Learning Minimax Estimators via Online Learning

Pradeep Ravikumar
Machine Learning Department
Carnegie Mellon University

Joint work with Arun Suggala, Kartik Gupta, Adarsh Prasad, Praneeth Netrapalli

Statistical Machine Learning

- Consider the following generic statistical machine learning task:

- Suppose we **observe** n i.i.d. samples $X_i \sim P_{\theta^*}$, for $i = 1, \ldots, n$
  - from some distribution or **statistical model** P_{θ^*} from the family $\{P_\theta\}_{\theta \in \Theta}$.
- The goal of **learning** is to infer $\theta^* \in \Theta$ from the samples $\{X_i\}_{i=1}^n$

Estimators

- We do so via **estimators** $\widehat{\theta} : \mathcal{X}^* \mapsto \Theta$, where $\mathcal{X}$ is the domain of the samples, so that $X_i \in \mathcal{X},\ i = 1, \ldots, n$.

Loss & Risk

- We evaluate estimators using a **loss function** $\mathcal{L} : \Theta \times \Theta \mapsto \mathbb{R}$
- Since the estimator $\widehat{\theta}(X_1^n)$ is random, so is $\mathcal{L}(\widehat{\theta}(X_1^n), \theta^*)$.
- So it is more natural to use the **risk function** $R(\widehat{\theta}, \theta^*) = \mathbb{E}_{X_1^n \sim P_{\theta^*}} \mathcal{L}(\widehat{\theta}(X_1^n), \theta^*)$.

Optimality

- So it is more natural to use the **risk function** $R(\widehat{\theta}, \theta^*) = \mathbb{E}_{X_1^n \sim P_{\theta^*}} \mathcal{L}(\widehat{\theta}(X_1^n), \theta^*)$.
- But there exists a very simple estimator with the least risk; the constant estimator: $\widehat{\theta}(X_1^n) = \theta^*$.

Global Optimality

- We are thus interested in "global" notions of optimality
- A natural notion is **minimax optimality**, requiring that the estimator is optimal (or with a constant factor approx.) with respect to:

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta} R(\widehat{\theta}, \theta^*).$$

Minimax Optimality

- **Minimax Optimality**:

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta} R(\widehat{\theta}, \theta^*).$$

- Widely studied (Ibragimov-Has'minskii 81, Vaart 98, Tsybakov 08, ...)
- Minimax estimators obtained either via problem specific approaches (with no general recipe), or by showing post-hoc that estimators derived from other principles (e.g. likelihood principle) are (asymptotically) minimax using information-theoretic tools

Learning Minimax Estimators

- **Minimax Optimality**:

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta} R(\widehat{\theta}, \theta^*).$$

- But would it be possible to **algorithmically** derive the optimal minimax-estimator?

Statistical Games

- **Minimax Optimality**:

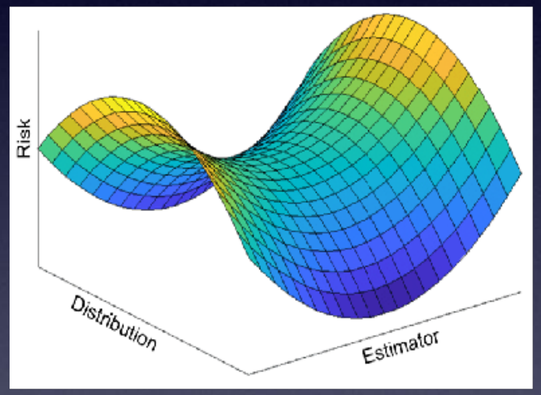
$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta} R(\widehat{\theta}, \theta^*).$$

- But would it be possible to **algorithmically** derive the optimal minimax-estimator?
- Towards this, note that the above can be viewed as a zero-sum **statistical game**, between a learner, whose actions are estimators $\widehat{\theta}$, and nature, whose actions are statistical models, with parameters $\theta^* \in \Theta$.
- Can we directly solve this statistical game?

Statistical Game

- Suppose the statistical game is convex-concave: $R(\widehat{\theta}, \theta)$ is convex in $\widehat{\theta}$ and concave in θ
- It then naturally follows (Sion 58) that:

$$\min_{\widehat{\theta}} \max_{\theta^* \in \Theta} R(\widehat{\theta}, \theta^*) = \max_{\theta^* \in \Theta} \min_{\widehat{\theta}} R(\widehat{\theta}, \theta^*).$$

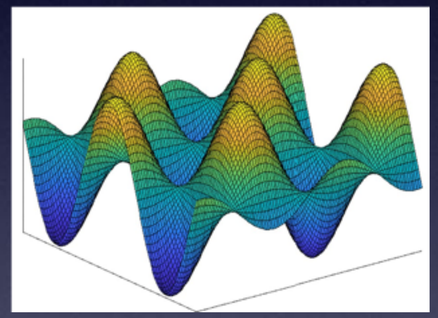


- The optimal solution $(\widehat{\theta}, \theta^*)$ is called Nash Equilibrium (NE)
- Many efficient approaches for finding NE, including gradient descent ascent

Statistical Game: Difficulties

- But the statistical game is non (convex-concave): $R(\widehat{\theta}, \theta^*)$ is non-concave in θ^*
- Thus, for most statistical games:

$$\min_{\widehat{\theta}} \max_{\theta^* \in \Theta} R(\widehat{\theta}, \theta^*) \neq \max_{\theta^* \in \Theta} \min_{\widehat{\theta}} R(\widehat{\theta}, \theta^*).$$



- Moreover, the set of *all possible estimators* $\widehat{\theta}$ is too large for typical algorithms to solve zero-sum games

Statistical Game: Difficulties

- We could thus aim to solve the **linearized game**:

$$\min_{P_{\widehat{\theta}}} \max_{P_{\theta^*}} \mathbb{E}_{\widehat{\theta} \sim P_{\widehat{\theta}}, \theta^* \sim P_{\theta^*}} R(\widehat{\theta}, \theta^*)$$

- Learner's actions: distributions over all possible estimators; Nature's actions: distributions over all possible parameters
- This is a bilinear (and hence convex-concave): so minimax theorem holds. The NE of linearized game is also referred to as a mixed NE of the original game
- Bottleneck: if the space of all possible estimators is large, distributions over all possible estimators is even larger. Modern game solving algorithms do not scale!

Statistical Game: Reduction to Online Learning

- One approach to solve zero-sum games is via reduction to online learning
- Let us briefly recall online learning: which could be viewed as a repeated game between a learner and an adversary
- In each round, the learner plays an action $\theta_t \in \Theta$, and the adversary chooses the loss f_t, so that the learner suffers loss $f_t(\theta_t)$
- The goal of the learner is to minimize **regret** defined as:

$$\sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^{T} f_t(\theta).$$

- There exist many algorithms to achieve regret that is sub-linear in T: Online Mirror Descent, Follow The Regularized Leader (FTRL; McMahan 11), Follow the Perturbed Leader (FTPL; Kalai & Vempala 05, Suggala & Netrapalli 20)

Statistical Game: Reduction to Online Learning

- Proposition (Gupta, Suggala, Prasad, Netrapalli, Ravikumar, 21): Suppose the learner and nature play $\epsilon_\ell(T)$ and $\epsilon_n(T)$ regret online learning strategies to choose their actions at each round t, given previous round actions $\{P_{\widehat{\theta};s}; P_{\theta^*;s}\}_{s=1}^{t-1}$
- Consider the randomized estimator

$$\widehat{\theta}_{\mathrm{RND}} = \frac{1}{T}\sum_{t=1}^{T} P_{\widehat{\theta};t}$$

  and mixture distribution

$$P_{\mathrm{AVG}} = \frac{1}{T}\sum_{t=1}^{T} P_{\theta^*;t}.$$

- Then $(\widehat{\theta}_{\mathrm{RND}}, P_{\mathrm{AVG}})$ are an $\epsilon(T) := \epsilon_\ell(T) + \epsilon_n(T)$ approximate mixed NE of the statistical game, so that:

$$\sup_{\theta^* \in \Theta} R(\widehat{\theta}_{\mathrm{RND}}, \theta^*) - \epsilon(T) \leq R(\widehat{\theta}_{\mathrm{RND}}, P_{\mathrm{AVG}}) \leq \inf_{\widehat{\theta}} R(\widehat{\theta}, P_{\mathrm{AVG}}) + \epsilon(T).$$

Computational Caveats

- The key caveat is that sub-linear regret online learning strategies have computational complexity that scale with the size of the domain: infeasible when the domain of learner actions is (distributions over) all possible estimators!
- There is however a particular choice of sub-linear regret online learning strategies that map to much more tractable well-known problems

Solving Statistical Games

- The learner plays **Best Response** to nature action at round t, so that:

$$\widehat{\theta}_t = \arg\min_{\widehat{\theta}} R(\widehat{\theta}, P_{\theta^*;t}).$$

- This is simply the well-known **Bayes Estimation** Problem: find the optimal estimator with respect to the prior $P_{\theta^*;t}$

Solving Statistical Games

- Nature plays FTPL:

$$\theta^*(\sigma) = \arg\max_{\theta \in \Theta} \sum_{s=1}^{t-1} R(\widehat{\theta}_s, \theta) + \langle \sigma, \theta \rangle,$$

  for $\{\sigma_j\} \sim^{iid} \text{Exp}(\eta)$

- Note that nature is playing a mixed strategy P_{θ^*} that is implicitly specified via $\theta^*(\sigma)$, given a random noise vector σ

- This is a finite-dimensional maximization program, which while non-convex, is a standard sub-routine in modern ML

Oracles

- We assume access to the following two oracles
- An approximate **Bayes Estimator** oracle $\mathcal{O}_{\alpha}^{\text{Bayes}}$ that given any distribution P over Θ outputs $\widehat{\theta}$ s.t.

$$R(\widehat{\theta}, P) \leq \inf_{\widehat{\theta}} R(\widehat{\theta}, P) + \alpha.$$

- An approximate **Maximization** oracle $\mathcal{O}_{\beta}^{\text{Max}}$, that provides a β-approx solution θ to the following max. program, given a set of estimators $\{\widehat{\theta}_s\}_{s=1}^{T}$:

$$\sum_{t=1}^{T} R(\widehat{\theta}_t, \theta) + \langle \sigma, \theta \rangle \geq \sup_{\theta' \in \Theta} \sum_{t=1}^{T} R(\widehat{\theta}_t, \theta') + \langle \sigma, \theta' \rangle - \beta.$$

Regularity Assumptions

- Suppose the parameter domain Θ is compact, with bounded ℓ_∞ diameter $D = \sup_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|_\infty$.
- Suppose $R(\widehat{\theta}, \theta^*)$ is L-Lipschitz in its second argument wrt ℓ_1 norm:

$$|R(\widehat{\theta}, \theta_1) - R(\widehat{\theta}, \theta_2)| \leq L\|\theta_1 - \theta_2\|_1.$$

- Suppose nature plays FTPL with noise drawn from $\mathrm{Exp}(\sqrt{\frac{1}{dL^2T}})$

Algorithmic Minimax Estimators

- Theorem (Gupta, Suggala, Prasad, Netrapalli, Ravikumar 21): Suppose that in each of T rounds of the statistical game, the learner plays best response, and nature plays FTPL, using the earlier oracles, and suppose the earlier regularity assumptions hold.
- Then the randomized estimator $\widehat{\theta}_{\mathrm{RND}}$ and the mixture distribution P_{AVG} are ϵ-approximate mixed NE of the statistical game, for $\epsilon = O(d^{3/2}LT^{-1/2} + \alpha + \beta)$.

Deterministic Minimax Estimators

- The earlier theorem provided a randomized estimator, whereas the typical analysis of minimax estimators focuses on deterministic estimators

Deterministic Minimax Estimators

- The earlier theorem provided a randomized estimator, whereas the typical analysis of minimax estimators focuses on deterministic estimators
- When the risk function $R(\widehat{\theta}, \theta^*)$ is convex in the first argument, we could use the deterministic estimator: $\widehat{\theta}_{\text{AVG}} = \mathbb{E}_{\theta \sim \widehat{\theta}_{\text{RND}}}[\theta]$, which can again be shown to be an approximate NE.
- For general non-convex risk functions however, even with access to our oracles, finding a deterministic estimator will be NP-Hard (drawing from (Chen, Lucier, Singer, Syrgkanis 17))

Near Minimax Optimality

- Suppose the true minimax risk is $R^* := \min_{\widehat{\theta}} \max_{\theta^* \in \Theta} R(\widehat{\theta}, \theta^*)$.
- Our algorithm outputs a (randomized) minimax estimator with worst-case risk $(1 + o(1))R^*$.
- This is contrast to most *rate-optimal* minimax estimators which have worst-case risk $O(1)R^*$

Invariant Statistical Games

- A statistical game is invariant wrt group transformations G if:
  - If $\theta \in \Theta$, then $g\theta \in \Theta$
  - If $X \sim P_\theta$, then $gX \sim P_{g\theta}$
  - $\mathcal{L}(g\theta_1, g\theta_2) = \mathcal{L}(\theta_1, \theta_2)$.
- An estimator $\widehat{\theta}$ is invariant wrt group transformations G if for all $g \in G$,

$$\widehat{\theta}(g\, X_1^n) = g\, \widehat{\theta}(X_1^n).$$

Invariant Statistical Games

- Theorem (Gupta, Suggala, Prasad, Netrapalli, Ravikumar 21): Suppose the statistical game is invariant wrt group transformations G. And suppose the risk function $R(\widehat{\theta}, \theta^*)$ is convex in its first argument. Then:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\widehat{\theta}, \theta^*) = \inf_{\widehat{\theta;G}} \sup_{\theta_G \in \Theta/G} R(\widehat{\theta}, \theta^*),$$

  where in the latter reduced game, for the learner we restrict to invariant estimators, and for nature we restrict to the smaller quotient space Θ/G.

- Moreover, from any approximate mixed NE of the reduced game, we can easily reconstruct an approximate mixed NE of the original statistical game.

Application: Finite Gaussian Sequence Model

- Suppose we are given a single sample $X \in \mathbb{R}^d$ drawn from $N(\theta^*, I)$
- $\theta^* \in \Theta := \{\theta : \|\theta\|_2 \leq B\}$
- With the squared loss, we get the following statistical game:

$$\min_{\widehat{\theta}} \sup_{\|\theta^*\|_2 \leq B} \mathbb{E}_{X \sim N(\theta^*, I)}[\|\widehat{\theta} - \theta^*\|_2^2].$$

- Inspite of its simplicity, a fundamental problem with consequences for general non-parametric regression
- A well studied problem (Bickel et al, 81; Berry 90, Marchand & Perron 02,...)
- Exact minimax estimator not known for $B \geq 1.16\sqrt{d}$; our work (algorithmically) resolves this

Application: Finite Gaussian Sequence Model

- The finite Gaussian sequence statistical game is invariant with respect to group $\mathcal{O}(d)$ of orthonormal matrices with matrix mult. as group operation. And where for $g \in \mathcal{O}(d)$, the group action is given as: $g(X, \theta^*) = (gX, g\theta^*)$.
- The quotient space $\Theta/\mathcal{O}(d)$ is homeomorphic to the real interval $[0, B]$
- The reduced statistical game is given as:

$$\inf_{\widehat{\theta}} \sup_{b \in [0,B]} R(\widehat{\theta}, b\mathbf{e}_1).$$

Application: Finite Gaussian Sequence Model

- The maximization oracle involves simple 1D optimization over the bounded interval $[0, B]$
- The Bayes estimation oracle is available in closed form:

$$\hat{\theta}(X) = \left(\frac{\mathbb{E}_{b\sim P}\left[b^{3-d/2} e^{-b^2/2} I_{d/2}(b\|X\|_2) \right]}{\mathbb{E}_{b\sim P}\left[b^{2-d/2} e^{-b^2/2} I_{d/2-1}(b\|X\|_2) \right]} \right) \frac{X}{\|X\|_2},$$

where I_ν is the modified Bessel function of first kind of order ν.

Application: Finite Gaussian Sequence Model

Estimator	Worst-case Risk								
	$\mathbf{B = \sqrt{d}}$			$\mathbf{B = 1.5\sqrt{d}}$			$\mathbf{B = 2\sqrt{d}}$		
	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$
Standard	10	20	30	10	20	30	10	20	30
James Stein	6.0954	11.2427	16.073	7.9255	15.0530	21.3410	8.7317	16.6971	24.7261
Projection	8.3076	17.4788	26.7873	10.3308	20.3784	30.2464	10.1656	20.2360	30.3805
Bayes estimator for uniform prior on boundary	**4.8559**	**9.9909**	**14.8690**	11.7509	23.4726	35.2481	24.5361	49.0651	73.3158
Averaged Estimator	**4.7510** (0.1821)	**9.7299** (0.2973)	**14.8790** (0.0935)	**6.7990** (0.0733)	**13.8084** (0.2442)	**20.5704** (0.0087)	**7.8504** (0.3046)	**15.6686** (0.2878)	**23.8758** (0.6820)
Bayes estimator for avg. prior	**4.9763**	**10.1273**	**14.8128**	**6.7866**	**13.8200**	**20.3043**	**7.8772**	**15.6333**	**23.5954**

- The worst-case risk of our algorithmic estimators (last two rows) smaller than baselines (numbers in brackets indicate duality gap)

Application: Finite Gaussian Sequence Model

	$\mathbf{B=1}$	$\mathbf{B=2}$	$\mathbf{B=3}$	$\mathbf{B=4}$
Worst case risk of Averaged Estimator	0.456	0.688	0.799	0.869
Lower bound	0.449	0.644	0.750	0.814

- (Donoho 90) derived lower bounds for d = 1
- Table shows comparison with our algorithmic estimator

Application: Linear Regression

- In Linear Regression with random design, we are given n samples $D_n = \{(X_i, Y_i)\}_{i=1}^n$ generated as: $Y_i = X_i^T \theta^* + \epsilon_i$, $X_i \sim N(0, I)$, and $\epsilon_i \sim N(0, 1)$.
- We assume the parameters $\theta^* \in \Theta := \{\theta : \|\theta\|_2 \leq B\}$
- Given the squared loss, we get the following statistical game:

$$\min_{\widehat{\theta}} \max_{\|\theta^*\|_2 \leq B} \mathbb{E}_{D_n} \left[\|\widehat{\theta}(D_n) - \theta^*\|_2^2 \right].$$

Application: Linear Regression

- The linear regression statistical game is invariant with respect to group $\mathcal{O}(d)$ of orthonormal matrices with matrix mult. as group operation. And where for $g \in \mathcal{O}(d)$, the group action is given as: $g((X, Y), \theta^*) = ((gX, Y), g\theta^*)$.
- The quotient space $\Theta/\mathcal{O}(d)$ is homeomorphic to the real interval $[0, B]$
- The reduced statistical game is given as:

$$\inf_{\widehat{\theta}} \sup_{b \in [0,B]} R(\widehat{\theta}, b\mathbf{e}_1).$$

Application: Linear Regression

- The maximization oracle involves simple 1D optimization over the bounded interval $[0, B]$
- Suppose $C(A, \gamma)$ is the normalization constant of a Fisher-Bingham distribution:

$$p(Z; A, \gamma) \propto \exp(-z^T AZ + \langle \gamma, Z \rangle).$$

- The Bayes estimation oracle is available in closed form in terms of the normalization constant above:

$$\hat{\theta}(D_n) = \frac{\mathbb{E}_{b\sim P}\left[b^2 \frac{\partial}{\partial \gamma} C\left(2^{-1}b^2 \mathbf{X}^T\mathbf{X}, \gamma\right)\Big|_{\gamma = b\mathbf{X}^T\mathbf{Y}}\right]}{\mathbb{E}_{b\sim P}\left[b\, C\left(2^{-1}b^2\mathbf{X}^T\mathbf{X}, b\mathbf{X}^T\mathbf{Y}\right)\right]},$$

where $\mathbf{X} = [X_1, X_2 \ldots X_n]^T$ and $\mathbf{Y} = [Y_1, Y_2 \ldots Y_n]$.

Application: Linear Regression

Estimator	Worst-case Risk							
	$\mathbf{n = 1.5 \times d, B = 0.5 \times \sqrt{d}}$				$\mathbf{n = 1.5 \times d, B = \sqrt{d}}$			
	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 15}$	$\mathbf{d = 20}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 15}$	$\mathbf{d = 20}$
OLS	5.0000	2.5000	2.5000	2.2222	5.0000	2.5000	2.5000	2.2222
Ridge regression	0.6637	0.9048	1.1288	1.1926	1.3021	1.4837	1.6912	1.6704
Averaged Estimator	**0.5827** (0.0003)	**0.8275** (0.0052)	**0.9839** (0.0187)	**1.0946** (0.0404)	**1.2030** (0.0981)	1.4615 (0.1145)	1.6178 (0.1768)	1.6593 (0.1863)
Bayes estimator for avg. prior	**0.5827**	**0.8275**	**0.9844**	**1.0961**	**1.1750**	1.4621	1.6265	1.6674

- The worst-case risk of our algorithmic estimators (last two rows) smaller than baselines (numbers in brackets indicate duality gap)

Application: Entropy Estimation

- Consider Entropy Estimation, given n samples $X_1^n = \{X_i\}_{i=1}^n$ from a discrete distribution $P^* = (p_1, \ldots, p_d)$, and where $X_i \sim \{1, \ldots, d\}$.
- The goal is to estimate entropy of P^*: $\theta^*(P^*) = -\sum_{i=1}^d p_i^* \log p_i^*$.
- The domain of discrete distributions is the d-simplex Δ_d
- Given the squared loss, we get the following statistical game:

$$\min_{\widehat{\theta}} \max_{P^* \in \Delta_d} \mathbb{E}_{X_1^n} \left[(\widehat{\theta}(X_1^n) - \theta^*(P^*))^2 \right].$$

Application: Entropy Estimation

Estimator	Worst-case Risk									
	$\mathbf{d = 10}$		$\mathbf{d = 20}$		$\mathbf{d = 40}$			$\mathbf{d = 80}$		
	$\mathbf{n = 10}$	$\mathbf{n = 20}$	$\mathbf{n = 20}$	$\mathbf{n = 40}$	$\mathbf{n = 10}$	$\mathbf{n = 20}$	$\mathbf{n = 40}$	$\mathbf{n = 20}$	$\mathbf{n = 40}$	$\mathbf{n = 80}$
Plugin MLE	0.2895	0.1178	0.2512	0.0347	2.1613	0.8909	0.2710	2.2424	0.9142	0.2899
JVHW	0.3222	0.0797	**0.1322**	0.0489	0.6788	0.2699	0.0648	**0.3751**	**0.1755**	0.0974
Averaged Estimator	**0.1382**	0.0723	0.1680	0.0439	**0.5392**	**0.2320**	0.0822	0.5084	0.2539	0.0672

- Plugin MLE: entropy of MLE estimate of distribution
- JVHW: rate-optimal minimax estimator of (Jiao, Venkat, Han, Weissman 15)

Summary

- Analyzed **minimax estimation** as a computationally challenging **zero-sum statistical game**
- Developed algorithmic approaches for solving this challenging zero-sum statistical game: “learning to learn”
- Showed that efficient computation of algorithmic estimator possible given access to two well-studied oracles (finite-dim. maximization and Bayes Estimation)
- Can further leverage natural invariances in statistical game to provide scalable implementations of oracles, and hence algorithmic minimax estimator
- Open Questions:
  - Other constraints on statistical games so that oracles are tractable?
  - Would allowing for constant factor approximations allow for easier oracles?