MLSS 2020 note

Thanh Tang Nguyen *

July 10, 2020

Contents

*Email: nguyent2792@gmail.com; Link: https://thanhnguyentang.github.io/blogs/mlss20.pdf

Forewords

I took this minimal note while attending the Machine Learning Summer School (MLSS) 2020 (http://mlss.tuebingen.mpg.de/2020/) for two weeks (June 28 - July 11, 2020). The goal of this note is to capture most salient ideas and key concepts in each of the MLSS20 lectures, and to keep relevant resources in place for future reference. Thus, I omitted most of the technical details which could be referred to in the talk's slides or Youtube link provided at the end of each section. Note that the note presentation is not optimized and I might have also missed some interesting points from the talks (but hopefully not many). Finally notice that the kernel method section is particularly short (because I studied Arthur's lectures before the MLSS, thus very familiar with the topic and skipped many details in this note).

1 Symbolic, Statistical and Causal AI by Bernhard Schölkopf

cybernetics 40-50s

Wiener: information processing rather than energy in animals and machines (paradiam shift),

John von Neumann, Alan Turing, Claude Shannon

Percepttron limitations by Rosenblatt

Perceptron convergence thereom by Novikoff 1962.

- Symbolic AI: Dartmouth summer school 1956.

- Intelligence = manipulate discrete symbols (John McCarthy, Allen newell, herb simon, Minsky)

- Minsky Papert recall 1988/89

Minsky and Papert 1969: parity problem as an argument for limitation of perceptron.

The end of perceptrons (1996) –> symbolic AI gained popularity

Optimistic predictions of symbolic AI (Herb Simon 1957) but not correct but gave birth to computer science.

Moravec's paradox 1980s: Machines will be capable within 20 years of doing any work a man can do.

The return of neural nets:

- Symbolic AI: did poorly at speech and vision
- Boltzmann (Hinton)
- Back-propropagation mid 1980s
- Minsky & Papert 1988 Perceptrons 2nd ed.: not impressed with backpropagation.
- PAC (VC 1968-1982)
- Ising model (82)
- Expert systems, knowledge repres (Pearl 88)
- first UAI (85)
- first Neurips (87)
- Probabilistic foundsations (90s): MacKay, Neal, Jordan, Hinton, Bishop
- SVM, kernels (90)

Classic AI: rules by humans

ML: rules by learning.

Big data (06): optimality in the limit - i.i.d. data.

DQN Nature paper (2015) AI probems becomes large-scale pattern recognition.

Human-level object recognition: ML uses correlation rather than causality.

Adversarial vulnerability ("Intriguing properties of neural networks" '13)

Difference btw dependence and causation.

> Reichenbach's common cause principle: if X and Y are dependent, there is Z causally influencing both and X and Y are independent given Z.

Industrial revolution (energy related) -> digital revolution (information related): we human special in information processing, not so pround in energy processing.

Summary:

- 30: prob theory, stats, cuber,
- 60: Symbolic AI: birth of CS, XOR problem, Moravec's paradox 80
- 90: Statistical AI.

lecture: https://youtu.be/8staJlMbAig

2 Causality by Bernhard Schölkopf and Stefan Bauer

2.1 Day 1 by Bernhard Schölkopf

Roadmap:

- structural causal models
- independent mechanisms and disentanged factorizations
- do-calculus
- confounding
- causal discovery: 2-variable
-cause M:: 2-variable case
- algorithmic SCM and the narrow of time
- Confounder modelling in exoplanet discovery
...

$$(\forall f, g \in F, cov(f(X), g(Y)) = 0) \implies X \perp Y.$$

intervention on a. $p(t|a) \perp p(a)$ We expect $p(t|a)$ is invariant across different domains.

Independent mechanism: the conditional distributions of each varibale given its causes does not inform or influence other conditional distributions.

Reichenbach's common cause principle.

Structural Causal Model (Pearl):

- directed acyclic graph G with vertices X_i.
- vertices = observables
- arrows = direc causation
- $X_i = f_i(PA_i, U_i)$
- U_i: independent, unexplained r.v. (noise) -> causal sufficiency
- [graph here]

Entailed distribution: $X_i = f_i(PA_i, U_i$, a joint distribution of X_i is the observational distribution.

Questions:

- What we can say about it?
- Recover G from p?

MArkov conditions:

- existence of SCM
- local causal markov condition
- global causal markov condition: "d-seperation"
- Factorization $p(X_1, ..., X_n) = \prod_i p(X_i|PA_i)$ where $P(X_i|PA_i)$ causal conditional (causal Markov kernel).

Markov condition: Conditional indepdendent in G -> conditional indepdence in p. The revserse direction is "faithfulness".

> Intervention = replace $X_i = f_i(PA_i, U_i)$ by another assignment.

The entailed distribution is called interventional distribution. Some special cases: domain shift distribution, covariate shift distribution. General intervention: change $p(X_i|PA_i)$.

> Disentangled factorization: independent factorization $p(X_1, ..., X_n) = \prod_i p(X_i|PA_i)$ + mechanisms (i.e., causal conditionals) $p(X_i|PA_i)$ are independent.

> Causal shift hypothesis: a change in a distribution <- a sparse change in mechanisms.

Entangled factorization: $p(X_1, ..., X_n) = \prod p(X_i|X_{i+1}, ..., X_n)$; here changes are not local (changing 1 term affects other terms).

Counterfactuals: missing data problem (?)

> Does it make sense to talk about causality or stats without mentioning time?

From ODE to SCM: the deterministic case (UAI 16)

[table between causal and statistc and ode here Table 1]

> Does it make sense to talk about causality or stats without mentioning time?

A Modeling Taxonomy

	statistical model	causal model	differential equation model
i.i.d. prediction, pattern recognition, *"generalization"*	y	y	y
Predict under shift & intervention, *"horizontal generalization"*	n	y	y
Provide physical insight, understand predictions	n	(y)	y
Think/Reason, *"act in an imagined space"* (K. Lorenz)	n	?	?
Learn from data	y	(y)	n

Figure 1: Modeling taxonomy

Pearl's do calculus:

Goal of cuasality: infer the effect of interventions

p(y|x) is different from p(y |do x).

Controling for confounding/adjustment formula (it explains why p(y|x) is different from p(y |do x)) - Simpson's paradox.

Mediation analysis.

2-variable case: determining which is cause and which is effect.

Causal inference model: find causal direction.

Infer deterministic causal relations.

Benchmark dataset for cause-effect.

> Covariate shift and semi-supervised learning
> - Covariate shift: p(X) change btw train and test
> - SSL: improve estimate by more data from p(X) (high-order SSL UAI 20).

Algorithimc SCM.

Dynamic narrow of time paper.

Causal representation learning problem.

Causal mechanisms in ML.

2.2 Day 2 by Stefan Bauer

Chocolate vs Nobel prize example

Causal models as possets of distribution

Main questions today: what to do when do not know the graph? Connections with ML?

Key problem: many SCMs generate same graph

Assumptions for causal discovery:

- faithfulness
- Independent mechanism
- Additive noise.
- Linear non-Gaussian models

These assumptions: more data cannot improve inference (?)

Causal structure learning

- computational infeasible for large G, local minima
- indepdendent testing

Indepdendent component analysis

LiNGAM: Linear non-Gaussian acyclic models for causal discovery.

Structure learning: time series.

Time series and Granger causality

Confounded learning:

- intervention invariance

SCMs for ODEs/SDEs: Recover the network structure from data

- Causal approach: ultilize causal structure (invariance across environments)
- Measure invariance of an ODE: OOD generalization

A causal perspective on deep representation learning

- disentangled representation - disentangled generative factors

> Challenging common assumptions in the UL of disentangled repr (ICML 19)
> - Theory: for arbitrary data ,UL of DR is impossible - Random sees and hyperparameters seem matter mroe than the model
> - **inductive biases** is a key role
> - UL still require access to labels for model selection
> - Figure 2

Fairness and disentangled representation

Disentangled representation in real-world environments "Is Independence all you need? Pn the generalization of representation learned from correlated data" arxiv 2006.07886

Structural causal autoencoders (with structural decoder)

Insights:

- Structuring architecture of encoder-decoder -> can discover the underlying mechanism

Outlook: Causal world models

- learn a multi-task/multi-environment model
- reusable components across tasks/environments ("Learning independent mechanisms" ICML '18, "Recurrent independent mechanisms" by Anirudh Goyal et al. '19)

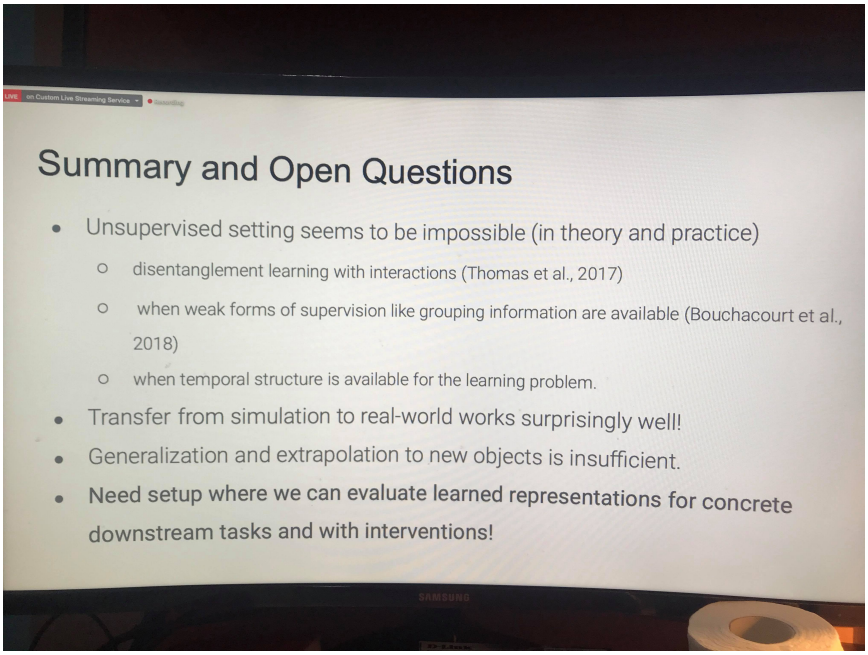


Figure 2: UL challenge

Summary:
- causal inference: requires assumptions
- Key problems: scale and computational efficiency; benchmark and evaluatons, data quantity; many applications variable are not directly observed
- Connection to deep learning (e.g.,) [biases-invariance-generalization.github.io (ICML '20)]
- Figure 3

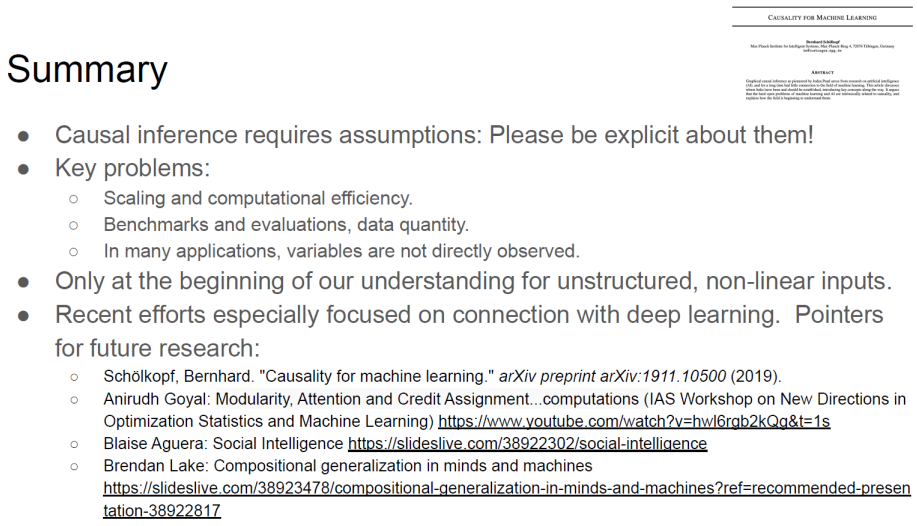


Figure 3: Summary of causality

[Open internship positions at MPI]

2.3 Resources

Day 1:
- lecture: https://youtu.be/btmJtThWmhA
- slides: shorturl.at/cxCJ6

Day 2:
- lecture: https://youtu.be/9DJWJpn0DmU
- slides: shorturl.at/dMNW8

3 Learning theory by Nicoló Cesa-Bianchi

3.1 Day 1

Outline:
- brief intro to statistical learning.
- from statistical learning to sequential decision making
- prediction with expert advice and multi-armed bandits
- Online convex optimization
- Contextual bandits
- Some short proofs

Statistical learning: pioneered by Vladimir Vapnik

Main contributions
- mathematical model of learning and conditions characterizing what can be learned
- Guidelines to practioners (e.g., choice of learning bias, control of overfitting)
- Principled and successful algorithms (SVM, Boosting)

Statistical learning setting:
- IID assumption and $S \sim D$.
- Statistical risk $l_D(f) = \mathbb{E}[l(Y, f(X))]$
- Bayes optimal predictor $f^*(x) = \arg\min_{\hat{y} \in \mathcal{Y}} \mathbb{E}[l(Y, \hat{y}|X = x)]$
- Bayes risk $l_D(f^*)$.

> The bias-variance decomposition: loss = estimation error + approximation error + Bayes risk
> - estimation error -> overfitting
> - approximation error -> underfitting
> - underfitting control: large hypothesis space
> - overfitting control: uniform convergence (ensure h is close to $l_D(h)$ for all h); regularized training error (stability); compress the training set.

Characterization of sample complexity (PAC-bound?) (VC dimension comes in a place here).

Statiscal efficiency. [skipped]

Online learning.
- Setting: data stream
- History: by Nick Littlestone and Manfred Warmuth '89
- The online learning protoco;
- Regret (seqquential risk) = sequentail counterpart of variance error in statistical learning.
- Online learning as a repated game (by James Hanna and DAvid BLackwell); replace data stream with a sequence of loss function;
- online learning in the simplex problem: lower bound $\sqrt{T \log d}$ (minimax rate).
- Hedging (exponetially weighted forecaster) -> bound $\sqrt{T \log d/2}$

Bandit problem
- limited feedback (bandit feedback)
- Applications: ad placement, dynamic content, real time bidding, recommender system, clinical trials, network protocol optimization.

3.2 Day 2

Hedge (exponential weights) on an observability graph.
Feedback models: experts, bandits, cops and robbers, revealing actions.
Online convex optimization problem:
- Projected descent (proximal form)
- Projected online GD
- Online mirror descent (OMD): use Bregman divergence rather than Euclidean distance to measure the "slow update" term.
- ψ in the Bregman divergence is μ-strongly convex.
Analysing regret: linearization step
AdaGrad:
- scale invariant to coordinate
- Useful in NN where grad varies across layers
- Analysis: apply OMD on each coordinate
Exploiting curvature of the losses
- OGD: ignore curvature
- Convex losses or strongly convex losses
Notions of regret:
- if the loss sequences does not have small cumulative value at all point, then regret bounds are meaningless
- if highly nonstationary data sequence, lack of single good minimizer -> think of more robust measures rather than regret
- dynamic regret
Contextual bandits
- actions have features (contexts)

3.3 Resources

Day 1:
- lecture: https://youtu.be/RflfnCsgNyI
- slides: shorturl.at/beiE3
Day 2:
- lecture: https://youtu.be/XA0iScv27W0

4 Fairness by Moritz Hardt

4.1 Day 1

Fairness through unawareness fails!
Fairness in supervised learning setting
Focus: decision at population level, not the training process
Decision theory 101: TN, TP, FP, FN, TP rate, TN rate, FP rate, FN rate.
Statistical fairness criteria
- r.v. encoding membership in protected class
- equalize statistical quantities regarding group (e.g., acceptance rate, error rate equal in all groups)

Fair representation: make it independent of group membership while representing original data as well as possible

Error rate parity is post-hoc

Group calibration often follows unconstrained learning.

Subgroup fairness

Machine bias

Prediction: narrow perspective (illustration: failure to appear to court, people fail to appear to court due to many reasons such as lack of child care, transportation; thus failing to attend court does not imply high risk)
Broader perspective
- Current statistical fairness criteria: only care about the joint statistics
- How to take salient social facts and context into account

4.2 Day 2

Broader perspective: causal fairness criteria

Ontological problem: what is the thing node in a causual graph reference?

Epistemological problem: how do we know facts about this thing

Ontological instability:
- Taylor instability
- Hacking instability

Summary for causality for fairness:
- Causality: confounding and mediation but not question of what is fair
- No causal fairness definition
- Causal modeling: suffers ontological and epistemological problems

4.3 Resources

Day 1:
- lecture: https://youtu.be/Igq_S_7IfOU

Day 2:
- lecture: https://youtu.be/9oNVFQ9llPc
- slides: shorturl.at/jRSZ0

5 Computational neuroscience in ML by Peter Dayan

5.1 Day 1 (only)

Outline:
- bio learning
- Conditioning: Pavlovian, prediction, TD learning, dopamine
- Bayesian conditioning: Kalman filtering, Chinese restaurant extinction

Marrian Cognition: 3 levels - computation (goal, logic), algorithm (procedure, repr) and impl (neural realization).

Neuroscience of learning:
- action/learning at synapse level

Conventional psychobio of leanring: procedure(implicit) (nonassociative + associative) <- Learning (memory) -> declarative (explicit) (episodic + semantic)

Ubiquitous learning of pred: require model-based/declarative control

Brain as forward/inverse models: motor, MDPs+policy, graphics/vision

Representation learning

Bakery LeCun: SSL, SL, RL (but RL is important)
Animals learn pred: acquisition phase + extinction phase
Formulation
- prediction = sum of binary stimuli
- average pred error
- learning rule as grad descent
Prediction: exponential weight of past rewards (RW rule)
Choice: [my connection lost :'(]
Dopamine and prediction error (distributional TD Nature paper)
Using prediction (as surrogate reward) for control
Direct algorithms: Actor-critic
Indirect alg: model-based, model-free

5.2 Resources

- slides: shorturl.at/mnzTU

6 Bayesian prediction with streaming data by Sonia Petrone

6.1 Day 1 (only)

Outline
- Bayes prediction
- Streaming data (inference and prediction have to be sequentially updated)
- From an alg to statistical methods
Stats: from data, estimate parameters and uncertainty quantification
ML: from data, train and predict
Core of Bayes: incomplete info via prob + learning though conditional prob
Bayes prediction: predictive distribution (via assigning a prior)
Example: unsupervised sequential learning and classification - predictive distribution of model parameters given data
Bayes approach: good in modeling heterogeneous data and complex dependence structures
Bayesian hierarchical models
Bayesian inference: predictive distribution of latent variables given the data
With streaming data, inference: e.g., sequential MC, sequential versions of variational Bayes
Important problem: trade-off btw statistical efficiency and computational efficiency

6.2 Resources

- lecture: https://youtu.be/ucj9341hSyQ
- slides: shorturl.at/cnNYZ

7 Game theory in ML by Constantinos Daskalakis

7.1 Day 1

Future of AI: learning + strategic reasoning
Minimization: learning/decison-making in a stationary env
Min-max optimization/equilibrium learning: changing env due to
- noise/adversaries

- presence of other agents

Focus: min-max optimization

Best scenario: convex-concave

Practical settings (e.g., GAN): non convex-concave

Main challenges:

- high-dimensional vars (statistical challenges)
- training oscillations [a simplified example where convex-concave, known function,]

Outline

- convex-concave: minmax, remove oscill via negative momentum
- nonconvex-nonconcave: musing, comp complex, multi-agent

Min-max theorem: min max f = max min f if f is convex-concave, unique optimal point

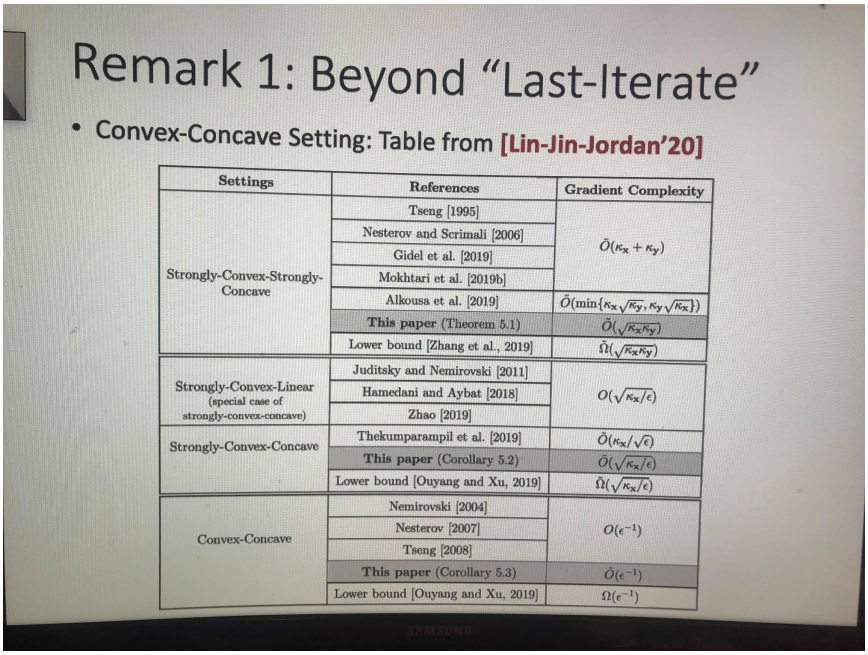
Fictitious play: play the best action w.r.t. the historical data of the opponent -> convergence guaranteed in bilinear case

Setting:

- learner: choose zt
- env choose L-Lipschitz convex loss
- learner observed the loss
- goal: no-regret
- e.g., Follow-the-regularized-leader

Negative momentum: do grad to today, but undo a bit of grad of yesterday (optimistic gradient descent)

Monotone variational inequalities



Settings	References	Gradient Complexity
Strongly-Convex-Strongly-Concave	Tseng [1995]	$\tilde{O}(\kappa_x+\kappa_y)$
	Nesterov and Scrimali [2006]	
	Gidel et al. [2019]	
	Mokhtari et al. [2019b]	
	Alkousa et al. [2019]	$\tilde{O}(\min\{\kappa_x\sqrt{\kappa_y},\kappa_y\sqrt{\kappa_x}\})$
	This paper (Theorem 5.1)	$\tilde{O}(\sqrt{\kappa_x\kappa_y})$
	Lower bound [Zhang et al., 2019]	$\tilde{\Omega}(\sqrt{\kappa_x\kappa_y})$
Strongly-Convex-Linear (special case of strongly-convex-concave)	Juditsky and Nemirovski [2011]	$O(\sqrt{\kappa_x}/\epsilon)$
	Hamedani and Aybat [2018]	
	Zhao [2019]	
Strongly-Convex-Concave	Thekumparampil et al. [2019]	$\tilde{O}(\kappa_x/\sqrt{\epsilon})$
	This paper (Corollary 5.2)	$\tilde{O}(\sqrt{\kappa_x}/\epsilon)$
	Lower bound [Ouyang and Xu, 2019]	$\tilde{\Omega}(\sqrt{\kappa_x}/\epsilon)$
Convex-Concave	Nemirovski [2004]	$O(\epsilon^{-1})$
	Nesterov [2007]	
	Tseng [2008]	
	This paper (Corollary 5.3)	$\tilde{O}(\epsilon^{-1})$
	Lower bound [Ouyang and Xu, 2019]	$\Omega(\epsilon^{-1})$

Figure 4: Convex-concave

7.2 Day 2

Wisdom: Minimization is to current AI what min-max optimization is to future AI (or more broadly, multi-agent equilibrium learning).

Focus: min-max optimization

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y)$$

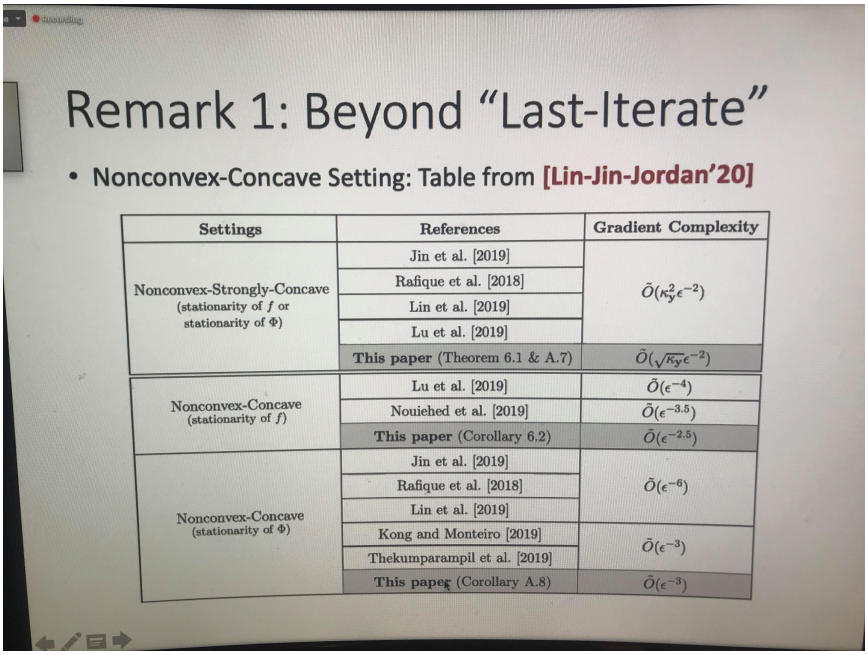


Figure 5: Nonconvex-concave

where x, y high dimensional and potentially constrained

Applications: mathematics, optimization, game theory

Best-case scenario: f is cont., convex-concave

Nonconvex-nonconcave optimization:

- under some f, min max is not max min
- under some other f, min max = max min but not unique solutions

Solution concept 1 for nonconvex-nonconcave objective: local min - global max solution

$$f(x^*, y) \leq f(x^*, y^*) \leq \max_{y' \in S(x,.)} f(x, y'), \forall y \in S(x^*, .), \forall x \in N_\delta(x^*) \cap S_{\mathcal{X}}$$

Solution concept 2: local minimax solution [Jin-Netrapali-Jordan ICML'20]

Solution concept 3: local min-max equilibrium

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*), \forall y \in S(x^*, .) \cap N_\delta(y^*), \forall x \in N_\delta(x^*) \cap S(., y^*)$$

1,2: sequential move notion, 3: simultaneous move notion

1: guaranteed to exist, but too strong (Np-hard); 2,3: don't always exist

GDA could a good solution to the nonconvex-nonconcave optimization:

Find GDA fixed points via Brouwer's fixed point theorems

7.3 Resources

Day 1:

- lecture: https://youtu.be/ks84JKokmqg
- slides: shorturl.at/lINQ2

Day 2:

- lecture: https://youtu.be/yNsrHLpjkvo

8 Optimization by Francis Bach

8.1 Day 1

large-scale: large n, large d

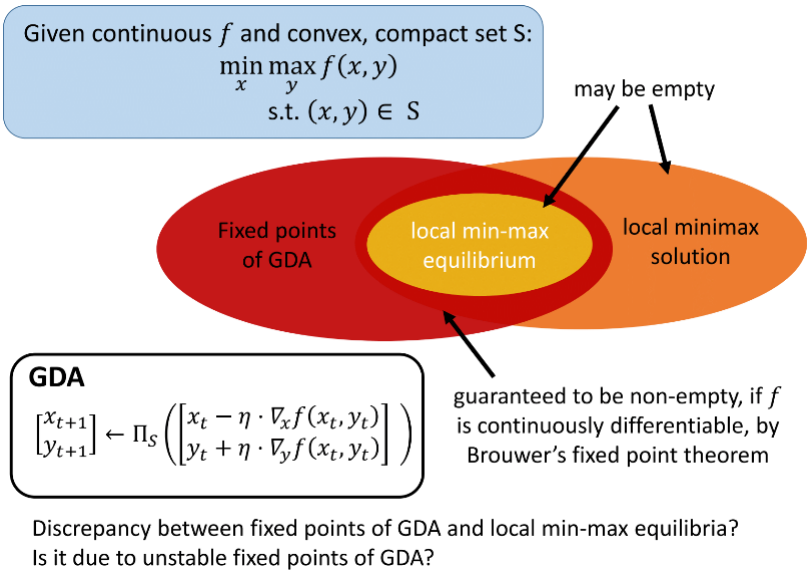


Figure 6: GDA

ml = opt of finite sums
variance reduction
global convergence for over-parameterized neural nets
Classical assumptions: smoothness and (strong) convexity
- L-smooth: twice diff and eigenvalue of second-order derivative bounded by L
- mu-strongly convex: twice diff and eigenvalues of second-order derivative low bounded by mu
- condition number: kappa = L / mu

8.2 Day 2

Non-strongly-convex case
Linearly convergent stochastic gradient algorithms
Robust averaged stochastic gradient: constant-step-size SGD is convergent for least-squares
Linearly-convergent SG methods
- provable and precise rates
- Improves on two known lower-bounds
- Several extensions, interpretations, accelerations
Beyond convex problems
Optimization for multi-layer neural nets: What could go wrong?
- Local minima
- Stationary points
- Plateaux
- Bad init
Generic local theoretical guarantees (convergent to stationary points or local minima)
General global performance guarantees impossible to obtain
GD for a single hidden layer
Optimization on measures
- Measure approximated by its empirical measures
- Many particle limit and global convergence
- Gradient flow: If you cannot do gradient flow, it is likely that you cannot do SGD

[Chizat and Bach, 2018a] "On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport":

- N to infinity, grad flow converges to global optimum)
- Two key ingredients: homogeneity and init

Mean-field limit

[Chiza and Bach, 2018b]

From lazy training to neural tangent kernel

Healthy interactions btw theory, applications, and hype:

- cannot ignore empirical success of DL
- should not lower scientific standards: critics and limits of theoretical and empirical results , rigor beyond mathematical guarantees

Conclusions

- Well understood: convex case w/ single machine, match lb and ub for variants of SGD, non-convex case of SGD for local risk minimization
- Not well understood: step-size schedule and acceleration, con-convexity, distributed learning

Wisdom from Francis Bach: Figure 7

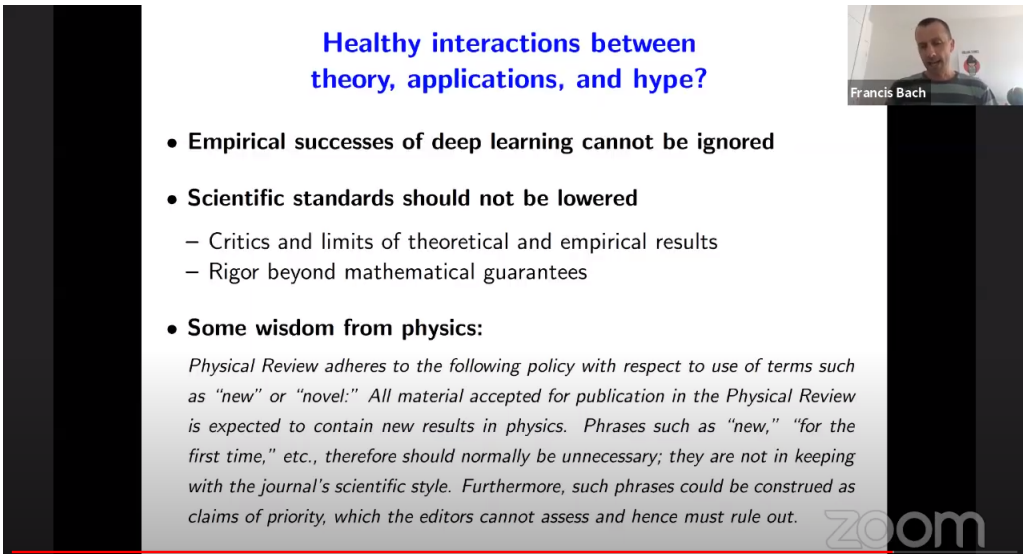


Figure 7: Bach's wisdom

8.3 Resources

Day 1:

- lecture: https://youtu.be/0MeNygohD6c
- slides: shorturl.at/qzAST

Day 2:

- lecture: https://youtu.be/hg2h53bU5ic

9 Optimal Transport by Marco Cuturi

9.1 Day 1

OT: natural geometry for probability measures

Outline

- Intro

- Exact OT
- Computing OT for data science
- Selected applications
Monge's problem
Figalli's work (Fields medalist)
Kantorovich problem (dual formulation of Wassterstein)
- (Cuturi derives this dual form in slide 39-40)
- Easier to store two functions than to store a coupling
Wasserstein distance: not just distance but interpolation between probability measures
Exact solvers for discrete measures

9.2 Day 2

How to compute OT?
Easy case 1: Univariate measures
Application of univariate measures: Sliced Wasserstein distance
- Project high-dimensional measures into line
- Fast and easy in practice
- But not clear what it computes, it is not OT
Easy case 2: Gaussian measures
Easy case 3: Elliptical measures
OT on empirical measures
- unstable P^* and not always unique
- cubic complexity and not paralleled
- Wasserstein distance is not differentiable
Computing OT for practioners
- computational properties
- statistical properties $\mathbb{E}[|W_p(\mu,\nu) - W_p(\hat{\mu}_n, \hat{\nu}_m)|] \leq f(n,m)$
where $f(n,n) = O(n^{-1/d})$
-> Regularize it!
Wishlist: faster, scalable, more stable, and differentiable
Sinkhorn's algorithm: $O(nm)$

9.3 Resources

Day 1:
- lecture: https://youtu.be/jgrkhZ8ovVc
- slides: shorturl.at/gCDGQ
Day 2:
- lecture: https://youtu.be/Bl8ZDN3Dbwk

10 Meta learning by Yee Whye Teh

10.1 Day 1

Small data problems
Role of meta-learning: learning to learn (inductive biases)
Multi-task learning: each task has its own data distribution
- Shared parameters + task-specific parameters
- Shared parameters: hope to capture the common structure of the tasks
Meta-learning is different from multi-task learning:
- meta-learning's goal: optimize shared parameters for test performance for each task

- multi-task learning's goal: learn all the tasks at the same time (?)

Optimization perspective on meta learning:

- Two level optimization problem

- Challenges: sensitive to neural architectures, can be expensive and unstable, back-propagating through many base learner iterations is hard

Black-box meta-learning:

- learn a diff. function from train data to test predictions

- challenges: no inductive biases -> hard to learn than optimization-based; less generalizable

- e.g., base learner can simply memorize data and match test data to memory

- can use matching networks, prototypical networks, memory-augmented networks (on the shared parameters) for memory matching

Permutation invariance is an inductive bias:

- data within each task should iid

- learner shoud invariant to permutation of training data

Datasets for meta-learning: Omniglot, mini-ImageNet, Meta Dataset.

10.2 Day 2

Conditional neural processes: observe -> aggregate (for task representation) -> predict

- characterize permutation-invariant

Probabilistic perspective on meta-learning

Stochastic process: a joint distribution over an infinite collection of r.v.s

Exchangeability and consistency

Bayesian nonparametrics

Latent variable model: generative model + variational objective

Meta-learning as learning a stochastic process

- meta-learn a prior over functions -> meta-learn inductive biases from meta-training set

- base learner can be as amortized learning

Uncertainties are important in meta-learning applications:

- active learning

- Bayesian optimization

- RL

Adversarial testing of RL agents

noise outsourcing:

$$(X, Y) \overset{a.s.}{=} (X, h(\eta, X)),$$

where $\eta \sim U[0, 1], \eta \perp X$

Combine invariant and equivalent modules

Probabilistic symmetries

10.3 Resources

Day 1:

- lecture: https://youtu.be/A0a1M61gjgI

- slides: shorturl.at/ghAPR

Day 2:

- lecture: https://youtu.be/WCREWkWGd6s

- slides: shorturl.at/lGI08

11 Deep learning by Yoshua Bengio

11.1 Day 1

Bengio's motivations to study AI: few principles give rise to intelligence

Neural net approach to AI

- brain-inspired
- distributed representation
- intelligence = objective or reward + approximate optimizer (learning rule) + architecture/parameterization
- end-to-end learning

Curse of dimensionality

- compositionality

DL: learn an internal representation

Exponential advantages of depth

Disentangle the factors of variation

Multi-task learning: learn intermediate representations that can be shared across tasks

Curriculum learning: present simple examples before harder examples (optimization in RL is much harder than that in standard DL)

The ultimate problem we need to solve is RL

11.2 Day 2

From attention to System 2 DL

Memory-extended networks

Multi-head attention

Attention is an internal action, needs a learned attention policy.

Systematic generalization

Missing from current ML: understanding and generalization beyond training distribution

- Learning theory only deals with generalization within the same distribution
- Models learn but do not generalize

System 1:

- intuitive, fast, unconscious, habitual, non-linguistic, implicit knowledge, current DL

System 2:

- slow, logical, sequential, conscious, linguistic, algorithmic, planning, reasoning, explicit knowledge, DL 2.0

From attention to consciousness: Global workspace theory

- bottleneck of conscious processing
- short-term memory, conditions perception and action
- system 2 -like sequential processing, conscious reasoning, planning and imagination
- only run 1 simulation at a time, and only few abstract concepts involved at each step

Implicit vs verbalizable knowledge

- most knowledge in our brain is implicit
- some is verbalizable where we can reason and plan explicitly
- concepts in verbalizable knowledge can be named with our language
- the joint distribution btw these concepts and the way the distribution can change over time satisfies special assumptions, exploited in system 2 tasks and conscious processing
- we want to clarify these assumptions as priors to be embedded into ML architectures and training frameworks, for better in- and out-of distribution

Some system 2 inductive bias:

- sparse factor graph in space of high-level semantic vars
- semantic variables are causal: agents, intentions, controllable objects

- distributional changes due to localized causal interventions (in semantic space)
- simple mapping btw high-level semnatic vars/thoughts/words/sentences
- shared 'rules' across instance tuples, requiring vars and indirection
- meaning is stable and robust wrt to changes in distribution
- credit assignment is only over short causal chains

Consciousness prior: sparse factor graph

- high-level vars we can manipulate with language: we can predict some given very few others
- disentangled factors: not marginally independent
- prior: sparse factor graph joint distribution btw high-level vars
- inference: involve few vars at a time, selected by attention and memory retrieval

DL objective: discover causal presentation

Sparse change in abstract latent space

Discovering cause and effect: A meta-transfer objective for learning to disentangle causal mechanisms

High-level representations = language

High-level concepts: meaning anchored in low-level perception and action -> tie system 1 and 2

Grounded language learning

Recurrent independent mechanisms (arXiv:1909.10893)

- modularize computation and operate on sets of named and typed ojects (object-centric representation)

Relation of different concepts: Fig. 8

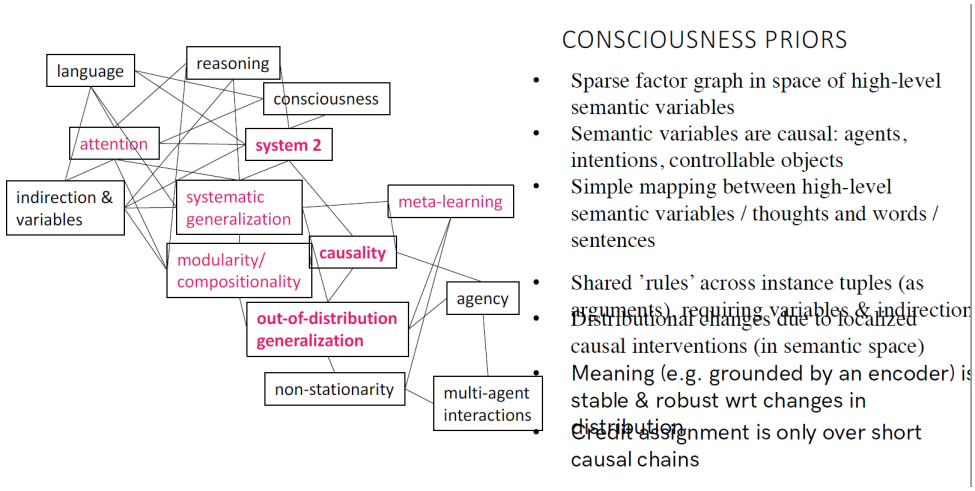


Figure 8: Relation of different concepts

11.3 Resources

Day 1:

- lecture: https://youtu.be/c_U4THknoHE
- slides: shorturl.at/zFSZ0

Day 2:

- lecture: https://youtu.be/PDPdIDihPvc

12 Kernel methods by Arthur Gretton

12.1 Day 1

Two-sample test
Testing goodness of fit
Kernel to enforce smoothness
Hilbert space: inner product space containing Cauchy sequence limits
RKHS def

12.2 Day 2

Mean maximum discrepancy (MMD) as an integral probability metric
Different divergences: MMD, Wasserstein, KL, chi-squared, Hellinger
Train GAN by MMD: MMD as critic, a scaled version of MMD, regularization

12.3 Resources

Day 1:
- lecture: https://youtu.be/alrKls6BORc
- slides: shorturl.at/jvyzK
Day 2:
- lecture: https://youtu.be/eANiXrWO1dM
- slides: shorturl.at/ejpY9

13 ML for healthcare by Mihaela van der Schaar

13.1 Day 1

Challenges:
- augment clinical decision making
- support and inspire clinical discovery
- new problem formulations, new ML models, new ML techniques
Opportunities: ML can transform healthcare
- precision medicine at patient-level
- ?
Plan
- AutoML for clinical analytics at scale
- interpretable, explanable and trustworth ML
- dynamic forecasting
- personalized monitoring and screening
- inividualized treatment effects
- how to get data?
Why AutoML for healthcare
- many algs to choose from
- many hyper-parameters
ML cannot do medicine but can provide actionable information: we cannot craft models for each disease -> make ML do the crafting!
Bayesian optimization for AutoML
BO does not work well for D > 10
- structured kernel: group correlated inputs (algorithms in healthcare) into a group
- grouping to be learned by using hierarchical Bayesian prior
Ensemble: create a linear combination of pipelines

Goal from AutoPrognosis: provide reliable evidence to assist difficult decisions to save lives

Radial GAN for translating data from one hospital to another

Interpretable, explainable, trustworthy ML

- interpret: why a prediction is made by the model
- explain: what can we learn from model

Our aims:

- understand what models discovered: feature importance, instance-wise feature important, feature/stats interactions - ?

Demystify black-box models using symbolic metamodels

- Metamodesl to map black-box models to white-box models
- Meijer functions (a very general function class)

How clinicians collaborate with ML healthcare systems?

13.2 Day 2

Dynamic forecasting:

- goal: Data-driven, accurate, interpretable model for a patient trajectory thatcan be used for risk scoring, prognosis and decision making
- challenges: multiple measurements, multiple outcomes, sparse measurements, unobserved states

Hidden absorbing semi-Markov model (HASMM)

Two goals of longitudinal models:

- accurate forecasting of individual-level disease trajectories
- understand disease progression mechanism

Attentive state space models:

- capture complex, non-stationary representations for patient-level trajectories

Dynamic and personalized comorbidity networks

Personalized monitoring and screening

Deep sensing: active sensing using multi-directional RNN [ICLR'18]

Temporal phenotyping using deep predicting clustering of disease progression [ICML'20]

Learning individualized treatement/causal effects from observational data

Modeling individualized treatment effect (ITEs):

- Causal modeling is not predictive modeling
- first theory for causal inference for ITE [ICML'18]
- Bayesian non-parametric ITE estimation
- Minimax rate for ITE estimation
- ITE using non-stationary GPs
- using GANs
- automating causal inference

Counterfactual RNNs:

= treatment invariant representations using domain adversarial training

- predicts counterfactuals using a novel sequence-to-sequence architecture

Can we have an end-to-end pipeline?

Differential privacy for healthcare

ML for healthcare: vision Fig. 9

13.3 Resources

Day 1:

- lecture: https://youtu.be/7gvBKP61Jus
- slides: shorturl.at/imnA0

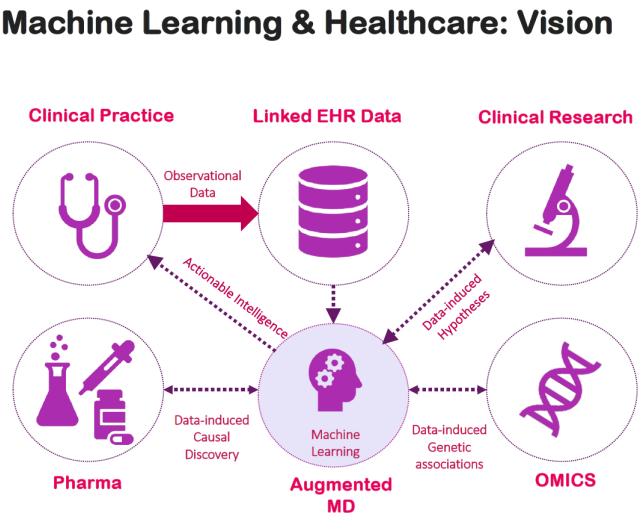


Figure 9: ML for healthcare: vision

Day 2:
- lecture: https://youtu.be/OVYNkIOPHnw

14 Geometric Deep Learning

14.1 Day 1 (only)

Inductive biases: translation equivariance, group equivariance

Examples of geometry: social networks, interaction networks, molecules, meshes, functional networks

Prototypical objects

Graph Fourier transform: use eigenvalues of the adjacency matrix A as the analogy of Fourier transform

Spatial graph convolution

Graph coarsening

Deep graph neural networks

Depth for graphs?

Message passing neural network

Weisfeiler-Lehman test

Graph substructure network

What's next?

- data + compute + SW
- Need ImageNEt for graphs (Open Graph Benchmark - OGB)
- software
- efficiency and scalability
- dynamic graphs (e.g., continuous-time graphs for Twitter)
- higher-order structures
- latent graphs
- theoretical understanding, robustness, performance guarantees

graphs = systems of relations and interactions

Applications

- recommender systems and link prediction
- high-energy physics
- neutrio detection

- computational chemistry and drug design
- hyperfoods
- combinatorial drug therapy
- protein science and cencer immunotherapy
- 3D vision and graphics

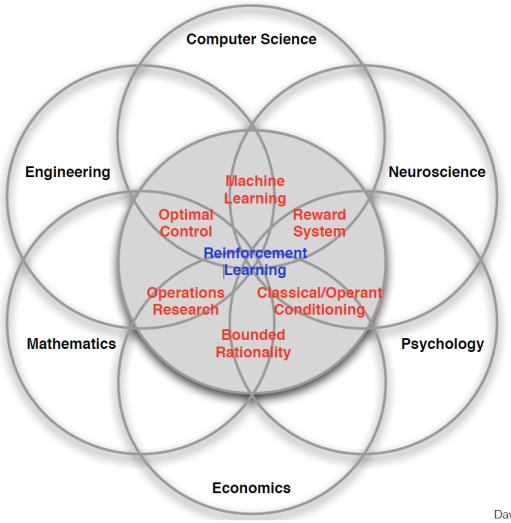
14.2 Resources

- lecture: https://youtu.be/8kTxTX0eBRA
- slides: shorturl.at/wIL35

15 Deep Reinforcement Learning by Doina Precup

15.1 Day 1

Key features of RL:
- not told which actions to take
- env is stochastic
- reward migh be delayed
- explore-exploit balance

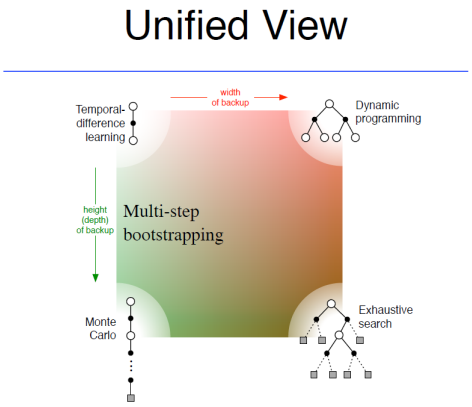


Some RL successes:
- world's best player of Backgammon
- acrobatic helicopter autopilots
- widely used in placement and selection of ad and web pages (e.g., A-B tests)
- make strategic decisions in Jeopardy
- human-level performance in Atari games
- in all these case, performance was better than any other methods, and was obtained without human instructions

Episodic tasks vs continuing tasks
- continuing tasks: discounted factor

Value function approximation (VFA):
- goal: minimize mean squared value error
- state aggregation is the simplest kind of VFA: states are grouped into subsets

Value of a state: $V(s)$
Value of an action in a state: $Q(s, a)$
MDP



Optimal value functions
Dynamic programming: bread first
Monte Carlo: depth first
TD learning

15.2 Day 2

For finite MDP, there is partial order among polices by their values
Bellman optimality equation
Value iteration
Policy iteration
Value function approximation (VFA) for control
Exploration
- randomization
- optimism in the face of uncertainty (e.g., UCB)
- probability matching (e.g., TS)
- plan to explore (e.g., Bayes-adaptive MDP)
SARSA (on-policy)
n-step semi-grad SARSA
Q-learning (off-policy)
- theoretical properties with function approximations: not convergent
- but empirical results are good
DQN
- target networks
- experience replay
Prioritized experience replay
- rank samples by TD error
Maximization bias:
- Double Q-learning
Agent57:
- episodic short-term novelty + long-term novelty
Approaches to control:
- action-value methods
- policy-gradient methods
Why approximate policies rather than values?
- policy can be simpler to approximate than value functions
- many problems, optimal policy is stochastic (e.g., POMDPs)
- enable smoother change in polices
- avoid a search on every step (max step)

- ?
Gradient-bandit algorithm
- store action preferences rather than action-value
- pick action by exponential soft-max
- store sample average of rewards
General policy-gradient:
- policy parameterization
- average reward as objective
- gradient of the objective based on policy-gradient theorem
Actor-critic architecture:
- A2C
- A3C
TRPO
PPO
Deep Deterministic Policy Gradient (DDPG)
AlphaGo:
- policy - procedural knowledge
- value - predictive knowledge
Two types of knowledge:
- procedural: skills, goal-driven behavior
- predictive, emprical: analogous to laws of physics, predicting effects of actions
Knowledge must be:
- expressive
- learnable
- composable
Option models:
- a procedural knowledge
- [subjective] model-based should learn temporal abstraction, not yet an example where model-based is better than model-free
- option-critic architecture
Option models can be GVFs
GVFs for synthesizing new behaviors
Ope questions:
- huge gap btw theory and practice
- more stable function approximators? (e.g., kernels, averages)
- policy or value-based? depend on application
- improve stability of deep RL
- exploration-exploitation

15.3 Resources

Day 1:
- lecture: https://youtu.be/hoHXV4Ujavg
- slides: shorturl.at/FLP29
Day 2:
- lecture: https://youtu.be/pN3hfZ6WaE0
- slides: shorturl.at/ptyVY

16 Bayesian learning by Shakir Mohamed

16.1 Day 1

Basics - computation - approximation - futures

Bayesian theory - manipulating probabilities - ethics and social impact

Probability definitions

- statistical
- logical
- subjective
- ?

Probabilistic models: write models in the language of probability

Probabilistic quantities

Bayesian statistics

Infinite exchangeable: joint probability is invariant to permutation of indices

De Finette's theorem: joint distribution = averaging out model parameters

Model-based Bayesian: likelihood model

Bayesian analysis

- integrate out unobserved variables
- integration is the central operation
- intractable integral

Regression and classification: probabilistic models ovef functions

Density estimation: learn probability distributions over data

Decision-making: probabilistic models over environments and actions

Marr's levels: computational -> algorithmic -> implementation

A poetics of ML: Fig 10

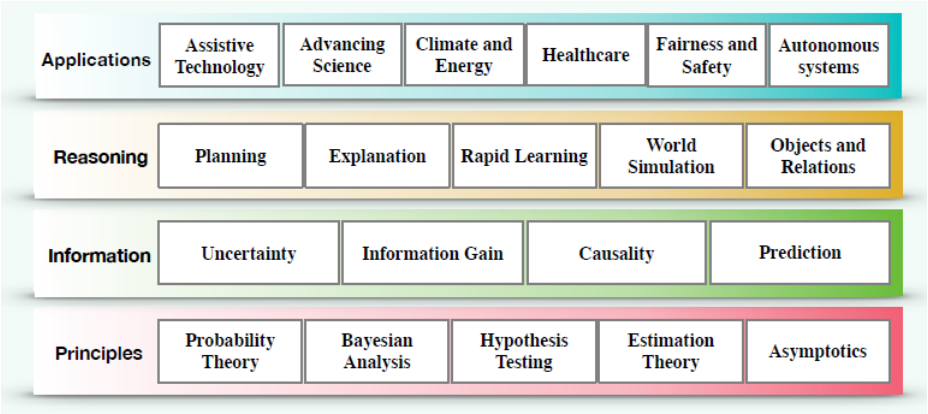


Figure 10: A poetics of ML

Statistical operations: Fig. 11

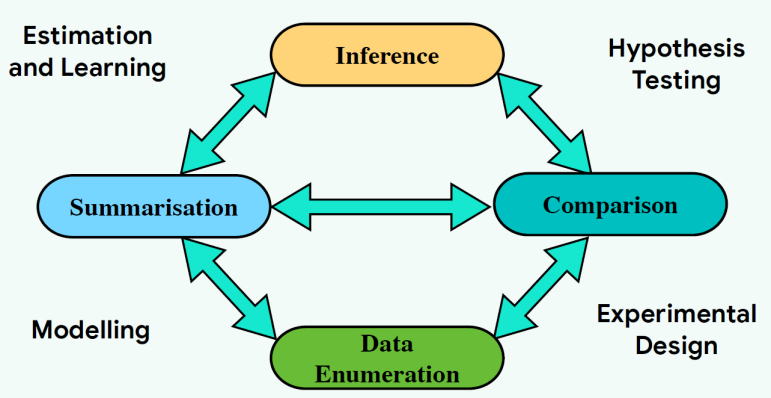


Figure 11: Statistical operations

Model-inference-algorithm:
- models
- learning principles
- algorithms
Statistical inference: Fig 12

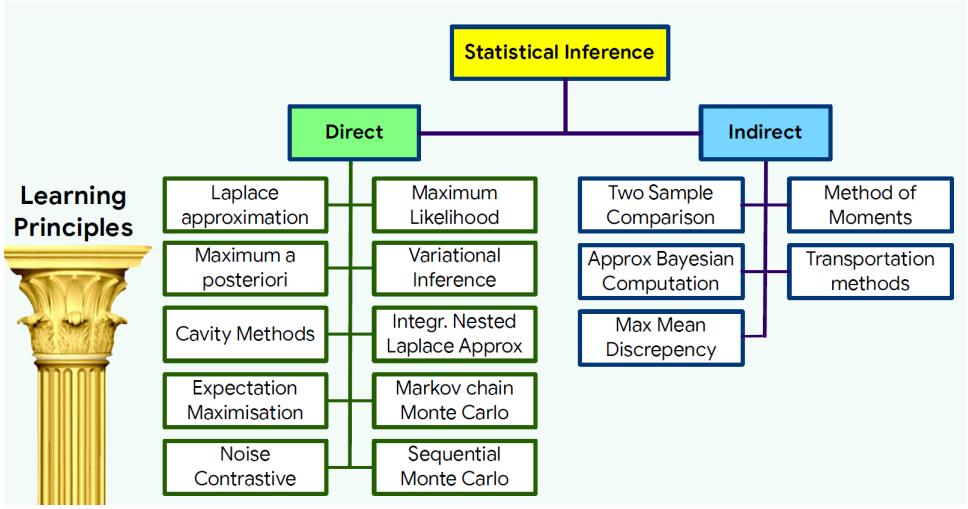


Figure 12: Statistical inference

Part 2: Bayesian computation
Probabilistic models and priors:
- linear regression
- deep nets
- deep and hierarchical: eg. information bottleneck
- Bayesian DL
Likelihood functions
Estimation theory
- regularization
- MAP estimation
- invariant MAP: used a modified probabilistic model to remove sensitivity (e.g., use Fisher information, information geometry)
Bayesian inference
Integral approximations
- reformulate to energy function

- Laplace approximation

Learning and inference:

- in Statistics, learning = inference = estimation (called inference)
- ML: inference - reason about unknown probability distributions; learning - estimate quantities in the model
- prediction

Bayes factors: to compare two models (hypothesis testing)

Inferential questions: Fig. 13

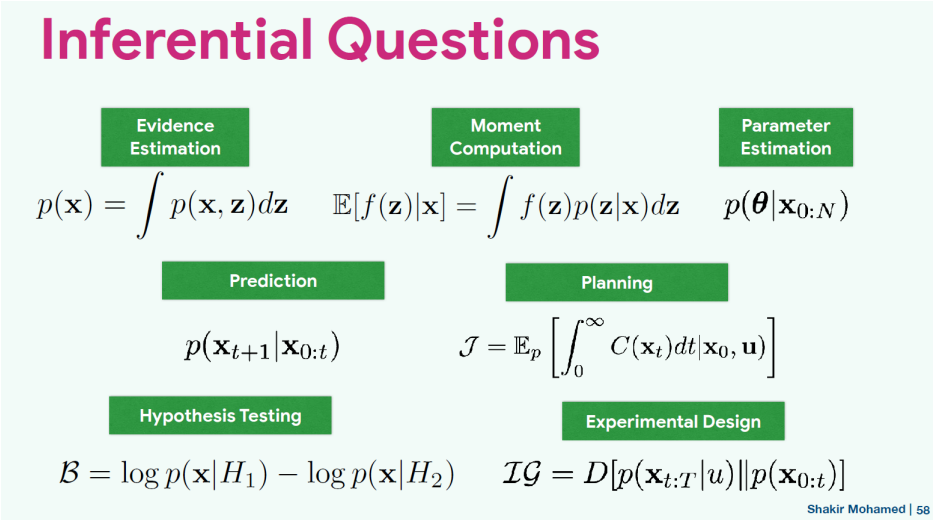


Figure 13: Inferential questions

16.2 Day 2

Represent distributions

- closed-form
- approximation

Monte Carlo methods:

- consistency
- unbiased
- low variance
- computation

Evaluating integrals

Importance sampling:

- identity trick: $w = p(z)/q(z)$

Markov Chain Monte Carlo:

- irreducible: cant from any state to any other state
- aperiodic: dont loop btw states
- ergodic: has a stationary distribution

Other MCMC methods:

- rejection sampling
- MH
- Gibbs
- Slice
- Metroplis with Gibbs
- Hamiltonian MC
- Seuqnetial MC

- Reversible Jump MC
- Non-Markovian (e.g., Stein, nested methods)
Limitations
- can be computationally expensive and slow
- can be difficult to evaluate
- Markov property ignore where samples come from in the past
- frequentist approach of large-sample behaviour
Approximation methods: Fig. 14

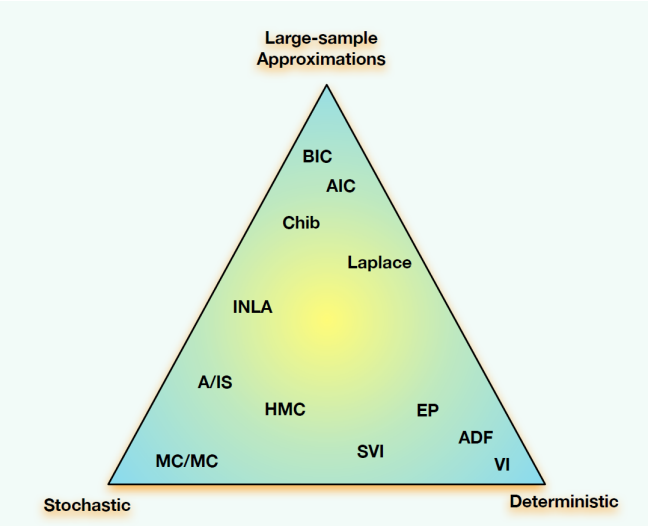


Figure 14: Approximation methods

Latent Variable Models:
- prescribed models: observer likelihood
- implicit models
Bounds in expectation:
- Jensen's inequality
Variational inference:
- integration is now optimization
- variational EM
- stochastic VI
- doubly stochastic VI
- amortized inference
Posterior approximations: Fig. 15
Stochastic VI:
- gradient before expectation (e.g., reparameterization trick)
- score function gradient using log-derivative trick and control variate
Log-derivative trick:
$E_{q(z)}[\nabla_\phi \log q_\phi(z)] = 0$
Amortized inference:
- instead of solving for every observation, amortize using a model
- use inference network as in VAE
Colonialism:
- be more critical and skeptical
Inference vs. decision-making:
- Inference: what we can know about our data
- Decision-making: what we can do with our data

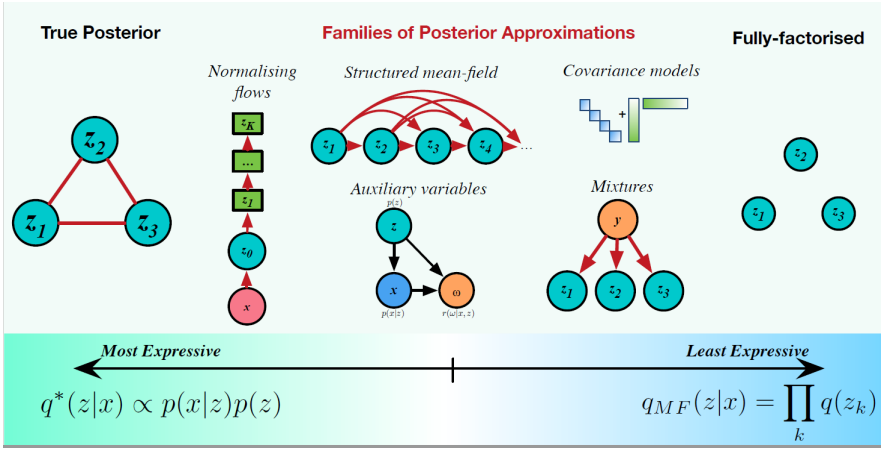


Figure 15: Posterior approximations

Bayesian RL:
- env as a generative process: unknown likelihood, not known, only able to observe its outcomes

Planning-as-inference:
- posteror distribution over actions?
- maximize probability of the return
- policy search
- hiararchical planning

Bayesian RL and control: expansion as probabilistic models
- Bellman's equation as different writing of message passing
- EM to policy search
- variational methods
- model-free and model-based

Bayesian deep learning:
- VI, MC dropout, variational dropout
- analyze of infinite width models using neural tangent kernels
- new MCMC methods

Bayesian optimization
Probabilistic dualities
Bayesian nonparametrics:
- fixed-dimension models
- adaptive-dimension models (e.g., GP)
- model complexity grows with data
- flexible and robust to overfitting

Mixture models
Dirichlet process mixture
Model relation: Fig. 16
Bayesian numerics:
- Numerical computation as Bayesain learning methods

Contextual values

16.3 Resources

Day 1:
- lecture: https://youtu.be/x4Y90zPjbq0
- slides: shorturl.at/bjuvH

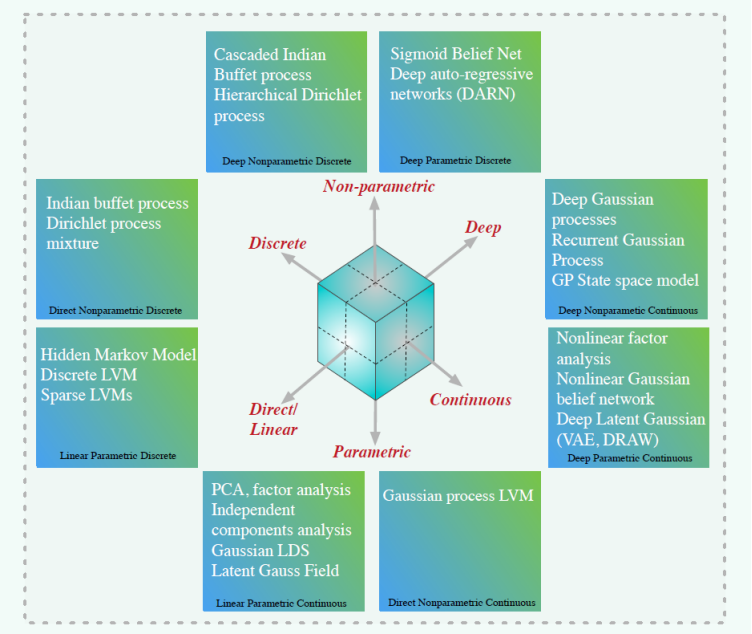


Figure 16: Relation of different models

Day 2:
- lecture: https://youtu.be/DItJbz2OH5U

17 Quantum machine learning by Maria Schuld

17.1 Day 1 (only)

Quantum computing is an emerging technology
Hardware for AI:
- CPU
- GPU
- QPU?
Data: classical and quantum
Algorithm: classical and quantum
AI = data/distributions + algs/hardware + models
Quantum computing:
- natural for ML
- like linear algebra(?)
Quantum theory predicts expectation of measurements
- quantum state: lives in Hilbert space H with scalar product
- observable represented by a Hermitian operator O
Quantum computers perform linear algebra
- quantum state: complex number $a + b * j$
- $|1|^2 = p(0...00)$
- $|0|^2 = p(0...01)$
Quantum computers cannot learn from "less than exponential" data
Quantum learning theory:
- in terms of queries, quantum is just polynomially better than classical
- in terms of time, quantum is much better than classical
Quantum computers can invert exponentially large matrices
Quantum computers can train Boltzmann machines

We can train quantum computations
- can do gradient descent on variational circuits
Quantum circuits are unitary neural nets in feature space
Quantum circuits are kernel methods:
- data encoding defines a quantum kernel
Open questions: Fig. 17

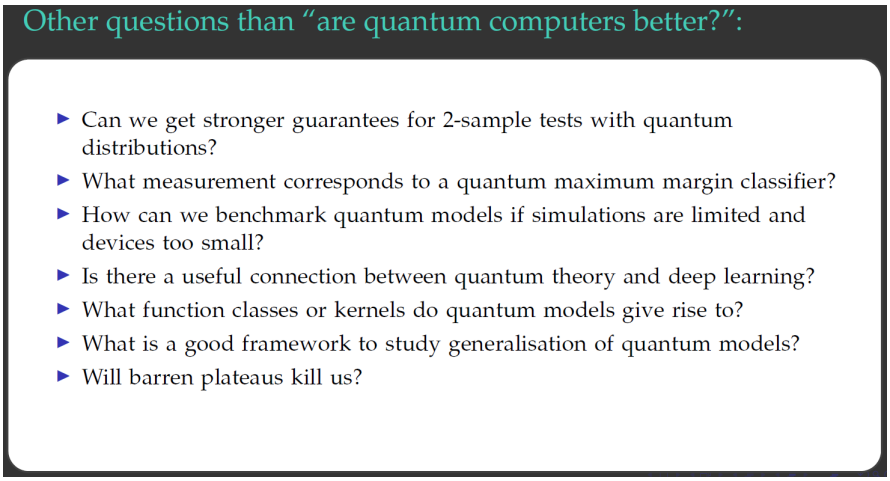


Figure 17: Open questions in quantum ml

17.2 Resources

- lecture: https://youtu.be/C_lBYKV_pJo