Machine Learning Theory (CS 6783)

Lecture 2: Bit Prediction Cover's Result

1 Bit Prediction Problem

Recap: n round bit prediction game where Player I (the learner) on each round t produces (possibly randomized) prediction $\hat{y}_t \in \{\pm 1\}$, to predict bit $y_t \in \{\pm 1\}$ produced by Player II (nature or adversary). We started with a simple goal of developing an algorithm that minimizes the following notion of regret against any adversary (strategy for player II).

$$\sum_{t=1}^{n} \mathbf{1}\{\hat{y}_t \neq y_t\} - \min_{b \in \{\pm 1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\}$$

We have the following facts about the above goal:

1. No algorithm can guarantee expected regret better than $\sqrt{n}$. (If y's were generated by random coin flips, any algorithm would suffer the regret of $\sqrt{n}$).

2. Any deterministic algorithm will suffer a regret of order n

3. Randomized algorithm that produces $\hat{y}_t$ by sampling according to past frequency of $+1$ does not work and suffers regret of order n.

So whats the best we can do?

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

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{b \in \{\pm 1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\} \leq \frac{\sqrt{n}}{2}$$

against any adversary!

Specifically this means that we have a strategy that never looses worse than $\sqrt{n}$ against any adversary (which is the best we could hope for even for optimal) and further, if we have uneven number of heads than tails, we can win significantly more.

To prove the above claim and much more, we first prove this following lemma, a result by Thomas Cover (all the way back in 1965). In fact, the more general question we will answer will be roughly in the form: For what function ϕ's is it possible to ensure that, there exists forecaster s.t.,

```
for any sequence,
        number of mistakes made by forecaster ≤ ϕ(sequence).
```

The function ϕ controlling the number of mistakes is a measure of "complexity" or "predictiveness" of the sequence. It captures our prior belief of what kinds of patterns might appear. For the Penny-Matching game, ϕ may be related to the frequency of heads vs tails, or more fine-grained statistics, such as predictability of the next outcome based on the last three outcomes. In fact, Shannon's mind reading machine was based on only 8 such states. Which ϕ can one choose? How to develop an efficient algorithm for a given ϕ?

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

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

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

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

if and only if,

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

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

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

Once we have the above lemma, using

$$\phi(y_1, \ldots, y_n) = \min_{b \in \{\pm 1\}} \sum_{t=1}^{n} \mathbb{1}_{\{b \neq y_t\}} + \frac{\sqrt{n}}{2}$$

we will conclude the result

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

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

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

To see this, note that the regret bound implies that

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

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,

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

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

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,

$$q_n = \frac{1}{2} + \frac{1}{2}\left(\phi(y_1,\ldots,y_{n-1},-1) - \phi(y_1,\ldots,y_{n-1},+1)\right)$$

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

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

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

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

Now recursively we continue just as above for $n-1$ to 0. Let us do the $n-1$th step and the rest follows. To this end, note that just as earlier, if $y_{n-1} = +1$ then $\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\left[\mathbb{1}_{\{\hat{y}_{n-1}\neq y_{n-1}\}}\right] = \mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\left[\mathbb{1}_{\{\hat{y}_{n-1}=-1\}}\right] = 1 - q_{n-1}$ and if $y_{n-1} = -1$ then $\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\left[\mathbb{1}_{\{\hat{y}_{n-1}\neq y_{n-1}\}}\right] = q_{n-1}$ and hence for the choice of $q_{n-1} = \frac{1}{2} + \frac{n}{2}\,\mathbb{E}_{\epsilon_n}\left[\phi(y_1,\ldots,y_{n-2},-1,\epsilon_n) - \phi(y_1,\ldots,y_{n-2},+1,\epsilon_n)\right]$, we have

$$\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\left[\mathbb{1}_{\{\hat{y}_{n-1}\neq y_{n-1}\}}\right] = \tfrac{1}{2} - \tfrac{y_{n-1}}{2}\left(\mathbb{E}_{\epsilon_n}\phi(y_1,\ldots,y_{n-2},-1,\epsilon_n) - \mathbb{E}_{\epsilon_n}\phi(y_1,\ldots,y_{n-2},+1,\epsilon_n)\right)$$

Thus we can conclude that,

$$
\begin{aligned}
&\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\left[\,\mathbb{1}_{\{\hat{y}_{n-1}\neq y_{n-1}\}}\right]+\tfrac{1}{n}\mathbb{E}_{\hat{y}_n\sim q_n}\left[\,\mathbb{1}_{\{\hat{y}_n\neq y_n\}}\right]-\phi(y_1,\dots,y_n)\\
&=\frac{1}{2}+\tfrac{1}{n}\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\left[\,\mathbb{1}_{\{\hat{y}_{n-1}\neq y_{n-1}\}}\right]-\mathbb{E}_{\epsilon_n}\phi(y_1,\dots,y_{n-1},\epsilon_n)\qquad\text{(From Eq.2)}\\
&=1-\tfrac{y_{n-1}}{2}\left(\mathbb{E}_{\epsilon_n}\phi(y_1,\dots,y_{n-2},-1,\epsilon_n)-\mathbb{E}_{\epsilon_n}\phi(y_1,\dots,y_{n-2},+1,\epsilon_n)\right)-\mathbb{E}_{\epsilon_n}\phi(y_1,\dots,y_{n-1},\epsilon_n)\\
&=1-\frac{1}{2}\left(\mathbb{E}_{\epsilon_n}\phi(y_1,\dots,y_{n-2},+1,\epsilon_n)+\mathbb{E}_{\epsilon_n}\phi(y_1,\dots,y_{n-2},-1,\epsilon_n)\right)\\
&=1-\mathbb{E}_{\epsilon_{n-1},\epsilon_n}\phi(y_1,\dots,y_{n-2},\epsilon_{n-1},\epsilon_n)
\end{aligned}
$$

Proceeding in similar way we conclude that,

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

Hence, if $\mathbb{E}_{\epsilon_1,\dots,\epsilon_n}\phi(\epsilon_1,\dots,\epsilon_n)\ge n/2$ then we can conclude that, $\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\left[\,\mathbb{1}_{\{\hat{y}_t\neq y_t\}}\right]\le\phi(y_1,\dots,y_n)$ as desired. Thus we conclude the proof of this lemma. □

Cover's result is remarkable because it says that if a bound on error of some stable ϕ is possible under random coin flips, then such bound is possible under any adversary!

So using this result, we will as promised prove our claim, but more generally we can ask what are examples of stable ϕ's and how far can we take this machinery. While we set our initial goal low, now its time to consider more general case of regret against a class of prefixed models $\mathcal{F}$. Instead of majority (which is either all $+1$ or all -1 in hindsight), consider an arbitrary set $\mathcal{F}\subset\{\pm1\}^n$. We would like to come up with a strategy that minimizes regret with respect to any choice in $\mathcal{F}$ as:

$$
\mathrm{Reg}_n(\mathcal{F}):=\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\left[\,\mathbb{1}_{\{\hat{y}_t\neq y_t\}}\right]-\min_{f\in\mathcal{F}}\sum_{t=1}^{n}\,\mathbb{1}_{\{f_t\neq y_t\}}
$$

Our goal specifically is to answer, what is $C_n(\mathcal{F})>0$ such that, there exists a prediction algorithm such that against any adversary this algorithm can guarantee that:

$$
\mathrm{Reg}_n(\mathcal{F})\le C_n(\mathcal{F})
$$

To this end we can use Cover's lemma above with

$$
\phi(y_1,\dots,y_n)=\min_{f\in\mathcal{F}}\sum_{t=1}^{n}\,\mathbb{1}_{\{f_t\neq y_t\}}+C_n(\mathcal{F})
$$

We will soon show that this function is stable but for now if you trust me that it is stable, note that cover's result tells us the smallest value of $C_n(\mathcal{F})$. It is given by

$$\begin{aligned}
C_n(\mathcal{F}) &= \frac{n}{2} - \mathbb{E}_\epsilon \left[\min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq \epsilon_t\}} \right] \\
&= \frac{n}{2} + \mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} - \mathbb{1}_{\{f_t \neq \epsilon_t\}} \right] \\
&= \frac{n}{2} + \mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} -\frac{1}{2}(1 - f_t \epsilon_t) \right] \\
&= \frac{n}{2} + \frac{1}{2}\mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t - 1 \right] \\
&= \frac{1}{2}\mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t \right]
\end{aligned}$$

Claim 3. *For any $\mathcal{F} \subseteq \{\pm 1\}^n$, the function*

$$\phi(y_1, \ldots, y_n) = \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq y_t\}} + C_n(\mathcal{F})$$

is stable.

Proof. W.l.o.g. pick the last bit to flip, we find,

$$\begin{aligned}
&\phi(y_1, \ldots, y_{n-1}, +1) - \phi(y_1, \ldots, y_{n-1}, -1) \\
&\quad = \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq +1\}} \right\} - \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq -1\}} \right\} \\
&\quad = \max_{f' \in \mathcal{F}} \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq +1\}} - \sum_{t=1}^{n-1} \mathbb{1}_{\{f'_t \neq y_t\}} - \mathbb{1}_{\{f'_n \neq -1\}} \right\} \\
&\quad \leq \max_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq +1\}} - \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} - \mathbb{1}_{\{f_n \neq -1\}} \right\} \\
&\quad = \max_{f \in \mathcal{F}} \left\{ \mathbb{1}_{\{f_n \neq +1\}} - \mathbb{1}_{\{f_n \neq -1\}} \right\} \leq 1
\end{aligned}$$

Similarly we have, $\phi(y_1, \ldots, y_{n-1}, -1) - \phi(y_1, \ldots, y_{n-1}, +1) \leq 1$ and so we have stability. □

Putting this together we have:

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

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{\hat{y}_t \neq y_t\} \right] - \min_{b \in \{\pm 1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\} \leq \frac{1}{2}\mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f_t \right]$$

against any adversary and any $\mathcal{F} \subset \{\pm 1\}^n$