6.7800 Spring 2025: Bonus Lecture 01

Minimax Optimal Bit Prediction

Jongha (Jon) Ryu
jongha@mit.edu

November 13, 2025

Contents

1 Problem Setting

In the lecture, we learned how to design the minimax optimal hypothesis testing rule, and learned the concept of "equalizer". In this note, we will think about these concepts from the sequential *bit prediction* problem, which dates back to [3].

The problem setting is as follows.

- A player is asked to predict a coin flip. After the prediction, the coin flip is revealed.
- This prediction game repeats for n times; hence, the player at round t can use the history of the previous $t-1$ coin flips.

- A player can use a randomized rule.

The goal of the player is to minimize the total number of wrong predictions. To proceed, let's mathematically formalize the problem. Let $y_t \in \{+1, -1\}$ be the coin flip at time t. The player's strategy $\mathbf{a}$ at time t can be represented as a single number $a_t \colon \mathcal{Y}^{t-1} \to [-1, 1]$, and the player's prediction $\hat{y}_t$ is drawn from $\mathsf{Rad}(a_t(y^{t-1}))$. The loss at round t is $\mathbb{1}\{\hat{y}_t \neq y_t\}$, and thus the cumulative loss after round n is

$$\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}.$$

This loss $\ell(\hat{y}, y) := \mathbb{1}\{\hat{y} \neq y\}$ is often called 0-1 loss or Hamming loss. Since the cumulative loss is random (where the randomness comes from the player's randomized strategy), we are interested in the *expected* cumulative loss:

$$\begin{aligned}
\overline{L}(\mathbf{a}, y^n) &:= \mathbb{E}\Big[\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}\Big] \\
&= \sum_{t=1}^{n} \Pr\{\hat{y}_t \neq y_t\} \\
&= \frac{1}{2} \sum_{t=1}^{n} (1 - a_t(y^{t-1}) y_t) \\
&= \frac{n}{2} - \frac{1}{2} \sum_{t=1}^{n} a_t(y^{t-1}) y_t.
\end{aligned}$$

In what follows, we will consider this as the performance measure of a strategy $\mathbf{a}$.

Exercise 1. *Show that* $\bar{\ell}(a, y) := \mathbb{E}_{\hat{y} \sim \mathsf{Rad}(a)}[\mathbb{1}\{\hat{y} \neq y\}] = \frac{1}{2}(1 - ay)$.

The first formal question we can ask is whether there exists a strategy that performs well for *all possible* coin flips. The answer is no, since we can show that *any* causal predictor will err on half of the predictions, on average over all possible sequences:

Exercise 2. *Prove that*

$$\frac{1}{2^n} \sum_{y^n \in \{+1,-1\}^n} \overline{L}(\mathbf{a}, y^n) = \frac{n}{2}.$$

Hence, we cannot expect to come up with a strategy that performs uniformly well over all sequences. Given this, we instead consider a set of *reference* strategies $\mathcal{F}$ to compete against, and aim to *track* the best performance in hindsight! Then, we wish to minimize the performance gap between our strategy $\mathbf{a}$ and any reference strategy $\mathbf{f}$ from $\mathcal{F}$, which we call the regret:

$$\mathsf{Reg}_{\mathcal{F}}(\mathbf{a}, y^n) := \overline{L}(\mathbf{a}, y^n) - \min_{\mathbf{f} \in \mathcal{F}} \overline{L}(\mathbf{f}, y^n).$$

We wish to minimize the worst-case regret:

$$V_{\mathcal{F}}^{(n)} := \min_{\mathbf{a}} \max_{y^n \in \{+1,-1\}^n} \mathsf{Reg}_{\mathcal{F}}(\mathbf{a}, y^n). \tag{1}$$

Here, the minimum over $\mathbf{a}$ is for all possible causal strategies. We call $V_{\mathcal{F}}^{(n)}$ the minimax value of the prediction game for length n.

The most simplest choices of such competitors are the constant predictors $-\mathbf{1}$ and $+\mathbf{1}$, which always output -1 and $+1$, respectively; formally, we consider $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$. In this case, the best performance of $\mathcal{F}$ can be written as follows:

Exercise 3. *For $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$, show that*

$$\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^n) = \frac{n}{2} + \frac{1}{2}\min\Big\{\sum_{t=1}^n y_t, -\sum_{t=1}^n y_t\Big\} = \frac{n}{2} - \frac{1}{2}\Big|\sum_{t=1}^n y_t\Big|. \tag{2}$$

This tells us that unless the underlying sequence has equal number of $+1$'s and -1's, the best performance is better than a random guess, and it is thus a reasonable predictor class to compete against.

Now, the question is: what is the minimax optimal strategy that exactly attains $V_{\mathcal{F}}^{(n)}$? It is a folklore that T. Cover proposed the following elegant, randomized algorithm.

Input: horizon $n \in \mathbb{N}$.

For each $t = 1, \dots, n$:

- Draw $n - t$ i.i.d. Rademacher random variables $\varepsilon_{t+1}, \dots, \varepsilon_n \sim \mathsf{Rad}(0)$ and set

$$\hat{y}_t^\star(y^{t-1}) := \mathsf{majority}\{y^{t-1}, \varepsilon_{t+1}^n\}. \tag{3}$$

- In case of tie for odd n, choose one from $\{\pm 1\}$ at random.

Theorem 4. *The randomized forecaster $\mathbf{a}^\star$ is minimax optimal.*

In the rest of this note, we will show why this strategy is minimax optimal.

2 Equalizer is Minimax Optimal

One of the most fundamental themes in game theory is to find a minimax optimal predictor given a sequential game. In general, as one may expect, constructing an optimal predictor in an explicit form is nontrivial. Moreover, it is even unclear how to check if a given predictor is indeed minimax optimal. One naive approach is to analyze and provide an upper bound on the regret of the predictor and show that it attains the minimax value of the game, which also needs to be computed. Under a mild condition on the loss function, however, there exists a sufficient condition on a predictor which guarantees its minimax optimality.

Definition 5 (Equalizer)**.** A predictor $\mathbf{a}$ is said to be an *equalizer* with respect to $\mathcal{F}$ if $\mathsf{Reg}_{\mathcal{F}}(\mathbf{a}, y^n)$ is independent of y^n.

Definition 6 (Balanced loss)**.** A loss function ℓ is said to be *balanced* if any predictors $\mathbf{a}$ performs exactly same on average, that is,

$$\sum_{y^n\in\mathcal{Y}^n} \overline{L}(\mathbf{a}, y^n)$$

is not a function of $\mathbf{a}$, but only a function of n.

Proposition 7. *If ℓ is balanced, then an equalizer $\mathbf{a}^*$, if exists, is minimax optimal.*

Proof. Suppose that an equalizer strategy $\mathbf{a}^*$ exists. Then, for any predictor $\mathbf{a}$ and for any $y^n \in \mathcal{Y}^n$, we have

$$\begin{aligned}\mathsf{Reg}_{\mathcal{F}}(\mathbf{a}^*, y^n) &\overset{(a)}{=} \frac{1}{|\mathcal{Y}|^n} \sum_{\tilde{y}^n \in \mathcal{Y}^n} \mathsf{Reg}_{\mathcal{F}}(\mathbf{a}^*, \tilde{y}^n) \\ &\overset{(b)}{=} \frac{1}{|\mathcal{Y}|^n} \sum_{\tilde{y}^n \in \mathcal{Y}^n} \mathsf{Reg}_{\mathcal{F}}(\mathbf{a}, \tilde{y}^n) \\ &\le \max_{\tilde{y}^n \in \mathcal{Y}^n} \mathsf{Reg}_{\mathcal{F}}(\mathbf{a}, \tilde{y}^n).\end{aligned}$$

Note that (a) follows from the definition of equalizer, and (b) follows from the balancedness of the loss function. By taking maximum over y^n and minimum over $\mathbf{a}$, we have

$$\max_{y^n} \mathsf{Reg}_{\mathcal{F}}(\mathbf{a}^*, y^n) \le \min_{\mathbf{a}} \max_{\tilde{y}^n \in \mathcal{Y}^n} \mathsf{Reg}_{\mathcal{F}}(\mathbf{a}, \tilde{y}^n),$$

which implies that $\mathbf{a}^*$ is minimax optimal. □

Since the Hamming loss is balanced as shown in Exercise 2, it suffices to show that the strategy is equalizer.

3 Minimax Optimal Predictor via Equalizer

3.1 Recursive Formulation of Minimax Optimal Predictor

In this section, we show how a minimax optimal strategy can be computed in a recursive manner from the best achievable loss by the given expert class.

The form of minimax regret (1) does not reflect the nature of the sequential game. In order to make it explicit, we can rewrite the minimax value of the game as

$$\begin{aligned}V_{\mathcal{F}}^{(n)} &= \min_{a_1} \max_{y_1 \in \mathcal{Y}} \mathbb{E}_{\hat{y}_1 \sim a_1} \dots \min_{a_n} \max_{y_n \in \mathcal{Y}} \mathbb{E}_{\hat{y}_n \sim a_n} \left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \min_{\mathbf{f} \in \mathcal{F}} \overline{L}(\mathbf{f}, y^n) \right] \\ &= \min_{a_1} \max_{y_1 \in \mathcal{Y}} \dots \min_{a_n} \max_{y_n \in \mathcal{Y}} \left[\sum_{t=1}^{n} \bar{\ell}(a_t, y_t) - \min_{\mathbf{f} \in \mathcal{F}} \overline{L}(\mathbf{f}, y^n) \right]. \end{aligned} \tag{4}$$

At a glimpse, this expression might seem to encode a harder game, but one can show the equivalence of the two expressions (1) and (4).[1]

For each history y^{t-1} ($2 \le t \le n$), we define the conditional value of the game as

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) := \min_{a_t} \max_{y_t \in \mathcal{Y}} \dots \min_{a_n} \max_{y_n \in \mathcal{Y}} \left[\sum_{s=t}^{n} \bar{\ell}(a_s, y_s) - \min_{\mathbf{f} \in \mathcal{F}} \overline{L}(\mathbf{f}, y^n) \right].$$

[1] Cesa-Bianchi and Lugosi [1, Sec. 8.2] has an equivalent derivation for binary prediction under Hamming loss.

Here, we define $V_{\mathcal{F}}^{(n)}(\varnothing) := V_{\mathcal{F}}^{(n)}$ and

$$V_{\mathcal{F}}^{(n)}(y^n) := -\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^n). \tag{5}$$

We have the following recursive relation: for $s = 1, \dots, n$, we have

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) = \min_{a_t} \max_{y_t\in\mathcal{Y}} \left\{ \bar{\ell}(a_t, y_t) + V_{\mathcal{F}}^{(n)}(y^t) \right\}.$$

Hence, we can write, for any $0 \le t \le n$,

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) := \min_{a_1} \max_{y_1\in\mathcal{Y}} \dots \min_{a_t} \max_{y_t\in\mathcal{Y}} \left[\sum_{s=1}^{t} \bar{\ell}(a_s, y_s) + V_{\mathcal{F}}^{(n)}(y^t) \right].$$

Given a history y^{t-1}, the minimax optimal action is the minimum achieving action of $V_{\mathcal{F}}^{(n)}(y^{t-1})$, i.e.,

$$a_t^\star(\cdot|y^{t-1}) := \arg\min_{a_t} \max_{y_t\in\mathcal{Y}} \left\{ \bar{\ell}(a_t, y_t) + V_{\mathcal{F}}^{(n)}(y^t) \right\}. \tag{13}$$

For the special case of binary prediction under Hamming loss, the value and the optimal rule can be somewhat simplified.

$$\begin{aligned}
V_{\mathcal{F}}^{(n)}(y^{t-1}) &= \min_{a_t} \max_{y_t\in\{\pm1\}} \left\{ \frac{1}{2}(1 - a_t y_t) + V_{\mathcal{F}}^{(n)}(y^t) \right\} \\
&= \min_{a_t} \max \left\{ \frac{1}{2}(1 - a_t) + V_{\mathcal{F}}^{(n)}(y^{t-1}1), \frac{1}{2}(1 + a_t) + V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}) \right\} \\
&= \frac{1}{2} + \frac{1}{2}(V_{\mathcal{F}}^{(n)}(y^{t-1}1) + V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})) + \frac{1}{2} \min_{a_t} |a_t + V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}) - V_{\mathcal{F}}^{(n)}(y^{t-1}1)| \\
&= \begin{cases} V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}) & \text{if } V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}) < -1, \\ \mathbb{E}_{\varepsilon_t}[V_{\mathcal{F}}^{(n)}(y^{t-1}\varepsilon_t)] + \frac{1}{2} & \text{if } |V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})| \le 1, \\ V_{\mathcal{F}}^{(n)}(y^{t-1}1) & \text{if } V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}) > 1, \end{cases}
\end{aligned} \tag{6}$$

and the optimal action is

$$a_t^\star(y^{t-1}) = \tau(V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})),$$

where the function $\tau : \mathbb{R} \to [-1, 1]$ is a thresholding function defined as

$$\tau(y) = \begin{cases} -1 & \text{if } y < -1, \\ y & \text{if } |y| \le 1, \\ 1 & \text{if } y > 1. \end{cases} \tag{7}$$

Note that $V_{\mathcal{F}}^{(n)}(y^t)$ acts as a "score" function. In the ambiguous case, where $V_{\mathcal{F}}^{(n)}(y^{t-1}1)$ and $V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})$ are close enough, we can play a randomized action to make the value of the game equal, being an equalizer. Otherwise, for example, if $V_{\mathcal{F}}^{(n)}(y^{t-1}1)$ is substantially larger than $V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})$, then we can guess that $y_t = 1$ is the choice the adversary will make, and thus the optimal strategy is to set $a_t = 1$, so that $\hat{y}_t = 1$ deterministically.

In the next two sections, we will show that $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$ admits an equalizer. We provide a general treatment in Section 3.2* and consider the special case of binary prediction in Section 3.3*. However, an inexperienced reader can skip these two sections and directly go to Section 3.4, regarding that $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$ admits an equalizer is a given fact.

3.2* When Does Equalizer Exist?

We are now ready to characterize a condition on the value of the game with $\mathcal{F}$ for an equalizer to exist with a general class of experts.

Recall the recursive formulation of the minimax optimal forecaster $a^\star$ for the time horizon n is: for $t = n, n-1, \dots, 1$,

$$a_t^\star(\cdot|y^{t-1}) := \arg\min_{a_t \in \Delta(\mathcal{Y})} \max_{y_t \in \mathcal{Y}} \{\bar{\ell}(a_t, y_t) + V_{\mathcal{F}}^{(n)}(y^t)\}.$$

That is, $a_t^\star(\cdot|y^{t-1})$ is an optimal strategy for the single-round game

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) = \min_{a_t \in \Delta(\mathcal{Y})} \max_{y_t \in \mathcal{Y}} \{\bar{\ell}(a_t, y_t) + V_{\mathcal{F}}^{(n)}(y^t)\}. \tag{8}$$

Theorem 8. *Assume that an equalizer is guaranteed to be minimax optimal. Then, a strategy $a^\star = (a_t^\star(\cdot|y^{t-1}))_{t=1}^n$ is an equalizer if and only if $a_t^\star(\cdot|y^{t-1})$ is an equalizer for every single-round of the game (8) for any history y^{t-1}, that is,*

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) = \bar{\ell}(a_t^\star(\cdot|y^{t-1}), y_t) + V_{\mathcal{F}}^{(n)}(y^t)$$

for any y^t, for all $t = n, n-1, \dots, 1$.

Proof. Assume that $a^\star$ is a minimax optimal equalizer. Observe that for each round of the game (8), $a_t^\star(\cdot|y^{t-1})$ must be the minimax optimal solution, and thus

$$\begin{aligned} V_{\mathcal{F}}^{(n)}(y^{t-1}) &= \max_{y_t \in \mathcal{Y}} \{\bar{\ell}(a_t^\star(\cdot|y^{t-1}), y_t) + V_{\mathcal{F}}^{(n)}(y^t)\} \\ &\geq \bar{\ell}(a_t^\star(\cdot|y^{t-1}), y_t) + V_{\mathcal{F}}^{(n)}(y^t) \end{aligned}$$

for any y_t and y^{t-1}. Hence, by definition, for any y^n, we have

$$\begin{aligned} \mathsf{Reg}_{\mathcal{F}}(a^\star, y^n) &= L_{a^\star}(y^n) - \min_{\mathbf{f} \in \mathcal{F}} \overline{L}(\mathbf{f}, y^n) \\ &= \sum_{t=1}^{n} \bar{\ell}(a_t^\star(\cdot|y^{t-1}), y_t) + V_{\mathcal{F}}^{(n)}(y^n) \\ &= \sum_{t=1}^{n-1} \bar{\ell}(a_t^\star(\cdot|y^{t-1}), y_t) + \underbrace{\bar{\ell}(a_n^\star(\cdot|y^{n-1}), y_n) + V_{\mathcal{F}}^{(n)}(y^n)}_{\leq V_{\mathcal{F}}^{(n)}(y^{n-1})} \\ &\leq \sum_{t=1}^{n-2} \bar{\ell}(a_t^\star(\cdot|y^{t-1}), y_t) + \underbrace{\bar{\ell}(a_{n-1}^\star(\cdot|y^{n-2}), y_{n-1}) + V_{\mathcal{F}}^{(n)}(y^{n-1})}_{\leq V_{\mathcal{F}}^{(n)}(y^{n-2})} \end{aligned}$$

$$
\begin{aligned}
&\le \dots \\
&\le \bar{\ell}(a_1^\star(\cdot|\varnothing), y_1) + V_{\mathcal{F}}^{(n)}(y_1) \\
&\le V_{\mathcal{F}}^{(n)}(\varnothing).
\end{aligned}
$$

Note, however, that $a^\star$ is an minimax optimal equalizer implies that $\mathsf{Reg}_{\mathcal{F}}(a^\star; y^n) = V_{\mathcal{F}}^{(n)}(\varnothing)$. Hence, all inequalities must hold with equality, which concludes that $a_t^\star(\cdot|y^{t-1})$ is an equalizer for every t and y^{t-1}.

Conversely, if we assume that $a_t^\star(\cdot|y^{t-1})$ is a minimax optimal equalizer for the single-round game (8), then the same reasoning leads to $\mathsf{Reg}_{\mathcal{F}}(a^\star; y^n) = V_{\mathcal{F}}^{(n)}(\varnothing)$ for any y^n, which implies that $a^\star$ is a minimax optimal equalizer. □

Hence, an expert class $\mathcal{F}$ admits an equalizer if and only if the sequential game admits an equalizer in every single round for all possible observations y^n.

3.3* Special Case: Binary Prediction Under Hamming Loss

Recall that a single-round equalizer for binary prediction under Hamming loss exists if and only if $|V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})| \le 1$. Therefore,

Corollary 9. *For binary prediction under Hamming loss, an expert class $\mathcal{F}$ admits an equalizer if and only if*

$$
|V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})| \le 1 \tag{9}
$$

for all y^{t-1} for any $t = n, n-1, \dots, 1$.

Equipped with the equivalent condition (9), it is now relatively easy to check if a given expert class admits an equalizer or not. The following is a sufficient condition for the competitor class $\mathcal{F}$ to admit an equalizer.

Corollary 10. *If for any $\mathbf{f} \in \mathcal{F}$ satisfies*

$$
\left|\overline{L}(\mathbf{f}, y^n) - \overline{L}(\mathbf{f}, \tilde{y}^n)\right| \le 1 \tag{10}
$$

for any y^n and $\tilde{y}^n$ within Hamming distance 1, i.e., $d_H(y^n, \tilde{y}^n) \le 1$, then $\mathcal{F}$ admits an equalizer.

Proof. Let $y^n = y^{t-1}1y_{t+1}^n$ and $\tilde{y}^n = y^{t-1}\bar{1}y_{t+1}^n$, and assume that

$$
\begin{aligned}
V_{\mathcal{F}}^{(n)}(y^n) &= -\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^n) = -\overline{L}(\mathbf{f}, y^n), \\
V_{\mathcal{F}}^{(n)}(\tilde{y}^n) &= -\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, \tilde{y}^n) = -L_{\tilde{f}}(\tilde{y}^n).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
-1 &\le \overline{L}(\mathbf{f}, y^n) - \overline{L}(\mathbf{f}, \tilde{y}^n) \\
&\le \overline{L}(\mathbf{f}, y^n) - L_{\tilde{f}}(\tilde{y}^n) \\
&\le L_{\tilde{f}}(y^n) - L_{\tilde{f}}(\tilde{y}^n) \le 1,
\end{aligned}
$$

which implies that $|V_{\mathcal{F}}^{(n)}(y^n) - V_{\mathcal{F}}^{(n)}(\tilde{y}^n)| \le 1$.

Now we use backward induction: assume that $|V_{\mathcal{F}}^{(n)}(y^{t'-1}1) - V_{\mathcal{F}}^{(n)}(y^{t'-1}\bar{1})| \le 1$ for $t' = t+1, \dots, n$ for $t \le n-1$. Then, by the recursive formula (6), we have $V_{\mathcal{F}}^{(n)}(y^t) = \mathbb{E}_{\varepsilon_{t+1}}[V_{\mathcal{F}}^{(n)}(y^t\varepsilon_{t+1})] + \frac{1}{2} = \mathbb{E}_{\varepsilon_{t+1}^n}[V_{\mathcal{F}}^{(n)}(y^t\varepsilon_{t+1}^n)] + \frac{n-t}{2}$. Hence,

$$\begin{aligned}|V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1})| &= \left|\mathbb{E}_{\varepsilon_{t+1}^n}[V_{\mathcal{F}}^{(n)}(y^{t-1}1\varepsilon_{t+1}^n) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}\varepsilon_{t+1}^n)]\right| \\ &\le \mathbb{E}_{\varepsilon_{t+1}^n}[|V_{\mathcal{F}}^{(n)}(y^{t-1}1\varepsilon_{t+1}^n) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}\varepsilon_{t+1}^n)|] \le 1,\end{aligned}$$

which completes the induction. □

Fortunately, a class of *constant* (or static) predictors such as $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$ admit an equalizer.

Corollary 11 (Cesa-Bianchi and Lugosi [1, Lemma 8.1])**.** *A class of static experts satisfies* (10)*, so an equalizer always exists.*

Proof. Note that $|d_H(a^n, y^n) - d_H(a^n, \tilde{y}^n)| \le d_H(y^n, \tilde{y}^n) = 1$ for any a^n. □

We remark, however, that not all classes of competitors admit an equalizer, as shown by the following example.

Example 12. *Let $n = 2$, and consider a class $\mathcal{F}$ with the following two experts:*

$$\begin{aligned}f(y_1y_2) &= f_1(\cdot)f_2(y_1) = \bar{1}\bar{y}_1, \\ g(y_1y_2) &= g_1(\cdot)g_2(y_1) = \bar{1}\bar{1}.\end{aligned}$$

y_1y_2	$f(y_1y_2)$	$g(y_1y_2)$	$\overline{L}(\mathbf{f}, y_1y_2)$	$L_g(y_1y_2)$	$V_2(\mathcal{F}\|y_1y_2)$
$\bar{1}\bar{1}$	$\bar{1}1$	$\bar{1}\bar{1}$	1	$\bar{1}$	0
$\bar{1}1$	$\bar{1}1$	$\bar{1}\bar{1}$	$\bar{1}$	1	0
$1\bar{1}$	$\bar{1}\bar{1}$	$\bar{1}\bar{1}$	1	1	-1
11	$\bar{1}\bar{1}$	$\bar{1}\bar{1}$	2	2	-2

Here, since $|V_2(\mathcal{F}|\bar{1}1) - V_2(\mathcal{F}|11)| = 2 > 1$, this class violates the stability condition. That is, in this case, some competitor is strictly better than the other and we the optimal prediction does not need to make a randomized action.

3.4 When Equalizer Exists

Now we assume that an expert class $\mathcal{F}$ admits an equalizer and draw some consequences. One can observe that for any $t = n, n-1, \dots, 1$ and for all y^{t-1}, the conditional value can be written as

$$\begin{aligned}V_{\mathcal{F}}^{(n)}(y^{t-1}) &= \mathbb{E}[V_{\mathcal{F}}^{(n)}(y^{t-1}\varepsilon_t)] + \frac{1}{2} \\ &= \dots \\ &= \mathbb{E}[V_{\mathcal{F}}^{(n)}(y^{t-1}\varepsilon_t^n)] + \frac{n-t+1}{2}\end{aligned}$$

$$= -\mathbb{E}\Big[\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^{t-1}\varepsilon_t^n)\Big] + \frac{n-t+1}{2}, \tag{11}$$

and the optimal equalizer is thus

$$\begin{aligned} a_t^\star(y^{t-1}) &= V_{\mathcal{F}}^{(n)}(y^{t-1}1) - V_{\mathcal{F}}^{(n)}(y^{t-1}\bar{1}) \\ &= \mathbb{E}\Big[\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^{t-1}\bar{1}\varepsilon_{t+1}^n) - \min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^{t-1}1\varepsilon_{t+1}^n)\Big]. \end{aligned} \tag{12}$$

Cesa-Bianchi and Lugosi [1, Exercise 8.4] and Cesa-Bianchi and Shamir [2] established this expression for the class of static experts.

Finally, the value of the game $V_{\mathcal{F}}^{(n)}$ can be written as

$$\begin{aligned} V_{\mathcal{F}}^{(n)} &= \frac{n}{2} - \mathbb{E}\Big[\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, \varepsilon^n)\Big] \\ &= \frac{n}{2} - \mathbb{E}\Big[\min_{\mathbf{f}\in\mathcal{F}} \sum_{t=1}^n \frac{1 - f_t(\varepsilon^{t-1})\varepsilon_t}{2}\Big] \\ &= \frac{1}{2}\mathbb{E}\Big[\max_{\mathbf{f}\in\mathcal{F}} \sum_{t=1}^n f_t(\varepsilon^{t-1})\varepsilon_t\Big]. \end{aligned}$$

We note that [1, Chap. 8] is devoted to studying $V_{\mathcal{F}}^{(n)}$ based on this expression.

4 Derivation of Cover's Strategy

Given the implication of the existence of equalizer in Section 3.4 and that $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$ admits an equalizer, we are now ready to derive Cover's strategy (3) as the equalizer strategy for $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$. Let $R_t := y_1 + \ldots + y_{t-1} + \varepsilon_{t+1} + \ldots + \varepsilon_n$. Then, for $y \in \{\pm 1\}$,

$$\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^{t-1}y\varepsilon_{t+1}^n) = \frac{n}{2} - \frac{1}{2}|R_t + y|,$$

and thus by combining (2) and (12), we have

$$\begin{aligned} a_t^\star(y^{t-1}) &= \frac{1}{2}(\mathbb{E}[|R_t + 1| - |R_t - 1|]) \\ &= \mathsf{P}(R_t \geq 1) - \mathsf{P}(R_t \leq -1) \\ &= 2\,\mathsf{P}(R_t \geq 1) + \mathsf{P}(R_t = 0) - 1. \end{aligned}$$

It is easy to see that $\{R_t \geq 1\} = \{\mathsf{majority}\{y_1, \ldots, y_{t-1}, \varepsilon_{t+1}, \ldots, \varepsilon_n\} = +1\}$, which confirms the equivalence.

For the special case of $\mathcal{F} = \{+\mathbf{1}, -\mathbf{1}\}$, the value of the game simplifies to

$$V_{\mathcal{F}}^{(n)} = \frac{1}{2}\mathbb{E}\Big[\max\Big\{\sum_{t=1}^n \varepsilon_t, -\sum_{t=1}^n \varepsilon_t\Big\}\Big] = \frac{1}{2}\mathbb{E}\Big[\Big|\sum_{t=1}^n \varepsilon_t\Big|\Big],$$

which is the expected value of the deviation of random walks from the origin. Hence, $V_{\mathcal{F}}^{(n)} \sim \Theta(\sqrt{n})$ (Exercise: find the leading constant). Note the well-known regret bound $\frac{\sqrt{n}}{2\sqrt{2}} \leq V_{\mathcal{F}}^{(n)} \leq \frac{\sqrt{n}}{2}$ [1,

Sec. 8.4]. This implies that if the cumulative regret is minimized by the length of horizon n, it diminishes as n goes to infinity, implying that Cover's strategy performs as well as the best predictor in $\mathcal{F}$, with diminishing regret per step.

5 Concluding Remarks

In general, if a sequential predictor has diminishing normalized regret with respect to a class of competitors, then we say that the predictor is *universal* with respect to the class. The notion of universality will be also covered in future lectures. Cover's strategy is not only universal, but exactly minimax optimal, in the sense that it achieves the minimum possible worst-case regret. Since it is hard to obtain an exactly minimax optimal strategy in general, especially for an indefinite horizon, researchers often aim to find a *nearly* minimax optimal strategy.

This bit prediction problem was studied in depth in a seminal paper [4] under the terminology of *universal prediction*. Unlike the finite, known horizon setting in this note, Feder et al. [4] proposed an efficient *universal* predictor with unknown horizon for $\mathcal{F} = \{-\mathbf{1}, +\mathbf{1}\}$,

$$a_t^{\mathsf{FMG}}(y^{t-1}) = \tau\Big(\frac{1}{\varepsilon_t}\frac{y_1 + \ldots + y_{t-1}}{t-1}\Big) \in [-1, 1],$$

for some sequence of vanishing, nonnegative sequence $(\varepsilon_t)_{t=1}^{\infty}$. (Recall the definition of the thresholding function τ in (7).) They showed that if $\varepsilon_t = 1/\sqrt{t+2}$, then

$$\mathsf{Reg}_{\mathcal{F}}(\mathbf{a}^{\mathsf{FMG}}, y^n) \le \sqrt{n+1} + \frac{1}{2}.$$

The philosophy of the FMG predictor is randomizing only if there are no clear single winner (i.e., arg max) in the empirical counts. It is quite remarkable that this simple randomized strategy is universal for an unknown horizon, while the underlying idea is quite similar to the exact minimax optimal strategy of Cover.

Rakhlin and Sridharan [5] wrote an excellent tutorial on the same bit prediction problem, but from a slightly different perspective of *achievable scores*, which was originally studied by Cover [3]. We refer an interested reader to this survey, for the alternative aspect and its application to node classification in social network. With a different emphasis, the current lecture focuses on deriving the minimax optimal prediction when we know that an equalizer exists, via the recursive characterization of a minimax optimal predictor. One can directly attempt to show that Cover's strategy is an equalizer, but it would require a rather complicated argument. (Exercise: prove that Cover's strategy is an equalizer without using the recursive characterization.)

A natural extension is to consider the m-ary prediction under Hamming loss for $m \ge 2$. We discuss the derivation of the minimax optimal rule in Appendix A.

A Multiary Extension

Let m denote the alphabet size $|\mathcal{Y}|$. The main question of this section is if a natural extension of the Cover's random-coin strategy remains optimal for $m > 2$ for static constant experts in general.

Recall the minimax forecaster can be recursively computed as

$$a_t^\star(\cdot|y^{t-1}) = \arg\min_{a_t\in\Delta(\mathcal{Y})}\max_{y_t\in\mathcal{Y}}\Big\{\ell(a_t, y_t) + V_{\mathcal{F}}^{(n)}(y^t)\Big\}$$
$$= \arg\min_{a_t\in\Delta(\mathcal{Y})}\max_{y_t\in\mathcal{Y}}\Big\{-a_t(y_t) + V_{\mathcal{F}}^{(n)}(y^t)\Big\}. \tag{13}$$

This expression can be further simplified by the following proposition:

Proposition 13. *For a given vector $\mathbf{v} = (v_1, \ldots, v_m) \in \mathbb{R}^m$, the minimax solution of the following game*

$$\min_{a\in\Delta([m])}\max_{y\in[m]}(v_y - a(y))$$

is given by the reverse water-filling solution $a_{\texttt{water}} \in \Delta([m])$ which is defined as

$$a_{\texttt{water}}(y|\mathbf{v}) = (v_y - \eta(\mathbf{v}))^+,$$

where $(u)^+$ denotes the positive part of u, i.e.,

$$(u)^+ := \begin{cases} u & \text{if } u \geq 0, \\ 0 & \text{if } u < 0, \end{cases}$$

and $\eta(\mathbf{v}) \in \mathbb{R}$ is chosen so that

$$\sum_{y\in[m]} (v_y - \eta(\mathbf{v}))^+ = 1.$$

In particular, the minimax value of the game is $\eta(\mathbf{v})$. Moreover, the reverse water-filling solution can be explicitly expressed as follows. Let $v_{[1]} \geq \ldots \geq v_{[m]}$ be the ordering of the coordinates of $\mathbf{v}$. Define

$$k^\star := \max\Big\{k \in [m]\colon \sum_{i=1}^{k}(v_{[i]} - v_{[k]}) < 1\Big\}.$$

Then,

$$\eta(\mathbf{v}) = \frac{1}{k^\star}(v_{[1]} + \ldots + v_{[k^\star]} - 1),$$

and the optimal water distribution is

$$a_{\texttt{water}}(y|\mathbf{v}) = \begin{cases} v_y - \eta(\mathbf{v}) & \text{if } v_y = v_{[i]} \text{ for some } i \in [k^\star], \\ 0 & \text{o.w.} \end{cases}$$

Here, note that the game is translation-invariant with respect to $\mathbf{v}$, $\eta(\mathbf{v})$ is the reverse water level, and $(v_y - \eta(\mathbf{v}))_+$ is the volume of water filled in slot y. The solution is an equalizer if and only if $\eta(\mathbf{v}) \leq v_y$ for all y, that is, the (reverse) water level $\eta(\mathbf{v})$ is not higher than the final water levels for all slots.

Hence, for each observation y^{t-1} and a vector $\mathbf{v}_n(\cdot|y^{t-1}) \in \mathbb{R}^m$ defined as $v_n(y|y^{t-1}) := V_{\mathcal{F}}^{(n)}(y^{t-1}y)$, the optimal forecaster (13) is given by the reverse-water-filling solution

$$a_t^\star(\cdot|y^{t-1}) = a_{\texttt{water}}(\cdot|\mathbf{v}_n(\cdot|y^{t-1}))$$

and the conditional value is

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) = 1 + \eta(\mathbf{v}_n(\cdot|y^{t-1})).$$

For an equalizer to exist, $\mathbf{v}_n(\cdot|y^{t-1})$ must satisfy that $k^\star = m$, in which case

$$\eta(\mathbf{v}) = \frac{1}{m}\Big(\sum_{i=1}^{m} v_i - 1\Big).$$

In this case, the conditional value is

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) = 1 - \frac{1}{m} + \frac{1}{m}\sum_{y=1}^{m} V_{\mathcal{F}}^{(n)}(y^{t-1}y).$$

and the equalizer is

$$a_t^\star(\cdot|y^{t-1}) = (V_{\mathcal{F}}^{(n)}(y^{t-1}y))_{y=1}^m + \frac{\mathbb{1}}{m}\Big(1 - \sum_{i=1}^{m} V_{\mathcal{F}}^{(n)}(y^{t-1}i)\Big).$$

Repeatedly applying the recursive relation, we obtain

$$V_{\mathcal{F}}^{(n)}(y^{t-1}) = (n-t+1)\Big(1 - \frac{1}{m}\Big) - \mathbb{E}_{\varepsilon_t^n \sim \mathrm{Unif}([m])^{n-t+1}}\Big[\min_{\mathbf{f}\in\mathcal{F}} \overline{L}(\mathbf{f}, y^{t-1}\varepsilon_t^n)\Big],$$

which implies that

$$a_t^\star(y|y^{t-1}) = \frac{1}{m} + \mathbb{E}_{\varepsilon_t^n}\Big[\min_{f\in\mathcal{F}} L_f(y^{t-1}\varepsilon_t^n)\Big] - \mathbb{E}_{\varepsilon_{t+1}^n}\Big[\min_{f\in\mathcal{F}} L_f(y^{t-1}y\varepsilon_{t+1}^n)\Big]. \tag{14}$$

References

[1] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[2] Nicolo Cesa-Bianchi and Ohad Shamir. Efficient online learning via randomized rounding. In *Adv. Neural Inf. Proc. Syst.*, volume 24, 2011.

[3] Thomas M Cover. Behavior of sequential predictors of binary sequences. In *Proc. 4th Prague Conf. Inf. Theory, Stat. Decis. Funct. Random Process.*, pages 263–272. Prague: Publishing House of the Czechoslovak Academy of Sciences, 1967.

[4] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Trans. Inf. Theory*, 38(4):1258–1270, 1992.

[5] Alexander Rakhlin and Karthik Sridharan. A tutorial on online supervised learning with applications to node classification in social networks. *arXiv preprint arXiv:1608.09014*, 2016.