduce T-Basis, a novel concept for a compact representation of a set of tensors, each of an arbitrary shape, which is often seen in Neural Networks. Each of the tensors in the set is modeled using Tensor Rings, though the concept applies to other Tensor Networks. Owing its name to the T-shape of nodes in diagram notation of Tensor Rings, T-Basis is simply a list of equally shaped three-dimensional tensors, used to represent Tensor Ring nodes. Such representation allows us to parameterize the tensor set with a small number of parameters (coefficients of the T-Basis tensors), scaling logarithmically with each tensor’s size in the set and linearly with the dimensionality of T-Basis. We evaluate the proposed approach on the task of neural network compression and demonstrate that it reaches high compression rates at acceptable performance drops. Finally, we analyze memory and operation requirements of the compressed networks and conclude that T-Basis networks are equally well suited for training and inference in resource-constrained environments and usage on the edge devices. Project website: obukhov.ai/tbasis.

Eliminating the Invariance on the Loss Landscape of Linear Autoencoders

- Reza Oftadeh, Jiayi Shen, Zhangyang Wang, Dylan Shell
- abstract: This paper proposes a new loss function for linear autoencoders (LAEs) and analytically identifies the structure of the associated loss surface. Optimizing the conventional Mean Square Error (MSE) loss results in a decoder matrix that spans the principal subspace of the sample covariance of the data, but, owing to an invariance that cancels out in the global map, it will fail to identify the exact eigenvectors. We show here that our proposed loss function eliminates this issue, so the decoder converges to the exact ordered unnormalized eigenvectors of the sample covariance matrix. We characterize the full structure of the new loss landscape by establishing an analytical expression for the set of all critical points, showing that it is a subset of critical points of MSE, and that all local minima are still global. Specifically, the invariant global minima under MSE are shown to become saddle points under the new loss. Additionally, the computational complexity of the loss and its gradients are the same as MSE and, thus, the new loss is not only of theoretical importance but is of practical value, e.g., for low-rank approximation.

On the (In)tractability of Computing Normalizing Constants for the Product of Determinantal Point Processes

- Naoto Ohsaka, Tatsuya Matsuoka
- abstract: We consider the product of determinantal point processes (DPPs), a point process whose probability mass is proportional to the product of principal minors of multiple matrices as a natural, promising generalization of DPPs. We study the computational complexity of computing its normalizing constant, which is among the most essential probabilistic inference tasks. Our complexity-theoretic results (almost) rule out the existence of efficient algorithms for this task, unless input matrices are forced to have favorable structures. In particular, we prove the following: (1) Computing $\sum_{S} \det(\mathbf{A}*{S,S})^p$ *exactly for every (fixed) positive even integer p is $\textsf{UP}$-hard and $\textsf{Mod}_3\textsf{P}$-hard, which gives a negative answer to an open question posed by Kulesza and Taskar (2012). (2) $\sum*{S} \det(\mathbf{A}{S,S}) \det(\mathbf{B}*{S,S}) \det(\mathbf{C}{S,S})$ is $\textsf{NP}$-hard to approximate within a factor of $ 2^{\mathcal{O}(|\mathcal{I}|^{1-\epsilon})} $ for any $\epsilon > 0$, where $|\mathcal{I}|$ is the input size. This result is stronger than $\sharp\textsf{P}$-hardness for the case of two matrices by Gillenwater (2014). (3) There exists a $ k^{\mathcal{O}(k)} |\mathcal{I}|^{\mathcal{O}(1)} $-time algorithm for computing $\sum*{S} \det(\mathbf{A}{S,S}) \det(\mathbf{B}*{S,S})$*, where k is "the maximum rank of $\mathbf{A}$ and $\mathbf{B}$" or "the treewidth of the graph formed by nonzero entries of $\mathbf{A}$ and $\mathbf{B}$." Such parameterized algorithms are said to be fixed-parameter tractable.

Can Increasing Input Dimensionality Improve Deep Reinforcement Learning?

- Kei Ota, Tomoaki Oiki, Devesh Jha, Toshisada Mariyama, Daniel Nikovski
- abstract: Deep reinforcement learning (RL) algorithms have recently achieved remarkable successes in various sequential decision making tasks, leveraging advances in methods for training large deep networks. However, these methods usually require large amounts of training data, which is often a big problem for real-world applications. One natural question to ask is whether learning good representations for states and using larger networks helps in learning better policies. In this paper, we try to study if increasing input dimensionality helps improve performance and sample efficiency of model-free deep RL algorithms. To do so, we propose an online feature extractor network (OFENet) that uses neural nets to produce \emph{good} representations to be used as inputs to an off-policy RL algorithm. Even though the high dimensionality of input is usually thought to make learning of RL agents more difficult, we show that the RL agents in fact learn more efficiently with the high-dimensional representation than with the lower-dimensional state observations. We believe that stronger feature propagation together with larger networks allows RL agents to learn more complex functions of states and thus improves the sample efficiency. Through numerical experiments, we show that the proposed method achieves much higher sample efficiency and better performance. Codes for the proposed method are available at http://www.merl.com/research/license/OFENet

Interferometric Graph Transform: a Deep Unsupervised Graph Representation

- Edouard Oyallon
- abstract: We propose the Interferometric Graph Transform (IGT), which is a new class of deep unsupervised graph convolutional neural network for building graph representations. Our first contribution is to propose a generic, complex-valued spectral graph architecture obtained from a generalization of the Euclidean Fourier transform. We show that our learned representation consists of both discriminative and invariant features, thanks to a novel greedy concave objective. From our experiments, we conclude that our learning procedure exploits the topology of the spectral domain, which is normally a flaw of spectral methods, and in particular our method can recover an analytic operator for vision tasks. We test our algorithm on various and challenging tasks such as image classification (MNIST, CIFAR-10), community detection (Authorship, Facebook graph) and action recognition from 3D skeletons videos (SBU, NTU), exhibiting a new state-of-the-art in spectral graph unsupervised settings.

Learning to Score Behaviors for Guided Policy Optimization

- Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Krzysztof Choromanski, Anna Choromanska, Michael Jordan
- abstract: We introduce a new approach for comparing reinforcement learning policies, using Wasserstein distances (WDs) in a newly defined latent behavioral space. We show that by utilizing the dual formulation of the WD, we can learn score functions over policy behaviors that can in turn be used to lead policy optimization towards (or away from) (un)desired behaviors. Combined with smoothed WDs, the dual formulation allows us to devise efficient algorithms that take stochastic gradient descent steps through WD regularizers. We incorporate these regularizers into two novel on-policy algorithms, Behavior-Guided Policy Gradient and Behavior-Guided Evolution Strategies, which we demonstrate can outperform existing methods in a variety of challenging environments. We also provide an open source demo.

Neural Clustering Processes

- Ari Pakman, Yueqi Wang, Catalin Mitelut, Jinhyung Lee, Liam Paninski
- abstract: Probabilistic clustering models (or equivalently, mixture models) are basic building blocks in countless statistical models and involve latent random variables over discrete spaces. For these models, posterior inference methods can be inaccurate and/or very slow. In this work we introduce deep network architectures trained with labeled samples from any generative model of clustered datasets. At test time, the networks generate approximate posterior samples of cluster labels for any new dataset of arbitrary size. We develop two complementary approaches to this task, requiring either O(N) or O(K) network forward passes per dataset, where N is the dataset size and K the number of clusters. Unlike previous approaches, our methods sample the labels of all the data points from a well-defined posterior, and can learn nonparametric Bayesian posteriors since they do not limit the number of mixture components. As a scientific application, we present a novel approach to neural spike sorting for high-density multielectrode arrays.

Recovery of Sparse Signals from a Mixture of Linear Samples

- Soumyabrata Pal, Arya Mazumdar
- abstract: Mixture of linear regressions is a popular learning theoretic model that is used widely to represent heterogeneous data. In the simplest form, this model assumes that the labels are generated from either of two different linear models and mixed together. Recent works of Yin et al. and Krishnamurthy et al., 2019, focus on an experimental design setting of model recovery for this problem. It is assumed that the features can be designed and queried with to obtain their label. When queried, an oracle randomly selects one of the two different sparse linear models and generates a label accordingly. How many such oracle queries are needed to recover both of the models simultaneously? This question can also be thought of as a generalization of the well-known

compressed sensing problem (Candès and Tao, 2005, Donoho, 2006). In this work we address this query complexity problem and provide efficient algorithms that improves on the previously best known results.

Adversarial Mutual Information for Text Generation

- Boyuan Pan, Yazheng Yang, Kaizhao Liang, Bhavya Kailkhura, Zhongming Jin, Xian-Sheng Hua, Deng Cai, Bo Li
- abstract: Recent advances in maximizing mutual information (MI) between the source and target have demonstrated its effectiveness in text generation. However, previous works paid little attention to modeling the backward network of MI (i.e., dependency from the target to the source), which is crucial to the tightness of the variational information maximization lower bound. In this paper, we propose Adversarial Mutual Information (AMI): a text generation framework which is formed as a novel saddle point (min-max) optimization aiming to identify joint interactions between the source and target. Within this framework, the forward and backward networks are able to iteratively promote or demote each other's generated instances by comparing the real and synthetic data distributions. We also develop a latent noise sampling strategy that leverages random variations at the high-level semantic space to enhance the long term dependency in the generation process. Extensive experiments based on different text generation tasks demonstrate that the proposed AMI framework can significantly outperform several strong baselines, and we also show that AMI has potential to lead to a tighter lower bound of maximum mutual information for the variational information maximization problem.

Stabilizing Transformers for Reinforcement Learning

- Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphaël Lopez Kaufman, Aidan Clark, Seb Noury, Matthew Botvinick, Nicolas Heess, Raia Hadsell
- abstract: Owing to their ability to both effectively integrate information over long time horizons and scale to massive amounts of data, self-attention architectures have recently shown breakthrough success in natural language processing (NLP). Harnessing the transformer's ability to process long time horizons of information could provide a similar performance boost in partially observable reinforcement learning (RL) domains, but the large-scale transformers used in NLP have yet to be successfully applied to the RL setting. In this work we demonstrate that the standard transformer architecture is difficult to optimize, which was previously observed in the supervised learning setting but becomes especially pronounced with RL objectives. We propose architectural modifications that substantially improve the stability and learning speed of the original Transformer and XL variant. The proposed architecture, the Gated Transformer-XL (GTrXL), surpasses LSTMs on challenging memory environments and achieves state-of-the-art results on the multi-task DMLab-30 benchmark suite, exceeding the performance of an external memory architecture. We show that the GTrXL has stability and performance that consistently matches or exceeds a competitive LSTM baseline, including on more reactive tasks where memory is less critical.

Multiresolution Tensor Learning for Efficient and Interpretable Spatial Analysis

- Jung Yeon Park, Kenneth Carr, Stephan Zheng, Yisong Yue, Rose Yu
- abstract: Efficient and interpretable spatial analysis is crucial in many fields such as geology, sports, and climate science. Tensor latent factor models can describe higher-order correlations for spatial data. However, they are computationally expensive to train and are sensitive to initialization, leading to spatially incoherent, uninterpretable results. We develop a novel Multiresolution Tensor Learning (MRTL) algorithm for efficiently learning interpretable spatial patterns. MRTL initializes the latent factors from an approximate full-rank tensor model for improved interpretability and progressively learns from a coarse resolution to the fine resolution to reduce computation. We also prove the theoretical convergence and computational complexity of MRTL. When applied to two real-world datasets, MRTL demonstrates 4 5x speedup compared to a fixed resolution approach while yielding accurate and interpretable latent factors.

Meta Variance Transfer: Learning to Augment from the Others

- Seong-Jin Park, Seungju Han, Ji-Won Baek, Insoo Kim, Juhwan Song, Hae Beom Lee, Jae-Joon Han, Sung Ju Hwang
- abstract: Humans have the ability to robustly recognize objects with various factors of variations such as nonrigid transformations, background noises, and changes in lighting conditions. However, training deep learning models generally require huge amount of data instances under diverse variations, to ensure its robustness. To alleviate the need of collecting large amount of data and better learn to generalize with scarce data instances, we propose a novel meta-learning method which learns to transfer factors of variations from one class to another, such that it can improve the classification performance on unseen examples. Transferred variations generate virtual samples that augment the feature space of the target class during training, simulating upcoming query samples with similar variations. By sharing the factors of variations across different classes, the model becomes more robust to variations in the unseen examples and tasks using small number of examples per class. We validate our model on multiple benchmark datasets for few-shot classification and face recognition, on which our model significantly improves the performance of the base model, outperforming relevant baselines.

Structured Policy Iteration for Linear Quadratic Regulator

- Youngsuk Park, Ryan Rossi, Zheng Wen, Gang Wu, Handong Zhao
- abstract: Linear quadratic regulator (LQR) is one of the most popular frameworks to tackle continuous Markov decision process tasks. With its fundamental theory and tractable optimal policy, LQR has been revisited and analyzed in recent years, in terms of reinforcement learning scenarios such as the model-free or model-based setting. In this paper, we introduce the Structured Policy Iteration (S-PI) for LQR, a method capable of deriving a structured linear policy. Such a structured policy with (block) sparsity or low-rank can have significant advantages over the standard LQR policy: more interpretable, memory-efficient, and well-suited for the distributed setting. In order to derive such a policy, we first cast a regularized LQR problem when the model is known. Then, our Structured Policy Iteration (S-PI) algorithm, which takes a policy evaluation step and a policy improvement step in an iterative manner, can solve this regularized LQR efficiently. We further extend the S-PI algorithm to the model-free setting where a smoothing procedure is adopted to estimate the gradient. In both the known-model and model-free setting, we prove convergence analysis under the proper choice of parameters. Finally, the experiments demonstrate the advantages of S-PI in terms of balancing the LQR performance and level of structure by varying the weight parameter.

Regularized Optimal Transport is Ground Cost Adversarial

- François-Pierre Paty, Marco Cuturi
- abstract: Regularizing the optimal transport (OT) problem has proven crucial for OT theory to impact the field of machine learning. For instance, it is known that regularizing OT problems with entropy leads to faster computations and better differentiation using the Sinkhorn algorithm, as well as better sample complexity bounds than classic OT. In this work we depart from this practical perspective and propose a new interpretation of regularization as a robust mechanism, and show using Fenchel duality that any convex regularization of OT can be interpreted as ground cost adversarial. This incidentally gives access to a robust dissimilarity measure on the ground space, which can in turn be used in other applications. We propose algorithms to compute this robust cost, and illustrate the interest of this approach empirically.

Reducing Sampling Error in Batch Temporal Difference Learning

- Brahma Pavse, Ishan Durugkar, Josiah Hanna, Peter Stone

- abstract: Temporal difference (TD) learning is one of the main foundations of modern reinforcement learning. This paper studies the use of TD(0), a canonical TD algorithm, to estimate the value function of a given policy from a batch of data. In this batch setting, we show that TD(0) may converge to an inaccurate value function because the update following an action is weighted according to the number of times that action occurred in the batch – not the true probability of the action under the given policy. To address this limitation, we introduce \emph{policy sampling error corrected}-TD(0) (PSEC-TD(0)). PSEC-TD(0) first estimates the empirical distribution of actions in each state in the batch and then uses importance sampling to correct for the mismatch between the empirical weighting and the correct weighting for updates following each action. We refine the concept of a certainty-equivalence estimate and argue that PSEC-TD(0) is a more data efficient estimator than TD(0) for a fixed batch of data. Finally, we conduct an empirical evaluation of PSEC-TD(0) on three batch value function learning tasks, with a hyperparameter sensitivity analysis, and show that PSEC-TD(0) produces value function estimates with lower mean squared error than TD(0).

Acceleration through spectral density estimation

- Fabian Pedregosa, Damien Scieur
- abstract: We develop a framework for the average-case analysis of random quadratic problems and derive algorithms that are optimal under this analysis. This yields a new class of methods that achieve acceleration given a model of the Hessian's eigenvalue distribution. We develop explicit algorithms for the uniform, Marchenko-Pastur, and exponential distributions. These methods have a simple momentum-like update, in which each update only makes use on the current gradient and previous two iterates. Furthermore, the momentum and step-size parameters can be estimated without knowledge of the Hessian's smallest singular value, in contrast with classical accelerated methods like Nesterov acceleration and Polyak momentum. Through empirical benchmarks on quadratic and logistic regression problems, we identify regimes in which the the proposed methods improve over classical (worst-case) accelerated methods.

Einsum Networks: Fast and Scalable Learning of Tractable Probabilistic Circuits

- Robert Peharz, Steven Lang, Antonio Vergari, Karl Stelzner, Alejandro Molina, Martin Trapp, Guy Van Den Broeck, Kristian Kersting, Zoubin Ghahramani
- abstract: Probabilistic circuits (PCs) are a promising avenue for probabilistic modeling, as they permit a wide range of exact and efficient inference routines. Recent "deep-learning-style" implementations of PCs strive for a better scalability, but are still difficult to train on real-world data, due to their sparsely connected computational graphs. In this paper, we propose Einsum Networks (EiNets), a novel implementation design for PCs, improving prior art in several regards. At their core, EiNets combine a large number of arithmetic operations in a single monolithic einsum-operation, leading to speedups and memory savings of up to two orders of magnitude, in comparison to previous implementations. As an algorithmic contribution, we show that the implementation of Expectation-Maximization (EM) can be simplified for PCs, by leveraging automatic differentiation. Furthermore, we demonstrate that EiNets scale well to datasets which were previously out of reach, such as SVHN and CelebA, and that they can be used as faithful generative image models.

Learning Selection Strategies in Buchberger's Algorithm

- Dylan Peifer, Michael Stillman, Daniel Halpern-Leistner
- abstract: Studying the set of exact solutions of a system of polynomial equations largely depends on a single iterative algorithm, known as Buchberger's algorithm. Optimized versions of this algorithm are crucial for many computer algebra systems (e.g., Mathematica, Maple, Sage). We introduce a new approach to Buchberger's algorithm that uses reinforcement learning agents to perform S-pair selection, a key step in the algorithm. We then study how the difficulty of the problem depends on the choices of domain and distribution of polynomials, about which little is known. Finally, we train a policy model using proximal policy optimization (PPO) to learn S-pair selection strategies for random systems of binomial equations. In certain domains, the trained model outperforms state-of-the-art selection heuristics in total number of polynomial additions performed, which provides a proof-of-concept that recent developments in machine learning have the potential to improve performance of algorithms in symbolic computation.

Non-Autoregressive Neural Text-to-Speech

- Kainan Peng, Wei Ping, Zhao Song, Kexin Zhao
- abstract: In this work, we propose ParaNet, a non-autoregressive seq2seq model that converts text to spectrogram. It is fully convolutional and brings 46.7 times speed-up over the lightweight Deep Voice 3 at synthesis, while obtaining reasonably good speech quality. ParaNet also produces stable alignment between text and speech on the challenging test sentences by iteratively improving the attention in a layer-by-layer manner. Furthermore, we build the parallel text-to-speech system by applying various parallel neural vocoders, which can synthesize speech from text through a single feed-forward pass. We also explore a novel VAE-based approach to train the inverse autoregressive flow (IAF) based parallel vocoder from scratch, which avoids the need for distillation from a separately trained WaveNet as previous work.

Performative Prediction

- Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, Moritz Hardt
- abstract: When predictions support decisions they may influence the outcome they aim to predict. We call such predictions performative; the prediction influences the target. Performativity is a well-studied phenomenon in policy-making that has so far been neglected in supervised learning. When ignored, performativity surfaces as undesirable distribution shift, routinely addressed with retraining. We develop a risk minimization framework for performative prediction bringing together concepts from statistics, game theory, and causality. A conceptual novelty is an equilibrium notion we call performative stability. Performative stability implies that the predictions are calibrated not against past outcomes, but against the future outcomes that manifest from acting on the prediction. Our main results are necessary and sufficient conditions for the convergence of retraining to a performatively stable point of nearly minimal loss. In full generality, performative prediction strictly subsumes the setting known as strategic classification. We thus also give the first sufficient conditions for retraining to overcome strategic feedback effects.

Constructive Universal High-Dimensional Distribution Generation through Deep ReLU Networks

- Dmytro Perekrestenko, Stephan Müller, Helmut Bölcskei
- abstract: We present an explicit deep neural network construction that transforms uniformly distributed one-dimensional noise into an arbitrarily close approximation of any two-dimensional Lipschitz-continuous target distribution. The key ingredient of our design is a generalization of the "space-filling" property of sawtooth functions discovered in (Bailey & Telgarsky, 2018). We elicit the importance of depth - in our neural network construction - in driving the Wasserstein distance between the target distribution and the approximation realized by the network to zero. An extension to output distributions of arbitrary dimension is outlined. Finally, we show that the proposed construction does not incur a cost - in terms of error measured in Wasserstein-distance - relative to generating d-dimensional target distributions from d independent random variables.

Budgeted Online Influence Maximization

- Pierre Perrault, Jennifer Healey, Zheng Wen, Michal Valko

- abstract: We introduce a new budgeted framework for online influence maximization, considering the total cost of an advertising campaign instead of the common cardinality constraint on a chosen influencer set. Our approach models better the real-world setting where the cost of influencers varies and advertizers want to find the best value for their overall social advertising budget. We propose an algorithm assuming an independent cascade diffusion model and edge-level semi-bandit feedback, and provide both theoretical and experimental results. Our analysis is also valid for the cardinality-constraint setting and improves the state of the art regret bound in this case.

Low Bias Low Variance Gradient Estimates for Boolean Stochastic Networks

- Adeel Pervez, Taco Cohen, Efstratios Gavves
- abstract: Stochastic neural networks with discrete random variables are an important class of models for their expressiveness and interpretability. Since direct differentiation and backpropagation is not possible, Monte Carlo gradient estimation techniques are a popular alternative. Efficient stochastic gradient estimators, such Straight-Through and Gumbel-Softmax, work well for shallow stochastic models. Their performance, however, suffers with hierarchical, more complex models. We focus on stochastic networks with Boolean latent variables. To analyze such networks, we introduce the framework of harmonic analysis for Boolean functions to derive an analytic formulation for the bias and variance in the Straight-Through estimator. Exploiting these formulations, we propose \emph{FouST}, a low-bias and low-variance gradient estimation algorithm that is just as efficient. Extensive experiments show that FouST performs favorably compared to state-of-the-art biased estimators and is much faster than unbiased ones.

On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent

- Scott Pesme, Aymeric Dieuleveut, Nicolas Flammarion
- abstract: Constant step-size Stochastic Gradient Descent exhibits two phases: a transient phase during which iterates make fast progress towards the optimum, followed by a stationary phase during which iterates oscillate around the optimal point. In this paper, we show that efficiently detecting this transition and appropriately decreasing the step size can lead to fast convergence rates. We analyse the classical statistical test proposed by Pflug (1983), based on the inner product between consecutive stochastic gradients. Even in the simple case where the objective function is quadratic we show that this test cannot lead to an adequate convergence diagnostic. We then propose a novel and simple statistical procedure that accurately detects stationarity and we provide experimental results showing state-of-the-art performance on synthetic and real-word datasets.

Sample Factory: Egocentric 3D Control from Pixels at 100000 FPS with Asynchronous Reinforcement Learning

- Aleksei Petrenko, Zhehui Huang, Tushar Kumar, Gaurav Sukhatme, Vladlen Koltun
- abstract: Increasing the scale of reinforcement learning experiments has allowed researchers to achieve unprecedented results in both training sophisticated agents for video games, and in sim-to-real transfer for robotics. Typically such experiments rely on large distributed systems and require expensive hardware setups, limiting wider access to this exciting area of research. In this work we aim to solve this problem by optimizing the efficiency and resource utilization of reinforcement learning algorithms instead of relying on distributed computation. We present the "Sample Factory", a high-throughput training system optimized for a single-machine setting. Our architecture combines a highly efficient, asynchronous, GPU-based sampler with off-policy correction techniques, allowing us to achieve throughput higher than 10^5 environment frames/second on non-trivial control problems in 3D without sacrificing sample efficiency. We extend Sample Factory to support self-play and population-based training and apply these techniques to train highly capable agents for a multiplayer first-person shooter game. Github: https://github.com/alex-petrenko/sample-factory

IPBoost – Non-Convex Boosting via Integer Programming

- Marc Pfetsch, Sebastian Pokutta
- abstract: Recently non-convex optimization approaches for solving machine learning problems have gained significant attention. In this paper we explore non-convex boosting in classification by means of integer programming and demonstrate real-world practicability of the approach while circumvent- ing shortcomings of convex boosting approaches. We report results that are comparable to or better than the current state-of-the-art.

On Unbalanced Optimal Transport: An Analysis of Sinkhorn Algorithm

- Khiem Pham, Khang Le, Nhat Ho, Tung Pham, Hung Bui
- abstract: We provide a computational complexity analysis for the Sinkhorn algorithm that solves the entropic regularized Unbalanced Optimal Transport (UOT) problem between two measures of possibly different masses with at most n components. We show that the complexity of the Sinkhorn algorithm for finding an ε-approximate solution to the UOT problem is of order $\widetilde{\mathcal{O}}(n^2/ \varepsilon)$. To the best of our knowledge, this complexity is better than the best known complexity upper bound of the Sinkhorn algorithm for solving the Optimal Transport (OT) problem, which is of order $\widetilde{\mathcal{O}}(n^2/\varepsilon^2)$. Our proof technique is based on the geometric convergence rate of the Sinkhorn updates to the optimal dual solution of the entropic regularized UOT problem and scaling properties of the primal solution. It is also different from the proof technique used to establish the complexity of the Sinkhorn algorithm for approximating the OT problem since the UOT solution does not need to meet the marginal constraints of the measures.

Scalable Differential Privacy with Certified Robustness in Adversarial Learning

- Hai Phan, My T. Thai, Han Hu, Ruoming Jin, Tong Sun, Dejing Dou
- abstract: In this paper, we aim to develop a scalable algorithm to preserve differential privacy (DP) in adversarial learning for deep neural networks (DNNs), with certified robustness to adversarial examples. By leveraging the sequential composition theory in DP, we randomize both input and latent spaces to strengthen our certified robustness bounds. To address the trade-off among model utility, privacy loss, and robustness, we design an original adversarial objective function, based on the post-processing property in DP, to tighten the sensitivity of our model. A new stochastic batch training is proposed to apply our mechanism on large DNNs and datasets, by bypassing the vanilla iterative batch-by-batch training in DP DNNs. An end-to-end theoretical analysis and evaluations show that our mechanism notably improves the robustness and scalability of DP DNNs.

Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks

- Mert Pilanci, Tolga Ergen
- abstract: We develop exact representations of training two-layer neural networks with rectified linear units (ReLUs) in terms of a single convex program with number of variables polynomial in the number of training samples and the number of hidden neurons. Our theory utilizes semi-infinite duality and minimum norm regularization. We show that ReLU networks trained with standard weight decay are equivalent to block ℓ_1 penalized convex models. Moreover, we show that certain standard convolutional linear networks are equivalent semi-definite programs which can be simplified to ℓ_1 regularized linear models in a polynomial sized discrete Fourier feature space

WaveFlow: A Compact Flow-based Model for Raw Audio

- Wei Ping, Kainan Peng, Kexin Zhao, Zhao Song

- abstract: In this work, we propose WaveFlow, a small-footprint generative flow for raw audio, which is directly trained with maximum likelihood. It handles the long-range structure of 1-D waveform with a dilated 2-D convolutional architecture, while modeling the local variations using expressive autoregressive functions. WaveFlow provides a unified view of likelihood-based models for 1-D data, including WaveNet and WaveGlow as special cases. It generates high-fidelity speech as WaveNet, while synthesizing several orders of magnitude faster as it only requires a few sequential steps to generate very long waveforms with hundreds of thousands of time-steps. Furthermore, it can significantly reduce the likelihood gap that has existed between autoregressive models and flow-based models for efficient synthesis. Finally, our small-footprint WaveFlow has only 5.91M parameters, which is 15{\texttimes} smaller than WaveGlow. It can generate 22.05 kHz high-fidelity audio 42.6{\texttimes} faster than real-time (at a rate of 939.3 kHz) on a V100 GPU without engineered inference kernels.

Randomization matters How to defend against strong adversarial attacks

- Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, Jamal Atif
- abstract: \emph{Is there a classifier that ensures optimal robustness against all adversarial attacks?} This paper tackles this question by adopting a game-theoretic point of view. We present the adversarial attacks and defenses problem as an \emph{infinite} zero-sum game where classical results (\emph{e.g.} Nash or Sion theorems) do not apply. We demonstrate the non-existence of a Nash equilibrium in our game when the classifier and the Adversary are both deterministic, hence giving a negative answer to the above question in the deterministic regime. Nonetheless, the question remains open in the randomized regime. We tackle this problem by showing that any deterministic classifier can be outperformed by a randomized one. This gives arguments for using randomization, and leads us to a simple method for building randomized classifiers that are robust to state-or-the-art adversarial attacks. Empirical results validate our theoretical analysis, and show that our defense method considerably outperforms Adversarial Training against strong adaptive attacks, by achieving 0.55 accuracy under adaptive PGD-attack on CIFAR10, compared to 0.42 for Adversarial training.

Efficient Domain Generalization via Common-Specific Low-Rank Decomposition

- Vihari Piratla, Praneeth Netrapalli, Sunita Sarawagi
- abstract: Domain generalization refers to the task of training a model which generalizes to new domains that are not seen during training. We present CSD (Common Specific Decomposition), for this setting, which jointly learns a common component (which generalizes to new domains) and a domain specific component (which overfits on training domains). The domain specific components are discarded after training and only the common component is retained. The algorithm is extremely simple and involves only modifying the final linear classification layer of any given neural network architecture. We present a principled analysis to understand existing approaches, provide identifiability results of CSD, and study the effect of low-rank on domain generalization. We show that CSD either matches or beats state of the art approaches for domain generalization based on domain erasure, domain perturbed data augmentation, and meta-learning. Further diagnostics on rotated MNIST, where domains are interpretable, confirm the hypothesis that CSD successfully disentangles common and domain specific components and hence leads to better domain generalization; moreover, our code and dataset are publicly available at the following URL: \url{https://github.com/vihari/csd}.

Dissecting Non-Vacuous Generalization Bounds based on the Mean-Field Approximation

- Konstantinos Pitas
- abstract: Explaining how overparametrized neural networks simultaneously achieve low risk and zero empirical risk on benchmark datasets is an open problem. PAC-Bayes bounds optimized using variational inference (VI) have been recently proposed as a promising direction in obtaining non-vacuous bounds. We show empirically that this approach gives negligible gains when modelling the posterior as a Gaussian with diagonal covariance—known as the mean-field approximation. We investigate common explanations, such as the failure of VI due to problems in optimization or choosing a suboptimal prior. Our results suggest that investigating richer posteriors is the most promising direction forward.

Maximum Entropy Gain Exploration for Long Horizon Multi-goal Reinforcement Learning

- Silviu Pitis, Harris Chan, Stephen Zhao, Bradly Stadie, Jimmy Ba
- abstract: What goals should a multi-goal reinforcement learning agent pursue during training in long-horizon tasks? When the desired (test time) goal distribution is too distant to offer a useful learning signal, we argue that the agent should not pursue unobtainable goals. Instead, it should set its own intrinsic goals that maximize the entropy of the historical achieved goal distribution. We propose to optimize this objective by having the agent pursue past achieved goals in sparsely explored areas of the goal space, which focuses exploration on the frontier of the achievable goal set. We show that our strategy achieves an order of magnitude better sample efficiency than the prior state of the art on long-horizon multi-goal tasks including maze navigation and block stacking.

Explaining Groups of Points in Low-Dimensional Representations

- Gregory Plumb, Jonathan Terhorst, Sriram Sankararaman, Ameet Talwalkar
- abstract: A common workflow in data exploration is to learn a low-dimensional representation of the data, identify groups of points in that representation, and examine the differences between the groups to determine what they represent. We treat this workflow as an interpretable machine learning problem by leveraging the model that learned the low-dimensional representation to help identify the key differences between the groups. To solve this problem, we introduce a new type of explanation, a Global Counterfactual Explanation (GCE), and our algorithm, Transitive Global Translations (TGT), for computing GCEs. TGT identifies the differences between each pair of groups using compressed sensing but constrains those pairwise differences to be consistent among all of the groups. Empirically, we demonstrate that TGT is able to identify explanations that accurately explain the model while being relatively sparse, and that these explanations match real patterns in the data.

On the Unreasonable Effectiveness of the Greedy Algorithm: Greedy Adapts to Sharpness

- Sebastian Pokutta, Mohit Singh, Alfredo Torrico
- abstract: It is well known that the standard greedy algorithm guarantees a worst-case approximation factor of $1-1/e$ when maximizing a monotone submodular function under a cardinality constraint. However, empirical studies show that its performance is substantially better in practice. This raises a natural question of explaining this improved performance of the greedy algorithm. In this work, we define sharpness for submodular functions as a candidate explanation for this phenomenon. We show that the greedy algorithm provably performs better as the sharpness of the submodular function increases. This improvement ties in closely with the faster convergence rates of first order methods for sharp functions in convex optimization.

Skew-Fit: State-Covering Self-Supervised Reinforcement Learning

- Vitchyr Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, Sergey Levine
- abstract: Autonomous agents that must exhibit flexible and broad capabilities will need to be equipped with large repertoires of skills. Defining each skill with a manually-designed reward function limits this repertoire and imposes a manual engineering burden. Self-supervised agents that set their own goals can automate this process, but designing appropriate goal setting objectives can be difficult, and often involves heuristic design decisions. In this paper, we propose a formal exploration objective for goal-reaching policies that maximizes state coverage. We show that this objective is equivalent to maximizing goal reaching performance together with the entropy of the goal distribution, where goals correspond to full state observations. To instantiate this

principle, we present an algorithm called Skew-Fit for learning a maximum-entropy goal distributions. We prove that, under regularity conditions, Skew-Fit converges to a uniform distribution over the set of valid states, even when we do not know this set beforehand. Our experiments show that combining Skew-Fit for learning goal distributions with existing goal-reaching methods outperforms a variety of prior methods on open-sourced visual goal-reaching tasks. Moreover, we demonstrate that Skew-Fit enables a real-world robot to learn to open a door, entirely from scratch, from pixels, and without any manually-designed reward function.

SoftSort: A Continuous Relaxation for the argsort Operator

- Sebastian Prillo, Julian Eisenschlos
- abstract: While sorting is an important procedure in computer science, the argsort operator - which takes as input a vector and returns its sorting permutation - has a discrete image and thus zero gradients almost everywhere. This prohibits end-to-end, gradient-based learning of models that rely on the argsort operator. A natural way to overcome this problem is to replace the argsort operator with a continuous relaxation. Recent work has shown a number of ways to do this, but the relaxations proposed so far are computationally complex. In this work we propose a simple continuous relaxation for the argsort operator which has the following qualities: it can be implemented in three lines of code, achieves state-of-the-art performance, is easy to reason about mathematically - substantially simplifying proofs - and is faster than competing approaches. We open source the code to reproduce all of the experiments and results.

Graph-based Nearest Neighbor Search: From Practice to Theory

- Liudmila Prokhorenkova, Aleksandr Shekhovtsov
- abstract: Graph-based approaches are empirically shown to be very successful for the nearest neighbor search (NNS). However, there has been very little research on their theoretical guarantees. We fill this gap and rigorously analyze the performance of graph-based NNS algorithms, specifically focusing on the low-dimensional ($d \ll \log n$) regime. In addition to the basic greedy algorithm on nearest neighbor graphs, we also analyze the most successful heuristics commonly used in practice: speeding up via adding shortcut edges and improving accuracy via maintaining a dynamic list of candidates. We believe that our theoretical insights supported by experimental analysis are an important step towards understanding the limits and benefits of graph-based NNS algorithms.

Adversarial Risk via Optimal Transport and Optimal Couplings

- Muni Sreenivas Pydi, Varun Jog
- abstract: The accuracy of modern machine learning algorithms deteriorates severely on adversarially manipulated test data. Optimal adversarial risk quantifies the best error rate of any classifier in the presence of adversaries, and optimal adversarial classifiers are sought that minimize adversarial risk. In this paper, we investigate the optimal adversarial risk and optimal adversarial classifiers from an optimal transport perspective. We present a new and simple approach to show that the optimal adversarial risk for binary classification with $0-1$ loss function is completely characterized by an optimal transport cost between the probability distributions of the two classes, for a suitably defined cost function. We propose a novel coupling strategy that achieves the optimal transport cost for several univariate distributions like Gaussian, uniform and triangular. Using the optimal couplings, we obtain the optimal adversarial classifiers in these settings and show how they differ from optimal classifiers in the absence of adversaries. Based on our analysis, we evaluate algorithm-independent fundamental limits on adversarial risk for CIFAR-10, MNIST, Fashion-MNIST and SVHN datasets, and Gaussian mixtures based on them.

Deep Isometric Learning for Visual Recognition

- Haozhi Qi, Chong You, Xiaolong Wang, Yi Ma, Jitendra Malik
- abstract: Initialization, normalization, and skip connections are believed to be three indispensable techniques for training very deep convolutional neural networks and obtaining state-of-the-art performance. This paper shows that deep vanilla ConvNets without normalization nor skip connections can also be trained to achieve surprisingly good performance on standard image recognition benchmarks. This is achieved by enforcing the convolution kernels to be near isometric during initialization and training, as well as by using a variant of ReLU that is shifted towards being isometric. Further experiments show that if combined with skip connections, such near isometric networks can achieve performances on par with (for ImageNet) and better than (for COCO) the standard ResNet, even without normalization at all. Our code is available at https://github.com/HaozhiQi/ISONet.

Unsupervised Speech Decomposition via Triple Information Bottleneck

- Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, David Cox
- abstract: Speech information can be roughly decomposed into four components: language content, timbre, pitch, and rhythm. Obtaining disentangled representations of these components is useful in many speech analysis and generation applications. Recently, state-of-the-art voice conversion systems have led to speech representations that can disentangle speaker-dependent and independent information. However, these systems can only disentangle timbre, while information about pitch, rhythm and content is still mixed together. Further disentangling the remaining speech components is an under-determined problem in the absence of explicit annotations for each component, which are difficult and expensive to obtain. In this paper, we propose SpeechSplit, which can blindly decompose speech into its four components by introducing three carefully designed information bottlenecks. SpeechSplit is among the first algorithms that can separately perform style transfer on timbre, pitch and rhythm without text labels. Our code is publicly available at https://github.com/auspicious3000/SpeechSplit.

Scalable Differentiable Physics for Learning and Control

- Yi-Ling Qiao, Junbang Liang, Vladlen Koltun, Ming Lin
- abstract: Differentiable physics is a powerful approach to learning and control problems that involve physical objects and environments. While notable progress has been made, the capabilities of differentiable physics solvers remain limited. We develop a scalable framework for differentiable physics that can support a large number of objects and their interactions. To accommodate objects with arbitrary geometry and topology, we adopt meshes as our representation and leverage the sparsity of contacts for scalable differentiable collision handling. Collisions are resolved in localized regions to minimize the number of optimization variables even when the number of simulated objects is high. We further accelerate implicit differentiation of optimization with nonlinear constraints. Experiments demonstrate that the presented framework requires up to two orders of magnitude less memory and computation in comparison to recent particle-based methods. We further validate the approach on inverse problems and control scenarios, where it outperforms derivative-free and model-free baselines by at least an order of magnitude.

Robust One-Bit Recovery via ReLU Generative Networks: Near-Optimal Statistical Rate and Global Landscape Analysis

- Shuang Qiu, Xiaohan Wei, Zhuoran Yang
- abstract: We study the robust one-bit compressed sensing problem whose goal is to design an algorithm that faithfully recovers any sparse target vector $\theta_0\in\mathbb{R}^d$ \emph{uniformly} via m quantized noisy measurements. Specifically, we consider a new framework for this problem where the sparsity is implicitly enforced via mapping a low dimensional representation $x_0 \in \mathbb{R}^k$ through a known n-layer ReLU generative network $G:\mathbb{R}^k\rightarrow\mathbb{R}^d$ such that $\theta_0 = G(x_0)$. Such a framework poses low-dimensional priors on θ_0

without a known sparsity basis. We propose to recover the target $G(x_0)$ solving an unconstrained empirical risk minimization (ERM). Under a weak \emph{sub-exponential measurement assumption}, we establish a joint statistical and computational analysis. In particular, we prove that the ERM estimator in this new framework achieves a statistical rate of $m=\widetilde{\mathcal{O}}(kn \log d /\varepsilon^2)$ recovering any $G(x_0)$ uniformly up to an error ε. When the network is shallow (i.e., n is small), we show this rate matches the information-theoretic lower bound up to logarithm factors of ε^{-1}. From the lens of computation, we prove that under proper conditions on the network weights, our proposed empirical risk, despite non-convexity, has no stationary point outside of small neighborhoods around the true representation x_0 and its negative multiple; furthermore, we show that the global minimizer of the empirical risk stays within the neighborhood around x_0 rather than its negative multiple under further assumptions on weights.

Few-shot Relation Extraction via Bayesian Meta-learning on Relation Graphs

- Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, Jian Tang
- abstract: This paper studies few-shot relation extraction, which aims at predicting the relation for a pair of entities in a sentence by training with a few labeled examples in each relation. To more effectively generalize to new relations, in this paper we study the relationships between different relations and propose to leverage a global relation graph. We propose a novel Bayesian meta-learning approach to effectively learn the posterior distribution of the prototype vectors of relations, where the initial prior of the prototype vectors is parameterized with a graph neural network on the global relation graph. Moreover, to effectively optimize the posterior distribution of the prototype vectors, we propose to use the stochastic gradient Langevin dynamics, which is related to the MAML algorithm but is able to handle the uncertainty of the prototype vectors. The whole framework can be effectively and efficiently optimized in an end-to-end fashion. Experiments on two benchmark datasets prove the effectiveness of our proposed approach against competitive baselines in both the few-shot and zero-shot settings.

DeepCoDA: personalized interpretability for compositional health data

- Thomas Quinn, Dang Nguyen, Santu Rana, Sunil Gupta, Svetha Venkatesh
- abstract: Abstract Interpretability allows the domain-expert to directly evaluate the model’s relevance and reliability, a practice that offers assurance and builds trust. In the healthcare setting, interpretable models should implicate relevant biological mechanisms independent of technical factors like data pre-processing. We define personalized interpretability as a measure of sample-specific feature attribution, and view it as a minimum requirement for a precision health model to justify its conclusions. Some health data, especially those generated by high-throughput sequencing experiments, have nuances that compromise precision health models and their interpretation. These data are compositional, meaning that each feature is conditionally dependent on all other features. We propose the Deep Compositional Data Analysis (DeepCoDA) framework to extend precision health modelling to high-dimensional compositional data, and to provide personalized interpretability through patient-specific weights. Our architecture maintains state-of-the-art performance across 25 real-world data sets, all while producing interpretations that are both personalized and fully coherent for compositional data.

Fast and Private Submodular and k-Submodular Functions Maximization with Matroid Constraints

- Akbar Rafiey, Yuichi Yoshida
- abstract: The problem of maximizing nonnegative monotone submodular functions under a certain constraint has been intensively studied in the last decade, and a wide range of efficient approximation algorithms have been developed for this problem. Many machine learning problems, including data summarization and influence maximization, can be naturally modeled as the problem of maximizing monotone submodular functions. However, when such applications involve sensitive data about individuals, their privacy concerns should be addressed. In this paper, we study the problem of maximizing monotone submodular functions subject to matroid constraints in the framework of differential privacy. We provide $(1-\frac{1}{\mathrm{e}})$-approximation algorithm which improves upon the previous results in terms of approximation guarantee. This is done with an almost cubic number of function evaluations in our algorithm. Moreover, we study k-submodularity, a natural generalization of submodularity. We give the first $\frac{1}{2}$-approximation algorithm that preserves differential privacy for maximizing monotone k-submodular functions subject to matroid constraints. The approximation ratio is asymptotically tight and is obtained with an almost linear number of function evaluations.

Transparency Promotion with Model-Agnostic Linear Competitors

- Hassan Rafique, Tong Wang, Qihang Lin, Arshia Singhani
- abstract: We propose a novel type of hybrid model for multi-class classification, which utilizes competing linear models to collaborate with an existing black-box model, promoting transparency in the decision-making process. Our proposed hybrid model, Model-Agnostic Linear Competitors (MALC), brings together the interpretable power of linear models and the good predictive performance of the state-of-the-art black-box models. We formulate the training of a MALC model as a convex optimization problem, optimizing the predictive accuracy and transparency (defined as the percentage of data captured by the linear models) in the objective function. Experiments show that MALC offers more model flexibility for users to balance transparency and accuracy, in contrast to the currently available choice of either a pure black-box model or a pure interpretable model. The human evaluation also shows that more users are likely to choose MALC for this model flexibility compared with interpretable models and black-box models.

Understanding and Mitigating the Tradeoff between Robustness and Accuracy

- Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, Percy Liang
- abstract: Adversarial training augments the training set with perturbations to improve the robust error (over worst-case perturbations), but it often leads to an increase in the standard error (on unperturbed test inputs). Previous explanations for this tradeoff rely on the assumption that no predictor in the hypothesis class has low standard and robust error. In this work, we precisely characterize the effect of augmentation on the standard error in linear regression when the optimal linear predictor has zero standard and robust error. In particular, we show that the standard error could increase even when the augmented perturbations have noiseless observations from the optimal linear predictor. We then prove that the recently proposed robust self-training (RST) estimator improves robust error without sacrificing standard error for noiseless linear regression. Empirically, for neural networks, we find that RST with different adversarial training methods improves both standard and robust error for random and adversarial rotations and adversarial l_infty perturbations in CIFAR-10.

Fast Adaptation to New Environments via Policy-Dynamics Value Functions

- Roberta Raileanu, Max Goldstein, Arthur Szlam, Rob Fergus
- abstract: Standard RL algorithms assume fixed environment dynamics and require a significant amount of interaction to adapt to new environments. We introduce Policy-Dynamics Value Functions (PD-VF), a novel approach for rapidly adapting to dynamics different from those previously seen in training. PD-VF explicitly estimates the cumulative reward in a space of policies and environments. An ensemble of conventional RL policies is used to gather experience on training environments, from which embeddings of both policies and environments can be learned. Then, a value function conditioned on both embeddings is trained. At test time, a few actions are sufficient to infer the environment embedding, enabling a policy to be selected by maximizing the learned value function (which requires no additional environment interaction). We show that our method can rapidly adapt to new dynamics on a set of MuJoCo domains.

Improving Robustness of Deep-Learning-Based Image Reconstruction

- Ankit Raj, Yoram Bresler, Bo Li
- abstract: Deep-learning-based methods for various applications have been shown vulnerable to adversarial examples. Here we address the use of deep-learning networks as inverse problem solvers, which has generated much excitement and even adoption efforts by the main equipment vendors for medical imaging including computed tomography (CT) and MRI. However, the recent demonstration that such networks suffer from a similar vulnerability to adversarial attacks potentially undermines their future. We propose to modify the training strategy of end-to-end deep-learning-based inverse problem solvers to improve robustness. To this end, we introduce an auxiliary net-work to generate adversarial examples, which is used in a min-max formulation to build robust image reconstruction networks. Theoretically, we argue that for such inverse problem solvers, one should analyze and study the effect of adversaries in the measurement-space, instead of in the signal-space used in previous work. We show for a linear reconstruction scheme that our min-max formulation results in a singular-value filter regularized solution, which suppresses the effect of adversarial examples. Numerical experiments using the proposed min-max scheme confirm convergence to this solution. We complement the theory by experiments on non-linear Compressive Sensing(CS) reconstruction by a deep neural network on two standard datasets, and, using anonymized clinical data, on a state-of-the-art published algorithm for low-dose x-ray CT reconstruction. We show a significant improvement in robustness over other methods for deep network-based reconstruction, by using the proposed approach.

Multi-Precision Policy Enforced Training (MuPPET) : A Precision-Switching Strategy for Quantised Fixed-Point Training of CNNs

- Aditya Rajagopal, Diederik Vink, Stylianos Venieris, Christos-Savvas Bouganis
- abstract: Large-scale convolutional neural networks (CNNs) suffer from very long training times, spanning from hours to weeks, limiting the productivity and experimentation of deep learning practitioners. As networks grow in size and complexity, training time can be reduced through low-precision data representations and computations, however, in doing so the final accuracy suffers due to the problem of vanishing gradients. Existing state-of-the-art methods combat this issue by means of a mixed-precision approach utilising two different precision levels, FP32 (32-bit floating-point) and FP16/FP8 (16-/8-bit floating-point), leveraging the hardware support of recent GPU architectures for FP16 operations to obtain performance gains. This work pushes the boundary of quantised training by employing a multilevel optimisation approach that utilises multiple precisions including low-precision fixed-point representations resulting in a novel training strategy MuPPET; it combines the use of multiple number representation regimes together with a precision-switching mechanism that decides at run time the transition point between precision regimes. Overall, the proposed strategy tailors the training process to the hardware-level capabilities of the target hardware architecture and yields improvements in training time and energy efficiency compared to state-of-the-art approaches. Applying MuPPET on the training of AlexNet, ResNet18 and GoogLeNet on ImageNet (ILSVRC12) and targeting an NVIDIA Turing GPU, MuPPET achieves the same accuracy as standard full-precision training with training-time speedup of up to 1.84x and an average speedup of 1.58x across the networks.

A Game Theoretic Framework for Model Based Reinforcement Learning

- Aravind Rajeswaran, Igor Mordatch, Vikash Kumar
- abstract: Designing stable and efficient algorithms for model-based reinforcement learning (MBRL) with function approximation has remained challenging despite growing interest in the field. To help expose the practical challenges in MBRL and simplify algorithm design from the lens of abstraction, we develop a new framework that casts MBRL as a game between: (1) a policy player, which attempts to maximize rewards under the learned model; (2) a model player, which attempts to fit the real-world data collected by the policy player. We show that a near-optimal policy for the environment can be obtained by finding an approximate equilibrium for aforementioned game, and we develop two families of algorithms to find the game equilibrium by drawing upon ideas from Stackelberg games. Experimental studies suggest that the proposed algorithms achieve state of the art sample efficiency, match the asymptotic performance of model-free policy gradient, and scale gracefully to high-dimensional tasks like dexterous hand manipulation. Project page: \url{https://sites.google.com/view/mbrl-game}.

Closing the convergence gap of SGD without replacement

- Shashank Rajput, Anant Gupta, Dimitris Papailiopoulos
- abstract: Stochastic gradient descent without replacement sampling is widely used in practice for model training. However, the vast majority of SGD analyses assumes data is sampled with replacement, and when the function minimized is strongly convex, an $\mathcal{O}\left(\frac{1}{T}\right)$ rate can be established when SGD is run for T iterations. A recent line of breakthrough works on SGD without replacement (SGDo) established an $\mathcal{O}\left(\frac{n}{T^2}\right)$ convergence rate when the function minimized is strongly convex and is a sum of n smooth functions, and an $\mathcal{O}\left(\frac{1}{T^2}+\frac{n^3}{T^3}\right)$ rate for sums of quadratics. On the other hand, the tightest known lower bound postulates an $\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$ rate, leaving open the possibility of better SGDo convergence rates in the general case. In this paper, we close this gap and show that SGD without replacement achieves a rate of $\mathcal{O}\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$ when the sum of the functions is a quadratic, and offer a new lower bound of $\Omega\left(\frac{n}{T^2}\right)$ for strongly convex functions that are sums of smooth functions.

Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning

- Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, Adish Singla
- abstract: We study a security threat to reinforcement learning where an attacker poisons the learning environment to force the agent into executing a target policy chosen by the attacker. As a victim, we consider RL agents whose objective is to find a policy that maximizes average reward in undiscounted infinite-horizon problem settings. The attacker can manipulate the rewards or the transition dynamics in the learning environment at training-time and is interested in doing so in a stealthy manner. We propose an optimization framework for finding an \emph{optimal stealthy attack} for different measures of attack cost. We provide sufficient technical conditions under which the attack is feasible and provide lower/upper bounds on the attack cost. We instantiate our attacks in two settings: (i) an \emph{offline} setting where the agent is doing planning in the poisoned environment, and (ii) an \emph{online} setting where the agent is learning a policy using a regret-minimization framework with poisoned feedback. Our results show that the attacker can easily succeed in teaching any target policy to the victim under mild conditions and highlight a significant security threat to reinforcement learning agents in practice.

Implicit Generative Modeling for Efficient Exploration

- Neale Ratzlaff, Qinxun Bai, Li Fuxin, Wei Xu
- abstract: Efficient exploration remains a challenging problem in reinforcement learning, especially for those tasks where rewards from environments are sparse. In this work, we introduce an exploration approach based on a novel implicit generative modeling algorithm to estimate a Bayesian uncertainty of the agent's belief of the environment dynamics. Each random draw from our generative model is a neural network that instantiates the dynamic function, hence multiple draws would approximate the posterior, and the variance in the predictions based on this posterior is used as an intrinsic reward for exploration. We design a training algorithm for our generative model based on the amortized Stein Variational Gradient Descent. In experiments, we demonstrate the effectiveness of this exploration algorithm in both pure exploration tasks and a downstream task, comparing with state-of-the-art intrinsic reward-based exploration approaches, including two recent approaches based on an ensemble of dynamic models. In challenging exploration tasks, our implicit generative model consistently outperforms competing approaches regarding data efficiency in exploration.

Universal Equivariant Multilayer Perceptrons

- Siamak Ravanbakhsh
- abstract: Group invariant and equivariant Multilayer Perceptrons (MLP), also known as Equivariant Networks and Group Group Convolutional Neural Networks (G-CNN) have achieved remarkable success in learning on a variety of data structures, such as sequences, images, sets, and graphs. This paper proves the universality of a broad class of equivariant MLPs with a single hidden layer. In particular, it is shown that having a hidden layer on which the group acts regularly is sufficient for universal equivariance (invariance). For example, some types of steerable-CNN's become universal. Another corollary is the unconditional universality of equivariant MLPs for all Abelian groups. A third corollary is the universality of equivariant MLPs with a high-order hidden layer, where we give both group-agnostic bounds and group-specific bounds on the order of the hidden layer that guarantees universal equivariance.

AutoML-Zero: Evolving Machine Learning Algorithms From Scratch

- Esteban Real, Chen Liang, David So, Quoc Le
- abstract: Machine learning research has advanced in multiple aspects, including model structures and learning methods. The effort to automate such research, known as AutoML, has also made significant progress. However, this progress has largely focused on the architecture of neural networks, where it has relied on sophisticated expert-designed layers as building blocks—or similarly restrictive search spaces. Our goal is to show that AutoML can go further: it is possible today to automatically discover complete machine learning algorithms just using basic mathematical operations as building blocks. We demonstrate this by introducing a novel framework that significantly reduces human bias through a generic search space. Despite the vastness of this space, evolutionary search can still discover two-layer neural networks trained by backpropagation. These simple neural networks can then be surpassed by evolving directly on tasks of interest, e.g. CIFAR-10 variants, where modern techniques emerge in the top algorithms, such as bilinear interactions, normalized gradients, and weight averaging. Moreover, evolution adapts algorithms to different task types: e.g., dropout-like techniques appear when little data is available. We believe these preliminary successes in discovering machine learning algorithms from scratch indicate a promising new direction for the field.

Learning Human Objectives by Evaluating Hypothetical Behavior

- Siddharth Reddy, Anca Dragan, Sergey Levine, Shane Legg, Jan Leike
- abstract: We seek to align agent behavior with a user's objectives in a reinforcement learning setting with unknown dynamics, an unknown reward function, and unknown unsafe states. The user knows the rewards and unsafe states, but querying the user is expensive. We propose an algorithm that safely and efficiently learns a model of the user's reward function by posing 'what if?' questions about hypothetical agent behavior. We start with a generative model of initial states and a forward dynamics model trained on off-policy data. Our method uses these models to synthesize hypothetical behaviors, asks the user to label the behaviors with rewards, and trains a neural network to predict the rewards. The key idea is to actively synthesize the hypothetical behaviors from scratch by maximizing tractable proxies for the value of information, without interacting with the environment. We call this method reward query synthesis via trajectory optimization (ReQueST). We evaluate ReQueST with simulated users on a state-based 2D navigation task and the image-based Car Racing video game. The results show that ReQueST significantly outperforms prior methods in learning reward models that transfer to new environments with different initial state distributions. Moreover, ReQueST safely trains the reward model to detect unsafe states, and corrects reward hacking before deploying the agent.

Optimistic Bounds for Multi-output Learning

- Henry Reeve, Ata Kaban
- abstract: We investigate the challenge of multi-output learning, where the goal is to learn a vector-valued function based on a supervised data set. This includes a range of important problems in Machine Learning including multi-target regression, multi-class classification and multi-label classification. We begin our analysis by introducing the self-bounding Lipschitz condition for multi-output loss functions, which interpolates continuously between a classical Lipschitz condition and a multi-dimensional analogue of a smoothness condition. We then show that the self-bounding Lipschitz condition gives rise to optimistic bounds for multi-output learning, which attain the minimax optimal rate up to logarithmic factors. The proof exploits local Rademacher complexity combined with a powerful minoration inequality due to Srebro, Sridharan and Tewari. As an application we derive a state-of-the-art generalisation bound for multi-class gradient boosting.

Active Learning on Attributed Graphs via Graph Cognizant Logistic Regression and Preemptive Query Generation

- Florence Regol, Soumyasundar Pal, Yingxue Zhang, Mark Coates
- abstract: Node classification in attributed graphs is an important task in multiple practical settings, but it can often be difficult or expensive to obtain labels. Active learning can improve the achieved classification performance for a given budget on the number of queried labels. The best existing methods are based on graph neural networks, but they often perform poorly unless a sizeable validation set of labelled nodes is available in order to choose good hyperparameters. We propose a novel graph-based active learning algorithm for the task of node classification in attributed graphs; our algorithm uses graph cognizant logistic regression, equivalent to a linearized graph-convolutional neural network (GCN), for the prediction phase and maximizes the expected error reduction in the query phase. To reduce the delay experienced by a labeller interacting with the system, we derive a preemptive querying system that calculates a new query during the labelling process, and to address the setting where learning starts with almost no labelled data, we also develop a hybrid algorithm that performs adaptive model averaging of label propagation and linearized GCN inference. We conduct experiments on five public benchmark datasets, demonstrating a significant improvement over state-of-the-art approaches and illustrate the practical value of the method by applying it to a private microwave link network dataset.

The Sample Complexity of Best-k Items Selection from Pairwise Comparisons

- Wenbo Ren, Jia Liu, Ness Shroff
- abstract: This paper studies the sample complexity (aka number of comparisons) bounds for the active best-k items selection from pairwise comparisons. From a given set of items, the learner can make pairwise comparisons on every pair of items, and each comparison returns an independent noisy result about the preferred item. At any time, the learner can adaptively choose a pair of items to compare according to past observations (i.e., active learning). The learner's goal is to find the (approximately) best-k items with a given confidence, while trying to use as few comparisons as possible. In this paper, we study two problems: (i) finding the probably approximately correct (PAC) best-k items and (ii) finding the exact best-k items, both under strong stochastic transitivity and stochastic triangle inequality. For PAC best-k items selection, we first show a lower bound and then propose an algorithm whose sample complexity upper bound matches the lower bound up to a constant factor. For the exact best-k items selection, we first prove a worst-instance lower bound. We then propose two algorithms based on our PAC best items selection algorithms: one works for $k=1$ and is sample complexity optimal up to a loglog factor, and the other works for all values of k and is sample complexity optimal up to a log factor.

NetGAN without GAN: From Random Walks to Low-Rank Approximations

- Luca Rendsburg, Holger Heidrich, Ulrike Von Luxburg
- abstract: A graph generative model takes a graph as input and is supposed to generate new graphs that "look like" the input graph. While most classical models focus on few, hand-selected graph statistics and are too simplistic to reproduce real-world graphs, NetGAN recently emerged as an attractive alternative: by training a GAN to learn the random walk distribution of the input graph, the algorithm is able to reproduce a large number of important network patterns simultaneously, without explicitly specifying any of them. In this paper, we investigate the implicit bias of NetGAN. We find that the root

of its generalization properties does not lie in the GAN architecture, but in an inconspicuous low-rank approximation of the logits random walk transition matrix. Step by step we can strip NetGAN of all unnecessary parts, including the GAN, and obtain a highly simplified reformulation that achieves comparable generalization results, but is orders of magnitudes faster and easier to adapt. Being much simpler on the conceptual side, we reveal the implicit inductive bias of the algorithm — an important step towards increasing the interpretability, transparency and acceptance of machine learning systems.

Normalizing Flows on Tori and Spheres

- Danilo Jimenez Rezende, George Papamakarios, Sebastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, Kyle Cranmer
- abstract: Normalizing flows are a powerful tool for building expressive distributions in high dimensions. So far, most of the literature has concentrated on learning flows on Euclidean spaces. Some problems however, such as those involving angles, are defined on spaces with more complex geometries, such as tori or spheres. In this paper, we propose and compare expressive and numerically stable flows on such spaces. Our flows are built recursively on the dimension of the space, starting from flows on circles, closed intervals or spheres.

Overfitting in adversarially robust deep learning

- Leslie Rice, Eric Wong, Zico Kolter
- abstract: It is common practice in deep learning to use overparameterized networks and train for as long as possible; there are numerous studies that show, both theoretically and empirically, that such practices surprisingly do not unduly harm the generalization performance of the classifier. In this paper, we empirically study this phenomenon in the setting of adversarially trained deep networks, which are trained to minimize the loss under worst-case adversarial perturbations. We find that overfitting to the training set does in fact harm robust performance to a very large degree in adversarially robust training across multiple datasets (SVHN, CIFAR-10, CIFAR-100, and ImageNet) and perturbation models (L-infinity and L-2). Based upon this observed effect, we show that the performance gains of virtually all recent algorithmic improvements upon adversarial training can be matched by simply using early stopping. We also show that effects such as the double descent curve do still occur in adversarially trained models, yet fail to explain the observed overfitting. Finally, we study several classical and modern deep learning remedies for overfitting, including regularization and data augmentation, and find that no approach in isolation improves significantly upon the gains achieved by early stopping. All code for reproducing the experiments as well as pretrained model weights and training logs can be found at https://github.com/ locuslab/robust_overfitting.

Decentralised Learning with Random Features and Distributed Gradient Descent

- Dominic Richards, Patrick Rebeschini, Lorenzo Rosasco
- abstract: We investigate the generalisation performance of Distributed Gradient Descent with implicit regularisation and random features in the homogenous setting where a network of agents are given data sampled independently from the same unknown distribution. Along with reducing the memory footprint, random features are particularly convenient in this setting as they provide a common parameterisation across agents that allows to overcome previous difficulties in implementing decentralised kernel regression. Under standard source and capacity assumptions, we establish high probability bounds on the predictive performance for each agent as a function of the step size, number of iterations, inverse spectral gap of the communication matrix and number of random features. By tuning these parameters, we obtain statistical rates that are minimax optimal with respect to the total number of samples in the network. The algorithm provides a linear improvement over single-machine gradient descent in memory cost and, when agents hold enough data with respect to the network size and inverse spectral gap, a linear speed up in computational run-time for any network topology. We present simulations that show how the number of random features, iterations and samples impact predictive performance.

Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge

- Laura Rieger, Chandan Singh, William Murdoch, Bin Yu
- abstract: For an explanation of a deep learning model to be effective, it must provide both insight into a model and suggest a corresponding action in order to achieve some objective. Too often, the litany of proposed explainable deep learning methods stop at the first step, providing practitioners with insight into a model, but no way to act on it. In this paper, we propose contextual decomposition explanation penalization (CDEP), a method which enables practitioners to leverage existing explanation methods to increase the predictive accuracy of a deep learning model. In particular, when shown that a model has incorrectly assigned importance to some features, CDEP enables practitioners to correct these errors by inserting domain knowledge into the model via explanations. We demonstrate the ability of CDEP to increase performance on an array of toy and real datasets.

Strength from Weakness: Fast Learning Using Weak Supervision

- Joshua Robinson, Stefanie Jegelka, Suvrit Sra
- abstract: We study generalization properties of weakly supervised learning, that is, learning where only a few "strong" labels (the actual target for prediction) are present but many more "weak" labels are available. In particular, we show that pretraining using weak labels and finetuning using strong can accelerate the learning rate for the strong task to the fast rate of O(1/n), where n is the number of strongly labeled data points. This acceleration can happen even if, by itself, the strongly labeled data admits only the slower O(1/\sqrt{n}) rate. The acceleration depends continuously on the number of weak labels available, and on the relation between the two tasks. Our theoretical results are reflected empirically across a range of tasks and illustrate how weak labels speed up learning on the strong task.

On Semi-parametric Inference for BART

- Veronika Rockova
- abstract: There has been a growing realization of the potential of Bayesian machine learning as a platform that can provide both flexible modeling, accurate predictions as well as coherent uncertainty statements. In particular, Bayesian Additive Regression Trees (BART) have emerged as one of today's most effective general approaches to predictive modeling under minimal assumptions. Statistical theoretical developments for machine learning have been mostly concerned with approximability or rates of estimation when recovering infinite dimensional objects (curves or densities). Despite the impressive array of available theoretical results, the literature has been largely silent about uncertainty quantification. In this work, we continue the theoretical investigation of BART initiated recently by Rockova and van der Pas (2017). We focus on statistical inference questions. In particular, we study the Bernstein-von Mises (BvM) phenomenon (i.e. asymptotic normality) for smooth linear functionals of the regression surface within the framework of non-parametric regression with fixed covariates. Our semi-parametric BvM results show that, beyond rate-optimal estimation, BART can be also used for valid statistical inference.

FR-Train: A Mutual Information-Based Approach to Fair and Robust Training

- Yuji Roh, Kangwook Lee, Steven Whang, Changho Suh
- abstract: Trustworthy AI is a critical issue in machine learning where, in addition to training a model that is accurate, one must consider both fair and robust training in the presence of data bias and poisoning. However, the existing model fairness techniques mistakenly view poisoned data as an additional bias to be fixed, resulting in severe performance degradation. To address this problem, we propose FR-Train, which holistically performs fair and robust model training. We provide a mutual information-based interpretation of an existing adversarial training-based fairness-only method, and apply this idea to architect an additional discriminator that can identify poisoned data using a clean validation set and reduce its influence. In our experiments, FR-Train

shows almost no decrease in fairness and accuracy in the presence of data poisoning by both mitigating the bias and defending against poisoning. We also demonstrate how to construct clean validation sets using crowdsourcing, and release new benchmark datasets.

Balancing Competing Objectives with Noisy Data: Score-Based Classifiers for Welfare-Aware Machine Learning

- Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T. Liu, Daniel Bjorkegren, Moritz Hardt, Joshua Blumenstock
- abstract: While real-world decisions involve many competing objectives, algorithmic decisions are often evaluated with a single objective function. In this paper, we study algorithmic policies which explicitly trade off between a private objective (such as profit) and a public objective (such as social welfare). We analyze a natural class of policies which trace an empirical Pareto frontier based on learned scores, and focus on how such decisions can be made in noisy or data-limited regimes. Our theoretical results characterize the optimal strategies in this class, bound the Pareto errors due to inaccuracies in the scores, and show an equivalence between optimal strategies and a rich class of fairness-constrained profit-maximizing policies. We then present empirical results in two different contexts — online content recommendation and sustainable abalone fisheries — to underscore the generality of our approach to a wide range of practical decisions. Taken together, these results shed light on inherent trade-offs in using machine learning for decisions that impact social welfare.

Double-Loop Unadjusted Langevin Algorithm

- Paul Rolland, Armin Eftekhari, Ali Kavis, Volkan Cevher
- abstract: A well-known first-order method for sampling from log-concave probability distributions is the Unadjusted Langevin Algorithm (ULA). This work proposes a new annealing step-size schedule for ULA, which allows to prove new convergence guarantees for sampling from a smooth log-concave distribution, which are not covered by existing state-of-the-art convergence guarantees. To establish this result, we derive a new theoretical bound that relates the Wasserstein distance to total variation distance between any two log-concave distributions that complements the reach of Talagrand T_2 inequality. Moreover, applying this new step size schedule to an existing constrained sampling algorithm, we show state-of-the-art convergence rates for sampling from a constrained log-concave distribution, as well as improved dimension dependence.

Reverse-engineering deep ReLU networks

- David Rolnick, Konrad Kording
- abstract: The output of a neural network depends on its architecture and weights in a highly nonlinear way, and it is often assumed that a network's parameters cannot be recovered from its output. Here, we prove that, in fact, it is frequently possible to reconstruct the architecture, weights, and biases of a deep ReLU network by observing only its output. We leverage the fact that every ReLU network defines a piecewise linear function, where the boundaries between linear regions correspond to inputs for which some neuron in the network switches between inactive and active ReLU states. By dissecting the set of region boundaries into components associated with particular neurons, we show both theoretically and empirically that it is possible to recover the weights of neurons and their arrangement within the network, up to isomorphism.

Attentive Group Equivariant Convolutional Networks

- David Romero, Erik Bekkers, Jakub Tomczak, Mark Hoogendoorn
- abstract: Although group convolutional networks are able to learn powerful representations based on symmetry patterns, they lack explicit means to learn meaningful relationships among them (e.g., relative positions and poses). In this paper, we present attentive group equivariant convolutions, a generalization of the group convolution, in which attention is applied during the course of convolution to accentuate meaningful symmetry combinations and suppress non-plausible, misleading ones. We indicate that prior work on visual attention can be described as special cases of our proposed framework and show empirically that our attentive group equivariant convolutional networks consistently outperform conventional group convolutional networks on benchmark image datasets. Simultaneously, we provide interpretability to the learned concepts through the visualization of equivariant attention maps.

Finite-Time Convergence in Continuous-Time Optimization

- Orlando Romero, Mouhacine Benosman
- abstract: In this paper, we investigate a Lyapunov-like differential inequality that allows us to establish finite-time stability of a continuous-time state-space dynamical system represented via a multivariate ordinary differential equation or differential inclusion. Equipped with this condition, we successfully synthesize first and second-order dynamical systems that achieve finite-time convergence to the minima of a given sufficiently regular cost function. As a byproduct, we show that the p-rescaled gradient flow (p-RGF) proposed by Wibisono et al. (2016) is indeed finite-time convergent, provided the cost function is gradient dominated of order q in (1,p). Thus, we effectively bridge a gap between the p-RGF and the normalized gradient flow (NGF) (p=\infty) proposed by Cortes (2006) in his seminal paper in the context of multi-agent systems. We discuss strategies to discretize our proposed flows and conclude by conducting some numerical experiments to illustrate our results.

Near-optimal Regret Bounds for Stochastic Shortest Path

- Aviv Rosenberg, Alon Cohen, Yishay Mansour, Haim Kaplan
- abstract: Stochastic shortest path (SSP) is a well-known problem in planning and control, in which an agent has to reach a goal state in minimum total expected cost. In the learning formulation of the problem, the agent is unaware of the environment dynamics (i.e., the transition function) and has to repeatedly play for a given number of episodes, while learning the problem's optimal solution. Unlike other well-studied models in reinforcement learning (RL), the length of an episode is not predetermined (or bounded) and is influenced by the agent's actions. Recently, \cite{tarbouriech2019noregret} studied this problem in the context of regret minimization, and provided an algorithm whose regret bound is inversely proportional to the square root of the minimum instantaneous cost. In this work we remove this dependence on the minimum cost—we give an algorithm that guarantees a regret bound of $\widetilde{O}(B^{3/2} S \sqrt{A K})$, where B is an upper bound on the expected cost of the optimal policy, S is the number of states, A is the number of actions and K is the total number of episodes. We additionally show that any learning algorithm must have at least $\Omega(B \sqrt{S A K})$ regret in the worst case.

Predicting Choice with Set-Dependent Aggregation

- Nir Rosenfeld, Kojin Oshiba, Yaron Singer
- abstract: Providing users with alternatives to choose from is an essential component of many online platforms, making the accurate prediction of choice vital to their success. A renewed interest in learning choice models has led to improved modeling power, but most current methods are either limited in the type of choice behavior they capture, cannot be applied to large-scale data, or both. Here we propose a learning framework for predicting choice that is accurate, versatile, and theoretically grounded. Our key modeling point is that to account for how humans choose, predictive models must be expressive enough to accommodate complex choice patterns but structured enough to retain statistical efficiency. Building on recent results in economics, we derive a class of models that achieves this balance, and propose a neural implementation that allows for scalable end-to-end training. Experiments on three large choice datasets demonstrate the utility of our approach.

Certified Robustness to Label-Flipping Attacks via Randomized Smoothing

- Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, Zico Kolter
- abstract: Machine learning algorithms are known to be susceptible to data poisoning attacks, where an adversary manipulates the training data to degrade performance of the resulting classifier. In this work, we present a unifying view of randomized smoothing over arbitrary functions, and we leverage this novel characterization to propose a new strategy for building classifiers that are pointwise-certifiably robust to general data poisoning attacks. As a specific instantiation, we utilize our framework to build linear classifiers that are robust to a strong variant of label flipping, where each test example is targeted independently. In other words, for each test point, our classifier includes a certification that its prediction would be the same had some number of training labels been changed adversarially. Randomized smoothing has previously been used to guarantee—with high probability—test-time robustness to adversarial manipulation of the input to a classifier; we derive a variant which provides a deterministic, analytical bound, sidestepping the probabilistic certificates that traditionally result from the sampling subprocedure. Further, we obtain these certified bounds with minimal additional runtime complexity over standard classification and no assumptions on the train or test distributions. We generalize our results to the multi-class case, providing the first multi-class classification algorithm that is certifiably robust to label-flipping attacks.

Revisiting Training Strategies and Generalization Performance in Deep Metric Learning

- Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, Joseph Paul Cohen
- abstract: Deep Metric Learning (DML) is arguably one of the most influential lines of research for learning visual similarities with many proposed approaches every year. Although the field benefits from the rapid progress, the divergence in training protocols, architectures, and parameter choices make an unbiased comparison difficult. To provide a consistent reference point, we revisit the most widely used DML objective functions and conduct a study of the crucial parameter choices as well as the commonly neglected mini-batch sampling process. Under consistent comparison, DML objectives show much higher saturation than indicated by literature. Further based on our analysis, we uncover a correlation between the embedding space density and compression to the generalization performance of DML models. Exploiting these insights, we propose a simple, yet effective, training regularization to reliably boost the performance of ranking-based DML models on various standard benchmark datasets. Code and a publicly accessible WandB-repo are available at https://github.com/Confusezius/Revisiting_Deep_Metric_Learning_PyTorch.

FetchSGD: Communication-Efficient Federated Learning with Sketching

- Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, Raman Arora
- abstract: Existing approaches to federated learning suffer from a communication bottleneck as well as convergence issues due to sparse client participation. In this paper we introduce a novel algorithm,called FetchSGD, to overcome these challenges. FetchSGD compresses model updates using a Count Sketch, and then takes advantage of the mergeability of sketches to combine model updates from many workers. A key insight in the design of FetchSGD is that, because the Count Sketch is linear, momentum and error accumulation can both be carried out within the sketch.This allows the algorithm to move momentum and error accumulation from clients to the central aggregator, overcoming the challenges of sparse client participation while still achieving high compression rates and good convergence. We prove that FetchSGD has favorable convergence guarantees, and we demonstrate its empirical effectiveness by training two residual networks and a transformer model.

Simple and sharp analysis of k-means||

- Václav Rozhoň
- abstract: We present a simple analysis of k-means|| (Bahmani et al., PVLDB 2012) - a distributed variant of the k-means++ algorithm (Arthur and Vassilvitskii, SODA 2007). Moreover, the bound on the number of rounds is improved from $O(\log n)$ to $O(\log n / \log\log n)$, which we show to be tight.

Bayesian Optimisation over Multiple Continuous and Categorical Inputs

- Binxin Ru, Ahsan Alvi, Vu Nguyen, Michael A. Osborne, Stephen Roberts
- abstract: Efficient optimisation of black-box problems that comprise both continuous and categorical inputs is important, yet poses significant challenges. Current approaches, like one-hot encoding, severely increase the dimension of the search space, while separate modelling of category-specific data is sample-inefficient. Both frameworks are not scalable to practical applications involving multiple categorical variables, each with multiple possible values. We propose a new approach, Continuous and Categorical Bayesian Optimisation (CoCaBO), which combines the strengths of multi-armed bandits and Bayesian optimisation to select values for both categorical and continuous inputs. We model this mixed-type space using a Gaussian Process kernel, designed to allow sharing of information across multiple categorical variables; this allows CoCaBO to leverage all available data efficiently. We extend our method to the batch setting and propose an efficient selection procedure that dynamically balances exploration and exploitation whilst encouraging batch diversity. We demonstrate empirically that our method outperforms existing approaches on both synthetic and real-world optimisation tasks with continuous and categorical inputs.

Inter-domain Deep Gaussian Processes

- Tim G. J. Rudner, Dino Sejdinovic, Yarin Gal
- abstract: Inter-domain Gaussian processes (GPs) allow for high flexibility and low computational cost when performing approximate inference in GP models. They are particularly suitable for modeling data exhibiting global structure but are limited to stationary covariance functions and thus fail to model non-stationary data effectively. We propose Inter-domain Deep Gaussian Processes, an extension of inter-domain shallow GPs that combines the advantages of inter-domain and deep Gaussian processes (DGPs), and demonstrate how to leverage existing approximate inference methods to perform simple and scalable approximate inference using inter-domain features in DGPs. We assess the performance of our method on a range of regression tasks and demonstrate that it outperforms inter-domain shallow GPs and conventional DGPs on challenging large-scale real-world datasets exhibiting both global structure as well as a high-degree of non-stationarity.

Bio-Inspired Hashing for Unsupervised Similarity Search

- Chaitanya Ryali, John Hopfield, Leopold Grinberg, Dmitry Krotov
- abstract: The fruit fly Drosophila’s olfactory circuit has inspired a new locality sensitive hashing (LSH) algorithm, FlyHash. In contrast with classical LSH algorithms that produce low dimensional hash codes, FlyHash produces sparse high-dimensional hash codes and has also been shown to have superior empirical performance compared to classical LSH algorithms in similarity search. However, FlyHash uses random projections and cannot learn from data. Building on inspiration from FlyHash and the ubiquity of sparse expansive representations in neurobiology, our work proposes a novel hashing algorithm BioHash that produces sparse high dimensional hash codes in a data-driven manner. We show that BioHash outperforms previously published benchmarks for various hashing methods. Since our learning algorithm is based on a local and biologically plausible synaptic plasticity rule, our work provides evidence for the proposal that LSH might be a computational reason for the abundance of sparse expansive motifs in a variety of biological systems. We also propose a convolutional variant BioConvHash that further improves performance. From the perspective of computer science, BioHash and BioConvHash are fast, scalable and yield compressed binary representations that are useful for similarity search.

Adversarial Attacks on Copyright Detection Systems

- Parsa Saadatpanah, Ali Shafahi, Tom Goldstein
- abstract: It is well-known that many machine learning models are susceptible to adversarial attacks, in which an attacker evades a classifier by making small perturbations to inputs. This paper discusses how industrial copyright detection tools, which serve a central role on the web, are susceptible to adversarial attacks. As proof of concept, we describe a well-known music identification method and implement this system in the form of a neural net. We then attack this system using simple gradient methods and show that it is easily broken with white-box attacks. By scaling these perturbations up, we can create transfer attacks on industrial systems, such as the AudioTag copyright detector and YouTube's Content ID system, using perturbations that are audible but significantly smaller than a random baseline. Our goal is to raise awareness of the threats posed by adversarial examples in this space and to highlight the importance of hardening copyright detection systems to attacks.

Bounding the fairness and accuracy of classifiers from population statistics

- Sivan Sabato, Elad Yom-Tov
- abstract: We consider the study of a classification model whose properties are impossible to estimate using a validation set, either due to the absence of such a set or because access to the classifier, even as a black-box, is impossible. Instead, only aggregate statistics on the rate of positive predictions in each of several sub-populations are available, as well as the true rates of positive labels in each of these sub-populations. We show that these aggregate statistics can be used to lower-bound the discrepancy of a classifier, which is a measure that balances inaccuracy and unfairness. To this end, we define a new measure of unfairness, equal to the fraction of the population on which the classifier behaves differently, compared to its global, ideally fair behavior, as defined by the measure of equalized odds. We propose an efficient and practical procedure for finding the best possible lower bound on the discrepancy of the classifier, given the aggregate statistics, and demonstrate in experiments the empirical tightness of this lower bound, as well as its possible uses on various types of problems, ranging from estimating the quality of voting polls to measuring the effectiveness of patient identification from internet search queries. The code and data are available at https://github.com/sivansabato/bfa.

Radioactive data: tracing through training

- Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Herve Jegou
- abstract: Data tracing determines whether particular data samples have been used to train a model. We propose a new technique, radioactive data, that makes imperceptible changes to these samples such that any model trained on them will bear an identifiable mark. Given a trained model, our technique detects the use of radioactive data and provides a level of confidence (p-value). Experiments on large-scale benchmarks (Imagenet), with standard architectures (Resnet-18, VGG-16, Densenet-121) and training procedures, show that we detect radioactive data with high confidence ($p<0.0001$) when only 1% of the data used to train a model is radioactive. Our radioactive mark is resilient to strong data augmentations and variations of the model architecture. As a result, it offers a much higher signal-to-noise ratio than data poisoning and backdoor methods.

Causal Structure Discovery from Distributions Arising from Mixtures of DAGs

- Basil Saeed, Snigdha Panigrahi, Caroline Uhler
- abstract: We consider distributions arising from a mixture of causal models, where each model is represented by a directed acyclic graph (DAG). We provide a graphical representation of such mixture distributions and prove that this representation encodes the conditional independence relations of the mixture distribution. We then consider the problem of structure learning based on samples from such distributions. Since the mixing variable is latent, we consider causal structure discovery algorithms such as FCI that can deal with latent variables. We show that such algorithms recover a "union" of the component DAGs and can identify variables whose conditional distribution across the component DAGs vary. We demonstrate our results on synthetic and real data showing that the inferred graph identifies nodes that vary between the different mixture components. As an immediate application, we demonstrate how retrieval of this causal information can be used to cluster samples according to each mixture component.

An Investigation of Why Overparameterization Exacerbates Spurious Correlations

- Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, Percy Liang
- abstract: We study why overparameterization—increasing model size well beyond the point of zero training error—can hurt test error on minority groups despite improving average test error when there are spurious correlations in the data. Through simulations and experiments on two image datasets, we identify two key properties of the training data that drive this behavior: the proportions of majority versus minority groups, and the signal-to-noise ratio of the spurious correlations. We then analyze a linear setting and theoretically show how the inductive bias of models towards "memorizing" fewer examples can cause overparameterization to hurt. Our analysis leads to a counterintuitive approach of subsampling the majority group, which empirically achieves low minority error in the overparameterized regime, even though the standard approach of upweighting the minority fails. Overall, our results suggest a tension between using overparameterized models versus using all the training data for achieving low worst-group error.

Improved Sleeping Bandits with Stochastic Action Sets and Adversarial Rewards

- Aadirupa Saha, Pierre Gaillard, Michal Valko
- abstract: In this paper, we consider the problem of sleeping bandits with stochastic action sets and adversarial rewards. In this setting, in contrast to most work in bandits, the actions may not be available at all times. For instance, some products might be out of stock in item recommendation. The best existing efficient (i.e., polynomial-time) algorithms for this problem only guarantee an $O(T^{2/3})$ upper-bound on the regret. Yet, inefficient algorithms based on EXP4 can achieve $O(\sqrt{T})$. In this paper, we provide a new computationally efficient algorithm inspired by EXP3 satisfying a regret of order $O(\sqrt{T})$ when the availabilities of each action $i \in \cA$ are independent. We then study the most general version of the problem where at each round available sets are generated from some unknown arbitrary distribution (i.e., without the independence assumption) and propose an efficient algorithm with $O(\sqrt {2^K T})$ regret guarantee. Our theoretical results are corroborated with experimental evaluations.

From PAC to Instance-Optimal Sample Complexity in the Plackett-Luce Model

- Aadirupa Saha, Aditya Gopalan
- abstract: We consider PAC learning a good item from k-subsetwise feedback sampled from a Plackett-Luce probability model, with instance-dependent sample complexity performance. In the setting where subsets of a fixed size can be tested and top-ranked feedback is made available to the learner, we give an optimal instance-dependent algorithm with a sample complexity bound for PAC best arm identification algorithm of $O\bigg(\frac{\Theta_{[k]}}{k}\sum_{i = 2}^n\max\Big(1,\frac{1}{\Delta_i^2}\Big) \ln\frac{k}{\delta}\Big(\ln \frac{1}{\Delta_i}\Big)\bigg)$, Δ_i being the Plackett-Luce parameter gap between the best and the i^{th} best item, and $\Theta_{[k]}$ is the sum of the Plackett-Luce parameters for top-k items. The algorithm is based on a wrapper around a PAC winner-finding algorithm with weaker performance guarantees to adapt to the hardness of the input instance. The sample complexity is also shown to be multiplicatively better depending on the length of rank-ordered feedback available in each subset-wise play. We show optimality of our algorithms with matching sample complexity lower bounds. We next address the winner-finding problem in Plackett-Luce models in the fixed-budget setting with instance dependent upper and lower bounds on the misidentification probability, of $\Omega\left(\exp(-2 \tilde \Delta Q) \right)$ for a given budget Q, where $\tilde \Delta$ is an explicit instance-dependent problem complexity parameter. Numerical performance results are also reported for the algorithms.

Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics

- Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, Michael Tschantz
- abstract: Bias in machine learning has manifested injustice in several areas, such as medicine, hiring, and criminal justice. In response, computer scientists have developed myriad definitions of fairness to correct this bias in fielded algorithms. While some definitions are based on established legal and ethical norms, others are largely mathematical. It is unclear whether the general public agrees with these fairness definitions, and perhaps more importantly, whether they understand these definitions. We take initial steps toward bridging this gap between ML researchers and the public, by addressing the question: does a lay audience understand a basic definition of ML fairness? We develop a metric to measure comprehension of three such definitions– demographic parity, equal opportunity, and equalized odds. We evaluate this metric using an online survey, and investigate the relationship between comprehension and sentiment, demographics, and the definition itself.

From Sets to Multisets: Provable Variational Inference for Probabilistic Integer Submodular Models

- Aytunc Sahin, Yatao Bian, Joachim Buhmann, Andreas Krause
- abstract: Submodular functions have been studied extensively in machine learning and data mining. In particular, the optimization of submodular functions over the integer lattice (integer submodular functions) has recently attracted much interest, because this domain relates naturally to many practical problem settings, such as multilabel graph cut, budget allocation and revenue maximization with discrete assignments. In contrast, the use of these functions for probabilistic modeling has received surprisingly little attention so far. In this work, we firstly propose the Generalized Multilinear Extension, a continuous DR-submodular extension for integer submodular functions. We study central properties of this extension and formulate a new probabilistic model which is defined through integer submodular functions. Then, we introduce a block-coordinate ascent algorithm to perform approximate inference for this class of models and finally, we demonstrate its effectiveness and viability on several real-world social connection graph datasets with integer submodular objectives.

Counterfactual Cross-Validation: Stable Model Selection Procedure for Causal Inference Models

- Yuta Saito, Shota Yasui
- abstract: We study the model selection problem in \emph{conditional average treatment effect} (CATE) prediction. Unlike previous works on this topic, we focus on preserving the rank order of the performance of candidate CATE predictors to enable accurate and stable model selection. To this end, we analyze the model performance ranking problem and formulate guidelines to obtain a better evaluation metric. We then propose a novel metric that can identify the ranking of the performance of CATE predictors with high confidence. Empirical evaluations demonstrate that our metric outperforms existing metrics in both model selection and hyperparameter tuning tasks.

Inferring DQN structure for high-dimensional continuous control

- Andrey Sakryukin, Chedy Raissi, Mohan Kankanhalli
- abstract: Despite recent advancements in the field of Deep Reinforcement Learning, Deep Q-network (DQN) models still show lackluster performance on problems with high-dimensional action spaces. The problem is even more pronounced for cases with high-dimensional continuous action spaces due to a combinatorial increase in the number of the outputs. Recent works approach the problem by dividing the network into multiple parallel or sequential (action) modules responsible for different discretized actions. However, there are drawbacks to both the parallel and the sequential approaches. Parallel module architectures lack coordination between action modules, leading to extra complexity in the task, while a sequential structure can result in the vanishing gradients problem and exploding parameter space. In this work, we show that the compositional structure of the action modules has a significant impact on model performance. We propose a novel approach to infer the network structure for DQN models operating with high-dimensional continuous actions. Our method is based on the uncertainty estimation techniques introduced in the paper. Our approach achieves state-of-the-art performance on MuJoCo environments with high-dimensional continuous action spaces. Furthermore, we demonstrate the improvement of the introduced approach on a realistic AAA sailing simulator game.

The Performance Analysis of Generalized Margin Maximizers on Separable Data

- Fariborz Salehi, Ehsan Abbasi, Babak Hassibi
- abstract: Logistic models are commonly used for binary classification tasks. The success of such models has often been attributed to their connection to maximum-likelihood estimators. It has been shown that gradient descent algorithm, when applied on the logistic loss, converges to the max-margin classifier (a.k.a. hard-margin SVM). The performance of the max-margin classifier has been recently analyzed in \cite{montanari2019generalization, deng2019model}. Inspired by these results, in this paper, we present and study a more general setting, where the underlying parameters of the logistic model possess certain structures (sparse, block-sparse, low-rank, etc.) and introduce a more general framework (which is referred to as “Generalized Margin Maximizer”, GMM). While classical max-margin classifiers minimize the 2-norm of the parameter vector subject to linearly separating the data, GMM minimizes any arbitrary convex function of the parameter vector. We provide a precise analysis of the performance of GMM via the solution of a system of nonlinear equations. We also provide a detailed study for three special cases: (1) ℓ_2-GMM that is the max-margin classifier, (2) ℓ_1-GMM which encourages sparsity, and (3) ℓ_{∞}-GMM which is often used when the parameter vector has binary entries. Our theoretical results are validated by extensive simulation results across a range of parameter values, problem instances, and model structures.

Stochastic Coordinate Minimization with Progressive Precision for Stochastic Convex Optimization

- Sudeep Salgia, Qing Zhao, Sattar Vakili
- abstract: A framework based on iterative coordinate minimization (CM) is developed for stochastic convex optimization. Given that exact coordinate minimization is impossible due to the unknown stochastic nature of the objective function, the crux of the proposed optimization algorithm is an optimal control of the minimization precision in each iteration. We establish the optimal precision control and the resulting order-optimal regret performance for strongly convex and separably nonsmooth functions. An interesting finding is that the optimal progression of precision across iterations is independent of the low-dimension CM routine employed, suggesting a general framework for extending low-dimensional optimization routines to high-dimensional problems. The proposed algorithm is amenable to online implementation and inherits the scalability and parallelizability properties of CM for large-scale optimization. Requiring only a sublinear order of message exchanges, it also lends itself well to distributed computing as compared with the alternative approach of coordinate gradient descent.

A Quantile-based Approach for Hyperparameter Transfer Learning

- David Salinas, Huibin Shen, Valerio Perrone
- abstract: Bayesian optimization (BO) is a popular methodology to tune the hyperparameters of expensive black-box functions. Traditionally, BO focuses on a single task at a time and is not designed to leverage information from related functions, such as tuning performance objectives of the same algorithm across multiple datasets. In this work, we introduce a novel approach to achieve transfer learning across different datasets as well as different objectives. The main idea is to regress the mapping from hyperparameter to objective quantiles with a semi-parametric Gaussian Copula distribution, which provides robustness against different scales or outliers that can occur in different tasks. We introduce two methods to leverage this estimation: a Thompson sampling strategy as well as a Gaussian Copula process using such quantile estimate as a prior. We show that these strategies can combine the estimation of multiple objectives such as latency and accuracy, steering the optimization toward faster predictions for the same level of accuracy. Experiments on an extensive set of hyperparameter tuning tasks demonstrate significant improvements over state-of-the-art methods for both hyperparameter optimization and neural architecture search.

Spectral Subsampling MCMC for Stationary Time Series

- Robert Salomone, Matias Quiroz, Robert Kohn, Mattias Villani, Minh-Ngoc Tran
- abstract: Bayesian inference using Markov Chain Monte Carlo (MCMC) on large datasets has developed rapidly in recent years. However, the underlying methods are generally limited to relatively simple settings where the data have specific forms of independence. We propose a novel technique for speeding up MCMC for time series data by efficient data subsampling in the frequency domain. For several challenging time series models, we demonstrate a speedup of up to two orders of magnitude while incurring negligible bias compared to MCMC on the full dataset. We also propose alternative control variates for variance reduction based on data grouping and coreset constructions.

Learning to Simulate Complex Physics with Graph Networks

- Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, Peter Battaglia
- abstract: Here we present a machine learning framework and model implementation that can learn to simulate a wide variety of challenging physical domains, involving fluids, rigid solids, and deformable materials interacting with one another. Our framework—which we term "Graph Network-based Simulators" (GNS)—represents the state of a physical system with particles, expressed as nodes in a graph, and computes dynamics via learned message-passing. Our results show that our model can generalize from single-timestep predictions with thousands of particles during training, to different initial conditions, thousands of timesteps, and at least an order of magnitude more particles at test time. Our model was robust to hyperparameter choices across various evaluation metrics: the main determinants of long-term performance were the number of message-passing steps, and mitigating the accumulation of error by corrupting the training data with noise. Our GNS framework advances the state-of-the-art in learned physical simulation, and holds promise for solving a wide range of complex forward and inverse problems.

The Impact of Neural Network Overparameterization on Gradient Confusion and Stochastic Gradient Descent

- Karthik Abinav Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, Tom Goldstein
- abstract: This paper studies how neural network architecture affects the speed of training. We introduce a simple concept called gradient confusion to help formally analyze this. When gradient confusion is high, stochastic gradients produced by different data samples may be negatively correlated, slowing down convergence. But when gradient confusion is low, data samples interact harmoniously, and training proceeds quickly. Through theoretical and experimental results, we demonstrate how the neural network architecture affects gradient confusion, and thus the efficiency of training. Our results show that, for popular initialization techniques, increasing the width of neural networks leads to lower gradient confusion, and thus faster model training. On the other hand, increasing the depth of neural networks has the opposite effect. Our results indicate that alternate initialization techniques or networks using both batch normalization and skip connections help reduce the training burden of very deep networks.

Explicit Gradient Learning for Black-Box Optimization

- Elad Sarafian, Mor Sinay, Yoram Louzoun, Noa Agmon, Sarit Kraus
- abstract: Black-Box Optimization (BBO) methods can find optimal policies for systems that interact with complex environments with no analytical representation. As such, they are of interest in many Artificial Intelligence (AI) domains. Yet classical BBO methods fall short in high-dimensional non-convex problems. They are thus often overlooked in real-world AI tasks. Here we present a BBO method, termed Explicit Gradient Learning (EGL), that is designed to optimize high-dimensional ill-behaved functions. We derive EGL by finding weak spots in methods that fit the objective function with a parametric Neural Network (NN) model and obtain the gradient signal by calculating the parametric gradient. Instead of fitting the function, EGL trains a NN to estimate the objective gradient directly. We prove the convergence of EGL to a stationary point and its robustness in the optimization of integrable functions. We evaluate EGL and achieve state-of-the-art results in two challenging problems: (1) the COCO test suite against an assortment of standard BBO methods; and (2) in a high-dimensional non-convex image generation task.

Detecting Out-of-Distribution Examples with Gram Matrices

- Chandramouli Shama Sastry, Sageev Oore
- abstract: When presented with Out-of-Distribution (OOD) examples, deep neural networks yield confident, incorrect predictions; detecting OOD examples is challenging, and the potential risks are high. In this paper, we propose to detect OOD examples by identifying inconsistencies between activity patterns and predicted class. We find that characterizing activity patterns by Gram matrices and identifying anomalies in Gram matrix values can yield high OOD detection rates. We identify anomalies in the Gram matrices by simply comparing each value with its respective range observed over the training data. Unlike many approaches, this can be used with any pre-trained softmax classifier and neither requires access to OOD data for fine-tuning hyperparameters, nor does it require OOD access for inferring parameters. We empirically demonstrate applicability across a variety of architectures and vision datasets and, for the important and surprisingly hard task of detecting far out-of-distribution examples, it generally performs better than or equal to state-of-the-art OOD detection methods (including those that do assume access to OOD examples).

Constrained Markov Decision Processes via Backward Value Functions

- Harsh Satija, Philip Amortila, Joelle Pineau
- abstract: Although Reinforcement Learning (RL) algorithms have found tremendous success in simulated domains, they often cannot directly be applied to physical systems, especially in cases where there are hard constraints to satisfy (e.g. on safety or resources). In standard RL, the agent is incentivized to explore any behavior as long as it maximizes rewards, but in the real world, undesired behavior can damage either the system or the agent in a way that breaks the learning process itself. In this work, we model the problem of learning with constraints as a Constrained Markov Decision Process and provide a new on-policy formulation for solving it. A key contribution of our approach is to translate cumulative cost constraints into state-based constraints. Through this, we define a safe policy improvement method which maximizes returns while ensuring that the constraints are satisfied at every step. We provide theoretical guarantees under which the agent converges while ensuring safety over the course of training. We also highlight the computational advantages of this approach. The effectiveness of our approach is demonstrated on safe navigation tasks and in safety-constrained versions of MuJoCo environments, with deep neural networks.

A Sample Complexity Separation between Non-Convex and Convex Meta-Learning

- Nikunj Saunshi, Yi Zhang, Mikhail Khodak, Sanjeev Arora
- abstract: One popular trend in meta-learning is to learn from many training tasks a common initialization that a gradient-based method can use to solve a new task with few samples. The theory of meta-learning is still in its early stages, with several recent learning-theoretic analyses of methods such as Reptile [Nichol et al., 2018] being for \emph{convex models}. This work shows that convex-case analysis might be insufficient to understand the success of meta-learning, and that even for non-convex models it is important to look inside the optimization black-box, specifically at properties of the optimization trajectory. We construct a simple meta-learning instance that captures the problem of one-dimensional subspace learning. For the convex formulation of linear regression on this instance, we show that the new task sample complexity of any \emph{initialization-based meta-learning} algorithm is $\Omega(d)$, where d is the input dimension. In contrast, for the non-convex formulation of a two layer linear network on the same instance, we show that both Reptile and multi-task representation learning can have new task sample complexity of $O(1)$, demonstrating a separation

from convex meta-learning. Crucially, analyses of the training dynamics of these methods reveal that they can meta-learn the correct subspace onto which the data should be projected.

Harmonic Decompositions of Convolutional Networks

- Meyer Scetbon, Zaid Harchaoui
- abstract: We present a description of the function space and the smoothness class associated with a convolutional network using the machinery of reproducing kernel Hilbert spaces. We show that the mapping associated with a convolutional network expands into a sum involving elementary functions akin to spherical harmonics. This functional decomposition can be related to the functional ANOVA decomposition in nonparametric statistics. Building off our functional characterization of convolutional networks, we obtain statistical bounds highlighting an interesting trade-off between the approximation error and the estimation error.

Implicit competitive regularization in GANs

- Florian Schaefer, Hongkai Zheng, Animashree Anandkumar
- abstract: The success of GANs is usually attributed to properties of the divergence obtained by an optimal discriminator. In this work we show that this approach has a fundamental flaw:\{If} we do not impose regularity of the discriminator, it can exploit visually imperceptible errors of the generator to always achieve the maximal generator loss. In practice, gradient penalties are used to regularize the discriminator. However, this needs a metric on the space of images that captures visual similarity. Such a metric is not known, which explains the limited success of gradient penalties in stabilizing GANs.\{Instead}, we argue that the implicit competitive regularization (ICR) arising from the simultaneous optimization of generator and discriminator enables GANs performance. We show that opponent-aware modelling of generator and discriminator, as present in competitive gradient descent (CGD), can significantly strengthen ICR and thus stabilize GAN training without explicit regularization. In our experiments, we use an existing implementation of WGAN-GP and show that by training it with CGD without any explicit regularization, we can improve the inception score (IS) on CIFAR10, without any hyperparameter tuning.

Off-Policy Actor-Critic with Shared Experience Replay

- Simon Schmitt, Matteo Hessel, Karen Simonyan
- abstract: We investigate the combination of actor-critic reinforcement learning algorithms with a uniform large-scale experience replay and propose solutions for two ensuing challenges: (a) efficient actor-critic learning with experience replay (b) the stability of off-policy learning where agents learn from other agents behaviour. To this end we analyze the bias-variance tradeoffs in V-trace, a form of importance sampling for actor-critic methods. Based on our analysis, we then argue for mixing experience sampled from replay with on-policy experience, and propose a new trust region scheme that scales effectively to data distributions where V-trace becomes unstable. We provide extensive empirical validation of the proposed solutions on DMLab-30 and further show the benefits of this setup in two training regimes for Atari: (1) a single agent is trained up until 200M environment frames per game (2) a population of agents is trained up until 200M environment frames each and may share experience. We demonstrate state-of-the-art data efficiency among model-free agents in both regimes.

Discriminative Adversarial Search for Abstractive Summarization

- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano
- abstract: We introduce a novel approach for sequence decoding, Discriminative Adversarial Search (DAS), which has the desirable properties of alleviating the effects of exposure bias without requiring external metrics. Inspired by Generative Adversarial Networks (GANs), wherein a discriminator is used to improve the generator, our method differs from GANs in that the generator parameters are not updated at training time and the discriminator is used to drive sequence generation at inference time. We investigate the effectiveness of the proposed approach on the task of Abstractive Summarization: the results obtained show that a naive application of DAS improves over the state-of-the-art methods, with further gains obtained via discriminator retraining. Moreover, we show how DAS can be effective for cross-domain adaptation. Finally, all results reported are obtained without additional rule-based filtering strategies, commonly used by the best performing systems available: this indicates that DAS can effectively be deployed without relying on post-hoc modifications of the generated outputs.

Universal Average-Case Optimality of Polyak Momentum

- Damien Scieur, Fabian Pedregosa
- abstract: Polyak momentum (PM), also known as the heavy-ball method, is a widely used optimization method that enjoys an asymptotic optimal worst-case complexity on quadratic objectives. However, its remarkable empirical success is not fully explained by this optimality, as the worst-case analysis – contrary to the average-case– is not representative of the expected complexity of an algorithm. In this work we establish a novel link between PM and the average-case analysis. Our main contribution is to prove that any optimal average-case method converges in the number of iterations to PM, under mild assumptions. This brings a new perspective on this classical method, showing that PM is asymptotically both worst-case and average-case optimal.

Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures

- Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, Romain Couillet
- abstract: This paper shows that deep learning (DL) representations of data produced by generative adversarial nets (GANs) are random vectors which fall within the class of so-called \emph{concentrated} random vectors. Further exploiting the fact that Gram matrices, of the type $G = X^\intercal X$ with $X=[x_1,\ldots,x_n]\in \mathbb{R}^{p\times n}$ and x_i independent concentrated random vectors from a mixture model, behave asymptotically (as $n,p\to \infty$) as if the x_i were drawn from a Gaussian mixture, suggests that DL representations of GAN-data can be fully described by their first two statistical moments for a wide range of standard classifiers. Our theoretical findings are validated by generating images with the BigGAN model and across different popular deep representation networks.

Planning to Explore via Self-Supervised World Models

- Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, Deepak Pathak
- abstract: Reinforcement learning allows solving complex tasks, however, the learning tends to be task-specific and the sample efficiency remains a challenge. We present Plan2Explore, a self-supervised reinforcement learning agent that tackles both these challenges through a new approach to self-supervised exploration and fast adaptation to new tasks, which need not be known during exploration. During exploration, unlike prior methods which retrospectively compute the novelty of observations after the agent has already reached them, our agent acts efficiently by leveraging planning to seek out expected future novelty. After exploration, the agent quickly adapts to multiple downstream tasks in a zero or a few-shot manner. We evaluate on challenging control tasks from high-dimensional image inputs. Without any training supervision or task-specific interaction, Plan2Explore outperforms prior self-supervised exploration methods, and in fact, almost matches the performances oracle which has access to rewards. Videos and code: https://ramanans1.github.io/plan2explore/

An Explicitly Relational Neural Network Architecture

- Murray Shanahan, Kyriacos Nikiforou, Antonia Creswell, Christos Kaplanis, David Barrett, Marta Garnelo
- abstract: With a view to bridging the gap between deep learning and symbolic AI, we present a novel end-to-end neural network architecture that learns to form propositional representations with an explicitly relational structure from raw pixel data. In order to evaluate and analyse the architecture, we introduce a family of simple visual relational reasoning tasks of varying complexity. We show that the proposed architecture, when pre-trained on a curriculum of such tasks, learns to generate reusable representations that better facilitate subsequent learning on previously unseen tasks when compared to a number of baseline architectures. The workings of a successfully trained model are visualised to shed some light on how the architecture functions.

Optimistic Policy Optimization with Bandit Feedback

- Lior Shani, Yonathan Efroni, Aviv Rosenberg, Shie Mannor
- abstract: Policy optimization methods are one of the most widely used classes of Reinforcement Learning (RL) algorithms. Yet, so far, such methods have been mostly analyzed from an optimization perspective, without addressing the problem of exploration, or by making strong assumptions on the interaction with the environment. In this paper we consider model-based RL in the tabular finite-horizon MDP setting with unknown transitions and bandit feedback. For this setting, we propose an optimistic trust region policy optimization (TRPO) algorithm for which we establish $\tilde O(\sqrt{S^2 A H^4 K})$ regret for stochastic rewards. Furthermore, we prove $\tilde O(\sqrt{ S^2 A H^4 } K^{2/3}) $ regret for adversarial rewards. Interestingly, this result matches previous bounds derived for the bandit feedback case, yet with known transitions. To the best of our knowledge, the two results are the first sub-linear regret bounds obtained for policy optimization algorithms with unknown transitions and bandit feedback.

Neural Kernels Without Tangents

- Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Jonathan Ragan-Kelley, Ludwig Schmidt, Benjamin Recht
- abstract: We investigate the connections between neural networks and simple building blocks in kernel space. In particular, using well established feature space tools such as direct sum, averaging, and moment lifting, we present an algebra for creating “compositional” kernels from bags of features. We show that these operations correspond to many of the building blocks of “neural tangent kernels (NTK)”. Experimentally, we show that there is a correlation in test error between neural network architectures and the associated kernels. We construct a simple neural network architecture using only 3x3 convolutions, 2x2 average pooling, ReLU, and optimized with SGD and MSE loss that achieves 96% accuracy on CIFAR10, and whose corresponding compositional kernel achieves 90% accuracy. We also use our constructions to investigate the relative performance of neural networks, NTKs, and compositional kernels in the small dataset regime. In particular, we find that compositional kernels outperform NTKs and neural networks outperform both kernel methods.

Learning Robot Skills with Temporal Variational Inference

- Tanmay Shankar, Abhinav Gupta
- abstract: In this paper, we address the discovery of robotic options from demonstrations in an unsupervised manner. Specifically, we present a framework to jointly learn low-level control policies and higher-level policies of how to use them from demonstrations of a robot performing various tasks. By representing options as continuous latent variables, we frame the problem of learning these options as latent variable inference. We then present a temporally causal variant of variational inference based on a temporal factorization of trajectory likelihoods, that allows us to infer options in an unsupervised manner. We demonstrate the ability of our framework to learn such options across three robotic demonstration datasets, and provide our code.

Evaluating Machine Accuracy on ImageNet

- Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, Ludwig Schmidt
- abstract: We evaluate a wide range of ImageNet models with five trained human labelers. In our year-long experiment, trained humans first annotated 40,000 images from the ImageNet and ImageNetV2 test sets with multi-class labels to enable a semantically coherent evaluation. Then we measured the classification accuracy of the five trained humans on the full task with 1,000 classes. Only the latest models from 2020 are on par with our best human labeler, and human accuracy on the 590 object classes is still 4% and 10% higher than the best model on ImageNet and ImageNetV2, respectively. Moreover, humans achieve the same accuracy on ImageNet and ImageNetV2, while all models see a consistent accuracy drop. Overall, our results show that there is still substantial room for improvement on ImageNet and direct accuracy comparisons between humans and machines may overstate machine performance.

Channel Equilibrium Networks for Learning Deep Representation

- Wenqi Shao, Shitao Tang, Xingang Pan, Ping Tan, Xiaogang Wang, Ping Luo
- abstract: Convolutional Neural Networks (CNNs) are typically constructed by stacking multiple building blocks, each of which contains a normalization layer such as batch normalization (BN) and a rectified linear function such as ReLU. However, this work shows that the combination of normalization and rectified linear function leads to inhibited channels, which have small magnitude and contribute little to the learned feature representation, impeding the generalization ability of CNNs. Unlike prior arts that simply removed the inhibited channels, we propose to “wake them up” during training by designing a novel neural building block, termed Channel Equilibrium (CE) block, which enables channels at the same layer to contribute equally to the learned representation. We show that CE is able to prevent inhibited channels both empirically and theoretically. CE has several appealing benefits. (1) It can be integrated into many advanced CNN architectures such as ResNet and MobileNet, outperforming their original networks. (2) CE has an interesting connection with the Nash Equilibrium, a well-known solution of a non-cooperative game. (3) Extensive experiments show that CE achieves state-of-the-art performance on various challenging benchmarks such as ImageNet and COCO.

ControlVAE: Controllable Variational Autoencoder

- Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, Tarek Abdelzaher
- abstract: Variational Autoencoders (VAE) and their variants have been widely used in a variety of applications, such as dialog generation, image generation and disentangled representation learning. However, the existing VAE models may suffer from KL vanishing in language modeling and low reconstruction quality for disentangling. To address these issues, we propose a novel controllable variational autoencoder framework, ControlVAE, that combines a controller, inspired by automatic control theory, with the basic VAE to improve the performance of resulting generative models. Specifically, we design a new non-linear PI controller, a variant of the proportional-integral-derivative (PID) control, to automatically tune the hyperparameter (weight) added in the VAE objective using the output KL-divergence as feedback during model training. The framework is evaluated using three applications; namely, language modeling, disentangled representation learning, and image generation. The results show that ControlVAE can achieve much better reconstruction quality than the competitive methods for the comparable disentanglement performance. For language modeling, it not only averts the KL-vanishing, but also improves the diversity of generated text. Finally, we also demonstrate that ControlVAE improves the reconstruction quality for image generation compared to the original VAE.

Lookahead-Bounded Q-learning

- Ibrahim El Shar, Daniel Jiang

- abstract: We introduce the lookahead-bounded Q-learning (LBQL) algorithm, a new, provably convergent variant of Q-learning that seeks to improve the performance of standard Q-learning in stochastic environments through the use of “lookahead” upper and lower bounds. To do this, LBQL employs previously collected experience and each iteration’s state-action values as dual feasible penalties to construct a sequence of sampled information relaxation problems. The solutions to these problems provide estimated upper and lower bounds on the optimal value, which we track via stochastic approximation. These quantities are then used to constrain the iterates to stay within the bounds at every iteration. Numerical experiments on benchmark problems show that LBQL exhibits faster convergence and more robustness to hyperparameters when compared to standard Q-learning and several related techniques. Our approach is particularly appealing in problems that require expensive simulations or real-world interactions.

Causal Strategic Linear Regression

- Yonadav Shavit, Benjamin Edelman, Brian Axelrod
- abstract: In many predictive decision-making scenarios, such as credit scoring and academic testing, a decision-maker must construct a model that accounts for agents’ propensity to “game” the decision rule by changing their features so as to receive better decisions. Whereas the strategic classification literature has previously assumed that agents’ outcomes are not causally affected by their features (and thus that strategic agents’ goal is deceiving the decision-maker), we join concurrent work in modeling agents’ outcomes as a function of their changeable attributes. As our main contribution, we provide efficient algorithms for learning decision rules that optimize three distinct decision-maker objectives in a realizable linear setting: accurately predicting agents’ post-gaming outcomes (prediction risk minimization), incentivizing agents to improve these outcomes (agent outcome maximization), and estimating the coefficients of the true underlying model (parameter estimation). Our algorithms circumvent a hardness result of Miller et al. (2019) by allowing the decision maker to test a sequence of decision rules and observe agents’ responses, in effect performing causal interventions through the decision rules.

Adaptive Sampling for Estimating Probability Distributions

- Shubhanshu Shekhar, Tara Javidi, Mohammad Ghavamzadeh
- abstract: We consider the problem of allocating a fixed budget of samples to a finite set of discrete distributions to learn them uniformly well (minimizing the maximum error) in terms of four common distance measures: ℓ_2^2, ℓ_1, f-divergence, and separation distance. To present a unified treatment of these distances, we first propose a general \emph{optimistic tracking algorithm} and analyze its sample allocation performance w.r.t. an oracle. We then instantiate this algorithm for the four distance measures and derive bounds on their regret. We also show that the allocation performance of the proposed algorithm cannot, in general, be improved, by deriving lower-bounds on the expected deviation from the oracle allocation for any adaptive scheme. We verify our theoretical findings through some experiments. Finally, we show that the techniques developed in the paper can be easily extended to learn some classes of continuous distributions as well as to the related setting of minimizing the average error (in terms of the four distances) in learning a set of distributions.

PDO-eConvs: Partial Differential Operator Based Equivariant Convolutions

- Zhengyang Shen, Lingshen He, Zhouchen Lin, Jinwen Ma
- abstract: Recent research has shown that incorporating equivariance into neural network architectures is very helpful, and there have been some works investigating the equivariance of networks under group actions. However, as digital images and feature maps are on the discrete meshgrid, corresponding equivariance-preserving transformation groups are very limited. In this work, we deal with this issue from the connection between convolutions and partial differential operators (PDOs). In theory, assuming inputs to be smooth, we transform PDOs and propose a system which is equivariant to a much more general continuous group, the n-dimension Euclidean group. In implementation, we discretize the system using the numerical schemes of PDOs, deriving approximately equivariant convolutions (PDO-eConvs). Theoretically, the approximation error of PDO-eConvs is of the quadratic order. It is the first time that the error analysis is provided when the equivariance is approximate. Extensive experiments on rotated MNIST and natural image classification show that PDO-eConvs perform competitively yet use parameters much more efficiently. Particularly, compared with Wide ResNets, our methods result in better results using only 12.6% parameters.

Deep Reinforcement Learning with Robust and Smooth Policy

- Qianli Shen, Yan Li, Haoming Jiang, Zhaoran Wang, Tuo Zhao
- abstract: Deep reinforcement learning (RL) has achieved great empirical successes in various domains. However, the large search space of neural networks requires a large amount of data, which makes the current RL algorithms not sample efficient. Motivated by the fact that many environments with continuous state space have smooth transitions, we propose to learn a smooth policy that behaves smoothly with respect to states. We develop a new framework — \textbf{S}mooth \textbf{R}egularized \textbf{R}einforcement \textbf{L}earning ($\textbf{SR}^2\textbf{L}$), where the policy is trained with smoothness-inducing regularization. Such regularization effectively constrains the search space, and enforces smoothness in the learned policy. Moreover, our proposed framework can also improve the robustness of policy against measurement error in the state space, and can be naturally extended to distribubutionally robust setting. We apply the proposed framework to both on-policy (TRPO) and off-policy algorithm (DDPG). Through extensive experiments, we demonstrate that our method achieves improved sample efficiency and robustness.

Educating Text Autoencoders: Latent Representation Guidance via Denoising

- Tianxiao Shen, Jonas Mueller, Dr.Regina Barzilay, Tommi Jaakkola
- abstract: Generative autoencoders offer a promising approach for controllable text generation by leveraging their learned sentence representations. However, current models struggle to maintain coherent latent spaces required to perform meaningful text manipulations via latent vector operations. Specifically, we demonstrate by example that neural encoders do not necessarily map similar sentences to nearby latent vectors. A theoretical explanation for this phenomenon establishes that high-capacity autoencoders can learn an arbitrary mapping between sequences and associated latent representations. To remedy this issue, we augment adversarial autoencoders with a denoising objective where original sentences are reconstructed from perturbed versions (referred to as DAAE). We prove that this simple modification guides the latent space geometry of the resulting model by encouraging the encoder to map similar texts to similar latent representations. In empirical comparisons with various types of autoencoders, our model provides the best trade-off between generation quality and reconstruction capacity. Moreover, the improved geometry of the DAAE latent space enables \emph{zero-shot} text style transfer via simple latent vector arithmetic.

Learning for Dose Allocation in Adaptive Clinical Trials with Safety Constraints

- Cong Shen, Zhiyang Wang, Sofia Villar, Mihaela Van Der Schaar
- abstract: Phase I dose-finding trials are increasingly challenging as the relationship between efficacy and toxicity of new compounds (or combination of them) becomes more complex. Despite this, most commonly used methods in practice focus on identifying a Maximum Tolerated Dose (MTD) by learning only from toxicity events. We present a novel adaptive clinical trial methodology, called Safe Efficacy Exploration Dose Allocation (SEEDA), that aims at maximizing the cumulative efficacies while satisfying the toxicity safety constraint with high probability. We evaluate performance objectives that have operational meanings in practical clinical trials, including cumulative efficacy, recommendation/allocation success probabilities, toxicity violation probability, and sample efficiency. An extended SEEDA-Plateau algorithm that is tailored for the increase-then-plateau efficacy behavior of molecularly targeted agents (MTA) is also presented. Through numerical experiments using both synthetic and real-world datasets, we show that SEEDA outperforms state-of-the-art clinical trial designs by finding the optimal dose with higher success rate and fewer patients.

PowerNorm: Rethinking Batch Normalization in Transformers

- Sheng Shen, Zhewei Yao, Amir Gholami, Michael Mahoney, Kurt Keutzer
- abstract: The standard normalization method for neural network (NN) models used in Natural Language Processing (NLP) is layer normalization (LN).This is different than batch normalization (BN), which is widely-adopted in Computer Vision. The preferred use of LN in NLP is principally due to the empirical observation that a (naive/vanilla) use of BN leads to significant performance degradation for NLP tasks; however, a thorough understanding of the underlying reasons for this is not always evident. In this paper, we perform a systematic study of NLP transformer models to understand why BN has a poor performance, as compared to LN. We find that the statistics of NLP data across the batch dimension exhibit large fluctuations throughout training. This results in instability, if BN is naively implemented. To address this, we propose Power Normalization (PN), a novel normalization scheme that resolves this issue by (i) relaxing zero-mean normalization in BN, (ii) incorporating a running quadratic mean instead of per batch statistics to stabilize fluctuations, and (iii) using an approximate backpropagation for incorporating the running statistics in the forward pass. We show theoretically, under mild assumptions, that PN leads to a smaller Lipschitz constant for the loss, compared with BN. Furthermore, we prove that the approximate backpropagation scheme leads to bounded gradients. We extensively test PN for transformers on a range of NLP tasks, and we show that it significantly outperforms both LN and BN. In particular, PN outperforms LN by 0.4/0.6 BLEU on IWSLT14/WMT14 and 5.6/3.0 PPL on PTB/WikiText-103. We make our code publicly available at https://github.com/sIncerass/powernorm.

Extreme Multi-label Classification from Aggregated Labels

- Yanyao Shen, Hsiang-Fu Yu, Sujay Sanghavi, Inderjit Dhillon
- abstract: Extreme multi-label classification (XMC) is the problem of finding the relevant labels for an input, from a very large universe of possible labels. We consider XMC in the setting where labels are available only for groups of samples - but not for individual ones. Current XMC approaches are not built for such multi-instance multi-label (MIML) training data, and MIML approaches do not scale to XMC sizes. We develop a new and scalable algorithm to impute individual-sample labels from the group labels; this can be paired with any existing XMC method to solve the aggregated label problem. We characterize the statistical properties of our algorithm under mild assumptions, and provide a new end-to-end framework for MIML as an extension. Experiments on both aggregated label XMC and MIML tasks show the advantages over existing approaches.

One-shot Distributed Ridge Regression in High Dimensions

- Yue Sheng, Edgar Dobriban
- abstract: To scale up data analysis, distributed and parallel computing approaches are increasingly needed. Here we study a fundamental problem in this area: How to do ridge regression in a distributed computing environment? We study one-shot methods constructing weighted combinations of ridge regression estimators computed on each machine. By analyzing the mean squared error in a high dimensional model where each predictor has a small effect, we discover several new phenomena including that the efficiency depends strongly on the signal strength, but does not degrade with many workers, the risk decouples over machines, and the unexpected consequence that the optimal weights do not sum to unity. We also propose a new optimally weighted one-shot ridge regression algorithm. Our results are supported by simulations and real data analysis.

Landscape Connectivity and Dropout Stability of SGD Solutions for Over-parameterized Neural Networks

- Alexander Shevchenko, Marco Mondelli
- abstract: The optimization of multilayer neural networks typically leads to a solution with zero training error, yet the landscape can exhibit spurious local minima and the minima can be disconnected. In this paper, we shed light on this phenomenon: we show that the combination of stochastic gradient descent (SGD) and over-parameterization makes the landscape of multilayer neural networks approximately connected and thus more favorable to optimization. More specifically, we prove that SGD solutions are connected via a piecewise linear path, and the increase in loss along this path vanishes as the number of neurons grows large. This result is a consequence of the fact that the parameters found by SGD are increasingly dropout stable as the network becomes wider. We show that, if we remove part of the neurons (and suitably rescale the remaining ones), the change in loss is independent of the total number of neurons, and it depends only on how many neurons are left. Our results exhibit a mild dependence on the input dimension: they are dimension-free for two-layer networks and require the number of neurons to scale linearly with the dimension for multilayer networks. We validate our theoretical findings with numerical experiments for different architectures and classification tasks.

Incremental Sampling Without Replacement for Sequence Models

- Kensen Shi, David Bieber, Charles Sutton
- abstract: Sampling is a fundamental technique, and sampling without replacement is often desirable when duplicate samples are not beneficial. Within machine learning, sampling is useful for generating diverse outputs from a trained model. We present an elegant procedure for sampling without replacement from a broad class of randomized programs, including generative neural models that construct outputs sequentially. Our procedure is efficient even for exponentially-large output spaces. Unlike prior work, our approach is incremental, i.e., samples can be drawn one at a time, allowing for increased flexibility. We also present a new estimator for computing expectations from samples drawn without replacement. We show that incremental sampling without replacement is applicable to many domains, e.g., program synthesis and combinatorial optimization.

Message Passing Least Squares Framework and its Application to Rotation Synchronization

- Yunpeng Shi, Gilad Lerman
- abstract: We propose an efficient algorithm for solving group synchronization under high levels of corruption and noise, while we focus on rotation synchronization. We first describe our recent theoretically guaranteed message passing algorithm that estimates the corruption levels of the measured group ratios. We then propose a novel reweighted least squares method to estimate the group elements, where the weights are initialized and iteratively updated using the estimated corruption levels. We demonstrate the superior performance of our algorithm over state-of-the-art methods for rotation synchronization using both synthetic and real data.

Does the Markov Decision Process Fit the Data: Testing for the Markov Property in Sequential Decision Making

- Chengchun Shi, Runzhe Wan, Rui Song, Wenbin Lu, Ling Leng
- abstract: The Markov assumption (MA) is fundamental to the empirical validity of reinforcement learning. In this paper, we propose a novel Forward-Backward Learning procedure to test MA in sequential decision making. The proposed test does not assume any parametric form on the joint distribution of the observed data and plays an important role for identifying the optimal policy in high-order Markov decision processes (MDPs) and partially observable MDPs. Theoretically, we establish the validity of our test. Empirically, we apply our test to both synthetic datasets and a real data example from mobile health studies to illustrate its usefulness.

A Graph to Graphs Framework for Retrosynthesis Prediction

- Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, Jian Tang

- abstract: A fundamental problem in computational chemistry is to find a set of reactants to synthesize a target molecule, a.k.a. retrosynthesis prediction. Existing state-of-the-art methods rely on matching the target molecule with a large set of reaction templates, which are very computationally expensive and also suffer from the problem of coverage. In this paper, we propose a novel template-free approach called G2Gs by transforming a target molecular graph into a set of reactant molecular graphs. G2Gs first splits the target molecular graph into a set of synthons by identifying the reaction centers, and then translates the synthons to the final reactant graphs via a variational graph translation framework. Experimental results show that G2Gs significantly outperforms existing template-free approaches by up to 63% in terms of the top-1 accuracy and achieves a performance close to that of state-of-the-art template-based approaches, but does not require domain knowledge and is much more scalable.

Informative Dropout for Robust Representation Learning: A Shape-bias Perspective

- Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, Jingdong Wang
- abstract: Convolutional Neural Networks (CNNs) are known to rely more on local texture rather than global shape when making decisions. Recent work also indicates a close relationship between CNN's texture-bias and its robustness against distribution shift, adversarial perturbation, random corruption, etc. In this work, we attempt at improving various kinds of robustness universally by alleviating CNN's texture bias. With inspiration from the human visual system, we propose a light-weight model-agnostic method, namely Informative Dropout (InfoDrop), to improve interpretability and reduce texture bias. Specifically, we discriminate texture from shape based on local self-information in an image, and adopt a Dropout-like algorithm to decorrelate the model output from the local texture. Through extensive experiments, we observe enhanced robustness under various scenarios (domain generalization, few-shot classification, image corruption, and adversarial perturbation). To the best of our knowledge, this work is one of the earliest attempts to improve different kinds of robustness in a unified model, shedding new light on the relationship between shape-bias and robustness, also on new approaches to trustworthy machine learning algorithms. Code is available at https://github.com/bfshi/InfoDrop.

Dispersed Exponential Family Mixture VAEs for Interpretable Text Generation

- Wenxian Shi, Hao Zhou, Ning Miao, Lei Li
- abstract: Deep generative models are commonly used for generating images and text. Interpretability of these models is one important pursuit, other than the generation quality. Variational auto-encoder (VAE) with Gaussian distribution as prior has been successfully applied in text generation, but it is hard to interpret the meaning of the latent variable. To enhance the controllability and interpretability, one can replace the Gaussian prior with a mixture of Gaussian distributions (GM-VAE), whose mixture components could be related to hidden semantic aspects of data. In this paper, we generalize the practice and introduce DEM-VAE, a class of models for text generation using VAEs with a mixture distribution of exponential family. Unfortunately, a standard variational training algorithm fails due to the \emph{mode-collapse} problem. We theoretically identify the root cause of the problem and propose an effective algorithm to train DEM-VAE. Our method penalizes the training with an extra \emph{dispersion term} to induce a well-structured latent space. Experimental results show that our approach does obtain a meaningful space, and it outperforms strong baselines in text generation benchmarks. The code is available at \url{https://github.com/wenxianxian/demvae}.

On Conditional Versus Marginal Bias in Multi-Armed Bandits

- Jaehyeok Shin, Aaditya Ramdas, Alessandro Rinaldo
- abstract: The bias of the sample means of the arms in multi-armed bandits is an important issue in adaptive data analysis that has recently received considerable attention in the literature. Existing results relate in precise ways the sign and magnitude of the bias to various sources of data adaptivity, but do not apply to the conditional inference setting in which the sample means are computed only if some specific conditions are satisfied. In this paper, we characterize the sign of the conditional bias of monotone functions of the rewards, including the sample mean. Our results hold for arbitrary conditioning events and leverage natural monotonicity properties of the data collection policy. We further demonstrate, through several examples from sequential testing and best arm identification, that the sign of the conditional and marginal bias of the sample mean of an arm can be different, depending on the conditioning event. Our analysis offers new and interesting perspectives on the subtleties of assessing the bias in data adaptive settings.

Predictive Coding for Locally-Linear Control

- Rui Shu, Tung Nguyen, Yinlam Chow, Tuan Pham, Khoat Than, Mohammad Ghavamzadeh, Stefano Ermon, Hung Bui
- abstract: High-dimensional observations and unknown dynamics are major challenges when applying optimal control to many real-world decision making tasks. The Learning Controllable Embedding (LCE) framework addresses these challenges by embedding the observations into a lower dimensional latent space, estimating the latent dynamics, and then performing control directly in the latent space. To ensure the learned latent dynamics are predictive of next-observations, all existing LCE approaches decode back into the observation space and explicitly perform next-observation prediction—a challenging high-dimensional task that furthermore introduces a large number of nuisance parameters (i.e., the decoder) which are discarded during control. In this paper, we propose a novel information-theoretic LCE approach and show theoretically that explicit next-observation prediction can be replaced with predictive coding. We then use predictive coding to develop a decoder-free LCE model whose latent dynamics are amenable to locally-linear control. Extensive experiments on benchmark tasks show that our model reliably learns a controllable latent space that leads to superior performance when compared with state-of-the-art LCE baselines.

A Markov Decision Process Model for Socio-Economic Systems Impacted by Climate Change

- Salman Sadiq Shuvo, Yasin Yilmaz, Alan Bush, Mark Hafen
- abstract: Coastal communities are at high risk of natural hazards due to unremitting global warming and sea level rise. Both the catastrophic impacts, e.g., tidal flooding and storm surges, and the long-term impacts, e.g., beach erosion, inundation of low lying areas, and saltwater intrusion into aquifers, cause economic, social, and ecological losses. Creating policies through appropriate modeling of the responses of stakeholders, such as government, businesses, and residents, to climate change and sea level rise scenarios can help to reduce these losses. In this work, we propose a Markov decision process (MDP) formulation for an agent (government) which interacts with the environment (nature and residents) to deal with the impacts of climate change, in particular sea level rise. Through theoretical analysis we show that a reasonable government's policy on infrastructure development ought to be proactive and based on detected sea levels in order to minimize the expected total cost, as opposed to a straightforward government that reacts to observed costs from nature. We also provide a deep reinforcement learning-based scenario planning tool considering different government and resident types in terms of cooperation, and different sea level rise projections by the National Oceanic and Atmospheric Administration (NOAA).

Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits

- Nian Si, Fan Zhang, Zhengyuan Zhou, Jose Blanchet
- abstract: Policy learning using historical observational data is an important problem that has found widespread applications. However, existing literature rests on the crucial assumption that the future environment where the learned policy will be deployed is the same as the past environment that has generated the data{–}an assumption that is often false or too coarse an approximation. In this paper, we lift this assumption and aim to learn a distributionally robust policy with bandit observational data. We propose a novel learning algorithm that is able to learn a robust policy to adversarial perturbations and unknown covariate shifts. We first present a policy evaluation procedure in the ambiguous environment and also give a heuristic algorithm to solve the distributionally robust policy learning problems efficiently. Additionally, we provide extensive simulations to demonstrate the robustness of our policy.

Piecewise Linear Regression via a Difference of Convex Functions

- Ali Siahkamari, Aditya Gangrade, Brian Kulis, Venkatesh Saligrama
- abstract: We present a new piecewise linear regression methodology that utilises fitting a \emph{difference of convex} functions (DC functions) to the data. These are functions f that may be represented as the difference $\phi_1 - \phi_2$ for a choice of \emph{convex} functions ϕ_1, ϕ_2. The method proceeds by estimating piecewise-liner convex functions, in a manner similar to max-affine regression, whose difference approximates the data. The choice of the function is regularised by a new seminorm over the class of DC functions that controls the ℓ_∞ Lipschitz constant of the estimate. The resulting methodology can be efficiently implemented via Quadratic programming \emph{even in high dimensions}, and is shown to have close to minimax statistical risk. We empirically validate the method, showing it to be practically implementable, and to outperform existing regression methods in accuracy on real-world datasets.

Learning Fair Policies in Multi-Objective (Deep) Reinforcement Learning with Average and Discounted Rewards

- Umer Siddique, Paul Weng, Matthieu Zimmer
- abstract: As the operations of autonomous systems generally affect simultaneously several users, it is crucial that their designs account for fairness considerations. In contrast to standard (deep) reinforcement learning (RL), we investigate the problem of learning a policy that treats its users equitably. In this paper, we formulate this novel RL problem, in which an objective function, which encodes a notion of fairness that we formally define, is optimized. For this problem, we provide a theoretical discussion where we examine the case of discounted rewards and that of average rewards. During this analysis, we notably derive a new result in the standard RL setting, which is of independent interest: it states a novel bound on the approximation error with respect to the optimal average reward of that of a policy optimal for the discounted reward. Since learning with discounted rewards is generally easier, this discussion further justifies finding a fair policy for the average reward by learning a fair policy for the discounted reward. Thus, we describe how several classic deep RL algorithms can be adapted to our fair optimization problem, and we validate our approach with extensive experiments in three different domains.

Deep Gaussian Markov Random Fields

- Per Sidén, Fredrik Lindsten
- abstract: Gaussian Markov random fields (GMRFs) are probabilistic graphical models widely used in spatial statistics and related fields to model dependencies over spatial structures. We establish a formal connection between GMRFs and convolutional neural networks (CNNs). Common GMRFs are special cases of a generative model where the inverse mapping from data to latent variables is given by a 1-layer linear CNN. This connection allows us to generalize GMRFs to multi-layer CNN architectures, effectively increasing the order of the corresponding GMRF in a way which has favorable computational scaling. We describe how well-established tools, such as autodiff and variational inference, can be used for simple and efficient inference and learning of the deep GMRF. We demonstrate the flexibility of the proposed model and show that it outperforms the state-of-the-art on a dataset of satellite temperatures, in terms of prediction and predictive uncertainty.

Collaborative Machine Learning with Incentive-Aware Model Rewards

- Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, Bryan Kian Hsiang Low
- abstract: Collaborative machine learning (ML) is an appealing paradigm to build high-quality ML models by training on the aggregated data from many parties. However, these parties are only willing to share their data when given enough incentives, such as a guaranteed fair reward based on their contributions. This motivates the need for measuring a party's contribution and designing an incentive-aware reward scheme accordingly. This paper proposes to value a party's reward based on Shapley value and information gain on model parameters given its data. Subsequently, we give each party a model as a reward. To formally incentivize the collaboration, we define some desirable properties (e.g., fairness and stability) which are inspired by cooperative game theory but adapted for our model reward that is uniquely freely replicable. Then, we propose a novel model reward scheme to satisfy fairness and trade off between the desirable properties via an adjustable parameter. The value of each party's model reward determined by our scheme is attained by injecting Gaussian noise to the aggregated training data with an optimized noise variance. We empirically demonstrate interesting properties of our scheme and evaluate its performance using synthetic and real-world datasets.

Naive Exploration is Optimal for Online LQR

- Max Simchowitz, Dylan Foster
- abstract: We consider the problem of online adaptive control of the linear quadratic regulator, where the true system parameters are unknown. We prove new upper and lower bounds demonstrating that the optimal regret scales as $\tilde{\Theta} (\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T})$, where T is the number of time steps, $d_{\mathbf{u}}$ is the dimension of the input space, and $d_{\mathbf{x}}$ is the dimension of the system state. Notably, our lower bounds rule out the possibility of a $\mathrm{poly}(\log{T})$-regret algorithm, which had been conjectured due to the apparent strong convexity of the problem. Our upper bound is attained by a simple variant of certainty equivalent control, where the learner selects control inputs according to the optimal controller for their estimate of the system while injecting exploratory random noise. While this approach was shown to achieve $\sqrt{T}$ regret by Mania et al. (2019), we show that if the learner continually refines their estimates of the system matrices, the method attains optimal dimension dependence as well. Central to our upper and lower bounds is a new approach for controlling perturbations of Riccati equations called the self-bounding ODE method, which we use to derive suboptimality bounds for the certainty equivalent controller synthesized from estimated system dynamics. This in turn enables regret upper bounds which hold for any stabilizable instance and scale with natural control-theoretic quantities.

A Generative Model for Molecular Distance Geometry

- Gregor Simm, Jose Miguel Hernandez-Lobato
- abstract: Great computational effort is invested in generating equilibrium states for molecular systems using, for example, Markov chain Monte Carlo. We present a probabilistic model that generates statistically independent samples for molecules from their graph representations. Our model learns a low-dimensional manifold that preserves the geometry of local atomic neighborhoods through a principled learning representation that is based on Euclidean distance geometry. In a new benchmark for molecular conformation generation, we show experimentally that our generative model achieves state-of-the-art accuracy. Finally, we show how to use our model as a proposal distribution in an importance sampling scheme to compute molecular properties.

Reinforcement Learning for Molecular Design Guided by Quantum Mechanics

- Gregor Simm, Robert Pinsler, Jose Miguel Hernandez-Lobato
- abstract: Automating molecular design using deep reinforcement learning (RL) holds the promise of accelerating the discovery of new chemical compounds. Existing approaches work with molecular graphs and thus ignore the location of atoms in space, which restricts them to 1) generating single organic molecules and 2) heuristic reward functions. To address this, we present a novel RL formulation for molecular design in Cartesian coordinates, thereby extending the class of molecules that can be built. Our reward function is directly based on fundamental physical properties such as the energy, which we approximate via fast quantum-chemical methods. To enable progress towards de-novo molecular design, we introduce MolGym, an RL environment comprising several challenging molecular design tasks along with baselines. In our experiments, we show that our agent can efficiently learn to solve these tasks from scratch by working in a translation and rotation invariant state-action space.

Fractional Underdamped Langevin Dynamics: Retargeting SGD with Momentum under Heavy-Tailed Gradient Noise

- Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, Mert Gurbuzbalaban
- abstract: Stochastic gradient descent with momentum (SGDm) is one of the most popular optimization algorithms in deep learning. While there is a rich theory of SGDm for convex problems, the theory is considerably less developed in the context of deep learning where the problem is non-convex and the gradient noise might exhibit a heavy-tailed behavior, as empirically observed in recent studies. In this study, we consider a \emph{continuous-time} variant of SGDm, known as the underdamped Langevin dynamics (ULD), and investigate its asymptotic properties under heavy-tailed perturbations. Supported by recent studies from statistical physics, we argue both theoretically and empirically that the heavy-tails of such perturbations can result in a bias even when the step-size is small, in the sense that \emph{the optima of stationary distribution} of the dynamics might not match \emph{the optima of the cost function to be optimized}. As a remedy, we develop a novel framework, which we coin as \emph{fractional} ULD (FULD), and prove that FULD targets the so-called Gibbs distribution, whose optima exactly match the optima of the original cost. We observe that the Euler discretization of FULD has noteworthy algorithmic similarities with \emph{natural gradient} methods and \emph{gradient clipping}, bringing a new perspective on understanding their role in deep learning. We support our theory with experiments conducted on a synthetic model and neural networks.

Second-Order Provable Defenses against Adversarial Attacks

- Sahil Singla, Soheil Feizi
- abstract: A robustness certificate against adversarial examples is the minimum distance of a given input to the decision boundary of the classifier (or its lower bound). For \emph{any} perturbation of the input with a magnitude smaller than the certificate value, the classification output will provably remain unchanged. Computing exact robustness certificates for neural networks is difficult in general since it requires solving a non-convex optimization. In this paper, we provide computationally-efficient robustness certificates for neural networks with differentiable activation functions in two steps. First, we show that if the eigenvalues of the Hessian of the network (curvatures of the network) are bounded (globally or locally), we can compute a robustness certificate in the l_2 norm efficiently using convex optimization. Second, we derive a computationally-efficient differentiable upper bound on the curvature of a deep network. We also use the curvature bound as a regularization term during the training of the network to boost its certified robustness. Putting these results together leads to our proposed {\bf C}urvature-based {\bf R}obustness {\bf C}ertificate (CRC) and {\bf C}urvature-based {\bf R}obust {\bf T}raining (CRT). Our numerical results show that CRT leads to significantly higher certified robust accuracy compared to interval-bound propagation based training.

FormulaZero: Distributionally Robust Online Adaptation via Offline Population Synthesis

- Aman Sinha, Matthew O'Kelly, Hongrui Zheng, Rahul Mangharam, John Duchi, Russ Tedrake
- abstract: Balancing performance and safety is crucial to deploying autonomous vehicles in multi-agent environments. In particular, autonomous racing is a domain that penalizes safe but conservative policies, highlighting the need for robust, adaptive strategies. Current approaches either make simplifying assumptions about other agents or lack robust mechanisms for online adaptation. This work makes algorithmic contributions to both challenges. First, to generate a realistic, diverse set of opponents, we develop a novel method for self-play based on replica-exchange Markov chain Monte Carlo. Second, we propose a distributionally robust bandit optimization procedure that adaptively adjusts risk aversion relative to uncertainty in beliefs about opponents' behaviors. We rigorously quantify the tradeoffs in performance and robustness when approximating these computations in real-time motion-planning, and we demonstrate our methods experimentally on autonomous vehicles that achieve scaled speeds comparable to Formula One racecars.

Small-GAN: Speeding up GAN Training using Core-Sets

- Samarth Sinha, Han Zhang, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, Augustus Odena
- abstract: Recent work suggests that Generative Adversarial Networks (GANs) benefit disproportionately from large mini-batch sizes. This finding is interesting but also discouraging – large batch sizes are slow and expensive to emulate on conventional hardware. Thus, it would be nice if there were some trick by which we could generate batches that were effectively big though small in practice. In this work, we propose such a trick, inspired by the use of Coreset-selection in active learning. When training a GAN, we draw a large batch of samples from the prior and then compress that batch using Coreset-selection. To create effectively large batches of real images, we create a cached dataset of Inception activations of each training image, randomly project them down to a smaller dimension, and then use Coreset-selection on those projected embeddings at training time. We conduct experiments showing that this technique substantially reduces training time and memory usage for modern GAN variants, that it reduces the fraction of dropped modes in a synthetic dataset, and that it helps us use GANs to reach a new state of the art in anomaly detection.

Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure

- John Sipple
- abstract: In this paper we propose a scalable, unsupervised approach for detecting anomalies in the Internet of Things (IoT). Complex devices are connected daily and eagerly generate vast streams of multidimensional telemetry. These devices often operate in distinct modes based on external conditions (day/night, occupied/vacant, etc.), and to prevent complete or partial system outage, we would like to recognize as early as possible when these devices begin to operate outside the normal modes. We propose an unsupervised anomaly detection method that creates a negative sample from the positive, observed sample, and trains a classifier to distinguish between positive and negative samples. Using the Concentration Phenomenon, we explain why such a classifier ought to establish suitable decision boundaries between normal and anomalous regions, and show how Integrated Gradients can attribute the anomaly to specific dimensions within the anomalous state vector. We have demonstrated that negative sampling with random forest or neural network classifiers yield significantly higher AUC scores compared to state-of-the-art approaches against benchmark anomaly detection datasets, and a multidimensional, multimodal dataset from real climate control devices. Finally, we describe how negative sampling with neural network classifiers have been successfully deployed at large scale to predict failures in real time in over 15,000 climate-control and power meter devices in 145 office buildings within the California Bay Area.

Structured Linear Contextual Bandits: A Sharp and Geometric Smoothed Analysis

- Vidyashankar Sivakumar, Steven Wu, Arindam Banerjee
- abstract: Bandit learning algorithms typically involve the balance of exploration and exploitation. However, in many practical applications, worst-case scenarios needing systematic exploration are seldom encountered. In this work, we consider a smoothed setting for structured linear contextual bandits where the adversarial contexts are perturbed by Gaussian noise and the unknown parameter $\theta^$ *has structure, e.g., sparsity, group sparsity, low rank, etc. We propose simple greedy algorithms for both the single- and multi-parameter (i.e., different parameter for each context) settings and provide a unified regret analysis for* $\theta^$ with any assumed structure. The regret bounds are expressed in terms of geometric quantities such as Gaussian widths associated with the structure of $\theta^$. *We also obtain sharper regret bounds compared to earlier work for the unstructured* $\theta^$ setting as a consequence of our improved analysis. We show there is implicit exploration in the smoothed setting where a simple greedy algorithm works.

Optimizer Benchmarking Needs to Account for Hyperparameter Tuning

- Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, François Fleuret

- abstract: The performance of optimizers, particularly in deep learning, depends considerably on their chosen hyperparameter configuration. The efficacy of optimizers is often studied under near-optimal problem-specific hyperparameters, and finding these settings may be prohibitively costly for practitioners. In this work, we argue that a fair assessment of optimizers' performance must take the computational cost of hyperparameter tuning into account, i.e., how easy it is to find good hyperparameter configurations using an automatic hyperparameter search. Evaluating a variety of optimizers on an extensive set of standard datasets and architectures, our results indicate that Adam is the most practical solution, particularly in low-budget scenarios.

When Explanations Lie: Why Many Modified BP Attributions Fail

- Leon Sixt, Maximilian Granz, Tim Landgraf
- abstract: Attribution methods aim to explain a neural network's prediction by highlighting the most relevant image areas. A popular approach is to backpropagate (BP) a custom relevance score using modified rules, rather than the gradient. We analyze an extensive set of modified BP methods: Deep Taylor Decomposition, Layer-wise Relevance Propagation (LRP), Excitation BP, PatternAttribution, DeepLIFT, Deconv, RectGrad, and Guided BP. We find empirically that the explanations of all mentioned methods, except for DeepLIFT, are independent of the parameters of later layers. We provide theoretical insights for this surprising behavior and also analyze why DeepLIFT does not suffer from this limitation. Empirically, we measure how information of later layers is ignored by using our new metric, cosine similarity convergence (CSC). The paper provides a framework to assess the faithfulness of new and existing modified BP methods theoretically and empirically.

On the Generalization Benefit of Noise in Stochastic Gradient Descent

- Samuel Smith, Erich Elsen, Soham De
- abstract: It has long been argued that minibatch stochastic gradient descent can generalize better than large batch gradient descent in deep neural networks. However recent papers have questioned this claim, arguing that this effect is simply a consequence of suboptimal hyperparameter tuning or insufficient compute budgets when the batch size is large. In this paper, we perform carefully designed experiments and rigorous hyperparameter sweeps on a range of popular models, which verify that small or moderately large batch sizes can substantially outperform very large batches on the test set. This occurs even when both models are trained for the same number of iterations and large batches achieve smaller training losses. Our results confirm that the noise in stochastic gradients can enhance generalization. We study how the optimal learning rate schedule changes as the epoch budget grows, and we provide a theoretical account of our observations based on the stochastic differential equation perspective of SGD dynamics.

Multiclass Neural Network Minimization via Tropical Newton Polytope Approximation

- Georgios Smyrnis, Petros Maragos
- abstract: The field of tropical algebra is closely linked with the domain of neural networks with piecewise linear activations, since their output can be described via tropical polynomials in the max-plus semiring. In this work, we attempt to make use of methods stemming from a form of approximate division of such polynomials, which relies on the approximation of their Newton Polytopes, in order to minimize networks trained for multiclass classification problems. We make theoretical contributions in this domain, by proposing and analyzing methods which seek to reduce the size of such networks. In addition, we make experimental evaluations on the MNIST and Fashion-MNIST datasets, with our results demonstrating a significant reduction in network size, while retaining adequate performance.

Bridging the Gap Between f-GANs and Wasserstein GANs

- Jiaming Song, Stefano Ermon
- abstract: Generative adversarial networks (GANs) variants approximately minimize divergences between the model and the data distribution using a discriminator. Wasserstein GANs (WGANs) enjoy superior empirical performance, however, unlike in f-GANs, the discriminator does not provide an estimate for the ratio between model and data densities, which is useful in applications such as inverse reinforcement learning. To overcome this limitation, we propose an new training objective where we additionally optimize over a set of importance weights over the generated samples. By suitably constraining the feasible set of importance weights, we obtain a family of objectives which includes and generalizes the original f-GAN and WGAN objectives. We show that a natural extension outperforms WGANs while providing density ratios as in f-GAN, and demonstrate empirical success on distribution modeling, density ratio estimation and image generation.

Provably Efficient Model-based Policy Adaptation

- Yuda Song, Aditi Mavalankar, Wen Sun, Sicun Gao
- abstract: The high sample complexity of reinforcement learning challenges its use in practice. A promising approach is to quickly adapt pre-trained policies to new environments. Existing methods for this policy adaptation problem typically rely on domain randomization and meta-learning, by sampling from some distribution of target environments during pre-training, and thus face difficulty on out-of-distribution target environments. We propose new model-based mechanisms that are able to make online adaptation in unseen target environments, by combining ideas from no-regret online learning and adaptive control. We prove that the approach learns policies in the target environment that can quickly recover trajectories from the source environment, and establish the rate of convergence in general settings. We demonstrate the benefits of our approach for policy adaptation in a diverse set of continuous control tasks, achieving the performance of state-of-the-art methods with much lower sample complexity. Our project website, including code, can be found at https://yudasong.github.io/PADA.

Hypernetwork approach to generating point clouds

- Przemysław Spurek, Sebastian Winczowski, Jacek Tabor, Maciej Zamorski, Maciej Zieba, Tomasz Trzcinski
- abstract: In this work, we propose a novel method for generating 3D point clouds that leverage properties of hyper networks. Contrary to the existing methods that learn only the representation of a 3D object, our approach simultaneously finds a representation of the object and its 3D surfaces. The main idea of our HyperCloud method is to build a hyper network that returns weights of a particular neural network (target network) trained to map points from a uniform unit ball distribution into a 3D shape. As a consequence, a particular 3D shape can be generated using point-by-point sampling from the assumed prior distribution and transforming sampled points with the target network. Since the hyper network is based on an auto-encoder architecture trained to reconstruct realistic 3D shapes, the target network weights can be considered a parametrisation of the surface of a 3D shape, and not a standard representation of point cloud usually returned by competitive approaches. The proposed architecture allows to find mesh-based representation of 3D objects in a generative manner, while providing point clouds en pair in quality with the state-of-the-art methods.

Robustness to Spurious Correlations via Human Annotations

- Megha Srivastava, Tatsunori Hashimoto, Percy Liang
- abstract: The reliability of machine learning systems critically assumes that the associations between features and labels remain similar between training and test distributions. However, unmeasured variables, such as confounders, break this assumption—useful correlations between features and labels at training time can become useless or even harmful at test time. For example, high obesity is generally predictive for heart disease, but this relation may not hold for smokers who generally have lower rates of obesity and higher rates of heart disease. We present a framework for making models robust to spurious correlations by leveraging humans' common sense knowledge of causality. Specifically, we use human annotation to augment each training

example with a potential unmeasured variable (i.e. an underweight patient with heart disease may be a smoker), reducing the problem to a covariate shift problem. We then introduce a new distributionally robust optimization objective over unmeasured variables (UV-DRO) to control the worst-case loss over possible test- time shifts. Empirically, we show improvements of 5–10% on a digit recognition task confounded by rotation, and 1.5–5% on the task of analyzing NYPD Police Stops confounded by location.

Which Tasks Should Be Learned Together in Multi-task Learning?

- Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, Silvio Savarese
- abstract: Many computer vision applications require solving multiple tasks in real-time. A neural network can be trained to solve multiple tasks simultaneously using multi-task learning. This can save computation at inference time as only a single network needs to be evaluated. Unfortunately, this often leads to inferior overall performance as task objectives can compete, which consequently poses the question: which tasks should and should not be learned together in one network when employing multi-task learning? We study task cooperation and competition in several different learning settings and propose a framework for assigning tasks to a few neural networks such that cooperating tasks are computed by the same neural network, while competing tasks are computed by different networks. Our framework offers a time-accuracy trade-off and can produce better accuracy using less inference time than not only a single large multi-task neural network but also many single-task networks.

Responsive Safety in Reinforcement Learning by PID Lagrangian Methods

- Adam Stooke, Joshua Achiam, Pieter Abbeel
- abstract: Lagrangian methods are widely used algorithms for constrained optimization problems, but their learning dynamics exhibit oscillations and overshoot which, when applied to safe reinforcement learning, leads to constraint-violating behavior during agent training. We address this shortcoming by proposing a novel Lagrange multiplier update method that utilizes derivatives of the constraint function. We take a controls perspective, wherein the traditional Lagrange multiplier update behaves as \emph{integral} control; our terms introduce \emph{proportional} and \emph{derivative} control, achieving favorable learning dynamics through damping and predictive measures. We apply our PID Lagrangian methods in deep RL, setting a new state of the art in Safety Gym, a safe RL benchmark. Lastly, we introduce a new method to ease controller tuning by providing invariance to the relative numerical scales of reward and cost. Our extensive experiments demonstrate improved performance and hyperparameter robustness, while our algorithms remain nearly as simple to derive and implement as the traditional Lagrangian approach.

Learning Discrete Structured Representations by Adversarially Maximizing Mutual Information

- Karl Stratos, Sam Wiseman
- abstract: We propose learning discrete structured representations from unlabeled data by maximizing the mutual information between a structured latent variable and a target variable. Calculating mutual information is intractable in this setting. Our key technical contribution is an adversarial objective that can be used to tractably estimate mutual information assuming only the feasibility of cross entropy calculation. We develop a concrete realization of this general formulation with Markov distributions over binary encodings. We report critical and unexpected findings on practical aspects of the objective such as the choice of variational priors. We apply our model on document hashing and show that it outperforms current best baselines based on discrete and vector quantized variational autoencoders. It also yields highly compressed interpretable representations.

Confidence-Calibrated Adversarial Training: Generalizing to Unseen Attacks

- David Stutz, Matthias Hein, Bernt Schiele
- abstract: Adversarial training yields robust models against a specific threat model, e.g., L_∞ adversarial examples. Typically robustness does not generalize to previously unseen threat models, e.g., other L_p norms, or larger perturbations. Our confidence-calibrated adversarial training (CCAT) tackles this problem by biasing the model towards low confidence predictions on adversarial examples. By allowing to reject examples with low confidence, robustness generalizes beyond the threat model employed during training. CCAT, trained only on L_∞ adversarial examples, increases robustness against larger L_∞, L_2, L_1 and L_0 attacks, adversarial frames, distal adversarial examples and corrupted examples and yields better clean accuracy compared to adversarial training. For thorough evaluation we developed novel white- and black-box attacks directly attacking CCAT by maximizing confidence. For each threat model, we use 7 attacks with up to 50 restarts and 5000 iterations and report worst-case robust test error, extended to our confidence-thresholded setting, across all attacks.

Doubly robust off-policy evaluation with shrinkage

- Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, Miroslav Dudik
- abstract: We propose a new framework for designing estimators for off-policy evaluation in contextual bandits. Our approach is based on the asymptotically optimal doubly robust estimator, but we shrink the importance weights to minimize a bound on the mean squared error, which results in a better bias-variance tradeoff in finite samples. We use this optimization-based framework to obtain three estimators: (a) a weight-clipping estimator, (b) a new weight-shrinkage estimator, and (c) the first shrinkage-based estimator for combinatorial action sets. Extensive experiments in both standard and combinatorial bandit benchmark problems show that our estimators are highly adaptive and typically outperform state-of-the-art methods.

Task Understanding from Confusing Multi-task Data

- Xin Su, Yizhou Jiang, Shangqi Guo, Feng Chen
- abstract: Beyond machine learning’s success in the specific tasks, research for learning multiple tasks simultaneously is referred to as multi-task learning. However, existing multi-task learning needs manual definition of tasks and manual task annotation. A crucial problem for advanced intelligence is how to understand the human task concept using basic input-output pairs. Without task definition, samples from multiple tasks are mixed together and result in a confusing mapping challenge. We propose Confusing Supervised Learning (CSL) that takes these confusing samples and extracts task concepts by differentiating between these samples. We theoretically proved the feasibility of the CSL framework and designed an iterative algorithm to distinguish between tasks. The experiments demonstrate that our CSL methods could achieve a human-like task understanding without task labeling in multi-function regression problems and multi-task recognition problems.

ConQUR: Mitigating Delusional Bias in Deep Q-Learning

- Dijia Su, Jayden Ooi, Tyler Lu, Dale Schuurmans, Craig Boutilier
- abstract: Delusional bias is a fundamental source of error in approximate Q-learning. To date, the only techniques that explicitly address delusion require comprehensive search using tabular value estimates. In this paper, we develop efficient methods to mitigate delusional bias by training Q-approximators with labels that are "consistent" with the underlying greedy policy class. We introduce a simple penalization scheme that encourages Q-labels used across training batches to remain (jointly) consistent with the expressible policy class. We also propose a search framework that allows multiple Q-approximators to be generated and tracked, thus mitigating the effect of premature (implicit) policy commitments. Experimental results demonstrate that these methods can improve the performance of Q-learning in a variety of Atari games, sometimes dramatically.

Adaptive Estimator Selection for Off-Policy Evaluation

- Yi Su, Pavithra Srinath, Akshay Krishnamurthy
- abstract: We develop a generic data-driven method for estimator selection in off-policy policy evaluation settings. We establish a strong performance guarantee for the method, showing that it is competitive with the oracle estimator, up to a constant factor. Via in-depth case studies in contextual bandits and reinforcement learning, we demonstrate the generality and applicability of the method. We also perform comprehensive experiments, demonstrating the empirical efficacy of our approach and comparing with related approaches. In both case studies, our method compares favorably with existing methods.

Generative Teaching Networks: Accelerating Neural Architecture Search by Learning to Generate Synthetic Training Data

- Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, Jeffrey Clune
- abstract: This paper investigates the intriguing question of whether we can create learning algorithms that automatically generate training data, learning environments, and curricula in order to help AI agents rapidly learn. We show that such algorithms are possible via Generative Teaching Networks (GTNs), a general approach that is, in theory, applicable to supervised, unsupervised, and reinforcement learning, although our experiments only focus on the supervised case. GTNs are deep neural networks that generate data and/or training environments that a learner (e.g. a freshly initialized neural network) trains on for a few SGD steps before being tested on a target task. We then differentiate \emph{through the entire learning process} via meta-gradients to update the GTN parameters to improve performance on the target task. This paper introduces GTNs, discusses their potential, and showcases that they can substantially accelerate learning. We also demonstrate a practical and exciting application of GTNs: accelerating the evaluation of candidate architectures for neural architecture search (NAS). GTN-NAS improves the NAS state of the art, finding higher performing architectures when controlling for the search proposal mechanism. GTN-NAS also is competitive with the overall state of the art approaches, which achieve top performance while using orders of magnitude less computation than typical NAS methods. Speculating forward, GTNs may represent a first step toward the ambitious goal of algorithms that generate their own training data and, in doing so, open a variety of interesting new research questions and directions.

Improving the Sample and Communication Complexity for Decentralized Non-Convex Optimization: Joint Gradient Estimation and Tracking

- Haoran Sun, Songtao Lu, Mingyi Hong
- abstract: Many modern large-scale machine learning problems benefit from decentralized and stochastic optimization. Recent works have shown that utilizing both decentralized computing and local stochastic gradient estimates can outperform state-of-the-art centralized algorithms, in applications involving highly non-convex problems, such as training deep neural networks. In this work, we propose a decentralized stochastic algorithm to deal with certain smooth non-convex problems where there are m nodes in the system, and each node has a large number of samples (denoted as n). Differently from the majority of the existing decentralized learning algorithms for either stochastic or finite-sum problems, our focus is given to \emph{both} reducing the total communication rounds among the nodes, while accessing the minimum number of local data samples. In particular, we propose an algorithm named D-GET (decentralized gradient estimation and tracking), which jointly performs decentralized gradient estimation (which estimates the local gradient using a subset of local samples) \emph{and} gradient tracking (which tracks the global full gradient using local estimates). We show that to achieve certain ϵ stationary solution of the deterministic finite sum problem, the proposed algorithm achieves an $\mathcal{O}(mn^{1/2}\epsilon^{-1})$ sample complexity and an $\mathcal{O}(\epsilon^{-1})$ communication complexity. These bounds significantly improve upon the best existing bounds of $\mathcal{O}(mn\epsilon^{-1})$ and $\mathcal{O}(\epsilon^{-1})$, respectively. Similarly, for online problems, the proposed method achieves an $\mathcal{O}(m \epsilon^{-3/2})$ sample complexity and an $\mathcal{O}(\epsilon^{-1})$ communication complexity.

Test-Time Training with Self-Supervision for Generalization under Distribution Shifts

- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, Moritz Hardt
- abstract: In this paper, we propose Test-Time Training, a general approach for improving the performance of predictive models when training and test data come from different distributions. We turn a single unlabeled test sample into a self-supervised learning problem, on which we update the model parameters before making a prediction. This also extends naturally to data in an online stream. Our simple approach leads to improvements on diverse image classification benchmarks aimed at evaluating robustness to distribution shifts.

An EM Approach to Non-autoregressive Conditional Sequence Generation

- Zhiqing Sun, Yiming Yang
- abstract: Autoregressive (AR) models have been the dominating approach to conditional sequence generation, but are suffering from the issue of high inference latency. Non-autoregressive (NAR) models have been recently proposed to reduce the latency by generating all output tokens in parallel but could only achieve inferior accuracy compared to their autoregressive counterparts, primarily due to a difficulty in dealing with the multi-modality in sequence generation. This paper proposes a new approach that jointly optimizes both AR and NAR models in a unified Expectation-Maximization (EM) framework. In the E-step, an AR model learns to approximate the regularized posterior of the NAR model. In the M-step, the NAR model is updated on the new posterior and selects the training examples for the next AR model. This iterative process can effectively guide the system to remove the multi-modality in the output sequences. To our knowledge, this is the first EM approach to NAR sequence generation. We evaluate our method on the task of machine translation. Experimental results on benchmark data sets show that the proposed approach achieves competitive, if not better, performance with existing NAR models and significantly reduces the inference latency.

The Shapley Taylor Interaction Index

- Mukund Sundararajan, Kedar Dhamdhere, Ashish Agarwal
- abstract: The attribution problem, that is the problem of attributing a model's prediction to its base features, is well-studied. We extend the notion of attribution to also apply to feature interactions. The Shapley value is a commonly used method to attribute a model's prediction to its base features. We propose a generalization of the Shapley value called Shapley-Taylor index that attributes the model's prediction to interactions of subsets of features up to some size k. The method is analogous to how the truncated Taylor Series decomposes the function value at a certain point using its derivatives at a different point. In fact, we show that the Shapley Taylor index is equal to the Taylor Series of the multilinear extension of the set-theoretic behavior of the model. We axiomatize this method using the standard Shapley axioms—linearity, dummy, symmetry and efficiency—and an additional axiom that we call the interaction distribution axiom. This new axiom explicitly characterizes how interactions are distributed for a class of functions that model pure interaction. We contrast the Shapley-Taylor index against the previously proposed Shapley Interaction index from the cooperative game theory literature. We also apply the Shapley Taylor index to three models and identify interesting qualitative insights.

The Many Shapley Values for Model Explanation

- Mukund Sundararajan, Amir Najmi
- abstract: The Shapley value has become the basis for several methods that attribute the prediction of a machine-learning model on an input to its base features. The use of the Shapley value is justified by citing the uniqueness result from \cite{Shapley53}, which shows that it is the only method that satisfies certain good properties (\emph{axioms}). There are, however, a multiplicity of ways in which the Shapley value is operationalized for model explanation. These differ in how they reference the model, the training data, and the explanation context. Hence they differ in output, rendering the uniqueness result inapplicable. Furthermore, the techniques that rely on they training data produce non-intuitive attributions, for instance unused features can still receive attribution. In this paper, we use the axiomatic approach to study the differences between some of the many operationalizations of the

Shapley value for attribution. We discuss a technique called Baseline Shapley (BShap), provide a proper uniqueness result for it, and contrast it with two other techniques from prior literature, Integrated Gradients \cite{STY17} and Conditional Expectation Shapley \cite{Lundberg2017AUA}.

Multi-objective Bayesian Optimization using Pareto-frontier Entropy

- Shinya Suzuki, Shion Takeno, Tomoyuki Tamura, Kazuki Shitara, Masayuki Karasuyama
- abstract: This paper studies an entropy-based multi-objective Bayesian optimization (MBO). Existing entropy-based MBO methods need complicated approximations to evaluate entropy or employ over-simplification that ignores trade-off among objectives. We propose a novel entropy-based MBO called Pareto-frontier entropy search (PFES), which is based on the information gain of Pareto-frontier. We show that our entropy evaluation can be reduced to a closed form whose computation is quite simple while capturing the trade-off relation in Pareto-frontier. We further propose an extension for the "decoupled" setting, in which each objective function can be observed separately, and show that the PFES-based approach derives a natural extension of the original acquisition function which can also be evaluated simply. Our numerical experiments show effectiveness of PFES through several benchmark datasets, and real-word datasets from materials science.

The k-tied Normal Distribution: A Compact Parameterization of Gaussian Mean Field Posteriors in Bayesian Neural Networks

- Jakub Swiatkowski, Kevin Roth, Bastiaan Veeling, Linh Tran, Joshua Dillon, Jasper Snoek, Stephan Mandt, Tim Salimans, Rodolphe Jenatton, Sebastian Nowozin
- abstract: Variational Bayesian Inference is a popular methodology for approximating posterior distributions over Bayesian neural network weights. Recent work developing this class of methods has explored ever richer parameterizations of the approximate posterior in the hope of improving performance. In contrast, here we share a curious experimental finding that suggests instead restricting the variational distribution to a more compact parameterization. For a variety of deep Bayesian neural networks trained using Gaussian mean-field variational inference, we find that the posterior standard deviations consistently exhibit strong low-rank structure after convergence. This means that by decomposing these variational parameters into a low-rank factorization, we can make our variational approximation more compact without decreasing the models' performance. Furthermore, we find that such factorized parameterizations improve the signal-to-noise ratio of stochastic gradient estimates of the variational lower bound, resulting in faster convergence.

Multi-Agent Routing Value Iteration Network

- Quinlan Sykora, Mengye Ren, Raquel Urtasun
- abstract: In this paper we tackle the problem of routing multiple agents in a coordinated manner. This is a complex problem that has a wide range of applications in fleet management to achieve a common goal, such as mapping from a swarm of robots and ride sharing. Traditional methods are typically not designed for realistic environments which contain sparsely connected graphs and unknown traffic, and are often too slow in runtime to be practical. In contrast, we propose a graph neural network based model that is able to perform multi-agent routing based on learned value iteration in a sparsely connected graph with dynamically changing traffic conditions. Moreover, our learned communication module enables the agents to coordinate online and adapt to changes more effectively. We created a simulated environment to mimic realistic mapping performed by autonomous vehicles with unknown minimum edge coverage and traffic conditions; our approach significantly outperforms traditional solvers both in terms of total cost and runtime. We also show that our model trained with only two agents on graphs with a maximum of 25 nodes can easily generalize to situations with more agents and/or nodes.

Distinguishing Cause from Effect Using Quantiles: Bivariate Quantile Causal Discovery

- Natasa Tagasovska, Valérie Chavez-Demoulin, Thibault Vatter
- abstract: Causal inference using observational data is challenging, especially in the bivariate case. Through the minimum description length principle, we link the postulate of independence between the generating mechanisms of the cause and of the effect given the cause to quantile regression. Based on this theory, we develop Bivariate Quantile Causal Discovery (bQCD), a new method to distinguish cause from effect assuming no confounding, selection bias or feedback. Because it uses multiple quantile levels instead of the conditional mean only, bQCD is adaptive not only to additive, but also to multiplicative or even location-scale generating mechanisms. To illustrate the effectiveness of our approach, we perform an extensive empirical comparison on both synthetic and real datasets. This study shows that bQCD is robust across different implementations of the method (i.e., the quantile regression), computationally efficient, and compares favorably to state-of-the-art methods.

Quantized Decentralized Stochastic Learning over Directed Graphs

- Hossein Taheri, Aryan Mokhtari, Hamed Hassani, Ramtin Pedarsani
- abstract: We consider a decentralized stochastic learning problem where data points are distributed among computing nodes communicating over a directed graph. As the model size gets large, decentralized learning faces a major bottleneck that is the heavy communication load due to each node transmitting large messages (model updates) to its neighbors. To tackle this bottleneck, we propose the quantized decentralized stochastic learning algorithm over directed graphs that is based on the push-sum algorithm in decentralized consensus optimization. We prove that our algorithm achieves the same convergence rates of the decentralized stochastic learning algorithm with exact-communication for both convex and non-convex losses. Numerical evaluations corroborate our main theoretical results and illustrate significant speed-up compared to the exact-communication methods.

Multi-fidelity Bayesian Optimization with Max-value Entropy Search and its Parallelization

- Shion Takeno, Hitoshi Fukuoka, Yuhki Tsukada, Toshiyuki Koyama, Motoki Shiga, Ichiro Takeuchi, Masayuki Karasuyama
- abstract: In a standard setting of Bayesian optimization (BO), the objective function evaluation is assumed to be highly expensive. Multi-fidelity Bayesian optimization (MFBO) accelerates BO by incorporating lower fidelity observations available with a lower sampling cost. We propose a novel information-theoretic approach to MFBO, called multi-fidelity max-value entropy search (MF-MES), that enables us to obtain a more reliable evaluation of the information gain compared with existing information-based methods for MFBO. Further, we also propose a parallelization of MF-MES mainly for the asynchronous setting because queries typically occur asynchronously in MFBO due to a variety of sampling costs. We show that most of computations in our acquisition functions can be derived analytically, except for at most only two dimensional numerical integration that can be performed efficiently by simple approximations. We demonstrate effectiveness of our approach by using benchmark datasets and a real-world application to materials science data.

Fiedler Regularization: Learning Neural Networks with Graph Sparsity

- Edric Tam, David Dunson
- abstract: We introduce a novel regularization approach for deep learning that incorporates and respects the underlying graphical structure of the neural network. Existing regularization methods often focus on penalizing weights in a global/uniform manner that ignores the connectivity structure of the neural network. We propose to use the Fiedler value of the neural network's underlying graph as a tool for regularization. We provide theoretical support for this approach via spectral graph theory. We show several useful properties of the Fiedler value that make it suitable for regularization. We provide an approximate, variational approach for faster computation during training. We provide an alternative formulation of this framework in the form of a

structurally weighted L1 penalty, thus linking our approach to sparsity induction. We performed experiments on datasets that compare Fiedler regularization with traditional regularization methods such as Dropout and weight decay. Results demonstrate the efficacy of Fiedler regularization.

DropNet: Reducing Neural Network Complexity via Iterative Pruning

- Chong Min John Tan, Mehul Motani
- abstract: Modern deep neural networks require a significant amount of computing time and power to train and deploy, which limits their usage on edge devices. Inspired by the iterative weight pruning in the Lottery Ticket Hypothesis, we propose DropNet, an iterative pruning method which prunes nodes/filters to reduce network complexity. DropNet iteratively removes nodes/filters with the lowest average post-activation value across all training samples. Empirically, we show that DropNet is robust across a wide range of scenarios, including MLPs and CNNs using the MNIST, CIFAR-10 and Tiny ImageNet datasets. We show that up to 90% of the nodes/filters can be removed without any significant loss of accuracy. The final pruned network performs well even with reinitialisation of the weights and biases. DropNet also achieves similar accuracy to an oracle which greedily removes nodes/filters one at a time to minimise training loss, highlighting its effectiveness.

Reinforcement Learning for Integer Programming: Learning to Cut

- Yunhao Tang, Shipra Agrawal, Yuri Faenza
- abstract: Integer programming is a general optimization framework with a wide variety of applications, e.g., in scheduling, production planning, and graph optimization. As Integer Programs (IPs) model many provably hard to solve problems, modern IP solvers rely on heuristics. These heuristics are often human-designed, and tuned over time using experience and data. The goal of this work is to show that the performance of those solvers can be greatly enhanced using reinforcement learning (RL). In particular, we investigate a specific methodology for solving IPs, known as the Cutting Plane Method. This method is employed as a subroutine by all modern IP solvers. We present a deep RL formulation, network architecture, and algorithms for intelligent adaptive selection of cutting planes (aka cuts). Across a wide range of IP tasks, we show that our trained RL agent significantly outperforms human-designed heuristics, and effectively generalizes to larger instances and across IP problem classes. The trained agent is also demonstrated to benefit the popular downstream application of cutting plane methods in Branch-and-Cut algorithm, which is the backbone of state-of-the-art commercial IP solvers.

The Buckley-Osthus model and the block preferential attachment model: statistical analysis and application

- Wenpin Tang, Xin Guo, Fengmin Tang
- abstract: This paper is concerned with statistical estimation of two preferential attachment models: the Buckley-Osthus model and the block preferential attachment model. We prove that the maximum likelihood estimates for both models are consistent. We perform simulation studies to corroborate our theoretical findings. We also apply both models to study the evolution of a real-world network. A list of open problems are presented.

Clinician-in-the-Loop Decision Making: Reinforcement Learning with Near-Optimal Set-Valued Policies

- Shengpu Tang, Aditya Modi, Michael Sjoding, Jenna Wiens
- abstract: Standard reinforcement learning (RL) aims to find an optimal policy that identifies the best action for each state. However, in healthcare settings, many actions may be near-equivalent with respect to the reward (e.g., survival). We consider an alternative objective – learning set-valued policies to capture near-equivalent actions that lead to similar cumulative rewards. We propose a model-free algorithm based on temporal difference learning and a near-greedy heuristic for action selection. We analyze the theoretical properties of the proposed algorithm, providing optimality guarantees and demonstrate our approach on simulated environments and a real clinical task. Empirically, the proposed algorithm exhibits good convergence properties and discovers meaningful near-equivalent actions. Our work provides theoretical, as well as practical, foundations for clinician/human-in-the-loop decision making, in which humans (e.g., clinicians, patients) can incorporate additional knowledge (e.g., side effects, patient preference) when selecting among near-equivalent actions.

Taylor Expansion Policy Optimization

- Yunhao Tang, Michal Valko, Remi Munos
- abstract: In this work, we investigate the application of Taylor expansions in reinforcement learning. In particular, we propose Taylor Expansion Policy Optimization, a policy optimization formalism that generalizes prior work as a first-order special case. We also show that Taylor expansions intimately relate to off-policy evaluation. Finally, we show that this new formulation entails modifications which improve the performance of several state-of-the-art distributed algorithms.

Variational Imitation Learning with Diverse-quality Demonstrations

- Voot Tangkaratt, Bo Han, Mohammad Emtiyaz Khan, Masashi Sugiyama
- abstract: Learning from demonstrations can be challenging when the quality of demonstrations is diverse, and even more so when the quality is unknown and there is no additional information to estimate the quality. We propose a new method for imitation learning in such scenarios. We show that simple quality-estimation approaches might fail due to compounding error, and fix this issue by jointly estimating both the quality and reward using a variational approach. Our method is easy to implement within reinforcement-learning frameworks and also achieves state-of-the-art performance on continuous-control benchmarks.Our work enables scalable and data-efficient imitation learning under more realistic settings than before.

Learning disconnected manifolds: a no GAN’s land

- Ugo Tanielian, Thibaut Issenhuth, Elvis Dohmatob, Jeremie Mary
- abstract: Typical architectures of Generative Adversarial Networks make use of a unimodal latent/input distribution transformed by a continuous generator. Consequently, the modeled distribution always has connected support which is cumbersome when learning a disconnected set of manifolds. We formalize this problem by establishing a "no free lunch" theorem for the disconnected manifold learning stating an upper-bound on the precision of the targeted distribution. This is done by building on the necessary existence of a low-quality region where the generator continuously samples data between two disconnected modes. Finally, we derive a rejection sampling method based on the norm of generator’s Jacobian and show its efficiency on several generators including BigGAN.

No-Regret Exploration in Goal-Oriented Reinforcement Learning

- Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, Alessandro Lazaric
- abstract: Many popular reinforcement learning problems (e.g., navigation in a maze, some Atari games, mountain car) are instances of the episodic setting under its stochastic shortest path (SSP) formulation, where an agent has to achieve a goal state while minimizing the cumulative cost. Despite the popularity of this setting, the exploration-exploitation dilemma has been sparsely studied in general SSP problems, with most of the theoretical literature focusing on different problems (i.e., fixed-horizon and infinite-horizon) or making the restrictive loop-free SSP assumption (i.e., no state can be visited twice during an episode). In this paper, we study the general SSP problem with no assumption on its dynamics (some policies may actually never reach the goal). We introduce UC-SSP, the first no-regret algorithm in this setting, and prove a regret bound scaling as $\widetilde{\mathcal{O}}(D S \sqrt{ A D

K})$ after K episodes for any unknown SSP with S states, A actions, positive costs and SSP-diameter D, defined as the smallest expected hitting time from any starting state to the goal. We achieve this result by crafting a novel stopping rule, such that UC-SSP may interrupt the current policy if it is taking too long to achieve the goal and switch to alternative policies that are designed to rapidly terminate the episode.

Sparse Sinkhorn Attention

- Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, Da-Cheng Juan
- abstract: We propose Sparse Sinkhorn Attention, a new efficient and sparse method for learning to attend. Our method is based on differentiable sorting of internal representations. Concretely, we introduce a meta sorting network that learns to generate latent permutations over sequences. Given sorted sequences, we are then able to compute quasi-global attention with only local windows, improving the memory efficiency of the attention module. To this end, we propose new algorithmic innovations such as Causal Sinkhorn Balancing and SortCut, a dynamic sequence truncation method for tailoring Sinkhorn Attention for encoding and/or decoding purposes. Via extensive experiments on algorithmic seq2seq sorting, language modeling, pixel-wise image generation, document classification and natural language inference, we demonstrate that our memory efficient Sinkhorn Attention method is competitive with vanilla attention and consistently outperforms recently proposed efficient Transformer models such as Sparse Transformers.

Inductive Relation Prediction by Subgraph Reasoning

- Komal Teru, Etienne Denis, Will Hamilton
- abstract: The dominant paradigm for relation prediction in knowledge graphs involves learning and operating on latent representations (i.e., embeddings) of entities and relations. However, these embedding-based methods do not explicitly capture the compositional logical rules underlying the knowledge graph, and they are limited to the transductive setting, where the full set of entities must be known during training. Here, we propose a graph neural network based relation prediction framework, GraIL, that reasons over local subgraph structures and has a strong inductive bias to learn entity-independent relational semantics. Unlike embedding-based models, GraIL is naturally inductive and can generalize to unseen entities and graphs after training. We provide theoretical proof and strong empirical evidence that GraIL can rep-resent a useful subset of first-order logic and show that GraIL outperforms existing rule-induction baselines in the inductive setting. We also demonstrate significant gains obtained by ensembling GraIL with various knowledge graph embedding methods in the transductive setting, highlighting the complementary inductive bias of our method.

Few-shot Domain Adaptation by Causal Mechanism Transfer

- Takeshi Teshima, Issei Sato, Masashi Sugiyama
- abstract: We study few-shot supervised domain adaptation (DA) for regression problems, where only a few labeled target domain data and many labeled source domain data are available. Many of the current DA methods base their transfer assumptions on either parametrized distribution shift or apparent distribution similarities, e.g., identical conditionals or small distributional discrepancies. However, these assumptions may preclude the possibility of adaptation from intricately shifted and apparently very different distributions. To overcome this problem, we propose mechanism transfer, a meta-distributional scenario in which a data generating mechanism is invariant among domains. This transfer assumption can accommodate nonparametric shifts resulting in apparently different distributions while providing a solid statistical basis for DA. We take the structural equations in causal modeling as an example and propose a novel DA method, which is shown to be useful both theoretically and experimentally. Our method can be seen as the first attempt to fully leverage the invariance of structural causal models for DA.

Student Specialization in Deep Rectified Networks With Finite Width and Input Dimension

- Yuandong Tian
- abstract: We consider a deep ReLU / Leaky ReLU student network trained from the output of a fixed teacher network of the same depth, with Stochastic Gradient Descent (SGD). The student network is \emph{over-realized}: at each layer l, the number n_l of student nodes is more than that (m_l) of teacher. Under mild conditions on dataset and teacher network, we prove that when the gradient is small at every data sample, each teacher node is \emph{specialized} by at least one student node \emph{at the lowest layer}. For two-layer network, such specialization can be achieved by training on any dataset of \emph{polynomial} size $\mathcal{O}(K^{5/2} d^3 \epsilon^{-1})$. until the gradient magnitude drops to $\mathcal{O}(\epsilon/K^{3/2}\sqrt{d})$. Here d is the input dimension, $K = m_1 + n_1$ is the total number of neurons in the lowest layer of teacher and student. Note that we require a specific form of data augmentation and the sample complexity includes the additional data generated from augmentation. To our best knowledge, we are the first to give polynomial sample complexity for student specialization of training two-layer (Leaky) ReLU networks with finite depth and width in teacher-student setting, and finite complexity for the lowest layer specialization in multi-layer case, without parametric assumption of the input (like Gaussian). Our theory suggests that teacher nodes with large fan-out weights get specialized first when the gradient is still large, while others are specialized with small gradient, which suggests inductive bias in training. This shapes the stage of training as empirically observed in multiple previous works. Experiments on synthetic and CIFAR10 verify our findings. The code is released in \url{https://github.com/facebookresearch/luckmatters}.

Sequential Transfer in Reinforcement Learning with a Generative Model

- Andrea Tirinzoni, Riccardo Poiani, Marcello Restelli
- abstract: We are interested in how to design reinforcement learning agents that provably reduce the sample complexity for learning new tasks by transferring knowledge from previously-solved ones. The availability of solutions to related problems poses a fundamental trade-off: whether to seek policies that are expected to immediately achieve high (yet sub-optimal) performance in the new task or whether to seek information to quickly identify an optimal solution, potentially at the cost of poor initial behaviour. In this work, we focus on the second objective when the agent has access to a generative model of state-action pairs. First, given a set of solved tasks containing an approximation of the target one, we design an algorithm that quickly identifies an accurate solution by seeking the state-action pairs that are most informative for this purpose. We derive PAC bounds on its sample complexity which clearly demonstrate the benefits of using this kind of prior knowledge. Then, we show how to learn these approximate tasks sequentially by reducing our transfer setting to a hidden Markov model and employing spectral methods to recover its parameters. Finally, we empirically verify our theoretical findings in simple simulated domains.

Convolutional dictionary learning based auto-encoders for natural exponential-family distributions

- Bahareh Tolooshams, Andrew Song, Simona Temereanca, Demba Ba
- abstract: We introduce a class of auto-encoder neural networks tailored to data from the natural exponential family (e.g., count data). The architectures are inspired by the problem of learning the filters in a convolutional generative model with sparsity constraints, often referred to as convolutional dictionary learning (CDL). Our work is the first to combine ideas from convolutional generative models and deep learning for data that are naturally modeled with a non-Gaussian distribution (e.g., binomial and Poisson). This perspective provides us with a scalable and flexible framework that can be re-purposed for a wide range of tasks and assumptions on the generative model. Specifically, the iterative optimization procedure for solving CDL, an unsupervised task, is mapped to an unfolded and constrained neural network, with iterative adjustments to the inputs to account for the generative distribution. We also show that the framework can easily be extended for discriminative training, appropriate for a supervised task. We 1) demonstrate that fitting the generative model to learn, in an unsupervised fashion, the latent stimulus that underlies neural spiking data leads to better goodness-of-fit compared to other baselines, 2) show competitive performance compared to state-of-the-art algorithms for supervised Poisson image denoising, with significantly fewer parameters, and 3) characterize the gradient dynamics of the shallow binomial auto-encoder.

Multi-step Greedy Reinforcement Learning Algorithms

- Manan Tomar, Yonathan Efroni, Mohammad Ghavamzadeh
- abstract: Multi-step greedy policies have been extensively used in model-based reinforcement learning (RL), both when a model of the environment is available (e.g., in the game of Go) and when it is learned. In this paper, we explore their benefits in model-free RL, when employed using multi-step dynamic programming algorithms: κ-Policy Iteration (κ-PI) and κ-Value Iteration (κ-VI). These methods iteratively compute the next policy (κ-PI) and value function (κ-VI) by solving a surrogate decision problem with a shaped reward and a smaller discount factor. We derive model-free RL algorithms based on κ-PI and κ-VI in which the surrogate problem can be solved by any discrete or continuous action RL method, such as DQN and TRPO. We identify the importance of a hyper-parameter that controls the extent to which the surrogate problem is solved and suggest a way to set this parameter. When evaluated on a range of Atari and MuJoCo benchmark tasks, our results indicate that for the right range of κ, our algorithms outperform DQN and TRPO. This shows that our multi-step greedy algorithms are general enough to be applied over any existing RL algorithm and can significantly improve its performance.

Choice Set Optimization Under Discrete Choice Models of Group Decisions

- Kiran Tomlinson, Austin Benson
- abstract: The way that people make choices or exhibit preferences can be strongly affected by the set of available alternatives, often called the choice set. Furthermore, there are usually heterogeneous preferences, either at an individual level within small groups or within sub-populations of large groups. Given the availability of choice data, there are now many models that capture this behavior in order to make effective predictions—however, there is little work in understanding how directly changing the choice set can be used to influence the preferences of a collection of decision-makers. Here, we use discrete choice modeling to develop an optimization framework of such interventions for several problems of group influence, namely maximizing agreement or disagreement and promoting a particular choice. We show that these problems are NP-hard in general, but imposing restrictions reveals a fundamental boundary: promoting a choice can be easier than encouraging consensus or sowing discord. We design approximation algorithms for the hard problems and show that they work well on real-world choice data.

TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics

- Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, Smita Krishnaswamy
- abstract: It is increasingly common to encounter data in the form of cross-sectional population measurements over time, particularly in biomedical settings. Recent attempts to model individual trajectories from this data use optimal transport to create pairwise matchings between time points. However, these methods cannot model non-linear paths common in many underlying dynamic systems. We establish a link between continuous normalizing flows and dynamic optimal transport to model the expected paths of points over time. Continuous normalizing flows are generally under constrained, as they are allowed to take an arbitrary path from the source to the target distribution. We present \emph{TrajectoryNet}, which controls the continuous paths taken between distributions. We show how this is particularly applicable for studying cellular dynamics in data from single-cell RNA sequencing (scRNA-seq) technologies, and that TrajectoryNet improves upon recently proposed static optimal transport-based models that can be used for interpolating cellular distributions.

Alleviating Privacy Attacks via Causal Learning

- Shruti Tople, Amit Sharma, Aditya Nori
- abstract: Machine learning models, especially deep neural networks are known to be susceptible to privacy attacks such as membership inference where an adversary can detect whether a data point was used to train a model. Such privacy risks are exacerbated when a model is used for predictions on an unseen data distribution. To alleviate privacy attacks, we demonstrate the benefit of predictive models that are based on the causal relationships between input features and the outcome. We first show that models learnt using causal structure generalize better to unseen data, especially on data from different distributions than the train distribution. Based on this generalization property, we establish a theoretical link between causality and privacy: compared to associational models, causal models provide stronger differential privacy guarantees and are more robust to membership inference attacks. Experiments on simulated Bayesian networks and the colored-MNIST dataset show that associational models exhibit upto 80% attack accuracy under different test distributions and sample sizes whereas causal models exhibit attack accuracy close to a random guess.

Bayesian Learning from Sequential Data using Gaussian Processes with Signature Covariances

- Csaba Toth, Harald Oberhauser
- abstract: We develop a Bayesian approach to learning from sequential data by using Gaussian processes (GPs) with so-called signature kernels as covariance functions. This allows to make sequences of different length comparable and to rely on strong theoretical results from stochastic analysis. Signatures capture sequential structure with tensors that can scale unfavourably in sequence length and state space dimension. To deal with this, we introduce a sparse variational approach with inducing tensors. We then combine the resulting GP with LSTMs and GRUs to build larger models that leverage the strengths of each of these approaches and benchmark the resulting GPs on multivariate time series (TS) classification datasets.

Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations

- Florian Tramer, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, Joern-Henrik Jacobsen
- abstract: Adversarial examples are malicious inputs crafted to induce misclassification. Commonly studied \emph{sensitivity-based} adversarial examples introduce semantically-small changes to an input that result in a different model prediction. This paper studies a complementary failure mode, \emph{invariance-based} adversarial examples, that introduce minimal semantic changes that modify an input's true label yet preserve the model's prediction. We demonstrate fundamental tradeoffs between these two types of adversarial examples. We show that defenses against sensitivity-based attacks actively harm a model's accuracy on invariance-based attacks, and that new approaches are needed to resist both attack types. In particular, we break state-of-the-art adversarially-trained and \emph{certifiably-robust} models by generating small perturbations that the models are (provably) robust to, yet that change an input's class according to human labelers. Finally, we formally show that the existence of excessively invariant classifiers arises from the presence of \emph{overly-robust} predictive features in standard datasets.

Stochastic Gauss-Newton Algorithms for Nonconvex Compositional Optimization

- Quoc Tran-Dinh, Nhan Pham, Lam Nguyen
- abstract: We develop two new stochastic Gauss-Newton algorithms for solving a class of non-convex stochastic compositional optimization problems frequently arising in practice. We consider both the expectation and finite-sum settings under standard assumptions, and use both classical stochastic and SARAH estimators for approximating function values and Jacobians. In the expectation case, we establish $\BigO{\varepsilon^{-2}}$ iteration-complexity to achieve a stationary point in expectation and estimate the total number of stochastic oracle calls for both function value and its Jacobian, where ε is a desired accuracy. In the finite sum case, we also estimate $\BigO{\varepsilon^{-2}}$ iteration-complexity and the total oracle calls with high probability. To our best knowledge, this is the first time such global stochastic oracle complexity is established for stochastic Gauss-Newton methods. Finally, we illustrate our theoretical results via two numerical examples on both synthetic and real datasets.

Bayesian Differential Privacy for Machine Learning

- Aleksei Triastcyn, Boi Faltings
- abstract: Traditional differential privacy is independent of the data distribution. However, this is not well-matched with the modern machine learning context, where models are trained on specific data. As a result, achieving meaningful privacy guarantees in ML often excessively reduces accuracy. We propose Bayesian differential privacy (BDP), which takes into account the data distribution to provide more practical privacy guarantees. We also derive a general privacy accounting method under BDP, building upon the well-known moments accountant. Our experiments demonstrate that in-distribution samples in classic machine learning datasets, such as MNIST and CIFAR-10, enjoy significantly stronger privacy guarantees than postulated by DP, while models maintain high classification accuracy.

Single Point Transductive Prediction

- Nilesh Tripuraneni, Lester Mackey
- abstract: Standard methods in supervised learning separate training and prediction: the model is fit independently of any test points it may encounter. However, can knowledge of the next test point $\mathbf{x}_{\star}$ be exploited to improve prediction accuracy? We address this question in the context of linear prediction, showing how techniques from semi-parametric inference can be used transductively to combat regularization bias. We first lower bound the $\mathbf{x}_{\star}$ prediction error of ridge regression and the Lasso, showing that they must incur significant bias in certain test directions. We then provide non-asymptotic upper bounds on the $\mathbf{x}_{\star}$ prediction error of two transductive prediction rules. We conclude by showing the efficacy of our methods on both synthetic and real data, highlighting the improvements single point transductive prediction can provide in settings with distribution shift.

GraphOpt: Learning Optimization Models of Graph Formation

- Rakshit Trivedi, Jiachen Yang, Hongyuan Zha
- abstract: Formation mechanisms are fundamental to the study of complex networks, but learning them from observations is challenging. In real-world domains, one often has access only to the final constructed graph, instead of the full construction process, and observed graphs exhibit complex structural properties. In this work, we propose GraphOpt, an end-to-end framework that jointly learns an implicit model of graph structure formation and discovers an underlying optimization mechanism in the form of a latent objective function. The learned objective can serve as an explanation for the observed graph properties, thereby lending itself to transfer across different graphs within a domain. GraphOpt poses link formation in graphs as a sequential decision-making process and solves it using maximum entropy inverse reinforcement learning algorithm. Further, it employs a novel continuous latent action space that aids scalability. Empirically, we demonstrate that GraphOpt discovers a latent objective transferable across graphs with different characteristics. GraphOpt also learns a robust stochastic policy that achieves competitive link prediction performance without being explicitly trained on this task and further enables construction of graphs with properties similar to those of the observed graph.

Transfer Learning without Knowing: Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources

- Yun-Yun Tsai, Pin-Yu Chen, Tsung-Yi Ho
- abstract: Current transfer learning methods are mainly based on finetuning a pretrained model with target-domain data. Motivated by the techniques from adversarial machine learning (ML) that are capable of manipulating the model prediction via data perturbations, in this paper we propose a novel approach, black-box adversarial reprogramming (BAR), that repurposes a well-trained black-box ML model (e.g., a prediction API or a proprietary software) for solving different ML tasks, especially in the scenario with scarce data and constrained resources. The rationale lies in exploiting high-performance but unknown ML models to gain learning capability for transfer learning. Using zeroth order optimization and multi-label mapping techniques, BAR can reprogram a black-box ML model solely based on its input-output responses without knowing the model architecture or changing any parameter. More importantly, in the limited medical data setting, on autism spectrum disorder classification, diabetic retinopathy detection, and melanoma detection tasks, BAR outperforms state-of-the-art methods and yields comparable performance to the vanilla adversarial reprogramming method requiring complete knowledge of the target ML model. BAR also outperforms baseline transfer learning approaches by a significant margin, demonstrating cost-effective means and new insights for transfer learning.

From ImageNet to Image Classification: Contextualizing Progress on Benchmarks

- Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, Aleksander Madry
- abstract: Building rich machine learning datasets in a scalable manner often necessitates a crowd-sourced data collection pipeline. In this work, we use human studies to investigate the consequences of employing such a pipeline, focusing on the popular ImageNet dataset. We study how specific design choices in the ImageNet creation process impact the fidelity of the resulting dataset—including the introduction of biases that state-of-the-art models exploit. Our analysis pinpoints how a noisy data collection pipeline can lead to a systematic misalignment between the resulting benchmark and the real-world task it serves as a proxy for. Finally, our findings emphasize the need to augment our current model training and evaluation toolkit to take such misalignment into account.

Normalized Flat Minima: Exploring Scale Invariant Definition of Flat Minima for Neural Networks Using PAC-Bayesian Analysis

- Yusuke Tsuzuku, Issei Sato, Masashi Sugiyama
- abstract: The notion of flat minima has gained attention as a key metric of the generalization ability of deep learning models. However, current definitions of flatness are known to be sensitive to parameter rescaling. While some previous studies have proposed to rescale flatness metrics using parameter scales to avoid the scale dependence, the normalized metrics lose the direct theoretical connections between flat minima and generalization. In this paper, we first provide generalization error bounds using existing normalized flatness measures. Using the analysis, we then propose a novel normalized flatness metric. The proposed metric enjoys both direct theoretical connections and better empirical correlation to generalization error.

Approximating Stacked and Bidirectional Recurrent Architectures with the Delayed Recurrent Neural Network

- Javier Turek, Shailee Jain, Vy Vo, Mihai Capotă, Alexander Huth, Theodore Willke
- abstract: Recent work has shown that topological enhancements to recurrent neural networks (RNNs) can increase their expressiveness and representational capacity. Two popular enhancements are stacked RNNs, which increases the capacity for learning non-linear functions, and bidirectional processing, which exploits acausal information in a sequence. In this work, we explore the delayed-RNN, which is a single-layer RNN that has a delay between the input and output. We prove that a weight-constrained version of the delayed-RNN is equivalent to a stacked-RNN. We also show that the delay gives rise to partial acausality, much like bidirectional networks. Synthetic experiments confirm that the delayed-RNN can mimic bidirectional networks, solving some acausal tasks similarly, and outperforming them in others. Moreover, we show similar performance to bidirectional networks in a real-world natural language processing task. These results suggest that delayed-RNNs can approximate topologies including stacked RNNs, bidirectional RNNs, and stacked bidirectional RNNs – but with equivalent or faster runtimes for the delayed-RNNs.

Minimax Weight and Q-Function Learning for Off-Policy Evaluation

- Masatoshi Uehara, Jiawei Huang, Nan Jiang
- abstract: We provide theoretical investigations into off-policy evaluation in reinforcement learning using function approximators for (marginalized) importance weights and value functions. Our contributions include: (1) A new estimator, MWL, that directly estimates importance ratios over the state-action distributions, removing the reliance on knowledge of the behavior policy as in prior work (Liu et.al, 2018), (2) Another new estimator, MQL, obtained by swapping the roles of importance weights and value-functions in MWL. MQL has an intuitive interpretation of minimizing average Bellman errors and can be combined with MWL in a doubly robust manner, (3) Several additional results that offer further insights, including the sample complexities of MWL and MQL, their asymptotic optimality in the tabular setting, how the learned importance weights depend the choice of the discriminator class, and how our methods provide a unified view of some old and new algorithms in RL.

StochasticRank: Global Optimization of Scale-Free Discrete Functions

- Aleksei Ustimenko, Liudmila Prokhorenkova
- abstract: In this paper, we introduce a powerful and efficient framework for direct optimization of ranking metrics. The problem is ill-posed due to the discrete structure of the loss, and to deal with that, we introduce two important techniques: stochastic smoothing and novel gradient estimate based on partial integration. We show that classic smoothing approaches may introduce bias and present a universal solution for a proper debiasing. Importantly, we can guarantee global convergence of our method by adopting a recently proposed Stochastic Gradient Langevin Boosting algorithm. Our algorithm is implemented as a part of the CatBoost gradient boosting library and outperforms the existing approaches on several learning-to-rank datasets. In addition to ranking metrics, our framework applies to any scale-free discrete loss function.

Undirected Graphical Models as Approximate Posteriors

- Arash Vahdat, Evgeny Andriyash, William Macready
- abstract: The representation of the approximate posterior is a critical aspect of effective variational autoencoders (VAEs). Poor choices for the approximate posterior have a detrimental impact on the generative performance of VAEs due to the mismatch with the true posterior. We extend the class of posterior models that may be learned by using undirected graphical models. We develop an efficient method to train undirected approximate posteriors by showing that the gradient of the training objective with respect to the parameters of the undirected posterior can be computed by backpropagation through Markov chain Monte Carlo updates. We apply these gradient estimators for training discrete VAEs with Boltzmann machines as approximate posteriors and demonstrate that undirected models outperform previous results obtained using directed graphical models. Our implementation is publicly available.

Uncertainty Estimation Using a Single Deep Deterministic Neural Network

- Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, Yarin Gal
- abstract: We propose a method for training a deterministic deep model that can find and reject out of distribution data points at test time with a single forward pass. Our approach, deterministic uncertainty quantification (DUQ), builds upon ideas of RBF networks. We scale training in these with a novel loss function and centroid updating scheme and match the accuracy of softmax models. By enforcing detectability of changes in the input using a gradient penalty, we are able to reliably detect out of distribution data. Our uncertainty quantification scales well to large datasets, and using a single model, we improve upon or match Deep Ensembles in out of distribution detection on notable difficult dataset pairs such as FashionMNIST vs. MNIST, and CIFAR-10 vs. SVHN.

Deep Molecular Programming: A Natural Implementation of Binary-Weight ReLU Neural Networks

- Marko Vasic, Cameron Chalk, Sarfraz Khurshid, David Soloveichik
- abstract: Embedding computation in molecular contexts incompatible with traditional electronics is expected to have wide ranging impact in synthetic biology, medicine, nanofabrication and other fields. A key remaining challenge lies in developing programming paradigms for molecular computation that are well-aligned with the underlying chemical hardware and do not attempt to shoehorn ill-fitting electronics paradigms. We discover a surprisingly tight connection between a popular class of neural networks (binary-weight ReLU aka BinaryConnect) and a class of coupled chemical reactions that are absolutely robust to reaction rates. The robustness of rate-independent chemical computation makes it a promising target for bioengineering implementation. We show how a BinaryConnect neural network trained in silico using well-founded deep learning optimization techniques, can be compiled to an equivalent chemical reaction network, providing a novel molecular programming paradigm. We illustrate such translation on the paradigmatic IRIS and MNIST datasets. Toward intended applications of chemical computation, we further use our method to generate a chemical reaction network that can discriminate between different virus types based on gene expression levels. Our work sets the stage for rich knowledge transfer between neural network and molecular programming communities.

Linear bandits with Stochastic Delayed Feedback

- Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, Michael Brückner
- abstract: Stochastic linear bandits are a natural and well-studied model for structured exploration/exploitation problems and are widely used in applications such as on-line marketing and recommendation. One of the main challenges faced by practitioners hoping to apply existing algorithms is that usually the feedback is randomly delayed and delays are only partially observable. For example, while a purchase is usually observable some time after the display, the decision of not buying is never explicitly sent to the system. In other words, the learner only observes delayed positive events. We formalize this problem as a novel stochastic delayed linear bandit and propose OTFLinUCB and OTFLinTS, two computationally efficient algorithms able to integrate new information as it becomes available and to deal with the permanently censored feedback. We prove optimal O(d\sqrt{T}) bounds on the regret of the first algorithm and study the dependency on delay-dependent parameters. Our model, assumptions and results are validated by experiments on simulated and real data.

Non-Stationary Delayed Bandits with Intermediate Observations

- Claire Vernade, Andras Gyorgy, Timothy Mann
- abstract: Online recommender systems often face long delays in receiving feedback, especially when optimizing for some long-term metrics. While mitigating the effects of delays in learning is well-understood in stationary environments, the problem becomes much more challenging when the environment changes. In fact, if the timescale of the change is comparable to the delay, it is impossible to learn about the environment, since the available observations are already obsolete. However, the arising issues can be addressed if intermediate signals are available without delay, such that given those signals, the long-term behavior of the system is stationary. To model this situation, we introduce the problem of stochastic, non-stationary, delayed bandits with intermediate observations. We develop a computationally efficient algorithm based on UCRL, and prove sublinear regret guarantees for its performance. Experimental results demonstrate that our method is able to learn in non-stationary delayed environments where existing methods fail.

OPtions as REsponses: Grounding behavioural hierarchies in multi-agent reinforcement learning

- Alexander Vezhnevets, Yuhuai Wu, Maria Eckstein, Rémi Leblond, Joel Z Leibo

- abstract: This paper investigates generalisation in multi-agent games, where the generality of the agent can be evaluated by playing against opponents it hasn't seen during training. We propose two new games with concealed information and complex, non-transitive reward structure (think rock-paper-scissors). It turns out that most current deep reinforcement learning methods fail to efficiently explore the strategy space, thus learning policies that generalise poorly to unseen opponents. We then propose a novel hierarchical agent architecture, where the hierarchy is grounded in the game-theoretic structure of the game – the top level chooses strategic responses to opponents, while the low level implements them into policy over primitive actions. This grounding facilitates credit assignment across the levels of hierarchy. Our experiments show that the proposed hierarchical agent is capable of generalisation to unseen opponents, while conventional baselines fail to generalise whatsoever.

Born-Again Tree Ensembles

- Thibaut Vidal, Maximilian Schiffer
- abstract: The use of machine learning algorithms in finance, medicine, and criminal justice can deeply impact human lives. As a consequence, research into interpretable machine learning has rapidly grown in an attempt to better control and fix possible sources of mistakes and biases. Tree ensembles, in particular, offer a good prediction quality in various domains, but the concurrent use of multiple trees reduces the interpretability of the ensemble. Against this background, we study born-again tree ensembles, i.e., the process of constructing a single decision tree of minimum size that reproduces the exact same behavior as a given tree ensemble in its entire feature space. To find such a tree, we develop a dynamic-programming based algorithm that exploits sophisticated pruning and bounding rules to reduce the number of recursive calls. This algorithm generates optimal born-again trees for many datasets of practical interest, leading to classifiers which are typically simpler and more interpretable without any other form of compromise.

Private Reinforcement Learning with PAC and Regret Guarantees

- Giuseppe Vietri, Borja Balle, Akshay Krishnamurthy, Steven Wu
- abstract: Motivated by high-stakes decision-making domains like personalized medicine where user information is inherently sensitive, we design privacy preserving exploration policies for episodic reinforcement learning (RL). We first provide a meaningful privacy formulation using the notion of joint differential privacy (JDP)–a strong variant of differential privacy for settings where each user receives their own sets of output (e.g., policy recommendations). We then develop a private optimism-based learning algorithm that simultaneously achieves strong PAC and regret bounds, and enjoys a JDP guarantee. Our algorithm only pays for a moderate privacy cost on exploration: in comparison to the non-private bounds, the privacy parameter only appears in lower-order terms. Finally, we present lower bounds on sample complexity and regret for reinforcement learning subject to JDP.

New Oracle-Efficient Algorithms for Private Synthetic Data Release

- Giuseppe Vietri, Grace Tian, Mark Bun, Thomas Steinke, Steven Wu
- abstract: We present three new algorithms for constructing differentially private synthetic data—a sanitized version of a sensitive dataset that approximately preserves the answers to a large collection of statistical queries. All three algorithms are \emph{oracle-efficient} in the sense that they are computationally efficient when given access to an optimization oracle. Such an oracle can be implemented using many existing (non-private) optimization tools such as sophisticated integer program solvers. While the accuracy of the synthetic data is contingent on the oracle's optimization performance, the algorithms satisfy differential privacy even in the worst case. For all three algorithms, we provide theoretical guarantees for both accuracy and privacy. Through empirical evaluation, we demonstrate that our methods scale well with both the dimensionality of the data and the number of queries. Compared to the state-of-the-art method High-Dimensional Matrix Mechanism (McKenna et al. VLDB 2018), our algorithms provide better accuracy in the large workload and high privacy regime (corresponding to low privacy loss $\eps$).

Conditional gradient methods for stochastically constrained convex minimization

- Maria-Luiza Vladarean, Ahmet Alacaoglu, Ya-Ping Hsieh, Volkan Cevher
- abstract: We propose two novel conditional gradient-based methods for solving structured stochastic convex optimization problems with a large number of linear constraints. Instances of this template naturally arise from SDP-relaxations of combinatorial problems, which involve a number of constraints that is polynomial in the problem dimension. The most important feature of our framework is that only a subset of the constraints is processed at each iteration, thus gaining a computational advantage over prior works that require full passes. Our algorithms rely on variance reduction and smoothing used in conjunction with conditional gradient steps, and are accompanied by rigorous convergence guarantees. Preliminary numerical experiments are provided for illustrating the practical performance of the methods.

Unsupervised Discovery of Interpretable Directions in the GAN Latent Space

- Andrey Voynov, Artem Babenko
- abstract: The latent spaces of GAN models often have semantically meaningful directions. Moving in these directions corresponds to human-interpretable image transformations, such as zooming or recoloring, enabling a more controllable generation process. However, the discovery of such directions is currently performed in a supervised manner, requiring human labels, pretrained models, or some form of self-supervision. These requirements severely restrict a range of directions existing approaches can discover. In this paper, we introduce an unsupervised method to identify interpretable directions in the latent space of a pretrained GAN model. By a simple model-agnostic procedure, we find directions corresponding to sensible semantic manipulations without any form of (self-)supervision. Furthermore, we reveal several non-trivial findings, which would be difficult to obtain by existing methods, e.g., a direction corresponding to background removal. As an immediate practical benefit of our work, we show how to exploit this finding to achieve competitive performance for weakly-supervised saliency detection. The implementation of our method is available online.

Safe Reinforcement Learning in Constrained Markov Decision Processes

- Akifumi Wachi, Yanan Sui
- abstract: Safe reinforcement learning has been a promising approach for optimizing the policy of an agent that operates in safety-critical applications. In this paper, we propose an algorithm, SNO-MDP, that explores and optimizes Markov decision processes under unknown safety constraints. Specifically, we take a step-wise approach for optimizing safety and cumulative reward. In our method, the agent first learns safety constraints by expanding the safe region, and then optimizes the cumulative reward in the certified safe region. We provide theoretical guarantees on both the satisfaction of the safety constraint and the near-optimality of the cumulative reward under proper regularity assumptions. In our experiments, we demonstrate the effectiveness of SNO-MDP through two experiments: one uses a synthetic data in a new, openly-available environment named GP-Safety-Gym, and the other simulates Mars surface exploration by using real observation data.

Orthogonalized SGD and Nested Architectures for Anytime Neural Networks

- Chengcheng Wan, Henry Hoffmann, Shan Lu, Michael Maire
- abstract: We propose a novel variant of SGD customized for training network architectures that support anytime behavior: such networks produce a series of increasingly accurate outputs over time. Efficient architectural designs for these networks focus on re-using internal state; subnetworks must produce representations relevant for both imme- diate prediction as well as refinement by subse- quent network stages. We consider traditional branched networks as well as a new class of re- cursively nested networks. Our new optimizer, Orthogonalized SGD, dynamically re-balances task-specific gradients when

training a multitask network. In the context of anytime architectures, this optimizer projects gradients from later out- puts onto a parameter subspace that does not in- terfere with those from earlier outputs. Experi- ments demonstrate that training with Orthogonal- ized SGD significantly improves generalization accuracy of anytime networks.

Projection-free Distributed Online Convex Optimization with $O(\sqrtT)$ Communication Complexity

- Yuanyu Wan, Wei-Wei Tu, Lijun Zhang
- abstract: To deal with complicated constraints via locally light computations in distributed online learning, a recent study has presented a projection-free algorithm called distributed online conditional gradient (D-OCG), and achieved an $O(T^{3/4})$ regret bound, where T is the number of prediction rounds. However, in each round, the local learners of D-OCG need to communicate with their neighbors to share the local gradients, which results in a high communication complexity of $O(T)$. In this paper, we first propose an improved variant of D-OCG, namely D-BOCG, which enjoys an $O(T^{3/4})$ regret bound with only $O(\sqrt{T})$ communication complexity. The key idea is to divide the total prediction rounds into $\sqrt{T}$ equally-sized blocks, and only update the local learners at the beginning of each block by performing iterative linear optimization steps. Furthermore, to handle the more challenging bandit setting, in which only the loss value is available, we incorporate the classical one-point gradient estimator into D-BOCG, and obtain similar theoretical guarantees.

Logistic Regression for Massive Data with Rare Events

- Haiying Wang
- abstract: This paper studies binary logistic regression for rare events data, or imbalanced data, where the number of events (observations in one class, often called cases) is significantly smaller than the number of nonevents (observations in the other class, often called controls). We first derive the asymptotic distribution of the maximum likelihood estimator (MLE) of the unknown parameter, which shows that the asymptotic variance convergences to zero in a rate of the inverse of the number of the events instead of the inverse of the full data sample size, indicating that the available information in rare events data is at the scale of the number of events instead of the full data sample size. Furthermore, we prove that under-sampling a small proportion of the nonevents, the resulting under-sampled estimator may have identical asymptotic distribution to the full data MLE. This demonstrates the advantage of under-sampling nonevents for rare events data, because this procedure may significantly reduce the computation and/or data collection costs. Another common practice in analyzing rare events data is to over-sample (replicate) the events, which has a higher computational cost. We show that this procedure may even result in efficiency loss in terms of parameter estimation.

On the Global Optimality of Model-Agnostic Meta-Learning

- Lingxiao Wang, Qi Cai, Zhuoran Yang, Zhaoran Wang
- abstract: Model-agnostic meta-learning (MAML) formulates meta-learning as a bilevel optimization problem, where the inner level solves each subtask based on a shared prior, while the outer level searches for the optimal shared prior by optimizing its aggregated performance over all the subtasks. Despite its empirical success, MAML remains less understood in theory, especially in terms of its global optimality, due to the nonconvexity of the meta-objective (the outer-level objective). To bridge such a gap between theory and practice, we characterize the optimality gap of the stationary points attained by MAML for both reinforcement learning and supervised learning, where the inner-level and outer-level problems are solved via first-order optimization methods. In particular, our characterization connects the optimality gap of such stationary points with (i) the functional geometry of inner-level objectives and (ii) the representation power of function approximators, including linear models and neural networks. To the best of our knowledge, our analysis establishes the global optimality of MAML with nonconvex meta-objectives for the first time.

Towards Accurate Post-training Network Quantization via Bit-Split and Stitching

- Peisong Wang, Qiang Chen, Xiangyu He, Jian Cheng
- abstract: Network quantization is essential for deploying deep models to IoT devices due to its high efficiency. Most existing quantization approaches rely on the full training datasets and the time-consuming fine-tuning to retain accuracy. Post-training quantization does not have these problems, however, it has mainly been shown effective for 8-bit quantization due to the simple optimization strategy. In this paper, we propose a Bit-Split and Stitching framework (Bit-split) for lower-bit post-training quantization with minimal accuracy degradation. The proposed framework is validated on a variety of computer vision tasks, including image classification, object detection, instance segmentation, with various network architectures. Specifically, Bit-split can achieve near-original model performance even when quantizing FP32 models to INT3 without fine-tuning.

Self-Modulating Nonparametric Event-Tensor Factorization

- Zheng Wang, Xinqi Chu, Shandian Zhe
- abstract: Tensor factorization is a fundamental framework to analyze high-order interactions in data. Despite the success of the existing methods, the valuable temporal information are severely underused. The timestamps of the interactions are either ignored or discretized into crude steps. The recent work although formulates event-tensors to keep the timestamps in factorization and can capture mutual excitation effects among the interaction events, it overlooks another important type of temporal influence, inhibition. In addition, it uses a local window to exclude all the long-term dependencies. To overcome these limitations, we propose a self-modulating nonparametric Bayesian factorization model. We use the latent factors to construct mutually governed, general random point processes, which can capture various short-term/long-term, excitation/inhibition effects, so as to encode the complex temporal dependencies into factor representations. In addition, our model couples with a latent Gaussian process to estimate and fuse nonlinear yet static relationships between the entities. For efficient inference, we derive a fully decomposed model evidence lower bound to dispense with the huge kernel matrix and costly summations inside the rate and log rate functions. We then develop an efficient stochastic optimization algorithm. We show the advantage of our method in four real-world applications.

Upper bounds for Model-Free Row-Sparse Principal Component Analysis

- Guanyi Wang, Santanu Dey
- abstract: Sparse principal component analysis (PCA) is a widely-used dimensionality reduction tool in statistics and machine learning. Most methods mentioned in literature are either heuristics for good primal feasible solutions under statistical assumptions or ADMM-type algorithms with stationary/critical points convergence property for the regularized reformulation of sparse PCA. However, none of these methods can efficiently verify the quality of the solutions via comparing current objective values with their dual bounds, especially in model-free case. We propose a new framework that finds out upper (dual) bounds for the sparse PCA within polynomial time via solving a convex integer program (IP). We show that, in the worst-case, the dual bounds provided by the convex IP is within an affine function of the global optimal value. Moreover, in contrast to the semi-definition relaxation, this framework is much easier to scale on large cases. Numerical results on both artificial and real cases are reported to demonstrate the advantages of our method.

ROMA: Multi-Agent Reinforcement Learning with Emergent Roles

- Tonghan Wang, Heng Dong, Victor Lesser, Chongjie Zhang

- abstract: The role concept provides a useful tool to design and understand complex multi-agent systems, which allows agents with a similar role to share similar behaviors. However, existing role-based methods use prior domain knowledge and predefine role structures and behaviors. In contrast, multi-agent reinforcement learning (MARL) provides flexibility and adaptability, but less efficiency in complex tasks. In this paper, we synergize these two paradigms and propose a role-oriented MARL framework (ROMA). In this framework, roles are emergent, and agents with similar roles tend to share their learning and to be specialized on certain sub-tasks. To this end, we construct a stochastic role embedding space by introducing two novel regularizers and conditioning individual policies on roles. Experiments show that our method can learn specialized, dynamic, and identifiable roles, which help our method push forward the state of the art on the StarCraft II micromanagement benchmark. Demonstrative videos are available at https://sites.google.com/view/romarl/.

Non-separable Non-stationary random fields

- Kangrui Wang, Oliver Hamelijnck, Theodoros Damoulas, Mark Steel
- abstract: We describe a framework for constructing nonstationary nonseparable random fields based on an infinite mixture of convolved stochastic processes. When the mixing process is stationary but the convolution function is nonstationary we arrive at nonseparable kernels with constant non-separability that are available in closed form. When the mixing is nonstationary and the convolution function is stationary we arrive at nonseparable random fields that have varying nonseparability and better preserve local structure. These fields have natural interpretations through the spectral representation of stochastic differential equations (SDEs) and are demonstrated on a range of synthetic benchmarks and spatio-temporal applications in geostatistics and machine learning. We show how a single Gaussian process (GP) with these random fields can computationally and statistically outperform both separable and existing nonstationary nonseparable approaches such as treed GPs and deep GP constructions.

Continuously Indexed Domain Adaptation

- Hao Wang, Hao He, Dina Katabi
- abstract: Existing domain adaptation focuses on transferring knowledge between domains with categorical indices (e.g., between datasets A and B). However, many tasks involve continuously indexed domains. For example, in medical applications, one often needs to transfer disease analysis and prediction across patients of different ages, where age acts as a continuous domain index. Such tasks are challenging for prior domain adaptation methods since they ignore the underlying relation among domains. In this paper, we propose the first method for continuously indexed domain adaptation. Our approach combines traditional adversarial adaptation with a novel discriminator that models the encoding-conditioned domain index distribution. Our theoretical analysis demonstrates the value of leveraging the domain index to generate invariant features across a continuous range of domains. Our empirical results show that our approach outperforms the state-of-the-art domain adaption methods on both synthetic and real-world medical datasets.

Learning Efficient Multi-agent Communication: An Information Bottleneck Approach

- Rundong Wang, Xu He, Runsheng Yu, Wei Qiu, Bo An, Zinovi Rabinovich
- abstract: We consider the problem of the limited-bandwidth communication for multi-agent reinforcement learning, where agents cooperate with the assistance of a communication protocol and a scheduler. The protocol and scheduler jointly determine which agent is communicating what message and to whom. Under the limited bandwidth constraint, a communication protocol is required to generate informative messages. Meanwhile, an unnecessary communication connection should not be established because it occupies limited resources in vain. In this paper, we develop an Informative Multi-Agent Communication (IMAC) method to learn efficient communication protocols as well as scheduling. First, from the perspective of communication theory, we prove that the limited bandwidth constraint requires low-entropy messages throughout the transmission. Then inspired by the information bottleneck principle, we learn a valuable and compact communication protocol and a weight-based scheduler. To demonstrate the efficiency of our method, we conduct extensive experiments in various cooperative and competitive multi-agent tasks with different numbers of agents and different bandwidths. We show that IMAC converges faster and leads to efficient communication among agents under the limited bandwidth as compared to many baseline methods.

Frustratingly Simple Few-Shot Object Detection

- Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, Fisher Yu
- abstract: Detecting rare objects from a few examples is an emerging problem. Prior works show meta-learning is a promising approach. But, fine-tuning techniques have drawn scant attention. We find that fine-tuning only the last layer of existing detectors on rare classes is crucial to the few-shot object detection task. Such a simple approach outperforms the meta-learning methods by roughly 2 20 points on current benchmarks and sometimes even doubles the accuracy of the prior methods. However, the high variance in the few samples often leads to the unreliability of existing benchmarks. We revise the evaluation protocols by sampling multiple groups of training examples to obtain stable comparisons and build new benchmarks based on three datasets: PASCAL VOC, COCO and LVIS. Again, our fine-tuning approach establishes a new state of the art on the revised benchmarks. The code as well as the pretrained models are available at https://github.com/ucbdrive/few-shot-object-detection.

Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere

- Tongzhou Wang, Phillip Isola
- abstract: Contrastive representation learning has been outstandingly successful in practice. In this work, we identify two key properties related to the contrastive loss: (1) alignment (closeness) of features from positive pairs, and (2) uniformity of the induced distribution of the (normalized) features on the hypersphere. We prove that, asymptotically, the contrastive loss optimizes these properties, and analyze their positive effects on downstream tasks. Empirically, we introduce an optimizable metric to quantify each property. Extensive experiments on standard vision and language datasets confirm the strong agreement between both metrics and downstream task performance. Directly optimizing for these two metrics leads to representations with comparable or better performance at downstream tasks than contrastive learning.

Enhanced POET: Open-ended Reinforcement Learning through Unbounded Invention of Learning Challenges and their Solutions

- Rui Wang, Joel Lehman, Aditya Rawal, Jiale Zhi, Yulun Li, Jeffrey Clune, Kenneth Stanley
- abstract: Creating open-ended algorithms, which generate their own never-ending stream of novel and appropriately challenging learning opportunities, could help to automate and accelerate progress in machine learning. A recent step in this direction is the Paired Open-Ended Trailblazer (POET), an algorithm that generates and solves its own challenges, and allows solutions to goal-switch between challenges to avoid local optima. However, the original POET was unable to demonstrate its full creative potential because of limitations of the algorithm itself and because of external issues including a limited problem space and lack of a universal progress measure. Importantly, both limitations pose impediments not only for POET, but for the pursuit of open-endedness in general. Here we introduce and empirically validate two new innovations to the original algorithm, as well as two external innovations designed to help elucidate its full potential. Together, these four advances enable the most open-ended algorithmic demonstration to date. The algorithmic innovations are (1) a domain-general measure of how meaningfully novel new challenges are, enabling the system to potentially create and solve interesting challenges endlessly, and (2) an efficient heuristic for determining when agents should goal-switch from one problem to another (helping open-ended search better scale). Outside the algorithm itself, to enable a more definitive demonstration of open-endedness, we introduce (3) a novel, more flexible way to encode environmental challenges, and (4) a generic measure of the extent to which a system continues to exhibit open-ended innovation.

Enhanced POET produces a diverse range of sophisticated behaviors that solve a wide range of environmental challenges, many of which cannot be solved through other means.

Haar Graph Pooling

- Yu Guang Wang, Ming Li, Zheng Ma, Guido Montufar, Xiaosheng Zhuang, Yanan Fan
- abstract: Deep Graph Neural Networks (GNNs) are useful models for graph classification and graph-based regression tasks. In these tasks, graph pooling is a critical ingredient by which GNNs adapt to input graphs of varying size and structure. We propose a new graph pooling operation based on compressive Haar transforms — \emph{HaarPooling}. HaarPooling implements a cascade of pooling operations; it is computed by following a sequence of clusterings of the input graph. A HaarPooling layer transforms a given input graph to an output graph with a smaller node number and the same feature dimension; the compressive Haar transform filters out fine detail information in the Haar wavelet domain. In this way, all the HaarPooling layers together synthesize the features of any given input graph into a feature vector of uniform size. Such transforms provide a sparse characterization of the data and preserve the structure information of the input graph. GNNs implemented with standard graph convolution layers and HaarPooling layers achieve state of the art performance on diverse graph classification and regression problems.

Deep Streaming Label Learning

- Zhen Wang, Liu Liu, Dacheng Tao
- abstract: In multi-label learning, each instance can be associated with multiple and non-exclusive labels. Previous studies assume that all the labels in the learning process are fixed and static; however, they ignore the fact that the labels will emerge continuously in changing environments. In order to fill in these research gaps, we propose a novel deep neural network (DNN) based framework, Deep Streaming Label Learning (DSLL), to classify instances with newly emerged labels effectively. DSLL can explore and incorporate the knowledge from past labels and historical models to understand and develop emerging new labels. DSLL consists of three components: 1) a streaming label mapping to extract deep relationships between new labels and past labels with a novel label-correlation aware loss; 2) a streaming feature distillation propagating feature-level knowledge from the historical model to a new model; 3) a senior student network to model new labels with the help of knowledge learned from the past. Theoretically, we prove that DSLL admits tight generalization error bounds for new labels in the DNN framework. Experimentally, extensive empirical results show that the proposed method performs significantly better than the existing state-of-the-art multi-label learning methods to handle the continually emerging new labels.

BoXHED: Boosted eXact Hazard Estimator with Dynamic covariates

- Xiaochen Wang, Arash Pakbin, Bobak Mortazavi, Hongyu Zhao, Donald Lee
- abstract: The proliferation of medical monitoring devices makes it possible to track health vitals at high frequency, enabling the development of dynamic health risk scores that change with the underlying readings. Survival analysis, in particular hazard estimation, is well-suited to analyzing this stream of data to predict disease onset as a function of the time-varying vitals. This paper introduces the software package BoXHED (pronounced ‘box-head’) for nonparametrically estimating hazard functions via gradient boosting. BoXHED 1.0 is a novel tree-based implementation of the generic estimator proposed in Lee et al. (2017), which was designed for handling time-dependent covariates in a fully nonparametric manner. BoXHED is also the first publicly available software implementation for Lee et al. (2017). Applying it to a cardiovascular disease dataset from the Framingham Heart Study reveals novel interaction effects among known risk factors, potentially resolving an open question in clinical literature. BoXHED is available from GitHub: www.github.com/BoXHED.

Optimizing Data Usage via Differentiable Rewards

- Xinyi Wang, Hieu Pham, Paul Michel, Antonios Anastasopoulos, Jaime Carbonell, Graham Neubig
- abstract: To acquire a new skill, humans learn better and faster if a tutor, based on their current knowledge level, informs them of how much attention they should pay to particular content or practice problems. Similarly, a machine learning model could potentially be trained better with a scorer that “adapts” to its current learning state and estimates the importance of each training data instance. Training such an adaptive scorer efficiently is a challenging problem; in order to precisely quantify the effect of a data instance at a given time during the training, it is typically necessary to first complete the entire training process. To efficiently optimize data usage, we propose a reinforcement learning approach called Differentiable Data Selection (DDS). In DDS, we formulate a scorer network as a learnable function of the training data, which can be efficiently updated along with the main model being trained. Specifically, DDS updates the scorer with an intuitive reward signal: it should up-weigh the data that has a similar gradient with a dev set upon which we would finally like to perform well. Without significant computing overhead, DDS delivers strong and consistent improvements over several strong baselines on two very different tasks of machine translation and image classification.

Bandits for BMO Functions

- Tianyu Wang, Cynthia Rudin
- abstract: We study the bandit problem where the underlying expected reward is a Bounded Mean Oscillation (BMO) function. BMO functions are allowed to be discontinuous and unbounded, and are useful in modeling signals with singularities in the domain. We develop a toolset for BMO bandits, and provide an algorithm that can achieve poly-log δ-regret – a regret measured against an arm that is optimal after removing a δ-sized portion of the arm space.

When deep denoising meets iterative phase retrieval

- Yaotian Wang, Xiaohang Sun, Jason Fleischer
- abstract: Recovering a signal from its Fourier intensity underlies many important applications, including lensless imaging and imaging through scattering media. Conventional algorithms for retrieving the phase suffer when noise is present but display global convergence when given clean data. Neural networks have been used to improve algorithm robustness, but efforts to date are sensitive to initial conditions and give inconsistent performance. Here, we combine iterative methods from phase retrieval with image statistics from deep denoisers, via regularization-by-denoising. The resulting methods inherit the advantages of each approach and outperform other noise-robust phase retrieval algorithms. Our work paves the way for hybrid imaging methods that integrate machine-learned constraints in conventional algorithms.

Doubly Stochastic Variational Inference for Neural Processes with Hierarchical Latent Variables

- Qi Wang, Herke Van Hoof
- abstract: Neural processes (NPs) constitute a family of variational approximate models for stochastic processes with promising properties in computational efficiency and uncertainty quantification. These processes use neural networks with latent variable inputs to induce a predictive distribution. However, the expressiveness of vanilla NPs is limited as they only use a global latent variable, while target-specific local variation may be crucial sometimes. To address this challenge, we investigate NPs systematically and present a new variant of NP model that we call Doubly Stochastic Variational Neural Process (DSVNP). This model combines the global latent variable and local latent variables for prediction. We evaluate this model in several experiments, and our results demonstrate competitive prediction performance in multi-output regression and uncertainty estimation in classification.

Loss Function Search for Face Recognition

- Xiaobo Wang, Shuo Wang, Cheng Chi, Shifeng Zhang, Tao Mei
- abstract: In face recognition, designing margin-based (\emph{e.g.}, angular, additive, additive angular margins) softmax loss functions plays an important role to learn discriminative features. However, these hand-crafted heuristic methods may be sub-optimal because they require much effort to explore the large design space. Recently, an AutoML for loss function search method AM-LFS has been derived, which leverages reinforcement learning to search loss functions during the training process. But its search space is complex and unstable that hindering its superiority. In this paper, we first analyze that the key to enhance the feature discrimination is actually \textbf{how to reduce the softmax probability}. We then design a unified formulation for the current margin-based softmax losses. Accordingly, we define a novel search space and develop a reward-guided search method to automatically obtain the best candidate. Experimental results on a variety of face recognition benchmarks have demonstrated the effectiveness of our method over the state-of-the-art alternatives.

Sequential Cooperative Bayesian Inference

- Junqi Wang, Pei Wang, Patrick Shafto
- abstract: Cooperation is often implicitly assumed when learning from other agents. Cooperation implies that the agent selecting the data, and the agent learning from the data, have the same goal, that the learner infer the intended hypothesis. Recent models in human and machine learning have demonstrated the possibility of cooperation. We seek foundational theoretical results for cooperative inference by Bayesian agents through sequential data. We develop novel approaches analyzing consistency, rate of convergence and stability of Sequential Cooperative Bayesian Inference (SCBI). Our analysis of the effectiveness, sample efficiency and robustness show that cooperation is not only possible but theoretically well-founded. We discuss implications for human-human and human-machine cooperation.

Neural Network Control Policy Verification With Persistent Adversarial Perturbation

- Yuh-Shyang Wang, Lily Weng, Luca Daniel
- abstract: Deep neural networks are known to be fragile to small adversarial perturbations, which raises serious concerns when a neural network policy is interconnected with a physical system in a closed loop. In this paper, we show how to combine recent works on static neural network certification tools with robust control theory to certify a neural network policy in a control loop. We give a sufficient condition and an algorithm to ensure that the closed loop state and control constraints are satisfied when the persistent adversarial perturbation is l-infinity norm bounded. Our method is based on finding a positively invariant set of the closed loop dynamical system, and thus we do not require the continuity of the neural network policy. Along with the verification result, we also develop an effective attack strategy for neural network control systems that outperforms exhaustive Monte-Carlo search significantly. We show that our certification algorithm works well on learned models and could achieve 5 times better result than the traditional Lipschitz-based method to certify the robustness of a neural network policy on the cart-pole balance control problem.

Cost-effectively Identifying Causal Effects When Only Response Variable is Observable

- Tian-Zuo Wang, Xi-Zhu Wu, Sheng-Jun Huang, Zhi-Hua Zhou
- abstract: In many real tasks, we care about how to make decisions rather than mere predictions on an event, e.g. how to increase the revenue next month instead of merely knowing it will drop. The key is to identify the causal effects on the desired event. It is achievable with do-calculus if the causal structure is known; however, in many real tasks it is not easy to infer the whole causal structure with the observational data. Introducing external interventions is needed to achieve it. In this paper, we study the situation where only the response variable is observable under intervention. We propose a novel approach which is able to cost-effectively identify the causal effects, by an active strategy introducing limited interventions, and thus guide decision-making. Theoretical analysis and empirical studies validate the effectiveness of the proposed approach.

Striving for Simplicity and Performance in Off-Policy DRL: Output Normalization and Non-Uniform Sampling

- Che Wang, Yanqiu Wu, Quan Vuong, Keith Ross
- abstract: We aim to develop off-policy DRL algorithms that not only exceed state-of-the-art performance but are also simple and minimalistic. For standard continuous control benchmarks, Soft Actor-Critic (SAC), which employs entropy maximization, currently provides state-of-the-art performance. We first demonstrate that the entropy term in SAC addresses action saturation due to the bounded nature of the action spaces, with this insight, we propose a streamlined algorithm with a simple normalization scheme or with inverted gradients. We show that both approaches can match SAC's sample efficiency performance without the need of entropy maximization, we then propose a simple non-uniform sampling method for selecting transitions from the replay buffer during training. Extensive experimental results demonstrate that our proposed sampling scheme leads to state of the art sample efficiency on challenging continuous control tasks. We combine all of our findings into one simple algorithm, which we call Streamlined Off Policy with Emphasizing Recent Experience, for which we provide robust public-domain code.

On Differentially Private Stochastic Convex Optimization with Heavy-tailed Data

- Di Wang, Hanshen Xiao, Srinivas Devadas, Jinhui Xu
- abstract: In this paper, we consider the problem of designing Differentially Private (DP) algorithms for Stochastic Convex Optimization (SCO) on heavy-tailed data. The irregularity of such data violates some key assumptions used in almost all existing DP-SCO and DP-ERM methods, resulting in failure to provide the DP guarantees. To better understand this type of challenges, we provide in this paper a comprehensive study of DP-SCO under various settings. First, we consider the case where the loss function is strongly convex and smooth. For this case, we propose a method based on the sample-and-aggregate framework, which has an excess population risk of $\tilde{O}(\frac{d^3}{n\epsilon^4})$ (after omitting other factors), where n is the sample size and d is the dimensionality of the data. Then, we show that with some additional assumptions on the loss functions, it is possible to reduce the \emph{expected} excess population risk to $\tilde{O}(\frac{ d^2}{ n\epsilon^2 })$. To lift these additional conditions, we also provide a gradient smoothing and trimming based scheme to achieve excess population risks of $\tilde{O}(\frac{ d^2}{n\epsilon^2})$ and $\tilde{O}(\frac{d^\frac{2}{3}}{(n\epsilon^2)^\frac{1}{3}})$ for strongly convex and general convex loss functions, respectively, \emph{with high probability}. Experiments on both synthetic and real-world datasets suggest that our algorithms can effectively deal with the challenges caused by data irregularity.

Breaking the Curse of Many Agents: Provable Mean Embedding Q-Iteration for Mean-Field Reinforcement Learning

- Lingxiao Wang, Zhuoran Yang, Zhaoran Wang
- abstract: Multi-agent reinforcement learning (MARL) achieves significant empirical successes. However, MARL suffers from the curse of many agents. In this paper, we exploit the symmetry of agents in MARL. In the most generic form, we study a mean-field MARL problem. Such a mean-field MARL is defined on mean-field states, which are distributions that are supported on continuous space. Based on the mean embedding of the distributions, we propose MF-FQI algorithm, which solves the mean-field MARL and establishes a non-asymptotic analysis for MF-FQI algorithm. We highlight that MF-FQI algorithm enjoys a "blessing of many agents" property in the sense that a larger number of observed agents improves the performance of MF-FQI algorithm.

On Lp-norm Robustness of Ensemble Decision Stumps and Trees

- Yihan Wang, Huan Zhang, Hongge Chen, Duane Boning, Cho-Jui Hsieh
- abstract: Recent papers have demonstrated that ensemble stumps and trees could be vulnerable to small input perturbations, so robustness verification and defense for those models have become an important research problem. However, due to the structure of decision trees, where each node makes decision purely based on one feature value, all the previous works only consider the ℓ_∞ norm perturbation. To study robustness with respect to a general ℓ_p norm perturbation, one has to consider the correlation between perturbations on different features, which has not been handled by previous algorithms. In this paper, we study the problem of robustness verification and certified defense with respect to general ℓ_p norm perturbations for ensemble decision stumps and trees. For robustness verification of ensemble stumps, we prove that complete verification is NP-complete for $p\in(0, \infty)$ while polynomial time algorithms exist for $p=0$ or ∞. For $p\in(0, \infty)$ we develop an efficient dynamic programming based algorithm for sound verification of ensemble stumps. For ensemble trees, we generalize the previous multi-level robustness verification algorithm to ℓ_p norm. We demonstrate the first certified defense method for training ensemble stumps and trees with respect to ℓ_p norm perturbations, and verify its effectiveness empirically on real datasets.

Thompson Sampling via Local Uncertainty

- Zhendong Wang, Mingyuan Zhou
- abstract: Thompson sampling is an efficient algorithm for sequential decision making, which exploits the posterior uncertainty to address the exploration-exploitation dilemma. There has been significant recent interest in integrating Bayesian neural networks into Thompson sampling. Most of these methods rely on global variable uncertainty for exploration. In this paper, we propose a new probabilistic modeling framework for Thompson sampling, where local latent variable uncertainty is used to sample the mean reward. Variational inference is used to approximate the posterior of the local variable, and semi-implicit structure is further introduced to enhance its expressiveness. Our experimental results on eight contextual bandit benchmark datasets show that Thompson sampling guided by local uncertainty achieves state-of-the-art performance while having low computational complexity.

A Nearly-Linear Time Algorithm for Exact Community Recovery in Stochastic Block Model

- Peng Wang, Zirui Zhou, Anthony Man-Cho So
- abstract: Learning community structures in graphs that are randomly generated by stochastic block models (SBMs) has received much attention lately. In this paper, we focus on the problem of exactly recovering the communities in a binary symmetric SBM, where a graph of n vertices is partitioned into two equal-sized communities and the vertices are connected with probability $p = \alpha\log(n)/n$ within communities and $q = \beta\log(n)/n$ across communities for some $\alpha>\beta>0$. We propose a two-stage iterative algorithm for solving this problem, which employs the power method with a random starting point in the first-stage and turns to a generalized power method that can identify the communities in a finite number of iterations in the second-stage. It is shown that for any fixed α and β such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, which is known to be the information-theoretical limit for exact recovery, the proposed algorithm exactly identifies the underlying communities in $\tilde{O}(n)$ running time with probability tending to one as $n\rightarrow\infty$. We also present numerical results of the proposed algorithm to support and complement our theoretical development.

Learning Representations that Support Extrapolation

- Taylor Webb, Zachary Dulberg, Steven Frankland, Alexander Petrov, Randall O'Reilly, Jonathan Cohen
- abstract: Extrapolation – the ability to make inferences that go beyond the scope of one's experiences – is a hallmark of human intelligence. By contrast, the generalization exhibited by contemporary neural network algorithms is largely limited to interpolation between data points in their training corpora. In this paper, we consider the challenge of learning representations that support extrapolation. We introduce a novel visual analogy benchmark that allows the graded evaluation of extrapolation as a function of distance from the convex domain defined by the training data. We also introduce a simple technique, temporal context normalization, that encourages representations that emphasize the relations between objects. We find that this technique enables a significant improvement in the ability to extrapolate, considerably outperforming a number of competitive techniques.

Tuning-free Plug-and-Play Proximal Algorithm for Inverse Imaging Problems

- Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Carola-Bibiane Schönlieb, Hua Huang
- abstract: Plug-and-play (PnP) is a non-convex framework that combines ADMM or other proximal algorithms with advanced denoiser priors. Recently, PnP has achieved great empirical success, especially with the integration of deep learning-based denoisers. However, a key problem of PnP based approaches is that they require manual parameter tweaking. It is necessary to obtain high-quality results across the high discrepancy in terms of imaging conditions and varying scene content. In this work, we present a tuning-free PnP proximal algorithm, which can automatically determine the internal parameters including the penalty parameter, the denoising strength and the terminal time. A key part of our approach is to develop a policy network for automatic search of parameters, which can be effectively learned via mixed model-free and model-based deep reinforcement learning. We demonstrate, through numerical and visual experiments, that the learned policy can customize different parameters for different states, and often more efficient and effective than existing handcrafted criteria. Moreover, we discuss the practical considerations of the plugged denoisers, which together with our learned policy yield state-of-the-art results. This is prevalent on both linear and nonlinear exemplary inverse imaging problems, and in particular, we show promising results on Compressed Sensing MRI and phase retrieval.

Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes

- Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, Rahul Jain
- abstract: Model-free reinforcement learning is known to be memory and computation efficient and more amendable to large scale problems. In this paper, two model-free algorithms are introduced for learning infinite-horizon average-reward Markov Decision Processes (MDPs). The first algorithm reduces the problem to the discounted-reward version and achieves $\mathcal{O}(T^{2/3})$ regret after T steps, under the minimal assumption of weakly communicating MDPs. To our knowledge, this is the first model-free algorithm for general MDPs in this setting. The second algorithm makes use of recent advances in adaptive algorithms for adversarial multi-armed bandits and improves the regret to $\mathcal{O}(\sqrt{T})$, albeit with a stronger ergodic assumption. This result significantly improves over the $\mathcal{O}(T^{3/4})$ regret achieved by the only existing model-free algorithm by Abbasi-Yadkori et al. (2019) for ergodic MDPs in the infinite-horizon average-reward setting.

The Implicit and Explicit Regularization Effects of Dropout

- Colin Wei, Sham Kakade, Tengyu Ma
- abstract: Dropout is a widely-used regularization technique, often required to obtain state-of-the-art for a number of architectures. This work demonstrates that dropout introduces two distinct but entangled regularization effects: an explicit effect (also studied in prior work) which occurs since dropout modifies the expected training objective, and, perhaps surprisingly, an additional implicit effect from the stochasticity in the dropout training update. This implicit regularization effect is analogous to the effect of stochasticity in small mini-batch stochastic gradient descent. We disentangle these two effects through controlled experiments. We then derive analytic simplifications which characterize each effect in terms of the derivatives of the model and the loss, for deep neural networks. We demonstrate these simplified, analytic regularizers accurately capture the important aspects of dropout, showing they faithfully replace dropout in practice.

Online Control of the False Coverage Rate and False Sign Rate

- Asaf Weinstein, Aaditya Ramdas
- abstract: The reproducibility debate has caused a renewed interest in changing how one reports uncertainty, from p-value for testing a null hypothesis to a confidence interval (CI) for the corresponding parameter. When CIs for multiple selected parameters are being reported, the analog of the false discovery rate (FDR) is the false coverage rate (FCR), which is the expected ratio of number of reported CIs failing to cover their respective parameters to the total number of reported CIs. Here, we consider the general problem of FCR control in the online setting, where one encounters an infinite sequence of fixed unknown parameters ordered by time. We propose a novel solution to the problem which only requires the scientist to be able to construct marginal CIs. As special cases, our framework yields algorithms for online FDR control and online sign-classification procedures that control the false sign rate (FSR). All of our methodology applies equally well to prediction intervals, having particular implications for selective conformal inference.

Batch Stationary Distribution Estimation

- Junfeng Wen, Bo Dai, Lihong Li, Dale Schuurmans
- abstract: We consider the problem of approximating the stationary distribution of an ergodic Markov chain given a set of sampled transitions. Classical simulation-based approaches assume access to the underlying process so that trajectories of sufficient length can be gathered to approximate stationary sampling. Instead, we consider an alternative setting where a \emph{fixed} set of transitions has been collected beforehand, by a separate, possibly unknown procedure. The goal is still to estimate properties of the stationary distribution, but without additional access to the underlying system. We propose a consistent estimator that is based on recovering a correction ratio function over the given data. In particular, we develop a variational power method (VPM) that provides provably consistent estimates under general conditions. In addition to unifying a number of existing approaches from different subfields, we also find that VPM yields significantly better estimates across a range of problems, including queueing, stochastic differential equations, post-processing MCMC, and off-policy evaluation.

Domain Aggregation Networks for Multi-Source Domain Adaptation

- Junfeng Wen, Russell Greiner, Dale Schuurmans
- abstract: In many real-world applications, we want to exploit multiple source datasets to build a model for a different but related target dataset. Despite the recent empirical success, most existing research has used ad-hoc methods to combine multiple sources, leading to a gap between theory and practice. In this paper, we develop a finite-sample generalization bound based on domain discrepancy and accordingly propose a theoretically justified optimization procedure. Our algorithm, Domain AggRegation Network (DARN), can automatically and dynamically balance between including more data to increase effective sample size and excluding irrelevant data to avoid negative effects during training. We find that DARN can significantly outperform the state-of-the-art alternatives on multiple real-world tasks, including digit/object recognition and sentiment analysis.

Towards Understanding the Regularization of Adversarial Robustness on Neural Networks

- Yuxin Wen, Shuai Li, Kui Jia
- abstract: The problem of adversarial examples has shown that modern Neural Network (NN) models could be rather fragile. Among the more established techniques to solve the problem, one is to require the model to be \emph{ϵ-adversarially robust} (AR); that is, to require the model not to change predicted labels when any given input examples are perturbed within a certain range. However, it is observed that such methods would lead to standard performance degradation, i.e., the degradation on natural examples. In this work, we study the degradation through the regularization perspective. We identify quantities from generalization analysis of NNs; with the identified quantities we empirically find that AR is achieved by regularizing/biasing NNs towards less confident solutions by making the changes in the feature space (induced by changes in the instance space) of most layers smoother uniformly in all directions; so to a certain extent, it prevents sudden change in prediction w.r.t. perturbations. However, the end result of such smoothing concentrates samples around decision boundaries, resulting in less confident solutions, and leads to worse standard performance. Our studies suggest that one might consider ways that build AR into NNs in a gentler way to avoid the problematic regularization.

Amortised Learning by Wake-Sleep

- Li Wenliang, Theodore Moskovitz, Heishiro Kanagawa, Maneesh Sahani
- abstract: Models that employ latent variables to capture structure in observed data lie at the heart of many current unsupervised learning algorithms, but exact maximum-likelihood learning for powerful and flexible latent-variable models is almost always intractable. Thus, state-of-the-art approaches either abandon the maximum-likelihood framework entirely, or else rely on a variety of variational approximations to the posterior distribution over the latents. Here, we propose an alternative approach that we call amortised learning. Rather than computing an approximation to the posterior over latents, we use a wake-sleep Monte-Carlo strategy to learn a function that directly estimates the maximum-likelihood parameter updates. Amortised learning is possible whenever samples of latents and observations can be simulated from the generative model, treating the model as a “black box”. We demonstrate its effectiveness on a wide range of complex models, including those with latents that are discrete or supported on non-Euclidean spaces.

How Good is the Bayes Posterior in Deep Neural Networks Really?

- Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, Sebastian Nowozin
- abstract: During the past five years the Bayesian deep learning community has developed increasingly accurate and efficient approximate inference procedures that allow for Bayesian inference in deep neural networks. However, despite this algorithmic progress and the promise of improved uncertainty quantification and sample efficiency there are—as of early 2020—no publicized deployments of Bayesian neural networks in industrial practice. In this work we cast doubt on the current understanding of Bayes posteriors in popular deep neural networks: we demonstrate through careful MCMC sampling that the posterior predictive induced by the Bayes posterior yields systematically worse predictions when compared to simpler methods including point estimates obtained from SGD. Furthermore, we demonstrate that predictive performance is improved significantly through the use of a “cold posterior” that overcounts evidence. Such cold posteriors sharply deviate from the Bayesian paradigm but are commonly used as heuristic in Bayesian deep learning papers. We put forward several hypotheses that could explain cold posteriors and evaluate the hypotheses through experiments. Our work questions the goal of accurate posterior approximations in Bayesian deep learning: If the true Bayes posterior is poor, what is the use of more accurate approximations? Instead, we argue that it is timely to focus on understanding the origin of cold posteriors.

Predictive Sampling with Forecasting Autoregressive Models

- Auke Wiggers, Emiel Hoogeboom
- abstract: Autoregressive models (ARMs) currently hold state-of-the-art performance in likelihood-based modeling of image and audio data. Generally, neural network based ARMs are designed to allow fast inference, but sampling from these models is impractically slow. In this paper, we introduce the predictive sampling algorithm: a procedure that exploits the fast inference property of ARMs in order to speed up sampling, while keeping the model intact. We propose two variations of predictive sampling, namely sampling with ARM fixed-point iteration and learned forecasting modules. Their effectiveness is demonstrated in two settings: i) explicit likelihood modeling on binary MNIST, SVHN and CIFAR10, and ii) discrete latent modeling in

an autoencoder trained on SVHN, CIFAR10 and Imagenet32. Empirically, we show considerable improvements over baselines in number of ARM inference calls and sampling speed.

State Space Expectation Propagation: Efficient Inference Schemes for Temporal Gaussian Processes

- William Wilkinson, Paul Chang, Michael Andersen, Arno Solin
- abstract: We formulate approximate Bayesian inference in non-conjugate temporal and spatio-temporal Gaussian process models as a simple parameter update rule applied during Kalman smoothing. This viewpoint encompasses most inference schemes, including expectation propagation (EP), the classical (Extended, Unscented, etc.) Kalman smoothers, and variational inference. We provide a unifying perspective on these algorithms, showing how replacing the power EP moment matching step with linearisation recovers the classical smoothers. EP provides some benefits over the traditional methods via introduction of the so-called cavity distribution, and we combine these benefits with the computational efficiency of linearisation, providing extensive empirical analysis demonstrating the efficacy of various algorithms under this unifying framework. We provide a fast implementation of all methods in JAX.

Efficient nonparametric statistical inference on population feature importance using Shapley values

- Brian Williamson, Jean Feng
- abstract: The true population-level importance of a variable in a prediction task provides useful knowledge about the underlying data-generating mechanism and can help in deciding which measurements to collect in subsequent experiments. Valid statistical inference on this importance is a key component in understanding the population of interest. We present a computationally efficient procedure for estimating and obtaining valid statistical inference on the \textbf{S}hapley \textbf{P}opulation \textbf{V}ariable \textbf{I}mportance \textbf{M}easure (SPVIM). Although the computational complexity of the true SPVIM scales exponentially with the number of variables, we propose an estimator based on randomly sampling only $\Theta(n)$ feature subsets given n observations. We prove that our estimator converges at an asymptotically optimal rate. Moreover, by deriving the asymptotic distribution of our estimator, we construct valid confidence intervals and hypothesis tests. Our procedure has good finite-sample performance in simulations, and for an in-hospital mortality prediction task produces similar variable importance estimates when different machine learning algorithms are applied.

Efficiently sampling functions from Gaussian process posteriors

- James Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, Marc Deisenroth
- abstract: Gaussian processes are the gold standard for many real-world modeling problems, especially in cases where a model's success hinges upon its ability to faithfully represent predictive uncertainty. These problems typically exist as parts of larger frameworks, wherein quantities of interest are ultimately defined by integrating over posterior distributions. These quantities are frequently intractable, motivating the use of Monte Carlo methods. Despite substantial progress in scaling up Gaussian processes to large training sets, methods for accurately generating draws from their posterior distributions still scale cubically in the number of test locations. We identify a decomposition of Gaussian processes that naturally lends itself to scalable sampling by separating out the prior from the data. Building off of this factorization, we propose an easy-to-use and general-purpose approach for fast posterior sampling, which seamlessly pairs with sparse approximations to afford scalability both during training and at test time. In a series of experiments designed to test competing sampling schemes' statistical properties and practical ramifications, we demonstrate how decoupled sample paths accurately represent Gaussian process posteriors at a fraction of the usual cost.

Learning to Rank Learning Curves

- Martin Wistuba, Tejaswini Pedapati
- abstract: Many automated machine learning methods, such as those for hyperparameter and neural architecture optimization, are computationally expensive because they involve training many different model configurations. In this work, we present a new method that saves computational budget by terminating poor configurations early on in the training. In contrast to existing methods, we consider this task as a ranking and transfer learning problem. We qualitatively show that by optimizing a pairwise ranking loss and leveraging learning curves from other data sets, our model is able to effectively rank learning curves without having to observe many or very long learning curves. We further demonstrate that our method can be used to accelerate a neural architecture search by a factor of up to 100 without a significant performance degradation of the discovered architecture. In further experiments we analyze the quality of ranking, the influence of different model components as well as the predictive behavior of the model.

Causal Inference using Gaussian Processes with Structured Latent Confounders

- Sam Witty, Kenta Takatsu, David Jensen, Vikash Mansinghka
- abstract: Latent confounders—unobserved variables that influence both treatment and outcome—can bias estimates of causal effects. In some cases, these confounders are shared across observations, e.g. all students taking a course are influenced by the course's difficulty in addition to any educational interventions they receive individually. This paper shows how to semiparametrically model latent confounders that have this structure and thereby improve estimates of causal effects. The key innovations are a hierarchical Bayesian model, Gaussian processes with structured latent confounders (GP-SLC), and a Monte Carlo inference algorithm for this model based on elliptical slice sampling. GP-SLC provides principled Bayesian uncertainty estimates of individual treatment effect with minimal assumptions about the functional forms relating confounders, covariates, treatment, and outcome. Finally, this paper shows GP-SLC is competitive with or more accurate than widely used causal inference techniques on three benchmark datasets, including the Infant Health and Development Program and a dataset showing the effect of changing temperatures on state-wide energy consumption across New England.

Near Input Sparsity Time Kernel Embeddings via Adaptive Sampling

- David Woodruff, Amir Zandieh
- abstract: To accelerate kernel methods, we propose a near input sparsity time method for sampling the high-dimensional space implicitly defined by a kernel transformation. Our main contribution is an importance sampling method for subsampling the feature space of a degree q tensoring of data points in almost input sparsity time, improving the recent oblivious sketching of (Ahle et al., 2020) by a factor of $q^{5/2}/\epsilon^2$. This leads to a subspace embedding for the polynomial kernel as well as the Gaussian kernel with a target dimension that is only linearly dependent on the statistical dimension of the kernel and in time which is only linearly dependent on the sparsity of the input dataset. We show how our subspace embedding bounds imply new statistical guarantees for kernel ridge regression. Furthermore, we empirically show that in large-scale regression tasks, our algorithm outperforms state-of-the-art kernel approximation methods.

Is Local SGD Better than Minibatch SGD?

- Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, Nathan Srebro
- abstract: We study local SGD (also known as parallel SGD and federated SGD), a natural and frequently used distributed optimization method. Its theoretical foundations are currently lacking and we highlight how all existing error guarantees in the convex setting are dominated by a simple baseline, minibatch SGD. (1) For quadratic objectives we prove that local SGD strictly dominates minibatch SGD and that accelerated local SGD is minmax

optimal for quadratics; (2) For general convex objectives we provide the first guarantee that at least \emph{sometimes} improves over minibatch SGD, but our guarantee does not always improve over, nor even match, minibatch SGD; (3) We show that indeed local SGD does \emph{not} dominate minibatch SGD by presenting a lower bound on the performance of local SGD that is worse than the minibatch SGD guarantee.

Obtaining Adjustable Regularization for Free via Iterate Averaging

- Jingfeng Wu, Vladimir Braverman, Lin Yang
- abstract: Regularization for optimization is a crucial technique to avoid overfitting in machine learning. In order to obtain the best performance, we usually train a model by tuning the regularization parameters. It becomes costly, however, when a single round of training takes significant amount of time. Very recently, Neu and Rosasco show that if we run stochastic gradient descent (SGD) on linear regression problems, then by averaging the SGD iterates properly, we obtain a regularized solution. It left open whether the same phenomenon can be achieved for other optimization problems and algorithms. In this paper, we establish an averaging scheme that provably converts the iterates of SGD on an arbitrary strongly convex and smooth objective function to its regularized counterpart with an adjustable regularization parameter. Our approaches can be used for accelerated and preconditioned optimization methods as well. We further show that the same methods work empirically on more general optimization objectives including neural networks. In sum, we obtain adjustable regularization for free for a large class of optimization problems and resolve an open question raised by Neu and Rosasco.

DeltaGrad: Rapid retraining of machine learning models

- Yinjun Wu, Edgar Dobriban, Susan Davidson
- abstract: Machine learning models are not static and may need to be retrained on slightly changed datasets, for instance, with the addition or deletion of a set of data points. This has many applications, including privacy, robustness, bias reduction, and uncertainty quantifcation. However, it is expensive to retrain models from scratch. To address this problem, we propose the DeltaGrad algorithm for rapid retraining machine learning models based on information cached during the training phase. We provide both theoretical and empirical support for the effectiveness of DeltaGrad, and show that it compares favorably to the state of the art.

On the Noisy Gradient Descent that Generalizes as SGD

- Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, Zhanxing Zhu
- abstract: The gradient noise of SGD is considered to play a central role in the observed strong generalization abilities of deep learning. While past studies confirm that the magnitude and the covariance structure of gradient noise are critical for regularization, it remains unclear whether or not the class of noise distributions is important. In this work we provide negative results by showing that noises in classes different from the SGD noise can also effectively regularize gradient descent. Our finding is based on a novel observation on the structure of the SGD noise: it is the multiplication of the gradient matrix and a sampling noise that arises from the mini-batch sampling procedure. Moreover, the sampling noises unify two kinds of gradient regularizing noises that belong to the Gaussian class: the one using (scaled) Fisher as covariance and the one using the gradient covariance of SGD as covariance. Finally, thanks to the flexibility of choosing noise class, an algorithm is proposed to perform noisy gradient descent that generalizes well, the variant of which even benefits large batch SGD training without hurting generalization.

Stronger and Faster Wasserstein Adversarial Attacks

- Kaiwen Wu, Allen Wang, Yaoliang Yu
- abstract: Deep models, while being extremely flexible and accurate, are surprisingly vulnerable to “small, imperceptible” perturbations known as adversarial attacks. While the majority of existing attacks focus on measuring perturbations under the ℓ_p metric, Wasserstein distance, which takes geometry in pixel space into account, has long been known to be a suitable metric for measuring image quality and has recently risen as a compelling alternative to the ℓ_p metric in adversarial attacks. However, constructing an effective attack under the Wasserstein metric is computationally much more challenging and calls for better optimization algorithms. We address this gap in two ways: (a) we develop an exact yet efficient projection operator to enable a stronger projected gradient attack; (b) we show that the Frank-Wolfe method equipped with a suitable linear minimization oracle works extremely fast under Wasserstein constraints. Our algorithms not only converge faster but also generate much stronger attacks. For instance, we decrease the accuracy of a residual network on CIFAR-10 to $3.4%$ within a Wasserstein perturbation ball of radius 0.005, in contrast to $65.6%$ using the previous Wasserstein attack based on an \emph{approximate} projection operator. Furthermore, employing our stronger attacks in adversarial training significantly improves the robustness of adversarially trained models.

Sequence Generation with Mixed Representations

- Lijun Wu, Shufang Xie, Yingce Xia, Yang Fan, Jian-Huang Lai, Tao Qin, Tieyan Liu
- abstract: Tokenization is the first step of many natural language processing (NLP) tasks and plays an important role for neural NLP models. Tokenizaton method such as byte-pair encoding (BPE), which can greatly reduce the large vocabulary and deal with out-of-vocabulary words, has shown to be effective and is widely adopted for sequence generation tasks. While various tokenization methods exist, there is no common acknowledgement which is the best. In this work, we propose to leverage the mixed representations from different tokenization methods for sequence generation tasks, in order to boost the model performance with unique characteristics and advantages of individual tokenization methods. Specifically, we introduce a new model architecture to incorporate mixed representations and a co-teaching algorithm to better utilize the diversity of different tokenization methods. Our approach achieves significant improvements on neural machine translation (NMT) tasks with six language pairs (e.g., English$\leftrightarrow$German, English$\leftrightarrow$Romanian), as well as an abstractive summarization task.

Adversarial Robustness via Runtime Masking and Cleansing

- Yi-Hsuan Wu, Chia-Hung Yuan, Shan-Hung Wu
- abstract: Deep neural networks are shown to be vulnerable to adversarial attacks. This motivates robust learning techniques, such as the adversarial training, whose goal is to learn a network that is robust against adversarial attacks. However, the sample complexity of robust learning can be significantly larger than that of “standard” learning. In this paper, we propose improving the adversarial robustness of a network by leveraging the potentially large test data seen at runtime. We devise a new defense method, called runtime masking and cleansing (RMC), that adapts the network at runtime before making a prediction to dynamically mask network gradients and cleanse the model of the non-robust features inevitably learned during the training process due to the size limit of the training set. We conduct experiments on real-world datasets and the results demonstrate the effectiveness of RMC empirically.

On the Generalization Effects of Linear Transformations in Data Augmentation

- Sen Wu, Hongyang Zhang, Gregory Valiant, Christopher Re
- abstract: Data augmentation is a powerful technique to improve performance in applications such as image and text classification tasks. Yet, there is little rigorous understanding of why and how various augmentations work. In this work, we consider a family of linear transformations and study their effects on the ridge estimator in an over-parametrized linear regression setting. First, we show that transformations which preserve the labels of the data can improve estimation by enlarging the span of the training data. Second, we show that transformations which mix data can improve estimation by playing a regularization effect. Finally, we validate our theoretical insights on MNIST. Based on the insights, we propose an augmentation scheme that searches

over the space of transformations by how \emph{uncertain} the model is about the transformed data. We validate our proposed scheme on image and text datasets. For example, our method outperforms RandAugment by 1.24% on CIFAR-100 using Wide-ResNet-28-10. Furthermore, we achieve comparable accuracy to the SoTA Adversarial AutoAugment on CIFAR datasets.

Amortized Population Gibbs Samplers with Neural Sufficient Statistics

- Hao Wu, Heiko Zimmermann, Eli Sennesh, Tuan Anh Le, Jan-Willem Van De Meent
- abstract: We develop amortized population Gibbs (APG) samplers, a class of scalable methods that frame structured variational inference as adaptive importance sampling. APG samplers construct high-dimensional proposals by iterating over updates to lower-dimensional blocks of variables. We train each conditional proposal by minimizing the inclusive KL divergence with respect to the conditional posterior. To appropriately account for the size of the input data, we develop a new parameterization in terms of neural sufficient statistics. Experiments show that APG samplers can be used to train highly-structured deep generative models in an unsupervised manner, and achieve substantial improvements in inference accuracy relative to standard autoencoding variational methods.

Continuous Graph Neural Networks

- Louis-Pascal Xhonneux, Meng Qu, Jian Tang
- abstract: This paper builds on the connection between graph neural networks and traditional dynamical systems. We propose continuous graph neural networks (CGNN), which generalise existing graph neural networks with discrete dynamics in that they can be viewed as a specific discretisation scheme. The key idea is how to characterise the continuous dynamics of node representations, i.e. the derivatives of node representations, w.r.t. time.Inspired by existing diffusion-based methods on graphs (e.g. PageRank and epidemic models on social networks), we define the derivatives as a combination of the current node representations,the representations of neighbors, and the initial values of the nodes. We propose and analyse two possible dynamics on graphs{—}including each dimension of node representations (a.k.a. the feature channel) change independently or interact with each other{—}both with theoretical justification. The proposed continuous graph neural net-works are robust to over-smoothing and hence allow us to build deeper networks, which in turn are able to capture the long-range dependencies between nodes. Experimental results on the task of node classification demonstrate the effectiveness of our proposed approach over competitive baselines.

A Flexible Framework for Nonparametric Graphical Modeling that Accommodates Machine Learning

- Yunhua Xiang, Noah Simon
- abstract: Graphical modeling has been broadly useful for exploring the dependence structure among features in a dataset. However, the strength of graphical modeling hinges on our ability to encode and estimate conditional dependencies. In particular, commonly used measures such as partial correlation are only meaningful under strongly parametric (in this case, multivariate Gaussian) assumptions. These assumptions are unverifiable, and there is often little reason to believe they hold in practice. In this paper, we instead consider 3 non-parametric measures of conditional dependence. These measures are meaningful without structural assumptions on the multivariate distribution of the data. In addition, we show that for 2 of these measures there are simple, strong plug-in estimators that require only the estimation of a conditional mean. These plug-in estimators (1) are asymptotically linear and non-parametrically efficient, (2) allow incorporation of flexible machine learning techniques for conditional mean estimation, and (3) enable the construction of valid Wald-type confidence intervals. In addition, by leveraging the influence function of these estimators, one can obtain intervals with simultaneous coverage guarantees for all pairs of features.

Generative Flows with Matrix Exponential

- Changyi Xiao, Ligang Liu
- abstract: Generative flows models enjoy the properties of tractable exact likelihood and efficient sampling, which are composed of a sequence of invertible functions. In this paper, we incorporate matrix exponential into generative flows. Matrix exponential is a map from matrices to invertible matrices, this property is suitable for generative flows. Based on matrix exponential, we propose matrix exponential coupling layers that are a general case of affine coupling layers and matrix exponential invertible 1 x 1 convolutions that do not collapse during training. And we modify the networks architecture to make training stable and significantly speed up the training process. Our experiments show that our model achieves great performance on density estimation amongst generative flows models.

Disentangling Trainability and Generalization in Deep Neural Networks

- Lechao Xiao, Jeffrey Pennington, Samuel Schoenholz
- abstract: A longstanding goal in the theory of deep learning is to characterize the conditions under which a given neural network architecture will be trainable, and if so, how well it might generalize to unseen data. In this work, we provide such a characterization in the limit of very wide and very deep networks, for which the analysis simplifies considerably. For wide networks, the trajectory under gradient descent is governed by the Neural Tangent Kernel (NTK), and for deep networks the NTK itself maintains only weak data dependence. By analyzing the spectrum of the NTK, we formulate necessary conditions for trainability and generalization across a range of architectures, including Fully Connected Networks (FCNs) and Convolutional Neural Networks (CNNs). We identify large regions of hyperparameter space for which networks can memorize the training set but completely fail to generalize. We find that CNNs without global average pooling behave almost identically to FCNs, but that CNNs with pooling have markedly different and often better generalization performance. These theoretical results are corroborated experimentally on CIFAR10 for a variety of network architectures. We include a \href{https://colab.research.google.com/github/google/neural-tangents/blob/master/notebooks/disentangling_trainability_and_generalization.ipynb}{colab} notebook that reproduces the essential results of the paper.

Optimally Solving Two-Agent Decentralized POMDPs Under One-Sided Information Sharing

- Yuxuan Xie, Jilles Dibangoye, Olivier Buffet
- abstract: Optimally solving decentralized partially observable Markov decision processes under either full or no information sharing received significant attention in recent years. However, little is known about how partial information sharing affects existing theory and algorithms. This paper addresses this question for a team of two agents, with one-sided information sharing—\ie both agents have imperfect information about the state of the world, but only one has access to what the other sees and does. From the perspective of a central planner, we show that the original problem can be reformulated into an equivalent information-state Markov decision process and solved as such. Besides, we prove that the optimal value function exhibits a specific form of uniform continuity. We also present a heuristic search algorithm utilizing this property and providing the first results for this family of problems.

Maximum-and-Concatenation Networks

- Xingyu Xie, Hao Kong, Jianlong Wu, Wayne Zhang, Guangcan Liu, Zhouchen Lin
- abstract: While successful in many fields, deep neural networks (DNNs) still suffer from some open problems such as bad local minima and unsatisfactory generalization performance. In this work, we propose a novel architecture called Maximum-and-Concatenation Networks (MCN) to try eliminating bad local minima and improving generalization ability as well. Remarkably, we prove that MCN has a very nice property; that is, every local minimum of an (l+1)-layer MCN can be better than, at least as good as, the global minima of the network consisting of its first l layers. In other words, by increasing the

network depth, MCN can autonomously improve its local minima's goodness, what is more, it is easy to plug MCN into an existing deep model to make it also have this property. Finally, under mild conditions, we show that MCN can approximate certain continuous function arbitrarily well with high efficiency; that is, the covering number of MCN is much smaller than most existing DNNs such as deep ReLU. Based on this, we further provide a tight generalization bound to guarantee the inference ability of MCN when dealing with testing samples.

Zeno++: Robust Fully Asynchronous SGD

- Cong Xie, Sanmi Koyejo, Indranil Gupta
- abstract: We propose Zeno++, a new robust asynchronous Stochastic Gradient Descent(SGD) procedure, intended to tolerate Byzantine failures of workers. In contrast to previous work, Zeno++ removes several unrealistic restrictions on worker-server communication, now allowing for fully asynchronous updates from anonymous workers, for arbitrarily stale worker updates, and for the possibility of an unbounded number of Byzantine workers. The key idea is to estimate the descent of the loss value after the candidate gradient is applied, where large descent values indicate that the update results in optimization progress. We prove the convergence of Zeno++ for non-convex problems under Byzantine failures. Experimental results show that Zeno++ outperforms existing Byzantine-tolerant asynchronous SGD algorithms.

Lower Complexity Bounds for Finite-Sum Convex-Concave Minimax Optimization Problems

- Guangzeng Xie, Luo Luo, Yijiang Lian, Zhihua Zhang
- abstract: This paper studies the lower bound complexity for minimax optimization problem whose objective function is the average of n individual smooth convex-concave functions. We consider the algorithm which gets access to gradient and proximal oracle for each individual component. For the strongly-convex-strongly-concave case, we prove such an algorithm can not reach an ε-suboptimal point in fewer than $\Omega\left((n+\kappa)\log(1/\varepsilon)\right)$ iterations, where κ is the condition number of the objective function. This lower bound matches the upper bound of the existing incremental first-order oracle algorithm stochastic variance-reduced extragradient. We develop a novel construction to show the above result, which partitions the tridiagonal matrix of classical examples into n groups. This construction is friendly to the analysis of incremental gradient and proximal oracle and we also extend the analysis to general convex-concave cases.

On the Number of Linear Regions of Convolutional Neural Networks

- Huan Xiong, Lei Huang, Mengyang Yu, Li Liu, Fan Zhu, Ling Shao
- abstract: One fundamental problem in deep learning is understanding the outstanding performance of deep Neural Networks (NNs) in practice. One explanation for the superiority of NNs is that they can realize a large class of complicated functions, i.e., they have powerful expressivity. The expressivity of a ReLU NN can be quantified by the maximal number of linear regions it can separate its input space into. In this paper, we provide several mathematical results needed for studying the linear regions of CNNs, and use them to derive the maximal and average numbers of linear regions for one-layer ReLU CNNs. Furthermore, we obtain upper and lower bounds for the number of linear regions of multi-layer ReLU CNNs. Our results suggest that deeper CNNs have more powerful expressivity than their shallow counterparts, while CNNs have more expressivity than fully-connected NNs per parameter.

On Layer Normalization in the Transformer Architecture

- Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, Tieyan Liu
- abstract: The Transformer is widely used in natural language processing tasks. To train a Transformer however, one usually needs a carefully designed learning rate warm-up stage, which is shown to be crucial to the final performance but will slow down the optimization and bring more hyper-parameter tunings. In this paper, we first study theoretically why the learning rate warm-up stage is essential and show that the location of layer normalization matters. Specifically, we prove with mean field theory that at initialization, for the original-designed Post-LN Transformer, which places the layer normalization between the residual blocks, the expected gradients of the parameters near the output layer are large. Therefore, using a large learning rate on those gradients makes the training unstable. The warm-up stage is practically helpful for avoiding this problem. On the other hand, our theory also shows that if the layer normalization is put inside the residual blocks (recently proposed as Pre-LN Transformer), the gradients are well-behaved at initialization. This motivates us to remove the warm-up stage for the training of Pre-LN Transformers. We show in our experiments that Pre-LN Transformers without the warm-up stage can reach comparable results with baselines while requiring significantly less training time and hyper-parameter tuning on a wide range of applications.

On Variational Learning of Controllable Representations for Text without Supervision

- Peng Xu, Jackie Chi Kit Cheung, Yanshuai Cao
- abstract: The variational autoencoder (VAE) can learn the manifold of natural images on certain datasets, as evidenced by meaningful interpolating or extrapolating in the continuous latent space. However, on discrete data such as text, it is unclear if unsupervised learning can discover similar latent space that allows controllable manipulation. In this work, we find that sequence VAEs trained on text fail to properly decode when the latent codes are manipulated, because the modified codes often land in holes or vacant regions in the aggregated posterior latent space, where the decoding network fails to generalize. Both as a validation of the explanation and as a fix to the problem, we propose to constrain the posterior mean to a learned probability simplex, and performs manipulation within this simplex. Our proposed method mitigates the latent vacancy problem and achieves the first success in unsupervised learning of controllable representations for text. Empirically, our method outperforms unsupervised baselines and strong supervised approaches on text style transfer, and is capable of performing more flexible fine-grained control over text generation than existing methods.

Class-Weighted Classification: Trade-offs and Robust Approaches

- Ziyu Xu, Chen Dan, Justin Khim, Pradeep Ravikumar
- abstract: We consider imbalanced classification, the problem in which a label may have low marginal probability relative to other labels, by weighting losses according to the correct class. First, we examine the convergence rates of the expected excess weighted risk of plug-in classifiers where the weighting for the plug-in classifier and the risk may be different. This leads to irreducible errors that do not converge to the weighted Bayes risk, which motivates our consideration of robust risks. We define a robust risk that minimizes risk over a set of weightings, show excess risk bounds for this problem, and demonstrate that particular choices of the weighting set leads to a special instance of conditional value at risk (CVaR) from stochastic programming, which we call label conditional value at risk (LCVaR). Additionally, we generalize this weighting to derive a new robust risk problem that we call label heterogeneous conditional value at risk (LHCVaR). Finally, we empirically demonstrate the efficacy of LCVaR and LHCVaR on improving class conditional risks.

A Finite-Time Analysis of Q-Learning with Neural Network Function Approximation

- Pan Xu, Quanquan Gu
- abstract: Q-learning with neural network function approximation (neural Q-learning for short) is among the most prevalent deep reinforcement learning algorithms. Despite its empirical success, the non-asymptotic convergence rate of neural Q-learning remains virtually unknown. In this paper, we present a finite-time analysis of a neural Q-learning algorithm, where the data are generated from a Markov decision process, and the action-value function is

approximated by a deep ReLU neural network. We prove that neural Q-learning finds the optimal policy with an $O(1/\sqrt{T})$ convergence rate if the neural function approximator is sufficiently overparameterized, where T is the number of iterations. To our best knowledge, our result is the first finite-time analysis of neural Q-learning under non-i.i.d. data assumption.

Understanding and Stabilizing GANs' Training Dynamics Using Control Theory

- Kun Xu, Chongxuan Li, Jun Zhu, Bo Zhang
- abstract: Generative adversarial networks (GANs) are effective in generating realistic images but the training is often unstable. There are existing efforts that model the training dynamics of GANs in the parameter space but the analysis cannot directly motivate practically effective stabilizing methods. To this end, we present a conceptually novel perspective from control theory to directly model the dynamics of GANs in the frequency domain and provide simple yet effective methods to stabilize GAN's training. We first analyze the training dynamic of a prototypical Dirac GAN and adopt the widely-used closed-loop control (CLC) to improve its stability. We then extend CLC to stabilize the training dynamic of normal GANs, which can be implemented as an L2 regularizer on the output of the discriminator. Empirical results show that our method can effectively stabilize the training and obtain state-of-the-art performance on data generation tasks.

Learning Autoencoders with Relational Regularization

- Hongteng Xu, Dixin Luo, Ricardo Henao, Svati Shah, Lawrence Carin
- abstract: We propose a new algorithmic framework for learning autoencoders of data distributions. In this framework, we minimize the discrepancy between the model distribution and the target one, with relational regularization on learnable latent prior. This regularization penalizes the fused Gromov-Wasserstein (FGW) distance between the latent prior and its corresponding posterior, which allows us to learn a structured prior distribution associated with the generative model in a flexible way. Moreover, it helps us co-train multiple autoencoders even if they are with heterogeneous architectures and incomparable latent spaces. We implement the framework with two scalable algorithms, making it applicable for both probabilistic and deterministic autoencoders. Our relational regularized autoencoder (RAE) outperforms existing methods, e.g., variational autoencoder, Wasserstein autoencoder, and their variants, on generating images. Additionally, our relational co-training strategy of autoencoders achieves encouraging results in both synthesis and real-world multi-view learning tasks.

Learning Factorized Weight Matrix for Joint Filtering

- Xiangyu Xu, Yongrui Ma, Wenxiu Sun
- abstract: Joint filtering is a fundamental problem in computer vision with applications in many different areas. Most existing algorithms solve this problem with a weighted averaging process to aggregate input pixels. However, the weight matrix of this process is often empirically designed and not robust to complex input. In this work, we propose to learn the weight matrix for joint image filtering. This is a challenging problem, as directly learning a large weight matrix is computationally intractable. To address this issue, we introduce the correlation of deep features to approximate the aggregation weights. However, this strategy only uses inner product for the weight matrix estimation, which limits the performance of the proposed algorithm. Therefore, we further propose to learn a nonlinear function to predict sparse residuals of the feature correlation matrix. Note that the proposed method essentially factorizes the weight matrix into a low-rank and a sparse matrix and then learn both of them simultaneously with deep neural networks. Extensive experiments show that the proposed algorithm compares favorably against the state-of-the-art approaches on a wide variety of joint filtering tasks.

Variational Label Enhancement

- Ning Xu, Jun Shu, Yun-Peng Liu, Xin Geng
- abstract: Label distribution covers a certain number of labels, representing the degree to which each label describes the instance. When dealing with label ambiguity, label distribution could describe the supervised information in a fine-grained way. Unfortunately, many training sets only contain simple logical labels rather than label distributions due to the difficulty of obtaining label distributions directly. To solve this problem, we consider the label distributions as the latent vectors and infer them from the logical labels in the training datasets by using variational inference. After that, we induce a predictive model to train the label distribution data by employing the multi-output regression technique. The recovery experiment on thirteen real-world LDL datasets and the predictive experiment on ten multi-label learning datasets validate the advantage of our approach over the state-of-the-art approaches.

Prediction-Guided Multi-Objective Reinforcement Learning for Continuous Robot Control

- Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, Wojciech Matusik
- abstract: Many real-world control problems involve conflicting objectives where we desire a dense and high-quality set of control policies that are optimal for different objective preferences (called Pareto-optimal). While extensive research in multi-objective reinforcement learning (MORL) has been conducted to tackle such problems, multi-objective optimization for complex continuous robot control is still under-explored. In this work, we propose an efficient evolutionary learning algorithm to find the Pareto set approximation for continuous robot control problems, by extending a state-of-the-art RL algorithm and presenting a novel prediction model to guide the learning process. In addition to efficiently discovering the individual policies on the Pareto front, we construct a continuous set of Pareto-optimal solutions by Pareto analysis and interpolation. Furthermore, we design seven multi-objective RL environments with continuous action space, which is the first benchmark platform to evaluate MORL algorithms on various robot control problems. We test the previous methods on the proposed benchmark problems, and the experiments show that our approach is able to find a much denser and higher-quality set of Pareto policies than the existing algorithms.

MetaFun: Meta-Learning with Iterative Functional Updates

- Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam Kosiorek, Yee Whye Teh
- abstract: We develop a functional encoder-decoder approach to supervised meta-learning, where labeled data is encoded into an infinite-dimensional functional representation rather than a finite-dimensional one. Furthermore, rather than directly producing the representation, we learn a neural update rule resembling functional gradient descent which iteratively improves the representation. The final representation is used to condition the decoder to make predictions on unlabeled data. Our approach is the first to demonstrates the success of encoder-decoder style meta-learning methods like conditional neural processes on large-scale few-shot classification benchmarks such as miniImageNet and tieredImageNet, where it achieves state-of-the-art performance.

Video Prediction via Example Guidance

- Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, Trevor Darrell
- abstract: In video prediction tasks, one major challenge is to capture the multi-modal nature of future contents and dynamics. In this work, we propose a simple yet effective framework that can efficiently predict plausible future states, where the key insight is that the potential distribution of a sequence could be approximated with analogous ones in a repertoire of training pool, namely, expert examples. By further incorporating a novel optimization scheme into the training procedure, plausible predictions can be sampled efficiently from distribution constructed from the retrieved examples. Meanwhile, our method could be seamlessly integrated with existing stochastic predictive models; significant enhancement is observed with comprehensive experiments in both quantitative and qualitative aspects. We also demonstrate the generalization ability to predict the motion of unseen

class, i.e., without access to corresponding data during training phase. Project Page: \hyperlink{https://sites.google.com/view/vpeg-supp/home.} {https://sites.google.com/view/vpeg-supp/home.}

Amortized Finite Element Analysis for Fast PDE-Constrained Optimization

- Tianju Xue, Alex Beatson, Sigrid Adriaenssens, Ryan Adams
- abstract: Optimizing the parameters of partial differential equations (PDEs), i.e., PDE-constrained optimization (PDE-CO), allows us to model natural systems from observations or perform rational design of structures with complicated mechanical, thermal, or electromagnetic properties. However, PDE-CO is often computationally prohibitive due to the need to solve the PDE—typically via finite element analysis (FEA)—at each step of the optimization procedure. In this paper we propose amortized finite element analysis (AmorFEA), in which a neural network learns to produce accurate PDE solutions, while preserving many of the advantages of traditional finite element methods. This network is trained to directly minimize the potential energy from which the PDE and finite element method are derived, avoiding the need to generate costly supervised training data by solving PDEs with traditional FEA. As FEA is a variational procedure, AmorFEA is a direct analogue to popular amortized inference approaches in latent variable models, with the finite element basis acting as the variational family. AmorFEA can perform PDE-CO without the need to repeatedly solve the associated PDE, accelerating optimization when compared to a traditional workflow using FEA and the adjoint method.

Feature Selection using Stochastic Gates

- Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, Yuval Kluger
- abstract: Feature selection problems have been extensively studied in the setting of linear estimation (e.g. LASSO), but less emphasis has been placed on feature selection for non-linear functions. In this study, we propose a method for feature selection in neural network estimation problems. The new procedure is based on probabilistic relaxation of the ℓ_0 norm of features, or the count of the number of selected features. Our ℓ_0-based regularization relies on a continuous relaxation of the Bernoulli distribution; such relaxation allows our model to learn the parameters of the approximate Bernoulli distributions via gradient descent. The proposed framework simultaneously learns either a nonlinear regression or classification function while selecting a small subset of features. We provide an information-theoretic justification for incorporating Bernoulli distribution into feature selection. Furthermore, we evaluate our method using synthetic and real-life data to demonstrate that our approach outperforms other commonly used methods in both predictive performance and feature selection.

Stochastic Optimization for Non-convex Inf-Projection Problems

- Yan Yan, Yi Xu, Lijun Zhang, Wang Xiaoyu, Tianbao Yang
- abstract: In this paper, we study a family of non-convex and possibly non-smooth inf-projection minimization problems, where the target objective function is equal to minimization of a joint function over another variable. This problem include difference of convex (DC) functions and a family of bi-convex functions as special cases. We develop stochastic algorithms and establish their first-order convergence for finding a (nearly) stationary solution of the target non-convex function under different conditions of the component functions. To the best of our knowledge, this is the first work that comprehensively studies stochastic optimization of non-convex inf-projection minimization problems with provable convergence guarantee. Our algorithms enable efficient stochastic optimization of a family of non-decomposable DC functions and a family of bi-convex functions. To demonstrate the power of the proposed algorithms we consider an important application in variance-based regularization. Experiments verify the effectiveness of our inf-projection based formulation and the proposed stochastic algorithm in comparison with previous stochastic algorithms based on the min-max formulation for achieving the same effect.

Variational Bayesian Quantization

- Yibo Yang, Robert Bamler, Stephan Mandt
- abstract: We propose a novel algorithm for quantizing continuous latent representations in trained models. Our approach applies to deep probabilistic models, such as variational autoencoders (VAEs), and enables both data and model compression. Unlike current end-to-end neural compression methods that cater the model to a fixed quantization scheme, our algorithm separates model design and training from quantization. Consequently, our algorithm enables “plug-and-play” compression with variable rate-distortion trade-off, using a single trained model. Our algorithm can be seen as a novel extension of arithmetic coding to the continuous domain, and uses adaptive quantization accuracy based on estimates of posterior uncertainty. Our experimental results demonstrate the importance of taking into account posterior uncertainties, and show that image compression with the proposed algorithm outperforms JPEG over a wide range of bit rates using only a single standard VAE. Further experiments on Bayesian neural word embeddings demonstrate the versatility of the proposed method.

Energy-Based Processes for Exchangeable Data

- Mengjiao Yang, Bo Dai, Hanjun Dai, Dale Schuurmans
- abstract: Recently there has been growing interest in modeling sets with exchangeability such as point clouds. A shortcoming of current approaches is that they restrict the cardinality of the sets considered or can only express limited forms of distribution over unobserved data. To overcome these limitations, we introduce Energy-Based Processes (EBPs), which extend energy based models to exchangeable data while allowing neural network parameterizations of the energy function. A key advantage of these models is the ability to express more flexible distributions over sets without restricting their cardinality. We develop an efficient training procedure for EBPs that demonstrates state-of-the-art performance on a variety of tasks such as point cloud generation, classification, denoising, and image completion

Randomized Smoothing of All Shapes and Sizes

- Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, Jerry Li
- abstract: Randomized smoothing is the current state-of-the-art defense with provable robustness against ℓ_2 adversarial attacks. Many works have devised new randomized smoothing schemes for other metrics, such as ℓ_1 or ℓ_∞; however, substantial effort was needed to derive such new guarantees. This begs the question: can we find a general theory for randomized smoothing? We propose a novel framework for devising and analyzing randomized smoothing schemes, and validate its effectiveness in practice. Our theoretical contributions are: (1) we show that for an appropriate notion of "optimal", the optimal smoothing distributions for any "nice" norms have level sets given by the norm’s *Wulff Crystal*; (2) we propose two novel and complementary methods for deriving provably robust radii for any smoothing distribution; and, (3) we show fundamental limits to current randomized smoothing techniques via the theory of *Banach space cotypes*. By combining (1) and (2), we significantly improve the state-of-the-art certified accuracy in ℓ_1 on standard datasets. Meanwhile, we show using (3) that with only label statistics under random input perturbations, randomized smoothing cannot achieve nontrivial certified accuracy against perturbations of ℓ_p-norm $\Omega(\min(1, d^{\frac{1}{p} - \frac{1}{2}}))$, when the input dimension d is large. We provide code in github.com/tonyduan/rs4a.

Q-value Path Decomposition for Deep Multiagent Reinforcement Learning

- Yaodong Yang, Jianye Hao, Guangyong Chen, Hongyao Tang, Yingfeng Chen, Yujing Hu, Changjie Fan, Zhongyu Wei

- abstract: Recently, deep multiagent reinforcement learning (MARL) has become a highly active research area as many real-world problems can be inherently viewed as multiagent systems. A particularly interesting and widely applicable class of problems is the partially observable cooperative multiagent setting, in which a team of agents learns to coordinate their behaviors conditioning on their private observations and commonly shared global reward signals. One natural solution is to resort to the centralized training and decentralized execution paradigm and during centralized training, one key challenge is the multiagent credit assignment: how to allocate the global rewards for individual agent policies for better coordination towards maximizing system-level's benefits. In this paper, we propose a new method called Q-value Path Decomposition (QPD) to decompose the system's global Q-values into individual agents' Q-values. Unlike previous works which restrict the representation relation of the individual Q-values and the global one, we leverage the integrated gradient attribution technique into deep MARL to directly decompose global Q-values along trajectory paths to assign credits for agents. We evaluate QPD on the challenging StarCraft II micromanagement tasks and show that QPD achieves the state-of-the-art performance in both homogeneous and heterogeneous multiagent scenarios compared with existing cooperative MARL algorithms.

Improving Molecular Design by Stochastic Iterative Target Augmentation

- Kevin Yang, Wengong Jin, Kyle Swanson, Dr.Regina Barzilay, Tommi Jaakkola
- abstract: Generative models in molecular design tend to be richly parameterized, data-hungry neural models, as they must create complex structured objects as outputs. Estimating such models from data may be challenging due to the lack of sufficient training data. In this paper, we propose a surprisingly effective self-training approach for iteratively creating additional molecular targets. We first pre-train the generative model together with a simple property predictor. The property predictor is then used as a likelihood model for filtering candidate structures from the generative model. Additional targets are iteratively produced and used in the course of stochastic EM iterations to maximize the log-likelihood that the candidate structures are accepted. A simple rejection (re-weighting) sampler suffices to draw posterior samples since the generative model is already reasonable after pre-training. We demonstrate significant gains over strong baselines for both unconditional and conditional molecular design. In particular, our approach outperforms the previous state-of-the-art in conditional molecular design by over 10% in absolute gain. Finally, we show that our approach is useful in other domains as well, such as program synthesis.

On the consistency of top-k surrogate losses

- Forest Yang, Sanmi Koyejo
- abstract: The top-k error is often employed to evaluate performance for challenging classification tasks in computer vision as it is designed to compensate for ambiguity in ground truth labels. This practical success motivates our theoretical analysis of consistent top-k classification. To this end, we provide a characterization of Bayes optimality by defining a top-k preserving property, which is new and fixes a non-uniqueness gap in prior work. Then, we define top-k calibration and show it is necessary and sufficient for consistency. Based on the top-k calibration analysis, we propose a rich class of top-k calibrated Bregman divergence surrogates. Our analysis continues by showing previously proposed hinge-like top-k surrogate losses are not top-k calibrated and thus inconsistent. On the other hand, we propose two new hinge-like losses, one which is similarly inconsistent, and one which is consistent. Our empirical results highlight theoretical claims, confirming our analysis of the consistency of these losses.

Interpolation between Residual and Non-Residual Networks

- Zonghan Yang, Yang Liu, Chenglong Bao, Zuoqiang Shi
- abstract: Although ordinary differential equations (ODEs) provide insights for designing network architectures, its relationship with the non-residual convolutional neural networks (CNNs) is still unclear. In this paper, we present a novel ODE model by adding a damping term. It can be shown that the proposed model can recover both a ResNet and a CNN by adjusting an interpolation coefficient. Therefore, the damped ODE model provides a unified framework for the interpretation of residual and non-residual networks. The Lyapunov analysis reveals better stability of the proposed model, and thus yields robustness improvement of the learned networks. Experiments on a number of image classification benchmarks show that the proposed model substantially improves the accuracy of ResNet and ResNeXt over the perturbed inputs from both stochastic noise and adversarial attack methods. Moreover, the loss landscape analysis demonstrates the improved robustness of our method along the attack direction.

Reinforcement Learning in Feature Space: Matrix Bandit, Kernels, and Regret Bound

- Lin Yang, Mengdi Wang
- abstract: Exploration in reinforcement learning (RL) suffers from the curse of dimensionality when the state-action space is large. A common practice is to parameterize the high-dimensional value and policy functions using given features. However existing methods either have no theoretical guarantee or suffer a regret that is exponential in the planning horizon H.In this paper, we propose an online RL algorithm, namely the MatrixRL, that leverages ideas from linear bandit to learn a low-dimensional representation of the probability transition model while carefully balancing the exploitation-exploration tradeoff. We show that MatrixRL achieves a regret bound ${O}\big(H^2d\log T\sqrt{T}\big)$ where d is the number of features, independent with the number of state-action pairs. MatrixRL has an equivalent kernelized version, which is able to work with an arbitrary kernel Hilbert space without using explicit features. In this case, the kernelized MatrixRL satisfies a regret bound ${O}\big(H^2\wt{d}\log T\sqrt{T}\big)$, where $\wt{d}$ is the effective dimension of the kernel space.

Multi-Agent Determinantal Q-Learning

- Yaodong Yang, Ying Wen, Jun Wang, Liheng Chen, Kun Shao, David Mguni, Weinan Zhang
- abstract: Centralized training with decentralized execution has become an important paradigm in multi-agent learning. Though practical, current methods rely on restrictive assumptions to decompose the centralized value function across agents for execution. In this paper, we eliminate this restriction by proposing multi-agent determinantal Q-learning. Our method is established on Q-DPP, a novel extension of determinantal point process (DPP) to multi-agent setting. Q-DPP promotes agents to acquire diverse behavioral models; this allows a natural factorization of the joint Q-functions with no need for \emph{a priori} structural constraints on the value function or special network architectures. We demonstrate that Q-DPP generalizes major solutions including VDN, QMIX, and QTRAN on decentralizable cooperative tasks. To efficiently draw samples from Q-DPP, we develop a linear-time sampler with theoretical approximation guarantee. Our sampler also benefits exploration by coordinating agents to cover orthogonal directions in the state space during training. We evaluate our algorithm on multiple cooperative benchmarks; its effectiveness has been demonstrated when compared with the state-of-the-art.

Rethinking Bias-Variance Trade-off for Generalization of Neural Networks

- Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, Yi Ma
- abstract: The classical bias-variance trade-off predicts that bias decreases and variance increase with model complexity, leading to a U-shaped risk curve. Recent work calls this into question for neural networks and other over-parameterized models, for which it is often observed that larger models generalize better. We provide a simple explanation of this by measuring the bias and variance of neural networks: while the bias is \emph{monotonically decreasing} as in the classical theory, the variance is \emph{unimodal} or bell-shaped: it increases then decreases with the width of the network. We vary the network architecture, loss function, and choice of dataset and confirm that variance unimodality occurs robustly for all models we considered. The risk curve is the sum of the bias and variance curves and displays different qualitative shapes depending on the relative scale of bias and variance, with the double descent in the recent literature as a special case. We corroborate these empirical results with a theoretical analysis of two-layer linear networks with random first

layer. Finally, evaluation on out-of-distribution data shows that most of the drop in accuracy comes from increased bias while variance increases by a relatively small amount. Moreover, we find that deeper models decrease bias and increase variance for both in-distribution and out-of-distribution data.

Unsupervised Transfer Learning for Spatiotemporal Predictive Networks

- Zhiyu Yao, Yunbo Wang, Mingsheng Long, Jianmin Wang
- abstract: This paper explores a new research problem of unsupervised transfer learning across multiple spatiotemporal prediction tasks. Unlike most existing transfer learning methods that focus on fixing the discrepancy between supervised tasks, we study how to transfer knowledge from a zoo of unsupervisedly learned models towards another predictive network. Our motivation is that models from different sources are expected to understand the complex spatiotemporal dynamics from different perspectives, thereby effectively supplementing the new task, even if the task has sufficient training samples. Technically, we propose a differentiable framework named transferable memory. It adaptively distills knowledge from a bank of memory states of multiple pretrained RNNs, and applies it to the target network via a novel recurrent structure called the Transferable Memory Unit (TMU). Compared with finetuning, our approach yields significant improvements on three benchmarks for spatiotemporal prediction, and benefits the target task even from less relevant pretext ones.

Searching to Exploit Memorization Effect in Learning with Noisy Labels

- Quanming Yao, Hansi Yang, Bo Han, Gang Niu, James Tin-Yau Kwok
- abstract: Sample selection approaches are popular in robust learning from noisy labels. However, how to properly control the selection process so that deep networks can benefit from the memorization effect is a hard problem. In this paper, motivated by the success of automated machine learning (AutoML), we model this issue as a function approximation problem. Specifically, we design a domain-specific search space based on general patterns of the memorization effect and propose a novel Newton algorithm to solve the bi-level optimization problem efficiently. We further provide a theoretical analysis of the algorithm, which ensures a good approximation to critical points. Experiments are performed on both benchmark and real-world data sets. Results demonstrate that the proposed method is much better than the state-of-the-art noisy-label-learning approaches, and also much more efficient than existing AutoML algorithms.

Graph-based, Self-Supervised Program Repair from Diagnostic Feedback

- Michihiro Yasunaga, Percy Liang
- abstract: We consider the problem of learning to repair programs from diagnostic feedback (e.g., compiler error messages). Program repair is challenging for two reasons: First, it requires reasoning and tracking symbols across source code and diagnostic feedback. Second, labeled datasets available for program repair are relatively small. In this work, we propose novel solutions to these two challenges. First, we introduce a program-feedback graph, which connects symbols relevant to program repair in source code and diagnostic feedback, and then apply a graph neural network on top to model the reasoning process. Second, we present a self-supervised learning paradigm for program repair that leverages unlabeled programs available online to create a large amount of extra program repair examples, which we use to pre-train our models. We evaluate our proposed approach on two applications: correcting introductory programming assignments (DeepFix dataset) and correcting the outputs of program synthesis (SPoC dataset). Our final system, DrRepair, significantly outperforms prior work, achieving 68.2% full repair rate on DeepFix (+22.9% over the prior best), and 48.4% synthesis success rate on SPoC (+3.7% over the prior best).

Pretrained Generalized Autoregressive Model with Adaptive Probabilistic Label Clusters for Extreme Multi-label Text Classification

- Hui Ye, Zhiyu Chen, Da-Han Wang, Brian Davison
- abstract: Extreme multi-label text classification (XMTC) is a task for tagging a given text with the most relevant labels from an extremely large label set. We propose a novel deep learning method called APLC-XLNet. Our approach fine-tunes the recently released generalized autoregressive pretrained model (XLNet) to learn a dense representation for the input text. We propose Adaptive Probabilistic Label Clusters (APLC) to approximate the cross entropy loss by exploiting the unbalanced label distribution to form clusters that explicitly reduce the computational time. Our experiments, carried out on five benchmark datasets, show that our approach has achieved new state-of-the-art results on four benchmark datasets. Our source code is available publicly at https://github.com/huiyegit/APLC_XLNet.

Good Subnetworks Provably Exist: Pruning via Greedy Forward Selection

- Mao Ye, Chengyue Gong, Lizhen Nie, Denny Zhou, Adam Klivans, Qiang Liu
- abstract: Recent empirical works show that large deep neural networks are often highly redundant and one can find much smaller subnetworks without a significant drop of accuracy. However, most existing methods of network pruning are empirical and heuristic, leaving it open whether good subnetworks provably exist, how to find them efficiently, and if network pruning can be provably better than direct training using gradient descent. We answer these problems positively by proposing a simple greedy selection approach for finding good subnetworks, which starts from an empty network and greedily adds important neurons from the large network. This differs from the existing methods based on backward elimination, which remove redundant neurons from the large network. Theoretically, applying the greedy selection strategy on sufficiently large {pre-trained} networks guarantees to find small subnetworks with lower loss than networks directly trained with gradient descent. Our results also apply to pruning randomly weighted networks. Practically, we improve prior arts of network pruning on learning compact neural architectures on ImageNet, including ResNet, MobilenetV2/V3, and ProxylessNet. Our theory and empirical results on MobileNet suggest that we should fine-tune the pruned subnetworks to leverage the information from the large model, instead of re-training from new random initialization as suggested in \citet{liu2018rethinking}.

It’s Not What Machines Can Learn, It’s What We Cannot Teach

- Gal Yehuda, Moshe Gabel, Assaf Schuster
- abstract: Can deep neural networks learn to solve any task, and in particular problems of high complexity? This question attracts a lot of interest, with recent works tackling computationally hard tasks such as the traveling salesman problem and satisfiability. In this work we offer a different perspective on this question. Given the common assumption that NP != coNP we prove that any polynomial-time sample generator for an NP-hard problem samples, in fact, from an easier sub-problem. We empirically explore a case study, Conjunctive Query Containment, and show how common data generation techniques generate biased data-sets that lead practitioners to over-estimate model accuracy. Our results suggest that machine learning approaches that require training on a dense uniform sampling from the target distribution cannot be used to solve computationally hard problems, the reason being the difficulty of generating sufficiently large and unbiased training sets.

Data Valuation using Reinforcement Learning

- Jinsung Yoon, Sercan Arik, Tomas Pfister
- abstract: Quantifying the value of data is a fundamental problem in machine learning and has multiple important use cases: (1) building insights about the dataset and task, (2) domain adaptation, (3) corrupted sample discovery, and (4) robust learning. We propose Data Valuation using Reinforcement Learning (DVRL), to adaptively learn data values jointly with the predictor model. DVRL uses a data value estimator (DVE) to learn how likely each

datum is used in training of the predictor model. DVE is trained using a reinforcement signal that reflects performance on the target task. We demonstrate that DVRL yields superior data value estimates compared to alternative methods across numerous datasets and application scenarios. The corrupted sample discovery performance of DVRL is close to optimal in many regimes (i.e. as if the noisy samples were known apriori), and for domain adaptation and robust learning DVRL significantly outperforms state-of-the-art by 14.6% and 10.8%, respectively.

XtarNet: Learning to Extract Task-Adaptive Representation for Incremental Few-Shot Learning

- Sung Whan Yoon, Do-Yeon Kim, Jun Seo, Jaekyun Moon
- abstract: Learning novel concepts while preserving prior knowledge is a long-standing challenge in machine learning. The challenge gets greater when a novel task is given with only a few labeled examples, a problem known as incremental few-shot learning. We propose XtarNet, which learns to extract task-adaptive representation (TAR) for facilitating incremental few-shot learning. The method utilizes a backbone network pretrained on a set of base categories while also employing additional modules that are meta-trained across episodes. Given a new task, the novel feature extracted from the meta-trained modules is mixed with the base feature obtained from the pretrained model. The process of combining two different features provides TAR and is also controlled by meta-trained modules. The TAR contains effective information for classifying both novel and base categories. The base and novel classifiers quickly adapt to a given task by utilizing the TAR. Experiments on standard image datasets indicate that XtarNet achieves state-of-the-art incremental few-shot learning performance. The concept of TAR can also be used in conjunction with existing incremental few-shot learning methods; extensive simulation results in fact show that applying TAR enhances the known methods significantly.

Robustifying Sequential Neural Processes

- Jaesik Yoon, Gautam Singh, Sungjin Ahn
- abstract: When tasks change over time, meta-transfer learning seeks to improve the efficiency of learning a new task via both meta-learning and transfer-learning. While the standard attention has been effective in a variety of settings, we question its effectiveness in improving meta-transfer learning since the tasks being learned are dynamic and the amount of context can be substantially smaller. In this paper, using a recently proposed meta-transfer learning model, Sequential Neural Processes (SNP), we first empirically show that it suffers from a similar underfitting problem observed in the functions inferred by Neural Processes. However, we further demonstrate that unlike the meta-learning setting, the standard attention mechanisms are not effective in meta-transfer setting. To resolve, we propose a new attention mechanism, Recurrent Memory Reconstruction (RMR), and demonstrate that providing an imaginary context that is recurrently updated and reconstructed with interaction is crucial in achieving effective attention for meta-transfer learning. Furthermore, incorporating RMR into SNP, we propose Attentive Sequential Neural Processes-RMR (ASNP-RMR) and demonstrate in various tasks that ASNP-RMR significantly outperforms the baselines.

When Does Self-Supervision Help Graph Convolutional Networks?

- Yuning You, Tianlong Chen, Zhangyang Wang, Yang Shen
- abstract: Self-supervision as an emerging technique has been employed to train convolutional neural networks (CNNs) for more transferrable, generalizable, and robust representation learning of images. Its introduction to graph convolutional networks (GCNs) operating on graph data is however rarely explored. In this study, we report the first systematic exploration and assessment of incorporating self-supervision into GCNs. We first elaborate three mechanisms to incorporate self-supervision into GCNs, analyze the limitations of pretraining & finetuning and self-training, and proceed to focus on multi-task learning. Moreover, we propose to investigate three novel self-supervised learning tasks for GCNs with theoretical rationales and numerical comparisons. Lastly, we further integrate multi-task self-supervision into graph adversarial training. Our results show that, with properly designed task forms and incorporation mechanisms, self-supervision benefits GCNs in gaining more generalizability and robustness. Our codes are available at https://github.com/Shen-Lab/SS-GCNs.

Graph Structure of Neural Networks

- Jiaxuan You, Jure Leskovec, Kaiming He, Saining Xie
- abstract: Neural networks are often represented as graphs of connections between neurons. However, despite their wide use, there is currently little understanding of the relationship between the graph structure of the neural network and its predictive performance. Here we systematically investigate how does the graph structure of neural networks affect their predictive performance. To this end, we develop a novel graph-based representation of neural networks called relational graph, where layers of neural network computation correspond to rounds of message exchange along the graph structure. Using this representation we show that: (1) a “sweet spot” of relational graphs leads to neural networks with significantly improved predictive performance; (2) neural network’s performance is approximately a smooth function of the clustering coefficient and average path length of its relational graph; (3) our findings are consistent across many different tasks and datasets; (4) the sweet spot can be identified efficiently; (5) top-performing neural networks have graph structure surprisingly similar to those of real biological neural networks. Our work opens new directions for the design of neural architectures and the understanding on neural networks in general.

Simultaneous Inference for Massive Data: Distributed Bootstrap

- Yang Yu, Shih-Kang Chao, Guang Cheng
- abstract: In this paper, we propose a bootstrap method applied to massive data processed distributedly in a large number of machines. This new method is computationally efficient in that we bootstrap on the master machine without over-resampling, typically required by existing methods (Kleiner et al., 2014; Sengupta et al., 2016), while provably achieving optimal statistical efficiency with minimal communication. Our method does not require repeatedly re-fitting the model but only applies multiplier bootstrap in the master machine on the gradients received from the worker machines. Simulations validate our theory.

Graphical Models Meet Bandits: A Variational Thompson Sampling Approach

- Tong Yu, Branislav Kveton, Zheng Wen, Ruiyi Zhang, Ole J. Mengshoel
- abstract: We propose a novel framework for structured bandits, which we call an influence diagram bandit. Our framework uses a graphical model to capture complex statistical dependencies between actions, latent variables, and observations; and thus unifies and extends many existing models, such as combinatorial semi-bandits, cascading bandits, and low-rank bandits. We develop novel online learning algorithms that learn to act efficiently in our models. The key idea is to track a structured posterior distribution of model parameters, either exactly or approximately. To act, we sample model parameters from their posterior and then use the structure of the influence diagram to find the most optimistic action under the sampled parameters. We empirically evaluate our algorithms in three structured bandit problems, and show that they perform as well as or better than problem-specific state-of-the-art baselines.

Label-Noise Robust Domain Adaptation

- Xiyu Yu, Tongliang Liu, Mingming Gong, Kun Zhang, Kayhan Batmanghelich, Dacheng Tao
- abstract: Domain adaptation aims to correct the classifiers when faced with distribution shift between source (training) and target (test) domains. State-of-the-art domain adaptation methods make use of deep networks to extract domain-invariant representations. However, existing methods assume that all the

instances in the source domain are correctly labeled; while in reality, it is unsurprising that we may obtain a source domain with noisy labels. In this paper, we are the first to comprehensively investigate how label noise could adversely affect existing domain adaptation methods in various scenarios. Further, we theoretically prove that there exists a method that can essentially reduce the side-effect of noisy source labels in domain adaptation. Specifically, focusing on the generalized target shift scenario, where both label distribution P_Y and the class-conditional distribution $P_{X|Y}$ can change, we discover that the denoising Conditional Invariant Component (DCIC) framework can provably ensures (1) extracting invariant representations given examples with noisy labels in the source domain and unlabeled examples in the target domain and (2) estimating the label distribution in the target domain with no bias. Experimental results on both synthetic and real-world data verify the effectiveness of the proposed method.

Intrinsic Reward Driven Imitation Learning via Generative Model

- Xingrui Yu, Yueming Lyu, Ivor Tsang
- abstract: Imitation learning in a high-dimensional environment is challenging. Most inverse reinforcement learning (IRL) methods fail to outperform the demonstrator in such a high-dimensional environment, e.g., Atari domain. To address this challenge, we propose a novel reward learning module to generate intrinsic reward signals via a generative model. Our generative method can perform better forward state transition and backward action encoding, which improves the module's dynamics modeling ability in the environment. Thus, our module provides the imitation agent both the intrinsic intention of the demonstrator and a better exploration ability, which is critical for the agent to outperform the demonstrator. Empirical results show that our method outperforms state-of-the-art IRL methods on multiple Atari games, even with one-life demonstration. Remarkably, our method achieves performance that is up to 5 times the performance of the demonstration.

Graph Convolutional Network for Recommendation with Low-pass Collaborative Filters

- Wenhui Yu, Zheng Qin
- abstract: \textbf{G}raph \textbf{C}onvolutional \textbf{N}etwork (\textbf{GCN}) is widely used in graph data learning tasks such as recommendation. However, when facing a large graph, the graph convolution is very computationally expensive thus is simplified in all existing GCNs, yet is seriously impaired due to the oversimplification. To address this gap, we leverage the \emph{original graph convolution} in GCN and propose a \textbf{L}ow-pass \textbf{C}ollaborative \textbf{F}ilter (\textbf{LCF}) to make it applicable to the large graph. LCF is designed to remove the noise caused by exposure and quantization in the observed data, and it also reduces the complexity of graph convolution in an unscathed way. Experiments show that LCF improves the effectiveness and efficiency of graph convolution and our GCN outperforms existing GCNs significantly. Codes are available on \url{https://github.com/Wenhui-Yu/LCFN}.

Federated Learning with Only Positive Labels

- Felix Yu, Ankit Singh Rawat, Aditya Menon, Sanjiv Kumar
- abstract: We consider learning a multi-class classification model in the federated setting, where each user has access to the positive data associated with only a single class. As a result, during each federated learning round, the users need to locally update the classifier without having access to the features and the model parameters for the negative classes. Thus, naively employing conventional decentralized learning such as distributed SGD or Federated Averaging may lead to trivial or extremely poor classifiers. In particular, for embedding based classifiers, all the class embeddings might collapse to a single point. To address this problem, we propose a generic framework for training with only positive labels, namely Federated Averaging with Spreadout (FedAwS), where the server imposes a geometric regularizer after each round to encourage classes to be spreadout in the embedding space. We show, both theoretically and empirically, that FedAwS can almost match the performance of conventional learning where users have access to negative labels. We further extend the proposed method to settings with large output spaces.

Training Deep Energy-Based Models with f-Divergence Minimization

- Lantao Yu, Yang Song, Jiaming Song, Stefano Ermon
- abstract: Deep energy-based models (EBMs) are very flexible in distribution parametrization but computationally challenging because of the intractable partition function. They are typically trained via maximum likelihood, using contrastive divergence to approximate the gradient of the KL divergence between data and model distribution. While KL divergence has many desirable properties, other f-divergences have shown advantages in training implicit density generative models such as generative adversarial networks. In this paper, we propose a general variational framework termed f-EBM to train EBMs using any desired f-divergence. We introduce a corresponding optimization algorithm and prove its local convergence property with non-linear dynamical systems theory. Experimental results demonstrate the superiority of f-EBM over contrastive divergence, as well as the benefits of training EBMs using f-divergences other than KL.

Graph Random Neural Features for Distance-Preserving Graph Representations

- Daniele Zambon, Cesare Alippi, Lorenzo Livi
- abstract: We present Graph Random Neural Features (GRNF), a novel embedding method from graph-structured data to real vectors based on a family of graph neural networks. The embedding naturally deals with graph isomorphism and preserves the metric structure of the graph domain, in probability. In addition to being an explicit embedding method, it also allows us to efficiently and effectively approximate graph metric distances (as well as complete kernel functions); a criterion to select the embedding dimension trading off the approximation accuracy with the computational cost is also provided. GRNF can be used within traditional processing methods or as a training-free input layer of a graph neural network. The theoretical guarantees that accompany GRNF ensure that the considered graph distance is metric, hence allowing to distinguish any pair of non-isomorphic graphs.

Learning Near Optimal Policies with Low Inherent Bellman Error

- Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, Emma Brunskill
- abstract: We study the exploration problem with approximate linear action-value functions in episodic reinforcement learning under the notion of low inherent Bellman error, a condition normally employed to show convergence of approximate value iteration. First we relate this condition to other common frameworks and show that it is strictly more general than the low rank (or linear) MDP assumption of prior work. Second we provide an algorithm with a high probability regret bound $\widetilde O(\sum_{t=1}^H d_t \sqrt{K} + \sum_{t=1}^H \sqrt{d_t} \IBE K)$ where H is the horizon, K is the number of episodes, $\IBE$ is the value if the inherent Bellman error and d_t is the feature dimension at timestep t. In addition, we show that the result is unimprovable beyond constants and logs by showing a matching lower bound. This has two important consequences: 1) it shows that exploration is possible using only \emph{batch assumptions} with an algorithm that achieves the optimal statistical rate for the setting we consider, which is more general than prior work on low-rank MDPs 2) the lack of closedness (measured by the inherent Bellman error) is only amplified by $\sqrt{d_t}$ despite working in the online setting. Finally, the algorithm reduces to the celebrated \textsc{LinUCB} when $H=1$ but with a different choice of the exploration parameter that allows handling misspecified contextual linear bandits. While computational tractability questions remain open for the MDP setting, this enriches the class of MDPs with a linear representation for the action-value function where statistically efficient reinforcement learning is possible.

Scaling up Hybrid Probabilistic Inference with Logical and Arithmetic Constraints via Message Passing

- Zhe Zeng, Paolo Morettin, Fanqi Yan, Antonio Vergari, Guy Van Den Broeck
- abstract: Weighted model integration (WMI) is an appealing framework for probabilistic inference: it allows for expressing the complex dependencies in real-world problems, where variables are both continuous and discrete, via the language of Satisfiability Modulo Theories (SMT), as well as to compute probabilistic queries with complex logical and arithmetic constraints. Yet, existing WMI solvers are not ready to scale to these problems. They either ignore the intrinsic dependency structure of the problem entirely, or they are limited to overly restrictive structures. To narrow this gap, we derive a factorized WMI computation enabling us to devise a scalable WMI solver based on message passing, called MP-WMI. Namely, MP-WMI is the first WMI solver that can (i) perform exact inference on the full class of tree-structured WMI problems, and (ii) perform inter-query amortization, e.g., to compute all marginal densities simultaneously. Experimental results show that our solver dramatically outperforms the existingWMI solvers on a large set of benchmarks.

Learning Calibratable Policies using Programmatic Style-Consistency

- Eric Zhan, Albert Tseng, Yisong Yue, Adith Swaminathan, Matthew Hausknecht
- abstract: We study the problem of controllable generation of long-term sequential behaviors, where the goal is to calibrate to multiple behavior styles simultaneously. In contrast to the well-studied areas of controllable generation of images, text, and speech, there are two questions that pose significant challenges when generating long-term behaviors: how should we specify the factors of variation to control, and how can we ensure that the generated behavior faithfully demonstrates combinatorially many styles? We leverage programmatic labeling functions to specify controllable styles, and derive a formal notion of style-consistency as a learning objective, which can then be solved using conventional policy learning approaches. We evaluate our framework using demonstrations from professional basketball players and agents in the MuJoCo physics environment, and show that existing approaches that do not explicitly enforce style-consistency fail to generate diverse behaviors whereas our learned policies can be calibrated for up to $4^5 (1024)$ distinct style combinations.

Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach

- Junzhe Zhang
- abstract: A dynamic treatment regime (DTR) consists of a sequence of decision rules, one per stage of intervention, that dictates how to determine the treatment assignment to patients based on evolving treatments and covariates' history. These regimes are particularly effective for managing chronic disorders and is arguably one of the critical ingredients underlying more personalized decision-making systems. All reinforcement learning algorithms for finding the optimal DTR in online settings will suffer O(\sqrt{|D_{X, S}|T}) regret on some environments, where T is the number of experiments, and D_{X, S} is the domains of treatments X and covariates S. This implies T = O (|D_{X, S}|) trials to generate an optimal DTR. In many applications, domains of X and S could be so enormous that the time required to ensure appropriate learning may be unattainable. We show that, if the causal diagram of the underlying environment is provided, one could achieve regret that is exponentially smaller than D_{X, S}. In particular, we develop two online algorithms that satisfy such regret bounds by exploiting the causal structure underlying the DTR; one is based on the principle of optimism in the face of uncertainty (OFU-DTR), and the other uses the posterior sampling learning (PS-DTR). Finally, we introduce efficient methods to accelerate these online learning procedures by leveraging the abundant, yet biased observational (non-experimental) data.

Robustness to Programmable String Transformations via Augmented Abstract Training

- Yuhao Zhang, Aws Albarghouthi, Loris D'Antoni
- abstract: Deep neural networks for natural language processing tasks are vulnerable to adversarial input perturbations. In this paper, we present a versatile language for programmatically specifying string transformations—e.g., insertions, deletions, substitutions, swaps, etc.—that are relevant to the task at hand. We then present an approach to adversarially training models that are robust to such user-defined string transformations. Our approach combines the advantages of search-based techniques for adversarial training with abstraction-based techniques. Specifically, we show how to decompose a set of user-defined string transformations into two component specifications, one that benefits from search and another from abstraction. We use our technique to train models on the AG and SST2 datasets and show that the resulting models are robust to combinations of user-defined transformations mimicking spelling mistakes and other meaning-preserving transformations.

Converging to Team-Maxmin Equilibria in Zero-Sum Multiplayer Games

- Youzhi Zhang, Bo An
- abstract: Efficiently computing equilibria for multiplayer games is still an open challenge in computational game theory. This paper focuses on computing Team-Maxmin Equilibria (TMEs), which is an important solution concept for zero-sum multiplayer games where players in a team having the same utility function play against an adversary independently. Existing algorithms are inefficient to compute TMEs in large games, especially when the strategy space is too large to be represented due to limited memory. In two-player games, the Incremental Strategy Generation (ISG) algorithm is an efficient approach to avoid enumerating all pure strategies. However, the study of ISG for computing TMEs is completely unexplored. To fill this gap, we first study the properties of ISG for multiplayer games, showing that ISG converges to a Nash Equilibrium (NE) but may not converge to a TME. Second, we design an ISG variant for TMEs (ISGT) by exploiting that a TME is an NE maximizing the team's utility and show that ISGT converges to a TME and the impossibility of relaxing conditions in ISGT. Third, to further improve the scalability, we design an ISGT variant (CISGT) by using the strategy space for computing an equilibrium that is close to a TME but is easier to be computed as the initial strategy space of ISGT. Finally, extensive experimental results show that CISGT is orders of magnitude faster than ISGT and the state-of-the-art algorithm to compute TMEs in large games.

Generative Adversarial Imitation Learning with Neural Network Parameterization: Global Optimality and Convergence Rate

- Yufeng Zhang, Qi Cai, Zhuoran Yang, Zhaoran Wang
- abstract: Generative adversarial imitation learning (GAIL) demonstrates tremendous success in practice, especially when combined with neural networks. Different from reinforcement learning, GAIL learns both policy and reward function from expert (human) demonstration. Despite its empirical success, it remains unclear whether GAIL with neural networks converges to the globally optimal solution. The major difficulty comes from the nonconvex-nonconcave minimax optimization structure. To bridge the gap between practice and theory, we analyze a gradient-based algorithm with alternating updates and establish its sublinear convergence to the globally optimal solution. To the best of our knowledge, our analysis establishes the global optimality and convergence rate of GAIL with neural networks for the first time.

Cautious Adaptation For Reinforcement Learning in Safety-Critical Settings

- Jesse Zhang, Brian Cheung, Chelsea Finn, Sergey Levine, Dinesh Jayaraman
- abstract: Reinforcement learning (RL) in real-world safety-critical target settings like urban driving is hazardous, imperiling the RL agent, other agents, and the environment. To overcome this difficulty, we propose a "safety-critical adaptation" task setting: an agent first trains in non-safety-critical "source" environments such as in a simulator, before it adapts to the target environment where failures carry heavy costs. We propose a solution approach, CARL, that builds on the intuition that prior experience in diverse environments equips an agent to estimate risk, which in turn enables relative safety through risk-averse, cautious adaptation. CARL first employs model-based RL to train a probabilistic model to capture uncertainty about transition dynamics and catastrophic states across varied source environments. Then, when exploring a new safety-critical environment with unknown dynamics, the CARL agent plans to avoid actions that could lead to catastrophic states. In experiments on car driving, cartpole balancing, and half-cheetah locomotion, CARL successfully acquires cautious exploration behaviors, yielding higher rewards with fewer failures than strong RL adaptation baselines.

Learning the Valuations of a k-demand Agent

- Hanrui Zhang, Vincent Conitzer
- abstract: We study problems where a learner aims to learn the valuations of an agent by observing which goods he buys under varying price vectors. More specifically, we consider the case of a k-demand agent, whose valuation over the goods is additive when receiving up to k goods, but who has no interest in receiving more than k goods. We settle the query complexity for the active-learning (preference elicitation) version, where the learner chooses the prices to post, by giving a \emph{biased binary search} algorithm, generalizing the classical binary search procedure. We complement our query complexity upper bounds by lower bounds that match up to lower-order terms. We also study the passive-learning version in which the learner does not control the prices, and instead they are sampled from some distribution. We show that in the PAC model for passive learning, any \emph{empirical risk minimizer} has a sample complexity that is optimal up to a factor of $\widetilde{O}(k)$.

A Tree-Structured Decoder for Image-to-Markup Generation

- Jianshu Zhang, Jun Du, Yongxin Yang, Yi-Zhe Song, Si Wei, Lirong Dai
- abstract: Recent encoder-decoder approaches typically employ string decoders to convert images into serialized strings for image-to-markup. However, for tree-structured representational markup, string representations can hardly cope with the structural complexity. In this work, we first show via a set of toy problems that string decoders struggle to decode tree structures, especially as structural complexity increases, we then propose a tree-structured decoder that specifically aims at generating a tree-structured markup. Our decoders works sequentially, where at each step a child node and its parent node are simultaneously generated to form a sub-tree. This sub-tree is consequently used to construct the final tree structure in a recurrent manner. Key to the success of our tree decoder is twofold, (i) it strictly respects the parent-child relationship of trees, and (ii) it explicitly outputs trees as oppose to a linear string. Evaluated on both math formula recognition and chemical formula recognition, the proposed tree decoder is shown to greatly outperform strong string decoder baselines.

Approximation Capabilities of Neural ODEs and Invertible Residual Networks

- Han Zhang, Xi Gao, Jacob Unterman, Tom Arodz
- abstract: Recent interest in invertible models and normalizing flows has resulted in new architectures that ensure invertibility of the network model. Neural ODEs and i-ResNets are two recent techniques for constructing models that are invertible, but it is unclear if they can be used to approximate any continuous invertible mapping. Here, we show that out of the box, both of these architectures are limited in their approximation capabilities. We then show how to overcome this limitation: we prove that any homeomorphism on a p-dimensional Euclidean space can be approximated by a Neural ODE or an i-ResNet operating on a $2p$-dimensional Euclidean space. We conclude by showing that capping a Neural ODE or an i-ResNet with a single linear layer is sufficient to turn the model into a universal approximator for non-invertible continuous functions.

Random Hypervolume Scalarizations for Provable Multi-Objective Black Box Optimization

- Richard Zhang, Daniel Golovin
- abstract: Single-objective black box optimization (also known as zeroth-order optimization) is the process of minimizing a scalar objective $f(x)$, given evaluations at adaptively chosen inputs x. In this paper, we consider multi-objective optimization, where $f(x)$ outputs a vector of possibly competing objectives and the goal is to converge to the Pareto frontier. Quantitatively, we wish to maximize the standard \emph{hypervolume indicator} metric, which measures the dominated hypervolume of the entire set of chosen inputs. In this paper, we introduce a novel scalarization function, which we term the \emph{hypervolume scalarization}, and show that drawing random scalarizations from an appropriately chosen distribution can be used to efficiently approximate the \emph{hypervolume indicator} metric. We utilize this connection to show that Bayesian optimization with our scalarization via common acquisition functions, such as Thompson Sampling or Upper Confidence Bound, provably converges to the whole Pareto frontier by deriving tight \emph{hypervolume regret} bounds on the order of $\widetilde{O}(\sqrt{T})$. Furthermore, we highlight the general utility of our scalarization framework by showing that any provably convergent single-objective optimization process can be converted to a multi-objective optimization process with provable convergence guarantees.

Spread Divergence

- Mingtian Zhang, Peter Hayes, Thomas Bird, Raza Habib, David Barber
- abstract: For distributions $\mathbb{P}$ and $\mathbb{Q}$ with different supports or undefined densities, the divergence $\textrm{D}(\mathbb{P}||\mathbb{Q})$ may not exist. We define a Spread Divergence $\tilde{\textrm{D}}(\mathbb{P}||\mathbb{Q})$ on modified $\mathbb{P}$ and $\mathbb{Q}$ and describe sufficient conditions for the existence of such a divergence. We demonstrate how to maximize the discriminatory power of a given divergence by parameterizing and learning the spread. We also give examples of using a Spread Divergence to train implicit generative models, including linear models (Independent Components Analysis) and non-linear models (Deep Generative Networks).

Mix-n-Match : Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning

- Jize Zhang, Bhavya Kailkhura, T. Yong-Jin Han
- abstract: This paper studies the problem of post-hoc calibration of machine learning classifiers. We introduce the following desiderata for uncertainty calibration: (a) accuracy-preserving, (b) data-efficient, and (c) high expressive power. We show that none of the existing methods satisfy all three requirements, and demonstrate how Mix-n-Match calibration strategies (i.e., ensemble and composition) can help achieve remarkably better data-efficiency and expressive power while provably maintaining the classification accuracy of the original classifier. Mix-n-Match strategies are generic in the sense that they can be used to improve the performance of any off-the-shelf calibrator. We also reveal potential issues in standard evaluation practices. Popular approaches (e.g., histogram-based expected calibration error (ECE)) may provide misleading results especially in small-data regime. Therefore, we propose an alternative data-efficient kernel density-based estimator for a reliable evaluation of the calibration performance and prove its asymptotically unbiasedness and consistency. Our approaches outperform state-of-the-art solutions on both the calibration as well as the evaluation tasks in most of the experimental settings. Our codes are available at https://github.com/zhang64- llnl/Mix-n-Match-Calibration.

Privately Learning Markov Random Fields

- Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, Steven Wu
- abstract: We consider the problem of learning Markov Random Fields (including the prototypical example, the Ising model) under the constraint of differential privacy. Our learning goals include both \emph{structure learning}, where we try to estimate the underlying graph structure of the model, as well as the harder goal of \emph{parameter learning}, in which we additionally estimate the parameter on each edge. We provide algorithms and lower bounds for both problems under a variety of privacy constraints – namely pure, concentrated, and approximate differential privacy. While non-privately, both learning goals enjoy roughly the same complexity, we show that this is not the case under differential privacy. In particular, only structure learning under approximate differential privacy maintains the non-private logarithmic dependence on the dimensionality of the data, while a change in either the learning goal or the privacy notion would necessitate a polynomial dependence. As a result, we show that the privacy constraint imposes a strong separation between these two learning problems in the high-dimensional data regime.

Learning Structured Latent Factors from Dependent Data:A Generative Model Framework from Information-Theoretic Perspective

- Ruixiang Zhang, Masanori Koyama, Katsuhiko Ishiguro
- abstract: Learning controllable and generalizable representation of multivariate data with desired structural properties remains a fundamental problem in machine learning. In this paper, we present a novel framework for learning generative models with various underlying structures in the latent space. Learning controllable and generalizable representation of multivariate data with desired structural properties remains a fundamental problem in machine learning. In this paper, we present a novel framework for learning generative models with various underlying structures in the latent space. We represent the inductive bias in the form of mask variables to model the dependency structure in the graphical model and extend the theory of multivariate information bottleneck (Friedman et al., 2001) to enforce it. Our model provides a principled approach to learn a set of semantically meaningful latent factors that reflect various types of desired structures like capturing correlation or encoding invariance, while also offering the flexibility to automatically estimate the dependency structure from data. We show that our framework unifies many existing generative models and can be applied to a variety of tasks, including multimodal data modeling, algorithmic fairness, and out-of-distribution generalization.

Optimal Estimator for Unlabeled Linear Regression

- Hang Zhang, Ping Li
- abstract: Unlabeled linear regression, or “linear regression with an unknown permutation”, has attracted increasing attentions due to its applications in (e.g.,) linkage record and de-anonymization. However, the computation of unlabeled linear regression proves to be cumbersome and existing algorithms typically require considerable time, especially in the high dimensional regime. In this paper, we propose a one-step estimator which is optimal from both the computational and the statistical aspects. From the computational perspective, our estimator exhibits the same order of computational complexity as that of the oracle case (which means the regression coefficients are known in advance and only the permutation needs recovery). From the statistical perspective, when comparing with the necessary conditions for permutation recovery, our requirement on the \emph{signal-to-noise ratio} (SNR) agrees up to merely $\Omega\left(\log \log n\right)$ difference when the stable rank of the regression coefficients $\ensuremath{\mathbf{B}}^{\natural}$ is much less than $\log n/\log \log n$.

Dual-Path Distillation: A Unified Framework to Improve Black-Box Attacks

- Yonggang Zhang, Ya Li, Tongliang Liu, Xinmei Tian
- abstract: We study the problem of constructing black-box adversarial attacks, where no model information is revealed except for the feedback knowledge of the given inputs. To obtain sufficient knowledge for crafting adversarial examples, previous methods query the target model with inputs that are perturbed with different searching directions. However, these methods suffer from poor query efficiency since the employed searching directions are sampled randomly. To mitigate this issue, we formulate the goal of mounting efficient attacks as an optimization problem in which the adversary tries to fool the target model with a limited number of queries. Under such settings, the adversary has to select appropriate searching directions to reduce the number of model queries. By solving the efficient-attack problem, we find that we need to distill the knowledge in both the path of the adversarial examples and the path of the searching directions. Therefore, we propose a novel framework, dual-path distillation, that utilizes the feedback knowledge not only to craft adversarial examples but also to alter the searching directions to achieve efficient attacks. Experimental results suggest that our framework can significantly increase the query efficiency.

Complexity of Finding Stationary Points of Nonconvex Nonsmooth Functions

- Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, Ali Jadbabaie
- abstract: We provide the first non-asymptotic analysis for finding stationary points of nonsmooth, nonconvex functions. In particular, we study the class of Hadamard semi-differentiable functions, perhaps the largest class of nonsmooth functions for which the chain rule of calculus holds. This class contains important examples such as ReLU neural networks and others with non-differentiable activation functions. First, we show that finding an epsilon-stationary point with first-order methods is impossible in finite time. Therefore, we introduce the notion of (delta, epsilon)-stationarity, a generalization that allows for a point to be within distance delta of an epsilon-stationary point and reduces to epsilon-stationarity for smooth functions. We propose a series of randomized first-order methods and analyze their complexity of finding a (delta, epsilon)-stationary point. Furthermore, we provide a lower bound and show that our stochastic algorithm has min-max optimal dependence on delta. Empirically, our methods perform well for training ReLU neural networks.

Self-Attentive Hawkes Process

- Qiang Zhang, Aldo Lipani, Omer Kirnap, Emine Yilmaz
- abstract: Capturing the occurrence dynamics is crucial to predicting which type of events will happen next and when. A common method to do this is through Hawkes processes. To enhance their capacity, recurrent neural networks (RNNs) have been incorporated due to RNNs’ successes in processing sequential data such as languages. Recent evidence suggests that self-attention is more competent than RNNs in dealing with languages. However, we are unaware of the effectiveness of self-attention in the context of Hawkes processes. This study aims to fill the gap by designing a self-attentive Hawkes process (SAHP). SAHP employs self-attention to summarise the influence of history events and compute the probability of the next event. One deficit of the conventional self-attention when applied to event sequences is that its positional encoding only considers the order of a sequence ignoring the time intervals between events. To overcome this deficit, we modify its encoding by translating time intervals into phase shifts of sinusoidal functions. Experiments on goodness-of-fit and prediction tasks show the improved capability of SAHP. Furthermore, SAHP is more interpretable than RNN-based counterparts because the learnt attention weights reveal contributions of one event type to the happening of another type. To the best of our knowledge, this is the first work that studies the effectiveness of self-attention in Hawkes processes.

GradientDICE: Rethinking Generalized Offline Estimation of Stationary Values

- Shangtong Zhang, Bo Liu, Shimon Whiteson
- abstract: We present GradientDICE for estimating the density ratio between the state distribution of the target policy and the sampling distribution in off-policy reinforcement learning. GradientDICE fixes several problems of GenDICE (Zhang et al., 2020), the current state-of-the-art for estimating such density ratios. Namely, the optimization problem in GenDICE is not a convex-concave saddle-point problem once nonlinearity in optimization variable parameterization is introduced to ensure positivity, so primal-dual algorithms are not guaranteed to find the desired solution. However, such nonlinearity is essential to ensure the consistency of GenDICE even with a tabular representation. This is a fundamental contradiction, resulting from GenDICE’s original formulation of the optimization problem. In GradientDICE, we optimize a different objective from GenDICE by using the Perron-Frobenius theorem and eliminating GenDICE’s use of divergence, such that nonlinearity in parameterization is not necessary for GradientDICE, which is provably convergent under linear function approximation.

Provably Convergent Two-Timescale Off-Policy Actor-Critic with Function Approximation

- Shangtong Zhang, Bo Liu, Hengshuai Yao, Shimon Whiteson

- abstract: We present the first provably convergent two-timescale off-policy actor-critic algorithm (COF-PAC) with function approximation. Key to COF-PAC is the introduction of a new critic, the emphasis critic, which is trained via Gradient Emphasis Learning (GEM), a novel combination of the key ideas of Gradient Temporal Difference Learning and Emphatic Temporal Difference Learning. With the help of the emphasis critic and the canonical value function critic, we show convergence for COF-PAC, where the critics are linear and the actor can be nonlinear.

Invariant Causal Prediction for Block MDPs

- Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, Doina Precup
- abstract: Generalization across environments is critical to the successful application of reinforcement learning (RL) algorithms to real-world challenges. In this work we propose a method for learning state abstractions which generalize to novel observation distributions in the multi-environment RL setting. We prove that for certain classes of environments, this approach outputs, with high probability, a state abstraction corresponding to the causal feature set with respect to the return. We give empirical evidence that analogous methods for the nonlinear setting can also attain improved generalization over single- and multi-task baselines. Lastly, we provide bounds on model generalization error in the multi-environment setting, in the process showing a connection between causal variable identification and the state abstraction framework for MDPs.

Adaptive Reward-Poisoning Attacks against Reinforcement Learning

- Xuezhou Zhang, Yuzhe Ma, Adish Singla, Xiaojin Zhu
- abstract: In reward-poisoning attacks against reinforcement learning (RL), an attacker can perturb the environment reward r_t into $r_t+\delta_t$ at each step, with the goal of forcing the RL agent to learn a nefarious policy. We categorize such attacks by the infinity-norm constraint on δ_t: We provide a lower threshold below which reward-poisoning attack is infeasible and RL is certified to be safe; we provide a corresponding upper threshold above which the attack is feasible. Feasible attacks can be further categorized as non-adaptive where δ_t depends only on (s_t,a_t, s_{t+1}), or adaptive where δ_t depends further on the RL agent's learning process at time t. Non-adaptive attacks have been the focus of prior works. However, we show that under mild conditions, adaptive attacks can achieve the nefarious policy in steps polynomial in state-space size $|S|$, whereas non-adaptive attacks require exponential steps. We provide a constructive proof that a Fast Adaptive Attack strategy achieves the polynomial rate. Finally, we show that empirically an attacker can find effective reward-poisoning attacks using state-of-the-art deep RL techniques.

CAUSE: Learning Granger Causality from Event Sequences using Attribution Methods

- Wei Zhang, Thomas Panum, Somesh Jha, Prasad Chalasani, David Page
- abstract: We study the problem of learning Granger causality between event types from asynchronous, interdependent, multi-type event sequences. Existing work suffers from either limited model flexibility or poor model explainability and thus fails to uncover Granger causality across a wide variety of event sequences with diverse event interdependency. To address these weaknesses, we propose CAUSE (Causality from AttribUtions on Sequence of Events), a novel framework for the studied task. The key idea of CAUSE is to first implicitly capture the underlying event interdependency by fitting a neural point process, and then extract from the process a Granger causality statistic using an axiomatic attribution method. Across multiple datasets riddled with diverse event interdependency, we demonstrate that CAUSE achieves superior performance on correctly inferring the inter-type Granger causality over a range of state-of-the-art methods.

Convex Calibrated Surrogates for the Multi-Label F-Measure

- Mingyuan Zhang, Harish Guruprasad Ramaswamy, Shivani Agarwal
- abstract: The F-measure is a widely used performance measure for multi-label classification, where multiple labels can be active in an instance simultaneously (e.g. in image tagging, multiple tags can be active in any image). In particular, the F-measure explicitly balances recall (fraction of active labels predicted to be active) and precision (fraction of labels predicted to be active that are actually so), both of which are important in evaluating the overall performance of a multi-label classifier. As with most discrete prediction problems, however, directly optimizing the F-measure is computationally hard. In this paper, we explore the question of designing convex surrogate losses that are calibrated for the F-measure – specifically, that have the property that minimizing the surrogate loss yields (in the limit of sufficient data) a Bayes optimal multi-label classifier for the F-measure. We show that the F-measure for an s-label problem, when viewed as a $2^s \times 2^s$ loss matrix, has rank at most s^2+1, and apply a result of Ramaswamy et al. (2014) to design a family of convex calibrated surrogates for the F-measure. The resulting surrogate risk minimization algorithms can be viewed as decomposing the multi-label F-measure learning problem into s^2+1 binary class probability estimation problems. We also provide a quantitative regret transfer bound for our surrogates, which allows any regret guarantees for the binary problems to be transferred to regret guarantees for the overall F-measure problem, and discuss a connection with the algorithm of Dembczynski et al. (2013). Our experiments confirm our theoretical findings.

Sparsified Linear Programming for Zero-Sum Equilibrium Finding

- Brian Zhang, Tuomas Sandholm
- abstract: Computational equilibrium finding in large zero-sum extensive-form imperfect-information games has led to significant recent AI breakthroughs. The fastest algorithms for the problem are new forms of counterfactual regret minimization (Brown & Sandholm, 2019). In this paper we present a totally different approach to the problem, which is competitive and often orders of magnitude better than the prior state of the art. The equilibrium-finding problem can be formulated as a linear program (LP) (Koller et al., 1994), but solving it as an LP has not been scalable due to the memory requirements of LP solvers, which can often be quadratically worse than CFR-based algorithms. We give an efficient practical algorithm that factors a large payoff matrix into a product of two matrices that are typically dramatically sparser. This allows us to express the equilibrium-finding problem as a linear program with size only a logarithmic factor worse than CFR, and thus allows linear program solvers to run on such games. With experiments on poker endgames, we demonstrate in practice, for the first time, that modern linear program solvers are competitive against even game-specific modern variants of CFR in solving large extensive-form games, and can be used to compute exact solutions unlike iterative algorithms like CFR.

Fast Learning of Graph Neural Networks with Guaranteed Generalizability: One-hidden-layer Case

- Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong
- abstract: Although graph neural networks (GNNs) have made great progress recently on learning from graph-structured data in practice, their theoretical guarantee on generalizability remains elusive in the literature. In this paper, we provide a theoretically-grounded generalizability analysis of GNNs with one hidden layer for both regression and binary classification problems. Under the assumption that there exists a ground-truth GNN model (with zero generalization error), the objective of GNN learning is to estimate the ground-truth GNN parameters from the training data. To achieve this objective, we propose a learning algorithm that is built on tensor initialization and accelerated gradient descent. We then show that the proposed learning algorithm converges to the ground-truth GNN model for the regression problem, and to a model sufficiently close to the ground-truth for the binary classification problem. Moreover, for both cases, the convergence rate of the proposed learning algorithm is proven to be linear and faster than the vanilla gradient descent algorithm. We further explore the relationship between the sample complexity of GNNs and their underlying graph properties. Lastly, we provide numerical experiments to demonstrate the validity of our analysis and the effectiveness of the proposed learning algorithm for GNNs.

Attacks Which Do Not Kill Training Make Adversarial Learning Stronger

- Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, Mohan Kankanhalli
- abstract: Adversarial training based on the minimax formulation is necessary for obtaining adversarial robustness of trained models. However, it is conservative or even pessimistic so that it sometimes hurts the natural generalization. In this paper, we raise a fundamental question{—}do we have to trade off natural generalization for adversarial robustness? We argue that adversarial training is to employ confident adversarial data for updating the current model. We propose a novel formulation of friendly adversarial training (FAT): rather than employing most adversarial data maximizing the loss, we search for least adversarial data (i.e., friendly adversarial data) minimizing the loss, among the adversarial data that are confidently misclassified. Our novel formulation is easy to implement by just stopping the most adversarial data searching algorithms such as PGD (projected gradient descent) early, which we call early-stopped PGD. Theoretically, FAT is justified by an upper bound of the adversarial risk. Empirically, early-stopped PGD allows us to answer the earlier question negatively{—}adversarial robustness can indeed be achieved without compromising the natural generalization.

A Flexible Latent Space Model for Multilayer Networks

- Xuefei Zhang, Songkai Xue, Ji Zhu
- abstract: Entities often interact with each other through multiple types of relations, which are often represented as multilayer networks. Multilayer networks among the same set of nodes usually share common structures, while each layer can possess its distinct node connecting behaviors. This paper proposes a flexible latent space model for multilayer networks for the purpose of capturing such characteristics. Specifically, the proposed model embeds each node with a latent vector shared among layers and a layer-specific effect for each layer; both elements together with a layer-specific connectivity matrix determine edge formations. To fit the model, we develop a projected gradient descent algorithm for efficient parameter estimation. We also establish theoretical properties of the maximum likelihood estimators and show that the upper bound of the common latent structure's estimation error is inversely proportional to the number of layers under mild conditions. The superior performance of the proposed model is demonstrated through simulation studies and applications to two real-world data examples.

Perceptual Generative Autoencoders

- Zijun Zhang, Ruixiang Zhang, Zongpeng Li, Yoshua Bengio, Liam Paull
- abstract: Modern generative models are usually designed to match target distributions directly in the data space, where the intrinsic dimension of data can be much lower than the ambient dimension. We argue that this discrepancy may contribute to the difficulties in training generative models. We therefore propose to map both the generated and target distributions to the latent space using the encoder of a standard autoencoder, and train the generator (or decoder) to match the target distribution in the latent space. Specifically, we enforce the consistency in both the data space and the latent space with theoretically justified data and latent reconstruction losses. The resulting generative model, which we call a perceptual generative autoencoder (PGA), is then trained with a maximum likelihood or variational autoencoder (VAE) objective. With maximum likelihood, PGAs generalize the idea of reversible generative models to unrestricted neural network architectures and arbitrary number of latent dimensions. When combined with VAEs, PGAs substantially improve over the baseline VAEs in terms of sample quality. Compared to other autoencoder-based generative models using simple priors, PGAs achieve state-of-the-art FID scores on CIFAR-10 and CelebA.

Variance Reduction in Stochastic Particle-Optimization Sampling

- Jianyi Zhang, Yang Zhao, Changyou Chen
- abstract: Stochastic particle-optimization sampling (SPOS) is a recently-developed scalable Bayesian sampling framework unifying stochastic gradient MCMC (SG-MCMC) and Stein variational gradient descent (SVGD) algorithms based on Wasserstein gradient flows. With a rigorous non-asymptotic convergence theory developed, SPOS can avoid the particle-collapsing pitfall of SVGD. However, the variance-reduction effect in SPOS has not been clear. In this paper, we address this gap by presenting several variance-reduction techniques for SPOS. Specifically, we propose three variants of variance-reduced SPOS, called SAGA particle-optimization sampling (SAGA-POS), SVRG particle-optimization sampling (SVRG-POS) and a variant of SVRG-POS which avoids full gradient computations, denoted as SVRG-POS$^+$. Importantly, we provide non-asymptotic convergence guarantees for these algorithms in terms of the 2-Wasserstein metric and analyze their complexities. The results show our algorithms yield better convergence rates than existing variance-reduced variants of stochastic Langevin dynamics, though more space is required to store the particles in training. Our theory aligns well with experimental results on both synthetic and real datasets.

Learning with Feature and Distribution Evolvable Streams

- Zhen-Yu Zhang, Peng Zhao, Yuan Jiang, Zhi-Hua Zhou
- abstract: In many real-world applications, data are collected in the form of a stream, whose feature space can evolve over time. For instance, in the environmental monitoring task, features can be dynamically vanished or augmented due to the existence of expired old sensors and deployed new sensors. Furthermore, besides the evolvable feature space, the data distribution is usually changing in the streaming scenario. When both feature space and data distribution are evolvable, it is quite challenging to design algorithms with guarantees, particularly theoretical understandings of generalization ability. To address this difficulty, we propose a novel discrepancy measure for data with evolving feature space and data distribution, named the \emph{evolving discrepancy}. Based on that, we present the generalization error analysis, and the theory motivates the design of a learning algorithm which is further implemented by deep neural networks. Empirical studies on synthetic data verify the rationale of our proposed discrepancy measure, and extensive experiments on real-world tasks validate the effectiveness of our algorithm.

PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

- Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter Liu
- abstract: Recent work pre-training Transformers with self-supervised objectives on large text corpora has shown great success when fine-tuned on downstream NLP tasks including text summarization. However, pre-training objectives tailored for abstractive text summarization have not been explored. Furthermore there is a lack of systematic evaluation across diverse domains. In this work, we propose pre-training large Transformer-based encoder-decoder models on massive text corpora with a new self-supervised objective. In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary. We evaluated our best PEGASUS model on 12 downstream summarization tasks spanning news, science, stories, instructions, emails, patents, and legislative bills. Experiments demonstrate it achieves state-of-the-art performance on all 12 downstream datasets measured by ROUGE scores. Our model also shows surprising performance on low-resource summarization, surpassing previous state-of-the-art results on 6 datasets with only 1000 examples. Finally we validated our results using human evaluation and show that our model summaries achieve human performance on multiple datasets.

On Leveraging Pretrained GANs for Generation with Limited Data

- Miaoyun Zhao, Yulai Cong, Lawrence Carin
- abstract: Recent work has shown generative adversarial networks (GANs) can generate highly realistic images, that are often indistinguishable (by humans) from real images. Most images so generated are not contained in the training dataset, suggesting potential for augmenting training sets with GAN-generated data. While this scenario is of particular relevance when there are limited data available, there is still the issue of training the GAN itself based on that limited data. To facilitate this, we leverage existing GAN models pretrained on large-scale datasets (like ImageNet) to introduce additional knowledge (which may not exist within the limited data), following the concept of transfer learning. Demonstrated by natural-image generation, we reveal that low-level filters (those close to observations) of both the generator and discriminator of pretrained GANs can be transferred to facilitate generation in

a perceptually-distinct target domain with limited training data. To further adapt the transferred filters to the target domain, we propose adaptive filter modulation (AdaFM). An extensive set of experiments is presented to demonstrate the effectiveness of the proposed techniques on generation with limited data.

On Learning Language-Invariant Representations for Universal Machine Translation

- Han Zhao, Junjie Hu, Andrej Risteski
- abstract: The goal of universal machine translation is to learn to translate between any pair of languages. Despite impressive empirical results and an increasing interest in massively multilingual models, theoretical analysis on translation errors made by such universal machine translation models is only nascent. In this paper, we formally prove certain impossibilities of this endeavour in general, as well as prove positive results in the presence of additional (but natural) structure of data. For the former, we derive a lower bound on the translation error in the many-to-many translation setting, which shows that any algorithm aiming to learn shared sentence representations among multiple language pairs has to make a large translation error on at least one of the translation tasks, if no assumption on the structure of the languages is made. For the latter, we show that if the paired documents in the corpus follow a natural \emph{encoder-decoder} generative process, we can expect a natural notion of “generalization”: a linear number of language pairs, rather than quadratic, suffices to learn a good representation. Our theory also explains what kinds of connection graphs between pairs of languages are better suited: ones with longer paths result in worse sample complexity. We believe our theoretical insights and implications contribute to the future algorithmic design of universal machine translation.

Do RNN and LSTM have Long Memory?

- Jingyu Zhao, Feiqing Huang, Jia Lv, Yanjie Duan, Zhen Qin, Guodong Li, Guangjian Tian
- abstract: The LSTM network was proposed to overcome the difficulty in learning long-term dependence, and has made significant advancements in applications. With its success and drawbacks in mind, this paper raises the question - do RNN and LSTM have long memory? We answer it partially by proving that RNN and LSTM do not have long memory from a statistical perspective. A new definition for long memory networks is further introduced, and it requires the model weights to decay at a polynomial rate. To verify our theory, we convert RNN and LSTM into long memory networks by making a minimal modification, and their superiority is illustrated in modeling long-term dependence of various datasets.

Feature Quantization Improves GAN Training

- Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, Changyou Chen
- abstract: The instability in GANs’ training has been a long-standing problem despite remarkable research efforts. We identify that instability issues stem from difficulties of performing feature matching with mini-batch statistics, due to a fragile balance between the fixed target distribution and the progressively generated distribution. In this work, we propose feature quantizatoin (FQ) for the discriminator, to embed both true and fake data samples into a shared discrete space. The quantized values of FQ are constructed as an evolving dictionary, which is consistent with feature statistics of the recent distribution history. Hence, FQ implicitly enables robust feature matching in a compact space. Our method can be easily plugged into existing GAN models, with little computational overhead in training. Extensive experimental results show that the proposed FQ-GAN can improve the FID scores of baseline methods by a large margin on a variety of tasks, including three representative GAN models on 10 benchmarks, achieving new state-of-the-art performance.

Individual Calibration with Randomized Forecasting

- Shengjia Zhao, Tengyu Ma, Stefano Ermon
- abstract: Machine learning applications often require calibrated predictions, e.g. a 90% credible interval should contain the true outcome 90% of the times. However, typical definitions of calibration only require this to hold on average, and offer no guarantees on predictions made on individual samples. Thus, predictions can be systematically over or under confident on certain subgroups, leading to issues of fairness and potential vulnerabilities. We show that calibration for individual samples is possible in the regression setup if and only if the predictions are randomized, i.e. outputting randomized credible intervals. Randomization removes systematic bias by trading off bias with variance. We design a training objective to enforce individual calibration and use it to train randomized regression functions. The resulting models are more calibrated for arbitrarily chosen subgroups of the data, and can achieve higher utility in decision making against adversaries that exploit miscalibrated predictions.

Smaller, more accurate regression forests using tree alternating optimization

- Arman Zharmagambetov, Miguel Carreira-Perpinan
- abstract: Regression forests, based on ensemble approaches such as bagging or boosting, have long been recognized as the leading off-the-shelf method for regression. However, forests rely on a greedy top-down procedure such as CART to learn each tree. We extend a recent algorithm for learning classification trees, Tree Alternating Optimization (TAO), to the regression case, and use it with bagging to construct regression forests of oblique trees, having hyperplane splits at the decision nodes. In a wide range of datasets, we show that the resulting forests exceed the accuracy of state-of-the-art algorithms such as random forests, AdaBoost or gradient boosting, often considerably, while yielding forests that have usually fewer and shallower trees and hence fewer parameters and faster inference overall. This result has an immense practical impact and advocates for the power of optimization in ensemble learning.

Learning to Learn Kernels with Variational Random Features

- Xiantong Zhen, Haoliang Sun, Yingjun Du, Jun Xu, Yilong Yin, Ling Shao, Cees Snoek
- abstract: We introduce kernels with random Fourier features in the meta-learning framework for few-shot learning. We propose meta variational random features (MetaVRF) to learn adaptive kernels for the base-learner, which is developed in a latent variable model by treating the random feature basis as the latent variable. We formulate the optimization of MetaVRF as a variational inference problem by deriving an evidence lower bound under the meta-learning framework. To incorporate shared knowledge from related tasks, we propose a context inference of the posterior, which is established by an LSTM architecture. The LSTM-based inference network can effectively integrate the context information of previous tasks with task-specific information, generating informative and adaptive features. The learned MetaVRF can produce kernels of high representational power with a relatively low spectral sampling rate and also enables fast adaptation to new tasks. Experimental results on a variety of few-shot regression and classification tasks demonstrate that MetaVRF delivers much better, or at least competitive, performance compared to existing meta-learning alternatives.

Sharp Composition Bounds for Gaussian Differential Privacy via Edgeworth Expansion

- Qinqing Zheng, Jinshuo Dong, Qi Long, Weijie Su
- abstract: Datasets containing sensitive information are often sequentially analyzed by many algorithms and, accordingly, a fundamental question in differential privacy is concerned with how the overall privacy bound degrades under composition. To address this question, we introduce a family of analytical and sharp privacy bounds under composition using the Edgeworth expansion in the framework of the recently proposed f-differential privacy. In short, whereas the existing composition theorem, for example, relies on the central limit theorem, our new privacy bounds under composition gain improved tightness by leveraging the refined approximation accuracy of the Edgeworth expansion. Our approach is easy to implement and

computationally efficient for any number of compositions. The superiority of these new bounds is confirmed by an asymptotic error analysis and an application to quantifying the overall privacy guarantees of noisy stochastic gradient descent used in training private deep neural networks.

What Can Learned Intrinsic Rewards Capture?

- Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado Van Hasselt, David Silver, Satinder Singh
- abstract: The objective of a reinforcement learning agent is to behave so as to maximise the sum of a suitable scalar function of state: the reward. These rewards are typically given and immutable. In this paper, we instead consider the proposition that the reward function itself can be a good locus of learned knowledge. To investigate this, we propose a scalable meta-gradient framework for learning useful intrinsic reward functions across multiple lifetimes of experience. Through several proof-of-concept experiments, we show that it is feasible to learn and capture knowledge about long-term exploration and exploitation into a reward function. Furthermore, we show that unlike policy transfer methods that capture “how” the agent should behave, the learned reward functions can generalise to other kinds of agents and to changes in the dynamics of the environment by capturing “what” the agent should strive to do.

Error-Bounded Correction of Noisy Labels

- Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, Chao Chen
- abstract: To collect large scale annotated data, it is inevitable to introduce label noise, i.e., incorrect class labels. To be robust against label noise, many successful methods rely on the noisy classifiers (i.e., models trained on the noisy training data) to determine whether a label is trustworthy. However, it remains unknown why this heuristic works well in practice. In this paper, we provide the first theoretical explanation for these methods. We prove that the prediction of a noisy classifier can indeed be a good indicator of whether the label of a training data is clean. Based on the theoretical result, we propose a novel algorithm that corrects the labels based on the noisy classifier prediction. The corrected labels are consistent with the true Bayesian optimal classifier with high probability. We incorporate our label correction algorithm into the training of deep neural networks and train models that achieve superior testing performance on multiple public datasets.

Robust Graph Representation Learning via Neural Sparsification

- Cheng Zheng, Bo Zong, Wei Cheng, Dongjin Song, Jingchao Ni, Wenchao Yu, Haifeng Chen, Wei Wang
- abstract: Graph representation learning serves as the core of important prediction tasks, ranging from product recommendation to fraud detection. Real-life graphs usually have complex information in the local neighborhood, where each node is described by a rich set of features and connects to dozens or even hundreds of neighbors. Despite the success of neighborhood aggregation in graph neural networks, task-irrelevant information is mixed into nodes’ neighborhood, making learned models suffer from sub-optimal generalization performance. In this paper, we present NeuralSparse, a supervised graph sparsification technique that improves generalization power by learning to remove potentially task-irrelevant edges from input graphs. Our method takes both structural and non-structural information as input, utilizes deep neural networks to parameterize sparsification processes, and optimizes the parameters by feedback signals from downstream tasks. Under the NeuralSparse framework, supervised graph sparsification could seamlessly connect with existing graph neural networks for more robust performance. Experimental results on both benchmark and private datasets show that NeuralSparse can yield up to 7.2% improvement in testing accuracy when working with existing graph neural networks on node classification tasks.

Bisection-Based Pricing for Repeated Contextual Auctions against Strategic Buyer

- Anton Zhiyanov, Alexey Drutsa
- abstract: We are interested in learning algorithms that optimize revenue in repeated contextual posted-price auctions where a single seller faces a single strategic buyer. In our setting, the buyer maximizes his expected cumulative discounted surplus, and his valuation of a good is assumed to be a fixed function of a d-dimensional context (feature) vector. We introduce a novel deterministic learning algorithm that is based on ideas of the Bisection method and has strategic regret upper bound of $O(\log^2 T)$. Unlike previous works, our algorithm does not require any assumption on the distribution of context information, and the regret guarantee holds for any realization of feature vectors (adversarial upper bound). To construct our algorithm we non-trivially adopted techniques of integral geometry to act against buyer strategicness and improved the penalization trick to work in contextual auctions.

Best Arm Identification for Cascading Bandits in the Fixed Confidence Setting

- Zixin Zhong, Wang Chi Cheung, Vincent Tan
- abstract: We design and analyze CascadeBAI, an algorithm for finding the best set of K items, also called an arm, within the framework of cascading bandits. An upper bound on the time complexity of CascadeBAI is derived by overcoming a crucial analytical challenge, namely, that of probabilistically estimating the amount of available feedback at each step. To do so, we define a new class of random variables (r.v.’s) which we term as left-sided sub-Gaussian r.v.’s; this class is a relaxed version of the sub-Gaussian r.v.’s. This enables the application of a sufficiently tight Bernstein-type concentration inequality. We show, through the derivation of a lower bound on the time complexity, that the performance of CascadeBAI is optimal in some practical regimes. Finally, extensive numerical simulations corroborate the efficacy of CascadeBAI as well as the tightness of our upper bound on its time complexity.

Neural Contextual Bandits with UCB-based Exploration

- Dongruo Zhou, Lihong Li, Quanquan Gu
- abstract: We study the stochastic contextual bandit problem, where the reward is generated from an unknown function with additive noise. No assumption is made about the reward function other than boundedness. We propose a new algorithm, NeuralUCB, which leverages the representation power of deep neural networks and uses a neural network-based random feature mapping to construct an upper confidence bound (UCB) of reward for efficient exploration. We prove that, under standard assumptions, NeuralUCB achieves $\tilde O(\sqrt{T})$ regret, where T is the number of rounds. To the best of our knowledge, it is the first neural network-based contextual bandit algorithm with a near-optimal regret guarantee. We also show the algorithm is empirically competitive against representative baselines in a number of benchmarks.

MoNet3D: Towards Accurate Monocular 3D Object Localization in Real Time

- Xichuan Zhou, Yicong Peng, Chunqiao Long, Fengbo Ren, Cong Shi
- abstract: Monocular multi-object detection and localization in 3D space has been proven to be a challenging task. The MoNet3D algorithm is a novel and effective framework that can predict the 3D position of each object in a monocular image, and draw a 3D bounding box on each object. The MoNet3D method incorporates the prior knowledge of spatial geometric correlation of neighboring objects into the deep neural network training process, in order to improve the accuracy of 3D object localization. Experiments over the KITTI data set show that the accuracy of predicting the depth and horizontal coordinate of the object in 3D space can reach 96.25% and 94.74%, respectively. Meanwhile, the method can realize the real-time image processing capability of 27.85 FPS. Our code is publicly available at https://github.com/CQUlearningsystemgroup/YicongPeng

Nonparametric Score Estimators

- Yuhao Zhou, Jiaxin Shi, Jun Zhu
- abstract: Estimating the score, i.e., the gradient of log density function, from a set of samples generated by an unknown distribution is a fundamental task in inference and learning of probabilistic models that involve flexible yet intractable densities. Kernel estimators based on Stein's methods or score matching have shown promise, however their theoretical properties and relationships have not been fully-understood. We provide a unifying view of these estimators under the framework of regularized nonparametric regression. It allows us to analyse existing estimators and construct new ones with desirable properties by choosing different hypothesis spaces and regularizers. A unified convergence analysis is provided for such estimators. Finally, we propose score estimators based on iterative regularization that enjoy computational benefits from curl-free kernels and fast convergence.

Time-Consistent Self-Supervision for Semi-Supervised Learning

- Tianyi Zhou, Shengjie Wang, Jeff Bilmes
- abstract: Semi-supervised learning (SSL) leverages unlabeled data when training a model with insufficient labeled data. A common strategy for SSL is to enforce the consistency of model outputs between similar samples, e.g., neighbors or data augmentations of the same sample. However, model outputs can vary dramatically on unlabeled data over different training stages, e.g., when using large learning rates. This can introduce harmful noises and inconsistent objectives over time that may lead to concept drift and catastrophic forgetting. In this paper, we study the dynamics of neural net outputs in SSL and show that selecting and using first the unlabeled samples with more consistent outputs over the course of training (i.e., "time-consistency") can improve the final test accuracy and save computation. Under the time-consistent data selection, we design an SSL objective composed of two self-supervised losses, i.e., a consistency loss between a sample and its augmentation, and a contrastive loss encouraging different samples to have different outputs. Our approach achieves SOTA on several SSL benchmarks with much fewer computations.

Divide, Conquer, and Combine: a New Inference Strategy for Probabilistic Programs with Stochastic Support

- Yuan Zhou, Hongseok Yang, Yee Whye Teh, Tom Rainforth
- abstract: Universal probabilistic programming systems (PPSs) provide a powerful framework for specifying rich probabilistic models. They further attempt to automate the process of drawing inferences from these models, but doing this successfully is severely hampered by the wide range of non–standard models they can express. As a result, although one can specify complex models in a universal PPS, the provided inference engines often fall far short of what is required. In particular, we show that they produce surprisingly unsatisfactory performance for models where the support varies between executions, often doing no better than importance sampling from the prior. To address this, we introduce a new inference framework: Divide, Conquer, and Combine, which remains efficient for such models, and show how it can be implemented as an automated and generic PPS inference engine. We empirically demonstrate substantial performance improvements over existing approaches on three examples.

Go Wide, Then Narrow: Efficient Training of Deep Thin Networks

- Denny Zhou, Mao Ye, Chen Chen, Tianjian Meng, Mingxing Tan, Xiaodan Song, Quoc Le, Qiang Liu, Dale Schuurmans
- abstract: For deploying a deep learning model into production, it needs to be both accurate and compact to meet the latency and memory constraints. This usually results in a network that is deep (to ensure performance) and yet thin (to improve computational efficiency). In this paper, we propose an efficient method to train a deep thin network with a theoretic guarantee. Our method is motivated by model compression. It consists of three stages. First, we sufficiently widen the deep thin network and train it until convergence. Then, we use this well-trained deep wide network to warm up (or initialize) the original deep thin network. This is achieved by layerwise imitation, that is, forcing the thin network to mimic the intermediate outputs of the wide network from layer to layer. Finally, we further fine tune this already well-initialized deep thin network. The theoretical guarantee is established by using the neural mean field analysis. It demonstrates the advantage of our layerwise imitation approach over backpropagation. We also conduct large-scale empirical experiments to validate the proposed method. By training with our method, ResNet50 can outperform ResNet101, and BERT base can be comparable with BERT large, when ResNet101 and BERT large are trained under the standard training procedures as in the literature.

Hybrid Stochastic-Deterministic Minibatch Proximal Gradient: Less-Than-Single-Pass Optimization with Nearly Optimal Generalization

- Pan Zhou, Xiao-Tong Yuan
- abstract: Stochastic variance-reduced gradient (SVRG) algorithms have been shown to work favorably in solving large-scale learning problems. Despite the remarkable success, the stochastic gradient complexity of SVRG-type algorithms usually scales linearly with data size and thus could still be expensive for huge data. To address this deficiency, we propose a hybrid stochastic-deterministic minibatch proximal gradient (\HSDAN) algorithm for strongly-convex problems that enjoys provably improved data-size-independent complexity guarantees. More precisely, for quadratic loss $F(\wm)$ of n components, we prove that \HSDAN can attain an ϵ-optimization-error $\EE[F(\wm)-F(\wms)] \leq \epsilon$ within $\mathcal{O}\Big(\frac{\kappa^{1.5}\epsilon^{0.75} \log^{1.5}(\frac{1}{\epsilon}) + 1}{\epsilon} \wedge \Big(\kappa \sqrt{n} \log^{1.5}\big(\frac{1}{\epsilon}\big) + n \log \big(\frac{1}{\epsilon}\big) \Big) \Big)$ stochastic gradient evaluations, where κ is condition number. For generic strongly convex loss functions, we prove a nearly identical complexity bound though at the cost of slightly increased logarithmic factors. For large-scale learning problems, our complexity bounds are superior to those of the prior state-of-the-art SVRG algorithms with or without dependence on data size. Particularly, in the case of $\epsilon!=!\mathcal{O}\big(1/\sqrt{n}\big)$ which is at the order of intrinsic excess error bound of a learning model and thus sufficient for generalization, the stochastic gradient complexity bounds of \HSDAN for quadratic and generic loss functions are respectively $\mathcal{O}(n^{0.875}\log^{1.5}(n))$ and $\mathcal{O}(n^{0.875}\log^{2.25}(n))$, which to our best knowledge, for the first time achieve optimal generalization in less than a single pass over data. Extensive numerical results demonstrate the computational advantages of our algorithm over the prior ones.

Robust Outlier Arm Identification

- Yinglun Zhu, Sumeet Katariya, Robert Nowak
- abstract: We study the problem of Robust Outlier Arm Identification (ROAI), where the goal is to identify arms whose expected rewards deviate substantially from the majority, by adaptively sampling from their reward distributions. We compute the outlier threshold using the median and median absolute deviation of the expected rewards. This is a robust choice for the threshold compared to using the mean and standard deviation, since it can identify outlier arms even in the presence of extreme outlier values. Our setting is different from existing pure exploration problems where the threshold is pre-specified as a given value or rank. This is useful in applications where the goal is to identify the set of promising items but the cardinality of this set is unknown, such as finding promising drugs for a new disease or identifying items favored by a population. We propose two δ-PAC algorithms for ROAI, which includes the first UCB-style algorithm for outlier detection, and derive upper bounds on their sample complexity. We also prove a matching, up to logarithmic factors, worst case lower bound for the problem, indicating that our upper bounds are generally unimprovable. Experimental results show that our algorithms are both robust and about 5x sample efficient compared to state-of-the-art.

Variance Reduction and Quasi-Newton for Particle-Based Variational Inference

- Michael Zhu, Chang Liu, Jun Zhu
- abstract: Particle-based Variational Inference methods (ParVIs), like Stein Variational Gradient Descent, are nonparametric variational inference methods that optimize a set of particles to best approximate a target distribution. ParVIs have been proposed as efficient approximate inference algorithms and as potential alternatives to MCMC methods. However, to our knowledge, the quality of the posterior approximation of particles from ParVIs has not been examined before for large-scale Bayesian inference problems. We conduct this analysis and evaluate the sample quality of particles produced by ParVIs,

and we find that existing ParVI approaches using stochastic gradients converge insufficiently fast under sample quality metrics. We propose a novel variance reduction and quasi-Newton preconditioning framework for ParVIs, by leveraging the Riemannian structure of the Wasserstein space and advanced Riemannian optimization algorithms. Experimental results demonstrate the accelerated convergence of variance reduction and quasi-Newton methods for ParVIs for accurate posterior inference in large-scale and ill-conditioned problems.

Causal Effect Estimation and Optimal Dose Suggestions in Mobile Health

- Liangyu Zhu, Wenbin Lu, Rui Song
- abstract: In this article, we propose novel structural nested models to estimate causal effects of continuous treatments based on mobile health data. To find the treatment regime which optimizes the short-term outcomes for the patients, we define the weighted lag K advantage. The optimal treatment regime is then defined to be the one which maximizes this advantage. This method imposes minimal assumptions on the data generating process. Statistical inference can also be provided for the estimated parameters. Simulation studies and an application to the Ohio type 1 diabetes dataset show that our method could provide meaningful insights for dose suggestions with mobile health data.

Thompson Sampling Algorithms for Mean-Variance Bandits

- Qiuyu Zhu, Vincent Tan
- abstract: The multi-armed bandit (MAB) problem is a classical learning task that exemplifies the exploration-exploitation tradeoff. However, standard formulations do not take into account risk. In online decision making systems, risk is a primary concern. In this regard, the mean-variance risk measure is one of the most common objective functions. Existing algorithms for mean-variance optimization in the context of MAB problems have unrealistic assumptions on the reward distributions. We develop Thompson Sampling-style algorithms for mean-variance MAB and provide comprehensive regret analyses for Gaussian and Bernoulli bandits with fewer assumptions. Our algorithms achieve the best known regret bounds for mean-variance MABs and also attain the information-theoretic bounds in some parameter regimes. Empirical simulations show that our algorithms significantly outperform existing LCB-based algorithms for all risk tolerances.

Learning Adversarially Robust Representations via Worst-Case Mutual Information Maximization

- Sicheng Zhu, Xiao Zhang, David Evans
- abstract: Training machine learning models that are robust against adversarial inputs poses seemingly insurmountable challenges. To better understand adversarial robustness, we consider the underlying problem of learning robust representations. We develop a notion of representation vulnerability that captures the maximum change of mutual information between the input and output distributions, under the worst-case input perturbation. Then, we prove a theorem that establishes a lower bound on the minimum adversarial risk that can be achieved for any downstream classifier based on its representation vulnerability. We propose an unsupervised learning method for obtaining intrinsically robust representations by maximizing the worst-case mutual information between the input and output distributions. Experiments on downstream classification tasks support the robustness of the representations found using unsupervised learning with our training principle.

Linear Convergence of Randomized Primal-Dual Coordinate Method for Large-scale Linear Constrained Convex Programming

- Daoli Zhu, Lei Zhao
- abstract: Linear constrained convex programming has many practical applications, including support vector machine and machine learning portfolio problems. We propose the randomized primal-dual coordinate (RPDC) method, a randomized coordinate extension of the first-order primal-dual method by Cohen and Zhu, 1984 and Zhao and Zhu, 2019, to solve linear constrained convex programming. We randomly choose a block of variables based on a uniform distribution, linearize, and apply a Bregman-like function (core function) to the selected block to obtain simple parallel primal-dual decomposition. We then establish almost surely convergence and expected O(1/t) convergence rate, and expected linear convergence under global strong metric subregularity. Finally, we discuss implementation details for the randomized primal-dual coordinate approach and present numerical experiments on support vector machine and machine learning portfolio problems to verify the linear convergence.

When Demands Evolve Larger and Noisier: Learning and Earning in a Growing Environment

- Feng Zhu, Zeyu Zheng
- abstract: We consider a single-product dynamic pricing problem under a specific non-stationary setting, where the underlying demand process grows over time in expectation and also possibly in the level of random fluctuation. The decision maker sequentially sets price in each time period and learns the unknown demand model, with the goal of maximizing expected cumulative revenue over a time horizon T. We prove matching upper and lower bounds on regret and provide near-optimal pricing policies, showing how the growth rate of random fluctuation over time affects the best achievable regret order and the near-optimal policy design. In the analysis, we show that whether the seller knows the length of time horizon T in advance or not surprisingly render different optimal regret orders. We then extend the demand model such that the optimal price may vary with time and present a novel and near-optimal policy for the extended model. Finally, we consider an analogous non-stationary setting in the canonical multi-armed bandit problem, and points out that knowing or not knowing the length of time horizon T render the same optimal regret order, in contrast to the non-stationary dynamic pricing problem.

Adaptive Checkpoint Adjoint Method for Gradient Estimation in Neural ODE

- Juntang Zhuang, Nicha Dvornek, Xiaoxiao Li, Sekhar Tatikonda, Xenophon Papademetris, James Duncan
- abstract: The empirical performance of neural ordinary differential equations (NODEs) is significantly inferior to discrete-layer models on benchmark tasks (e.g. image classification). We demonstrate an explanation is the inaccuracy of existing gradient estimation methods: the adjoint method has numerical errors in reverse-mode integration; the naive method suffers from a redundantly deep computation graph. We propose the Adaptive Checkpoint Adjoint (ACA) method: ACA applies a trajectory checkpoint strategy which records the forward- mode trajectory as the reverse-mode trajectory to guarantee accuracy; ACA deletes redundant components for shallow computation graphs; and ACA supports adaptive solvers. On image classification tasks, compared with the adjoint and naive method, ACA achieves half the error rate in half the training time; NODE trained with ACA outperforms ResNet in both accuracy and test-retest reliability. On time-series modeling, ACA outperforms competing methods. Furthermore, NODE with ACA can incorporate physical knowledge to achieve better accuracy.

Learning Optimal Tree Models under Beam Search

- Jingwei Zhuo, Ziru Xu, Wei Dai, Han Zhu, Han Li, Jian Xu, Kun Gai
- abstract: Retrieving relevant targets from an extremely large target set under computational limits is a common challenge for information retrieval and recommendation systems. Tree models, which formulate targets as leaves of a tree with trainable node-wise scorers, have attracted a lot of interests in tackling this challenge due to their logarithmic computational complexity in both training and testing. Tree-based deep models (TDMs) and probabilistic label trees (PLTs) are two representative kinds of them. Though achieving many practical successes, existing tree models suffer from the training-testing discrepancy, where the retrieval performance deterioration caused by beam search in testing is not considered in training. This leads to an intrinsic gap between the most relevant targets and those retrieved by beam search with even the optimally trained node-wise scorers. We take a first step towards

understanding and analyzing this problem theoretically, and develop the concept of Bayes optimality under beam search and calibration under beam search as general analyzing tools for this purpose. Moreover, to eliminate the discrepancy, we propose a novel algorithm for learning optimal tree models under beam search. Experiments on both synthetic and real data verify the rationality of our theoretical analysis and demonstrate the superiority of our algorithm compared to state-of-the-art methods.

Laplacian Regularized Few-Shot Learning

- Imtiaz Ziko, Jose Dolz, Eric Granger, Ismail Ben Ayed
- abstract: We propose a transductive Laplacian-regularized inference for few-shot tasks. Given any feature embedding learned from the base classes, we minimize a quadratic binary-assignment function containing two terms: (1) a unary term assigning query samples to the nearest class prototype, and (2) a pairwise Laplacian term encouraging nearby query samples to have consistent label assignments. Our transductive inference does not re-train the base model, and can be viewed as a graph clustering of the query set, subject to supervision constraints from the support set. We derive a computationally efficient bound optimizer of a relaxation of our function, which computes independent (parallel) updates for each query sample, while guaranteeing convergence. Following a simple cross-entropy training on the base classes, and without complex meta-learning strategies, we conducted comprehensive experiments over five few-shot learning benchmarks. Our LaplacianShot consistently outperforms state-of-the-art methods by significant margins across different models, settings, and data sets. Furthermore, our transductive inference is very fast, with computational times that are close to inductive inference, and can be used for large-scale few-shot tasks.

Influenza Forecasting Framework based on Gaussian Processes

- Christoph Zimmer, Reza Yaesoubi
- abstract: The seasonal epidemic of influenza costs thousands of lives each year in the US. While influenza epidemics occur every year, timing and size of the epidemic vary strongly from season to season. This complicates the public health efforts to adequately respond to such epidemics. Forecasting techniques to predict the development of seasonal epidemics such as influenza, are of great help to public health decision making. Therefore, the US Center for Disease Control and Prevention (CDC) has initiated a yearly challenge to forecast influenza-like illness. Here, we propose a new framework based on Gaussian process (GP) for seasonal epidemics forecasting and demonstrate its capability on the CDC reference data on influenza like illness: our framework leads to accurate forecasts with small but reliable uncertainty estimation. We compare our framework to several state of the art benchmarks and show competitive performance. We, therefore, believe that our GP based framework for seasonal epidemics forecasting will play a key role for future influenza forecasting and, lead to further research in the area.

A general recurrent state space framework for modeling neural dynamics during decision-making

- David Zoltowski, Jonathan Pillow, Scott Linderman
- abstract: An open question in systems and computational neuroscience is how neural circuits accumulate evidence towards a decision. Fitting models of decision-making theory to neural activity helps answer this question, but current approaches limit the number of these models that we can fit to neural data. Here we propose a general framework for modeling neural activity during decision-making. The framework includes the canonical drift-diffusion model and enables extensions such as multi-dimensional accumulators, variable and collapsing boundaries, and discrete jumps. Our framework is based on constraining the parameters of recurrent state space models, for which we introduce a scalable variational Laplace EM inference algorithm. We applied the modeling approach to spiking responses recorded from monkey parietal cortex during two decision-making tasks. We found that a two-dimensional accumulator better captured the responses of a set of parietal neurons than a single accumulator model, and we identified a variable lower boundary in the responses of a parietal neuron during a random dot motion task. We expect this framework will be useful for modeling neural dynamics in a variety of decision-making settings.

Transformer Hawkes Process

- Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, Hongyuan Zha
- abstract: Modern data acquisition routinely produce massive amounts of event sequence data in various domains, such as social media, healthcare, and financial markets. These data often exhibit complicated short-term and long-term temporal dependencies. However, most of the existing recurrent neural network based point process models fail to capture such dependencies, and yield unreliable prediction performance. To address this issue, we propose a Transformer Hawkes Process (THP) model, which leverages the self-attention mechanism to capture long-term dependencies and meanwhile enjoys computational efficiency. Numerical experiments on various datasets show that THP outperforms existing models in terms of both likelihood and event prediction accuracy by a notable margin. Moreover, THP is quite general and can incorporate additional structural knowledge. We provide a concrete example, where THP achieves improved prediction performance for learning multiple point processes when incorporating their relational information.