1 Bit Prediction

Claim 1. There exists a randomized prediction strategy that ensures that

$$\mathbb{E}[\text{Reg}_n] \leq \frac{1}{2n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t \right]$$

To prove the above claim we first prove this following lemma, a result by Thomas Cover.

Lemma 2 (T. Cover’65). Let $\phi : \{\pm 1\}^n \mapsto \mathbb{R}$ be a function such that, for any $i$, and any $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$,

$$|\phi(y_1, \ldots, y_{i-1}, +1, y_{i+1}, \ldots, y_n) - \phi(y_1, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_n)| \leq \frac{1}{n}, \text{(stability condition)}$$

then, there exists a randomized strategy such that for any sequence of bits,

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} [1\{\hat{y}_t \neq y_t\}] \leq \phi(y_1, \ldots, y_n)$$

if and only if,

$$\mathbb{E}_\epsilon \phi(\epsilon_1, \ldots, \epsilon_n) \geq \frac{1}{2}$$

and further, the strategy achieving this bound on expected error is given by:

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

Proof of Lemma.
We start by proving that if there exists an algorithm that guarantees that

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} [1\{\hat{y}_t \neq y_t\}] \leq \phi(y_1, \ldots, y_n)$$

then, $\mathbb{E}_\epsilon[\phi(\epsilon_1, \ldots, \epsilon_n)] \geq 1/2$.

To see this, note that the regret bound implies that

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} [1\{\hat{y}_t \neq y_t\}] - \phi(y_1, \ldots, y_n) \leq 0$$
for any \( y_1, \ldots, y_n \). Now simply let the adversary pick \( y_t = \epsilon_t \) as a Rademacher random variable. Thus, taking expectation, this implies that,

\[
0 \geq \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{y_t \sim q_t} \left[ \mathbb{E}_c \mathbf{1}\{\hat{y}_t \neq \epsilon_t\} \right] - \mathbb{E}_c \phi(\epsilon_1, \ldots, \epsilon_n) = \frac{1}{2} - \mathbb{E}_c \phi(\epsilon_1, \ldots, \epsilon_n)
\]

Next we prove that if \( \mathbb{E}_c \phi(\epsilon_1, \ldots, \epsilon_n) \geq \frac{1}{2} \), then \( \exists \) strategy s.t. \( \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{y_t \sim q_t} \left[ \mathbf{1}\{\hat{y}_t \neq y_t\} \right] \leq \phi(y_1, \ldots, y_n) \).

The basic idea is to prove this statement starting from \( n \) and moving backwards. Say we have already played rounds up until round \( n - 1 \) and have observed \( y_1, \ldots, y_{n-1} \). Now let us consider the last round. On the last round we use,

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

Now note that if \( y_n = +1 \) then \( \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n \neq y_n\} \right] = \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n = -1\} \right] = 1 - q_n \) and if \( y_n = -1 \) then \( \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n \neq y_n\} \right] = q_n \) and hence for the choice of \( q_n \) above, we can write

\[
\mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n \neq y_n\} \right] = \frac{1}{2n} - \frac{y_n}{2} (\phi(y_1, \ldots, y_{n-1}, -1) - \phi(y_1, \ldots, y_{n-1}, +1))
\]

Plugging in the above, note that for any \( y_n \) (possibly chosen adversarially looking at \( q_n \)), we have,

\[
\frac{1}{n} \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n \neq y_n\} \right] - \phi(y_1, \ldots, y_n)
\]

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

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

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

Thus we can conclude that,

\[
\frac{1}{n} \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n \neq y_n\} \right] + \frac{1}{n} \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n \neq y_n\} \right] - \phi(y_1, \ldots, y_n)
\]

\[
= \frac{1}{2n} + \frac{1}{n} \mathbb{E}_{\hat{y}_n \sim q_n} \left[ \mathbf{1}\{\hat{y}_n \neq y_n\} \right] - \mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-1}, \epsilon_n)
\]

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

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

\[
= \frac{2}{2n} - \mathbb{E}_{\epsilon_n-1, \epsilon_n} \phi(y_1, \ldots, y_{n-2}, \epsilon_{n-1}, \epsilon_n)
\]
Proceeding in similar way we conclude that,

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\tilde{y}_t \sim q_t} \left[ 1_{\{\tilde{y}_t \neq y_t\}} \right] - \phi(y_1, \ldots, y_n) \leq \frac{n}{2n} - \mathbb{E}_{\epsilon_1, \ldots, \epsilon_n} \phi(\epsilon_1, \ldots, \epsilon_n) = \frac{1}{2} - \mathbb{E}_{\epsilon_1, \ldots, \epsilon_n} \phi(\epsilon_1, \ldots, \epsilon_n)$$

Hence, if \(\mathbb{E}_{\epsilon_1, \ldots, \epsilon_n} \phi(\epsilon_1, \ldots, \epsilon_n) \geq 1/2\) then we can conclude that, \(\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\tilde{y}_t \sim q_t} \left[ 1_{\{\tilde{y}_t \neq y_t\}} \right] \leq \phi(y_1, \ldots, y_n)\) as desired.

Hence we conclude the proof of this lemma. \(\square\)

2 Application: Binary Node Classification

Let \(G = (V, E)\) be a known undirected graph representing a social network. At each time step \(t\), a user in the network opens her Facebook page, and the system needs to decide whether to classify the user as type “−1” or “+1”, say, in order to decide on an advertisement to display. We assume here that the feedback on the “correct” type is revealed to the system after the prediction is made. Suppose we have a hunch that the type of the user (+1 or −1) is correlated with the community to which she belongs. For simplicity, suppose there are two communities, more densely connected within than across. To capture the idea of correlating communities and labels, we set \(\phi\) to be small on labelings that assign homogenous values within each community. We make the following simplifying assumptions: (i) \(|V| = n\), (ii) we only predict the label of each node once, and (iii) the order in which the nodes are presented is fixed (this assumption is easily removed). Smoothness of a labeling \(f \in \{\pm 1\}^n\) with respect to the graph may be computed via

$$\text{Cut}(f) = \sum_{(u,v) \in E} 1_{\{f_u \neq f_v\}} = \frac{1}{4} \sum_{(u,v) \in E} (f_u - f_v)^2 = f^T L f$$  (3)

where \(L = D - A\), the diagonal matrix \(D\) contains degrees of the nodes, and \(A\) is the adjacency matrix and \(f_v \in \{\pm 1\}\) is the label in \(f\) that corresponds to vertex \(v \in V\). This function in (3) counts the number of disagreements in labels at the endpoints of each edge. The value is also known as the size of the cut induced by \(f\) (the smallest possible being \(\text{MinCut}\)). As desired, the function in (3) gives a smaller value to the labelings that are homogenous within the communities.

Unfortunately, the function \(\text{Cut}(f)\) is not stable. Further, the cut size is \(n - 1\) for a star graph, where \(n - 1\) nodes, labeled as +1, are connected to the center node, labeled as −1. The large value
of the cut does not capture the simplicity of this labeling, which is only one bit away from being a constant +1. Instead, we opt for the indirect definition:

$$ F_{\kappa} = \left\{ f \in \{\pm 1\}^n : f^T L f \leq \kappa \right\} $$

(4)

for \( \kappa \geq 0 \), and then set

$$ \phi(y_1, \ldots, y_n) = \inf_{f \in F_{\kappa}} \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}_{\{f_t \neq y_t\}} + \frac{1}{2n} \mathbb{E}_\epsilon \left[ \sup_{f \in F_{\kappa}} \sum_{t=1}^{n} f_t \epsilon_t \right] $$

(5)

Parameter \( \kappa \) should be larger than the value of \( \text{MinCut} \), for otherwise the set \( F_{\kappa} \) is empty. This gives an interesting algorithm for the prediction problem . . . . What does this look like?

Well we want to use the strategy

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

$$ = \frac{1}{2} + n \frac{1}{2} \mathbb{E}_{\epsilon_{t+1}, \ldots, \epsilon_n} \left[ \inf_{f \in F_{\kappa}} \left\{ \frac{1}{n} \sum_{j=1}^{t-1} \mathbf{1}_{\{f_j \neq y_j\}} + \mathbf{1}_{\{f_t \neq -1\}} + \sum_{j=t+1}^{n} \mathbf{1}_{\{f_j \neq \epsilon_j\}} \right\} \right] $$

$$ - \inf_{f \in F_{\kappa}} \left\{ \frac{1}{n} \sum_{j=1}^{t-1} \mathbf{1}_{\{f_j \neq y_j\}} + \mathbf{1}_{\{f_t \neq +1\}} + \sum_{j=t+1}^{n} \mathbf{1}_{\{f_j \neq \epsilon_j\}} \right\} $$

It turns out that by concentration inequalities, it even suffices to take a single new sample of \( \epsilon_{t+1}, \ldots, \epsilon_n \) for round \( t \) to compute \( q_t \) above. In this case the underlying strategy is peculiar: At time \( t \), to predict label for vertex \( v_t \), we fill seen entries by labels, unseen entries by random \( \epsilon_v \)’s and solve two optimization problems. One with labels set as mentioned and with label of \( v_t \) set to \(-1\) we solve for \( \inf_{f \in F_{\kappa}} \left\{ \frac{1}{n} \sum_{j=1}^{t-1} \mathbf{1}_{\{f_j \neq y_j\}} + \mathbf{1}_{\{f_t \neq -1\}} + \sum_{j=t+1}^{n} \mathbf{1}_{\{f_j \neq \epsilon_j\}} \right\} \). Now we do the optimization with only changing the label of \( v_t \) to a +1. We can then set \( q_t \) by equation above. Here once can view the random signs we draw as a kind of regularization or protection against worst case adversarial future.

Of course two natural questions follow. First, what if outcomes are not binary. We will see this in the following section. Second, what if we did not know the graph in advance or worse yet the graph evolves with time, or more generally what if we didnt have just bit prediction but rather prediction of bit given some input \( x_t \) like in the classification setting?

3 “Dice Prediction” or the Multi-class case

The key idea for a strategy for bit prediction was to work backward and design a \( q_n \) such that for any \( y_n \in \{\pm 1\} \), the value of

$$ \frac{1}{n} \mathbb{E}_{y_n \sim q_n} \left[ \mathbf{1}_{\{y_n \neq y_n\}} \right] - \phi(y_1, \ldots, y_n) $$

is the same and this is what stability condition helps ensure. Can we do this for the multi-class case?
Let’s try to visualize this. Say we have $K$ classes and say we have already played for $n - 1$ rounds and hence have received $y_1, \ldots, y_{n-1}$. In this case, for each value of $a \in [K]$ we have corresponding $\phi(y_1, \ldots, y_{n-1}, a)$. Further, for a given distribution $q_n \in \Delta([K])$, note that

$$\frac{1}{n} \mathbb{E}_{\hat{y}_n \sim q_n} [1_{\{\hat{y}_n \neq a\}}] = \frac{1}{n} (1 - q_n[a])$$

that is, given by probability of not choosing $a$ as label. Hence the $q_n$ we desire is one where for all $a \in [K]$, we have the same value for

$$\frac{1}{n} (1 - q_n[a]) - \phi(y_1, \ldots, y_{n-1}, a)$$

Say this value is $C$. That is, $q_n$ is such that, for every $a \in [K]$, $\frac{1}{n} (1 - q_n[a]) - \phi(y_1, \ldots, y_{n-1}, a) = C$. In this case, note that $q_n[a] = 1 - n\phi(y_1, \ldots, y_{n-1}, a) - nC$. Since $\sum_{a=1}^K q_n[a] = 1$ this implies that,

$$K - n \sum_{a=1}^K \phi(y_1, \ldots, y_{n-1}, a) - nKC = 1$$

Thus we conclude that

$$C = \frac{1}{n} - \frac{1}{nK} - \frac{1}{K} \sum_{a=1}^K \phi(y_1, \ldots, y_{n-1}, a)$$

and so for any $y_n \in [K]$, our randomized prediction is given by:

$$q_n[y_n] = \frac{1}{K} - n \left( \phi(y_1, \ldots, y_n) - \frac{1}{K} \sum_{a=1}^K \phi(y_1, \ldots, y_{n-1}, a) \right)$$

Convince yourself that the above is exactly Cover’s algorithm for binary case!

The solution also gives us our stability condition for the multi-class case. The condition is basically the one that makes $q_n$ a valid probability distribution. We already ensured that it sums to 1. To ensure it is non-negative we require the stability condition that for all $y' \in [K]$,

$$\frac{1}{K} - n \left( \phi(y_1, \ldots, y_{n-1}, y') - \frac{1}{K} \sum_{a=1}^K \phi(y_1, \ldots, y_{n-1}, a) \right) \geq 0$$

or in other words,

$$\max_{y'} \phi(y_1, \ldots, y_{n-1}, y') - \frac{1}{K} \sum_{a=1}^K \phi(y_1, \ldots, y_{n-1}, a) \leq \frac{1}{Kn}$$

Again verify that this coincides with stability condition in the binary case.

We now present the multi-class extension of Cover’s result.

**Lemma 3.** Suppose $\phi : [K]^n \mapsto \mathbb{R}$ satisfies the stability condition that

$$\max_{y' \in [K]} \phi(\ldots, y', \ldots) - \frac{1}{K} \sum_{a=1}^K \phi(\ldots, a, \ldots) \leq \frac{1}{nK}$$
then there exists a randomized strategy such that for any sequence of \( y_1, \ldots, y_n \in [K] \) (possibly adversarially chosen),

\[
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{q_t \sim q_t} [ \mathbb{1}\{\hat{y}_t \neq y_t\}] \leq \phi(y_1, \ldots, y_n)
\]

if and only if,

\[
\mathbb{E}_{\varepsilon_1, \ldots, \varepsilon_n \sim \text{Unif}[K]^n} \phi(\varepsilon_1, \ldots, \varepsilon_n) \geq 1 - \frac{1}{K}
\]

and further, the strategy achieving this bound on expected error is given by:

\[
q_t[y] = \frac{1}{K} - \frac{n}{K} \mathbb{E}_{\varepsilon_{t+1}, \ldots, \varepsilon_n \sim \text{Unif}[K]^{n-t}} \left[ \phi(y_1, \ldots, y_{t-1}, y, \varepsilon_{t+1}, \ldots, \varepsilon_n) - \frac{1}{K} \sum_{a=1}^{K} \phi(y_1, \ldots, y_{t-1}, a, \varepsilon_{t+1}, \ldots, \varepsilon_n) \right]
\]

Proof. The “only if” direction is as before and simple. If one picks \( y \)'s from uniform distribution then irrespective of what algorithm we use, \( \mathbb{E}_{y_1, \ldots, y_n} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{q_t \sim q_t} [ \mathbb{1}\{\hat{y}_t \neq y_t\}] = 1 - \frac{1}{K} \). and so if the bound on average error holds for any algorithm then,

\[
1 - \frac{1}{K} \leq \mathbb{E}_{\varepsilon_1, \ldots, \varepsilon_n \sim \text{Unif}[K]^n} \phi(\varepsilon_1, \ldots, \varepsilon_n)
\]

Now lets look at the other direction. For the last step we already showed that for any \( y_n \in [K] \), for the \( q_n \) proposed,

\[
\frac{1}{n} \mathbb{E}_{y_n \sim q_n} [ \mathbb{1}\{\hat{y}_n \neq y_n\}] - \phi(y_1, \ldots, y_n) = C = \frac{1}{n} - \frac{1}{nK} - \frac{1}{K} \sum_{a=1}^{K} \phi(y_1, \ldots, y_{n-1}, a) = \frac{1}{n} \left( 1 - \frac{1}{K} \right) - \mathbb{E}_{\varepsilon_n \sim \text{Unif}[K]} \left[ \phi(y_1, \ldots, y_{n-1}, \varepsilon_n) \right]
\]

Now for the next step, we need to choose a distribution \( q_{n-1} \) such that for any \( y_{n-1} \in [K] \),

\[
\frac{1}{n} \mathbb{E}_{y_{n-1} \sim q_{n-1}} [ \mathbb{1}\{\hat{y}_{n-1} \neq y_{n-1}\}] - \mathbb{E}_{\varepsilon_n \sim \text{Unif}[K]} \left[ \phi(y_1, \ldots, y_{n-2}, y_{n-1}, \varepsilon_n) \right] + \frac{1}{n} \left( 1 - \frac{1}{K} \right)
\]

has the same value. Just as we did for the case of \( n \) (think of it as \( \phi(y_1, \ldots, y_{n-1}, y_n) \) being replaced by \( \mathbb{E}_{\varepsilon_n} \left[ \phi(y_1, \ldots, y_{n-1}, \varepsilon_n) \right] \)), we find that this values is given by

\[
\frac{2}{n} \left( 1 - \frac{1}{K} \right) - \mathbb{E}_{\varepsilon_{n-1}, \varepsilon_n \sim \text{Unif}[K]^2} \left[ \phi(y_1, \ldots, y_{n-2}, y_{n-1}, \varepsilon_n) \right]
\]

and the corresponding choice of \( q_{n-1} \) is as given in lemma statement. Proceeding in similar way to \( n - 2 \) and so on till \( n - n \) we get that,

\[
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{q_t \sim q_t} [ \mathbb{1}\{\hat{y}_t \neq y_t\}] - \phi(y_1, \ldots, y_n) \leq \left( 1 - \frac{1}{K} \right) - \mathbb{E}_{\varepsilon_1, \ldots, \varepsilon_n \sim \text{Unif}[K]^n} \left[ \phi(\varepsilon_1, \ldots, \varepsilon_n) \right]
\]

Since we started with the premise that \( 1 - \frac{1}{K} \leq \mathbb{E}_{\varepsilon_1, \ldots, \varepsilon_n \sim \text{Unif}[K]^n} \phi(\varepsilon_1, \ldots, \varepsilon_n) \), this implies that

\[
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{q_t \sim q_t} [ \mathbb{1}\{\hat{y}_t \neq y_t\}] - \phi(y_1, \ldots, y_n) \leq 0
\]

and hence the Lemma is proved. \( \square \)
Once we have the above lemma, using $\phi(y_1, \ldots, y_n) = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} 1_{\{f_t \neq y_t\}} + \frac{1}{n} \mathbb{E} \varepsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t[\varepsilon_t] \right] - \frac{1}{K}$ which satisfies the stability condition. This yields the following Corollary.

**Corollary 4.** There exists a randomized prediction strategy that ensures that

$$\mathbb{E} [\text{Reg}_n] \leq \frac{2}{n} \mathbb{E}_{\varepsilon, \epsilon} \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \cdot f_t[\varepsilon_t] \right]$$

This for finite class of course yields the upper bound $O(\sqrt{\log |\mathcal{F}|/n})$. 