APPM 4490/5490
"Theory of Machine Learning"
Prof. Becker, spring 2022
Student projects

Generalized (Linear) Tikhonov Regularization and Noise Addition

(Grant Norman)

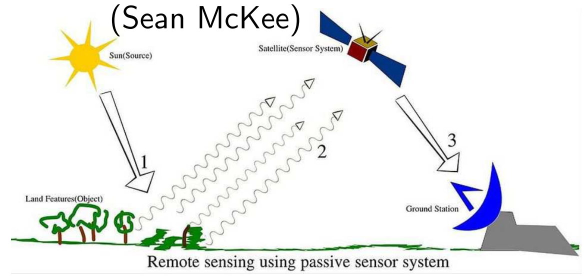
$$\arg\min_{\mathbf{w}\in\mathcal{H}} \tilde{L}_S = \arg\min_{\mathbf{w}\in\mathcal{H}} \left[\hat{L}_S(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{A}\mathbf{w}\|_2^2 \right]$$

$$\tilde{L}_{\mathcal{D}} = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[(h(x)-y)^2 + \eta^2 \left(\|\nabla h(x)\|^2 \right) \right]$$

$\sigma_{\min} = 0$
No Stability Guarantee…

Applications of machine learning to Inverse Scattering

(Sean McKee)

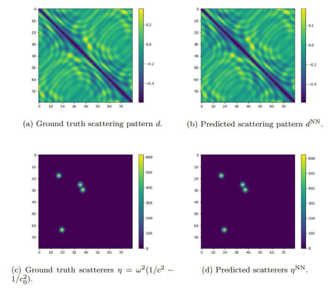


Physics-Assisted Learning Approach

Deconstruct the equations governing scattering, use this to inform the design of a minimally-sized neural network

"Switchnet: A neural network model for forward and inverse scattering problems."

Yuehaw Khoo and Lexing Ying



Sample Selection Bias in Machine Learning

(Robin Bowers)

Definition (Sample Selection Bias)

Our sample $S \sim \mathcal{D}' \neq \mathcal{D}$.

- Ancestry testing: samples overwhelmingly from North America/Europe/Australia
- Voluntary Survey participation: may be corellated with features
- Cancer screenings: only specific populations recommended for testing

Problem: We might ignore mistakes on training samples which are more common in the testing set.

Lemma ([1])

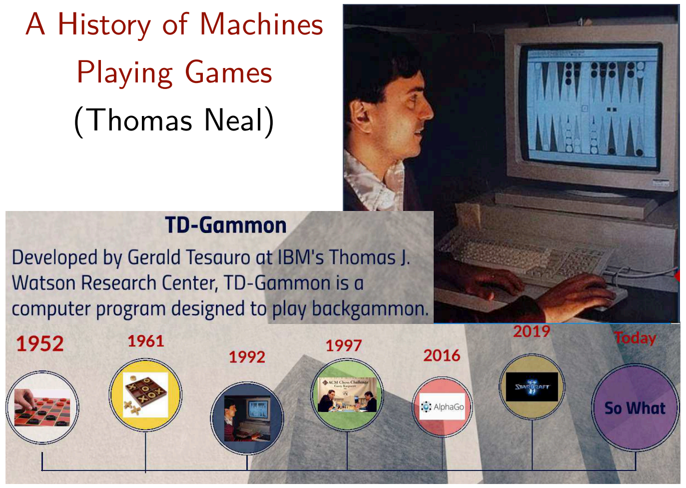
With probability at least $1-\delta$, *for all* $x \in S$:

$$\left|\mathbb{P}[s=1|x] - \widehat{\mathbb{P}}[s=1|x]\right| \leq \sqrt{\frac{\log 2|S| + \log\frac{1}{\delta}}{p_0|U|}}$$

where $p_0 = \min_{x\in U} \mathbb{P}[x] > 0$.

A History of Machines Playing Games

(Thomas Neal)



Universal Approximation Power of Deep Residual Networks: A review with applications to research in controls & optimization

Residual Networks $\iff$ Control Systems pt. 2 (Lily Cothren)

[1], [2] provide results from a control-theoretic view.[4]

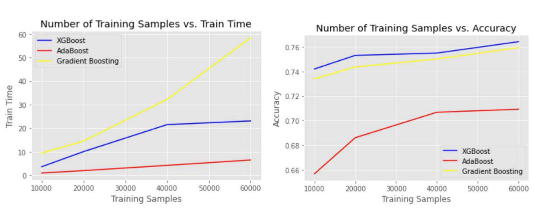
Discretized Differential Equation

To claim universal approximation of continuous functions, we need:

Assumption:

XGBoost

(Ben Sapper & Johnny Tamanaha)



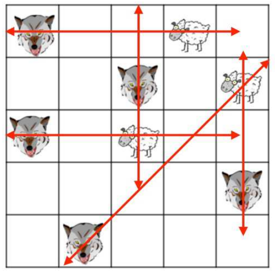
(a) Train Time (b) Accuracy

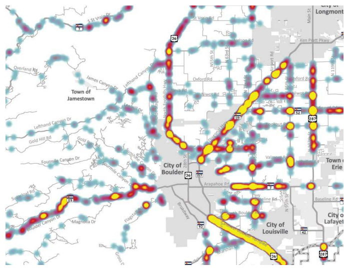
Figure 1: Multi-class Classification Boosting Comparison

Greedy and Risk-Averse Distributed Value Function Approximation via Convex Optimization

(Chi-Hui Lin)

$$V^n(x_s) = \max_{a^n\in A^n}[R^n(x_s,a^n) + \gamma \sum_{m\in neighbors(n)} f^{nm} \sum_{x_z\in X} p(x_z|a^n,x_s)V^m(x_z)]$$
$$\forall x_s \in X, \forall n \in [1, N]$$



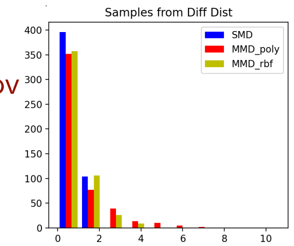


Student backgrounds:

- Applied Math (MS, PhD)
- Math (MS); Stat (BA)
- Computer Science (PhD)
- Electrical Engineering (MS, PhD)
- Aerospace Engineering (PhD)

Kernel Embedded Morozov Discrepancy Principle

(Noki Cheng)



The following proposition explains why MMD is a kernel embedded distance.

Proposition 2.1 Maximum mean discrepancy (MMD) between two distribution P and Q defined above can also be represented as

$$\mathrm{MMD}(P,Q) = \|w_P - w_Q\|_{\mathcal{H}},$$

where $w_P := \int_{\mathcal{X}} k(x,\cdot)d\mu_P$, and μ_P is the probability measure of P.

Symmetry, Invariance, and Machine Learning

(Ezzeddine AlSai)

How do symmetries arise?

Let G be the symmetry group of X, Y. A map $f: X \to Y$ is invariant if $f(T_g(x)) = f(x)$

it is equivariant if $f(T_g(x)) = T_g(f(x))$

Is there a generalization for more general G?

Yes if G is compact Lie

Kondor, Trivedi (2018)

The Benefit of Multitask Representation Learning: A Critical Review

(Killian Wood)

Single Task
Learn a predictor and representation for all schools all at once!
- Uses school-specific data poorly.

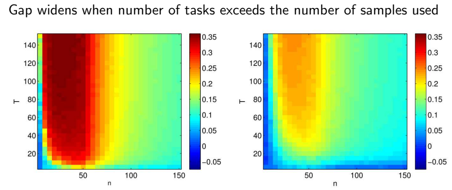
Multi-task
Learn a predictor for each school, and a representation for all schools. Train simultaneously.

Isolated Tasks
Learn a predictor *and representation* for each school separately. Train in isolation.
- Uses student-specific information poorly.

Theorem

Experimental Results

Binary Classification Problem: Dictionary representation using synthetic data.
Figures depict test error gap (0-1 loss) between MTL and ITL as a function of T and n.
Gap widens when number of tasks exceeds the number of samples used



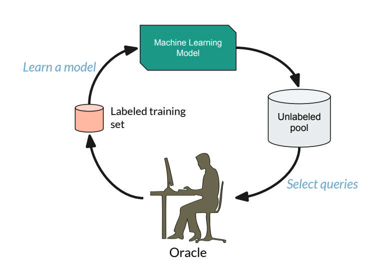
Review of Agnostic Active Learning

(Lauren Marsh)

Theorem 1 (Correctness Balcan et al.[3]). *For all H, for all (D, O), for all valid subroutines for computing UB and LB, for all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, with probability $1-\delta$, A^2 returns an ϵ-optimal hypothesis or does not terminate.*

Theorem 2 (Balcan et al.[3]). *For all H, for all (D, O), for all UB and LB satisfying $m(2\epsilon, \delta, H) \leq \frac{m(\epsilon,\delta,H)}{2}$, for all $0<\epsilon<1/2$ and $0<\delta<1/2$, the algorithm A^2 makes at most $2m(\epsilon, \delta', H)$ calls to oracle, O, where $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$ and $N(\epsilon,\delta,H)$ satisfies $N(\epsilon,\delta,H) \geq \ln\frac{1}{\epsilon}\ln m(\epsilon,\delta',H)$. Here $m(\epsilon,\delta,H)$ is the sample complexity of UB and LB.*

Valid for most "reasonable" subroutines for computing UB and LB, such as VC bound, Occam's Razor Bound or data-dependent generalization bounds based on Rademacher complexity



Consistent Region-Based Losses via Lovász Hinge

(Enrique Nueve)

Thm 23 [1]