CS281B/Stat241B. Statistical Learning Theory. Lecture 10.

Wouter M. Koolen

- The Minimax Algorithm for the Dot-loss Game
- Follow the Perturbed Leader (part 1)

Recall: Dot-loss game

Protocol:

- For $t = 1, 2, \ldots$
  - Learner chooses a distribution $\boldsymbol{w}_t$ on K “experts”.
  - Adversary reveals loss vector $\boldsymbol{\ell}_t \in [0, 1]^K$.
  - Learner’s loss is the **dot loss** $\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t$

Definition: *Regret* after T rounds:

$$R_T \;=\; \sum_{t=1}^{T} \boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \min_k \sum_{t=1}^{T} \ell_{t,k}$$

Goal: design an algorithm for Learner that guarantees low regret.

Today we take this goal seriously

Minimax regret with a *time horizon*:

$$\min_{\boldsymbol{w}_1} \max_{\boldsymbol{\ell}_1} \cdots \min_{\boldsymbol{w}_T} \max_{\boldsymbol{\ell}_T} R_T$$

turns out to be messy.

However, minimax regret with a *loss budget B of the best expert*:

$$\min_{\boldsymbol{w}_1} \max_{\boldsymbol{\ell}_1} \min_{\boldsymbol{w}_2} \max_{\boldsymbol{\ell}_2} \cdots \begin{cases} \sup_{T \geq 0} R_T & \text{if } \min_k L_T^k \leq B \text{ for all } T \geq 0 \\ -\infty & \text{o.w.} \end{cases}$$

results in an extremely elegant algorithm.

Sneak peek

The optimal algorithm:

Let $\boldsymbol{L}$ denote the current expert loss vector.

Start from $\boldsymbol{L}$. Repeatedly add uniformly drawn unit loss $\boldsymbol{\ell} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$. Play the last expert that goes over the budget B.

Menu

To keep things tractable we restrict to expert losses $0/1$.

We will perform the analysis in two stages:

1. Units $\boldsymbol{\ell} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$
2. Binary $\boldsymbol{\ell} \in \{0, 1\}^K$

Extending to arbitrary $[0, 1]$ *losses could be part of your project.*

Straightforward observations

A *dead* expert has loss $> B$.

Adversary may gratuitously assign max. loss 1 to dead experts.

Learner cannot benefit by putting weight on a dead expert.

Adversary cannot benefit by keeping the best expert loss $< B$.

So we might as well maximise Learners loss.

Backward induction

Let $V_B(\boldsymbol{L})$ be the amount of loss Adversary can inflict on the Learner

- from a starting point where experts have loss $\boldsymbol{L}$
- with the loss of the best expert at most the budget B.

Base case:

$$V_B(\boldsymbol{L}) \;=\; 0 \quad \text{if} \quad \min_k L_k > B$$

Recurrence:

$$V_B(\boldsymbol{L}) \;=\; \inf_{\boldsymbol{w}} \sup_{\boldsymbol{\ell}} \big\{ \boldsymbol{w}^\intercal \boldsymbol{\ell} + V_B(\boldsymbol{L} + \boldsymbol{\ell}) \big\}$$

where $\boldsymbol{w}$ ranges over distributions on live experts $\{k \mid L_k \leq B\}$ and $\boldsymbol{\ell} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$.

Units: main claim

By minimax theorem (Von Neumann)

$$\begin{aligned} V_B(\boldsymbol{L}) &= \inf_{\boldsymbol{w}} \sup_{\boldsymbol{\ell}} \boldsymbol{w}^\intercal \boldsymbol{\ell} + V_B(\boldsymbol{L} + \boldsymbol{\ell}) \\ &= \sup_{\boldsymbol{p}} \inf_{\boldsymbol{w}} \mathop{\mathbb{E}}_{k \sim \boldsymbol{p}} \left[\boldsymbol{w}^\intercal \boldsymbol{e}_k + V_B(\boldsymbol{L} + \boldsymbol{e}_k)\right] \end{aligned}$$

Both players have *equaliser* strategies:

- For $\boldsymbol{p}$, the equaliser is *uniform*: $p_k = \frac{1}{K}$.
- For $\boldsymbol{w}$, an equaliser satisfies:

$$w_k + V_B(\boldsymbol{L} + \boldsymbol{e}_k) \quad \text{is const in } k$$

solving for $\boldsymbol{w}$ (with $\sum_k w_k = 1$) results in

$$w_k = \frac{1 + \sum_j V_B(\boldsymbol{L} + \boldsymbol{e}_j)}{K} - V_B(\boldsymbol{L} + \boldsymbol{e}_k) \qquad \text{check} \geq 0$$

Value of the game

We get:

$$V_B(\boldsymbol{L}) \;=\; \frac{1}{K} + \frac{1}{K}\sum_{k=1}^{K} V_B(\boldsymbol{L}+\boldsymbol{e}_k),$$

showing that $V_B(\boldsymbol{L})$ is $\frac{1}{K}$ times the expected length of the game.

Optimal weights

Let $w_k\left[\boldsymbol{L}\right]$ denote the weight assigned to expert k in state $\boldsymbol{L}$. The value expression allows us to rewrite

$$\begin{aligned} w_k &= V_B(\boldsymbol{L}) - V_B(\boldsymbol{L} + \boldsymbol{e}_k) \\ &= \frac{1}{K} \sum_{j=1}^{K} \big(V_B(\boldsymbol{L} + \boldsymbol{e}_j) - V_B(\boldsymbol{L} + \boldsymbol{e}_j + \boldsymbol{e}_k) \big) \\ &= \frac{1}{K} \sum_{j=1}^{K} w_k\left[\boldsymbol{L} + \boldsymbol{e}_j\right] \end{aligned}$$

Idea: to sample $k \sim \boldsymbol{w}$, we may unroll this definition until we hit the base case of 1 surviving expert, where $V_B(\boldsymbol{L}) = 1$ and $V_B(\boldsymbol{L} + \boldsymbol{e}_k) = 0$.

Binary losses: monotonicity

Units $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ separately:

$$w_k\,[\boldsymbol{L}]^\intercal\,\boldsymbol{e}_1 + w_k\,[\boldsymbol{L}+\boldsymbol{e}_1]^\intercal\,\boldsymbol{e}_2$$

Combined $\boldsymbol{e}_1 + \boldsymbol{e}_2$:

$$w_k\,[\boldsymbol{L}]^\intercal\,(\boldsymbol{e}_1 + \boldsymbol{e}_2)$$

Which is bigger? Claim: separate. I.e.:

$$w_k\,[\boldsymbol{L}+\boldsymbol{e}_1]^\intercal\,\boldsymbol{e}_2 \;>\; w_k\,[\boldsymbol{L}]^\intercal\,\boldsymbol{e}_2$$

Every path in which expert 2 is the survivor from $\boldsymbol{L}$ also works from $\boldsymbol{L}+\boldsymbol{e}_1$. But there are more such paths.

Equalisation

The minimax algorithm for *mix loss* equalises the regret over all loss sequences in which all but one expert suffer infinite loss.

The minimax algorithm for *dot loss* equalises the regret over all sequences of unit losses.

Follow the Perturbed Leader

Follow-the-Leader is an intuitive algorithm. But its regret is horrible (Homework).

The reason is that FTL is overly sensitive to small loss differences.

In this lecture we see how FTL can be fixed by adding a pinch of randomness.

And we see that the solution extends to combinatorial prediction tasks (next lecture).

Follow the Perturbed Leader

The cumulative loss after t rounds: $\boldsymbol{L}_t = \boldsymbol{\ell}_1 + \ldots + \boldsymbol{\ell}_t$.

Definition: Let X_t^k be random. FPL with learning rate η plays in round t by choosing expert

$$\arg\min_k L_{t-1}^k + \frac{X_t^k}{\eta}$$

Question: how to choose the distribution of the perturbations X_t^k so that FPL guarantees low regret (*in expectation/with high probability*)?

We use i.i.d. negative-of-exponential distribution:

$$p(X_t^k = x) \;=\; e^x \qquad \text{for } x \leq 0.$$

FPL loss decomposition

In the Hedge analysis we decomposed dot loss in terms of *mix loss* and *mixability gap*.

Here we use the loss of *Infeasible Follow the Perturbed Leader*, which plays the leader *after* the upcoming loss.

$$\mathbb{E}\, L_T^{\text{FPL}} \;=\; \underbrace{\mathbb{E}\, L_T^{\text{IFPL}}}_{\substack{\text{close to best}\\ \text{for high } \eta}} + \underbrace{\mathbb{E}\, L_T^{\text{FPL}} - \mathbb{E}\, L_T^{\text{IFPL}}}_{\substack{\text{small}\\ \text{for low } \eta}}$$

IFPL close to best expert

Theorem: After $T \geq 0$ rounds:

$$\mathbb{E}\, L_T^{\text{IFPL}} \;\leq\; \min_k L_T^k + \frac{\ln K}{\eta}$$

We use the abbreviation $M(\boldsymbol{v}) := \boldsymbol{e}_{\arg\min_k v_k}$. So IFPL plays $M\left(\boldsymbol{L}_t + \frac{\boldsymbol{X}}{\eta}\right)$ in round t.

We first prove (result akin to telescoping for Hedge):

$$M\left(\frac{\boldsymbol{X}}{\eta}\right)^\intercal \frac{\boldsymbol{X}}{\eta} + \sum_{t=1}^T M\left(\boldsymbol{L}_t + \frac{\boldsymbol{X}}{\eta}\right)^\intercal \boldsymbol{\ell}_t \;\leq\; M\left(\boldsymbol{L}_T + \frac{\boldsymbol{X}}{\eta}\right)^\intercal \left(\boldsymbol{L}_T + \frac{\boldsymbol{X}}{\eta}\right)$$

By induction. Base case $T = 0$ holds by definition. For $T \geq 1$, we need to show:

$$M\left(\boldsymbol{L}_{T-1}+\frac{\boldsymbol{X}}{\eta}\right)^{\top}\left(\boldsymbol{L}_{T-1}+\frac{\boldsymbol{X}}{\eta}\right)+M\left(\boldsymbol{L}_{T}+\frac{\boldsymbol{X}}{\eta}\right)^{\top}\boldsymbol{\ell}_{T}$$
$$\leq\; M\left(\boldsymbol{L}_{T}+\frac{\boldsymbol{X}}{\eta}\right)^{\top}\left(\boldsymbol{L}_{T}+\frac{\boldsymbol{X}}{\eta}\right)$$

that is

$$M\left(\boldsymbol{L}_{T-1}+\frac{\boldsymbol{X}}{\eta}\right)^{\top}\left(\boldsymbol{L}_{T-1}+\frac{\boldsymbol{X}}{\eta}\right)$$
$$\leq\; M\left(\boldsymbol{L}_{T}+\frac{\boldsymbol{X}}{\eta}\right)^{\top}\left(\boldsymbol{L}_{T-1}+\frac{\boldsymbol{X}}{\eta}\right)$$

which follows from the definition of M.

Bringing the “round 0” term to the other side. The IFPL loss is at most

$$\sum_{t=1}^{T} M\left(\boldsymbol{L}_{t}+\frac{\boldsymbol{X}}{\eta}\right)^{\top}\boldsymbol{\ell}_{t}\;\leq\; M\left(\boldsymbol{L}_{T}+\frac{\boldsymbol{X}}{\eta}\right)^{\top}\left(\boldsymbol{L}_{T}+\frac{\boldsymbol{X}}{\eta}\right)-M\left(\frac{\boldsymbol{X}}{\eta}\right)^{\top}\frac{\boldsymbol{X}}{\eta}$$

We then use that $\boldsymbol{X} \leq 0$ to drop the middle perturbations, and observe that

$$-M\left(\frac{\boldsymbol{X}}{\eta}\right)^{\intercal}\frac{\boldsymbol{X}}{\eta} \;=\; \frac{\max_k -X_k}{\eta}$$

The expected maximum of K standard exponentials is $\leq 1 + \ln K$.

FPL close to IFPL

Theorem: In each round t:

$$\mathbb{E}\,\ell_t^{\text{FPL}} - \mathbb{E}\,\ell_t^{\text{IFPL}} \;\leq\; \eta$$

(Per-round bound, like mixability gap bound in Hedge analysis)

Crucial idea: Bound the maximal change in probability of choosing expert i under addition of one trial of losses:

$$\mathbb{P}\left(I_t^{\text{FPL}} = i\right) \;\leq\; e^{\eta}\,\mathbb{P}\left(I_t^{\text{IFPL}} = i\right)$$

(tedious but straightforward manipulation of exponential distributions)

So

$$\mathbb{E}\,\ell_t^{\text{FPL}} \;\leq\; e^{\eta}\,\mathbb{E}\,\ell_t^{\text{IFPL}}$$

And hence, using $e^{-\eta} \geq 1 - \eta$ and $\ell \in [0,1]$,

$$(1-\eta)\,\mathbb{E}\,\ell_t^{\mathrm{FPL}} \;\leq\; \mathbb{E}\,\ell_t^{\mathrm{IFPL}} \qquad \text{so that} \qquad \mathbb{E}\,\ell_t^{\mathrm{FPL}} - \mathbb{E}\,\ell_t^{\mathrm{IFPL}} \;\leq\; \eta.$$

Tuning FPL

Theorem: FPL with $\eta = \sqrt{\frac{1+\ln K}{T}}$ guarantees

$$\mathbb{E}\, R_T^{\text{FPL}} \;\leq\; 2\sqrt{T(1+\ln K)}$$

Constants not as good as tuned Hedge. This can be fixed by changing the perturbation distribution (homework).

FPL extends to combinatorial prediction spaces (next lecture).