CS281B/Stat241B. Statistical Learning Theory.
Lecture 11.

Wouter M. Koolen

- Follow the Perturbed Leader (part 2)
- Adaptive Regret and Tracking

Follow the Perturbed Leader

Today we look at *combinatorial* prediction tasks.

Sets	committee formation, advertising
Trees	spanning trees (networking), parse trees
Paths (source-sink)	route planning
Permutations	ordering

Crucial assumption: loss is linear

$$\text{Loss of a } \underbrace{\begin{cases}\text{set}\\ \text{tree}\\ \text{path}\\ \text{permutation}\\ \dots\end{cases}}_{\text{concepts}} \text{ is the \textit{sum} of the losses of its } \underbrace{\begin{cases}\text{elements}\\ \text{edges}\\ \text{edges}\\ \text{assignments}\\ \dots\end{cases}}_{\text{components}}$$

Represent *concept* as indicator $C \in \{0,1\}^d$ out of d components.

Combinatorial dot-loss game

Concept class: $\mathcal{C} = \{C_1, \ldots, C_D\} \subseteq \{0,1\}^d$.

Protocol:

- For $t = 1, 2, \ldots$
  - Learner chooses a distribution W_t on concepts $\mathcal{C}$.
  - Adversary reveals component loss vector $\boldsymbol{\ell}_t \in [0,1]^d$.
  - Learner incurs the dot loss $\mathbb{E}_{C \sim W_t}\left[C^\intercal \boldsymbol{\ell}_t\right]$.

Typically D is large, so spelling out $W_t = (w_1, \ldots, w_D)$ is intractable.

We allow Learner to randomise and analyse loss in expectation.

Expanded vs Collapsed

Expanded: perturb the loss of each **concept**, then pick best concept.
Analysis immediate from experts case, but intractable algorithm.

Collapsed: perturb the loss of each **component**, then pick best concept.

Follow the Perturbed Leader (Concept)

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

Definition: Let $X_t^1, \ldots, X_t^d$ be random. FPL with learning rate η plays in round t by choosing concept

$$\arg\min_{C \in \mathcal{C}} C^\intercal \left(\boldsymbol{L}_{t-1} + \frac{\boldsymbol{X}_t}{\eta} \right)$$

We have special-purpose linear optimisation algorithms:

- Sets: linear-time median
- Minimum spanning tree
- Shortest path
- Maximal weighted matching

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 concept

We use the abbreviation $M(\boldsymbol{v}) := \arg\min_{C \in \mathcal{C}} C^\intercal \boldsymbol{v}$. So IFPL plays $M\left(\boldsymbol{L}_t + \frac{\boldsymbol{X}}{\eta}\right)$ in round t.

Theorem: After $T \geq 0$ rounds:

$$\mathbb{E}\, L_T^{\text{IFPL}} \;\leq\; \min_{C \in \mathcal{C}} C^\intercal \boldsymbol{L}_T + \frac{U(1 + \ln d)}{\eta}$$

where $\mathcal{C} \subseteq \{0,1\}^d$ and $U = \max_{C \in \mathcal{C}} |C|_1$.

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:

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

that is

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

which follows from the definition of M.

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

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

We then use

$$M\left(\boldsymbol{L}_T + \frac{\boldsymbol{X}}{\eta}\right)^{\intercal}\left(\boldsymbol{L}_T + \frac{\boldsymbol{X}}{\eta}\right) \;\le\; M\left(\boldsymbol{L}_T\right)^{\intercal}\left(\boldsymbol{L}_T + \frac{\boldsymbol{X}}{\eta}\right)$$

$$= M\left(\boldsymbol{L}_T\right)^{\intercal}\boldsymbol{L}_T + \frac{1}{\eta}\underbrace{M\left(\boldsymbol{L}_T\right)^{\intercal}\boldsymbol{X}}_{\le 0 \text{ since } \boldsymbol{X} \le 0}.$$

We then continue to observe that

$$-M\left(\frac{\boldsymbol{X}}{\eta}\right)^{\intercal}\frac{\boldsymbol{X}}{\eta} \;\le\; \frac{1}{\eta}\left|M\left(\frac{\boldsymbol{X}}{\eta}\right)\right|_1 |\boldsymbol{X}|_\infty$$

$$= \frac{U\,|\boldsymbol{X}|_\infty}{\eta}$$

The expected maximum of d standard exponentials is $\le 1 + \ln d$.

FPL close to IFPL

Theorem: In each round t:

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

(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:

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

(tedious but straightforward manipulation of exponential distributions)

In the combinatorial concepts case we use $|\boldsymbol{\ell}|_1 \leq d$ to obtain

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

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

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

Tuning FPL

We proved

$$\mathbb{E}\, R_T^{\text{FPL}} \;\le\; T d U \eta + \frac{U(1+\ln d)}{\eta}$$

Theorem: FPL with $\eta = \sqrt{\frac{(1+\ln d)}{dT}}$ guarantees

$$\mathbb{E}\, R_T^{\text{FPL}} \;\le\; 2U\sqrt{Td(1+\ln d)}$$

Part 2: Adaptive Regret

Motivation: non-stationary data

Suppose the data are like this

	$T/2$ rounds	$T/2$ rounds
expert 1	loss 0	loss 1
expert 2	loss 1	loss 0

We want to be as good as expert 2 on the second half of the data.

The Aggregating Algorithm and Hedge do *not* accomplish this. They incur loss $\approx T/2$, not ≈ 0, on second half.

Diagnosis: Expert must be ahead in *cumulative* loss to receive substantial weight.

Recap: Mix-loss game

Protocol:

- For $t = 1, 2, \ldots$
  - Learner chooses a distribution $\boldsymbol{w}_t$ on K experts.
  - Adversary reveals loss vector $\boldsymbol{\ell}_t \in (-\infty, \infty]^K$.
  - Learner incurs the mix loss $-\ln\left(\sum_{k=1}^{K} w_{t,k} e^{-\ell_{t,k}}\right)$

New objective

Definition: The *adaptive regret* on time interval $[t_1, t_2]$ is given by

$$R_{[t_1,t_2]} = \sum_{t=t_1}^{t_2} \underbrace{-\ln\left(\sum_{k=1}^{K} w_t^k e^{-\ell_t^k}\right)}_{\text{Learner's mix loss in round } t} - \underbrace{\min_k \sum_{t=t_1}^{t_2} \ell_t^k}_{\text{best loss for interval}}$$

Goal: guarantee low adaptive regret on *any interval*.

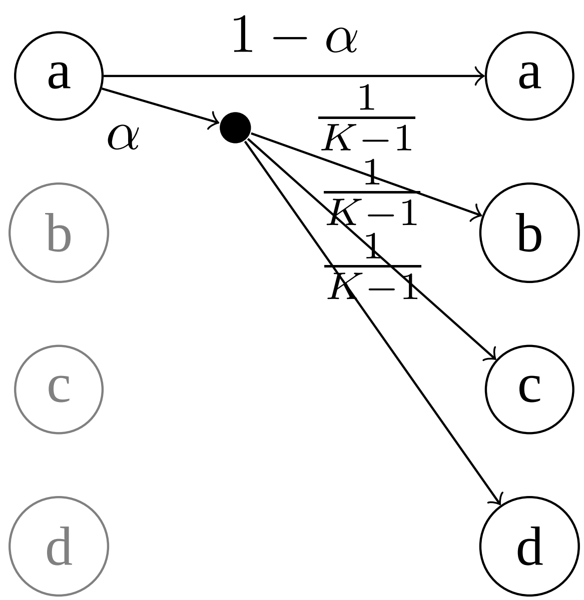
The Fixed Share Algorithm

Definition: *Fixed Share* with *switching rate sequence* $\alpha_2, \alpha_3, \ldots$ plays uniform $w_1^k = 1/K$ in round 1, and updates its weights as

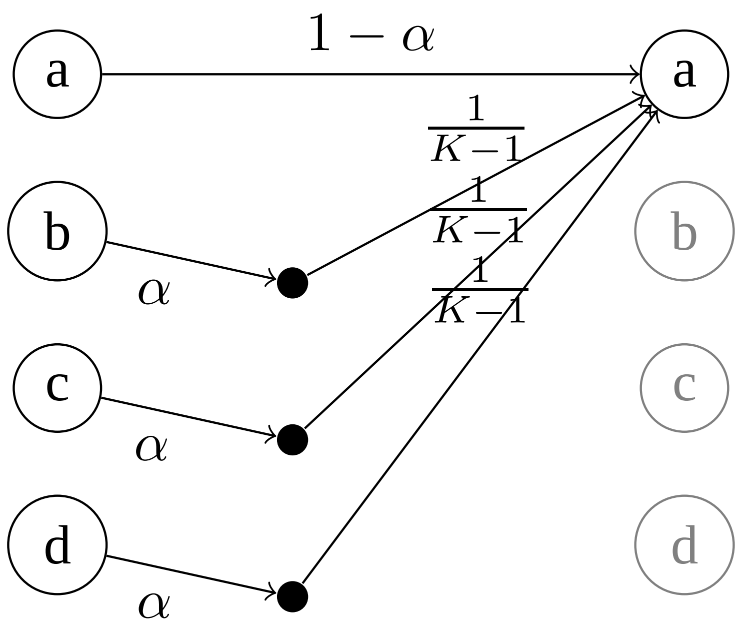
$$w_{t+1}^k \;:=\; \frac{\alpha_{t+1}}{K-1} + \left(1 - \frac{K}{K-1}\alpha_{t+1}\right) \frac{w_t^k e^{-\ell_t^k}}{\sum_{k=1}^K w_t^k e^{-\ell_t^k}}\,.$$

Fixed Share: weight going out

Fraction $1 - \alpha$ of weight stays put. The remainder fraction α is redistributed uniformly to the other experts.



Fixed Share: weight coming in



Adaptive regret of Fixed Share

Theorem: Fixed Share with switching rates $\alpha_2, \alpha_3, \ldots$ guarantees

$$R_{[t_1,t_2]} \;\leq\; -\ln\left(\frac{\alpha_{t_1}}{K-1}\prod_{t=t_1+1}^{t_2}(1-\alpha_t)\right)$$

Proof: The Fixed Share update can be written equivalently as

$$w_{t+1}^k \;=\; (1-\alpha_{t+1})\frac{w_t^k e^{-\ell_t^k}}{\sum_{k=1}^K w_t^k e^{-\ell_t^k}} + \frac{\alpha_{t+1}}{K-1}\left(1 - \frac{w_t^k e^{-\ell_t^k}}{\sum_{k=1}^K w_t^k e^{-\ell_t^k}}\right)$$

We next prove by induction that the mix loss telescopes (with overhead)

$$\sum_{t=t_1}^{t_2} -\ln\left(\sum_{k=1}^K w_t^k e^{-\ell_t^k}\right) \;\leq\; -\ln\left(\sum_{k=1}^K w_{t_1}^k e^{-\sum_{t=t_1}^{t_2}\ell_t^k}\right) - \ln\prod_{t=t_1+1}^{t_2}(1-\alpha_t)$$

Base case: $t_1 = t_2$ trivial. Induction step:

$$\sum_{t=t_1-1}^{t_2} -\ln\left(\sum_{k=1}^{K} w_t^k e^{-\ell_t^k}\right)$$

$$\leq\ -\ln\left(\sum_{k=1}^{K} w_{t_1-1}^k e^{-\ell_{t_1-1}^k}\right) - \ln\left(\sum_{k=1}^{K} w_{t_1}^k e^{-\sum_{t=t_1}^{t_2} \ell_t^k} \prod_{t=t_1+1}^{t_2} (1-\alpha_t)\right)$$

$$\leq\ -\ln\left(\sum_{k=1}^{K} \left((1-\alpha_{t_1})\left(w_{t_1-1}^k e^{-\ell_{t_1-1}^k}\right)\right) e^{-\sum_{t=t_1}^{t_2} \ell_t^k} \prod_{t=t_1+1}^{t_2} (1-\alpha_t)\right)$$

$$=\ -\ln\left(\sum_{k=1}^{K} w_{t_1-1}^k e^{-\sum_{t=t_1-1}^{t_2} \ell_t^k} \prod_{t=t_1}^{t_2} (1-\alpha_t)\right)$$

The proof of the theorem is concluded by observing that for any expert k

$$
\begin{aligned}
\sum_{t=t_1}^{t_2} - \ln\left(\sum_{k=1}^{K} w_t^k e^{-\ell_t^k}\right) \\
&\leq \; -\ln\left(\sum_{k=1}^{K} w_{t_1}^k e^{-\sum_{t=t_1}^{t_2} \ell_t^k} \prod_{t=t_1+1}^{t_2} (1-\alpha_t)\right) \\
&\leq \; \sum_{t=t_1}^{t_2} \ell_t^k - \ln\left(w_{t_1}^k \prod_{t=t_1+1}^{t_2} (1-\alpha_t)\right) \\
&\leq \; \sum_{t=t_1}^{t_2} \ell_t^k - \ln\left(\frac{\alpha_t}{K-1} \prod_{t=t_1+1}^{t_2} (1-\alpha_t)\right)
\end{aligned}
$$

where the last inequality results from

$$
w_{t_1}^k \;\geq\; \frac{\alpha_t}{K-1}.
$$

Tuning Fixed Share

A *constant* $\alpha_t = \alpha$ results in

$$R_{[t_1,t_2]} \leq \ln(K-1) - \ln\alpha - (t_2 - t_1)\ln(1-\alpha)$$

A *slowly decreasing* $\alpha_t = 1/t$ results in

$$R_{[t_1,t_2]} \leq \ln(K-1) + \ln t_2$$

A *quickly decreasing* $\alpha_t = 1/(t\ln t)$ results in

$$R_{[t_1,t_2]} \leq \ln(K-1) + \ln t_1 + \ln\ln t_2$$

A *sum-convergent* $\alpha_t = 1/t^2$ results in

$$R_{[t_1,t_2]} \leq \ln(K-1) + 2\ln t_1 + \ln 2$$

Note: for $t_1 = 1$ replace $\ln(K-1)$ by $\ln K$.

Fixed Share Wrap-up

Fixed Share (upgrade of Aggregating Algorithm) "tracks" the best expert, in the sense that it performs almost as well as the best expert *locally*.

We found a palette of adaptive regret guarantees, parametrised by the switching rate sequence $\alpha_2, \alpha_3, \ldots$.

It can be shown that Fixed Share is the definitive algorithm for adaptive regret (in the mix loss game): *any adaptive regret guarantee* $R_{[t_1,t_2]} \leq \phi(t_1, t_2)$ *— no matter how smart the strategy — is reproduced by Fixed Share (with particular switching rates depending on* ϕ*)*

Minimax replaced by *Pareto optimality*.