Online Learning under Logarithmic Loss

Changlong Wu & Wojciech Szpankowski

Center for Science of Information
Purdue University

October 21, 2024

Overview

- **Sequential Probability Assignment**
  - Weather forecasting, proper scoring, logarithmic loss
  - Bayesian algorithm
- **Minimax Regret under Log-loss**
  - Fixed design, Shtarkov sum
  - Truncated Bayesian Algorithm
  - Contextual Shtarkov sum
- **Application of Prediction with Log-loss**
  - Portfolio optimization
  - Converting prediction to investment strategy

Sequential Probability Assignment

Imagine a weather forecaster predicts the probability of rain tomorrow is 70%.

Sequential Probability Assignment

Imagine a weather forecaster predicts the probability of rain tomorrow is 70%.

But, it ends up being sunny the next day.

Sequential Probability Assignment

Imagine a weather forecaster predicts the probability of rain tomorrow is 70%.

But, it ends up being sunny the next day.

How should we meaningfully quantify the accuracy of this prediction?

Sequential Probability Assignment

Imagine a weather forecaster predicts the probability of rain tomorrow is 70%.

But, it ends up being sunny the next day.

How should we meaningfully quantify the accuracy of this prediction?

- The probability distribution for rain is different every day.
- We only observe one outcome (i.e., rain or no rain) for each distribution.

Sequential Probability Assignment

Imagine a weather forecaster predicts the probability of rain tomorrow is 70%.

But, it ends up being sunny the next day.

How should we meaningfully quantify the accuracy of this prediction?

- The probability distribution for rain is different every day.
- We only observe one outcome (i.e., rain or no rain) for each distribution.

Formally, we aim to find a loss function $\ell : \Delta(\{0,1\}) \times \{0,1\} \to \mathbb{R}$

Sequential Probability Assignment

Imagine a weather forecaster predicts the probability of rain tomorrow is 70%.

But, it ends up being sunny the next day.

How should we meaningfully quantify the accuracy of this prediction?

- The probability distribution for rain is different every day.
- We only observe one outcome (i.e., rain or no rain) for each distribution.

Formally, we aim to find a loss function $\ell : \Delta(\{0,1\}) \times \{0,1\} \to \mathbb{R}$ that satisfies the following minimal criteria:

1. It should penalize the true distribution minimally, i.e.,

$$\forall p, q \in \Delta(\{0,1\}),\ \mathbb{E}_{y \sim p}[\ell(p, y)] \leq \mathbb{E}_{y \sim p}[\ell(q, y)].$$

2. Ideally, the function ℓ should have a natural interpretation.

The Logarithmic Loss

It turns out that a natural choice is the so-called logarithmic loss (log-loss).

The Logarithmic Loss

It turns out that a natural choice is the so-called logarithmic loss (log-loss).

Let $\mathcal{Y}$ be a label space, and $\Delta(\mathcal{Y})$ be the set of all distributions over $\mathcal{Y}$.

The Logarithmic Loss

It turns out that a natural choice is the so-called logarithmic loss (log-loss).

Let $\mathcal{Y}$ be a label space, and $\Delta(\mathcal{Y})$ be the set of all distributions over $\mathcal{Y}$.

The logarithmic loss for any $p \in \Delta(\mathcal{Y})$ and $y \in \mathcal{Y}$ is defined as:

$$\ell^{\mathsf{log}}(p, y) = -\log p[y].$$

The Logarithmic Loss

It turns out that a natural choice is the so-called logarithmic loss (log-loss).

Let $\mathcal{Y}$ be a label space, and $\Delta(\mathcal{Y})$ be the set of all distributions over $\mathcal{Y}$.

The logarithmic loss for any $p \in \Delta(\mathcal{Y})$ and $y \in \mathcal{Y}$ is defined as:

$$\ell^{\mathsf{log}}(p, y) = -\log p[y].$$

Key properties of log-loss:

- It relates naturally to Shannon entropy and KL-divergence as: (verify it!)

$$\forall p, q \in \Delta(\mathcal{Y}),\ \mathbb{E}_{y \sim p}[\ell^{\mathsf{log}}(q, y)] = H(p) + \mathsf{KL}(p, q).$$

The Logarithmic Loss

It turns out that a natural choice is the so-called logarithmic loss (log-loss).

Let $\mathcal{Y}$ be a label space, and $\Delta(\mathcal{Y})$ be the set of all distributions over $\mathcal{Y}$.

The logarithmic loss for any $p \in \Delta(\mathcal{Y})$ and $y \in \mathcal{Y}$ is defined as:

$$\ell^{\text{log}}(p, y) = -\log p[y].$$

Key properties of log-loss:

- It relates naturally to Shannon entropy and KL-divergence as: (verify it!)

$$\forall p, q \in \Delta(\mathcal{Y}),\ \mathbb{E}_{y\sim p}[\ell^{\text{log}}(q, y)] = H(p) + \text{KL}(p, q).$$

- By the non-negativity of KL-divergence, this implies:

$$\mathbb{E}_{y\sim p}[\ell^{\text{log}}(p, y)] \leq \mathbb{E}_{y\sim p}[\ell^{\text{log}}(q, y)].$$

The Logarithmic Loss

It turns out that a natural choice is the so-called logarithmic loss (log-loss).

Let $\mathcal{Y}$ be a label space, and $\Delta(\mathcal{Y})$ be the set of all distributions over $\mathcal{Y}$.

The logarithmic loss for any $p \in \Delta(\mathcal{Y})$ and $y \in \mathcal{Y}$ is defined as:

$$\ell^{\log}(p, y) = -\log p[y].$$

Key properties of log-loss:

- It relates naturally to Shannon entropy and KL-divergence as: (verify it!)

$$\forall p, q \in \Delta(\mathcal{Y}), \; \mathbb{E}_{y\sim p}[\ell^{\log}(q, y)] = H(p) + \mathrm{KL}(p, q).$$

- By the non-negativity of KL-divergence, this implies:

$$\mathbb{E}_{y\sim p}[\ell^{\log}(p, y)] \leq \mathbb{E}_{y\sim p}[\ell^{\log}(q, y)].$$

- Equality is achieved when $p = q$.

Sequential Probability Assignment as Online Game

We now introduce the main learning paradigm of this lecture.

Sequential Probability Assignment as Online Game

We now introduce the main learning paradigm of this lecture.

Let $\mathcal{Y}$ be the label space, $\hat{\mathcal{Y}} := \Delta(\mathcal{Y})$ be the prediction space, $\mathcal{X}$ be the instance space and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be the hypothesis class.

Sequential Probability Assignment as Online Game

We now introduce the main learning paradigm of this lecture.

Let $\mathcal{Y}$ be the label space, $\hat{\mathcal{Y}} := \Delta(\mathcal{Y})$ be the prediction space, $\mathcal{X}$ be the instance space and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be the hypothesis class.

For $t = 1, \cdots, T$

- Nature selects an instance $\mathbf{x}_t \in \mathcal{X}$;
- Leaner predicts distribution $\hat{p}_t \in \hat{\mathcal{Y}}$;
- Nature selects true label $y_t \in \mathcal{Y}$;
- Learner suffers loss $\ell^{\log}(\hat{p}_t, y_t)$.

Sequential Probability Assignment as Online Game

We now introduce the main learning paradigm of this lecture.

Let $\mathcal{Y}$ be the label space, $\hat{\mathcal{Y}} := \Delta(\mathcal{Y})$ be the prediction space, $\mathcal{X}$ be the instance space and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be the hypothesis class.

For $t = 1, \cdots, T$

- Nature selects an instance $\mathbf{x}_t \in \mathcal{X}$;
- Leaner predicts distribution $\hat{p}_t \in \hat{\mathcal{Y}}$;
- Nature selects true label $y_t \in \mathcal{Y}$;
- Learner suffers loss $\ell^{\log}(\hat{p}_t, y_t)$.

Goal of Learner: Find predictor Φ that minimizes the worst-case regret:

$$\mathrm{reg}_T(\mathcal{H}, \Phi) = \sup_{\mathbf{x}^T, y^T} \left[\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Regret Bound for Finite Class: the Bayesian Algorithm

Recall from **Lecture 2** that for a finite class $\mathcal{H}$, the EWA algorithm Φ enjoys the worst-case regret for bounded convex:

$$\text{reg}_T(\mathcal{H}, \Phi) \leq O(\sqrt{T \log |\mathcal{H}|}).$$

Regret Bound for Finite Class: the Bayesian Algorithm

Recall from **Lecture 2** that for a finite class $\mathcal{H}$, the EWA algorithm Φ enjoys the worst-case regret for bounded convex:

$$\mathrm{reg}_T(\mathcal{H}, \Phi) \leq O(\sqrt{T \log |\mathcal{H}|}).$$

Unfortunately, this does not apply to log-loss, since $\ell^{\log}(\cdot, y)$ is not bounded.

Regret Bound for Finite Class: the Bayesian Algorithm

Recall from **Lecture 2** that for a finite class $\mathcal{H}$, the EWA algorithm Φ enjoys the worst-case regret for bounded convex:

$$\mathrm{reg}_T(\mathcal{H}, \Phi) \leq O(\sqrt{T \log |\mathcal{H}|}).$$

Unfortunately, this does not apply to log-loss, since $\ell^{\log}(\cdot, y)$ is not bounded.

Let $\mathcal{H} = \{h_1, \cdots, h_K\}$.

The Bayesian Algorithm:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step t, predict $\hat{p}_t := \sum_{k=1}^K \tilde{p}_t[k] \cdot h_k(\mathbf{x}_t)$, where

$$\forall k \in [K], \ \tilde{p}_t[k] = \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^K \mathbf{w}_k^{(t-1)}}.$$

3. Let y_t be the true label, and update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)} \cdot h_k(\mathbf{x}_t)[y_t]$.

Regret Bound for Finite Class: the Bayesian Algorithm

Recall from **Lecture 2** that for a finite class $\mathcal{H}$, the EWA algorithm Φ enjoys the worst-case regret for bounded convex:

$$\mathrm{reg}_T(\mathcal{H}, \Phi) \leq O(\sqrt{T \log |\mathcal{H}|}).$$

Unfortunately, this does not apply to log-loss, since $\ell^{\log}(\cdot, y)$ is not bounded.

Let $\mathcal{H} = \{h_1, \cdots, h_K\}$.

The Bayesian Algorithm:

1. Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
2. At each step t, predict $\hat{p}_t := \sum_{k=1}^{K} \tilde{p}_t[k] \cdot h_k(\mathbf{x}_t)$, where

$$\forall k \in [K], \ \tilde{p}_t[k] = \frac{\mathbf{w}_k^{(t-1)}}{\sum_{k=1}^{K} \mathbf{w}_k^{(t-1)}}.$$

3. Let y_t be the true label, and update $\mathbf{w}_k^{(t)} = \mathbf{w}_k^{(t-1)} \cdot h_k(\mathbf{x}_t)[y_t]$.

Observe that $h_k(\mathbf{x}_t)[y_t] = e^{-\ell^{\log}(h_k(\mathbf{x}_t), y_t)}$, i.e., the **Bayesian algorithm** is simply the EWA algorithm with a learning rate of $\eta = 1$.

Regret Bound for Bayesian Algorithm

Theorem 1: Let $\mathcal{H}$ be a finite class. The Bayesian algorithm Φ enjoys the worst-case regret under logarithmic loss:

$$\text{reg}_T(\mathcal{H}, \Phi) \leq \log |\mathcal{H}|.$$

Regret Bound for Bayesian Algorithm

Theorem 1: Let $\mathcal{H}$ be a finite class. The Bayesian algorithm Φ enjoys the worst-case regret under logarithmic loss:

$$\mathsf{reg}_T(\mathcal{H}, \Phi) \leq \log |\mathcal{H}|.$$

- Observe that the regret bound is tighter than the $O(\sqrt{T \log |\mathcal{H}|})$ regret bound for bounded Lipschitz losses.

Regret Bound for Bayesian Algorithm

Theorem 1: Let $\mathcal{H}$ be a finite class. The Bayesian algorithm Φ enjoys the worst-case regret under logarithmic loss:

$$\mathsf{reg}_T(\mathcal{H}, \Phi) \leq \log |\mathcal{H}|.$$

- Observe that the regret bound is tighter than the $O(\sqrt{T \log |\mathcal{H}|})$ regret bound for bounded Lipschitz losses.
- Although our predictions are probabilities, we do not assume any probabilistic mechanism for generating the data.

Regret Bound for Bayesian Algorithm

Theorem 1: Let $\mathcal{H}$ be a finite class. The Bayesian algorithm Φ enjoys the worst-case regret under logarithmic loss:

$$\text{reg}_T(\mathcal{H}, \Phi) \leq \log |\mathcal{H}|.$$

- Observe that the regret bound is tighter than the $O(\sqrt{T \log |\mathcal{H}|})$ regret bound for bounded Lipschitz losses.
- Although our predictions are probabilities, we do not assume any probabilistic mechanism for generating the data.
- The regret bound holds for any individual sequences $\mathbf{x}^T, y^T$.

Proof of Theorem 1

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

Proof of Theorem 1

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

Observe that

$$\log \frac{W^{(t)}}{W^{(t-1)}} = \log \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} h_k(\mathbf{x}_t)[y_t] = \log \hat{p}_t[y_t] = -\ell^{\log}(\hat{p}_t, y_t).$$

Proof of Theorem 1

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

Observe that

$$\log \frac{W^{(t)}}{W^{(t-1)}} = \log \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} h_k(\mathbf{x}_t)[y_t] = \log \hat{p}_t[y_t] = -\ell^{\log}(\hat{p}_t, y_t).$$

Summing from $t = 1$ to T, we have

$$\log \frac{W^{(T)}}{W^{(0)}} = -\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t).$$

Proof of Theorem 1

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

Observe that

$$\log \frac{W^{(t)}}{W^{(t-1)}} = \log \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} h_k(\mathbf{x}_t)[y_t] = \log \hat{p}_t[y_t] = -\ell^{\log}(\hat{p}_t, y_t).$$

Summing from $t = 1$ to T, we have

$$\log \frac{W^{(T)}}{W^{(0)}} = -\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t).$$

Note that

$$\log W^{(T)} \geq \sup_k \log \mathbf{w}_k^{(T)} = \sup_k \log \prod_{t=1}^{T} h_k(\mathbf{x}_t)[y_t] = -\inf_k \sum_{t=1}^{T} \ell^{\log}(h_k(\mathbf{x}_t), y_t).$$

Proof of Theorem 1

We again define the potential $W^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)}$ with $W^{(0)} = K$.

Observe that

$$\log \frac{W^{(t)}}{W^{(t-1)}} = \log \sum_{k=1}^{K} \frac{\mathbf{w}_k^{(t-1)}}{W^{(t-1)}} h_k(\mathbf{x}_t)[y_t] = \log \hat{p}_t[y_t] = -\ell^{\log}(\hat{p}_t, y_t).$$

Summing from $t = 1$ to T, we have

$$\log \frac{W^{(T)}}{W^{(0)}} = -\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t).$$

Note that

$$\log W^{(T)} \geq \sup_k \log \mathbf{w}_k^{(T)} = \sup_k \log \prod_{t=1}^{T} h_k(\mathbf{x}_t)[y_t] = -\inf_k \sum_{t=1}^{T} \ell^{\log}(h_k(\mathbf{x}_t), y_t).$$

Therefore,

$$\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_k \sum_{t=1}^{T} \ell^{\log}(h_k(\mathbf{x}_t), y_t) \leq \log K.$$

Overview

- **Sequential Probability Assignment**
  - Weather forecasting, proper scoring, logarithmic loss
  - Bayesian algorithm
- **Minimax Regret under Log-loss**
  - Fixed design, Shtarkov sum
  - Truncated Bayesian Algorithm
  - Contextual Shtarkov sum
- **Application of Prediction with Log-loss**
  - Portfolio optimization
  - Converting prediction to investment strategy

Minimax Regret under Log-loss

We have demonstrated that the Bayesian algorithm achieves $\log|\mathcal{H}|$ regret under log-loss for a finite class $\mathcal{H}$.

Minimax Regret under Log-loss

We have demonstrated that the Bayesian algorithm achieves $\log |\mathcal{H}|$ regret under log-loss for a finite class $\mathcal{H}$.

Several issues remain:

1. The Bayesian algorithm cannot be applied directly to infinite classes.
2. It is unclear whether the $\log |\mathcal{H}|$ bound is tight.

Minimax Regret under Log-loss

We have demonstrated that the Bayesian algorithm achieves $\log |\mathcal{H}|$ regret under log-loss for a finite class $\mathcal{H}$.

Several issues remain:

1. The Bayesian algorithm cannot be applied directly to infinite classes.
2. It is unclear whether the $\log |\mathcal{H}|$ bound is tight.

Problem 1: What intrinsic complexity measure of $\mathcal{H}$ determines the minimax regret $\text{reg}_T(\mathcal{H})$ under log-loss?

Minimax Regret under Log-loss

We have demonstrated that the Bayesian algorithm achieves $\log|\mathcal{H}|$ regret under log-loss for a finite class $\mathcal{H}$.

Several issues remain:

1. The Bayesian algorithm cannot be applied directly to infinite classes.
2. It is unclear whether the $\log|\mathcal{H}|$ bound is tight.

Problem 1: What intrinsic complexity measure of $\mathcal{H}$ determines the minimax regret $\mathsf{reg}_T(\mathcal{H})$ under log-loss?

Problem 2: What algorithm achieves the minimax regret?

Sequential vs. Fixed Design Regret

For simplicity, we will assume $\mathcal{Y}$ is finite in our following discussions.

Sequential vs. Fixed Design Regret

For simplicity, we will assume $\mathcal{Y}$ is finite in our following discussions.

For any given $\mathbf{x}^T$, we define the fixed design minimax regret as:

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \inf_{\Phi} \sup_{y^T} \left[\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Sequential vs. Fixed Design Regret

For simplicity, we will assume $\mathcal{Y}$ is finite in our following discussions.

For any given $\mathbf{x}^T$, we define the fixed design minimax regret as:

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \inf_{\Phi} \sup_{y^T} \left[\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Recall the (sequential) minimax regret is defined as:

$$\mathsf{reg}_T(\mathcal{H}) = \inf_{\Phi} \sup_{\mathbf{x}^T, y^T} \left[\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Sequential vs. Fixed Design Regret

For simplicity, we will assume $\mathcal{Y}$ is finite in our following discussions.

For any given $\mathbf{x}^T$, we define the fixed design minimax regret as:

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \inf_{\Phi} \sup_{y^T} \left[\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Recall the (sequential) minimax regret is defined as:

$$\mathsf{reg}_T(\mathcal{H}) = \inf_{\Phi} \sup_{\mathbf{x}^T, y^T} \left[\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

It is easy to observe that: (Why?)

$$\sup_{\mathbf{x}^T} \mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) \leq \mathsf{reg}_T(\mathcal{H}).$$

Characterizing Fixed-Design Minimax Regret: Shtarkov Sum

Shtarkov Sum: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class and $\mathbf{x}^T$ be any given instances. The *Shtarkov sum* of $\mathcal{H}$ conditioning on $\mathbf{x}^T$ is defined as

$$\mathrm{Sht}(\mathcal{H} \mid \mathbf{x}^T) = \sum_{y^T \in \mathcal{Y}^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t].$$

Characterizing Fixed-Design Minimax Regret: Shtarkov Sum

Shtarkov Sum: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class and $\mathbf{x}^T$ be any given instances. The *Shtarkov sum* of $\mathcal{H}$ conditioning on $\mathbf{x}^T$ is defined as

$$\mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T) = \sum_{y^T \in \mathcal{Y}^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t].$$

Example 1: Let $\mathcal{H}$ be a finite class, we have for any $\mathbf{x}^T$ that

$$\begin{aligned}
\mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T) &= \sum_{y^T \in \mathcal{Y}^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \\
&\le \sum_{y^T \in \mathcal{Y}^T} \sum_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \\
&= \sum_{h \in \mathcal{H}} \sum_{y^T \in \mathcal{Y}^T} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \overset{(\star)}{\le} \sum_{h \in \mathcal{H}} 1 = |\mathcal{H}|.
\end{aligned}$$

Characterizing Fixed-Design Minimax Regret: Shtarkov Sum

Theorem 2: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class, and let $\mathbf{x}^T$ be any given instances. Then

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T).$$

Characterizing Fixed-Design Minimax Regret: Shtarkov Sum

Theorem 2: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class, and let $\mathbf{x}^T$ be any given instances. Then

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T).$$

- These two quantities are exactly equal.
- For a finite class $\mathcal{H}$, we immediately have

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T) \leq \log |\mathcal{H}|.$$

- The Shtarkov sum forms a lower bound for the (sequential) minimax regret:

$$\mathsf{reg}_T(\mathcal{H}) \geq \sup_{\mathbf{x}^T} \mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) \geq \sup_{\mathbf{x}^T} \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T).$$

Proof of Theorem 2

We introduce the short-hand notations

$$P_h(y^T \mid \mathbf{x}^T) = \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t], \qquad \hat{Q}(y^T) = \prod_{t=1}^{T} \hat{p}_t[y_t].$$

Observe, by definition of log-loss, that

$$\begin{aligned}
\mathrm{reg}_T^{\mathrm{fix}}(\mathcal{H} \mid \mathbf{x}^T) &= \inf_{\hat{Q}} \sup_{y^T} \left[-\log \hat{Q}(y^T) + \log \sup_h P_h(y^T \mid \mathbf{x}^T) \right] \\
&= \inf_{\hat{Q}} \sup_{y^T} \left[-\log \hat{Q}(y^T) + \log P^*(y^T \mid \mathbf{x}^T) \right] + \log \sum_{y^T} \sup_h P_h(y^T \mid \mathbf{x}^T) \\
&\overset{(\star)}{=} \log \sum_{y^T} \sup_h P_h(y^T \mid \mathbf{x}^T) = \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T),
\end{aligned}$$

where $P^*(y^T \mid \mathbf{x}^T) := \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum \sup_h P_h(y^T|\mathbf{x}^T)}$ and $(\star)$ attains when $\hat{Q}(\cdot) \equiv P^*(\cdot \mid \mathbf{x}^T)$.

Minimax Optimal Predictor: Normalized Maximum Likelihood

A by-product of our previous proof shows that the minimax optimal predictor satisfies equality

$$\hat{Q}(\cdot) \equiv P^*(\cdot \mid \mathbf{x}^T),$$

Minimax Optimal Predictor: Normalized Maximum Likelihood

A by-product of our previous proof shows that the minimax optimal predictor satisfies equality

$$\hat{Q}(\cdot) \equiv P^*(\cdot \mid \mathbf{x}^T),$$

where $P^*(y^T \mid \mathbf{x}^T) := \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum_{y^T} \sup_h P_h(y^T|\mathbf{x}^T)}$. and $\hat{Q}(y^T) = \prod_{t=1}^T \hat{p}_t[y_t]$.

Minimax Optimal Predictor: Normalized Maximum Likelihood

A by-product of our previous proof shows that the minimax optimal predictor satisfies equality

$$\hat{Q}(\cdot) \equiv P^*(\cdot \mid \mathbf{x}^T),$$

where $P^*(y^T \mid \mathbf{x}^T) := \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum_{y^T} \sup_h P_h(y^T|\mathbf{x}^T)}$. and $\hat{Q}(y^T) = \prod_{t=1}^T \hat{p}_t[y_t]$.

To satisfy the equality, we can define (Why?)

$$\hat{p}_t[y] = \frac{\sum_{y^{T-t}} P^*(y^{t-1} y y^{T-t} \mid \mathbf{x}^T)}{\sum_{y^{T-t+1}} P^*(y^{t-1} y^{T-t+1} \mid \mathbf{x}^T)}$$

Minimax Optimal Predictor: Normalized Maximum Likelihood

A by-product of our previous proof shows that the minimax optimal predictor satisfies equality

$$\hat{Q}(\cdot) \equiv P^*(\cdot \mid \mathbf{x}^T),$$

where $P^*(y^T \mid \mathbf{x}^T) := \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum_{y^T} \sup_h P_h(y^T|\mathbf{x}^T)}$. and $\hat{Q}(y^T) = \prod_{t=1}^T \hat{p}_t[y_t]$.

To satisfy the equality, we can define (Why?)

$$\hat{p}_t[y] = \frac{\sum_{y^{T-t}} P^*(y^{t-1} y y^{T-t} \mid \mathbf{x}^T)}{\sum_{y^{T-t+1}} P^*(y^{t-1} y^{T-t+1} \mid \mathbf{x}^T)}$$

This predictor is known as the Normalized Maximum Likelihood (NML) predictor.

Bounding the (Sequential) Minimax Regret

We have shown that the fixed-design minimax regret is completely characterized by the Shtarkov sum.

Bounding the (Sequential) Minimax Regret

We have shown that the fixed-design minimax regret is completely characterized by the Shtarkov sum.

Moreover, the minimax optimal predictor is given by the NML predictor.

Bounding the (Sequential) Minimax Regret

We have shown that the fixed-design minimax regret is completely characterized by the Shtarkov sum.

Moreover, the minimax optimal predictor is given by the NML predictor.

What about the sequential minimax regret?

Bounding the (Sequential) Minimax Regret

We have shown that the fixed-design minimax regret is completely characterized by the Shtarkov sum.

Moreover, the minimax optimal predictor is given by the NML predictor.

What about the sequential minimax regret?

(Distribution) Sequential Cover: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class. We say a sequential function class $\mathcal{G} \subset \Delta(\mathcal{Y})^{\mathcal{X}^*}$ sequentially α-covers $\mathcal{H}$ up to step T if, for any $h \in \mathcal{H}$ and $\mathbf{x}^T$, there exists $g \in \mathcal{G}$ such that

$$\forall t \leq T, \; \|g(\mathbf{x}^t) - h(\mathbf{x}_t)\|_\infty \leq \alpha,$$

where $\|p - q\|_\infty = \sup_{y \in \mathcal{Y}} |p[y] - q[y]|$.

Bounding the (Sequential) Minimax Regret

We have shown that the fixed-design minimax regret is completely characterized by the Shtarkov sum.

Moreover, the minimax optimal predictor is given by the NML predictor.

What about the sequential minimax regret?

(Distribution) Sequential Cover: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class. We say a sequential function class $\mathcal{G} \subset \Delta(\mathcal{Y})^{\mathcal{X}^*}$ sequentially α-covers $\mathcal{H}$ up to step T if, for any $h \in \mathcal{H}$ and $\mathbf{x}^T$, there exists $g \in \mathcal{G}$ such that

$$\forall t \leq T, \ \|g(\mathbf{x}^t) - h(\mathbf{x}_t)\|_\infty \leq \alpha,$$

where $\|p - q\|_\infty = \sup_{y \in \mathcal{Y}} |p[y] - q[y]|$.

- Note that a crucial property when we apply the sequential cover for a Lipschitz loss ℓ is that: $\ell(\hat{y}_1, y) - \ell(\hat{y}_2, y) \leq L|\hat{y}_1 - \hat{y}_2|$.

Bounding the (Sequential) Minimax Regret

We have shown that the fixed-design minimax regret is completely characterized by the Shtarkov sum.

Moreover, the minimax optimal predictor is given by the NML predictor.

What about the sequential minimax regret?

(Distribution) Sequential Cover: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class. We say a sequential function class $\mathcal{G} \subset \Delta(\mathcal{Y})^{\mathcal{X}^*}$ sequentially α-covers $\mathcal{H}$ up to step T if, for any $h \in \mathcal{H}$ and $\mathbf{x}^T$, there exists $g \in \mathcal{G}$ such that

$$\forall t \leq T, \ \|g(\mathbf{x}^t) - h(\mathbf{x}_t)\|_\infty \leq \alpha,$$

where $\|p - q\|_\infty = \sup_{y \in \mathcal{Y}} |p[y] - q[y]|$.

- Note that a crucial property when we apply the sequential cover for a Lipschitz loss ℓ is that: $\ell(\hat{y}_1, y) - \ell(\hat{y}_2, y) \leq L|\hat{y}_1 - \hat{y}_2|$.
- Therefore, small regret on the cover $\mathcal{G}_\alpha$ automatically implies small regret on $\mathcal{H}$, offset by $\alpha L T$.

Bounding the (Sequential) Minimax Regret

We have shown that the fixed-design minimax regret is completely characterized by the Shtarkov sum.

Moreover, the minimax optimal predictor is given by the NML predictor.

What about the sequential minimax regret?

(Distribution) Sequential Cover: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class. We say a sequential function class $\mathcal{G} \subset \Delta(\mathcal{Y})^{\mathcal{X}^*}$ sequentially α-covers $\mathcal{H}$ up to step T if, for any $h \in \mathcal{H}$ and $\mathbf{x}^T$, there exists $g \in \mathcal{G}$ such that

$$\forall t \leq T, \ \|g(\mathbf{x}^t) - h(\mathbf{x}_t)\|_\infty \leq \alpha,$$

where $\|p - q\|_\infty = \sup_{y \in \mathcal{Y}} |p[y] - q[y]|$.

- Note that a crucial property when we apply the sequential cover for a Lipschitz loss ℓ is that: $\ell(\hat{y}_1, y) - \ell(\hat{y}_2, y) \leq L|\hat{y}_1 - \hat{y}_2|$.
- Therefore, small regret on the cover $\mathcal{G}_\alpha$ automatically implies small regret on $\mathcal{H}$, offset by $\alpha L T$.
- This, unfortunately, is not true for log-loss, e.g., $\ell^{\log}(0, y) - \ell^{\log}(\alpha, y) = \infty$.

From Covering to Dominance: The Smooth Truncation

Lemma 1: Let $\mathcal{G}$ be a sequential α-cover of $\mathcal{H}$. Then, for any $h \in \mathcal{H}$ and $\mathbf{x}^T, y^T$, there exists $g \in \mathcal{G}$ such that

$$\frac{\prod_{t=1}^{T} h(\mathbf{x}_t)[y_t]}{\prod_{t=1}^{T} g^{(\alpha)}(\mathbf{x}^t)[y_t]} \leq (1 + \alpha|\mathcal{Y}|)^T,$$

where $g^{(\alpha)} = \frac{g+\alpha}{1+\alpha|\mathcal{Y}|}$ is the smooth truncation of g.

From Covering to Dominance: The Smooth Truncation

Lemma 1: Let $\mathcal{G}$ be a sequential α-cover of $\mathcal{H}$. Then, for any $h \in \mathcal{H}$ and $\mathbf{x}^T, y^T$, there exists $g \in \mathcal{G}$ such that

$$\frac{\prod_{t=1}^{T} h(\mathbf{x}_t)[y_t]}{\prod_{t=1}^{T} g^{(\alpha)}(\mathbf{x}^t)[y_t]} \leq (1 + \alpha|\mathcal{Y}|)^T,$$

where $g^{(\alpha)} = \frac{g+\alpha}{1+\alpha|\mathcal{Y}|}$ is the smooth truncation of g.

Proof: For any $h \in \mathcal{H}$ and $\mathbf{x}^T, y^T$, we choose $g \in \mathcal{G}$ as the sequential α-cover of h on $\mathbf{x}^T$. This implies that, for all $t \leq T$ and $y \in \mathcal{Y}$,

$$h(\mathbf{x}_t)[y] \leq g(\mathbf{x}^t)[y] + \alpha.$$

Therefore, for any $t \leq T$, we have

$$\frac{h[y_t]}{g^{(\alpha)}[y_t]} = \frac{h[y_t]}{(g[y_t] + \alpha)/(1 + \alpha|\mathcal{Y}|)} \leq (1 + \alpha|\mathcal{Y}|).$$

Bounding sequential Minimax Regret via Sequential Cover

Theorem 2: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class that admits a sequential α-cover $\mathcal{G}_\alpha$ for all $\alpha \geq 0$. Then

$$\mathrm{reg}_T(\mathcal{H}) \leq \inf_{\alpha \geq 0} \{\alpha |\mathcal{Y}| T + \log |\mathcal{G}_\alpha|\}.$$

Bounding sequential Minimax Regret via Sequential Cover

Theorem 2: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class that admits a sequential α-cover $\mathcal{G}_\alpha$ for all $\alpha \geq 0$. Then

$$\mathsf{reg}_T(\mathcal{H}) \leq \inf_{\alpha \geq 0} \{\alpha |\mathcal{Y}| T + \log |\mathcal{G}_\alpha|\}.$$

Example 2: Let $\mathcal{Y} := \{0, 1\}$, $\mathcal{X} := B_2$ and

$$\mathcal{H}^{\mathsf{lin}} := \{h_{\mathbf{w}}(\mathbf{x}) := |\langle \mathbf{w}, \mathbf{x} \rangle| : \mathbf{w} \in B_2\} \subset [0, 1]^{\mathcal{X}}.$$

Here we interpreter $h(\mathbf{x}) \in [0, 1]$ as Bernoulli distribution with parameter $h(\mathbf{x})$.

From **lecture 3**, we know that $|\log \mathcal{G}_\alpha| \leq \tilde{O}(\alpha^{-2})$. This leads to the regret bound (verify it!)

$$\mathsf{reg}_T(\mathcal{H}^{\mathsf{lin}}) \leq \tilde{O}(T^{2/3}).$$

Proof of Theorem 2

Define $\mathcal{G}_{\alpha}^{(\alpha)} = \left\{ \frac{g+\alpha}{1+\alpha|\mathcal{Y}|} : g \in \mathcal{G}_{\alpha} \right\}$ as the smooth truncated class of $\mathcal{G}_{\alpha}$.

Proof of Theorem 2

Define $\mathcal{G}_{\alpha}^{(\alpha)} = \left\{ \frac{g+\alpha}{1+\alpha|\mathcal{Y}|} : g \in \mathcal{G}_{\alpha} \right\}$ as the smooth truncated class of $\mathcal{G}_{\alpha}$.

Let Φ be the predictor running the Bayesian algorithm over $\mathcal{G}_{\alpha}^{(\alpha)}$.

Proof of Theorem 2

Define $\mathcal{G}_\alpha^{(\alpha)} = \left\{ \frac{g+\alpha}{1+\alpha|\mathcal{Y}|} : g \in \mathcal{G}_\alpha \right\}$ as the smooth truncated class of $\mathcal{G}_\alpha$.

Let Φ be the predictor running the Bayesian algorithm over $\mathcal{G}_\alpha^{(\alpha)}$.

We have for any $\mathbf{x}^T, y^T$ that

$$\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{g \in \mathcal{G}_\alpha^{(\alpha)}} \sum_{t=1}^{T} \ell^{\log}(g(\mathbf{x}^t), y_t) \leq \log|\mathcal{G}_\alpha^{(\alpha)}| = \log|\mathcal{G}_\alpha|.$$

Proof of Theorem 2

Define $\mathcal{G}_\alpha^{(\alpha)} = \left\{ \frac{g+\alpha}{1+\alpha|\mathcal{Y}|} : g \in \mathcal{G}_\alpha \right\}$ as the smooth truncated class of $\mathcal{G}_\alpha$.

Let Φ be the predictor running the Bayesian algorithm over $\mathcal{G}_\alpha^{(\alpha)}$.

We have for any $\mathbf{x}^T, y^T$ that

$$\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{g \in \mathcal{G}_\alpha^{(\alpha)}} \sum_{t=1}^{T} \ell^{\log}(g(\mathbf{x}^t), y_t) \leq \log |\mathcal{G}_\alpha^{(\alpha)}| = \log |\mathcal{G}_\alpha|.$$

Invoking Lemma 1, we have (verify it!)

$$- \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \leq - \inf_{g \in \mathcal{G}_\alpha^{(\alpha)}} \sum_{t=1}^{T} \ell^{\log}(g(\mathbf{x}^t), y_t) + T \log(1 + \alpha|\mathcal{Y}|). \quad (1)$$

Proof of Theorem 2

Define $\mathcal{G}_\alpha^{(\alpha)} = \left\{ \frac{g+\alpha}{1+\alpha|\mathcal{Y}|} : g \in \mathcal{G}_\alpha \right\}$ as the smooth truncated class of $\mathcal{G}_\alpha$.

Let Φ be the predictor running the Bayesian algorithm over $\mathcal{G}_\alpha^{(\alpha)}$.

We have for any $\mathbf{x}^T, y^T$ that

$$\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{g \in \mathcal{G}_\alpha^{(\alpha)}} \sum_{t=1}^{T} \ell^{\log}(g(\mathbf{x}^t), y_t) \leq \log |\mathcal{G}_\alpha^{(\alpha)}| = \log |\mathcal{G}_\alpha|.$$

Invoking Lemma 1, we have (verify it!)

$$-\inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \leq -\inf_{g \in \mathcal{G}_\alpha^{(\alpha)}} \sum_{t=1}^{T} \ell^{\log}(g(\mathbf{x}^t), y_t) + T \log(1 + \alpha|\mathcal{Y}|). \quad (1)$$

The theorem follows by noting that $\log(1 + \alpha|\mathcal{Y}|) \leq \alpha|\mathcal{Y}|$.

Proof of Theorem 2

Define $\mathcal{G}_\alpha^{(\alpha)} = \left\{ \frac{g+\alpha}{1+\alpha|\mathcal{Y}|} : g \in \mathcal{G}_\alpha \right\}$ as the smooth truncated class of $\mathcal{G}_\alpha$.

Let Φ be the predictor running the Bayesian algorithm over $\mathcal{G}_\alpha^{(\alpha)}$.

We have for any $\mathbf{x}^T, y^T$ that

$$\sum_{t=1}^{T} \ell^{\log}(\hat{p}_t, y_t) - \inf_{g \in \mathcal{G}_\alpha^{(\alpha)}} \sum_{t=1}^{T} \ell^{\log}(g(\mathbf{x}^t), y_t) \leq \log |\mathcal{G}_\alpha^{(\alpha)}| = \log |\mathcal{G}_\alpha|.$$

Invoking Lemma 1, we have (verify it!)

$$-\inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \leq -\inf_{g \in \mathcal{G}_\alpha^{(\alpha)}} \sum_{t=1}^{T} \ell^{\log}(g(\mathbf{x}^t), y_t) + T \log(1 + \alpha|\mathcal{Y}|). \quad (1)$$

The theorem follows by noting that $\log(1 + \alpha|\mathcal{Y}|) \leq \alpha|\mathcal{Y}|$.

Note: The use of $\mathcal{G}_\alpha^{(\alpha)}$ instead of $\mathcal{G}_\alpha$ is crucial for (1) to work.

Sub-optimality of Covering-Based Bounds

We now mention the following theorem without proof.

Theorem 3: Let $\mathcal{Y} := \{0,1\}$, and assume $\hat{\mathcal{Y}} := [0,1]$, interpreted as Bernoulli distributions. Then for any class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ with a sequential α-cover $\mathcal{G}_\alpha$ of size $\log|\mathcal{G}_\alpha| \leq \tilde{O}(\alpha^{-p})$ for all $\alpha \geq 0$, we have

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p}{p+1}}).$$

Moreover, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \geq \tilde{\Omega}(T^{\frac{p}{p+1}}).$$

Furthermore, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p-1}{p}}).$$

Sub-optimality of Covering-Based Bounds

We now mention the following theorem without proof.

Theorem 3: Let $\mathcal{Y} := \{0, 1\}$, and assume $\hat{\mathcal{Y}} := [0, 1]$, interpreted as Bernoulli distributions. Then for any class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ with a sequential α-cover $\mathcal{G}_\alpha$ of size $\log |\mathcal{G}_\alpha| \leq \tilde{O}(\alpha^{-p})$ for all $\alpha \geq 0$, we have

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p}{p+1}}).$$

Moreover, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \geq \tilde{\Omega}(T^{\frac{p}{p+1}}).$$

Furthermore, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p-1}{p}}).$$

- Sequential α-covering characterizes minimax regret for the worst classes, but not for certain easy classes!

Sub-optimality of Covering-Based Bounds

We now mention the following theorem without proof.

Theorem 3: Let $\mathcal{Y} := \{0, 1\}$, and assume $\hat{\mathcal{Y}} := [0, 1]$, interpreted as Bernoulli distributions. Then for any class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ with a sequential α-cover $\mathcal{G}_\alpha$ of size $\log |\mathcal{G}_\alpha| \leq \tilde{O}(\alpha^{-p})$ for all $\alpha \geq 0$, we have

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p}{p+1}}).$$

Moreover, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \geq \tilde{\Omega}(T^{\frac{p}{p+1}}).$$

Furthermore, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p-1}{p}}).$$

- Sequential α-covering characterizes minimax regret for the worst classes, but not for certain easy classes!
  - For the proof, see Wu, Heidari, Grama, Szpankowski in (NeurIPS 2022).

Sub-optimality of Covering-Based Bounds

We now mention the following theorem without proof.

Theorem 3: Let $\mathcal{Y} := \{0, 1\}$, and assume $\hat{\mathcal{Y}} := [0, 1]$, interpreted as Bernoulli distributions. Then for any class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ with a sequential α-cover $\mathcal{G}_\alpha$ of size $\log |\mathcal{G}_\alpha| \leq \tilde{O}(\alpha^{-p})$ for all $\alpha \geq 0$, we have

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p}{p+1}}).$$

Moreover, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \geq \tilde{\Omega}(T^{\frac{p}{p+1}}).$$

Furthermore, for any $p \geq 2$, there exists a class that satisfies the above condition and

$$\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p-1}{p}}).$$

- Sequential α-covering characterizes minimax regret for the worst classes, but not for certain easy classes!
  - For the proof, see Wu, Heidari, Grama, Szpankowski in (NeurIPS 2022).
- We need a new complexity measure...

The Contextual Shtarkov Sum

Very recently, Liu, Attias, and Roy (to appear in NeurIPS 2024) demonstrated that a variant of the Shtarkov sum with context completely characterizes the (sequential) minimax regret...

The Contextual Shtarkov Sum

Very recently, Liu, Attias, and Roy (to appear in NeurIPS 2024) demonstrated that a variant of the Shtarkov sum with context completely characterizes the (sequential) minimax regret...

Contextual Shtarkov Sum: Let $\tau : \bigcup_{t=1}^{T} \mathcal{Y}^t \to \mathcal{X}$ be an $\mathcal{X}$-valued $|\mathcal{Y}|$-ary tree of depth T. The contextual Shtarkov sum w.r.t. τ is defined as

$$\mathsf{Sht}(\mathcal{H} \mid \tau) = \sum_{y^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\tau(y^{t-1}))[y_t].$$

The Contextual Shtarkov Sum

Very recently, Liu, Attias, and Roy (to appear in NeurIPS 2024) demonstrated that a variant of the Shtarkov sum with context completely characterizes the (sequential) minimax regret...

Contextual Shtarkov Sum: Let $\tau : \bigcup_{t=1}^{T} \mathcal{Y}^t \to \mathcal{X}$ be an $\mathcal{X}$-valued $|\mathcal{Y}|$-ary tree of depth T. The contextual Shtarkov sum w.r.t. τ is defined as

$$\mathsf{Sht}(\mathcal{H} \mid \tau) = \sum_{y^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\tau(y^{t-1}))[y_t].$$

Theorem 4: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class. Then:

$$\mathsf{reg}_T(\mathcal{H}) = \sup_{\tau} \log \mathsf{Sht}(\mathcal{H} \mid \tau).$$

The Contextual Shtarkov Sum

Very recently, Liu, Attias, and Roy (to appear in NeurIPS 2024) demonstrated that a variant of the Shtarkov sum with context completely characterizes the (sequential) minimax regret...

Contextual Shtarkov Sum: Let $\tau : \bigcup_{t=1}^{T} \mathcal{Y}^t \to \mathcal{X}$ be an $\mathcal{X}$-valued $|\mathcal{Y}|$-ary tree of depth T. The contextual Shtarkov sum w.r.t. τ is defined as

$$\mathsf{Sht}(\mathcal{H} \mid \tau) = \sum_{y^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\tau(y^{t-1}))[y_t].$$

Theorem 4: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class. Then:

$$\mathsf{reg}_T(\mathcal{H}) = \sup_{\tau} \log \mathsf{Sht}(\mathcal{H} \mid \tau).$$

- This result can be used to recover Theorem 2 using (smaller) local covers.

The Contextual Shtarkov Sum

Very recently, Liu, Attias, and Roy (to appear in NeurIPS 2024) demonstrated that a variant of the Shtarkov sum with context completely characterizes the (sequential) minimax regret...

Contextual Shtarkov Sum: Let $\tau : \bigcup_{t=1}^{T} \mathcal{Y}^t \to \mathcal{X}$ be an $\mathcal{X}$-valued $|\mathcal{Y}|$-ary tree of depth T. The contextual Shtarkov sum w.r.t. τ is defined as

$$\mathsf{Sht}(\mathcal{H} \mid \tau) = \sum_{y^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\tau(y^{t-1}))[y_t].$$

Theorem 4: Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class. Then:

$$\mathsf{reg}_T(\mathcal{H}) = \sup_{\tau} \log \mathsf{Sht}(\mathcal{H} \mid \tau).$$

- This result can be used to recover Theorem 2 using (smaller) local covers.
- It remains largely open how the contextual Shtarkov sum can be estimated for any non-trivial classes beyond covering methods...

Proof of Theorem 4

We provide only the high-level idea.

Proof of Theorem 4

We provide only the high-level idea.

Step One: Using the minimax switching trick (see **lecture 3**) to obtain the following Bayesian representation:

$$\sup_{\mathbf{x}_1, p_1} \mathbb{E}_{y_1 \sim p_1} \cdots \sup_{\mathbf{x}_T, p_T} \mathbb{E}_{y_T \sim p_T} \left[\sum_{t=1}^{T} \inf_{\hat{p}_t} \mathbb{E}_{y_t \sim p_t} \left[\ell^{\log}(\hat{p}_t, y_t) \right] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Proof of Theorem 4

We provide only the high-level idea.

Step One: Using the minimax switching trick (see **lecture 3**) to obtain the following Bayesian representation:

$$\sup_{\mathbf{x}_1, p_1} \mathbb{E}_{y_1 \sim p_1} \cdots \sup_{\mathbf{x}_T, p_T} \mathbb{E}_{y_T \sim p_T} \left[\sum_{t=1}^{T} \inf_{\hat{p}_t} \mathbb{E}_{y_t \sim p_t} \left[\ell^{\log}(\hat{p}_t, y_t) \right] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Step Two: Show that (recall from our previous slides):

$$\inf_{\hat{p}_t} \mathbb{E}_{y_t \sim p_t} \left[\ell^{\log}(\hat{p}_t, y_t) \right] = H(p_t),$$

where $H(p_t)$ is the Shannon entropy.

Proof of Theorem 4

We provide only the high-level idea.

Step One: Using the minimax switching trick (see **lecture 3**) to obtain the following Bayesian representation:

$$\sup_{\mathbf{x}_1, p_1} \mathbb{E}_{y_1 \sim p_1} \cdots \sup_{\mathbf{x}_T, p_T} \mathbb{E}_{y_T \sim p_T} \left[\sum_{t=1}^{T} \inf_{\hat{p}_t} \mathbb{E}_{y_t \sim p_t} \left[\ell^{\log}(\hat{p}_t, y_t) \right] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\mathbf{x}_t), y_t) \right].$$

Step Two: Show that (recall from our previous slides):

$$\inf_{\hat{p}_t} \mathbb{E}_{y_t \sim p_t} \left[\ell^{\log}(\hat{p}_t, y_t) \right] = H(p_t),$$

where $H(p_t)$ is the Shannon entropy.

Step Three: Show that via Skolemization the expression reduces to:

$$\sup_{\tau} \sup_{P} \mathbb{E}_{y^T \sim P} \left[H(P) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\log}(h(\tau(y^{t-1})), y_t) \right],$$

where τ runs over trees $\tau : \bigcup_{t=1}^{T} \mathcal{Y}^t \to \mathcal{X}$ and $P \in \Delta(\mathcal{Y}^T)$.

Proof of Theorem 4

Step Four: Denote $\mathbf{x}_t = \tau(y^{t-1})$, and let $P_h(y^T|\mathbf{x}^T) = \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t]$.

Proof of Theorem 4

Step Four: Denote $\mathbf{x}_t = \tau(y^{t-1})$, and let $P_h(y^T|\mathbf{x}^T) = \prod_{t=1}^T h(\mathbf{x}_t)[y_t]$. We have

$$\inf_{h\in\mathcal{H}} \sum_{t=1}^T \ell^{\log}(h(\tau(y^{t-1})), y_t) = \inf_h -\log P_h(y^T|\mathbf{x}^T) = -\sup_h \log P_h(y^T|\mathbf{x}^T).$$

Proof of Theorem 4

Step Four: Denote $\mathbf{x}_t = \tau(y^{t-1})$, and let $P_h(y^T|\mathbf{x}^T) = \prod_{t=1}^T h(\mathbf{x}_t)[y_t]$. We have

$$\inf_{h\in\mathcal{H}} \sum_{t=1}^T \ell^{\log}(h(\tau(y^{t-1})), y_t) = \inf_h -\log P_h(y^T|\mathbf{x}^T) = -\sup_h \log P_h(y^T|\mathbf{x}^T).$$

Therefore, we are reduced to

$$\sup_P \mathbb{E}_{y^T\sim P}\left[H(P) + \log\sup P_h(y^T|\mathbf{x}^T)\right] = \sup_P \mathbb{E}\left[-\log P(y^T) + \log\sup P_h(y^T|\mathbf{x}^T)\right]$$

$$= \sup_P \mathbb{E}\left[-\log P(y^T) + \log P^*(y^T|\mathbf{x}^T)\right] + \log\sum_{y^T}\sup_h \prod_{t=1}^T h(\mathbf{x}_t)[y_t]$$

$$= \underbrace{\sup_P -\mathsf{KL}(P, P^*)}_{=0} + \log\sum_{y^T}\sup_h\prod_{t=1}^T h(\mathbf{x}_t)[y_t].$$

Here, $P^*(y^T|\mathbf{x}^T) = \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum_{y^T}\sup_h P_h(y^T|\mathbf{x}^T)}$, and equality is attained at $P = P^*$.

Proof of Theorem 4

Step Four: Denote $\mathbf{x}_t = \tau(y^{t-1})$, and let $P_h(y^T|\mathbf{x}^T) = \prod_{t=1}^T h(\mathbf{x}_t)[y_t]$. We have

$$\inf_{h\in\mathcal{H}} \sum_{t=1}^T \ell^{\log}(h(\tau(y^{t-1})), y_t) = \inf_h - \log P_h(y^T|\mathbf{x}^T) = -\sup_h \log P_h(y^T|\mathbf{x}^T).$$

Therefore, we are reduced to

$$\sup_P \mathbb{E}_{y^T\sim P}\left[H(P) + \log\sup P_h(y^T|\mathbf{x}^T)\right] = \sup_P \mathbb{E}\left[-\log P(y^T) + \log\sup P_h(y^T|\mathbf{x}^T)\right]$$

$$= \sup_P \mathbb{E}\left[-\log P(y^T) + \log P^*(y^T|\mathbf{x}^T)\right] + \log\sum_{y^T}\sup_h \prod_{t=1}^T h(\mathbf{x}_t)[y_t]$$

$$= \underbrace{\sup_P -\mathsf{KL}(P, P^*)}_{=0} + \log\sum_{y^T}\sup_h \prod_{t=1}^T h(\mathbf{x}_t)[y_t].$$

Here, $P^*(y^T|\mathbf{x}^T) = \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum_{y^T}\sup_h P_h(y^T|\mathbf{x}^T)}$, and equality is attained at $P = P^*$.

Note: The distribution P^* is not a minimax optimal strategy; achieving this would require using the relaxation-based approach (c.f. **lecture 3**)...

Overview

- **Sequential Probability Assignment**
  - Weather forecasting, proper scoring, logarithmic loss
  - Bayesian algorithm
- **Minimax Regret under Log-loss**
  - Fixed design, Shtarkov sum
  - Truncated Bayesian Algorithm
  - Contextual Shtarkov sum
- **Application of Prediction with Log-loss**
  - Portfolio optimization
  - Converting prediction to investment strategy

Portfolio Optimization

Consider a (simplified) stock market that operates in discrete time steps.

Portfolio Optimization

Consider a (simplified) stock market that operates in discrete time steps.

Let $\mathcal{Y}$ be a set of assets (stocks) across which we want to allocate our investment.

Portfolio Optimization

Consider a (simplified) stock market that operates in discrete time steps.

Let $\mathcal{Y}$ be a set of assets (stocks) across which we want to allocate our investment.

At the beginning of each step t, we specify a distribution $\hat{p}_t \in \Delta(\mathcal{Y})$, such that $\hat{p}_t[y]$ determines the portion of our total wealth allocated to asset y.

Portfolio Optimization

Consider a (simplified) stock market that operates in discrete time steps.

Let $\mathcal{Y}$ be a set of assets (stocks) across which we want to allocate our investment.

At the beginning of each step t, we specify a distribution $\hat{p}_t \in \Delta(\mathcal{Y})$, such that $\hat{p}_t[y]$ determines the portion of our total wealth allocated to asset y.

Let $\mathbf{v}_t \in \mathbb{R}^{\mathcal{Y}}$ be the market vector, where $\mathbf{v}_t[y]$ represents the ratio of the market value of asset y at closing to its value at opening at step t.

Portfolio Optimization

Consider a (simplified) stock market that operates in discrete time steps.

Let $\mathcal{Y}$ be a set of assets (stocks) across which we want to allocate our investment.

At the beginning of each step t, we specify a distribution $\hat{p}_t \in \Delta(\mathcal{Y})$, such that $\hat{p}_t[y]$ determines the portion of our total wealth allocated to asset y.

Let $\mathbf{v}_t \in \mathbb{R}^{\mathcal{Y}}$ be the market vector, where $\mathbf{v}_t[y]$ represents the ratio of the market value of asset y at closing to its value at opening at step t.

Assuming the initial wealth is 1, the total wealth after T steps is given by:

$$\prod_{t=1}^{T} \left(\sum_{y \in \mathcal{Y}} \mathbf{v}_t[y] \cdot \hat{p}_t[y] \right) .$$

Portfolio Optimization

Consider a (simplified) stock market that operates in discrete time steps.

Let $\mathcal{Y}$ be a set of assets (stocks) across which we want to allocate our investment.

At the beginning of each step t, we specify a distribution $\hat{p}_t \in \Delta(\mathcal{Y})$, such that $\hat{p}_t[y]$ determines the portion of our total wealth allocated to asset y.

Let $\mathbf{v}_t \in \mathbb{R}^{\mathcal{Y}}$ be the market vector, where $\mathbf{v}_t[y]$ represents the ratio of the market value of asset y at closing to its value at opening at step t.

Assuming the initial wealth is 1, the total wealth after T steps is given by:

$$\prod_{t=1}^{T} \left(\sum_{y \in \mathcal{Y}} \mathbf{v}_t[y] \cdot \hat{p}_t[y] \right).$$

Goal: Find an investment strategy $\hat{p}^T$ that maximizes total wealth.

Investment Strategies

Let $\mathcal{X}$ be a feature space, representing all the side information we can use when specifying $\hat{p}_t$ (such as past market values).

Investment Strategies

Let $\mathcal{X}$ be a feature space, representing all the side information we can use when specifying $\hat{p}_t$ (such as past market values).

An investment strategy is a function mapping $\mathcal{X} \to \Delta(\mathcal{Y})$.

Investment Strategies

Let $\mathcal{X}$ be a feature space, representing all the side information we can use when specifying $\hat{p}_t$ (such as past market values).

An investment strategy is a function mapping $\mathcal{X} \to \Delta(\mathcal{Y})$.

Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class of investment strategies.

Investment Strategies

Let $\mathcal{X}$ be a feature space, representing all the side information we can use when specifying $\hat{p}_t$ (such as past market values).

An investment strategy is a function mapping $\mathcal{X} \to \Delta(\mathcal{Y})$.

Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class of investment strategies.

For any given investment strategy Φ, market vectors $\mathbf{v}^T$, and side information $\mathbf{x}^T$, we define its total wealth as

$$S_T(\mathbf{v}^T, \mathbf{x}^T, \Phi) = \prod_{t=1}^{T} \left(\sum_y \mathbf{v}_t[y] \cdot \Phi(\mathbf{x}_t)[y] \right).$$

Investment Strategies

Let $\mathcal{X}$ be a feature space, representing all the side information we can use when specifying $\hat{p}_t$ (such as past market values).

An investment strategy is a function mapping $\mathcal{X} \to \Delta(\mathcal{Y})$.

Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class of investment strategies.

For any given investment strategy Φ, market vectors $\mathbf{v}^T$, and side information $\mathbf{x}^T$, we define its total wealth as

$$S_T(\mathbf{v}^T, \mathbf{x}^T, \Phi) = \prod_{t=1}^{T} \left(\sum_y \mathbf{v}_t[y] \cdot \Phi(\mathbf{x}_t)[y] \right).$$

Here, we assume that $\mathbf{v}^{t-1} \subset \mathbf{x}_t$, i.e., the side information contains all the past market vectors, so that our investment strategy could rely solely on $\mathbf{x}^T$.

From Prediction to Investment

Recall that an online predictor is a function $\Phi : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{Y} \to \Delta(\mathcal{Y})$.

From Prediction to Investment

Recall that an online predictor is a function $\Phi : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{Y} \to \Delta(\mathcal{Y})$.

For any online predictor Φ, we can define the following investment strategy:

$$\Psi(\mathbf{x}_t) = \sum_{y^{t-1}} \Phi(\mathbf{x}^t, y^{t-1}) \frac{\prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}{\sum_{y^{t-1}} \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]},$$

where $\hat{p}_i := \Phi(\mathbf{x}^i, y^{i-1}) \in \Delta(\mathcal{Y})$.

From Prediction to Investment

Recall that an online predictor is a function $\Phi : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{Y} \to \Delta(\mathcal{Y})$.

For any online predictor Φ, we can define the following investment strategy:

$$\Psi(\mathbf{x}_t) = \sum_{y^{t-1}} \Phi(\mathbf{x}^t, y^{t-1}) \frac{\prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}{\sum_{y^{t-1}} \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]},$$

where $\hat{p}_i := \Phi(\mathbf{x}^i, y^{i-1}) \in \Delta(\mathcal{Y})$.

Theorem 5: Let Φ be an online predictor and Ψ be the induced investment strategy. Then, for any market vectors $\mathbf{v}^T$, side information $\mathbf{x}^T$, and hypothesis class $\mathcal{H}$, we have

$$\sup_{h \in \mathcal{H}} \log \frac{S_T(\mathbf{v}^T, \mathbf{x}^T, h)}{S_T(\mathbf{v}^T, \mathbf{x}^T, \Psi)} \leq \sup_{y^T} \sup_{h \in \mathcal{H}} \log \frac{\prod_{t=1}^T h(\mathbf{x}_t)[y_t]}{\prod_{t=1}^T \hat{p}_t[y_t]} \leq \mathrm{reg}_T(\mathcal{H}, \Phi),$$

where $\hat{p}_t := \Phi(\mathbf{x}^t, y^{t-1})$.

From Prediction to Investment

Recall that an online predictor is a function $\Phi : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{Y} \to \Delta(\mathcal{Y})$.

For any online predictor Φ, we can define the following investment strategy:

$$\Psi(\mathbf{x}_t) = \sum_{y^{t-1}} \Phi(\mathbf{x}^t, y^{t-1}) \frac{\prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}{\sum_{y^{t-1}} \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]},$$

where $\hat{p}_i := \Phi(\mathbf{x}^i, y^{i-1}) \in \Delta(\mathcal{Y})$.

Theorem 5: Let Φ be an online predictor and Ψ be the induced investment strategy. Then, for any market vectors $\mathbf{v}^T$, side information $\mathbf{x}^T$, and hypothesis class $\mathcal{H}$, we have

$$\sup_{h \in \mathcal{H}} \log \frac{S_T(\mathbf{v}^T, \mathbf{x}^T, h)}{S_T(\mathbf{v}^T, \mathbf{x}^T, \Psi)} \leq \sup_{y^T} \sup_{h \in \mathcal{H}} \log \frac{\prod_{t=1}^T h(\mathbf{x}_t)[y_t]}{\prod_{t=1}^T \hat{p}_t[y_t]} \leq \mathrm{reg}_T(\mathcal{H}, \Phi),$$

where $\hat{p}_t := \Phi(\mathbf{x}^t, y^{t-1})$.

- Any online predictor with low worst-case regret can be converted into an investment strategy that achieves a low logarithmic wealth ratio.

Proof of Theorem 5

Observe that

$$\begin{aligned} S_T(\mathbf{v}^T, \mathbf{x}^T, h) &= \prod_{t=1}^{T} \left(\sum_{y} \mathbf{v}_t[y] \cdot h(\mathbf{x}_t)[y] \right) \\ &= \sum_{y^T} \left(\prod_{t=1}^{T} \mathbf{v}_t[y_t] \right) \left(\prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \right). \end{aligned}$$

Proof of Theorem 5

Observe that

$$S_T(\mathbf{v}^T, \mathbf{x}^T, h) = \prod_{t=1}^{T} \left(\sum_{y} \mathbf{v}_t[y] \cdot h(\mathbf{x}_t)[y] \right)$$

$$= \sum_{y^T} \left(\prod_{t=1}^{T} \mathbf{v}_t[y_t] \right) \left(\prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \right).$$

Moreover, by the definition of Ψ, we have

$$S_T(\mathbf{v}^T, \mathbf{x}^T, \Psi) = \prod_{t=1}^{T} \frac{\sum_{y} \sum_{y^{t-1}} \hat{p}_t[y] \mathbf{v}_t[y] \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}{\sum_{y^{t-1}} \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}$$

$$= \prod_{t=1}^{T} \frac{\sum_{y^t} \prod_{i=1}^{t} \hat{p}_i[y_i] \prod_{i=1}^{t} \mathbf{v}_i[y_i]}{\sum_{y^{t-1}} \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}$$

$$= \sum_{y^T} \prod_{t=1}^{T} \hat{p}_t[y_t] \prod_{t=1}^{T} \mathbf{v}_t[y_t].$$

Proof of Theorem 5

Observe that

$$S_T(\mathbf{v}^T, \mathbf{x}^T, h) = \prod_{t=1}^{T} \left(\sum_{y} \mathbf{v}_t[y] \cdot h(\mathbf{x}_t)[y] \right)$$
$$= \sum_{y^T} \left(\prod_{t=1}^{T} \mathbf{v}_t[y_t] \right) \left(\prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \right).$$

Moreover, by the definition of Ψ, we have

$$S_T(\mathbf{v}^T, \mathbf{x}^T, \Psi) = \prod_{t=1}^{T} \frac{\sum_{y} \sum_{y^{t-1}} \hat{p}_t[y]\mathbf{v}_t[y] \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}{\sum_{y^{t-1}} \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}$$
$$= \prod_{t=1}^{T} \frac{\sum_{y^t} \prod_{i=1}^{t} \hat{p}_i[y_i] \prod_{i=1}^{t} \mathbf{v}_i[y_i]}{\sum_{y^{t-1}} \prod_{i=1}^{t-1} \hat{p}_i[y_i] \prod_{i=1}^{t-1} \mathbf{v}_i[y_i]}$$
$$= \sum_{y^T} \prod_{t=1}^{T} \hat{p}_t[y_t] \prod_{t=1}^{T} \mathbf{v}_t[y_t].$$

The theorem now follows from the inequality $\log \frac{\sum_i a_i}{\sum_i b_i} \le \sup_i \log \frac{a_i}{b_i}$. (Why?)

Concluding Remarks

- In this lecture, we introduced online learning under logarithmic loss.
- We provided several approaches, such as sequential covering and the Shtarkov sum, for characterizing the minimax regret under log-loss.
- We also introduced an application of prediction under log-loss in the context of portfolio optimization.
- There are also many other applications of log-loss across various domains, such as universal compression, interactive decision-making, and online distribution estimation, which we unfortunately could not cover.
  - We refer interested readers to "*Prediction, Learning, and Games*" by N. Cesa-Bianchi and G. Lugosi.