Regularization, Perturbations and Stability

Ambuj Tewari

Department of Statistics, and
Department of EECS,
University of Michigan, Ann Arbor

May 3, 2018

(based on papers with Jake Abernethy, Chansoo Lee, Zifan Li, Audra McMillan)

Outline

Outline

What is This Talk About?

$$\text{Data} \xrightarrow{\text{randomness}} \text{Perturbed Data} \xrightarrow{\text{optimization}} \text{Prediction Function}$$

Bagging

$$\text{Data} \xrightarrow{\text{Bootstrap}} \text{Bootstrapped Data} \xrightarrow{\text{Training}} \text{Decision Tree}$$

- In bagging, many trees are trained and their predictions averaged at test time

Dropout

$$\text{Data} \xrightarrow{\text{Add noise to features}} \text{Noisy Data} \xrightarrow{\text{Optimization}} \text{Prediction Function}$$

- Blankout noise: For any feature, keep it (rescaling it by $1/p$) with probability p, make it zero otherwise
- For GLMs, blankout noise is same as dropout noise

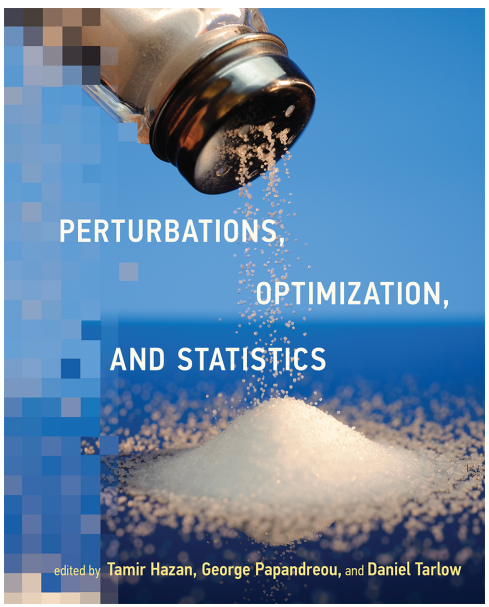
Perturb and Optimize

- Recent approach – still under active development!
- Assume that some underlying optimization problem is tractable
- Achieve regularization by perturbing the data

$$\text{Data} \overset{\text{randomness}}{\longrightarrow} \text{Perturbed Data} \overset{\text{optimization}}{\longrightarrow} \text{Prediction Function}$$

Snapshot of Current State-of-the-art

- NIPS workshops on Perturbations, Optimization, and Statistics in 2012, 2013 and 2014
- *Perturbations, Optimization, and Statistics*, Tamir Hazan, George Papandreou, and Daniel Tarlow, eds., MIT Press, 2016



The Gumbel Lemma

- Consider a finite space $\{1, \ldots, N\}$
- Given θ_i, want to draw $i \in [N]$ with probability $\propto e^{-\theta_i}$
- Let Z_i be iid draws from (min-stable) Gumbel with density

$$f(z) = \exp(-(-z + e^z))$$

- Minimizing the perturbed values $\tilde{\theta}_i = \theta_i + Z_i$ generates the correct probability distribution!
- That is,

$$\mathbb{P}\left(\operatorname*{argmin}_i \ (\theta_i + Z_i) = i_0\right) = \frac{e^{-\theta_{i_0}}}{\sum_i e^{-\theta_i}}$$

The Gumbel Lemma (max version)

- We want to draw $i \in [n]$ with probability $\propto e^{\theta_i}$
- Let Z_i be iid draws from (max-stable) Gumbel with density

$$f(z) = \exp(-(z + e^{-z}))$$

- Maximizing the perturbed values $\tilde{\theta}_i = \theta_i + Z_i$ generates the correct probability distribution!
- That is,

$$\mathbb{P}\left(\operatorname*{argmax}_i \left(\theta_i + Z_i\right) = i_0\right) = \frac{e^{\theta_{i_0}}}{\sum_i e^{\theta_i}}$$

Outline

Online Learning: A Simple Case

- Assume losses are bounded in $[0, 1]$
- At time $t = 1, \ldots, T$
  Learner plays $i_t \in \{1, \ldots, N\}$
  Nature plays $\ell_t \in [0, 1]^N$
  Learner suffers ℓ_{t,i_t}
- Performance of learning algorithm is measured by regret:

$$\sum_{t=1}^{T} \ell_{t,i_t} - \min_{i=1}^{N} \sum_{t=1}^{T} \ell_{t,i}$$

Hannan's Theorem

- A learning algorithm is Hannan consistent if its regret is $o(T)$
- Hannan was the first to propose a Hannan consistent algorithm when the learner has finitely many options

APPROXIMATION TO BAYES RISK IN REPEATED PLAY

James Hannan[1]

SUMMARY

This paper is concerned with the development of a dynamic theory of decision under uncertainty. The results obtained are directly applicable to the development of a dynamic theory of games in which at least one player is, at each stage, fully informed on the joint empirical distribution of the past choices of strategies of the rest. Since the decision problem can be imbedded in a sufficiently unspecified game theoretic model, the paper is written in the language and notation of the general two person game, in which, however, player I's motivation is completely unspecified.

Sections 2 - 7 consider a sequence game based on N successive plays of the same m by n game and culminate in Theorem 4 which exhibits a usable sequence-strategy for II, consisting in the use at the (k+1)-st play of a strategy Bayes against the perturbation of I's cumulative past choice by the addition of $[3n^2/2m]^{1/2}k^{1/2}z$, with z chosen at random from the unit m-cube.

Don't Just Follow the Leader!

- ERM is a good learning algorithm in the iid setting
- Thus it might be tempting to study the following algorithm

$$i_t = \operatorname*{argmin}_{i=1}^{N} \sum_{s=1}^{t-1} \ell_{s,i}$$

- This is called follow the leader (FTL) (aka fictitious play)
- Counterexample showing linear regret even with $N = 2$

$$\begin{array}{ccccccc} 0.5 & 0 & 1 & 0 & 1 & \dots \\ 0 & 1 & 0 & 1 & 0 & \dots \end{array}$$

Follow the Perturbed Leader!

- Hannan's idea was to stabilize FTL by adding random perturbations

$$i_t = \operatorname*{argmin}_{i=1}^{N} \sum_{s=1}^{t-1} \ell_{s,i} + \frac{Z_i}{\eta}$$

- This is called follow the perturbed leader (FTPL) (aka stochastic fictitious play)
- Hannan chose Z_i's to be independent draws from the uniform distribution over $[0, 1]$
- Doing and tuning η gives optimal $O(\sqrt{T})$ dependence on T
- However, dependence on N isn't optimal

Exponential Weights Algorithm (EWA)

- EWA maintains weights $w_{t,i}$
- Weights are initialized to 1 (i.e., $\forall i,\ w_{0,i} = 1$)
- Draws i_t with probability $\propto w_{t-1,i}$
- Updates $w_{t,i} = w_{t-1,i} \cdot \exp(-\eta \ell_{t,i})$
- EWA achieves $O(\sqrt{T \log N})$ regret which is optimal

EWA as FTPL

- Recall FTPL

$$i_t = \operatorname*{argmin}_{i=1}^{N} \sum_{s=1}^{t-1} \ell_{s,i} + \frac{Z_i}{\eta}$$

- Turns out we can recover EWA by drawing Z_i from the Gumbel distribution!
- Follows from Gumbel lemma: we will draw i proportional to $\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{s,i}\right)$
- EWA weight $w_{t-1,i} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{s,i}\right)$

EWA as FTRL

- Let p_t denotes the algorithm's probability distribution over the N choices
- Follow the regularized leader (FTRL) is defined as:

$$p_t = \operatorname*{argmin}_p \sum_{s=1}^{t-1} p^\top \ell_s + \frac{R(p)}{\eta}$$

- R is a regularizer enforcing stability of updates
- Choose $R(p) = -\mathrm{Ent}(p) = \sum_i p_i \log p_i$ also yields EWA!

Relationship between FTPL and FTRL

- We know FTPL with Gumbel perturbation = FTRL with negative entropy regularization
- Is there a deeper connection between FTPL and FTRL?
- For a given FTPL, can I find an R to get an equivalent FTRL?
- For a given FTRL, can I find a perturbation to get an equivalent FTPL?

Relationship between FTPL and FTRL

- We know FTPL with Gumbel perturbation = FTRL with negative entropy regularization
- Is there a deeper connection between FTPL and FTRL?
  Yes! By thinking of them as smoothings of non-smooth functions.
- For a given FTPL, can I find an R to get an equivalent FTRL?
  Yes! R can be found via Fenchel duality
- For a given FTRL, can I find a perturbation to get an equivalent FTPL?
  Not always!

Smoothing by Adding Regularization

- Consider the non-smooth function

$$\Phi(G) = \max_i G_i = \max_p p^\top G$$

- Adding a strictly convex regularizer R ensures smoothness:

$$R^\star(G) = \max_p p^\top G - R(p)$$

- In particular when $R(p) = -\mathrm{Ent}(p)$, we get

$$R^\star(G) = \log \sum_i e^{G_i} \quad , \quad [\nabla R^\star(G)]_i = \frac{e^{G_i}}{\sum_i e^{G_i}}$$

which is smooth

Smoothing by Adding Perturbations

- Consider the non-smooth function

$$\Phi(G) = \max_i G_i = \max_p p^\top G$$

- Add perturbations $Z = (Z_1, \ldots, Z_N)^\top$:

$$\widetilde{\Phi}(G) = \mathbb{E}\left[\phi(G + Z)\right]$$

- $\widetilde{\Phi}$ will be smooth, e.g., if Z_i's are iid from a distribution with a "nice" density
- By duality, there is some function $\widetilde{\Phi}^\star$ such that:

$$\widetilde{\Phi}(G) = \max_p p^\top G - \widetilde{\Phi}^\star(p)$$

which is the implicit regularizer being used

Outline

Follow the Sampled Leader

- Recall FTL:

$$\operatorname*{argmax}_{i=1}^{N} \sum_{s=1}^{t-1} \ell_{s,i}$$

- Consider the Rademacher (ϵ_s is ± 1 each with prob. 0.5) perturbation

$$\sum_{s=1}^{t-1} \epsilon_s \ell_{s,i}$$

- FTPL with this perturbation:

$$\operatorname*{argmin}_{i=1}^{N} \sum_{s=1}^{t-1} (1+\epsilon_s)\ell_{s,i} = \operatorname*{argmin}_{i=1}^{N} \sum_{s:\epsilon_s=+1}^{t-1} \ell_{s,i}$$

  is playing FTL on a subsample of the past!
- This algorithm is also known as sampled fictitious play

Hannan Consistency of FTSL

Theorem (Li & Tewari, 2016)

FTSL with Bernoulli sampling = FTPL with Rademacher perturbations is Hannan consistent.

- Just like Hannan's original algorithm, ours achieves optimal dependence on T but not on N
- There are no tuning parameters!
- Computational advantage for large sets of learner's choices: only need FTL blackbox

Why Does FTL Fail?

- Recall FTL counterexample: major reason for bad behavior is instability
- Consider the time indexed process $L_{t,1} - L_{t,2}$ corresponding to the cumulative losses of the two choices

$\ell_{t,1}$	0.5	0	1	0	1	$\ldots$
$\ell_{t,2}$	0	1	0	1	0	$\ldots$
$L_{t,1}$	0.5	0.5	1.5	1.5	2.5	$\ldots$
$L_{t,2}$	0	1	1	2	2	$\ldots$
$L_{t,2} - L_{t,1}$	-0.5	0.5	-0.5	0.5	-0.5	$\ldots$

- This process crosses zero $\Omega(T)$ times

Why does FTSL work?

Theorem (Littlewood-Offord 1943, Erdös, 1945)

Let $x_1, \ldots, x_T$ be such that $|x_i| \geq 1$. For any given radius $\Delta > 0$, the small ball probability satisfies

$$\sup_B P\left(\epsilon_1 x_1 + \ldots + \epsilon_T x_T \in B\right) \leq C \cdot \frac{\lfloor \Delta \rfloor + 1}{\sqrt{T}}$$

where B ranges over all intervals of length 2Δ and $C < 3$ is a universal constant.

- Consider the perturbed losses $\tilde{\ell}_{t,i} = (1 + \epsilon_t)\ell_{t,i}$
- And the corresponding cumulative sums $\tilde{L}_{t,i} = \sum_{s=1}^{t} \tilde{\ell}_{s,i}$
- The process $\tilde{L}_{t,2} - \tilde{L}_{t,1}$ does not cross zero too many times
- In fact, in expectation, it crosses zero only $O(\sqrt{T})$ times

Outline

Differential Privacy (DP)

- Suppose X and X' are two datasets that are "close"
- $\mathcal{A}$ is some randomized algorithm that operates on datasets
- $\mathcal{A}$ is (ϵ, δ)-differentially private iff

$$\mathbb{P}\left[\mathcal{A}(X) \in S\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}(X') \in S\right] + \delta$$

  for any subset S of the range of $\mathcal{A}$
- At its heart, DP is a stability notion
- Common DP mechanisms: add Laplace or Gaussian perturbations

Post Processing Immunity

- Suppose $\mathcal{A}$ is (ϵ, δ)-DP
- Let f be an arbitrary function
- Then $f \circ \mathcal{A}$ is (ϵ, δ)-DP

$$\text{Data} \xrightarrow{\text{DP mechanism}} \text{Perturbed Data} \xrightarrow{\text{optimization}} \text{Processed Data}$$

- DP techniques give us powerful tools to reason about the pipeline above
- We can systematically derive FTPL regret bounds (including new ones!) using this idea

Summary

- Regularization via random perturbations: feed randomly perturbed data into optimization routine
- In online learning, FTPL-FTRL can viewed as two different ways to smooth a non-smooth function
- For experts problem, Rademacher perturbations lead to a Hannan consistent algorithm
- Differential privacy offers a lens to examine the stability of FTPL algorithms

Thank You!

References

- Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Perturbation techniques in online learning and optimization. In *Perturbations, Optimization, and Statistics*, Neural Information Processing Series, chapter 8. MIT Press, 2016.
  Stochastic smoothing view of FTPL-FTRL
- Zifan Li and Ambuj Tewari. Sampled Fictitious Play is Hannan Consistent. *Games and Economic Behavior*, 2018.
  Rademacher perturbations
- Jacob Abernethy, Chansoo Lee, Audra McMillan, and Ambuj Tewari. Online Linear Optimization through the Differential Privacy Lens, 2018. *under review*, available upon request.
  DP techniques to derive FTPL guarantees