Online learning using Bregman Divergences

Yoav Freund

February 2, 2025

Outline

Hedge(η)Algorithm

Bound on total loss

The hedging problem

- N possible actions
- At each time step $t = 1, 2, \ldots, T$:
  - Algorithm chooses a distribution $\mathbf{p}^t$ over actions.
  - Losses $0 \leq \ell_i^t \leq 1$ of all actions $i = 1, \ldots, N$ are revealed.
  - Algorithm suffers **expected** loss $\mathbf{p}^t \cdot \ell_t$
- **Goal:** minimize total expected loss
- Here we have stochasticity - but only in **algorithm**, not in **outcome**

The **Hedge**(η)Algorithm

Consider action i at time t

- Total loss:

$$L_i^t = \sum_{s=1}^{t-1} \ell_i^s$$

- Weight:

$$w_i^t = w_i^1 e^{-\eta L_i^t}$$

  Note freedom to choose initial weight (w_i^1) $\sum_{i=1}^{n} w_i^1 = 1$.
- $\eta > 0$ is the learning rate parameter. Halving: $\eta \to \infty$
- Probability:

$$p_i^t = \frac{w_i^t}{\sum_{j=1}^{N} w_i^t}, \quad \mathbf{p}^t = \frac{\mathbf{w}^t}{\sum_{j=1}^{N} w_i^t}$$

Bound on the loss of **Hedge**(η)Algorithm

Theorem (main theorem)

For any sequence of loss vectors $\ell_1, \ldots, \ell_T$*, and for any* $i \in \{1, \ldots, N\}$*, we have*

$$L_{\mathbf{Hedge}(\eta)} \leq \frac{-\ln(w_i^1) + \eta L_i}{1 - e^{-\eta}}.$$

-
- Proof: by combining upper and lower bounds on $\sum_{i=1}^{N} w_i^{T+1}$

Comparing with the best distribution

- **Comparison class:** single experts. hindsite.
- Does not take advantage of multiple good experts.
- We will get tighter bounds by increasing the comparison class to include all **convex combinations** of the experts.

Recall Single step bound for **Hedge**(η)

The total weight has to decrease if the loss is large

$$\sum_{i=1}^{N} w_i^{t+1} \leq \left(\sum_{i=1}^{N} w_i^t\right) \left(1 - (1 - e^{-\eta})\mathbf{p}^t \cdot \ell_t\right)$$

Enlarging the comparison set

- Bound compares cumulative loss to that of best expert in hindsite.
- Does not take advantage of multiple good experts.
- We will get tighter bounds by comparing to the best convex combination of experts.

Comparing with the best distribution

- Denote by $\mathbf{q}$ an arbitrary distribution over N experts. $\mathbf{q} \in \Delta^N$. Distribution = convex combination.
- Compare loss of algorithm to loss of best convex combination of experts:

$$\sum_{t=1}^{T} L_A^t \leq +a \min_{\mathbf{q} \in \Delta^N} \sum_{t=1}^{T} \mathbf{q} \cdot \ell_t + cX$$

- When comparing to single best expert $X = \log N$
- **Intuition:** X should be small if best distribution $\mathbf{q}^*$ is close to initial distribution $\mathbf{p}^0$

Relative Entropy Bound

- KL-divergence or Relative Entropy: X
- For any distribution **q** and any iteration of **Hedge**(η):

Proof (from RE to ratio)

Proof (from ratio to bound)

Visual Intuition