Prediction with Expert Advice

Mehryar Mohri
Courant Institute and Google Research
mohri@cs.nyu.edu

Motivation

- On-line decisions, agent interacts with environment.
- Model: adversarial, no assumption about points being drawn from a distribution.
- Performance measure: regret, no risk or expected loss.

General On-Line Setting

- For $t = 1$ to T do
  - receive instance $x_t \in X$.
  - predict $\widehat{y}_t \in Y$.
  - receive label $y_t \in Y$.
  - incur loss $L(\widehat{y}_t, y_t)$.
- Classification: $Y = \{0, 1\}, L(y, y') = |y' - y|$.
- Regression: $Y \subseteq \mathbb{R}, L(y, y') = (y' - y)^2$.
- Objective: minimize total loss $\sum_{t=1}^{T} L(\widehat{y}_t, y_t)$.

Prediction with Experts

- For $t = 1$ to T do
  - receive instance $x_t \in X$ and advice $y_{t,i} \in Y, i \in [1, N]$.
  - predict $\widehat{y}_t \in Y$.
  - receive label $y_t \in Y$.
  - incur loss $L(\widehat{y}_t, y_t)$.
- Objective: minimize regret, i.e., difference of total loss incurred and that of best expert.

$$\text{Regret}(T) = \sum_{t=1}^{T} L(\widehat{y}_t, y_t) - \min_{i=1}^{N} L(\widehat{y}_{t,i}, y_t).$$

Weighted Majority Algorithm

- Algorithm: prediction with $N \geq 1$ experts, 0/1-loss.
  - at any time t, expert i has weight w_i^t.
  - originally, $w_i^0 = 1, \forall i \in [1, N]$.
  - prediction according to weighted majority.
  - weight of each wrong expert updated ($\epsilon > 0$):

$$w_i^{t+1} \leftarrow w_i^t (1 - \epsilon).$$

Weighted Majority - Bound

- Theorem (mistake bound): let m_i^t be the number of mistakes made by expert i till time t and m^t the total number of mistakes. Then, for all t and for any expert i (in particular best expert),

$$m^t \leq \frac{2 \log N}{\epsilon} + 2(1+\epsilon) m_i^t.$$

  - Thus, $m^t \leq O(\log N) + \text{constant} \times \text{best expert}.$
  - Realizable case: $m^t \leq O(\log N)$.

Weighted Majority - Proof

- Potential: $\Phi^t = \sum_{i=1}^N w_i^t$.
- Upper bound: after each error,

$$\Phi^{t+1} \leq \big[1/2 + 1/2\,(1-\epsilon)\big]\Phi^t = \big[1 - \epsilon/2\big]\Phi^t.$$

  Thus, $\Phi^t \leq \big(1-\epsilon/2\big)^{m^t} N$.
- Lower bound: for any expert i , $\Phi^t \geq w_i^t = (1-\epsilon)^{m_i^t}$.
- Comparison: $(1-\epsilon)^{m_i^t} \leq (1-\epsilon/2)^{m^t} N$

$$\Rightarrow m_i^t \log(1-\epsilon) \leq \log N + m^t \log(1-\epsilon/2)$$

$$\Rightarrow - \;\; m_i^t(\epsilon + \epsilon^2) \leq \log N - m^t \epsilon/2.$$

Exponential Weighted Average

total loss incurred by expert i up to time t

- Algorithm:
  - weight update: $w_{t+1,i} \leftarrow w_{t,i}\, e^{-\eta L(\widehat{y}_{t,i}, y_t)} = e^{-\eta L_{t,i}}$.
  - prediction: $\widehat{y}_t = \frac{\sum_{i=1}^N w_{t,i} y_{t,i}}{\sum_{i=1}^N w_{t,i}}$.
- Theorem: assume that L is convex in its first argument and takes values in $[0, 1]$. Then, for any $\eta > 0$ and any sequence $y_1, \ldots, y_T \in Y$, the regret at T satisfies

$$\text{Regret}(T) \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

For $\eta = \sqrt{8 \log N / T}$,

$$\boxed{\text{Regret}(T) \leq \sqrt{(T/2) \log N}}.$$

Exponential Weighted Avg - Proof

- Potential: $\Phi_t = \log \sum_{i=1}^N w_{t,i}$.
- Upper bound:

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \log \frac{\sum_{i=1}^N w_{t-1,i}\, e^{-\eta L(\widehat{y}_{t,i}, y_t)}}{\sum_{i=1}^N w_{t-1,i}} \\
&= \log \Big(\underset{w_{t-1}}{\mathrm{E}} [e^{-\eta L(\widehat{y}_{t,i}, y_t)}] \Big) \\
&\leq -\eta \underset{w_{t-1}}{\mathrm{E}} [L(\widehat{y}_{t,i}, y_t)] + \frac{\eta^2}{8} \quad \text{(Hoeffding's ineq.)} \\
&\leq -\eta L(\underset{w_{t-1}}{\mathrm{E}} [\widehat{y}_{t,i}], y_t) + \frac{\eta^2}{8} \quad \text{(convexity of first arg. of } L) \\
&= -\eta L(\widehat{y}_t, y_t) + \frac{\eta^2}{8}.
\end{aligned}$$

Exponential Weighted Avg - Proof

- **Upper bound:** summing up the inequalities yields

$$\Phi_T - \Phi_0 \leq -\eta \sum_{t=1}^{T} L(\widehat{y}_t, y_t) + \frac{\eta^2 T}{8}.$$

- **Lower bound:**

$$\Phi_T - \Phi_0 = \log \sum_{i=1}^{N} e^{-\eta L_{T,i}} - \log N \geq \log \max_{i=1}^{N} e^{-\eta L_{T,i}} - \log N$$

$$= -\eta \min_{i=1}^{N} L_{T,i} - \log N.$$

- **Comparison:**

$$-\eta \min_{i=1}^{N} L_{T,i} - \log N \leq -\eta \sum_{t=1}^{T} L(\widehat{y}_t, y_t) + \frac{\eta^2 T}{8}$$

$$\Rightarrow \sum_{t=1}^{T} L(\widehat{y}_t, y_t) - \min_{i=1}^{N} L_{T,i} \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

Exponential Weighted Avg - Notes

- Advantage: weight update does not depend on past predictions, but only on past performance.
- Disadvantage: choice of η requires knowledge of horizon T.

Doubling Trick

- Idea: divide time into periods $[2^k, 2^{k+1}-1]$ of length 2^k with $k=0, \ldots, n$, $T \geq 2^n - 1$, and choose $\eta_k = \sqrt{\frac{8 \log N}{2^k}}$ in each period.
- Theorem: with the same assumptions as before, for any T, the following holds:

$$\text{Regret}(T) \leq \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{(T/2) \log N} + \sqrt{\log N/2}.$$

Doubling Trick - Proof

- By the previous theorem, for any $I_k = [2^k, 2^{k+1} - 1]$,

$$L_{I_k} - \min_{i=1}^{N} L_{I_k,i} \leq \sqrt{2^k/2 \, \log N}.$$

Thus, $L_T = \sum_{k=0}^{n} L_{I_k} \leq \sum_{k=0}^{n} \min_{i=1}^{N} L_{I_k,i} + \sum_{k=0}^{n} \sqrt{2^k \, (\log N)/2}$

$$\leq \min_{i=1}^{N} L_{T,i} + \sum_{k=0}^{n} 2^{\frac{k}{2}} \sqrt{(\log N)/2}.$$

with

$$\sum_{i=0}^{n} 2^{\frac{k}{2}} = \frac{\sqrt{2}^{n+1} - 1}{\sqrt{2} - 1} = \frac{2^{(n+1)/2} - 1}{\sqrt{2} - 1} \leq \frac{\sqrt{2}\sqrt{T+1} - 1}{\sqrt{2} - 1} \leq \frac{\sqrt{2}(\sqrt{T} + 1) - 1}{\sqrt{2} - 1} \leq \frac{\sqrt{2}\sqrt{T}}{\sqrt{2} - 1} + 1.$$

Notes

- Doubling trick used in a variety of other contexts and proofs.
- More general method, learning parameter function of time: $\eta_t = \sqrt{(8 \log N)/t}$. Constant factor improvement:

$$\text{Regret}(T) \leq 2\sqrt{(T/2)\log N} + \sqrt{(1/8)\log N}.$$

Exp. Weighted Avg - Small Loss

- Cumulated loss: $L_T = \sum_{t=1}^T L_t = \sum_{t=1}^T L(\widehat{y}_t, y_t)$.
- Theorem: assume that L is convex in its first argument and takes values in $[0, 1]$. Then, for any $\eta > 0$ and any sequence $y_1, \ldots, y_T \in Y$, the cumulated loss L_T satisfies

$$L_T \leq \frac{\eta L_T^* + \log N}{1 - e^{-\eta}}.$$

For $\eta = 1$,

$$\boxed{L_T \leq \frac{L_T^* + \log N}{1 - 1/e}.}$$

Better bound when $L_T^* \leq O(\sqrt{T})$.

Small Loss - Proof

- **Potential:** $\Phi_t = \log \sum_{i=1}^N w_{t,i}$.
- **Upper bound:** use variance.

$$Y_t = X_t - \mathrm{E}[X_t]$$

X_t

$$\Phi_t - \Phi_{t-1} = \log \frac{\sum_{i=1}^N w_{t-1,i}\, e^{-\eta L(\widehat{y}_{t,i}, y_t)}}{\sum_{i=1}^N w_{t-1,i}} = \log \Big(\mathop{\mathrm{E}}_{w_{t-1}} [e^{-\eta L(\widehat{y}_{t,i}, y_t)}] \Big).$$

$$\begin{aligned}
\mathrm{E}[e^{-\eta Y_t}] &= 1 - \mathrm{E}[\eta Y_t] + \sum_{n=2}^{+\infty} \frac{(-\eta)^n}{n!} \mathrm{E}[Y_t^n] \\
&\leq 1 + \sigma^2 [e^{-\eta} - 1 + \eta] \\
&\leq 1 + \mathrm{E}[X_t](1 - \mathrm{E}[X_t])[e^{-\eta} - 1 + \eta] \\
&\leq 1 + \mathrm{E}[X_t][e^{-\eta} - 1 + \eta].
\end{aligned}$$

Small Loss - Proof

- Upper bound on difference of potential

$$\begin{aligned}\Phi_t - \Phi_{t-1} &= \log \operatorname*{E}_{w_{t-1}}[e^{-\eta X_t}] \\ &= \log \operatorname*{E}_{w_{t-1}}[e^{-\eta Y_t} e^{-\eta \, \mathrm{E}[X_t]}] \\ &\leq \mathrm{E}[X_t][e^{-\eta} - 1 + \eta] - \eta \, \mathrm{E}[X_t] \\ &= \mathrm{E}[X_t][e^{-\eta} - 1] \\ &\leq L(\widehat{y}_t, y_t)[e^{-\eta} - 1] \quad \text{(Jensen's ineq.)}.\end{aligned}$$

Thus, $\Phi_T - \Phi_0 \leq L_T[e^{-\eta} - 1]$.

- Lower bound (proof of a previous theorem):

$$\Phi_T - \Phi_0 \geq -\eta L_T^* - \log N.$$

Small Loss - Better Bound

- Corollary: assume that L is convex in its first argument and takes values in $[0, 1]$. Then, for the choice $\eta = \log\left(1 + \sqrt{(2 \log N)/L_T^*}\right)$ and any sequence $y_1, \ldots, y_T \in Y$, the regret satisfies

$$\text{Regret}(T) \leq \sqrt{2L_T^* \log N} + \log N.$$

Better bound when $L_T^* \leq O(T)$.

Better Bound - Proof

- Use inequality $\eta \leq (e^\eta - e^{-\eta})/2$ in theorem to bound η in the numerator:

$$
\begin{aligned}
L_T &\leq \frac{\eta L_T^* + \log N}{1 - e^{-\eta}} \\
&\leq \frac{e^\eta - e^{-\eta}}{1 - e^{-\eta}} L_T^*/2 + \frac{\log N}{1 - e^{-\eta}} \\
&= \frac{e^\eta - 1 + 1 - e^{-\eta}}{1 - e^{-\eta}} L_T^*/2 + \frac{\log N}{1 - e^{-\eta}} \\
&= (e^\eta + 1) L_T^*/2 + \frac{\log N}{1 - e^{-\eta}} \\
&= (1/u + 1) L_T^*/2 + \frac{\log N}{1 - u} = f(u). \qquad (u = e^{-\eta})
\end{aligned}
$$

Better Bound - Proof

- Differentiating f and setting it to zero gives:

$$f'(u) = -\frac{L_T^*}{2u^2} + \frac{\log N}{(u-1)^2} = 0$$

$$\Leftrightarrow u^2(2\log N/L_T^* - 1) + 2u - 1 = 0.$$

$$\Delta' = 1 + 2\log N/L_T^* - 1 = 2\log N/L_T^*.$$

Since $u = e^{-\eta} > 0,$ it is equal to the positive root:

$$u = \frac{-1 + \sqrt{(2\log N)/L_T^*}}{(2\log N)/L_T^* - 1} = \frac{1}{\sqrt{(2\log N)/L_T^*} + 1}.$$

General Case

- Potential Φ_t .
- Predictions:

$$\widehat{y}_t = \frac{\sum_{i=1}^{N} \nabla\Phi(L_{t-1} - L_{t-1,i}) y_{t,i}}{\sum_{i=1}^{N} \nabla\Phi(L_{t-1} - L_{t-1,i})}.$$