Classification vs regression in overparameterized regimes: Does the loss function matter?

We compare classification and regression tasks in an overparameterized linear model with Gaussian features. On the one hand, we show that with sufficient overparameterization all training points are support vectors: solutions obtained by least-squares minimum-norm interpolation, typically used for regression, are identical to those produced by the hardmargin support vector machine (SVM) that minimizes the hinge loss, typically used for training classifiers. On the other hand, we show that there exist regimes where these interpolating solutions generalize well when evaluated by the 0-1 test loss function, but do not generalize if evaluated by the square loss function, i.e. they approach the null risk.

A general linear-time inference method for Gaussian Processes on one dimension

Gaussian Processes (GPs) provide powerful probabilistic frameworks for interpolation, forecasting, and smoothing, but have been hampered by computational scaling issues. Here we investigate data sampled on one dimension (e.g., a scalar or vector time series sampled at arbitrarily-spaced intervals), for which state-space models are popular due to their linearly-scaling computational costs. It has long been conjectured that state-space models are general, able to approximate any one-dimensional GP. We provide the first general proof of this conjecture, showing that any stationary GP on one dimension with vector-valued observations governed by a Lebesgue-integrable continuous kernel can be approximated to any desired precision using a specifically-chosen state-space model: the Latent Exponentially Generated (LEG) family. This new family offers several advantages compared to the general state-space model: it is always stable (no unbounded growth), the covariance can be computed in closed form, and its parameter space is unconstrained (allowing straightforward estimation via gradient descent). The theorem's proof also draws connections to Spectral Mixture Kernels, providing insight about this popular family of kernels. We develop parallelized algorithms for performing inference and learning in the LEG model, test the algorithm on real and synthetic data, and demonstrate scaling to datasets with billions of samples.

Distributed Bayesian Varying Coefficient Modeling Using a Gaussian Process Prior

Varying coefficient models (VCMs) are widely used for estimating nonlinear regression functions for functional data. Their Bayesian variants using Gaussian process priors on the functional coefficients, however, have received limited attention in massive data applications, mainly due to the prohibitively slow posterior computations using Markov chain Monte Carlo (MCMC) algorithms. We address this problem using a divide-and-conquer Bayesian approach. We first create a large number of data subsamples with much smaller sizes. Then, we formulate the VCM as a linear mixed-effects model and develop a data augmentation algorithm for obtaining MCMC draws on all the subsets in parallel. Finally, we aggregate the MCMC-based estimates of subset posteriors into a single Aggregated Monte Carlo (AMC) posterior, which is used as a computationally efficient alternative to the true posterior distribution. Theoretically, we derive minimax optimal posterior convergence rates for the AMC posteriors of both the varying coefficients and the mean regression function. We provide quantification on the orders of subset sample sizes and the number of subsets. The empirical results show that the combination schemes that satisfy our theoretical assumptions, including the AMC posterior, have better estimation performance than their main competitors across diverse simulations and in a real data analysis.

Oblivious Data for Fairness with Kernels

We investigate the problem of algorithmic fairness in the case where sensitive and non-sensitive features are available and one aims to generate new, ‘oblivious’, features that closely approximate the non-sensitive features, and are only minimally dependent on the sensitive ones. We study this question in the context of kernel methods. We analyze a relaxed version of the Maximum Mean Discrepancy criterion which does not guarantee full independence but makes the optimization problem tractable. We derive a closed-form solution for this relaxed optimization problem and complement the result with a study of the dependencies between the newly generated features and the sensitive ones. Our key ingredient for generating such oblivious features is a Hilbert-space-valued conditional expectation, which needs to be estimated from data. We propose a plug-in approach and demonstrate how the estimation errors can be controlled. While our techniques help reduce the bias, we would like to point out that no post-processing of any dataset could possibly serve as an alternative to well-designed experiments.

Structure Learning for Directed Trees

Knowing the causal structure of a system is of fundamental interest in many areas of science and can aid the design of prediction algorithms that work well under manipulations to the system. The causal structure becomes identifiable from the observational distribution under certain restrictions. To learn the structure from data, score-based methods evaluate different graphs according to the quality of their fits. However, for large, continuous, and nonlinear models, these rely on heuristic optimization approaches with no general guarantees of recovering the true causal structure. In this paper, we consider structure learning of directed trees. We propose a fast and scalable method based on Chu–Liu–Edmonds' algorithm we call causal additive trees (CAT). For the case of Gaussian errors, we prove consistency in an asymptotic regime with a vanishing identifiability gap. We also introduce two methods for testing substructure hypotheses with asymptotic family-wise error rate control that is valid post-selection and in unidentified settings. Furthermore, we study the identifiability gap, which quantifies how much better the true causal model fits the observational distribution, and prove that it is lower bounded by local properties of the causal model. Simulation studies demonstrate the favorable performance of CAT compared to competing structure learning methods.

Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits

We introduce GLR-klUCB, a novel algorithm for the piecewise i.i.d. non-stationary bandit problem with bounded rewards. This algorithm combines an efficient bandit algorithm, klUCB, with an efficient, parameter-free, change-point detector, the Bernoulli Generalized Likelihood Ratio Test, for which we provide new theoretical guarantees of independent interest. Unlike previous nonstationary bandit algorithms using a change-point detector, GLR-klUCB does not need to be calibrated based on prior knowledge on the arms' means. We prove that this algorithm can attain a O(p T AYT ln(T)) regret in T rounds on some "easy" instances in which there is sufficient delay between two change-points, where A is the number of arms and YT the number of change-points, without prior knowledge of YT . In contrast with recently proposed algorithms that are agnostic to YT , we perform a numerical study showing that GLR-klUCB is also very efficient in practice, beyond easy instances.

How Well Generative Adversarial Networks Learn Distributions

This paper studies the rates of convergence for learning distributions implicitly with the adversarial framework and Generative Adversarial Networks (GANs), which subsume Wasserstein, Sobolev, MMD GAN, and Generalized/Simulated Method of Moments (GMM/SMM) as special cases. We study a wide range of parametric and nonparametric target distributions under a host of objective evaluation metrics. We investigate how to obtain valid statistical guarantees for GANs through the lens of regularization. On the nonparametric end, we derive the optimal minimax rates for distribution estimation under the adversarial framework. On the parametric end, we establish a theory for general neural network classes (including deep leaky ReLU networks) that characterizes the interplay on the choice of generator and discriminator pair. We discover and isolate a new notion of regularization, called the generator-discriminator-pair regularization, that sheds light on the advantage of GANs compared to classical parametric and nonparametric approaches for explicit distribution estimation. We develop novel oracle inequalities as the main technical tools for analyzing GANs, which are of independent interest.

Regularized spectral methods for clustering signed networks

We study the problem of k-way clustering in signed graphs. Considerable attention in recent years has been devoted to analyzing and modeling signed graphs, where the affinity measure between nodes takes either positive or negative values. Recently, Cucuringu et al. (2019) proposed a spectral method, namely SPONGE (Signed Positive over Negative Generalized Eigenproblem), which casts the clustering task as a generalized eigenvalue problem optimizing a suitably defined objective function. This approach is motivated by social balance theory, where the clustering task aims to decompose a given network into disjoint groups, such that individuals within the same group are connected by as many positive edges as possible, while individuals from different groups are mainly connected by negative edges. Through extensive numerical experiments, SPONGE was shown to achieve state-of-the-art empirical performance. On the theoretical front, Cucuringu et al. (2019) analyzed SPONGE, as well as the popular Signed Laplacian based spectral method under the setting of a Signed Stochastic Block Model, for $k = 2$ equal-sized clusters, in the regime where the graph is moderately dense. In this work, we build on the results in Cucuringu et al. (2019) on two fronts for the normalized versions of SPONGE and the Signed Laplacian. Firstly, for both algorithms, we extend the theoretical analysis in Cucuringu et al. (2019) to the general setting of $k \geq 2$ unequal-sized clusters in the moderately dense regime. Secondly, we introduce regularized versions of both methods to handle sparse graphs – a regime where standard spectral methods are known to underperform – and provide theoretical guarantees under the same setting of a Signed Stochastic Block Model. To the best of our knowledge, regularized spectral methods have so far not been considered in the setting of clustering signed graphs. We complement our theoretical results with an extensive set of numerical experiments on synthetic data, and three real world data sets standard in the signed networks literature.

Deep Learning in Target Space

Deep learning uses neural networks which are parameterised by their weights. The neural networks are usually trained by tuning the weights to directly minimise a given loss function. In this paper we propose to re-parameterise the weights into targets for the firing strengths of the individual nodes in the network. Given a set of targets, it is possible to calculate the weights which make the firing strengths best meet those targets. It is argued that using targets for training addresses the problem of exploding gradients, by a process which we call cascade untangling, and makes the loss-function surface smoother to traverse, and so leads to easier, faster training, and also potentially better generalisation, of the neural network. It also allows for easier learning of deeper and recurrent network structures. The necessary conversion of targets to weights comes at an extra computational expense, which is in many cases manageable. Learning in target space can be combined with existing neural-network optimisers, for extra gain. Experimental results show the speed of using target space, and examples of improved generalisation, for fully-connected networks and convolutional networks, and the ability to recall and process long time sequences and perform naturallanguage processing with recurrent networks.

The Importance of Being Correlated: Implications of Dependence in Joint Spectral Inference across Multiple Networks

Spectral inference on multiple networks is a rapidly-developing subfield of graph statistics. Recent work has demonstrated that joint, or simultaneous, spectral embedding of multiple independent networks can deliver more accurate estimation than individual spectral decompositions of those same networks. Such inference procedures typically rely heavily on independence assumptions across the multiple network realizations, and even in this case, little attention has been paid to the induced network correlation that can be a consequence of such joint embeddings. In this paper, we present a generalized omnibus embedding methodology and we provide a detailed analysis of this embedding across both independent and correlated networks, the latter of which significantly extends the reach of such procedures, and we describe how this omnibus embedding can itself induce correlation. This leads us to distinguish between inherent correlation—that is, the correlation that arises naturally in multisample network data—and induced correlation, which is an artifice of the joint embedding methodology. We show that the generalized omnibus embedding procedure is flexible and robust, and we prove both consistency and a central limit theorem for the embedded points. We examine how induced and inherent correlation can impact inference for network time series data, and we provide network analogues of classical questions such as the effective sample size for more generally correlated data. Further, we show how an appropriately calibrated generalized omnibus embedding can detect changes in real biological networks that previous embedding procedures could not discern, confirming that the effect of inherent and induced correlation can be subtle and transformative. By allowing for and deconstructing both forms of correlation, our methodology widens the scope of spectral techniques for network inference, with import in theory and practice

CAT: Compression-Aware Training for Bandwidth Reduction

One major obstacle hindering the ubiquitous use of CNNs for inference is their relatively high memory bandwidth requirements, which can be the primary energy consumer and throughput bottleneck in hardware accelerators. Inspired by quantization-aware training approaches, we propose a compression-aware training (CAT) method that involves training the model to allow better compression of weights and feature maps during neural network deployment. Our method trains the model to achieve low-entropy feature maps, enabling efficient compression at inference time using classical transform coding methods. CAT significantly improves the state-of-the-art results reported for quantization evaluated on various vision and NLP tasks, such as image classification (ImageNet), image detection (Pascal VOC), sentiment analysis (CoLa) , and textual entailment (MNLI). For example, on ResNet-18, we achieve near baseline ImageNet accuracy with an average representation of only 1.5 bits per value with 5-bit quantization. Moreover, we show that entropy reduction of weights and activations can be applied together, further improving bandwidth reduction. Reference implementation is available.

Are All Layers Created Equal?

Understanding deep neural networks is a major research objective with notable experimental and theoretical attention in recent years. The practical success of excessively large networks underscores the need for better theoretical analyses and justifications. In this paper we focus on layer-wise functional structure and behavior in overparameterized deep models. To do so, we study empirically the layers’ robustness to post-training re-initialization and re-randomization of the parameters. We provide experimental results which give evidence for the heterogeneity of layers. Morally, layers of large deep neural networks can be categorized as either “robust” or “critical”. Resetting the robust layers to their initial values does not result in adverse decline in performance. In many cases, robust layers hardly change throughout training. In contrast, re-initializing critical layers vastly degrades the performance of the network with test error essentially dropping to random guesses. Our study provides further evidence that mere parameter counting or norm calculations are too coarse in studying generalization of deep models, and “flatness” and robustness analysis of trained models need to be examined while taking into account the respective network architectures.