CS281B/Stat241B. Statistical Learning Theory. Lecture 9.

Wouter M. Koolen

Mix loss and dot loss: one-and-a-half algorithm

Online learning focus

- Tight feedback loop (recurring prediction task)
- Continuous learning (no training/learning separation)
- Adversarial analysis (Prequential principle, individual sequence. There is only the data. Also establishes robustness of statistical estimators.)
- Emphasis on both computational and statistical performance
- Regret: relative notion of performance

Application domains

Truly sequential problems:

- electricity demand prediction (EDF, also Amazon)
- mobile device power management
- hybrid cars engine switching
- caching
- medical trials (bandits)
- online advertisement (bandits)
- weather forecasting
- data compression (CTW)
- statistical testing
- investment (Universal portfolios)
- input assistants (e.g. Dasher)
- prediction with expert advice (meld human and machine prediction)
- online convex optimisation

Wider application

- Big data sets (transport online algorithm state, online to batch conversion)
- Convex optimisation
- Game theory (online learning methods for approximate equilibrium)
- General understanding
  - Uncertainty and ways to manipulate it
  - Makeup of and patterns in data
  - Complexity of classes of strategies

The menu for today

Two fundamental and prototypical online learning problems

- The mix-loss game
- The dot-loss game

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's loss is the **mix loss** $-\ln\left(\sum_{k=1}^K w_{t,k} e^{-\ell_{t,k}}\right)$

Instances:

- Investment (loss is *negative log-growth*)
- Data compression (loss is *code length*)
- Probability forecasting (loss is *logarithmic loss*)

Mix-loss objective

Obviously we cannot guarantee small loss.

Idea: relative evaluation, i.e. performance close to best expert.

Definition: After T rounds of the mix-loss game, the *regret* is given by

$$R_T = \underbrace{\sum_{t=1}^{T} - \ln \left(\sum_{k=1}^{K} w_{t,k} e^{-\ell_{t,k}} \right)}_{\text{Learner's mix loss}} - \underbrace{\min_k \sum_{t=1}^{T} \ell_{t,k}}_{\text{loss of best expert}}$$

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

Mix-loss regret: lower bound (adversary)

Theorem: Any algorithm for Learner can be forced to incur regret $R_T \geq \ln K$, *already in* $T = 1$ *round.*

Idea: Look at $k_{\text{low}} = \arg\min_k w_{1,k}$ so that $w_{1,k_{\text{low}}} \leq \frac{1}{K}$.

Administer loss killing everyone but k_{low}

$$\ell_{1,k} \;=\; \begin{cases} \infty & k \neq k_{\text{low}} \\ 0 & k = k_{\text{low}} \end{cases}$$

Now Learner's mix loss equals

$$-\ln\left(\sum_{k=1}^{K} w_{1,k} e^{-\ell_{1,k}}\right) \;=\; -\ln\left(w_{1,k_{\text{low}}} e^{-\ell_{t,k_{\text{low}}}}\right) \;\geq\; \ln K + \ell_{t,k_{\text{low}}}$$

The Aggregating Algorithm for mix loss

Definition: The *Aggregating Algorithm* plays weights in round t:

$$w_{t,k} = \frac{e^{-\sum_{s=1}^{t-1} \ell_{s,k}}}{\sum_{j=1}^{K} e^{-\sum_{s=1}^{t-1} \ell_{s,j}}} \qquad \text{(AA)}$$

or, equivalently, $w_{1,k} = \frac{1}{K}$ and

$$w_{t+1,k} = \frac{w_{t,k} e^{-\ell_{t,k}}}{\sum_{j=1}^{K} w_{t,j} e^{-\ell_{t,j}}} \qquad \text{(AA, incremental)}$$

Many names

- (Generalisation of) Bayes rule
- Exponentially weighted average

Mix-loss regret: upper bound (algorithm)

Theorem: The regret of the Aggregating Algorithm does not exceed $R_T \leq \ln K$ for all $T \geq 0$.

Proof: Crucial observation is that mix loss *telescopes*

$$\sum_{t=1}^{T} -\ln\left(\sum_{k=1}^{K} w_{t,k} e^{-\ell_{t,k}}\right) = \sum_{t=1}^{T} -\ln\left(\sum_{k=1}^{K} \frac{e^{-\sum_{s=1}^{t-1} \ell_{s,k}}}{\sum_{j=1}^{K} e^{-\sum_{s=1}^{t-1} \ell_{s,j}}} e^{-\ell_{t,k}}\right)$$

$$= \sum_{t=1}^{T} -\ln\left(\frac{\sum_{k=1}^{K} e^{-\sum_{s=1}^{t} \ell_{s,k}}}{\sum_{j=1}^{K} e^{-\sum_{s=1}^{t-1} \ell_{s,j}}}\right)$$

$$= -\ln\left(\sum_{k=1}^{K} e^{-\sum_{t=1}^{T} \ell_{t,k}}\right) + \ln K$$

and is bounded for each k by

$$\leq \sum_{t=1}^{T} \ell_{t,k} + \ln K \qquad (1)$$

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** $\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t$

Many names:

- Decision Theoretic Online Learning
- Prediction with Expert Advice
- The Hedge setting

Dot-loss objective

Definition: *Regret* after T rounds:

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

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

Hedge algorithm

Idea: re-use AA for mix loss, now with learning rate η.

Definition: The *Hedge algorithm* with *learning rate* η plays weights in round t:

$$w_{t,k} = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_{s,k}}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \ell_{s,j}}}. \qquad \text{(Hedge)}$$

or, equivalently, $w_{1,k} = \frac{1}{K}$ and

$$w_{t+1,k} = \frac{w_{t,k} e^{-\eta \ell_{t,k}}}{\sum_{j=1}^{K} w_{t,j} e^{-\eta \ell_{t,j}}} \qquad \text{(Hedge, incremental)}$$

Hedge analysis

Lemma: The regret of Hedge is bounded by

$$R_T \;\le\; T\frac{\eta}{8} + \frac{\ln K}{\eta}$$

Proof:

$$\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t \;=\; \underbrace{\frac{-1}{\eta}\ln\left(\sum_{k=1}^K w_{t,k} e^{-\eta \ell_{t,k}}\right)}_{\text{mix loss}} + \underbrace{\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \frac{-1}{\eta}\ln\left(\sum_{k=1}^K w_{t,k} e^{-\eta \ell_{t,k}}\right)}_{\text{mixability gap}}$$

The mix loss telescopes, and is bounded by (1) by

$$\sum_{t=1}^T \frac{-1}{\eta}\ln\left(\sum_{k=1}^K w_{t,k} e^{-\eta \ell_{t,k}}\right) \;\le\; \sum_{t=1}^T \ell_{t,k} + \frac{\ln K}{\eta}. \tag{2}$$

The mixability gap is bounded by Hoeffding (recall $\ell_{t,k} \in [0,1]$) by

$$\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \frac{-1}{\eta} \ln\left(\sum_{k=1}^K w_{t,k} e^{-\eta \ell_{t,k}}\right) \ \leq \ \frac{\eta}{8} \tag{3}$$

(to think about: when is Hoeffding tight?)

And over T rounds this accumulates to $T\frac{\eta}{8}$.

Putting (2) and (3) together yields the desired result.

Hedge tuning

Theorem: The Hedge regret bound is minimised at $\eta = \sqrt{\frac{8 \ln K}{T}}$ where it states

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

Extensions/generalisations

- Other losses
  - *Mixable losses* naturally reduce to mix-loss game.
  - *Bounded convex losses* naturally reduce to dot-loss game.
- Luckiness (not all data are created equal)
  - Loss of best expert exceptionally small (or big)
  - Empirical variance of loss of best expert small
  - Many experts are good
  - Special/favoured expert is good
  - ERM/FTL has low regret
- Bandits
- …