Machine Learning Theory. Lecture 10.

Wouter M. Koolen

Recap:

- Mix loss (with non-uniform regret bounds)
- Specialists

Non-stationary environments

- Switching (Fixed Share algorithm)
- Long-term memory (Mixing Past Posteriors algorithm)

Mix loss

For $t = 1, 2, \ldots$

1. Play $\boldsymbol{w}_t \in \triangle_K$.
2. See $\boldsymbol{\ell}_t \in \mathbb{R}^K$.
3. Incur *mix loss* $\hat{\ell}_t := -\ln\left(\sum_{k=1}^K w_t^k e^{-\ell_t^k}\right)$.

Definition 1. *The* regret *w.r.t. expert* $k \in [K]$ *after* $T \geq 0$ *rounds is*

$$R_T^k := \sum_{t=1}^T \left(\hat{\ell}_t - \ell_t^k\right).$$

The Aggregating Algorithm

Definition 2. *The Aggregating Algorithm with* prior $\boldsymbol{\pi} \in \triangle_K$ *plays*

$$w_t^k \;=\; \frac{\pi^k e^{-\sum_{s=1}^{t-1} \ell_s^k}}{\sum_{j=1}^{K} \pi^j e^{-\sum_{s=1}^{t-1} \ell_s^j}} \qquad \text{(AA-}\boldsymbol{\pi}\text{)}$$

(so $\boldsymbol{w}_1 = \boldsymbol{\pi}$ *and* $w_{t+1}^k = \frac{w_t^k e^{-\ell_t^k}}{\sum_{j=1}^{K} w_t^j e^{-\ell_t^j}}$*)*

Theorem 3. *The regret of AA*$-\boldsymbol{\pi}$ *w.r.t. expert* $k \in [K]$ *satisfies*

$$R_T^k \;\le\; -\ln \pi^k$$

Proof. The cumulative loss of AA$-\boldsymbol{\pi}$ telescopes to

$$\begin{aligned}
\sum_{t=1}^{T} \hat{\ell}_t &= \sum_{t=1}^{T} -\ln\left(\sum_{k=1}^{K} \frac{\pi^k e^{-\sum_{s=1}^{t-1}\ell_s^k}}{\sum_{j=1}^{K}\pi^j e^{-\sum_{s=1}^{t-1}\ell_s^j}} e^{-\boldsymbol{\ell}_t^k}\right) \\
&= \sum_{t=1}^{T} -\ln\left(\frac{\sum_{k=1}^{K}\pi^k e^{-\sum_{s=1}^{t}\ell_s^k}}{\sum_{k=1}^{K}\pi^k e^{-\sum_{s=1}^{t-1}\ell_s^k}}\right) \\
&= -\ln\left(\frac{\sum_{k=1}^{K}\pi^k e^{-\sum_{s=1}^{T}\ell_s^k}}{\sum_{k=1}^{K}\pi^k}\right) \\
&= -\ln\left(\sum_{k=1}^{K}\pi^k e^{-\sum_{s=1}^{T}\ell_s^k}\right)
\end{aligned}$$

So for any $k \in [K]$

$$\sum_{t=1}^{T} \hat{\ell}_t \leq -\ln\left(\pi^k e^{-\sum_{t=1}^{T}\ell_t^k}\right) = -\ln\pi^k + \sum_{t=1}^{T}\ell_t^k \qquad \square$$

Mix loss and specialists

For $t = 1, 2, \ldots$

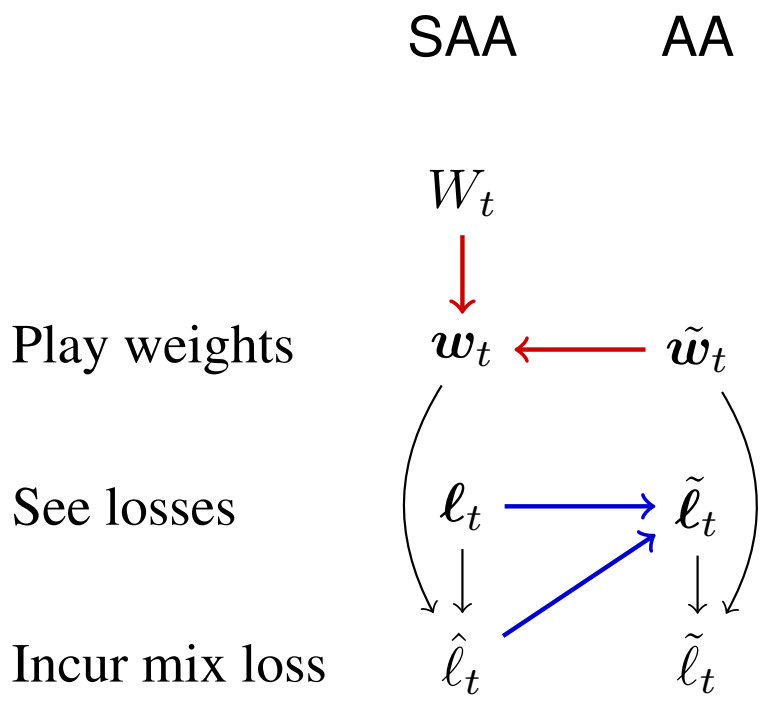
1. Observe set $W_t \subseteq [K]$ of available specialists.
2. Play $\boldsymbol{w}_t \in \triangle_{W_t}$.
3. See $\boldsymbol{\ell}_t \in \mathbb{R}^{W_t}$.
4. Incur *mix loss* $\hat{\ell}_t := -\ln\left(\sum_{k \in W_t} w_t^k e^{-\ell_t^k}\right)$.

Definition 4. *The* regret *w.r.t. specialist* $k \in [K]$ *after* $T \geq 0$ *rounds is*

$$R_T^k := \sum_{t \in [T] : k \in W_t} \left(\hat{\ell}_t - \ell_t^k\right)$$

Specialists Aggregating Algorithm (Reduction)

The Specialists Aggregating Algorithm (SAA-π) is defined by reduction to AA-π for experts. SAA sets AA's losses and transforms AA's weights.



$$w_t^k := \frac{\tilde{w}_t^k}{\sum_{j \in W_t} \tilde{w}_t^j}$$

$$\hat{\ell}_t = -\ln\left(\sum_{k \in W_t} w_t^k e^{-\ell_t^k}\right)$$

$$\tilde{\ell}_t^k := \begin{cases} \ell_t^k & k \in W_t \\ \hat{\ell}_t & k \notin W_t \end{cases}$$

$$\tilde{\ell}_t = -\ln\left(\sum_{k=1}^{K} \tilde{w}_t^k e^{-\tilde{\ell}_t^k}\right)$$

Specialists Crucial Equality

Lemma 5. *The SAA-$\boldsymbol{\pi}$ mix loss equals the AA-$\boldsymbol{\pi}$ mix loss:* $\hat{\ell}_t = \tilde{\ell}_t$.

Proof. Let's expand the AA mix loss

$$
\begin{aligned}
\tilde{\ell}_t &= -\ln\left(\sum_{k=1}^{K} \tilde{w}_t^k e^{-\tilde{\ell}_t^k}\right) \\
&\overset{\text{def } \tilde{\ell}_t^k}{=} -\ln\left(\sum_{k\in W_t} \tilde{w}_t^k e^{-\ell_t^k} + \sum_{k\notin W_t} \tilde{w}_t^k e^{-\hat{\ell}_t}\right) \\
&\overset{\text{def } w_t^k}{=} -\ln\left(\sum_{k\in W_t} \tilde{w}_t^k \sum_{k\in W_t} w_t^k e^{-\ell_t^k} + \sum_{k\notin W_t} \tilde{w}_t^k e^{-\hat{\ell}_t}\right) \\
&\overset{\text{def } \hat{\ell}_t}{=} -\ln\left(\sum_{k\in W_t} \tilde{w}_t^k e^{-\hat{\ell}_t} + \sum_{k\notin W_t} \tilde{w}_t^k e^{-\hat{\ell}_t}\right) = \hat{\ell}_t
\end{aligned}
$$

□

Specialists regret bound

Theorem 6. *SAA-$\boldsymbol{\pi}$ guarantees for each expert k at each time T*

$$R_T^k \;\leq\; -\ln \pi^k$$

Proof. By construction of $\tilde{\ell}_t^k$, we have

$$R_T^k \;=\; \sum_{t\in[T]:k\in W_t} \left(\hat{\ell}_t - \ell_t^k\right) \;=\; \sum_{t=1}^{T} \begin{cases} \hat{\ell}_t - \ell_t^k & k \in W_t \\ \hat{\ell}_t - \hat{\ell}_t & \text{o.w.} \end{cases} \;=\; \sum_{t=1}^{T} \left(\hat{\ell}_t - \tilde{\ell}_t^k\right)$$

Hence for each $k \in [K]$

$$R_T^k \;=\; \sum_{t=1}^{T} \left(\hat{\ell}_t - \tilde{\ell}_t^k\right) \overset{\text{L. 5}}{=} \sum_{t=1}^{T} \left(\tilde{\ell}_t - \tilde{\ell}_t^k\right) \overset{\text{Th. 3}}{\leq} -\ln \pi^k.$$

□

Non-stationary data

1. Switching, Tracking, Shifting
2. Long-term Memory

Non-stationary data

So far we have been looking at regret compared to a *fixed* expert/action.

$$R_T \;=\; \max_{k\in[K]} \sum_{t=1}^{T} \left(\hat{\ell}_t - \ell_t^k\right).$$

But what if we do not expect a single expert to be good for all data?

Definition 7. *The regret on interval* $[t_1, t_2]$ *is defined by*

$$R_{[t_1,t_2]} \;:=\; \max_{k\in[K]} \sum_{t\in[t_1,t_2]} \left(\hat{\ell}_k - \ell_t^k\right)$$

Question: Can we keep the interval regret small on *every interval*?

Fixed Share (Specialists Rendering)

Starting with K experts, create an "explosion" of specialists

$$\mathcal{S} \;:=\; \{(k,s) \mid k \in [K] \text{ and } s \geq 1\}.$$

Set the SAA prior, availability and losses

$$\begin{aligned} \pi^{(k,s)} &:= \frac{1}{K}\pi^s \qquad \text{with } \pi^s \text{ a prior on } \{1,2,\ldots\} \\ W_t &:= \{(k,s) \mid k \in [K] \text{ and } s \leq t\} \\ \ell_t^{(k,s)} &:= \ell_t^k \qquad \text{for } s \leq t \end{aligned}$$

Based on the SAA predictions $w_t^{(k,s)} \in \triangle_{W_T}$, *Fixed Share* plays

$$w_t^k \;=\; \sum_{s:(k,s)\in W_t} w_t^{(k,s)} \qquad \text{(FS)}$$

Fixed Share Regret Bound

Application of the SAA specialists regret bound, (Theorem 6) gives

Theorem 8. *Fixed Share ensures that the regret on each interval* $1 \leq t_1 \leq t_2$ *is at most*

$$R_{[t_1,t_2]} \;\leq\; \ln K - \ln \pi^{t_1}$$

Corollary 9. *For example, picking* $\pi^t = \frac{1}{t(t+1)}$*, we find*

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

Fixed Share: Computation Collapses

Seems we need to maintain infinitely many weights. But after t rounds

- $\{(k, s) \mid k \in [K], s > t\}$ were never available (so weight still prior).
- for each k, $\{(k, s) \mid s \leq t\}$ have same future (so can merge weights).

Conclusion: only need K weights. Same as AA!

Long-Term Memory

We saw Fixed Share pays a $\ln K$ term for each interval. So for a partition with B blocks, it pays $B \ln K$ for the per-block best expert.

Freund's problem (2000) What if there are many experts K, but only a few, say $M \ll K$ are useful in B blocks?

Can we pay $M \ln K$ once, and then $B \ln M$?

The method would need *long-term memory*. An expert that was useful in the past needs to be re-learned *faster/cheaper*.

Freund presented a computationally intractable solution.

Question: can one do this efficiently?

Mixing Past Posteriors (Specialists Rendering)

Idea: Create an even bigger explosion of experts.

Fix a horizon T (for simplicity). Create specialists

$$\mathcal{S} \;:=\; \{(k,S)|k \in [K] \text{ and } S \subseteq [T]\}$$

Set SAA prior, availability and losses

$$\begin{aligned}
\pi^{(k,S)} &:= \frac{1}{K}\pi^S \quad \text{where } \pi^S \text{ a prior on } S \subseteq [T] \\
W_t &:= \{(k,S) \in \mathcal{S} | t \in S\} \\
\ell_t^{(k,S)} &:= \ell_t^k
\end{aligned}$$

Now when SAA produces $w_t^{(k,S)} \in \triangle_{W_t}$, *Mixing Past Posteriors* plays

$$w_t^k \;=\; \sum_{S:(k,S)\in W_t} w_t^{(k,S)}. \qquad \text{(MPP)}$$

MPP Regret

Application of the SAA specialists regret bound, (Theorem 6) gives

Theorem 10. *Fix any sequence* $(k_1, S_1), \ldots, (k_M, S_M)$ *where* $k_m \in [K]$ *and* $\{S_1, \ldots, S_M\}$ *partition* $[T]$. *Let* $m(t)$ *be the index* m *such that* $t \in S_m$. *Then the* partition regret *of MPP is at most*

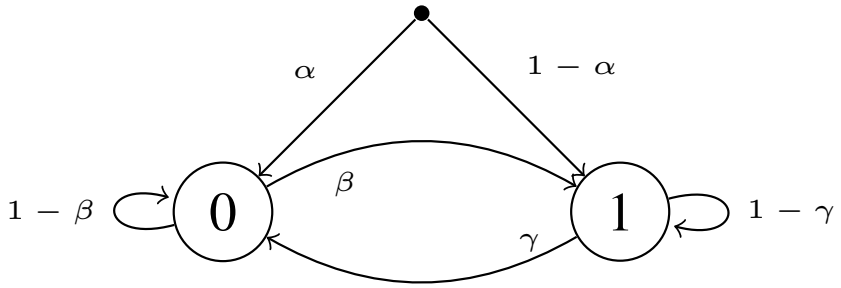
$$\sum_{t=1}^{T} \left(\hat{\ell}_t - \ell_t^{k_{m(t)}} \right) \;\leq\; M \ln K + \sum_{m=1}^{M} - \ln \pi^{S_m}.$$

Desiderata: prior π^S that

- Gives high likelihood to "simple" S: few/long stretches
- Simplifies the computations

A Suitable Prior for Mixing Past Posteriors

We need to pick a prior on $\{0,1\}^T$. Idea: Markov prior.



Example 11. *The set* $S = 111000011 \in \{0,1\}^9$ *has prior probability* $\pi^S = (1-\alpha)(1-\gamma)^2\gamma(1-\beta)^3\beta(1-\gamma)$.

Cost of a partition

Now a partition $S_1, \ldots, S_M$ of $[T]$ into B blocks costs

$$\sum_{m=1}^{M} -\ln \pi^{S_m} = \underbrace{-\ln(1-a)}_{\bullet\to 1} \; \underbrace{-(M-1)\ln\alpha}_{\bullet\to 0} \; \underbrace{-(B-1)\ln\gamma}_{1\to 0} \; \underbrace{-(B-1)\ln\beta}_{0\to 1} \; \underbrace{-(T-B)\ln(1-\gamma)}_{1\to 1} \; \underbrace{-((T-1)(M-1)-(B-1))\ln(1-\beta)}_{0\to 0}$$

Tuning MPP

With the optimal tuning

$$\alpha = \frac{1}{M} \qquad \beta = \frac{B-1}{(M-1)(T-1)} \qquad \gamma = \frac{B-1}{T-1}$$

We find partition cost (with $H(\theta) = -\theta \ln \theta - (1-\theta)\ln(1-\theta)$)

$$MH\left(\frac{1}{M}\right) + (M-1)(T-1)H\left(\frac{B-1}{(M-1)(T-1)}\right) + (T-1)H\left(\frac{B-1}{T-1}\right)$$

Which is at most (using $ZH(1/Z) \leq 1 + \ln Z$)

$$1 + \ln M + (B-1)(1 + \ln \frac{(M-1)(T-1)}{B-1}) + (B-1)(1 + \ln \frac{T-1}{B-1})$$

that is

$$B \ln M + 2B \ln T + O(1)$$

MPP Computation Collapses

1. Markov prior does not use T. MPP works for $T \to \infty$.
2. After t rounds, for each k, the prior is Markov. So future only depends on $t \in S$ or $t \notin S$.

Can implement MPP with only $2K$ weights.

Conclusion

Techniques for adapting to changing environment

- Fixed Share for switching between experts
- Mixing Past Posteriors for long-term memory

Conceptual message:

- Adapting to changing environment is not automatic
- Modelling with Specialists