CS281B/Stat241B. Statistical Learning Theory. Lecture 14.

Wouter M. Koolen

- Convex losses
- Exp-concave losses
- Mixable losses
- The gradient trick
- Specialists

Menu

Today we solve new online learning problems by reducing them to problems/algorithms/analyses we already cracked before.

Prediction with Expert Advice

Prediction with expert advice:

Protocol:

- For $t = 1, 2, \ldots$
  - Experts announce actions $a_t^1, \ldots, a_t^K \in \mathcal{A}$.
  - Learner chooses an action $a_t \in \mathcal{A}$.
  - Adversary reveals outcome $x_t \in \mathcal{X}$.
  - Learner incurs loss $\mathcal{L}(a_t, x_t)$.

Goal: small regret w.r.t. best expert.

Convex: Reduction to dot loss

Say $\mathcal{L}(a, x)$ is $[0, 1]$-bounded and convex in a for each x:

$$\sum_{k=1}^{K} w^k \mathcal{L}(a^k, x) \;\geq\; \mathcal{L}\left(\sum_{k=1}^{K} w^k a^k, x\right)$$

Then we can feed Hedge $\ell_t^k = \mathcal{L}(a_t^k, x_t)$.

Hedge outputs $\boldsymbol{w}_t$. Play the mean action $a_t \;=\; \sum_{k=1}^{K} w_t^k a_t^k$.

$$\underbrace{\sum_{k=1}^{K} w_t^k \mathcal{L}(a_t^k, x_t)}_{\text{Dot loss } \boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t} \;\geq\; \underbrace{\mathcal{L}\left(a_t, x_t\right)}_{\text{actual loss}}$$

Dot-loss bound translates to convex bounded loss $\mathcal{L}$.

$$R_T \;\leq\; \sqrt{T/2 \ln K}$$

Exp-concave: Reduction to mix loss

Definition: We say $\mathcal{L}(a, x)$ is η-*exp-concave* in a for each x if

$$\sum_{k=1}^{K} w^k e^{-\eta \mathcal{L}(a^k, x)} \;\leq\; e^{-\eta \mathcal{L}\left(\sum_{k=1}^{K} w^k a^k, x\right)}$$

Then we can feed the AA $\ell_t^k = \eta \mathcal{L}(a_t^k, x_t)$.

The AA outputs $\boldsymbol{w}_t$. Play the mean action $a_t \;=\; \sum_{k=1}^{K} w_t^k a_t^k$.

$$\underbrace{-\ln\left(\sum_{k=1}^{K} w_t^k e^{-\eta \mathcal{L}(a_t^k, x_t)}\right)}_{\text{Mix loss of } \boldsymbol{w}_t \text{ on } \eta \boldsymbol{\ell}_t} \;\geq\; \eta\, \underbrace{\mathcal{L}\left(\sum_{k=1}^{K} w_t^k a_t^k, x_t\right)}_{\text{actual loss}}$$

Mix-loss bound translates to exp-concave loss $\mathcal{L}$:

$$R_T \leq \frac{\ln K}{\eta}$$

Example: square loss is exp-concave

Let's consider

$$\mathcal{L}(a, x) = (a - x)^2$$

where $\mathcal{A} = \mathcal{X} = [-1, +1]$.

Find η such that $\mathcal{L}$ is η-exp-concave by testing negative second derivative:

$$\begin{aligned} \frac{\partial^2}{\partial a^2} e^{-\eta(a-x)^2} &= \frac{\partial}{\partial a} - 2e^{-\eta(a-x)^2}\eta(a-x) \\ &= e^{-\eta(a-x)^2}\eta\left(4\eta(a-x)^2 - 2\right) \end{aligned}$$

Highest η such that $4\eta(a-x)^2 - 2 \leq 0$ for all a, x. $\Rightarrow \eta = 1/8$.

For $\mathcal{X} = \mathcal{A} = [-Y, +Y]$ we find $\eta = \frac{1}{8Y^2}$.

Mixable loss: Reduction to mix loss

Crux: exp-concavity is convenient but *too strong*.

Definition: We say $\mathcal{L}(a, x)$ is η-*mixable* if

$$\forall \boldsymbol{w} \forall a^1, \ldots, a^K \exists a \forall x \quad \mathcal{L}(a, x) \leq \frac{-1}{\eta} \ln \left(\sum_{k=1}^{K} w^k e^{-\eta \mathcal{L}(a^k, x)} \right)$$

Mapping from $\boldsymbol{w}$, $a_1, \ldots, a_K$ to witness a called *substitution function*.

Mixable losses behave just enough like the mix loss to carry the AA regret bound through.

$$R_T \leq \frac{\ln K}{\eta}$$

Square loss is mixable

Square loss is mixable with $\eta = \frac{1}{2}$. The substitution function is

$$\boldsymbol{w}, a^1, \ldots, a^K \;\mapsto\; \frac{m_{\frac{1}{2}}(-1) - m_{\frac{1}{2}}(+1)}{4}$$

where $m_\eta(x) \;=\; \frac{-1}{\eta} \ln \sum_{k=1}^{K} w^k e^{-\eta(a^k - x)^2}$

See (Vovk 1990, Haussler, Kivinen, Warmuth, 1998)

Mixable loss list

Popular mixable losses:

- mix loss, log loss, entropic loss
- square loss, Brier loss
- Hellinger loss $\mathcal{A} = \mathcal{X} = [0, 1]$:

$$\mathcal{L}(a, x) \;:=\; \frac{1}{2}\left((\sqrt{1-x} - \sqrt{1-a})^2 + (\sqrt{x} - \sqrt{a})\right)$$

Characterisation of mixability: (Van Erven, Reid, Williamson 2012).

Gradient trick

Gradient trick

Abusing an algorithm that can compete with the best *expert* to in fact compete with the best *convex combination* (cf portfolios).

Assume convex loss $\mathcal{L}(\boldsymbol{w}, x)$:

$$\mathcal{L}(\boldsymbol{w}, x) \;\geq\; \underbrace{\mathcal{L}(\boldsymbol{w}_t, x) + (\boldsymbol{w} - \boldsymbol{w}_t)\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, x)}_{\text{First-order expansion around algorithm's action } \boldsymbol{w}_t}$$

Idea: feed Hedge $\boldsymbol{\ell}_t = \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, x)$ (may need restriction + translation + scaling to make this $[0, 1]$ bounded). Get $\boldsymbol{w}_t$. Play $\boldsymbol{w}_t$.

Imaginary regret upper bounds the actual regret:

$$\begin{aligned}
\sum_{t=1}^{T} \boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \min_k \sum_{t=1}^{T} \ell_t^k &= \max_k \sum_{t=1}^{T} \left(\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \ell_t^k \right) \\
&= \max_{\boldsymbol{w}} \sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{w})^\intercal \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_t, x) \\
&\geq \max_{\boldsymbol{w}} \sum_{t=1}^{T} \left(\mathcal{L}(\boldsymbol{w}_t, x) - \mathcal{L}(\boldsymbol{w}, x) \right) \\
&= \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{w}_t, x) - \max_{\boldsymbol{w}} \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{w}, x)
\end{aligned}$$

Caveat: even if original loss was nice (mixable/curved/...), the imagined loss $\boldsymbol{w} \mapsto \boldsymbol{w}^\intercal \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, x)$ is *linear*. Regret of order $\sqrt{T}$.

Specialists

Motivation

Not all expert predictions/actions available every round.

- Missing data
- Noise
- Too expensive ($/time/memory)

How to model missingness? Adversarial.

How to redefine the objective? New variant of regret.

How to still do something optimal? Upgrade of AA.

Mix loss game with specialists

Protocol:

- For $t = 1, 2, \ldots$
  - Adversary picks the subset $A_t \subseteq [K]$ of *awake* specialists.
  - Learner chooses a distribution $\boldsymbol{w}_t$ on awake specialists A_t.
  - Adversary reveals loss vector $\boldsymbol{\ell}_t \in (-\infty, \infty]^{A_t}$.
  - Learner's loss is the **mix loss** $-\ln\left(\sum_{k \in A_t} w_{t,k} e^{-\ell_{t,k}}\right)$

Objective

There is no loss when a specialist is asleep.

Regret w.r.t. specialist j: only measured during rounds where j is awake

$$R_T^j = \sum_{\substack{t\in[T]\\ j\in A_t}} -\ln\left(\sum_{k\in A_t} w_t^k e^{-\ell_t^k}\right) - \sum_{\substack{t\in[T]\\ j\in A_t}} \ell_t^j$$

Specialist AA

Definition: The *Specialist Aggregating Algorithm* (SAA) maintains a distribution $\boldsymbol{u}_t$. It starts uniform $\boldsymbol{u}_1^k = 1/K$.

In round t with awake experts A_t, SAA predict with

$$w_t^k \;=\; u_t(k|A_t) \;=\; \frac{u_t^k \mathbf{1}_{\{k\in A_t\}}}{\sum_{j\in A_t} u_t^j}$$

Update:

$$u_{t+1}^k \;=\; \begin{cases} \dfrac{u_t^k e^{-\ell_t^k}}{\sum_{j\in A_t} u_t^k e^{-\ell_t^k}} \sum_{j\in A_t} u_t^k & k \in A_t \\ u_t^k & k \notin A_t \end{cases}$$

AA update relative to awake set A_t

What makes this tick

Consider the sequence $\boldsymbol{\ell}'_1, \boldsymbol{\ell}'_2$ obtained by completing $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2$ by assigning in each round the SAA mix loss to all the asleep specialists.

Theorem: SAA on $\boldsymbol{\ell}$ and AA on $\boldsymbol{\ell}'$ produce identical weights $\boldsymbol{u}_t = \boldsymbol{w}'_t$ and suffer identical mix loss.

Proof: (homework)

Specialist regret bound for SAA

The AA has small regret w.r.t. expert j:

$$\begin{aligned}
\ln K &\geq \sum_{t=1}^{T} -\ln\left(\sum_{k=1}^{K} {w'_t}^k e^{-{\ell'_t}^k}\right) - \sum_{t=1}^{T} {\ell'_t}^j \\
&= \sum_{t=1}^{T} -\ln\left(\sum_{k\in A_t} w_t^k e^{-\ell_t^k}\right) - \sum_{\substack{t=[T]\\ t\in A_j}} \ell_t^j - \sum_{\substack{t=[T]\\ t\notin A_j}} -\ln\left(\sum_{k\in A_t} w_t^k e^{-\ell_t^k}\right) \\
&= \sum_{\substack{t=[T]\\ t\in A_j}} -\ln\left(\sum_{k\in A_t} w_t^k e^{-\ell_t^k}\right) - \sum_{\substack{t=[T]\\ t\in A_j}} \ell_t^j \\
&= R_T^j
\end{aligned}$$

Adversary more power (sleeping) but regret still $\ln K$: SAA minimax for specialist mix-loss regret game.

Discussion

- Convex bounded losses are easier than dot loss.
- Mixable losses are easier than mix loss.
- Gradient trick allows us to compete with mixtures (at a cost)
- Specialists extension deals with missing data (at no cost).