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

Loss of a

\[
\begin{align*}
\text{set} & \quad \text{elements} \\
\text{tree} & \quad \text{edges} \\
\text{path} & \quad \text{edges} \\
\text{permutation} & \quad \text{assignments} \\
\ldots & \quad \ldots \\
\text{concepts} & \quad \text{components}
\end{align*}
\]

is the sum of the losses of its elements, edges, and assignments.

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 \( \ell_t \in [0, 1]^d \).
  - Learner incurs the dot loss \( \mathbb{E}_{C \sim W_t} [C^\top \ell_t] \).

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: $L_t = \ell_1 + \ldots + \ell_t$.

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

$$\arg\min_{C \in \mathcal{C}} C^\top \left( L_{t-1} + \frac{X_t}{\eta} \right)$$

We have special-purpose linear optimisation algorithms:

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

\[
\mathbb{E} L_T^{\text{FPL}} = \mathbb{E} L_T^{\text{IFPL}} + \mathbb{E} L_T^{\text{FPL}} - \mathbb{E} L_T^{\text{IFPL}}
\]

- close to best for high \( \eta \)
- small for low \( \eta \)
IFPL close to best concept

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

**Theorem:** After $T \geq 0$ rounds:

$$\mathbb{E} L_T^{\text{IFPL}} \leq \min_{C \in \mathcal{C}} C^\top L_T + \frac{U(1 + \ln d)}{\eta}$$

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):

$$M \left( \frac{X}{\eta} \right)^\top \frac{X}{\eta} + \sum_{t=1}^{T} M \left( L_t + \frac{X}{\eta} \right)^\top \ell_t \leq M \left( L_T + \frac{X}{\eta} \right)^\top \left( L_T + \frac{X}{\eta} \right)$$

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

\[ M \left( L_{T-1} + \frac{X}{\eta} \right)^T \left( L_{T-1} + \frac{X}{\eta} \right) + M \left( L_T + \frac{X}{\eta} \right)^T \ell_T \leq M \left( L_T + \frac{X}{\eta} \right)^T \left( L_T + \frac{X}{\eta} \right) \]

that is

\[ M \left( L_{T-1} + \frac{X}{\eta} \right)^T \left( L_{T-1} + \frac{X}{\eta} \right) \leq M \left( L_T + \frac{X}{\eta} \right)^T \left( L_{T-1} + \frac{X}{\eta} \right) \]

which follows from the definition of \( M \).

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

\[ \sum_{t=1}^{T} M \left( L_t + \frac{X}{\eta} \right)^T \ell_t \leq M \left( L_T + \frac{X}{\eta} \right)^T \left( L_T + \frac{X}{\eta} \right) - M \left( \frac{X}{\eta} \right)^T \frac{X}{\eta} \]
We then use
\[
M \left( L_T + \frac{X}{\eta} \right)^T \left( L_T + \frac{X}{\eta} \right) \leq M (L_T)^T \left( L_T + \frac{X}{\eta} \right)
\]
\[
= M (L_T)^T L_T + \frac{1}{\eta} M (L_T)^T X.
\]
\leq 0 \text{ since } X \leq 0

We then continue to observe that
\[
-M \left( \frac{X}{\eta} \right)^T \frac{X}{\eta} \leq \frac{1}{\eta} \left| M \left( \frac{X}{\eta} \right) \right|_1 |X|_\infty
\]
\[
= \frac{U |X|_\infty}{\eta}
\]

The expected maximum of \(d\) standard exponentials is \(\leq 1 + \ln d\).
**FPL close to IFPL**

**Theorem:** In each round $t$:

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

(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:

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

(tedious but straightforward manipulation of exponential distributions)

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

$$
\mathbb{E} \ell_t^{\text{FPL}} \leq e^{\eta d} \mathbb{E} \ell_t^{\text{IFPL}}
$$
And hence, using $e^{-\eta d} \geq 1 - \eta d$ and $\ell \in [0, U]$,

$$(1 - \eta d) \mathbb{E} \ell_t^{FPL} \leq \mathbb{E} \ell_t^{IFPL}$$

so that

$$\mathbb{E} \ell_t^{FPL} - \mathbb{E} \ell_t^{IFPL} \leq \eta d U.$$
We proved

\[ \mathbb{E} R_T^{\text{FPL}} \leq TdU\eta + \frac{U(1 + \ln d)}{\eta} \]

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

\[ \mathbb{E} R_T^{\text{FPL}} \leq 2U \sqrt{Td(1 + \ln d)} \]
Part 2: Adaptive Regret
Motivation: non-stationary data

Suppose the data are like this

<table>
<thead>
<tr>
<th></th>
<th>T/2 rounds</th>
<th>T/2 rounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>expert 1</td>
<td>loss 0</td>
<td>loss 1</td>
</tr>
<tr>
<td>expert 2</td>
<td>loss 1</td>
<td>loss 0</td>
</tr>
</tbody>
</table>

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 $\approx 0$, 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 $w_t$ on $K$ experts.
  – Adversary reveals loss vector $\ell_t \in (-\infty, \infty]^K$.
  – 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

\[
R_{[t_1,t_2]} = \sum_{t=t_1}^{t_2} \left( \ln \left( \sum_{k=1}^{K} w_t^k e^{-\ell_t^k} \right) \right) - \min_k \sum_{t=t_1}^{t_2} \ell_t^k
\]

- Learner’s mix loss in round \(t\)
- 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 $w^k_1 = 1/K$ in round 1, and updates its weights as

$$w^k_{t+1} := \frac{\alpha_{t+1}}{K - 1} + \left(1 - \frac{K}{K - 1} \alpha_{t+1}\right) \frac{w^k_t e^{-\ell^k_t}}{\sum_{k=1}^K w^k_t e^{-\ell^k_t}}.$$
Fraction $1 - \alpha$ of weight stays put. The remainder fraction $\alpha$ is redistributed uniformly to the other experts.
Fixed Share: weight coming in

\[ a \quad 1 - \alpha \quad a \]

\[ b \quad \alpha \quad b \]

\[ c \quad \alpha \quad c \]

\[ d \quad \alpha \quad d \]
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

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

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:

$$
\sum_{t=t_1-1}^{t_2} - \ln \left( \sum_{k=1}^{K} w_t^k e^{-\ell_t^k} \right)
\leq \sum_{k=1}^{K} w_{t_1-1}^k e^{-\ell_{t_1-1}^k} - \ln \left( \sum_{k=1}^{K} w_{t_1}^k e^{-\sum_{t=t_1}^{t_2} \ell_t^k} \prod_{t=t+1}^{t_2} (1 - \alpha_t) \right)
\leq \sum_{k=1}^{K} \left( (1 - \alpha_{t_1}) \left( w_{t_1-1}^k e^{-\ell_{t_1-1}^k} \right) \right) e^{-\sum_{t=t_1}^{t_2} \ell_t^k} \prod_{t=t_1+1}^{t_2} (1 - \alpha_t)
= \sum_{k=1}^{K} w_{t_1-1}^k e^{-\sum_{t=t_1-1}^{t_2} \ell_t^k} \prod_{t=t_1}^{t_2} (1 - \alpha_t)
$$
The proof of the theorem is concluded by observing that for any expert $k$

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

where the last inequality results from

\[
w_{t_1}^k \geq \frac{\alpha_t}{K-1}.
\]
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 $\phi$).

Minimax replaced by Pareto optimality.