1 Online Learning

For \( t = 1 \) to \( n \)

Instance \( x_t \in X \) is provided

Learner picks \( \hat{y}_t \in Y \) (or randomized version \( q_t \in \Delta(Y) \))

True label \( y_t \in Y \) is revealed and learner pays loss \( \ell(\hat{y}_t, y_t) \)

end

\[
R_n = \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in F} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t)
\]

If we use randomized algorithm then, on each round, label \( \hat{y}_t \) is drawn from \( q_t \). In this case, we wish to bound regret defined as :

\[
R_n = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] - \inf_{f \in F} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t)
\]

A simple application of Hoeffding-Azuma can in fact turn the above statement into a high probability statement of form, for any \( \delta > 0 \) with probability at least \( 1 - \delta \) over the randomization of the learning algorithm,

\[
R_n \leq \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} [\ell(\hat{y}_t, y_t)] - \inf_{f \in F} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \sqrt{\frac{\log 1/\delta}{n}}
\]

1.1 Halving : Realizable Online Binary Classification, finite class \( F \)

Assume \( Y = \{ \pm 1 \} \). Also assume that \( y_t = f^*(x_t) \) where \( f^* \in F \) is unknown to the learner.

At round \( t \) given \( x_t \) predict with majority of consistent hypotheses. That is given past data define set of consistent hypotheses as

\[
F_t = \{ f \in F : \forall i < t, f(x_i) = y_i \}
\]

Given \( x_t \) we predict :

\[
\hat{y}_t = \text{sign} \left( \sum_{f \in F_t} f(x_t) \right)
\]
For the above procedure, we have that

$$\frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) \leq \frac{\log_2 |\mathcal{F}|}{n}$$

Why? Notice that if we make a mistake in our prediction at round $t$, $|\mathcal{F}_{t+1}| \leq \frac{1}{2} |\mathcal{F}_t|$. Hence total number of mistakes can’t be larger than $\log_2 |\mathcal{F}|$

2 **Experts/Exponential Weights Algorithm**

Algorithm: $q_1(f) = 1/|F|$. Further, each round we update the distribution over experts as,

$$q_{t+1}(f) \propto q_t(f)e^{-\eta \ell(f(x_t), y_t)}$$

Or in other words, $q_{t+1}(f) = \frac{e^{-\eta \sum_{i=1}^{t} \ell(f(x_i), y_i)}}{\sum_{f \in \mathcal{F}} e^{-\eta \sum_{i=1}^{t} \ell(f(x_i), y_i)}}$

**Claim 1.**

$$\sum_{t=1}^{n} \mathbb{E}_{f \sim q_t} [\ell(f(x_t), y_t)] - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \sqrt{\frac{2 \log |\mathcal{F}|}{n}}$$

**Proof.** We use the notation $L_t(f) = \sum_{i=1}^{t} \ell(f(x_i), y_i)$. Define $W_0 = |F|$ and define $W_t = \sum_{f \in \mathcal{F}} e^{-\eta L_t(f)}$. Note that

$$\log \left( \frac{W_n}{W_0} \right) = \log \left( \sum_{f \in \mathcal{F}} e^{-\eta L_n(f)} \right) - \log |\mathcal{F}|$$

$$\geq \log \left( \max_{f \in \mathcal{F}} e^{-\eta L_n(f)} \right) - \log |\mathcal{F}|$$

$$= -\eta \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \log |\mathcal{F}|$$
On the other hand,
\[
\log \left( \frac{W_n}{W_0} \right) = \sum_{t=1}^{n} \log \left( \frac{W_t}{W_{t-1}} \right) = \sum_{t=1}^{n} \log \left( \frac{\sum_{f \in \mathcal{F}} e^{-\eta L_t(f)}}{\sum_{f \in \mathcal{F}} e^{-\eta L_{t-1}(f)}} \right) \\
= \sum_{t=1}^{n} \log \left( \sum_{f \in \mathcal{F}} \sum_{f \in \mathcal{F}} e^{-\eta L_{t-1}(f)} e^{-\eta \ell(f(x_t), y_t)} \right) \\
= \sum_{t=1}^{n} \log \left( \mathbb{E}_{f \sim q_t} \left[ e^{-\eta \ell(f(x_t), y_t)} \right] \right) \\
= \sum_{t=1}^{n} \log \left( \mathbb{E}_{f \sim q_t} \left[ e^{-\eta \ell(f(x_t), y_t) - \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)]} - \eta \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)] \right] \right) \\
= \sum_{t=1}^{n} \log \left( \mathbb{E}_{f \sim q_t} \left[ e^{-\eta \ell(f(x_t), y_t) - \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)]} \right] \right) \times e^{-\eta \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)]} \\
= \sum_{t=1}^{n} \log \left( \mathbb{E}_{f \sim q_t} \left[ e^{-\eta \ell(f(x_t), y_t) - \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)]} \right] \right) - \eta \sum_{t=1}^{n} \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)]
\]

Thus we conclude that
\[
\sum_{t=1}^{n} \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)] - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \frac{\log |\mathcal{F}|}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \log \left( \mathbb{E}_{f \sim q_t} \left[ e^{-\eta \ell(f(x_t), y_t) - \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)]} \right] \right)
\]

Note that for any zero mean RV \(X\) in the range \([-1, 1]\), \(\mathbb{E}[e^{-\eta X}] \leq e^{-\eta^2 / 2}\). Hence,
\[
\sum_{t=1}^{n} \mathbb{E}_{f \sim q_t}[\ell(f(x_t), y_t)] - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \frac{\log |\mathcal{F}|}{\eta} + \frac{\eta}{2} \frac{n}{2}
\]

Picking \(\eta = \sqrt{2 \log |\mathcal{F}| / n}\) concludes the statement. \(\square\)

3 Learning Thresholds

Not learnable, (even in realizable case) why?

4 Predicting Bit-sequences

Think of the online learning problem where on each round \(t\) we predict the next bit \(y_t \in \{\pm 1\}\). Also say \(\mathcal{F} \subset \{\pm 1\}^n\) and we want to minimize regret (in expectation):

\[
\text{Reg}_n = \frac{1}{n} \sum_{t=1}^{n} 1_{\{\hat{y}_t \neq y_t\}} - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} 1_{\{f_t \neq y_t\}}
\]

When can we ensure \(\mathbb{E}[\text{Reg}_n] \to 0\)? Let us denote the minimax rate as

\[
V_n = \min_{\text{algorithms sequence}} \max_{\text{sequence}} \mathbb{E}[\text{Reg}_n]
\]
Claim 2.

\[ V_n = \frac{1}{2n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^{n} f_t \epsilon_t \right] \]

Proof. The basic idea is to write down the minimax rate in a recursive form and get a characterization for it. To this end, say you had already played rounds 1 to \( n-1 \) optimally, then, on the last two rounds, what are the optimal moves for both the players. We write this value given \( y_1, \ldots, y_{n-1} \) were already produced as:

\[ V_n(y_1, \ldots, y_{n-1}) = \min_{q_n \in [0,1]} \sup_{y_n \in \{\pm 1\}} \left\{ \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbb{1}_{\hat{y}_n \neq y_n} \right] - \inf_{f \in F} \sum_{t=1}^{n} \mathbb{1}_{f_t \neq y_t} \right\} \]

That is on the last round, the learner picks distribution \( q_n \) that minimizes loss at the last step while the adversary picks \( y_n \) that maximizes the loss at last step while also minimizes loss of the target we are comparing our regret against. In fact if we define \( V_n(y_1, \ldots, y_n) = - \inf_{f \in F} \sum_{t=1}^{n} \mathbb{1}_{f_t \neq y_t} \) then we see that

\[ V_n(y_1, \ldots, y_{n-1}) = \min_{q_n \in [0,1]} \sup_{y_n \in \{\pm 1\}} \left\{ \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbb{1}_{\hat{y}_n \neq y_n} \right] + V_n(y_1, \ldots, y_n) \right\} \]

Thus we see that,

\[ V_n(y_1, \ldots, y_{n-1}) = \min_{q_n \in [0,1]} \sup_{y_n \in \{\pm 1\}} \left\{ q_n \mathbb{1}_{1 \neq y_n} + (1 - q_n) \mathbb{1}_{1 = y_n} + V_n(y_1, \ldots, y_n) \right\} \\
= \min_{q_n \in [0,1]} \max \left\{ (1 - q_n) + V_n(y_1, \ldots, y_{n-1}, +1), q_n + V_n(y_1, \ldots, y_{n-1}, -1) \right\} \]

Solution is to pick \( q_n \) such that the two terms are equal. Hence

\[ V_n(y_1, \ldots, y_{n-1}) = \frac{1}{2} + \frac{V_n(y_1, \ldots, y_{n-1}, +1) + V_n(y_1, \ldots, y_{n-1}, -1)}{2} \]

\[ = \frac{1}{2} + \mathbb{E}_{\epsilon_n} [V_n(y_1, \ldots, y_{n-1}, \epsilon_n)] \]

Now recursively we continue as

\[ V_n(y_1, \ldots, y_{n-2}) = \min_{q_{n-1} \in [0,1]} \sup_{y_{n-1} \in \{\pm 1\}} \left\{ q_{n-1} \mathbb{1}_{1 \neq y_{n-1}} + (1 - q_{n-1}) \mathbb{1}_{1 = y_{n-1}} + V_n(y_1, \ldots, y_{n-1}) \right\} \\
= \frac{1}{2} + \mathbb{E}_{\epsilon_{n-1}} [V_n(y_1, \ldots, y_{n-2}, \epsilon_{n-1})] \]

Proceeding as follows we conclude that:

\[ V_n(\cdot) = \mathbb{E}_{\epsilon_1} [V_n(\epsilon_1)] = \ldots = \mathbb{E}_{\epsilon} [V_n(\epsilon_1, \ldots, \epsilon_n)] \]

Hence we conclude that:

\[ \text{Minimax}_n = \frac{V_n(\cdot)}{n} = \frac{1}{2} + \frac{1}{n} \mathbb{E}_{\epsilon} \left[ - \inf_{f \in F} \sum_{t=1}^{n} \mathbb{1}_{f_t \neq \epsilon_t} \right] = \frac{1}{2} + \frac{1}{2n} \mathbb{E}_{\epsilon} \left[ \sup_{f \in F} \sum_{t=1}^{n} f_t \epsilon_t \right] - \frac{1}{2} \]

\( \Box \)
Prediction algorithm: the prediction algorithm corresponding to the above analysis is exactly the $q_t$ that minimizes the recursion at each step and hence is given by

$$q_t = \arg\min_{q \in [0,1]} \max_{y_t \in \{\pm 1\}} \left\{ \mathbb{E}_{y_t \sim q} \left[ \mathbf{1}_{\{y_t \neq y_t\}} \right] + V_n(y_1, \ldots, y_t) \right\}$$

$$= \frac{1}{2} (1 + V_n(y_1, \ldots, y_t, +1) - V_n(y_1, \ldots, y_t, -1))$$

$$= \frac{1}{2} (1 + \mathbb{E}_{\epsilon_{t+1:n}} [V_n(y_1, \ldots, y_{t-1}, +1, \epsilon_{t+1}, \ldots, \epsilon_n)] - \mathbb{E}_{\epsilon_{t+1:n}} [V_n(y_1, \ldots, y_{t-1}, -1, \epsilon_{t+1}, \ldots, \epsilon_n)])$$

In fact, we can also show that the following randomized algorithm works. Draw $\epsilon_{t+1}, \ldots, \epsilon_n$ and set:

$$q_t = \frac{1}{2} \left( 1 + \inf_{f \in F} \left\{ \sum_{j=1}^{t-1} \mathbb{1}_{\{f_j \neq y_t\}} + \mathbb{1}_{\{f_t = 1\}} + \sum_{i=t+1}^{n} \mathbb{1}_{\{f_i \neq \epsilon_i\}} \right\} - \inf_{f \in F} \left\{ \sum_{j=1}^{t-1} \mathbb{1}_{\{f_j \neq y_t\}} + \mathbb{1}_{\{f_t = -1\}} + \sum_{i=t+1}^{n} \mathbb{1}_{\{f_i \neq \epsilon_i\}} \right\} \right)$$