Advanced Machine Learning

Follow-The-Perturbed Leader

MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

General Ideas

- Linear loss: decomposition as a sum along substructures.
  - sum of edge losses in a tree.
  - sum of edge losses along a path.
  - sum of other substructures losses in a discrete problem.
  - includes expert setting.

FPL

(Kalai and Vempala, 2004)

- General linear decision problem:
  - player selects $\mathbf{w}_t \in \mathcal{W} \subseteq \mathbb{R}^N$, $l_1\text{-diam}(\mathcal{W}) \leq W_1$.
  - player receives $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^N$, $\mathcal{X} \subseteq \{\mathbf{x} \colon \|\mathbf{x}\|_1 \leq X_1\}$.
  - player incurs loss $\mathbf{w}_t \cdot \mathbf{x}_t$, $\sup_{\mathbf{w} \in \mathcal{W}, \mathbf{x} \in \mathcal{X}} |\mathbf{w} \cdot \mathbf{x}| \leq R$.
- Objective: minimize cumulative loss or regret.
- Notation: $M(\mathbf{x}) = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \mathbf{w} \cdot \mathbf{x}$.

FL

- Follow the Leader (FL): use M at every round (aka fictitious play).

- FL problem: Suppose $N = 2$ and consider a sequence starting with $\binom{0}{1/2}$ and then alternating $\binom{1}{0}$ and $\binom{0}{1}$. Then,
  - FL incurs loss 1 at every round, T overall.
  - any single expert incurs loss $T/2$ overall.

FPL Algorithms

(Hannan 1957; Kalai and Vempala, 2004)

- Additive bound Follow the Perturbed Leader (FPL):
  - $\mathsf{p}_t \sim U([0, 1/\epsilon]^N)$.
  - $\mathbf{w}_t = \text{argmin}_{\mathbf{w} \in \mathcal{W}} \sum_{s=1}^{t-1} \mathbf{w} \cdot \mathbf{x}_s + \mathbf{w} \cdot \mathsf{p}_t$
    $= M(\mathbf{x}_{1:t-1} + \mathsf{p}_t)$.
- Multiplicative bound Follow the Perturbed Leader (FPL*):
  - $\mathsf{p}_t \sim$ Laplacian with density $f(\mathbf{x}) = \frac{\epsilon}{2} e^{-\epsilon \|\mathbf{x}\|_1}$.
  - $\mathbf{w}_t = \text{argmin}_{\mathbf{w} \in \mathcal{W}} \sum_{s=1}^{t-1} \mathbf{w} \cdot \mathbf{x}_s + \mathbf{w} \cdot \mathsf{p}_t$
    $= M(\mathbf{x}_{1:t-1} + \mathsf{p}_t)$.

FPL - Bound

- Theorem: fix $\epsilon > 0$. Then, the expected cumulative loss of additive FPL(ϵ) is bounded as follows

$$\mathrm{E}[\mathcal{L}_T] \leq \mathcal{L}_T^{\min} + \epsilon R X_1 T + \frac{W_1}{\epsilon}.$$

For $\epsilon = \sqrt{\frac{W_1}{R X_1 T}}$

$$\boxed{\mathrm{E}[\mathcal{L}_T] \leq \mathcal{L}_T^{\min} + 2\sqrt{X_1 W_1 R T}.}$$

FPL* - Bound

- Theorem: fix $\epsilon > 0$ and assume that $\mathcal{W}, \mathcal{X} \subseteq \mathbb{R}_+^N$. Then, the expected cumulative loss of (multiplicative) FPL*($\epsilon/2X_1$) is bounded as follows

$$\mathrm{E}[\mathcal{L}_T] \leq (1+\epsilon)\mathcal{L}_T^{\min} + \frac{2X_1W_1(1+\log N)}{\epsilon}.$$

For $\epsilon = \min\left(1/2X_1, \sqrt{W_1(1+\log N)/X_1\mathcal{L}_T^{\min}}\right)$

$$\mathrm{E}[\mathcal{L}_T] \leq \mathcal{L}_T^{\min} + 4\sqrt{\mathcal{L}_T^{\min}X_1W_1(1+\log N)} + 4X_1W_1(1+\log N).$$

Proof Outline

- Be the perturbed leader (BPL): $\mathbf{w}_t = M(\mathbf{x}_{1:t} + \mathsf{p}_t)$.

1. Bound on regret of BPL: $\mathrm{E}[R_T(\mathrm{BPL})] \leq \frac{W_1}{\epsilon}$.
2. Bound on difference of regrets of FPL and BPL:
$$\mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] - \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t].$$
3. Difference of expectations small because similar distributions.

Proof: BL Regret

- Lemma 1: $\sum_{t=1}^{T} M(\mathbf{x}_{1:t}) \cdot \mathbf{x}_t \leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T}$.
- Proof: case $T = 1$ is clear. By induction,

$$
\begin{aligned}
&\sum_{t=1}^{T+1} M(\mathbf{x}_{1:t}) \cdot \mathbf{x}_t \\
&\leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{T+1} && \text{(induction)} \\
&\leq M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{1:T} + M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{T+1} && \text{(def. of } M(\mathbf{x}_{1:T}) \text{ as minimizer)} \\
&= M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{1:T+1}.
\end{aligned}
$$

Proof: BPL Regret

- **Lemma 2**: let $\mathsf{p}_0 = 0$. Then, the following holds:

$$\sum_{t=1}^{T} M(\mathbf{x}_{1:t} + \mathsf{p}_t) \cdot \mathbf{x}_t \leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + W_1 \sum_{t=1}^{T} \|\mathsf{p}_t - \mathsf{p}_{t-1}\|_\infty.$$

- **Proof**: use Lemma 1 with $\mathbf{x}'_t = \mathbf{x}_t + \mathsf{p}_t - \mathsf{p}_{t-1}$, then

$$\begin{aligned}
\sum_{t=1}^{T} M(\mathbf{x}_{1:t} + \mathsf{p}_t) \cdot (\mathbf{x}_t + \mathsf{p}_t - \mathsf{p}_{t-1}) &\leq M(\mathbf{x}_{1:T} + \mathsf{p}_T) \cdot (\mathbf{x}_{1:T} + \mathsf{p}_T) \\
&\leq M(\mathbf{x}_{1:T}) \cdot (\mathbf{x}_{1:T} + \mathsf{p}_T) \\
&= M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + M(\mathbf{x}_{1:T}) \cdot \sum_{t=1}^{T} \mathsf{p}_t - \mathsf{p}_{t-1}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{t=1}^{T} M(\mathbf{x}_{1:t} + \mathsf{p}_t) \cdot \mathbf{x}_t &\leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + \sum_{t=1}^{T} \left[M(\mathbf{x}_{1:T}) - M(\mathbf{x}_{1:t} + \mathsf{p}_t)\right] \cdot \left[\mathsf{p}_t - \mathsf{p}_{t-1}\right] \\
&\leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + W_1 \sum_{t=1}^{T} \left\|\mathsf{p}_t - \mathsf{p}_{t-1}\right\|_\infty.
\end{aligned}$$

Proof: FPL vs. BPL Regrets

- Proof: for the expected loss, we can just choose $\mathsf{p}_t = \mathsf{p}_1$ for all $t > 0$, which yields:

$$\sum_{t=1}^{T} M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t \leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + W_1 \|\mathsf{p}_1\|_\infty.$$

- Thus,

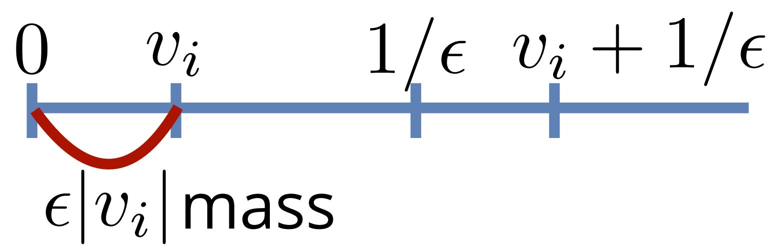
$$\begin{aligned}
&\sum_{t=1}^{T} \mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] \\
&= \sum_{t=1}^{T} \mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] - \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t] + \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t] \\
&\leq \sum_{t=1}^{T} \Big[\mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] - \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t] \Big] + \mathcal{L}_T^{\min} + W_1 \|\mathsf{p}_1\|_\infty.
\end{aligned}$$

Proof: FPL

- By definition of the perturbation, $\|\mathsf{p}_1\|_\infty \leq \frac{1}{\epsilon}$.
- Now, $\mathbf{x}_{1:t} + \mathsf{p}_1$ and $\mathbf{x}_{1:t-1} + \mathsf{p}_1$ both follow a uniform distribution over a cube. Thus,

$$\mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] - \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t] \leq R(1 - \text{fraction of overlap}).$$

- Two cubes $[0, 1/\epsilon]^N$ and $\mathbf{v} + [0, 1/\epsilon]^N$ overlap over at least the fraction $(1 - \epsilon\|\mathbf{v}\|_1)$:
  - if $\mathbf{x} \in [0, 1/\epsilon]^N$ but $\mathbf{x} \notin \mathbf{v} + [0, 1/\epsilon]^N$ then for at least one i, $x_i \notin v_i + [0, 1/\epsilon]^N$, which has probability at most $\epsilon|v_i|$.



Proof: FPL

- Thus,

$$\mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] - \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t] \leq R\epsilon \|\mathbf{x}_t\|_1 \leq R\epsilon X_1.$$

- And,

$$\mathrm{E}[R_T] \leq R\epsilon X_1 T + \frac{W_1}{\epsilon}.$$

Proof: FPL*

- Lemma 3:

$$\mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] \leq e^{\epsilon X_1}\, \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t].$$

- Proof:

$$\begin{aligned}
&\mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] \\
&= \int_{\mathbb{R}^N} M(\mathbf{x}_{1:t-1} + \mathbf{u}) \cdot \mathbf{x}_t \, d\mu(\mathbf{u}) \\
&= \int_{\mathbb{R}^N} M(\mathbf{x}_{1:t} + \mathbf{v}) \cdot \mathbf{x}_t \, d\mu(\mathbf{x}_t + \mathbf{v}) \quad (\text{change of var. } \mathbf{v} = \mathbf{u} + \mathbf{x}_t) \\
&= \int_{\mathbb{R}^N} M(\mathbf{x}_{1:t} + \mathbf{v}) \cdot \mathbf{x}_t \, \underbrace{e^{-\epsilon(\|\mathbf{x}_t + \mathbf{v}\|_1 - \|\mathbf{v}\|_1)}}_{\leq e^{\epsilon X_1}} d(\mathbf{v}) \\
&\leq e^{\epsilon X_1}\, \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t].
\end{aligned}$$

Proof: FPL*

- For $\epsilon \leq 1/X_1$, $e^{\epsilon X_1} \leq (1 + 2\epsilon X_1)$, thus,

$$\sum_{t=1}^{T} \mathrm{E}[M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) \cdot \mathbf{x}_t] \leq \sum_{t=1}^{T} (1 + 2\epsilon X_1) \mathrm{E}[M(\mathbf{x}_{1:t} + \mathsf{p}_1) \cdot \mathbf{x}_t]$$
$$\leq \sum_{t=1}^{T} (1 + 2\epsilon X_1)(\mathcal{L}_T^{\min} + W_1 \mathrm{E}[\|\mathsf{p}_1\|_\infty]).$$

Thus,

$$\begin{aligned}
\mathrm{E}[\|\mathsf{p}_1\|_\infty] = \mathrm{E}\Big[\max_{i \in [1,N]} |p_{1,i}|\Big] &= \int_0^{+\infty} \Pr\Big[\max_{i \in [1,N]} |p_{1,i}| > t\Big]\, dt \\
&\leq 2 \int_0^{+\infty} \Pr\Big[\max_{i \in [1,N]} p_{1,i} > t\Big]\, dt \\
&= 2 \int_0^{u} \Pr\Big[\max_{i \in [1,N]} p_{1,i} > t\Big]\, dt + \int_u^{+\infty} \Pr\Big[\max_{i \in [1,N]} p_{1,i} > t\Big]\, dt \\
&\leq 2u + N \int_u^{+\infty} \Pr\Big[p_{1,1} > t\Big]\, dt \\
&= 2u + N \frac{e^{-\epsilon u}}{\epsilon} \leq \frac{2(1 + \log N)}{\epsilon} \quad \text{(best choice of } u\text{).}
\end{aligned}$$

Expert Setting

- $W_1 = 1, X_1 = N$, and $R = 1$; for FLP*(ϵ) ,

$$\mathrm{E}[\mathcal{L}_T] \leq (1 + 2N\epsilon)\mathcal{L}_T^{\min} + \frac{2(1+\log(N))}{\epsilon}.$$

- More favorable bound:
  - $\mathbf{x}_t \rightarrow x_{t,1}\mathbf{e}_1 \ldots x_{t,N}\mathbf{e}_N$.
  - new $\mathcal{L}_{NT}^{\min}$ = old $\mathcal{L}_T^{\min}$.
  - $\mathrm{E}[\mathcal{L}_T^{\mathrm{old}}] \leq \mathrm{E}[\mathcal{L}_{TN}^{\mathrm{new}}]$.
  - new guarantee: for FLP*(ϵ) ,

$$\mathrm{E}[\mathcal{L}_T] \leq (1 + 2\epsilon)\mathcal{L}_T^{\min} + \frac{2(1+\log(NT))}{\epsilon}.$$

$$\longrightarrow \mathrm{E}[R_T] \leq 2\sqrt{2\mathcal{L}_T^{\min}(1 + \log(NT))}.$$

RWM = FPL

- Let FPL(η) be an instance of the general FPL algorithm with a perturbation defined by

$$\mathsf{p}_1 = \left[\frac{\log(-\log(u_1))}{\eta}, \ldots, \frac{\log(-\log(u_N))}{\eta}\right]^\top,$$

  - where u_j is drawn according to the uniform distribution over $[0, 1]$.

- Then, FPL(η) and RWM(η) coincide.

References

- Nicolò Cesa-Bianchi, Alex Conconi, Claudio Gentile: On the Generalization Ability of On-Line Learning Algorithms. *IEEE Transactions on Information Theory* 50(9): 2050-2057. 2004.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Yoav Freund and Robert Schapire. Large margin classification using the perceptron algorithm. In *Proceedings of COLT 1998*. ACM Press, 1998.
- Adam T. Kalai, Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.* 71(3): 291-307. 2005.
- Nick Littlestone. From On-Line to Batch Learning. COLT 1989: 269-284.
- Nick Littlestone. "Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm" *Machine Learning* 285-318(2). 1988.

References

- Nick Littlestone, Manfred K. Warmuth: The Weighted Majority Algorithm. *FOCS 1989*: 256-261.

- Tom Mitchell. *Machine Learning*, McGraw Hill, 1997.

- Novikoff, A. B. (1962). On convergence proofs on perceptrons. *Symposium on the Mathematical Theory of Automata, 12*, 615-622. Polytechnic Institute of Brooklyn.