Minimax Value of Online Learning Games: Part II

Changlong Wu & Wojciech Szpankowski

Center for Science of Information
Purdue University

October 20, 2024

Overview

- **Bayesian Representation of Minimax Regret**
  - The minimax switching trick
- **Bounding the Minimax Regret: Real-valued Case**
  - The sequential Rademacher complexity, symmetrization
  - The Sequential fat-shattering dimension
  - Regret bounds via Sequential fat-shattering dimension
- **From Value to Algorithm**
  - The relaxation framework
  - The hypbrid setting, random play-out

Bayesian Representation of Minimax Regret

Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$. The minimax regret for $\mathcal{H}$ can be expressed as (c.f. Fact 1 in **lecture 2**):

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right].$$

Bayesian Representation of Minimax Regret

Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$. The minimax regret for $\mathcal{H}$ can be expressed as (c.f. Fact 1 in **lecture 2**):

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right].$$

How can we make the iterated minimax operator manageable?

Bayesian Representation of Minimax Regret

Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$. The minimax regret for $\mathcal{H}$ can be expressed as (c.f. Fact 1 in **lecture 2**):

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right].$$

How can we make the iterated minimax operator manageable?

Theorem 1: Assume the loss ℓ is bounded and $\ell(\cdot, y)$ is convex and continuous, $\hat{\mathcal{Y}}$ is convex and $\Delta(\mathcal{X} \times \mathcal{Y})$ is compact. We have:

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mu \in \Delta(\mathcal{X} \times Y)^T} \mathbb{E}_{(\mathbf{x}^T, y^T) \sim \mu} \left[\sum_{t=1}^{T} \inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\mathbb{E}_t$ denotes the conditional distribution of μ on $\mathbf{x}^t, y^{t-1}$.

Bayesian Representation of Minimax Regret

Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$. The minimax regret for $\mathcal{H}$ can be expressed as (c.f. Fact 1 in **lecture 2**):

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right].$$

How can we make the iterated minimax operator manageable?

Theorem 1: Assume the loss ℓ is bounded and $\ell(\cdot, y)$ is convex and continuous, $\hat{\mathcal{Y}}$ is convex and $\Delta(\mathcal{X} \times \mathcal{Y})$ is compact. We have:

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mu \in \Delta(\mathcal{X} \times Y)^T} \mathbb{E}_{(\mathbf{x}^T, y^T) \sim \mu} \left[\sum_{t=1}^{T} \inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\mathbb{E}_t$ denotes the conditional distribution of μ on $\mathbf{x}^t, y^{t-1}$.

- The minimax regret is reduced to finding the Bayesian optimal strategy for a single hard data distribution μ.
- One can analyze the minimax regret without needing to design an algorithm!

Preparing for Proof: The Minimax Switching Trick

Minimax Switching Trick: Let A be a convex set, B be a set such that $\Delta(B)$ is compact, and let $f : A \times B \to \mathbb{R}$ be a bounded function such that $f(\cdot, b)$ is convex for all $b \in B$. Then:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{\mu \in \Delta(B)} \inf_{a \in A} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Preparing for Proof: The Minimax Switching Trick

Minimax Switching Trick: Let A be a convex set, B be a set such that $\Delta(B)$ is compact, and let $f : A \times B \to \mathbb{R}$ be a bounded function such that $f(\cdot, b)$ is convex for all $b \in B$. Then:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{\mu \in \Delta(B)} \inf_{a \in A} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Proof: Note that:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \inf_{a \in A} \sup_{\mu \in \Delta(B)} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Preparing for Proof: The Minimax Switching Trick

Minimax Switching Trick: Let A be a convex set, B be a set such that $\Delta(B)$ is compact, and let $f : A \times B \to \mathbb{R}$ be a bounded function such that $f(\cdot, b)$ is convex for all $b \in B$. Then:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{\mu \in \Delta(B)} \inf_{a \in A} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Proof: Note that:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \inf_{a \in A} \sup_{\mu \in \Delta(B)} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Denote $F(a, \mu) = \mathbb{E}_{b \sim \mu}[f(a, b)]$.

Preparing for Proof: The Minimax Switching Trick

Minimax Switching Trick: Let A be a convex set, B be a set such that $\Delta(B)$ is compact, and let $f : A \times B \to \mathbb{R}$ be a bounded function such that $f(\cdot, b)$ is convex for all $b \in B$. Then:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{\mu \in \Delta(B)} \inf_{a \in A} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Proof: Note that:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \inf_{a \in A} \sup_{\mu \in \Delta(B)} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Denote $F(a, \mu) = \mathbb{E}_{b \sim \mu}[f(a, b)]$. We have $F(\cdot, \mu)$ is convex over A, and $F(a, \cdot)$ is linear (therefore concave) over $\Delta(B)$. (Verify this!)

Preparing for Proof: The Minimax Switching Trick

Minimax Switching Trick: Let A be a convex set, B be a set such that $\Delta(B)$ is compact, and let $f : A \times B \to \mathbb{R}$ be a bounded function such that $f(\cdot, b)$ is convex for all $b \in B$. Then:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{\mu \in \Delta(B)} \inf_{a \in A} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Proof: Note that:

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \inf_{a \in A} \sup_{\mu \in \Delta(B)} \mathbb{E}_{b \sim \mu}[f(a, b)].$$

Denote $F(a, \mu) = \mathbb{E}_{b \sim \mu}[f(a, b)]$. We have $F(\cdot, \mu)$ is convex over A, and $F(a, \cdot)$ is linear (therefore concave) over $\Delta(B)$. (Verify this!)

By the Minimax Theorem (c.f. **Lecture 2**), we conclude:

$$\inf_{a \in A} \sup_{\mu \in \Delta(B)} F(a, \mu) = \sup_{\mu \in \Delta(B)} \inf_{a \in A} F(a, \mu).$$

Proof of Theorem 1

Observe that the iterated minimax formulation can be written as:

$$\sup_{\mathbf{z}_0} \inf_{\hat{y}_1} \sup_{\mathbf{z}_1} \cdots \inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\mathbf{z}_0 = \mathbf{x}_1$, $\mathbf{z}_t = (y_t, \mathbf{x}_{t+1})$ for $t < T$ and $\mathbf{z}_T = y_T$.

Proof of Theorem 1

Observe that the iterated minimax formulation can be written as:

$$\sup_{\mathbf{z}_0} \inf_{\hat{y}_1} \sup_{\mathbf{z}_1} \cdots \inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\mathbf{z}_0 = \mathbf{x}_1$, $\mathbf{z}_t = (y_t, \mathbf{x}_{t+1})$ for $t < T$ and $\mathbf{z}_T = y_T$. Consider the last layer:

$$\inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \underbrace{\inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)}_{F(\mathbf{z}^T)} \right]$$

$$= \sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T) \right].$$

Proof of Theorem 1

Observe that the iterated minimax formulation can be written as:

$$\sup_{\mathbf{z}_0} \inf_{\hat{y}_1} \sup_{\mathbf{z}_1} \cdots \inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\mathbf{z}_0 = \mathbf{x}_1$, $\mathbf{z}_t = (y_t, \mathbf{x}_{t+1})$ for $t < T$ and $\mathbf{z}_T = y_T$. Consider the last layer:

$$\inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \underbrace{\inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)}_{F(\mathbf{z}^T)} \right]$$
$$= \sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T) \right].$$

We now bound the second term. By the Minimax Switching Trick, we have:

$$\inf_{\hat{y}_T} \sup_{\mathbf{z}_T} \left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T) \right] = \sup_{\mu_T \in \Delta(\mathcal{X} \times \mathcal{Y})} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T \sim \mu_T} \left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T) \right].$$

Proof of Theorem 1

Moreover, observe that

$$\begin{aligned}
&\sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T \sim \mu_T}\left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T)\right] \\
&\qquad = \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T}\left[\mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T, \mathbf{z}_T)] - \mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)]\right]
\end{aligned}$$

Proof of Theorem 1

Moreover, observe that

$$
\begin{aligned}
\sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T \sim \mu_T} \left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T) \right] \\
&= \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T} \left[\mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T, \mathbf{z}_T)] - \mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)] \right] \\
&= \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \left[\inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T, \mathbf{z}_T)] - \mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)] \right]
\end{aligned}
$$

Proof of Theorem 1

Moreover, observe that

$$
\begin{aligned}
&\sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T \sim \mu_T} \left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T) \right] \\
&\qquad = \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T} \left[\mathbb{E}_{\mathbf{z}_T} [\ell(\hat{y}_T, \mathbf{z}_T)] - \mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)] \right] \\
&\qquad = \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \left[\inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T} [\ell(\hat{y}_T, \mathbf{z}_T)] - \mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)] \right] \\
&\qquad = \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{\mathbf{z}_T} \left[\inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T} [\ell(\hat{y}_T, \mathbf{z}_T)] - F(\mathbf{z}^T) \right].
\end{aligned}
$$

Proof of Theorem 1

Moreover, observe that

$$\begin{aligned}
&\sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T\sim\mu_T}\left[\ell(\hat{y}_T, \mathbf{z}_T) - F(\mathbf{z}^T)\right] \\
&\qquad = \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \inf_{\hat{y}_T} \left[\mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T, \mathbf{z}_T)] - \mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)]\right] \\
&\qquad = \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \left[\inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T, \mathbf{z}_T)] - \mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)]\right] \\
&\qquad = \sup_{\mu_T \in \Delta(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{\mathbf{z}_T}\left[\inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T, \mathbf{z}_T)] - F(\mathbf{z}^T)\right].
\end{aligned}$$

Applying this argument for another $T-1$ steps, we obtain:

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mu_1} \mathbb{E}_{\mathbf{z}_1\sim\mu_1} \cdots \sup_{\mu_T} \mathbb{E}_{\mathbf{z}_T\sim\mu_T}\left[\sum_{t=1}^{T} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_t}[\ell(\hat{y}_t, \mathbf{z}_t)] - F(\mathbf{z}^T)\right].$$

Proof of Theorem 1

Note that

$$\sup_{\mu_1} \mathbb{E}_{\mathbf{z}_1 \sim \mu_1} \cdots \sup_{\mu_T} \mathbb{E}_{\mathbf{z}_T \sim \mu_T} \overset{(\star)}{\equiv} \sup_{\mu \in \Delta((\mathcal{X} \times \mathcal{Y})^T)} \mathbb{E}_{\mathbf{z}^T \sim \mu},$$

where μ is a joint distribution over $(\mathcal{X} \times \mathcal{Y})^T$.

Proof of Theorem 1

Note that

$$\sup_{\mu_1} \mathbb{E}_{\mathbf{z}_1 \sim \mu_1} \cdots \sup_{\mu_T} \mathbb{E}_{\mathbf{z}_T \sim \mu_T} \overset{(\star)}{\equiv} \sup_{\mu \in \Delta((\mathcal{X} \times \mathcal{Y})^T)} \mathbb{E}_{\mathbf{z}^T \sim \mu},$$

where μ is a joint distribution over $(\mathcal{X} \times \mathcal{Y})^T$. We conclude:

$$\mathrm{reg}_T(\mathcal{H}) = \sup_{\mu \in \Delta((\mathcal{X} \times \mathcal{Y})^T)} \mathbb{E}_{\mathbf{z}^T \sim \mu} \left[\sum_{t=1}^{T} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_t} [\ell(\hat{y}_t, \mathbf{z}_t)] - F(\mathbf{z}^T) \right].$$

Proof of Theorem 1

Note that

$$\sup_{\mu_1} \mathbb{E}_{\mathbf{z}_1 \sim \mu_1} \cdots \sup_{\mu_T} \mathbb{E}_{\mathbf{z}_T \sim \mu_T} \overset{(\bigstar)}{\equiv} \sup_{\mu \in \Delta((\mathcal{X} \times \mathcal{Y})^T)} \mathbb{E}_{\mathbf{z}^T \sim \mu},$$

where μ is a joint distribution over $(\mathcal{X} \times \mathcal{Y})^T$. We conclude:

$$\mathsf{reg}_T(\mathcal{H}) = \sup_{\mu \in \Delta((\mathcal{X} \times \mathcal{Y})^T)} \mathbb{E}_{\mathbf{z}^T \sim \mu} \left[\sum_{t=1}^{T} \inf_{\hat{y}_T} \mathbb{E}_{\mathbf{z}_t} [\ell(\hat{y}_t, \mathbf{z}_t)] - F(\mathbf{z}^T) \right].$$

Homework: Prove that for any function $F : A \times B \to \mathbb{R}$ and any distribution μ over A, we have

$$\mathbb{E}_{a \sim \mu} \sup_{b \in B} F(a, b) = \sup_{g \in B^A} \mathbb{E}_{a \sim \mu} F(a, g(a)).$$

Consequently, $(\bigstar)$ holds. (**Hint**: Use the same argument as in Skolemization and switch the $\sup_{\mu_t}$ operators.)

Overview

- **Bayesian Representation of Minimax Regret**
  - The minimax switching trick
- **Bounding the Minimax Regret: Real-valued Case**
  - The sequential Rademacher complexity, symmetrization
  - The Sequential fat-shattering dimension
  - Regret bounds via Sequential fat-shattering dimension
- **From Value to Algorithm**
  - The relaxation framework
  - The hybrid setting, random play-out

The Sequential Rademacher Complexity

Sequential Rademacher Complexity: For any real-valued class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, we define the *sequential Rademacher complexity* of $\mathcal{H}$ as

$$\mathsf{sRad}_T(\mathcal{H}) = \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\tau(\epsilon^{t-1})) \right],$$

where $\tau : \bigcup_{i \leq T} \{0,1\}^i \to \mathcal{X}$ runs over all $\mathcal{X}$-valued binary trees of depth T, and ϵ^T is sampled uniformly over $\{-1,+1\}^T$.

The Sequential Rademacher Complexity

Sequential Rademacher Complexity: For any real-valued class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, we define the *sequential Rademacher complexity* of $\mathcal{H}$ as

$$\mathsf{sRad}_T(\mathcal{H}) = \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\tau(\epsilon^{t-1})) \right],$$

where $\tau : \bigcup_{i \leq T} \{0,1\}^i \to \mathcal{X}$ runs over all $\mathcal{X}$-valued binary trees of depth T, and ϵ^T is sampled uniformly over $\{-1,+1\}^T$.

- Similar to classical Rademacher complexity, except that the optimizing is over trees instead of sequences.

The Sequential Rademacher Complexity

Sequential Rademacher Complexity: For any real-valued class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, we define the *sequential Rademacher complexity* of $\mathcal{H}$ as

$$\mathrm{sRad}_T(\mathcal{H}) = \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\tau(\epsilon^{t-1})) \right],$$

where $\tau : \bigcup_{i \le T} \{0,1\}^i \to \mathcal{X}$ runs over all $\mathcal{X}$-valued binary trees of depth T, and ϵ^T is sampled uniformly over $\{-1,+1\}^T$.

- Similar to classical Rademacher complexity, except that the optimizing is over trees instead of sequences.

Example 1: Let $\mathcal{H}^{\mathrm{lin}} := \{h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in B_2\}$ be the class of linear functions with weight $\mathbf{w}$ lie in a unit L_2 ball. Let $\mathcal{X} := B_2$ as well, we have

$$\mathrm{sRad}_T(\mathcal{H}^{\mathrm{lin}}) \le \sqrt{T}.$$

Proof of Example 1

Fix any tree τ and denote $\mathbf{x}_t := \tau(\epsilon^{t-1})$, we have:

$$\mathsf{sRad}_T(\mathcal{H}^{\mathsf{lin}}) = \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \sum_{t=1}^{T} \epsilon_t \langle \mathbf{w}, \mathbf{x}_t) \rangle \right]$$

Proof of Example 1

Fix any tree τ and denote $\mathbf{x}_t := \tau(\epsilon^{t-1})$, we have:

$$
\begin{aligned}
\mathsf{sRad}_T(\mathcal{H}^{\mathsf{lin}}) &= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \sum_{t=1}^{T} \epsilon_t \langle \mathbf{w}, \mathbf{x}_t) \rangle \right] \\
&= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \left\langle \mathbf{w}, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t) \right\rangle \right]
\end{aligned}
$$

Proof of Example 1

Fix any tree τ and denote $\mathbf{x}_t := \tau(\epsilon^{t-1})$, we have:

$$
\begin{aligned}
\mathsf{sRad}_T(\mathcal{H}^{\mathsf{lin}}) &= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \sum_{t=1}^{T} \epsilon_t \langle \mathbf{w}, \mathbf{x}_t) \rangle \right] \\
&= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \left\langle \mathbf{w}, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t) \right\rangle \right] \\
&\leq \sup_{\tau} \mathbb{E}_{\epsilon^T} \sqrt{\left\langle \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle}, \quad (\text{Why?})
\end{aligned}
$$

Proof of Example 1

Fix any tree τ and denote $\mathbf{x}_t := \tau(\epsilon^{t-1})$, we have:

$$
\begin{aligned}
\mathsf{sRad}_T(\mathcal{H}^{\text{lin}}) &= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \sum_{t=1}^{T} \epsilon_t \langle \mathbf{w}, \mathbf{x}_t) \rangle \right] \\
&= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \left\langle \mathbf{w}, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t) \right\rangle \right] \\
&\leq \sup_{\tau} \mathbb{E}_{\epsilon^T} \sqrt{\left\langle \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle}, \quad \text{(Why?)} \\
&\leq \sup_{\tau} \sqrt{\mathbb{E}_{\epsilon^T} \left\langle \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle}, \quad \text{by Jensen's inequality}
\end{aligned}
$$

Proof of Example 1

Fix any tree τ and denote $\mathbf{x}_t := \tau(\epsilon^{t-1})$, we have:

$$\begin{aligned}
\mathrm{sRad}_T(\mathcal{H}^{\mathrm{lin}}) &= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \sum_{t=1}^{T} \epsilon_t \langle \mathbf{w}, \mathbf{x}_t \rangle \right] \\
&= \sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{\mathbf{w} \in B_2} \left\langle \mathbf{w}, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle \right] \\
&\le \sup_{\tau} \mathbb{E}_{\epsilon^T} \sqrt{\left\langle \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle}, \quad \text{(Why?)} \\
&\le \sup_{\tau} \sqrt{\mathbb{E}_{\epsilon^T} \left\langle \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle}, \quad \text{by Jensen's inequality} \\
&\le \sup_{\tau} \sqrt{\mathbb{E}_{\epsilon^T} \left[T + \sum_{i \neq j \le T} \epsilon_i \epsilon_j \mathbf{x}_i^T \mathbf{x}_j \right]}, \quad \text{by } \|\mathbf{x}_t\|_2 \le 1
\end{aligned}$$

Proof of Example 1

Fix any tree τ and denote $\mathbf{x}_t := \tau(\epsilon^{t-1})$, we have:

$$
\begin{aligned}
\mathrm{sRad}_T(\mathcal{H}^{\mathrm{lin}}) &= \sup_\tau \mathbb{E}_{\epsilon^T}\left[\sup_{\mathbf{w}\in B_2} \sum_{t=1}^T \epsilon_t \langle \mathbf{w}, \mathbf{x}_t\rangle\right] \\
&= \sup_\tau \mathbb{E}_{\epsilon^T}\left[\sup_{\mathbf{w}\in B_2} \left\langle \mathbf{w}, \sum_{t=1}^T \epsilon_t \mathbf{x}_t \right\rangle\right] \\
&\le \sup_\tau \mathbb{E}_{\epsilon^T} \sqrt{\left\langle \sum_{t=1}^T \epsilon_t \mathbf{x}_t, \sum_{t=1}^T \epsilon_t \mathbf{x}_t \right\rangle}, \quad \text{(Why?)} \\
&\le \sup_\tau \sqrt{\mathbb{E}_{\epsilon^T} \left\langle \sum_{t=1}^T \epsilon_t \mathbf{x}_t, \sum_{t=1}^T \epsilon_t \mathbf{x}_t \right\rangle}, \quad \text{by Jensen's inequality} \\
&\le \sup_\tau \sqrt{\mathbb{E}_{\epsilon^T}\left[T + \sum_{i\neq j \le T} \epsilon_i \epsilon_j \mathbf{x}_i^T \mathbf{x}_j \right]}, \quad \text{by } \|\mathbf{x}_t\|_2 \le 1 \\
&= \sqrt{T}. \quad \text{(Why?)}
\end{aligned}
$$

Reduction to Sequential Rademacher Complexity: Symmetrization

We now introduce a general approach for reducing the minimax regret to sequential Rademacher complexity.

Reduction to Sequential Rademacher Complexity: Symmetrization

We now introduce a general approach for reducing the minimax regret to sequential Rademacher complexity.

From Theorem 1, we know that the minimax regret can be expressed as

$$\begin{aligned}
&\sup_{\mu} \mathbb{E}\left[\sum_{t=1}^{T} \inf_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right] \\
&= \sup_{\mu} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left\{\sum_{t=1}^{T} \inf_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right\}\right] \\
&\leq \sup_{\mu} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left\{\sum_{t=1}^{T} \mathbb{E}_t[\ell(h(\mathbf{x}_t), y_t)] - \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right\}\right].
\end{aligned}$$

Reduction to Sequential Rademacher Complexity: Symmetrization

We now introduce a general approach for reducing the minimax regret to sequential Rademacher complexity.

From Theorem 1, we know that the minimax regret can be expressed as

$$\sup_{\mu} \mathbb{E}\left[\sum_{t=1}^{T} \inf_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right]$$
$$= \sup_{\mu} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left\{\sum_{t=1}^{T} \inf_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right\}\right]$$
$$\leq \sup_{\mu} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left\{\sum_{t=1}^{T} \mathbb{E}_t[\ell(h(\mathbf{x}_t), y_t)] - \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right\}\right].$$

Denote $h^{\ell}(\mathbf{z}_t) := \ell(h(\mathbf{x}_t), y_t)$ where $\mathbf{z}_t = (\mathbf{x}_t, y_t)$. We obtain upper bound

$$\sup_{\mu} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left\{\sum_{t=1}^{T} \mathbb{E}_t[h^{\ell}(\mathbf{z}_t)] - h^{\ell}(\mathbf{z}_t)\right\}\right].$$

Reduction to Sequential Rademacher Complexity: Symmetrization

We now introduce a tangent sequence $\mathbf{z}'_1, \cdots, \mathbf{z}'_T$ such that $\mathbf{z}'_t = (\mathbf{x}'_t, y'_t)$ with $\mathbf{x}'_t = \mathbf{x}_t$ and y'_t being an *i.i.d.* copy of y_t conditioning on $\mathbf{x}^t, y^{t-1}$.

Reduction to Sequential Rademacher Complexity: Symmetrization

We now introduce a tangent sequence $\mathbf{z}'_1, \cdots, \mathbf{z}'_T$ such that $\mathbf{z}'_t = (\mathbf{x}'_t, y'_t)$ with $\mathbf{x}'_t = \mathbf{x}_t$ and y'_t being an *i.i.d.* copy of y_t conditioning on $\mathbf{x}^t, y^{t-1}$.

The upper bound can be expresses as

$$\sup_{\mu} \mathbb{E}_{\mathbf{z}^T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \mathbb{E}_t[h^\ell(\mathbf{z}'_t)] - h^\ell(\mathbf{z}_t) \right\} \right], \text{ by definition of } \mathbf{z}'^T$$

$$\leq \sup_{\mu} \mathbb{E}_{\mathbf{z}^T} \mathbb{E}_{\mathbf{z}'^T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} h^\ell(\mathbf{z}'_t) - h^\ell(\mathbf{z}_t) \right\} \right], \text{ by } \sup \mathbb{E} \leq \mathbb{E} \sup$$

$$\stackrel{(\star)}{=} \sup_{\mu} \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{y_1, y'_1} \mathbb{E}_{\epsilon_1} \cdots \mathbb{E}_{\mathbf{x}_T} \mathbb{E}_{y_T, y'_T} \mathbb{E}_{\epsilon_T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \epsilon_t (h^\ell(\mathbf{z}'_t) - h^\ell(\mathbf{z}_t)) \right\} \right]$$

$$\stackrel{(\star\star)}{\leq} 2 \sup_{\mu} \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{y_1} \mathbb{E}_{\epsilon_1} \cdots \mathbb{E}_{\mathbf{x}_T} \mathbb{E}_{y_T} \mathbb{E}_{\epsilon_T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \epsilon_t h^\ell(\mathbf{z}_t) \right\} \right]$$

where ϵ_t is uniform over $\{\pm 1\}$ and is (conditional) independent of y_t, y'_t.

Reduction to Sequential Rademacher Complexity: Symmetrization

We now introduce a tangent sequence $\mathbf{z}'_1, \cdots, \mathbf{z}'_T$ such that $\mathbf{z}'_t = (\mathbf{x}'_t, y'_t)$ with $\mathbf{x}'_t = \mathbf{x}_t$ and y'_t being an *i.i.d.* copy of y_t conditioning on $\mathbf{x}^t, y^{t-1}$.

The upper bound can be expresses as

$$
\begin{aligned}
&\sup_{\mu} \mathbb{E}_{\mathbf{z}^T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \mathbb{E}_t[h^\ell(\mathbf{z}'_t)] - h^\ell(\mathbf{z}_t) \right\} \right], \text{ by definition of } \mathbf{z}'^T \\
&\quad \leq \sup_{\mu} \mathbb{E}_{\mathbf{z}^T} \mathbb{E}_{\mathbf{z}'^T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} h^\ell(\mathbf{z}'_t) - h^\ell(\mathbf{z}_t) \right\} \right], \text{ by } \sup \mathbb{E} \leq \mathbb{E} \sup \\
&\quad \stackrel{(\bigstar)}{=} \sup_{\mu} \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{y_1, y'_1} \mathbb{E}_{\epsilon_1} \cdots \mathbb{E}_{\mathbf{x}_T} \mathbb{E}_{y_T, y'_T} \mathbb{E}_{\epsilon_T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \epsilon_t (h^\ell(\mathbf{z}'_t) - h^\ell(\mathbf{z}_t)) \right\} \right] \\
&\quad \stackrel{(\star\star)}{\leq} 2 \sup_{\mu} \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{y_1} \mathbb{E}_{\epsilon_1} \cdots \mathbb{E}_{\mathbf{x}_T} \mathbb{E}_{y_T} \mathbb{E}_{\epsilon_T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \epsilon_t h^\ell(\mathbf{z}_t) \right\} \right]
\end{aligned}
$$

where ϵ_t is uniform over $\{\pm 1\}$ and is (conditional) independent of y_t, y'_t.

Here $(\bigstar)$ follows by the conditional symmetries of y_t, y'_t and $(\star\star)$ follows by $\sup(A+B) \leq \sup A + \sup B$ and symmetries between y_t, y'_t.

Reduction to Sequential Rademacher Complexity: Symmetrization

Note that, the following operator inequality holds (by $\mathbb{E} \le \sup$):

$$\mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{y_1} \mathbb{E}_{\epsilon_1} \cdots \mathbb{E}_{\mathbf{x}_T} \mathbb{E}_{y_T} \mathbb{E}_{\epsilon_T} \le \sup_{\mathbf{x}_1, y_1} \mathbb{E}_{\epsilon_1} \cdots \sup_{\mathbf{x}_T, y_T} \mathbb{E}_{\epsilon_T}.$$

Reduction to Sequential Rademacher Complexity: Symmetrization

Note that, the following operator inequality holds (by $\mathbb{E} \leq \sup$):

$$\mathbb{E}_{\mathbf{x}_1}\mathbb{E}_{y_1}\mathbb{E}_{\epsilon_1}\cdots\mathbb{E}_{\mathbf{x}_T}\mathbb{E}_{y_T}\mathbb{E}_{\epsilon_T} \leq \sup_{\mathbf{x}_1,y_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\mathbf{x}_T,y_T}\mathbb{E}_{\epsilon_T}.$$

By Skolemization again, the upper bound equals

$$\sup_{\mathbf{x}_1,y_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\mathbf{x}_T,y_T}\mathbb{E}_{\epsilon_T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^{T}\epsilon_t h^{\ell}(\mathbf{z}_t)\right\}\right] = \underbrace{\sup_{\tau}\mathbb{E}_{\epsilon^T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^{T}\epsilon_t h^{\ell}(\tau(\epsilon^{t-1}))\right\}\right]}_{\mathrm{sRad}(\mathcal{H}^{\ell})},$$

where τ runs over all $(\mathcal{X}\times\mathcal{Y})$-valued binary trees.

Reduction to Sequential Rademacher Complexity: Symmetrization

Note that, the following operator inequality holds (by $\mathbb{E} \le \sup$):

$$\mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{y_1} \mathbb{E}_{\epsilon_1} \cdots \mathbb{E}_{\mathbf{x}_T} \mathbb{E}_{y_T} \mathbb{E}_{\epsilon_T} \le \sup_{\mathbf{x}_1, y_1} \mathbb{E}_{\epsilon_1} \cdots \sup_{\mathbf{x}_T, y_T} \mathbb{E}_{\epsilon_T}.$$

By Skolemization again, the upper bound equals

$$\sup_{\mathbf{x}_1, y_1} \mathbb{E}_{\epsilon_1} \cdots \sup_{\mathbf{x}_T, y_T} \mathbb{E}_{\epsilon_T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \epsilon_t h^{\ell}(\mathbf{z}_t) \right\} \right] = \underbrace{\sup_{\tau} \mathbb{E}_{\epsilon^T} \left[\sup_{h \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \epsilon_t h^{\ell}(\tau(\epsilon^{t-1})) \right\} \right]}_{\mathsf{sRad}(\mathcal{H}^{\ell})},$$

where τ runs over all $(\mathcal{X} \times \mathcal{Y})$-valued binary trees.

Lemma 1: Putting everything together, we have proved that

$$\mathsf{reg}_T(\mathcal{H}) \le 2 \cdot \mathsf{sRad}_T(\mathcal{H}^{\ell}),$$

where $\mathcal{H}^{\ell} := \{\ell(h(\mathbf{x}), y) : h \in \mathcal{H}\} \in \hat{\mathcal{Y}}^{(\mathcal{X} \times \mathcal{Y})}$.

The Lipschitz Contraction Lemma

Lemma 2: Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{Z}}$ and $\phi : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$. If for all $\mathbf{z} \in \mathcal{Z}$, $\phi(\cdot, \mathbf{z})$ is a L-Lipschitz function, then

$$\mathrm{sRad}_T(\phi(\mathcal{H})) \leq L \cdot \mathrm{sRad}_T(\mathcal{H}),$$

where $\phi(\mathcal{H}) = \{\mathbf{z} \to \phi(h(\mathbf{z}), z) : h \in \mathcal{H}\}$.

The Lipschitz Contraction Lemma

Lemma 2: Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{Z}}$ and $\phi : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$. If for all $\mathbf{z} \in \mathcal{Z}$, $\phi(\cdot, \mathbf{z})$ is a L-Lipschitz function, then

$$\mathrm{sRad}_T(\phi(\mathcal{H})) \leq L \cdot \mathrm{sRad}_T(\mathcal{H}),$$

where $\phi(\mathcal{H}) = \{\mathbf{z} \to \phi(h(\mathbf{z}), z) : h \in \mathcal{H}\}$.

- This lemma mirrors Talagrand's contraction lemma for regular Rademacher complexity.
- Apply this lemma to $\mathcal{H}^{\ell} := \{\ell(h(\mathbf{x}), y) : h \in \mathcal{H}\}$ for Lipschitz loss ℓ, we have
$$\mathrm{sRad}_T(\mathcal{H}^{\ell}) \leq O(\mathrm{sRad}_T(\mathcal{H})).$$
- Therefore, by Lemma 1, we have
$$\mathrm{reg}_T(\mathcal{H}) \leq O(\mathrm{sRad}_T(\mathcal{H})).$$

Proof of Lemma 2

Fix any tree τ and denote $\mathbf{z}_t = \tau(\epsilon^{t-1})$. Let $S_t^h = \sum_{i=1}^t \epsilon_i \phi(h(\mathbf{z}_i), z_i)$.

$$
\begin{aligned}
&\mathbb{E}_{\epsilon^T}\left[\sup_h S_T^h\right] \\
&= \mathbb{E}_{\epsilon^{T-1}}\left[\frac{1}{2}\left\{\sup_h\{S_{T-1}^h + \phi(h(\mathbf{z}_T), \mathbf{z}_T)\} + \sup_h\{S_{T-1}^h - \phi(h(\mathbf{z}_T), \mathbf{z}_T)\}\right\}\right] \\
&= \mathbb{E}_{\epsilon^{T-1}}\left[\frac{1}{2}\sup_{h,h'}\left\{S_{T-1}^h + S_{T-1}^{h'} + \phi(h(\mathbf{z}_T), \mathbf{z}_T) - \phi(h'(\mathbf{z}_T), z_T)\right\}\right] \\
&\overset{(\star)}{\leq} \mathbb{E}_{\epsilon^{T-1}}\left[\frac{1}{2}\sup_{h,h'}\left\{S_{T-1}^h + S_{T-1}^{h'} + L|h(\mathbf{z}_T) - h'(\mathbf{z}_T)|\right\}\right], \text{ by Lipschitz of } \phi \\
&\overset{(\star\star)}{=} \mathbb{E}_{\epsilon^{T-1}}\left[\frac{1}{2}\sup_{h,h'}\left\{S_{T-1}^h + S_{T-1}^{h'} + L(h(\mathbf{z}_T) - h'(\mathbf{z}_T))\right\}\right], \text{ by symmetries} \\
&= \mathbb{E}_{\epsilon^T}\left[\sup_h S_{T-1}^h + L\epsilon_T h(\mathbf{z}_T)\right], \text{ by reversing of step one}
\end{aligned}
$$

Continue the same argument for another $T-1$ steps, the lemma follows.

Bounding the Minimax Regret via Sequential Rademacher Complexity

Theorem 2: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be a real-valued class. If the loss function ℓ is bounded, convex, and Lipschitz in its first argument, then:

$$\mathrm{reg}_T(\mathcal{H}) \leq O(\mathrm{sRad}_T(\mathcal{H})).$$

Moreover, for the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$, we have

$$\mathrm{reg}_T(\mathcal{H}) \geq \Omega(\mathrm{sRad}_T(\mathcal{H})).$$

Bounding the Minimax Regret via Sequential Rademacher Complexity

Theorem 2: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be a real-valued class. If the loss function ℓ is bounded, convex, and Lipschitz in its first argument, then:

$$\mathsf{reg}_T(\mathcal{H}) \leq O(\mathsf{sRad}_T(\mathcal{H})).$$

Moreover, for the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$, we have

$$\mathsf{reg}_T(\mathcal{H}) \geq \Omega(\mathsf{sRad}_T(\mathcal{H})).$$

- The upper bound follows by our previous discussions (c.f. Lemma 1 and 2).

Bounding the Minimax Regret via Sequential Rademacher Complexity

Theorem 2: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be a real-valued class. If the loss function ℓ is bounded, convex, and Lipschitz in its first argument, then:

$$\mathsf{reg}_T(\mathcal{H}) \leq O(\mathsf{sRad}_T(\mathcal{H})).$$

Moreover, for the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$, we have

$$\mathsf{reg}_T(\mathcal{H}) \geq \Omega(\mathsf{sRad}_T(\mathcal{H})).$$

- The upper bound follows by our previous discussions (c.f. Lemma 1 and 2).
- The lower bound follows by constructing a specific hard data distribution, which we prove below.

Bounding the Minimax Regret via Sequential Rademacher Complexity

Theorem 2: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be a real-valued class. If the loss function ℓ is bounded, convex, and Lipschitz in its first argument, then:

$$\mathsf{reg}_T(\mathcal{H}) \leq O(\mathsf{sRad}_T(\mathcal{H})).$$

Moreover, for the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$, we have

$$\mathsf{reg}_T(\mathcal{H}) \geq \Omega(\mathsf{sRad}_T(\mathcal{H})).$$

- The upper bound follows by our previous discussions (c.f. Lemma 1 and 2).
- The lower bound follows by constructing a specific hard data distribution, which we prove below.
- For linear functions, we have $\mathsf{reg}_T(\mathcal{H}^{\mathsf{lin}}) \leq O(\sqrt{T})$.

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \leq T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \le T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0,1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \le T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0,1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Note that $\inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[|\hat{y}_t - y_t|] = \frac{1}{2}$, since y_t is uniform over $\{0,1\}$ conditioning on $\mathbf{x}^t, y^{t-1}$. That is the Bayesian optimal risk equals $\frac{T}{2}$.

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \leq T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0,1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Note that $\inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[|\hat{y}_t - y_t|] = \frac{1}{2}$, since y_t is uniform over $\{0,1\}$ conditioning on $\mathbf{x}^t, y^{t-1}$. That is the Bayesian optimal risk equals $\frac{T}{2}$.

Moreover, $|h(\mathbf{x}_t) - y_t| = \epsilon_t h(\mathbf{x}_t) + (1 - \epsilon_t)/2$, where $\epsilon_t = 1 - 2y_t \in \{-1, +1\}$.

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \le T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0,1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Note that $\inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[|\hat{y}_t - y_t|] = \frac{1}{2}$, since y_t is uniform over $\{0,1\}$ conditioning on $\mathbf{x}^t, y^{t-1}$. That is the Bayesian optimal risk equals $\frac{T}{2}$.

Moreover, $|h(\mathbf{x}_t) - y_t| = \epsilon_t h(\mathbf{x}_t) + (1-\epsilon_t)/2$, where $\epsilon_t = 1 - 2y_t \in \{-1,+1\}$.

Therefore, by Theorem 1, we have

$$\mathrm{reg}_T(\mathcal{H}) \ge \mathbb{E}_{y^T}\left[\frac{T}{2} - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \left(\epsilon_t h(\mathbf{x}_t) + \frac{1-\epsilon_t}{2}\right)\right] = \mathbb{E}_{\epsilon^T}\left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\mathbf{x}_t)\right],$$

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \le T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0,1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Note that $\inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[|\hat{y}_t - y_t|] = \frac{1}{2}$, since y_t is uniform over $\{0,1\}$ conditioning on $\mathbf{x}^t, y^{t-1}$. That is the Bayesian optimal risk equals $\frac{T}{2}$.

Moreover, $|h(\mathbf{x}_t) - y_t| = \epsilon_t h(\mathbf{x}_t) + (1-\epsilon_t)/2$, where $\epsilon_t = 1 - 2y_t \in \{-1,+1\}$.

Therefore, by Theorem 1, we have

$$\mathrm{reg}_T(\mathcal{H}) \ge \mathbb{E}_{y^T}\left[\frac{T}{2} - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \left(\epsilon_t h(\mathbf{x}_t) + \frac{1-\epsilon_t}{2}\right)\right] = \mathbb{E}_{\epsilon^T}\left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\mathbf{x}_t)\right],$$

where the equality follows by $\mathbb{E}_{y_t}[(1-\epsilon_t)/2] = \frac{1}{2}$ and changing measure to ϵ^T.

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \le T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0,1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Note that $\inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[|\hat{y}_t - y_t|] = \frac{1}{2}$, since y_t is uniform over $\{0,1\}$ conditioning on $\mathbf{x}^t, y^{t-1}$. That is the Bayesian optimal risk equals $\frac{T}{2}$.

Moreover, $|h(\mathbf{x}_t) - y_t| = \epsilon_t h(\mathbf{x}_t) + (1 - \epsilon_t)/2$, where $\epsilon_t = 1 - 2y_t \in \{-1, +1\}$.

Therefore, by Theorem 1, we have

$$\mathrm{reg}_T(\mathcal{H}) \ge \mathbb{E}_{y^T}\left[\frac{T}{2} - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \left(\epsilon_t h(\mathbf{x}_t) + \frac{1 - \epsilon_t}{2}\right)\right] = \mathbb{E}_{\epsilon^T}\left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\mathbf{x}_t)\right],$$

where the equality follows by $\mathbb{E}_{y_t}[(1 - \epsilon_t)/2] = \frac{1}{2}$ and changing measure to ϵ^T.

Since τ is selected arbitrary, the inequality remain holds when taking $\sup_\tau$.

Proof of Regret Lower Bound

Let $\tau : \bigcup_{i \le T} \{0,1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T.

We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0,1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Note that $\inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[|\hat{y}_t - y_t|] = \frac{1}{2}$, since y_t is uniform over $\{0,1\}$ conditioning on $\mathbf{x}^t, y^{t-1}$. That is the Bayesian optimal risk equals $\frac{T}{2}$.

Moreover, $|h(\mathbf{x}_t) - y_t| = \epsilon_t h(\mathbf{x}_t) + (1-\epsilon_t)/2$, where $\epsilon_t = 1 - 2y_t \in \{-1,+1\}$.

Therefore, by Theorem 1, we have

$$\mathrm{reg}_T(\mathcal{H}) \ge \mathbb{E}_{y^T}\left[\frac{T}{2} - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T}\left(\epsilon_t h(\mathbf{x}_t) + \frac{1-\epsilon_t}{2}\right)\right] = \mathbb{E}_{\epsilon^T}\left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\mathbf{x}_t)\right],$$

where the equality follows by $\mathbb{E}_{y_t}[(1-\epsilon_t)/2] = \frac{1}{2}$ and changing measure to ϵ^T.

Since τ is selected arbitrary, the inequality remain holds when taking $\sup_\tau$. We conclude that $\mathrm{reg}_T(\mathcal{H}) \ge \mathrm{sRad}_T(\mathcal{H})$, as needed.

The Sequential Fat-Shattering Dimension

We have shown that for Lipschitz losses, the minimax regret is tightly characterized by the sequential Rademacher complexity.

The Sequential Fat-Shattering Dimension

We have shown that for Lipschitz losses, the minimax regret is tightly characterized by the sequential Rademacher complexity.

But: How can we bound the sequential Rademacher complexity?

The Sequential Fat-Shattering Dimension

We have shown that for Lipschitz losses, the minimax regret is tightly characterized by the sequential Rademacher complexity.

But: How can we bound the sequential Rademacher complexity?

Sequential Fat-Shattering: Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. We say a $\mathcal{X}$-valued binary tree $\tau : \bigcup_{i \leq d} \{0,1\}^i \to \mathcal{X}$ is α-fat-shattered by $\mathcal{H}$, witnessed by a $\mathbb{R}$-valued binary tree $s : \bigcup_{i \leq d} \{0,1\}^i \to \mathbb{R}$, if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that:

1. If $\epsilon_t = 0$, then $h(\tau(\epsilon^{t-1})) \leq s(\epsilon^{t-1}) - \alpha$;
2. If $\epsilon_t = 1$, then $h(\tau(\epsilon^{t-1})) \geq s(\epsilon^{t-1}) + \alpha$.

The Sequential Fat-Shattering Dimension

We have shown that for Lipschitz losses, the minimax regret is tightly characterized by the sequential Rademacher complexity.

But: How can we bound the sequential Rademacher complexity?

Sequential Fat-Shattering: Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. We say a $\mathcal{X}$-valued binary tree $\tau : \bigcup_{i \leq d} \{0,1\}^i \to \mathcal{X}$ is α-fat-shattered by $\mathcal{H}$, witnessed by a $\mathbb{R}$-valued binary tree $s : \bigcup_{i \leq d} \{0,1\}^i \to \mathbb{R}$, if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that:

1. If $\epsilon_t = 0$, then $h(\tau(\epsilon^{t-1})) \leq s(\epsilon^{t-1}) - \alpha$;
2. If $\epsilon_t = 1$, then $h(\tau(\epsilon^{t-1})) \geq s(\epsilon^{t-1}) + \alpha$.

Sequential Fat-Shattering Dimension: The *Sequential α-Fat-Shattering Dimension* $\mathrm{sfat}_\alpha(\mathcal{H})$ for a class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ is defined as the maximal number d such that $\mathcal{H}$ can α-fat-shatter certain trees τ, s of depth d.

The Sequential Fat-Shattering Dimension

We have shown that for Lipschitz losses, the minimax regret is tightly characterized by the sequential Rademacher complexity.

But: How can we bound the sequential Rademacher complexity?

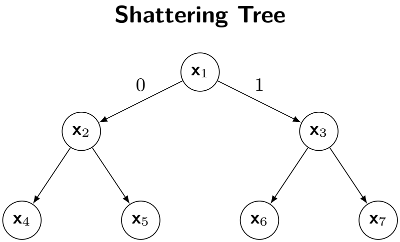
Sequential Fat-Shattering: Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. We say a $\mathcal{X}$-valued binary tree $\tau : \bigcup_{i \leq d} \{0,1\}^i \to \mathcal{X}$ is α-fat-shattered by $\mathcal{H}$, witnessed by a $\mathbb{R}$-valued binary tree $s : \bigcup_{i \leq d} \{0,1\}^i \to \mathbb{R}$, if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that:

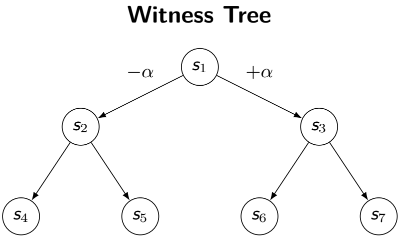
1. If $\epsilon_t = 0$, then $h(\tau(\epsilon^{t-1})) \leq s(\epsilon^{t-1}) - \alpha$;
2. If $\epsilon_t = 1$, then $h(\tau(\epsilon^{t-1})) \geq s(\epsilon^{t-1}) + \alpha$.

Sequential Fat-Shattering Dimension: The *Sequential α-Fat-Shattering Dimension* $\mathrm{sfat}_\alpha(\mathcal{H})$ for a class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ is defined as the maximal number d such that $\mathcal{H}$ can α-fat-shatter certain trees τ, s of depth d.

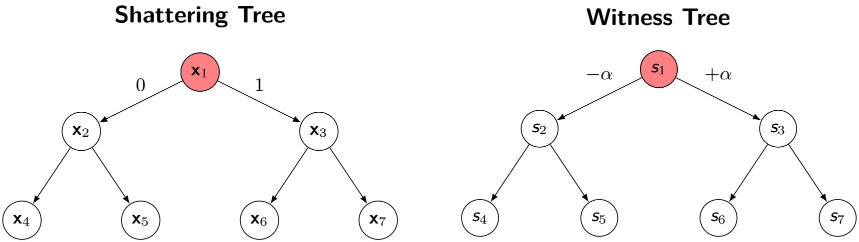
- Note that $\mathrm{sfat}_\alpha(\mathcal{H})$ mirrors the Littlestone dimension.

Shattering and Witness Trees





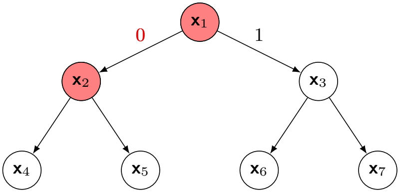
Shattering and Witness Trees



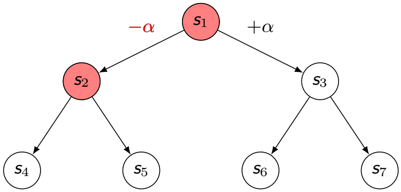
Consider a path $\{0, 1\}$, the α-fat shattering ensures $\exists h \in \mathcal{H}$ such that:

Shattering and Witness Trees

Shattering Tree



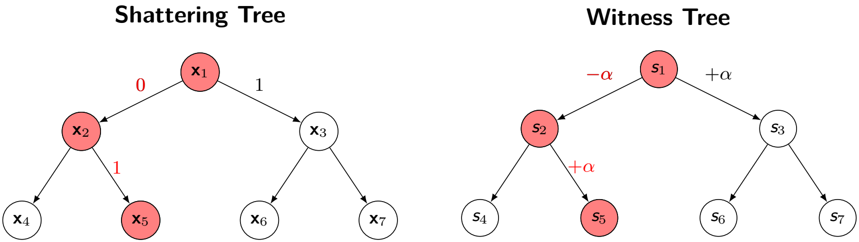
Witness Tree



Consider a path $\{0, 1\}$, the α-fat shattering ensures $\exists h \in \mathcal{H}$ such that:

1. $h(\mathbf{x}_1) \leq s_1 - \alpha$.

Shattering and Witness Trees



Consider a path $\{0, 1\}$, the α-fat shattering ensures $\exists h \in \mathcal{H}$ such that:

1. $h(\mathbf{x}_1) \leq s_1 - \alpha$.
2. $h(\mathbf{x}_2) \geq s_2 + \alpha$.

Sequential Covering for Real-valued Functions

(Real-valued) Sequential Cover: Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ and $\mathcal{G} \subset [0,1]^{\mathcal{X}^*}$ be a class mapping $\mathcal{X}^* \to [0,1]$. We say that the class $\mathcal{G}$ sequentially α-covers $\mathcal{H}$ up to step T if, for any $\mathbf{x}^T \in \mathcal{X}^T$ and $h \in \mathcal{H}$, there exists $g \in \mathcal{G}$ such that

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

Sequential Covering for Real-valued Functions

(Real-valued) Sequential Cover: Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ and $\mathcal{G} \subset [0,1]^{\mathcal{X}^*}$ be a class mapping $\mathcal{X}^* \to [0,1]$. We say that the class $\mathcal{G}$ sequentially α-covers $\mathcal{H}$ up to step T if, for any $\mathbf{x}^T \in \mathcal{X}^T$ and $h \in \mathcal{H}$, there exists $g \in \mathcal{G}$ such that

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

Similar to the binary-valued case, we can bound the (real-valued) sequential cover via the sequential fat-shattering dimension as follows:

Lemma 3: For any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ with sequential α-fat-shattering dimension $\mathsf{sfat}_\alpha(\mathcal{H})$, there exists a sequential α-cover $\mathcal{G}_\alpha$ of $\mathcal{H}$ such that

$$\log |\mathcal{G}_\alpha| \leq \tilde{O}(\mathsf{sfat}_{\alpha/3}(\mathcal{H})),$$

where $\tilde{O}$ hides poly-logarithmic factors in α and T.

Proof of Lemma 3

Let $K = \{2i\alpha : i \leq [1/(2\alpha)]\}$ be a discretization of $[0, 1]$ such that for any $a \in [0, 1]$, there exists $b \in K$ where $|a - b| \leq \alpha$.

Proof of Lemma 3

Let $K = \{2i\alpha : i \leq [1/(2\alpha)]\}$ be a discretization of $[0,1]$ such that for any $a \in [0,1]$, there exists $b \in K$ where $|a-b| \leq \alpha$.

For any $h \in \mathcal{H}$, we define a function $h' \in K^{\mathcal{X}}$ such that

$$h'(\mathbf{x}) = \arg\min_{\beta \in K} |h(\mathbf{x}) - \beta|.$$

Proof of Lemma 3

Let $K = \{2i\alpha : i \leq [1/(2\alpha)]\}$ be a discretization of $[0,1]$ such that for any $a \in [0,1]$, there exists $b \in K$ where $|a - b| \leq \alpha$.

For any $h \in \mathcal{H}$, we define a function $h' \in K^{\mathcal{X}}$ such that

$$h'(\mathbf{x}) = \arg\min_{\beta \in K} |h(\mathbf{x}) - \beta|.$$

Let $\mathcal{H}' = \{h' : h \in \mathcal{H}\} \subset K^{\mathcal{X}}$. It is easy to observe that any sequential 2α-cover of $\mathcal{H}'$ implies a sequential 3α-cover of $\mathcal{H}$. (Verify this!)

Proof of Lemma 3

Let $K = \{2i\alpha : i \leq [1/(2\alpha)]\}$ be a discretization of $[0, 1]$ such that for any $a \in [0, 1]$, there exists $b \in K$ where $|a - b| \leq \alpha$.

For any $h \in \mathcal{H}$, we define a function $h' \in K^{\mathcal{X}}$ such that

$$h'(\mathbf{x}) = \arg\min_{\beta \in K} |h(\mathbf{x}) - \beta|.$$

Let $\mathcal{H}' = \{h' : h \in \mathcal{H}\} \subset K^{\mathcal{X}}$. It is easy to observe that any sequential 2α-cover of $\mathcal{H}'$ implies a sequential 3α-cover of $\mathcal{H}$. (Verify this!)

Our primary goal is now reduced to bounding the 2α-covering set size of $\mathcal{H}'$.

Proof of Lemma 3

Let $K = \{2i\alpha : i \leq [1/(2\alpha)]\}$ be a discretization of $[0, 1]$ such that for any $a \in [0, 1]$, there exists $b \in K$ where $|a - b| \leq \alpha$.

For any $h \in \mathcal{H}$, we define a function $h' \in K^{\mathcal{X}}$ such that

$$h'(\mathbf{x}) = \arg\min_{\beta \in K} |h(\mathbf{x}) - \beta|.$$

Let $\mathcal{H}' = \{h' : h \in \mathcal{H}\} \subset K^{\mathcal{X}}$. It is easy to observe that any sequential 2α-cover of $\mathcal{H}'$ implies a sequential 3α-cover of $\mathcal{H}$. (Verify this!)

Our primary goal is now reduced to bounding the 2α-covering set size of $\mathcal{H}'$.

To achieve this, we introduce the following concept:

1-Shattering Dimension: The 1-shattering number of $\mathcal{H}'$ is defined as the maximum number d such that there exist a $\mathcal{X}$-valued tree τ and a K-valued tree s, both of depth d, such that $\forall \epsilon^d \in \{0, 1\}^d$, $\exists h' \in \mathcal{H}'$ we have:

1. If $\epsilon_t = 0$, then $h'(\tau(\epsilon^{t-1})) \leq s(\epsilon^{t-1}) - 2\alpha$;
2. If $\epsilon_t = 1$, then $h'(\tau(\epsilon^{t-1})) \geq s(\epsilon^{t-1}) + 2\alpha$.

We denote $\mathsf{FAT}_1(\mathcal{H}')$ as the 1-shattering dimension of $\mathcal{H}'$.

Proof of Lemma 3

It is easy to observe that $\mathsf{FAT}_1(\mathcal{H}') \leq \mathsf{sfat}_\alpha(\mathcal{H})$. (verify this!)

Proof of Lemma 3

It is easy to observe that $\mathsf{FAT}_1(\mathcal{H}') \le \mathsf{sfat}_\alpha(\mathcal{H})$. (verify this!)

The M-SOA Algorithm

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$, initially $\mathcal{H}^{(0)} = \mathcal{H}'$.
2. At time step t, for each $\beta \in K$, let: $\mathcal{H}^{(t)}_\beta = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = \beta\}$.
3. Predict $\hat{y}_t := \arg\max_{\beta \in K}\{\mathsf{FAT}_1(\mathcal{H}^{(t)}_\beta) : \beta \in K\}$.
4. Let y_t be the true label, and update:

$$\mathcal{H}^{(t)} = \begin{cases} \mathcal{H}^{(t)}_{y_t}, & \text{if } |\hat{y}_t - y_t| > 2\alpha, \\ \mathcal{H}^{(t-1)}, & \text{otherwise.} \end{cases}$$

Proof of Lemma 3

It is easy to observe that $\mathsf{FAT}_1(\mathcal{H}') \leq \mathsf{sfat}_\alpha(\mathcal{H})$. (verify this!)

The M-SOA Algorithm

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$, initially $\mathcal{H}^{(0)} = \mathcal{H}'$.
2. At time step t, for each $\beta \in K$, let: $\mathcal{H}_\beta^{(t)} = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = \beta\}$.
3. Predict $\hat{y}_t := \arg\max_{\beta \in K}\{\mathsf{FAT}_1(\mathcal{H}_\beta^{(t)}) : \beta \in K\}$.
4. Let y_t be the true label, and update:

$$\mathcal{H}^{(t)} = \begin{cases} \mathcal{H}_{y_t}^{(t)}, & \text{if } |\hat{y}_t - y_t| > 2\alpha, \\ \mathcal{H}^{(t-1)}, & \text{otherwise.} \end{cases}$$

Claim 1: The M-SOA algorithm enjoys the following realizable risk bound:

$$\sup_{\mathbf{x}^T} \sup_{h' \in \mathcal{H}'} \sum_{t=1}^{T} 1\{|\hat{y}_t - h'(\mathbf{x}_t)| > 2\alpha\} \leq \mathsf{FAT}_1(\mathcal{H}').$$

Proof of Lemma 3

It is easy to observe that $\mathsf{FAT}_1(\mathcal{H}') \leq \mathsf{sfat}_\alpha(\mathcal{H})$. (verify this!)

The M-SOA Algorithm

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$, initially $\mathcal{H}^{(0)} = \mathcal{H}'$.
2. At time step t, for each $\beta \in K$, let: $\mathcal{H}^{(t)}_\beta = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = \beta\}$.
3. Predict $\hat{y}_t := \arg\max_{\beta \in K}\{\mathsf{FAT}_1(\mathcal{H}^{(t)}_\beta) : \beta \in K\}$.
4. Let y_t be the true label, and update:

$$\mathcal{H}^{(t)} = \begin{cases} \mathcal{H}^{(t)}_{y_t}, & \text{if } |\hat{y}_t - y_t| > 2\alpha, \\ \mathcal{H}^{(t-1)}, & \text{otherwise.} \end{cases}$$

Claim 1: The M-SOA algorithm enjoys the following realizable risk bound:

$$\sup_{\mathbf{x}^T} \sup_{h' \in \mathcal{H}'} \sum_{t=1}^{T} 1\{|\hat{y}_t - h'(\mathbf{x}_t)| > 2\alpha\} \leq \mathsf{FAT}_1(\mathcal{H}').$$

Proof: Show that for any time step t where $|\hat{y}_t - y_t| > 2\alpha$ happens, $\mathsf{FAT}_1(\mathcal{H}^{(t)})$ is reduced by at least 1... (verify this!)

Proof of Lemma 3

Let Φ be the M-SOA algorithm.

Proof of Lemma 3

Let Φ be the M-SOA algorithm.

For any $I \subset [T]$ and $\{\beta_t\}_{t \in I} \in K^{|I|}$, we define a sequential function by simulating the M-SOA algorithm with the following modification at each step t:

1. If $t \in I$, update $\mathcal{H}^{(t)} = \mathcal{H}^{(t)}_{\beta_t}$;
2. If $t \notin I$, make no change.

Proof of Lemma 3

Let Φ be the M-SOA algorithm.

For any $I \subset [T]$ and $\{\beta_t\}_{t \in I} \in K^{|I|}$, we define a sequential function by simulating the M-SOA algorithm with the following modification at each step t:

1. If $t \in I$, update $\mathcal{H}^{(t)} = \mathcal{H}^{(t)}_{\beta_t}$;
2. If $t \notin I$, make no change.

Let $\mathcal{G}$ be the collection of all such sequential functions with $|I| \leq \mathrm{FAT}_1(\mathcal{H})$.

Proof of Lemma 3

Let Φ be the M-SOA algorithm.

For any $I \subset [T]$ and $\{\beta_t\}_{t \in I} \in K^{|I|}$, we define a sequential function by simulating the M-SOA algorithm with the following modification at each step t:

1. If $t \in I$, update $\mathcal{H}^{(t)} = \mathcal{H}^{(t)}_{\beta_t}$;
2. If $t \notin I$, make no change.

Let $\mathcal{G}$ be the collection of all such sequential functions with $|I| \leq \mathsf{FAT}_1(\mathcal{H})$.

Claim 2: The class $\mathcal{G}$ sequentially 2α-covers $\mathcal{H}'$, and

$$\log |\mathcal{G}| \leq O(\mathsf{FAT}_1(\mathcal{H}') \log(|K|T)).$$

- The covering follows from the risk bound in Claim 1. (Why?)
- The size follows by counting the number of such I's and $\{\beta_t\}_{t \in I}$'s.
- Lemma 3 follows by combining all of the previous results. (Verify this!)

Relating the Complexity Measures

Theorem 3: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$, and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. Assume the loss is bounded, convex, and Lipschitz. Then, the following statements are equivalent for a given $p \geq 2$ (the $\tilde{O}$ hides poly-logarithmic factors in α and T):

1. The *Sequential Fat-Shattering Dimension* $\mathrm{sfat}_\alpha(\mathcal{H}) = \tilde{\Theta}(\alpha^{-p})$;
2. There exists a *Sequential α-cover* $\mathcal{G}_\alpha$ with $\log|\mathcal{G}_\alpha| = \tilde{\Theta}(\alpha^{-p})$;
3. The *Sequential Rademacher Complexity* $\mathrm{sRad}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$;
4. The *minimax regret* $\mathrm{reg}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$.

Relating the Complexity Measures

Theorem 3: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$, and $\mathcal{H} \subset [0, 1]^{\mathcal{X}}$. Assume the loss is bounded, convex, and Lipschitz. Then, the following statements are equivalent for a given $p \geq 2$ (the $\tilde{O}$ hides poly-logarithmic factors in α and T):

1. The *Sequential Fat-Shattering Dimension* $\mathrm{sfat}_\alpha(\mathcal{H}) = \tilde{\Theta}(\alpha^{-p})$;
2. There exists a *Sequential* α*-cover* $\mathcal{G}_\alpha$ with $\log |\mathcal{G}_\alpha| = \tilde{\Theta}(\alpha^{-p})$;
3. The *Sequential Rademacher Complexity* $\mathrm{sRad}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$;
4. The *minimax regret* $\mathrm{reg}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$.

- In this lecture, we showed that $1 \Rightarrow 2$ (Lemma 3) and $3 \Leftrightarrow 4$ (Theorem 2). The other implications require more technical treatment.

Relating the Complexity Measures

Theorem 3: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$, and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. Assume the loss is bounded, convex, and Lipschitz. Then, the following statements are equivalent for a given $p \geq 2$ (the $\tilde{O}$ hides poly-logarithmic factors in α and T):

1. The *Sequential Fat-Shattering Dimension* $\mathrm{sfat}_\alpha(\mathcal{H}) = \tilde{\Theta}(\alpha^{-p})$;
2. There exists a *Sequential α-cover* $\mathcal{G}_\alpha$ with $\log |\mathcal{G}_\alpha| = \tilde{\Theta}(\alpha^{-p})$;
3. The *Sequential Rademacher Complexity* $\mathrm{sRad}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$;
4. The *minimax regret* $\mathrm{reg}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$.

- In this lecture, we showed that $1 \Rightarrow 2$ (Lemma 3) and $3 \Leftrightarrow 4$ (Theorem 2). The other implications require more technical treatment.
  - For these proofs, refer to (Rakhlin, Sridharan, Tewari, JMLR 2016)...

Relating the Complexity Measures

Theorem 3: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$, and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. Assume the loss is bounded, convex, and Lipschitz. Then, the following statements are equivalent for a given $p \geq 2$ (the $\tilde{O}$ hides poly-logarithmic factors in α and T):

1. The *Sequential Fat-Shattering Dimension* $\mathrm{sfat}_\alpha(\mathcal{H}) = \tilde{\Theta}(\alpha^{-p})$;
2. There exists a *Sequential α-cover* $\mathcal{G}_\alpha$ with $\log|\mathcal{G}_\alpha| = \tilde{\Theta}(\alpha^{-p})$;
3. The *Sequential Rademacher Complexity* $\mathrm{sRad}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$;
4. The *minimax regret* $\mathrm{reg}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$.

- In this lecture, we showed that $1 \Rightarrow 2$ (Lemma 3) and $3 \Leftrightarrow 4$ (Theorem 2). The other implications require more technical treatment.
  - For these proofs, refer to (Rakhlin, Sridharan, Tewari, JMLR 2016)...
- Note that a naïve implication $2 \Rightarrow 4$ can be obtained via the EWA algorithm, but with a bound of $\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p+1}{p+2}})$. (Proof left as **Homework**).

Relating the Complexity Measures

Theorem 3: Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$, and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. Assume the loss is bounded, convex, and Lipschitz. Then, the following statements are equivalent for a given $p \geq 2$ (the $\tilde{O}$ hides poly-logarithmic factors in α and T):

1. The *Sequential Fat-Shattering Dimension* $\mathrm{sfat}_\alpha(\mathcal{H}) = \tilde{\Theta}(\alpha^{-p})$;
2. There exists a *Sequential* α*-cover* $\mathcal{G}_\alpha$ with $\log|\mathcal{G}_\alpha| = \tilde{\Theta}(\alpha^{-p})$;
3. The *Sequential Rademacher Complexity* $\mathrm{sRad}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$;
4. The *minimax regret* $\mathrm{reg}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$.

- In this lecture, we showed that $1 \Rightarrow 2$ (Lemma 3) and $3 \Leftrightarrow 4$ (Theorem 2). The other implications require more technical treatment.
  - For these proofs, refer to (Rakhlin, Sridharan, Tewari, JMLR 2016)...
- Note that a naïve implication $2 \Rightarrow 4$ can be obtained via the EWA algorithm, but with a bound of $\mathrm{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p+1}{p+2}})$. (Proof left as **Homework**).
- The tighter $\tilde{O}(T^{\frac{p-1}{p}})$ regret bound arises from the benefit of chaining, through the path $2 \Rightarrow 3 \Rightarrow 4$.

Overview

- **Bayesian Representation of Minimax Regret**
  - The minimax switching trick
- **Bounding the Minimax Regret: Real-valued Case**
  - The sequential Rademacher complexity, symmetrization
  - The Sequential fat-shattering dimension
  - Regret bounds via Sequential fat-shattering dimension
- **From Value to Algorithm**
  - The relaxation framework
  - The hybrid setting, random play-out

From Value to Algorithm

So far, we have discussed various approaches to bound the minimax regret without designing an algorithm.

From Value to Algorithm

So far, we have discussed various approaches to bound the minimax regret without designing an algorithm.

What algorithm achieves such regret?

From Value to Algorithm

So far, we have discussed various approaches to bound the minimax regret without designing an algorithm.

What algorithm achieves such regret?

For any $\mathbf{x}^{t-1}$ and y^{t-1}, we define the partial minimax regret as:

$$\begin{aligned}&\mathrm{reg}_T^{(t)}(\mathcal{H}, \mathbf{x}^{t-1}, y^{t-1}) \\ &\quad = \mathcal{Q}_t \left[\ell(\hat{y}_t, y_t) + \mathcal{Q}_{t+1} \left[\ell(\hat{y}_{t+1}, y_{t+1}) + \cdots - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_T), y_T) \right] \right]\end{aligned}$$

where $\mathcal{Q}_t := \sup_{\mathbf{x}_t} \inf_{\hat{y}_t} \sup_{y_t}$.

From Value to Algorithm

So far, we have discussed various approaches to bound the minimax regret without designing an algorithm.

What algorithm achieves such regret?

For any $\mathbf{x}^{t-1}$ and y^{t-1}, we define the partial minimax regret as:

$$\begin{aligned}&\mathsf{reg}_T^{(t)}(\mathcal{H}, \mathbf{x}^{t-1}, y^{t-1})\\ &\quad = \mathcal{Q}_t \left[\ell(\hat{y}_t, y_t) + \mathcal{Q}_{t+1} \left[\ell(\hat{y}_{t+1}, y_{t+1}) + \cdots - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_T), y_T) \right] \right]\end{aligned}$$

where $\mathcal{Q}_t := \sup_{\mathbf{x}_t} \inf_{\hat{y}_t} \sup_{y_t}$.

It is easy to observe that the following naïve algorithm is minimax optimal:

$$\hat{y}_t = \arg\min_{\hat{y}} \sup_{y} \left[\ell(\hat{y}, y) + \mathsf{reg}_T^{(t+1)}(\mathcal{H}, \mathbf{x}^t, y^{t-1}y) \right].$$

From Value to Algorithm

So far, we have discussed various approaches to bound the minimax regret without designing an algorithm.

What algorithm achieves such regret?

For any $\mathbf{x}^{t-1}$ and y^{t-1}, we define the partial minimax regret as:

$$\begin{aligned}&\mathsf{reg}_T^{(t)}(\mathcal{H}, \mathbf{x}^{t-1}, y^{t-1})\\&\quad = \mathcal{Q}_t \left[\ell(\hat{y}_t, y_t) + \mathcal{Q}_{t+1}\left[\ell(\hat{y}_{t+1}, y_{t+1}) + \cdots - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_T), y_T)\right]\right]\end{aligned}$$

where $\mathcal{Q}_t := \sup_{\mathbf{x}_t} \inf_{\hat{y}_t} \sup_{y_t}$.

It is easy to observe that the following naïve algorithm is minimax optimal:

$$\hat{y}_t = \arg\min_{\hat{y}} \sup_{y} \left[\ell(\hat{y}, y) + \mathsf{reg}_T^{(t+1)}(\mathcal{H}, \mathbf{x}^t, y^{t-1}y)\right].$$

(**Hint**: Backward induction on $\mathsf{reg}_T^{(t)}(\mathcal{H}, \mathbf{x}^{t-1}, y^{t-1})$ from $t = T$ to 1.)

Relaxation

Note that the partial minimax regret involves complicated iterative minimax optimizations, which is generally not easy to compute.

Relaxation

Note that the partial minimax regret involves complicated iterative minimax optimizations, which is generally not easy to compute.

A natural approach is to replace the actual partial minimax regret with some more manageable functions.

Relaxation

Note that the partial minimax regret involves complicated iterative minimax optimizations, which is generally not easy to compute.

A natural approach is to replace the actual partial minimax regret with some more manageable functions.

We define the relaxation as a function: $\mathrm{Rel}_T : (\mathcal{X} \times \mathcal{Y})^* \to \mathbb{R}$.

Relaxation

Note that the partial minimax regret involves complicated iterative minimax optimizations, which is generally not easy to compute.

A natural approach is to replace the actual partial minimax regret with some more manageable functions.

We define the relaxation as a function: $\mathsf{Rel}_T : (\mathcal{X} \times \mathcal{Y})^* \to \mathbb{R}$.

A relaxation Rel is said to be admissible w.r.t. a class $\mathcal{H}$ if for any $\mathbf{x}^T, y^T$

1. $\mathsf{Rel}_T(\mathbf{x}^T, y^T) \geq -\inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)$.
2. For any $t < T$, we have

$$\sup_{\mathbf{x}} \inf_{\hat{y}} \sup_{y} \left[\ell(\hat{y}, y) + \mathsf{Rel}(\mathbf{x}^{t-1}\mathbf{x}, y^{t-1}y)\right] \leq \mathsf{Rel}(\mathbf{x}^{t-1}, y^{t-1}).$$

Relaxation

Note that the partial minimax regret involves complicated iterative minimax optimizations, which is generally not easy to compute.

A natural approach is to replace the actual partial minimax regret with some more manageable functions.

We define the relaxation as a function: $\mathrm{Rel}_T : (\mathcal{X} \times \mathcal{Y})^* \to \mathbb{R}$.

A relaxation Rel is said to be admissible w.r.t. a class $\mathcal{H}$ if for any $\mathbf{x}^T, y^T$

1. $\mathrm{Rel}_T(\mathbf{x}^T, y^T) \geq -\inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)$.
2. For any $t < T$, we have

$$\sup_{\mathbf{x}} \inf_{\hat{y}} \sup_{y} \left[\ell(\hat{y}, y) + \mathrm{Rel}(\mathbf{x}^{t-1}\mathbf{x}, y^{t-1}y)\right] \leq \mathrm{Rel}(\mathbf{x}^{t-1}, y^{t-1}).$$

Lemma 4: Let Rel_T be a relaxation that is admissible w.r.t. a class $\mathcal{H}$, then the following predictor Φ

$$\hat{y}_t = \arg\min_{\hat{y}} \sup_{y} \left[\ell(\hat{y}, y) + \mathrm{Rel}_T(\mathbf{x}^t, y^{t-1}y)\right]$$

achieves the worst-case regret $\mathrm{reg}_T(\mathcal{H}, \Phi) \leq \mathrm{Rel}_T(\emptyset)$.

Proof of Lemma 4

By condition 1 of admissibility, we have

$$\sup_{\mathbf{x}_1, y_1} \cdots \sup_{\mathbf{x}_{T-1}, y_{T-1}} \left[\sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t, y_t)) \right] \right]$$

$$\leq \sup_{\mathbf{x}_1, y_1} \cdots \sup_{\mathbf{x}_{T-1}, y_{T-1}} \left[\sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) + \mathrm{Rel}_T(\mathbf{x}^T, y^T) \right] \right]$$

Proof of Lemma 4

By condition 1 of admissibility, we have

$$\sup_{\mathbf{x}_1, y_1} \cdots \sup_{\mathbf{x}_{T-1}, y_{T-1}} \left[\sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t, y_t)) \right] \right]$$
$$\leq \sup_{\mathbf{x}_1, y_1} \cdots \sup_{\mathbf{x}_{T-1}, y_{T-1}} \left[\sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) + \mathrm{Rel}_T(\mathbf{x}^T, y^T) \right] \right]$$

Note that, by definition of $\hat{y}_T$, we have

$$\sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) + \mathrm{Rel}_T(\mathbf{x}^T, y^T) \right] = \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\ell(\hat{y}_T, y_T) + \mathrm{Rel}_T(\mathbf{x}^T, y^T) \right]$$
$$\overset{(\star)}{\leq} \mathrm{Rel}_T(\mathbf{x}^{T-1}, y^{T-1})$$

Proof of Lemma 4

By condition 1 of admissibility, we have

$$\sup_{\mathbf{x}_1,y_1}\cdots\sup_{\mathbf{x}_{T-1},y_{T-1}}\left[\sum_{t=1}^{T-1}\ell(\hat{y}_t,y_t)+\sup_{\mathbf{x}_T,y_T}\left[\ell(\hat{y}_T,y_T)-\inf_{h\in\mathcal{H}}\sum_{t=1}^{T}\ell(h(\mathbf{x}_t,y_t))\right]\right]$$
$$\leq \sup_{\mathbf{x}_1,y_1}\cdots\sup_{\mathbf{x}_{T-1},y_{T-1}}\left[\sum_{t=1}^{T-1}\ell(\hat{y}_t,y_t)+\sup_{\mathbf{x}_T,y_T}\left[\ell(\hat{y}_T,y_T)+\mathrm{Rel}_T(\mathbf{x}^T,y^T)\right]\right]$$

Note that, by definition of $\hat{y}_T$, we have

$$\sup_{\mathbf{x}_T,y_T}\left[\ell(\hat{y}_T,y_T)+\mathrm{Rel}_T(\mathbf{x}^T,y^T)\right]=\sup_{\mathbf{x}_T}\inf_{\hat{y}_T}\sup_{y_T}\left[\ell(\hat{y}_T,y_T)+\mathrm{Rel}_T(\mathbf{x}^T,y^T)\right]$$
$$\overset{(\star)}{\leq}\mathrm{Rel}_T(\mathbf{x}^{T-1},y^{T-1})$$

where $(\star)$ follows by condition 2 of admissibility.

Proof of Lemma 4

By condition 1 of admissibility, we have

$$\sup_{\mathbf{x}_1, y_1} \cdots \sup_{\mathbf{x}_{T-1}, y_{T-1}} \left[\sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t, y_t)) \right] \right]$$
$$\leq \sup_{\mathbf{x}_1, y_1} \cdots \sup_{\mathbf{x}_{T-1}, y_{T-1}} \left[\sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) + \mathrm{Rel}_T(\mathbf{x}^T, y^T) \right] \right]$$

Note that, by definition of $\hat{y}_T$, we have

$$\sup_{\mathbf{x}_T, y_T} \left[\ell(\hat{y}_T, y_T) + \mathrm{Rel}_T(\mathbf{x}^T, y^T) \right] = \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\ell(\hat{y}_T, y_T) + \mathrm{Rel}_T(\mathbf{x}^T, y^T) \right]$$
$$\overset{(\star)}{\leq} \mathrm{Rel}_T(\mathbf{x}^{T-1}, y^{T-1})$$

where $(\star)$ follows by condition 2 of admissibility.

Continue this argument for another $T-1$ steps, we have $\mathrm{reg}_T(\mathcal{H}, \Phi) \leq \mathrm{Rel}_T(\emptyset)$.

The Hybrid Setup

We have shown that a good relaxation automatically provides a good algorithm by solving an optimization problem with respect to the relaxation.

The Hybrid Setup

We have shown that a good relaxation automatically provides a good algorithm by solving an optimization problem with respect to the relaxation.

However, the computation is typically quite expensive.

The Hybrid Setup

We have shown that a good relaxation automatically provides a good algorithm by solving an optimization problem with respect to the relaxation.

However, the computation is typically quite expensive.

How can we construct a relaxation that leads to efficient algorithms?

The Hybrid Setup

We have shown that a good relaxation automatically provides a good algorithm by solving an optimization problem with respect to the relaxation.

However, the computation is typically quite expensive.

How can we construct a relaxation that leads to efficient algorithms?

- It turns out that a generic efficient algorithm is not possible for worst-case regret, even for finite classes.
  - See "The Computational Power of Optimization in Online Learning" by E. Hazan and T. Koren (STOC 2016).
- A workaround is to consider a weaker adversary/nature that generates data.

The Hybrid Setup

We have shown that a good relaxation automatically provides a good algorithm by solving an optimization problem with respect to the relaxation.

However, the computation is typically quite expensive.

How can we construct a relaxation that leads to efficient algorithms?

- It turns out that a generic efficient algorithm is not possible for worst-case regret, even for finite classes.
  - See "The Computational Power of Optimization in Online Learning" by E. Hazan and T. Koren (STOC 2016).
- A workaround is to consider a weaker adversary/nature that generates data.

Hybrid Regret: Let μ be a distribution over $\mathcal{X}$. The hybrid regret for a predictor Φ is defined as:

$$\tilde{\mathrm{reg}}_T(\mathcal{H}, \Phi, \mu) = \mathbb{E}_{\mathbf{x}_1} \sup_{y_1} \cdots \mathbb{E}_{\mathbf{x}_T} \sup_{y_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\mathbf{x}_t \sim \mu$ and are independent for different $t \leq T$.

Oracle Efficiency

Since we do not impose any structural assumptions on $\mathcal{H}$, a generic efficient algorithm is out of reach in the standard computational model.

Oracle Efficiency

Since we do not impose any structural assumptions on $\mathcal{H}$, a generic efficient algorithm is out of reach in the standard computational model.

Instead, we consider a weaker notion of oracle efficiency:

- Given any data $\mathbf{x}^t, y^t$, the Empirical Risk Minimization (ERM) oracle finds

$$\hat{h}_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{t} \ell(h(\mathbf{x}_i), y_i).$$

Oracle Efficiency

Since we do not impose any structural assumptions on $\mathcal{H}$, a generic efficient algorithm is out of reach in the standard computational model.

Instead, we consider a weaker notion of oracle efficiency:

- Given any data $\mathbf{x}^t, y^t$, the Empirical Risk Minimization (ERM) oracle finds

$$\hat{h}_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{t} \ell(h(\mathbf{x}_i), y_i).$$

- A prediction rule is oracle efficient if it runs in polynomial time by accessing the ERM oracle, with each oracle call counted as unit time.

Oracle Efficiency

Since we do not impose any structural assumptions on $\mathcal{H}$, a generic efficient algorithm is out of reach in the standard computational model.

Instead, we consider a weaker notion of oracle efficiency:

- Given any data $\mathbf{x}^t, y^t$, the Empirical Risk Minimization (ERM) oracle finds

$$\hat{h}_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{t} \ell(h(\mathbf{x}_i), y_i).$$

- A prediction rule is oracle efficient if it runs in polynomial time by accessing the ERM oracle, with each oracle call counted as unit time.
- The ERM oracle can often be computed efficiently in practice, even for non-convex classes like neural networks, using gradient-based methods.

Oracle Efficient Regret for Hybrid Regret

Theorem 3: For any given distribution μ over $\mathcal{X}$ and class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, if the loss function ℓ is convex and Lipschitz in its first argument, then there exists an oracle efficient predictor Φ such that:

$$\tilde{\mathrm{reg}}_T(\mathcal{H}, \Phi, \mu) \leq O(\mathrm{Rad}_T(\mathcal{H})),$$

where $\mathrm{Rad}_T(\mathcal{H})$ is the standard (non-sequential) Rademacher complexity of $\mathcal{H}$.

Oracle Efficient Regret for Hybrid Regret

Theorem 3: For any given distribution μ over $\mathcal{X}$ and class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, if the loss function ℓ is convex and Lipschitz in its first argument, then there exists an oracle efficient predictor Φ such that:

$$\tilde{\text{reg}}_T(\mathcal{H}, \Phi, \mu) \leq O(\text{Rad}_T(\mathcal{H})),$$

where $\text{Rad}_T(\mathcal{H})$ is the standard (non-sequential) Rademacher complexity of $\mathcal{H}$.

- The proof follows by finding an admissible relaxation Rel_T such that the induced predictor $\hat{y}_t = \arg\min_{\hat{y}} \sup_y \left[\ell(\hat{y}, y) + \text{Rel}_T(\mathbf{x}^t, y^{t-1}y)\right]$ can be computed in an oracle efficient manner.

Oracle Efficient Regret for Hybrid Regret

Theorem 3: For any given distribution μ over $\mathcal{X}$ and class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, if the loss function ℓ is convex and Lipschitz in its first argument, then there exists an oracle efficient predictor Φ such that:

$$\tilde{\mathsf{reg}}_T(\mathcal{H}, \Phi, \mu) \leq O(\mathsf{Rad}_T(\mathcal{H})),$$

where $\mathsf{Rad}_T(\mathcal{H})$ is the standard (non-sequential) Rademacher complexity of $\mathcal{H}$.

- The proof follows by finding an admissible relaxation Rel_T such that the induced predictor $\hat{y}_t = \arg\min_{\hat{y}} \sup_y \left[\ell(\hat{y}, y) + \mathsf{Rel}_T(\mathbf{x}^t, y^{t-1}y)\right]$ can be computed in an oracle efficient manner.
- The oracle efficiency follows by a random play-out approach that bypassed the estimation of Rel_T with a single random value.

Oracle Efficient Regret for Hybrid Regret

Theorem 3: For any given distribution μ over $\mathcal{X}$ and class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, if the loss function ℓ is convex and Lipschitz in its first argument, then there exists an oracle efficient predictor Φ such that:

$$\tilde{\mathrm{reg}}_T(\mathcal{H}, \Phi, \mu) \le O(\mathrm{Rad}_T(\mathcal{H})),$$

where $\mathrm{Rad}_T(\mathcal{H})$ is the standard (non-sequential) Rademacher complexity of $\mathcal{H}$.

- The proof follows by finding an admissible relaxation Rel_T such that the induced predictor $\hat{y}_t = \arg\min_{\hat{y}} \sup_y \left[\ell(\hat{y}, y) + \mathrm{Rel}_T(\mathbf{x}^t, y^{t-1}y)\right]$ can be computed in an oracle efficient manner.
- The oracle efficiency follows by a random play-out approach that bypassed the estimation of Rel_T with a single random value.
  - See "Relax and Randomize: From Value to Algorithms" by Rakhlin, Shamir, and Sridharan (NeurIPS 2012).

Oracle Efficient Regret for Hybrid Regret

Theorem 3: For any given distribution μ over $\mathcal{X}$ and class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, if the loss function ℓ is convex and Lipschitz in its first argument, then there exists an oracle efficient predictor Φ such that:

$$\tilde{\mathsf{reg}}_T(\mathcal{H}, \Phi, \mu) \leq O(\mathsf{Rad}_T(\mathcal{H})),$$

where $\mathsf{Rad}_T(\mathcal{H})$ is the standard (non-sequential) Rademacher complexity of $\mathcal{H}$.

- The proof follows by finding an admissible relaxation Rel_T such that the induced predictor $\hat{y}_t = \arg\min_{\hat{y}} \sup_y \left[\ell(\hat{y}, y) + \mathsf{Rel}_T(\mathbf{x}^t, y^{t-1}y)\right]$ can be computed in an oracle efficient manner.
- The oracle efficiency follows by a random play-out approach that bypassed the estimation of Rel_T with a single random value.
  - See "Relax and Randomize: From Value to Algorithms" by Rakhlin, Shamir, and Sridharan (NeurIPS 2012).
- It remains an active research area to explore oracle efficient predictors for more complex and unknown feature generation processes.

Oracle Efficient Regret for Hybrid Regret

Theorem 3: For any given distribution μ over $\mathcal{X}$ and class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, if the loss function ℓ is convex and Lipschitz in its first argument, then there exists an oracle efficient predictor Φ such that:

$$\tilde{\mathrm{reg}}_T(\mathcal{H}, \Phi, \mu) \le O(\mathrm{Rad}_T(\mathcal{H})),$$

where $\mathrm{Rad}_T(\mathcal{H})$ is the standard (non-sequential) Rademacher complexity of $\mathcal{H}$.

- The proof follows by finding an admissible relaxation Rel_T such that the induced predictor $\hat{y}_t = \arg\min_{\hat{y}} \sup_y \left[\ell(\hat{y}, y) + \mathrm{Rel}_T(\mathbf{x}^t, y^{t-1}y)\right]$ can be computed in an oracle efficient manner.
- The oracle efficiency follows by a random play-out approach that bypassed the estimation of Rel_T with a single random value.
  - See "Relax and Randomize: From Value to Algorithms" by Rakhlin, Shamir, and Sridharan (NeurIPS 2012).
- It remains an active research area to explore oracle efficient predictors for more complex and unknown feature generation processes.
  - See our recent paper "Oracle-Efficient Hybrid Online Learning with Unknown Distribution" by Wu, Sima, and Szpankowski (COLT 2024).

Concluding Remarks

- In this lecture, we introduced a general approach for bounding the minimax regret by converting it to a Bayesian representation.
- We showed that this Bayesian representation can be naturally bounded by the sequential Rademacher complexity through a symmetrization argument.
- We further demonstrated that the sequential Rademacher complexity can be effectively controlled by the sequential fat-shattering dimension.
- Finally, we discussed a principled way to construct prediction algorithms via the concept of admissible relaxation and addressed the issue of computational efficiency.
- A key assumption we made throughout this lecture is the Lipschitz condition of the loss, which is not always satisfied for certain natural losses. We will address this in the upcoming lecture.