Minimax Value of Online Learning Games: Part I

Changlong Wu & Wojciech Szpankowski

Center for Science of Information
Purdue University

October 21, 2024

Overview

- **Minimax Regret**
  - Pointwise, worst-case, and minimax regrets
  - The iterative minimax formulation
- **Bounding the Minimax Regret: Binary Labels**
  - The Littlestone dimension
  - Standard Optimal Algorithm
  - Sequential covering
- **The Minimax Theorem**
  - Proving minimax theorem via EWA algorithm

Minimax Regret

Let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ be the label space and $\hat{\mathcal{Y}}$ be a (convex) outcome space of predictors.

Unlike previous lecture, we define the hypothesis class as $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ and the learning rule (possibly improper) as:

$$\Phi : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \hat{\mathcal{Y}}.$$

For $t = 1, 2, \cdots, T$

1. Nature/Environment presents an instance $\mathbf{x}_t \in \mathcal{X}$
2. Learner predicts a label $\hat{y}_t \in \hat{\mathcal{Y}}$ via $\hat{y}_t := \Phi(\mathbf{x}^t, y^{t-1})$
3. Nature reveals true label $y_t \in \mathcal{Y}$
4. Learner suffers loss $\ell(\hat{y}_t, y_t)$, for certain function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$

Goal of Learner: Minimizes regret for the worst Nature.

Minimax Regret

For any given $\mathbf{x}^T \in \mathcal{X}$ and $y^T \in \mathcal{Y}^T$, the point-wise regret is defined as

$$R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T) := \sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)$$

Minimax Regret

For any given $\mathbf{x}^T \in \mathcal{X}$ and $y^T \in \mathcal{Y}^T$, the point-wise regret is defined as

$$R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T) := \sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)$$

The worst-case regret for give learning rule Φ is defined as

$$\mathsf{reg}_T(\mathcal{H}, \Phi) := \sup_{\mathbf{x}^T, y^T} R_T(\mathcal{H}, \Phi, x^T, y^T)$$

Minimax Regret

For any given $\mathbf{x}^T \in \mathcal{X}$ and $y^T \in \mathcal{Y}^T$, the point-wise regret is defined as

$$R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T) := \sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)$$

The worst-case regret for give learning rule Φ is defined as

$$\mathsf{reg}_T(\mathcal{H}, \Phi) := \sup_{\mathbf{x}^T, y^T} R_T(\mathcal{H}, \Phi, x^T, y^T)$$

The minimax regret for a hypothesis class $\mathcal{H}$ is defined as

$$\mathsf{reg}_T(\mathcal{H}) := \inf_{\Phi} \mathsf{reg}_T(\mathcal{H}, \Phi) = \inf_{\Phi} \sup_{\mathbf{x}^T, y^T} R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T)$$

Minimax Regret

For any given $\mathbf{x}^T \in \mathcal{X}$ and $y^T \in \mathcal{Y}^T$, the point-wise regret is defined as

$$R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T) := \sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)$$

The worst-case regret for give learning rule Φ is defined as

$$\mathrm{reg}_T(\mathcal{H}, \Phi) := \sup_{\mathbf{x}^T, y^T} R_T(\mathcal{H}, \Phi, x^T, y^T)$$

The minimax regret for a hypothesis class $\mathcal{H}$ is defined as

$$\mathrm{reg}_T(\mathcal{H}) := \inf_{\Phi} \mathrm{reg}_T(\mathcal{H}, \Phi) = \inf_{\Phi} \sup_{\mathbf{x}^T, y^T} R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T)$$

Fact 1: The minimax regret satisfies

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

Preparing for the Proof: Skolemization

Skolemization: Let A, B be two sets, and $F : A \times B \to \mathbb{R}$ be an arbitrary function, then

$$\sup_{b \in B} \inf_{a \in A} F(a, b) = \inf_{g \in \mathcal{G}} \sup_{b \in B} F(g(b), b),$$

where $\mathcal{G} := A^B$ is the class of all functions from $B \to A$.

Preparing for the Proof: Skolemization

Skolemization: Let A, B be two sets, and $F : A \times B \to \mathbb{R}$ be an arbitrary function, then

$$\sup_{b \in B} \inf_{a \in A} F(a, b) = \inf_{g \in \mathcal{G}} \sup_{b \in B} F(g(b), b),$$

where $\mathcal{G} := A^B$ is the class of all functions from $B \to A$.

- Define $\hat{g}(b) := \arg\inf_{a \in A} F(a, b)$ we have

$$\sup_{b} \inf_{a} F(a, b) = \sup_{b} F(\hat{g}(b), b) \geq \inf_{g} \sup_{b} F(g(b), b).$$

Preparing for the Proof: Skolemization

Skolemization: Let A, B be two sets, and $F : A \times B \to \mathbb{R}$ be an arbitrary function, then

$$\sup_{b \in B} \inf_{a \in A} F(a, b) = \inf_{g \in \mathcal{G}} \sup_{b \in B} F(g(b), b),$$

where $\mathcal{G} := A^B$ is the class of all functions from $B \to A$.

- Define $\hat{g}(b) := \arg\inf_{a \in A} F(a, b)$ we have

$$\sup_b \inf_a F(a, b) = \sup_b F(\hat{g}(b), b) \geq \inf_g \sup_b F(g(b), b).$$

- Moreover, let $g^* := \arg\min_{g \in \mathcal{G}} (\sup_b F(g(b), b))$ we have

$$\inf_g \sup_b F(g(b), b) = \sup_b F(g^*(b), b) \geq \sup_b \inf_a F(a, b).$$

Preparing for the Proof: Skolemization

Skolemization: Let A, B be two sets, and $F : A \times B \to \mathbb{R}$ be an arbitrary function, then

$$\sup_{b \in B} \inf_{a \in A} F(a, b) = \inf_{g \in \mathcal{G}} \sup_{b \in B} F(g(b), b),$$

where $\mathcal{G} := A^B$ is the class of all functions from $B \to A$.

- Define $\hat{g}(b) := \arg\inf_{a \in A} F(a, b)$ we have

$$\sup_{b} \inf_{a} F(a, b) = \sup_{b} F(\hat{g}(b), b) \geq \inf_{g} \sup_{b} F(g(b), b).$$

- Moreover, let $g^* := \arg\min_{g \in \mathcal{G}}(\sup_b F(g(b), b))$ we have

$$\inf_{g} \sup_{b} F(g(b), b) = \sup_{b} F(g^*(b), b) \geq \sup_{b} \inf_{a} F(a, b).$$

- Therefore, all inequalities become equality and the result follows.

Proof of Fact 1

We prove only the case for $T = 1$ to demonstrate the idea.

Proof of Fact 1

We prove only the case for $T = 1$ to demonstrate the idea. Define the function:

$$F(a, b) := \sup_{y_1} \left[\ell(a, y_1) - \inf_{h \in \mathcal{H}} \ell(h(b), y_1) \right].$$

Proof of Fact 1

We prove only the case for $T = 1$ to demonstrate the idea. Define the function:

$$F(a, b) := \sup_{y_1} \left[\ell(a, y_1) - \inf_{h \in \mathcal{H}} \ell(h(b), y_1) \right].$$

Note that:

$$\text{reg}_1(\mathcal{H}) := \inf_{\Phi} \sup_{\mathbf{x}_1} F(\Phi(\mathbf{x}_1), \mathbf{x}_1).$$

Proof of Fact 1

We prove only the case for $T = 1$ to demonstrate the idea. Define the function:

$$F(a, b) := \sup_{y_1} \left[\ell(a, y_1) - \inf_{h \in \mathcal{H}} \ell(h(b), y_1) \right].$$

Note that:

$$\text{reg}_1(\mathcal{H}) := \inf_{\Phi} \sup_{\mathbf{x}_1} F(\Phi(\mathbf{x}_1), \mathbf{x}_1).$$

By Skolemization, we have:

$$\inf_{\Phi} \sup_{\mathbf{x}_1} F(\Phi(\mathbf{x}_1), \mathbf{x}_1) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} F(\hat{y}_1, \mathbf{x}_1).$$

Proof of Fact 1

We prove only the case for $T = 1$ to demonstrate the idea. Define the function:

$$F(a, b) := \sup_{y_1} \left[\ell(a, y_1) - \inf_{h \in \mathcal{H}} \ell(h(b), y_1) \right].$$

Note that:

$$\mathrm{reg}_1(\mathcal{H}) := \inf_{\Phi} \sup_{\mathbf{x}_1} F(\Phi(\mathbf{x}_1), \mathbf{x}_1).$$

By Skolemization, we have:

$$\inf_{\Phi} \sup_{\mathbf{x}_1} F(\Phi(\mathbf{x}_1), \mathbf{x}_1) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} F(\hat{y}_1, \mathbf{x}_1).$$

Plugging back the expression of $F(a, b)$, we get:

$$\mathrm{reg}_1(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \left[\ell(\hat{y}_1, y_1) - \inf_{h \in \mathcal{H}} \ell(h(\mathbf{x}_1), y_1) \right].$$

Overview

- **Minimax Regret**
  - Pointwise, worst-case, and minimax regrets
  - The iterative minimax formulation
- **Bounding the Minimax Regret: Binary Labels**
  - The Littlestone dimension
  - Standard Optimal Algorithm
  - Sequential covering
- **The Minimax Theorem**
  - Proving minimax theorem via EWA algorithm

Preliminaries

We now consider the case when $\mathcal{Y} = \{0, 1\}$ and $\hat{\mathcal{Y}} = [0, 1]$, and consider also the specific loss function (i.e., the absolute loss):

$$\ell(\hat{y}, y) = |\hat{y} - y|.$$

Preliminaries

We now consider the case when $\mathcal{Y} = \{0, 1\}$ and $\hat{\mathcal{Y}} = [0, 1]$, and consider also the specific loss function (i.e., the absolute loss):

$$\ell(\hat{y}, y) = |\hat{y} - y|.$$

Observe that $|\hat{y} - y| = \mathbb{E}_{y' \sim \text{Bern}(\hat{y})}[1\{y' \neq y\}]$, i.e., it measures the *expected miss-classification loss* when sampling from a Bernoulli source of parameter $\hat{y}$.

Preliminaries

We now consider the case when $\mathcal{Y} = \{0, 1\}$ and $\hat{\mathcal{Y}} = [0, 1]$, and consider also the specific loss function (i.e., the absolute loss):

$$\ell(\hat{y}, y) = |\hat{y} - y|.$$

Observe that $|\hat{y} - y| = \mathbb{E}_{y' \sim \mathsf{Bern}(\hat{y})}[1\{y' \neq y\}]$, i.e., it measures the *expected miss-classification loss* when sampling from a Bernoulli source of parameter $\hat{y}$.

Recall from our last lecture:

Theorem 1: For any finite class $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$, the minimax regret of $\mathcal{H}$ under the absolute loss is upper bounded by

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

which is achieved by the (generalized) EWA algorithm.

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

What happens for infinite classes?

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

What happens for infinite classes?

- Consider the following **threshold functions**:

$$\mathcal{H}^{\text{thres}} := \{h_a(x) = 1\{x \geq a\} : a, x \in [0, 1]\}.$$

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

What happens for infinite classes?

- Consider the following **threshold functions**:

$$\mathcal{H}^{\text{thres}} := \{h_a(x) = 1\{x \geq a\} : a, x \in [0,1]\}.$$

- For any learner Φ, consider the following strategy for Nature:
  - At every step t, select label $y_t \in \{0,1\}$ such that $|y_t - \hat{y}_t| \geq \frac{1}{2}$.

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

What happens for infinite classes?

- Consider the following **threshold functions**:

$$\mathcal{H}^{\text{thres}} := \{h_a(x) = 1\{x \geq a\} : a, x \in [0, 1]\}.$$

- For any learner Φ, consider the following strategy for Nature:
  - At every step t, select label $y_t \in \{0, 1\}$ such that $|y_t - \hat{y}_t| \geq \frac{1}{2}$.
  - Select instances from the set of dyadic rationals, starting with $\mathbf{x}_1 = \frac{1}{2}$ and updating (according to learner's prediction $\hat{y}_{t-1}$) as:

$$\mathbf{x}_t = \begin{cases} \mathbf{x}_{t-1} + \frac{1}{2^t}, & \text{if } \hat{y}_{t-1} \geq 0.5, \\ \mathbf{x}_{t-1} - \frac{1}{2^t}, & \text{else.} \end{cases}$$

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

What happens for infinite classes?

- Consider the following **threshold functions**:

$$\mathcal{H}^{\text{thres}} := \{h_a(x) = 1\{x \geq a\} : a, x \in [0,1]\}.$$

- For any learner Φ, consider the following strategy for Nature:
  - At every step t, select label $y_t \in \{0,1\}$ such that $|y_t - \hat{y}_t| \geq \frac{1}{2}$.
  - Select instances from the set of dyadic rationals, starting with $\mathbf{x}_1 = \frac{1}{2}$ and updating (according to learner's prediction $\hat{y}_{t-1}$) as:

$$\mathbf{x}_t = \begin{cases} \mathbf{x}_{t-1} + \frac{1}{2^t}, & \text{if } \hat{y}_{t-1} \geq 0.5, \\ \mathbf{x}_{t-1} - \frac{1}{2^t}, & \text{else.} \end{cases}$$

- This ensures that:
  - The cumulative loss incurred by the learner is at least $T/2$.

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

What happens for infinite classes?

- Consider the following **threshold functions**:

$$\mathcal{H}^{\text{thres}} := \{h_a(x) = 1\{x \geq a\} : a, x \in [0,1]\}.$$

- For any learner Φ, consider the following strategy for Nature:
  - At every step t, select label $y_t \in \{0,1\}$ such that $|y_t - \hat{y}_t| \geq \frac{1}{2}$.
  - Select instances from the set of dyadic rationals, starting with $\mathbf{x}_1 = \frac{1}{2}$ and updating (according to learner's prediction $\hat{y}_{t-1}$) as:

$$\mathbf{x}_t = \begin{cases} \mathbf{x}_{t-1} + \frac{1}{2^t}, & \text{if } \hat{y}_{t-1} \geq 0.5, \\ \mathbf{x}_{t-1} - \frac{1}{2^t}, & \text{else.} \end{cases}$$

- This ensures that:
  - The cumulative loss incurred by the learner is at least $T/2$.
  - For all $t \leq T$, $h_{\mathbf{x}_{T+1}}(\mathbf{x}_t) = y_t$, i.e., the hypothesis $h_{\mathbf{x}_{T+1}}$ incurs zero loss.

Beyond Finite Classes

Observe that the regret bound based on the EWA algorithm applies only to a finite class $\mathcal{H}$, and it depends solely on the class size.

What happens for infinite classes?

- Consider the following **threshold functions**:

$$\mathcal{H}^{\text{thres}} := \{h_a(x) = 1\{x \geq a\} : a, x \in [0,1]\}.$$

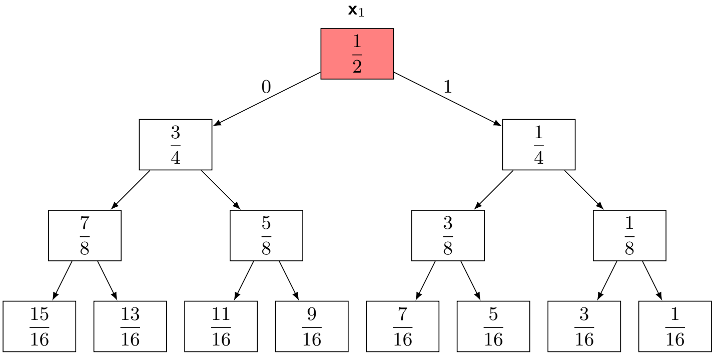
- For any learner Φ, consider the following strategy for Nature:
  - At every step t, select label $y_t \in \{0,1\}$ such that $|y_t - \hat{y}_t| \geq \frac{1}{2}$.
  - Select instances from the set of dyadic rationals, starting with $\mathbf{x}_1 = \frac{1}{2}$ and updating (according to learner's prediction $\hat{y}_{t-1}$) as:

$$\mathbf{x}_t = \begin{cases} \mathbf{x}_{t-1} + \frac{1}{2^t}, & \text{if } \hat{y}_{t-1} \geq 0.5, \\ \mathbf{x}_{t-1} - \frac{1}{2^t}, & \text{else.} \end{cases}$$

- This ensures that:
  - The cumulative loss incurred by the learner is at least $T/2$.
  - For all $t \leq T$, $h_{\mathbf{x}_{T+1}}(\mathbf{x}_t) = y_t$, i.e., the hypothesis $h_{\mathbf{x}_{T+1}}$ incurs zero loss.
  - Therefore, $\text{reg}_T(\mathcal{H}^{\text{thres}}) \geq T/2$.

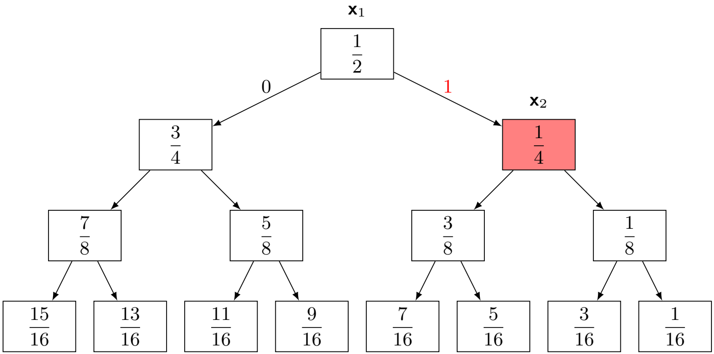
Demonstration of the Adversarial Process

Let learner's prediction be $\{0, 1, 1\}$, the strategy for Nature goes as follows:



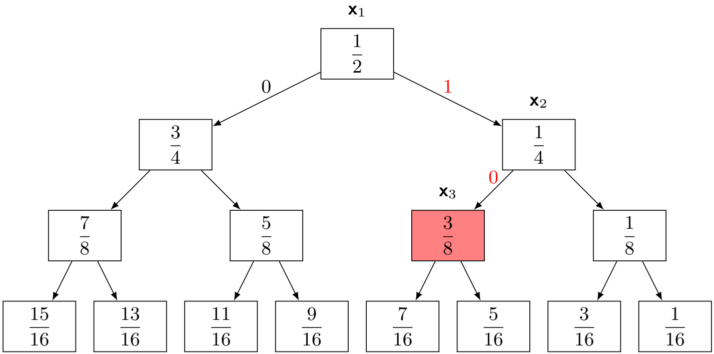
Demonstration of the Adversarial Process

Let learner's prediction be $\{0, 1, 1\}$, the strategy for Nature goes as follows:



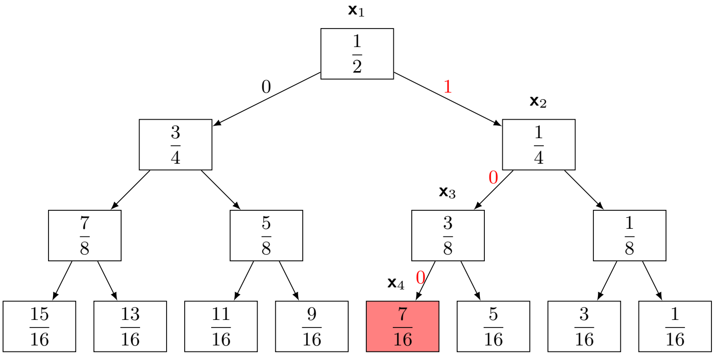
Demonstration of the Adversarial Process

Let learner's prediction be $\{0, 1, 1\}$, the strategy for Nature goes as follows:



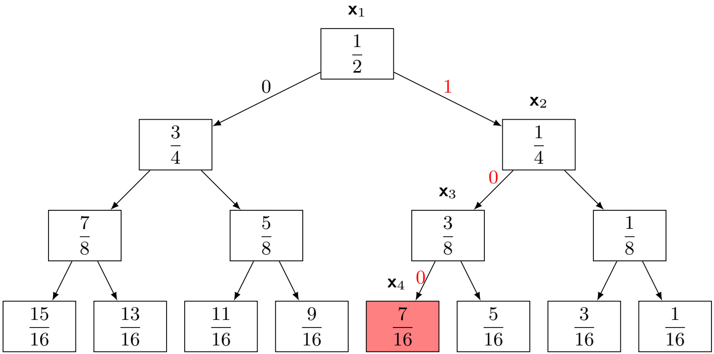
Demonstration of the Adversarial Process

Let learner's prediction be $\{0, 1, 1\}$, the strategy for Nature goes as follows:



Demonstration of the Adversarial Process

Let learner's prediction be $\{0, 1, 1\}$, the strategy for Nature goes as follows:



The function $h_{\mathbf{x}_4}(\mathbf{x}) := 1\{\mathbf{x} \geq \frac{7}{16}\}$ consistents with all true labels, but the learner errs at every step.

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

What intrinsic structure of $\mathcal{H}$ leads to this failure?

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

What intrinsic structure of $\mathcal{H}$ leads to this failure?

- Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued hypothesis class.

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

What intrinsic structure of $\mathcal{H}$ leads to this failure?

- Let $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ be any binary-valued hypothesis class.
- A $\mathcal{X}$-valued binary tree of depth d is defined as $\tau : \bigcup_{i \leq d} \{0, 1\}^i \to \mathcal{X}$.

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

What intrinsic structure of $\mathcal{H}$ leads to this failure?

- Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued hypothesis class.
- A $\mathcal{X}$-valued binary tree of depth d is defined as $\tau : \bigcup_{i \le d} \{0,1\}^i \to \mathcal{X}$.
- We say τ is shattered by $\mathcal{H}$ if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that

$$\forall i \le d, \; h(\tau(\epsilon^{i-1})) = \epsilon_i.$$

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

What intrinsic structure of $\mathcal{H}$ leads to this failure?

- Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued hypothesis class.
- A $\mathcal{X}$-valued binary tree of depth d is defined as $\tau : \bigcup_{i \leq d} \{0,1\}^i \to \mathcal{X}$.
- We say τ is shattered by $\mathcal{H}$ if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that

$$\forall i \leq d, \; h(\tau(\epsilon^{i-1})) = \epsilon_i.$$

- Note that, the tree formed by dyadic rationals is shattered by $\mathcal{H}^{\text{thres}}$.

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

What intrinsic structure of $\mathcal{H}$ leads to this failure?

- Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued hypothesis class.
- A $\mathcal{X}$-valued binary tree of depth d is defined as $\tau : \bigcup_{i \le d} \{0,1\}^i \to \mathcal{X}$.
- We say τ is shattered by $\mathcal{H}$ if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that
$$\forall i \le d, \; h(\tau(\epsilon^{i-1})) = \epsilon_i.$$
- Note that, the tree formed by dyadic rationals is shattered by $\mathcal{H}^{\text{thres}}$.

Fact 2: For any binary-valued class $\mathcal{H}$, if there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$, then: $\text{reg}_T(\mathcal{H}) \ge \frac{1}{2}\min\{d, T\}$.

The Shattering Trees

We have shown that even for simple threshold functions, achieving sublinear regret is not possible.

What intrinsic structure of $\mathcal{H}$ leads to this failure?

- Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued hypothesis class.
- A $\mathcal{X}$-valued binary tree of depth d is defined as $\tau : \bigcup_{i \leq d} \{0,1\}^i \to \mathcal{X}$.
- We say τ is shattered by $\mathcal{H}$ if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that

$$\forall i \leq d, \ h(\tau(\epsilon^{i-1})) = \epsilon_i.$$

- Note that, the tree formed by dyadic rationals is shattered by $\mathcal{H}^{\text{thres}}$.

Fact 2: For any binary-valued class $\mathcal{H}$, if there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$, then: $\text{reg}_T(\mathcal{H}) \geq \frac{1}{2}\min\{d, T\}$.

Proof: Select the labels opposite to learner's prediction, and the instances by following the shattering tree τ, similar to the threshold function case...

The Littlestone Dimension

Littlestone Dimension: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued hypothesis class. The *Littlestone dimension* of $\mathcal{H}$ is defined as the maximum number d such that there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$.

The Littlestone Dimension

Littlestone Dimension: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued hypothesis class. The *Littlestone dimension* of $\mathcal{H}$ is defined as the maximum number d such that there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$.

- We will denote $\mathsf{Ldim}(\mathcal{H})$ as the Littlestone dimension of $\mathcal{H}$.
- It is clear from our previous slides that $\mathsf{reg}_T(\mathcal{H}) \geq \frac{1}{2}\min\{\mathsf{Ldim}(\mathcal{H}), T\}$.
- Therefore, the Littlestone dimension forms an intrinsic barrier for the minimax regret.

The Littlestone Dimension

Littlestone Dimension: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued hypothesis class. The *Littlestone dimension* of $\mathcal{H}$ is defined as the maximum number d such that there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$.

- We will denote $\mathsf{Ldim}(\mathcal{H})$ as the Littlestone dimension of $\mathcal{H}$.
- It is clear from our previous slides that $\mathsf{reg}_T(\mathcal{H}) \geq \frac{1}{2}\min\{\mathsf{Ldim}(\mathcal{H}), T\}$.
- Therefore, the Littlestone dimension forms an intrinsic barrier for the minimax regret.

Example 1: For the threshold functions $\mathcal{H}^{\text{thres}}$, we have $\mathsf{Ldim}(\mathcal{H}^{\text{thres}}) = \infty$.

The Littlestone Dimension

Littlestone Dimension: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued hypothesis class. The *Littlestone dimension* of $\mathcal{H}$ is defined as the maximum number d such that there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$.

- We will denote $\mathsf{Ldim}(\mathcal{H})$ as the Littlestone dimension of $\mathcal{H}$.
- It is clear from our previous slides that $\mathsf{reg}_T(\mathcal{H}) \geq \frac{1}{2}\min\{\mathsf{Ldim}(\mathcal{H}), T\}$.
- Therefore, the Littlestone dimension forms an intrinsic barrier for the minimax regret.

Example 1: For the threshold functions $\mathcal{H}^{\text{thres}}$, we have $\mathsf{Ldim}(\mathcal{H}^{\text{thres}}) = \infty$.

Example 2: For any finite class $\mathcal{H}$, we have $\mathsf{Ldim}(\mathcal{H}) \leq \log|\mathcal{H}|$ (prove it!).

The Littlestone Dimension

Littlestone Dimension: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued hypothesis class. The *Littlestone dimension* of $\mathcal{H}$ is defined as the maximum number d such that there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$.

- We will denote $\mathsf{Ldim}(\mathcal{H})$ as the Littlestone dimension of $\mathcal{H}$.
- It is clear from our previous slides that $\mathsf{reg}_T(\mathcal{H}) \geq \frac{1}{2}\min\{\mathsf{Ldim}(\mathcal{H}), T\}$.
- Therefore, the Littlestone dimension forms an intrinsic barrier for the minimax regret.

Example 1: For the threshold functions $\mathcal{H}^{\mathsf{thres}}$, we have $\mathsf{Ldim}(\mathcal{H}^{\mathsf{thres}}) = \infty$.

Example 2: For any finite class $\mathcal{H}$, we have $\mathsf{Ldim}(\mathcal{H}) \leq \log|\mathcal{H}|$ (prove it!).

Example 3: Consider the following indicator functions

$$\mathcal{H}^{\mathsf{ind}} := \{h_a(x) := 1\{x = a\} : x, a \in [0,1]\}.$$

Then $\mathsf{Ldim}(\mathcal{H}^{\mathsf{ind}}) = 1$ (prove it!).

Upper Bounding Regret via Littlestone Dimension: Realizable case

We have shown that the Littlestone dimension forms a natural lower bound for the minimax regret. Can we achieve an upper bound as well?

Upper Bounding Regret via Littlestone Dimension: Realizable case

We have shown that the Littlestone dimension forms a natural lower bound for the minimax regret. Can we achieve an upper bound as well?

The Standard Optimal Algorithm (SOA):

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$, initially $\mathcal{H}^{(0)} = \mathcal{H}$.
2. At each time step t, we define, for $y \in \{0, 1\}$, that

$$\mathcal{H}_y^{(t)} = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = y\}.$$

3. Predict $\hat{y}_t := \arg\max_{y \in \{0,1\}} \{\mathrm{Ldim}(\mathcal{H}_y^{(t)}) : y \in \{0, 1\}\}$.
4. Let y_t be true label, update $\mathcal{H}^{(t)} = \mathcal{H}_{y_t}^{(t)}$.

Upper Bounding Regret via Littlestone Dimension: Realizable case

We have shown that the Littlestone dimension forms a natural lower bound for the minimax regret. Can we achieve an upper bound as well?

The Standard Optimal Algorithm (SOA):

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$, initially $\mathcal{H}^{(0)} = \mathcal{H}$.
2. At each time step t, we define, for $y \in \{0, 1\}$, that
$$\mathcal{H}_y^{(t)} = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = y\}.$$
3. Predict $\hat{y}_t := \arg\max_{y \in \{0,1\}} \{\mathsf{Ldim}(\mathcal{H}_y^{(t)}) : y \in \{0, 1\}\}$.
4. Let y_t be true label, update $\mathcal{H}^{(t)} = \mathcal{H}_{y_t}^{(t)}$.

Lemma 1: For any data $\mathbf{x}^T, y^T$ that is realizable w.r.t. a binary-valued class $\mathcal{H}$, i.e., $\exists h^* \in \mathcal{H}$ such that $\forall t \leq T$, $h^*(\mathbf{x}_t) = y_t$, the SOA predictor enjoys the following mistake bound

$$\sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\} \leq \mathsf{Ldim}(\mathcal{H}).$$

Upper Bounding Regret via Littlestone Dimension: Realizable case

We have shown that the Littlestone dimension forms a natural lower bound for the minimax regret. Can we achieve an upper bound as well?

The Standard Optimal Algorithm (SOA):

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$, initially $\mathcal{H}^{(0)} = \mathcal{H}$.
2. At each time step t, we define, for $y \in \{0, 1\}$, that

$$\mathcal{H}_y^{(t)} = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = y\}.$$

3. Predict $\hat{y}_t := \arg\max_{y \in \{0,1\}} \{\mathsf{Ldim}(\mathcal{H}_y^{(t)}) : y \in \{0, 1\}\}$.
4. Let y_t be true label, update $\mathcal{H}^{(t)} = \mathcal{H}_{y_t}^{(t)}$.

Lemma 1: For any data $\mathbf{x}^T, y^T$ that is realizable w.r.t. a binary-valued class $\mathcal{H}$, i.e., $\exists h^* \in \mathcal{H}$ such that $\forall t \leq T$, $h^*(\mathbf{x}_t) = y_t$, the SOA predictor enjoys the following mistake bound

$$\sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\} \leq \mathsf{Ldim}(\mathcal{H}).$$

Proof: Any mistake decreases Littlestone dimension by at least 1 (verify it!)...

The Sequential Covering

Sequential Cover: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, 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 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

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

The Sequential Covering

Sequential Cover: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, 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 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

$$\forall t \leq T, \;\; g(\mathbf{x}^t) = h(\mathbf{x}_t).$$

- The functions $g \in \mathcal{G}$ map finite sequences $\mathcal{X}^*$ of $\mathcal{X}$ to $\{0,1\}$.

The Sequential Covering

Sequential Cover: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, 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 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

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

- The functions $g \in \mathcal{G}$ map finite sequences $\mathcal{X}^*$ of $\mathcal{X}$ to $\{0,1\}$.
- The cover happens locally, depending on any given $\mathbf{x}^T$.
  - Unlike the classical uniform cover, where each h is covered by a fixed g.
  - Sequential cover allows the covering function g to depend on $\mathbf{x}^T$ as well.

The Sequential Covering

Sequential Cover: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, 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 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

$$\forall t \leq T, \; g(\mathbf{x}^t) = h(\mathbf{x}_t).$$

- The functions $g \in \mathcal{G}$ map finite sequences $\mathcal{X}^*$ of $\mathcal{X}$ to $\{0,1\}$.
- The cover happens locally, depending on any given $\mathbf{x}^T$.
  - Unlike the classical uniform cover, where each h is covered by a fixed g.
  - Sequential cover allows the covering function g to depend on $\mathbf{x}^T$ as well.
- Infinite classes $\mathcal{H}$ can be sequentially covered by a finite class $\mathcal{G}$.

The Sequential Covering

Sequential Cover: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, 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 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

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

- The functions $g \in \mathcal{G}$ map finite sequences $\mathcal{X}^*$ of $\mathcal{X}$ to $\{0,1\}$.
- The cover happens locally, depending on any given $\mathbf{x}^T$.
  - Unlike the classical uniform cover, where each h is covered by a fixed g.
  - Sequential cover allows the covering function g to depend on $\mathbf{x}^T$ as well.
- Infinite classes $\mathcal{H}$ can be sequentially covered by a finite class $\mathcal{G}$.
  - Consider the class $\mathcal{H}^{\text{ind}} := \{h_a(x) := 1\{x = a\} : x, a \in [0,1]\}$.

The Sequential Covering

Sequential Cover: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, 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 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

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

- The functions $g \in \mathcal{G}$ map finite sequences $\mathcal{X}^*$ of $\mathcal{X}$ to $\{0,1\}$.
- The cover happens locally, depending on any given $\mathbf{x}^T$.
  - Unlike the classical uniform cover, where each h is covered by a fixed g.
  - Sequential cover allows the covering function g to depend on $\mathbf{x}^T$ as well.
- Infinite classes $\mathcal{H}$ can be sequentially covered by a finite class $\mathcal{G}$.
  - Consider the class $\mathcal{H}^{\text{ind}} := \{h_a(x) := 1\{x = a\} : x, a \in [0,1]\}$.
  - For any $i \leq T$, define the sequential function:

$$g_i(\mathbf{x}^t) = \begin{cases} 1, & \text{if } t = i \\ 0, & \text{otherwise} \end{cases}.$$

The Sequential Covering

Sequential Cover: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, 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 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

$$\forall t \leq T, \; g(\mathbf{x}^t) = h(\mathbf{x}_t).$$

- The functions $g \in \mathcal{G}$ map finite sequences $\mathcal{X}^*$ of $\mathcal{X}$ to $\{0,1\}$.
- The cover happens locally, depending on any given $\mathbf{x}^T$.
  - Unlike the classical uniform cover, where each h is covered by a fixed g.
  - Sequential cover allows the covering function g to depend on $\mathbf{x}^T$ as well.
- Infinite classes $\mathcal{H}$ can be sequentially covered by a finite class $\mathcal{G}$.
  - Consider the class $\mathcal{H}^{\text{ind}} := \{h_a(x) := 1\{x = a\} : x, a \in [0,1]\}$.
  - For any $i \leq T$, define the sequential function:

$$g_i(\mathbf{x}^t) = \begin{cases} 1, & \text{if } t = i \\ 0, & \text{otherwise} \end{cases}.$$

  - The class $\mathcal{G} := \{g_i : i \in [T]\}$ sequentially covers $\mathcal{H}^{\text{ind}}$ (prove it!).

From Mistake Bound to Sequential Cover

Lemma 2: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued class. If there exists a predictor for $\mathcal{H}$ that achieves mistake bound err_T in the realizable case. Then there exists a sequential cover $\mathcal{G}$ of $\mathcal{H}$ up to step T such that

$$\log|\mathcal{G}| \leq \log \sum_{i=0}^{\mathrm{err}_T} \binom{T}{i} \leq O(\mathrm{err}_T \cdot \log T).$$

From Mistake Bound to Sequential Cover

Lemma 2: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued class. If there exists a predictor for $\mathcal{H}$ that achieves mistake bound err_T in the realizable case. Then there exists a sequential cover $\mathcal{G}$ of $\mathcal{H}$ up to step T such that

$$\log|\mathcal{G}| \leq \log \sum_{i=0}^{\mathsf{err}_T} \binom{T}{i} \leq O(\mathsf{err}_T \cdot \log T).$$

- Let Φ achieves err_T mistakes for $\mathcal{H}$ in the realizable case.

From Mistake Bound to Sequential Cover

Lemma 2: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued class. If there exists a predictor for $\mathcal{H}$ that achieves mistake bound err_T in the realizable case. Then there exists a sequential cover $\mathcal{G}$ of $\mathcal{H}$ up to step T such that

$$\log|\mathcal{G}| \leq \log \sum_{i=0}^{\mathsf{err}_T} \binom{T}{i} \leq O(\mathsf{err}_T \cdot \log T).$$

- Let Φ achieves err_T mistakes for $\mathcal{H}$ in the realizable case.
- For any $I \subset [T]$, we recursively define the sequential function

$$g_I(\mathbf{x}^t) = \begin{cases} \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), \text{ if } t \notin I \\ 1 - \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), \text{ if } t \in I \end{cases}.$$

From Mistake Bound to Sequential Cover

Lemma 2: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued class. If there exists a predictor for $\mathcal{H}$ that achieves mistake bound err_T in the realizable case. Then there exists a sequential cover $\mathcal{G}$ of $\mathcal{H}$ up to step T such that

$$\log|\mathcal{G}| \leq \log \sum_{i=0}^{\mathrm{err}_T} \binom{T}{i} \leq O(\mathrm{err}_T \cdot \log T).$$

- Let Φ achieves err_T mistakes for $\mathcal{H}$ in the realizable case.
- For any $I \subset [T]$, we recursively define the sequential function

$$g_I(\mathbf{x}^t) = \begin{cases} \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), \text{ if } t \notin I \\ 1 - \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), \text{ if } t \in I \end{cases}.$$

- The class $\mathcal{G} := \{g_I : I \subset [T],\ |I| \leq \mathrm{err}_T\}$ sequentially covers $\mathcal{H}$, since for any $\mathbf{x}^T$ and h we can pick I being the time steps where Φ errs...(why?)

From Mistake Bound to Sequential Cover

Lemma 2: Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued class. If there exists a predictor for $\mathcal{H}$ that achieves mistake bound err_T in the realizable case. Then there exists a sequential cover $\mathcal{G}$ of $\mathcal{H}$ up to step T such that

$$\log |\mathcal{G}| \le \log \sum_{i=0}^{\mathrm{err}_T} \binom{T}{i} \le O(\mathrm{err}_T \cdot \log T).$$

- Let Φ achieves err_T mistakes for $\mathcal{H}$ in the realizable case.
- For any $I \subset [T]$, we recursively define the sequential function

$$g_I(\mathbf{x}^t) = \begin{cases} \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), \text{ if } t \notin I \\ 1 - \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), \text{ if } t \in I \end{cases}.$$

- The class $\mathcal{G} := \{g_I : I \subset [T],\ |I| \le \mathrm{err}_T\}$ sequentially covers $\mathcal{H}$, since for any $\mathbf{x}^T$ and h we can pick I being the time steps where Φ errs...(why?)
- We have $|\mathcal{G}| \le \sum_{i=0}^{\mathrm{err}_T} \binom{T}{i}$ by counting the size of $\{I \subset [T] : |I| \le \mathrm{err}_T\}$.

Bounding Regret via Littlestone Dimension: Agnostic case

Theorem 2: For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$, the minimax regret of $\mathcal{H}$ satisfies

$$\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T}) \leq \mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T \log T}).$$

Bounding Regret via Littlestone Dimension: Agnostic case

Theorem 2: For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$, the minimax regret of $\mathcal{H}$ satisfies

$$\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T}) \leq \mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T \log T}).$$

- From our previous discussion (Lemma 1), we know that the class admits a mistake bound of $\mathsf{Ldim}(\mathcal{H})$ in the realizable case.

Bounding Regret via Littlestone Dimension: Agnostic case

Theorem 2: For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$, the minimax regret of $\mathcal{H}$ satisfies

$$\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T}) \leq \mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T \log T}).$$

- From our previous discussion (Lemma 1), we know that the class admits a mistake bound of $\mathsf{Ldim}(\mathcal{H})$ in the realizable case.
- This implies, by Lemma 2, a sequential cover $\mathcal{G}$ of size

$$\log |\mathcal{G}| \leq O(\mathsf{Ldim}(\mathcal{H}) \cdot \log T).$$

Bounding Regret via Littlestone Dimension: Agnostic case

Theorem 2: For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$, the minimax regret of $\mathcal{H}$ satisfies

$$\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T}) \leq \mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T \log T}).$$

- From our previous discussion (Lemma 1), we know that the class admits a mistake bound of $\mathsf{Ldim}(\mathcal{H})$ in the realizable case.
- This implies, by Lemma 2, a sequential cover $\mathcal{G}$ of size

$$\log |\mathcal{G}| \leq O(\mathsf{Ldim}(\mathcal{H}) \cdot \log T).$$

- Applying the EWA algorithm over $\mathcal{G}$ and using the property of sequential covering, we deduce, from Theorem 1, the upper bound $O(\sqrt{T \log |\mathcal{G}|})$.

Bounding Regret via Littlestone Dimension: Agnostic case

Theorem 2: For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with finite Littlestone dimension $\mathrm{Ldim}(\mathcal{H})$, the minimax regret of $\mathcal{H}$ satisfies

$$\Omega(\sqrt{\mathrm{Ldim}(\mathcal{H}) \cdot T}) \leq \mathrm{reg}_T(\mathcal{H}) \leq O(\sqrt{\mathrm{Ldim}(\mathcal{H}) \cdot T \log T}).$$

- From our previous discussion (Lemma 1), we know that the class admits a mistake bound of $\mathrm{Ldim}(\mathcal{H})$ in the realizable case.
- This implies, by Lemma 2, a sequential cover $\mathcal{G}$ of size

$$\log|\mathcal{G}| \leq O(\mathrm{Ldim}(\mathcal{H}) \cdot \log T).$$

- Applying the EWA algorithm over $\mathcal{G}$ and using the property of sequential covering, we deduce, from Theorem 1, the upper bound $O(\sqrt{T \log |\mathcal{G}|})$.
- We have shown a $\frac{1}{2}\min\{\mathrm{Ldim}(\mathcal{H}), T\}$ lower bound (c.f. Fact 2).

Bounding Regret via Littlestone Dimension: Agnostic case

Theorem 2: For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$, the minimax regret of $\mathcal{H}$ satisfies

$$\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T}) \leq \mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T \log T}).$$

- From our previous discussion (Lemma 1), we know that the class admits a mistake bound of $\mathsf{Ldim}(\mathcal{H})$ in the realizable case.
- This implies, by Lemma 2, a sequential cover $\mathcal{G}$ of size

$$\log |\mathcal{G}| \leq O(\mathsf{Ldim}(\mathcal{H}) \cdot \log T).$$

- Applying the EWA algorithm over $\mathcal{G}$ and using the property of sequential covering, we deduce, from Theorem 1, the upper bound $O(\sqrt{T \log |\mathcal{G}|})$.
- We have shown a $\frac{1}{2} \min\{\mathsf{Ldim}(\mathcal{H}), T\}$ lower bound (c.f. Fact 2). The lower bound $\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T})$ follows from a more technical argument...

Preparing for the Proof: Khinchine's Inequality

Khinchine's Inequality: Let $a_1, \cdots, a_T$ be real numbers and ϵ^T is uniformly distributed over $\{-1, +1\}^T$. Then

$$\mathbb{E}_{\epsilon^T}\left|\sum_{t=1}^{T} a_t \epsilon_t\right| \geq \frac{1}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} a_t^2}$$

Preparing for the Proof: Khinchine's Inequality

Khinchine's Inequality: Let $a_1, \cdots, a_T$ be real numbers and ϵ^T is uniformly distributed over $\{-1,+1\}^T$. Then

$$\mathbb{E}_{\epsilon^T}\left|\sum_{t=1}^{T} a_t \epsilon_t\right| \geq \frac{1}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} a_t^2}$$

Sketch of Proof: We give a short proof for sup-optimal constant $1/\sqrt{3}$.

Preparing for the Proof: Khinchine's Inequality

Khinchine's Inequality: Let $a_1, \cdots, a_T$ be real numbers and ϵ^T is uniformly distributed over $\{-1,+1\}^T$. Then

$$\mathbb{E}_{\epsilon^T}\left|\sum_{t=1}^{T} a_t\epsilon_t\right| \geq \frac{1}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} a_t^2}$$

Sketch of Proof: We give a short proof for sup-optimal constant $1/\sqrt{3}$. By Hölder's inequality, we have for any bounded random variable X

$$\mathbb{E}[X^2] = \mathbb{E}[|X|^{4/3}|X|^{2/3}] \leq (\mathbb{E}[X^4])^{1/3}(\mathbb{E}[|X|])^{2/3}.$$

Preparing for the Proof: Khinchine's Inequality

Khinchine's Inequality: Let $a_1, \cdots, a_T$ be real numbers and ϵ^T is uniformly distributed over $\{-1,+1\}^T$. Then

$$\mathbb{E}_{\epsilon^T}\left|\sum_{t=1}^{T} a_t\epsilon_t\right| \geq \frac{1}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} a_t^2}$$

Sketch of Proof: We give a short proof for sup-optimal constant $1/\sqrt{3}$. By Hölder's inequality, we have for any bounded random variable X

$$\mathbb{E}[X^2] = \mathbb{E}[|X|^{4/3}|X|^{2/3}] \leq (\mathbb{E}[X^4])^{1/3}(\mathbb{E}[|X|])^{2/3}.$$

Taking $X = \sum_{t=1}^{T} a_t\epsilon_t$, we have

$$\mathbb{E}_{\epsilon^T}\left|\sum_{t=1}^{T} a_t\epsilon_t\right| \geq \frac{(\sum_{t=1}^{T} a_t^2)^{3/2}}{\sqrt{\sum_{t=1}^{T} a_t^4 + 3\sum_{i\neq j} a_i^2 a_j^2}} \overset{(\star)}{\geq} \frac{1}{\sqrt{3}}\sqrt{\sum_{t=1}^{T} a_t^2},$$

Preparing for the Proof: Khinchine's Inequality

Khinchine's Inequality: Let $a_1, \cdots, a_T$ be real numbers and ϵ^T is uniformly distributed over $\{-1,+1\}^T$. Then

$$\mathbb{E}_{\epsilon^T}\left|\sum_{t=1}^{T} a_t\epsilon_t\right| \geq \frac{1}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} a_t^2}$$

Sketch of Proof: We give a short proof for sup-optimal constant $1/\sqrt{3}$. By Hölder's inequality, we have for any bounded random variable X

$$\mathbb{E}[X^2] = \mathbb{E}[|X|^{4/3}|X|^{2/3}] \leq (\mathbb{E}[X^4])^{1/3}(\mathbb{E}[|X|])^{2/3}.$$

Taking $X = \sum_{t=1}^{T} a_t\epsilon_t$, we have

$$\mathbb{E}_{\epsilon^T}\left|\sum_{t=1}^{T} a_t\epsilon_t\right| \geq \frac{(\sum_{t=1}^{T} a_t^2)^{3/2}}{\sqrt{\sum_{t=1}^{T} a_t^4 + 3\sum_{i\neq j} a_i^2 a_j^2}} \overset{(\star)}{\geq} \frac{1}{\sqrt{3}}\sqrt{\sum_{t=1}^{T} a_t^2},$$

where $(\star)$ follows by $\sum_{t=1}^{T} a_t^4 + 3\sum_{i\neq j} a_i^2 a_j^2 \leq 3(\sum_{t=1}^{T} a_t^2)^2$.

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$.

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}_t := \mathbf{x}$ for all $t \leq T$.

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}_t := \mathbf{x}$ for all $t \leq T$.

We have for any prediction rule Φ that $\mathbb{E}_{y^T}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] = \frac{T}{2}$.

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}_t := \mathbf{x}$ for all $t \leq T$.

We have for any prediction rule Φ that $\mathbb{E}_{y^T}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] = \frac{T}{2}$.

Let k be the number of 1's in y^T. We have

$$\inf_{h \in \{h_0, h_1\}} \sum_{t=1}^{T} |h(\mathbf{x}) - y_t| = \min\{k, T - k\}.$$

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}_t := \mathbf{x}$ for all $t \leq T$.

We have for any prediction rule Φ that $\mathbb{E}_{y^T}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] = \frac{T}{2}$.

Let k be the number of 1's in y^T. We have

$$\inf_{h \in \{h_0, h_1\}} \sum_{t=1}^{T} |h(\mathbf{x}) - y_t| = \min\{k, T-k\}.$$

Let ϵ^T be uniform over $\{\pm 1\}^T$, we have $\sum_{t=1}^{T} \epsilon_t$ distributed equally as $2k - T$.

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}_t := \mathbf{x}$ for all $t \leq T$.

We have for any prediction rule Φ that $\mathbb{E}_{y^T}\left[\sum_{t=1}^T |\hat{y}_t - y_t|\right] = \frac{T}{2}$.

Let k be the number of 1's in y^T. We have

$$\inf_{h\in\{h_0,h_1\}} \sum_{t=1}^{T} |h(\mathbf{x}) - y_t| = \min\{k, T-k\}.$$

Let ϵ^T be uniform over $\{\pm 1\}^T$, we have $\sum_{t=1}^T \epsilon_t$ distributed equally as $2k - T$.

Note that $|k - \frac{T}{2}| = \frac{T}{2} - \min\{k, T-k\}$, we have by Khinchine's Inequality that

$$\mathbb{E}[\min\{k, T-k\}] \leq \frac{T}{2} - \frac{1}{\sqrt{8}}\sqrt{T}.$$

Proof of Lower Bound

We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$.

Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}_t := \mathbf{x}$ for all $t \leq T$.

We have for any prediction rule Φ that $\mathbb{E}_{y^T}\left[\sum_{t=1}^T |\hat{y}_t - y_t|\right] = \frac{T}{2}$.

Let k be the number of 1's in y^T. We have

$$\inf_{h \in \{h_0, h_1\}} \sum_{t=1}^{T} |h(\mathbf{x}) - y_t| = \min\{k, T-k\}.$$

Let ϵ^T be uniform over $\{\pm 1\}^T$, we have $\sum_{t=1}^T \epsilon_t$ distributed equally as $2k - T$.

Note that $|k - \frac{T}{2}| = \frac{T}{2} - \min\{k, T-k\}$, we have by Khinchine's Inequality that

$$\mathbb{E}[\min\{k, T-k\}] \leq \frac{T}{2} - \frac{1}{\sqrt{8}}\sqrt{T}.$$

Therefore, the regret is lower bounded by $\sqrt{T/8}$.

Proof of Lower Bound

The $\Omega(\sqrt{\text{Ldim}(\mathcal{H})\,T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$.

Proof of Lower Bound

The $\Omega(\sqrt{\text{Ldim}(\mathcal{H})\,T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$.

Assume that T is divisible by $\text{Ldim}(\mathcal{H})$ (otherwise we truncate T).

Proof of Lower Bound

The $\Omega(\sqrt{\mathrm{Ldim}(\mathcal{H})\,T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$.

Assume that T is divisible by $\mathrm{Ldim}(\mathcal{H})$ (otherwise we truncate T).

We partition $\mathbf{x}^T, y^T$ into $\mathrm{Ldim}(\mathcal{H})$ blocks each of size $\frac{T}{\mathrm{Ldim}(\mathcal{H})}$, and denote k_i be the number of 1's in the i'th block of y^T.

Proof of Lower Bound

The $\Omega(\sqrt{\text{Ldim}(\mathcal{H})T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$.

Assume that T is divisible by $\text{Ldim}(\mathcal{H})$ (otherwise we truncate T).

We partition $\mathbf{x}^T, y^T$ into $\text{Ldim}(\mathcal{H})$ blocks each of size $\frac{T}{\text{Ldim}(\mathcal{H})}$, and denote k_i be the number of 1's in the i'th block of y^T.

Let τ be a $\mathcal{X}$-valued binary tree of depth $\text{Ldim}(\mathcal{H})$ that can be shattered by $\mathcal{H}$.

Proof of Lower Bound

The $\Omega(\sqrt{\text{Ldim}(\mathcal{H})T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$.

Assume that T is divisible by $\text{Ldim}(\mathcal{H})$ (otherwise we truncate T).

We partition $\mathbf{x}^T, y^T$ into $\text{Ldim}(\mathcal{H})$ blocks each of size $\frac{T}{\text{Ldim}(\mathcal{H})}$, and denote k_i be the number of 1's in the i'th block of y^T.

Let τ be a $\mathcal{X}$-valued binary tree of depth $\text{Ldim}(\mathcal{H})$ that can be shattered by $\mathcal{H}$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}^T$ by traversing τ:

1. We assign the same value within each block of $\mathbf{x}^T$, with the first block being the value of the root v_0 of τ.
2. Let v_i be the node in τ for the i's block. If $k_i \geq \frac{T}{2\text{Ldim}(\mathcal{H})}$ we set v_{i+1} being left child of v_i, and set to the right child otherwise.

Proof of Lower Bound

The $\Omega(\sqrt{\text{Ldim}(\mathcal{H})T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$.

Assume that T is divisible by $\text{Ldim}(\mathcal{H})$ (otherwise we truncate T).

We partition $\mathbf{x}^T, y^T$ into $\text{Ldim}(\mathcal{H})$ blocks each of size $\frac{T}{\text{Ldim}(\mathcal{H})}$, and denote k_i be the number of 1's in the i'th block of y^T.

Let τ be a $\mathcal{X}$-valued binary tree of depth $\text{Ldim}(\mathcal{H})$ that can be shattered by $\mathcal{H}$.

We now select y^T uniformly over $\{0, 1\}^T$ and select $\mathbf{x}^T$ by traversing τ:

1. We assign the same value within each block of $\mathbf{x}^T$, with the first block being the value of the root v_0 of τ.
2. Let v_i be the node in τ for the i's block. If $k_i \geq \frac{T}{2\text{Ldim}(\mathcal{H})}$ we set v_{i+1} being left child of v_i, and set to the right child otherwise.

By definition of shattering, $\exists h \in \mathcal{H}$ that achieves $\min\{k_i, \frac{T}{\text{Ldim}(\mathcal{H})} - k_i\}$ losses for all i simultaneously. (verify it!)

Proof of Lower Bound

The $\Omega(\sqrt{\mathrm{Ldim}(\mathcal{H})T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$.

Assume that T is divisible by $\mathrm{Ldim}(\mathcal{H})$ (otherwise we truncate T).

We partition $\mathbf{x}^T, y^T$ into $\mathrm{Ldim}(\mathcal{H})$ blocks each of size $\frac{T}{\mathrm{Ldim}(\mathcal{H})}$, and denote k_i be the number of 1's in the i'th block of y^T.

Let τ be a $\mathcal{X}$-valued binary tree of depth $\mathrm{Ldim}(\mathcal{H})$ that can be shattered by $\mathcal{H}$.

We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}^T$ by traversing τ:

1. We assign the same value within each block of $\mathbf{x}^T$, with the first block being the value of the root v_0 of τ.
2. Let v_i be the node in τ for the i's block. If $k_i \geq \frac{T}{2\mathrm{Ldim}(\mathcal{H})}$ we set v_{i+1} being left child of v_i, and set to the right child otherwise.

By definition of shattering, $\exists h \in \mathcal{H}$ that achieves $\min\{k_i, \frac{T}{\mathrm{Ldim}(\mathcal{H})} - k_i\}$ losses for all i simultaneously. (verify it!)

The regret is then lower bounded by

$$\Omega(\mathrm{Ldim}(\mathcal{H}) \cdot \sqrt{T/\mathrm{Ldim}(\mathcal{H})}) = \Omega(\sqrt{\mathrm{Ldim}(\mathcal{H})T}).$$

Overview

- **Minimax Regret**
  - Pointwise, worst-case, and minimax regrets
  - The iterative minimax formulation
- **Bounding the Minimax Regret: Binary Labels**
  - The Littlestone dimension
  - Standard Optimal Algorithm
  - Sequential covering
- **The Minimax Theorem**
  - Proving minimax theorem via EWA algorithm

The Minimax Theorem

Minimax Theorem: Let $f : A \times B \to \mathbb{R}$ be a bounded real-valued function, where both A and B are convex sets and A is compact. If $f(\cdot, b)$ is convex and continuous on A for any $b \in B$, and $f(a, \cdot)$ is concave on B for any $a \in A$, then

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{b \in B} \inf_{a \in A} f(a, b).$$

The Minimax Theorem

Minimax Theorem: Let $f : A \times B \to \mathbb{R}$ be a bounded real-valued function, where both A and B are convex sets and A is compact. If $f(\cdot, b)$ is convex and continuous on A for any $b \in B$, and $f(a, \cdot)$ is concave on B for any $a \in A$, then

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{b \in B} \inf_{a \in A} f(a, b).$$

- This theorem is stronger than von Neumann's minimax theorem, which specifically considers the case when f is a bi-linear function.
- It differs slightly from Sion's minimax theorem, which requires only semi-continuity and quasi-convexity (-concavity).

The Minimax Theorem

Minimax Theorem: Let $f : A \times B \to \mathbb{R}$ be a bounded real-valued function, where both A and B are convex sets and A is compact. If $f(\cdot, b)$ is convex and continuous on A for any $b \in B$, and $f(a, \cdot)$ is concave on B for any $a \in A$, then

$$\inf_{a \in A} \sup_{b \in B} f(a, b) = \sup_{b \in B} \inf_{a \in A} f(a, b).$$

- This theorem is stronger than von Neumann's minimax theorem, which specifically considers the case when f is a bi-linear function.
- It differs slightly from Sion's minimax theorem, which requires only semi-continuity and quasi-convexity (-concavity).

Interpretation: In a two-player game with actions from A and B, the minimax theorem shows that, under the stated conditions, player 1's best strategy yields the same value whether or not they know player 2's move.

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ for any f (why?).

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ for any f (why?).

For converse, by compactness of A, there exists a finite ϵ-net $A'_\epsilon \subset A$ of size N.

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ for any f (why?).

For converse, by compactness of A, there exists a finite ϵ-net $A'_\epsilon \subset A$ of size N.

We now view $\hat{\mathcal{Y}} := A$ as the prediction space and $\mathcal{Y} := B$ as the label space.

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a, b) \geq \sup_b \inf_a f(a, b)$ for any f (why?).

For converse, by compactness of A, there exists a finite ϵ-net $A'_\epsilon \subset A$ of size N.

We now view $\hat{\mathcal{Y}} := A$ as the prediction space and $\mathcal{Y} := B$ as the label space.

The set A'_ϵ is therefore a hypothesis class with constant-valued functions (i.e., with no features)

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ for any f (why?).

For converse, by compactness of A, there exists a finite ϵ-net $A'_\epsilon \subset A$ of size N.

We now view $\hat{\mathcal{Y}} := A$ as the prediction space and $\mathcal{Y} := B$ as the label space.

The set A'_ϵ is therefore a hypothesis class with constant-valued functions (i.e., with no features), and $f(a,b)$ is a loss function.

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ for any f (why?).

For converse, by compactness of A, there exists a finite ϵ-net $A'_\epsilon \subset A$ of size N.

We now view $\hat{\mathcal{Y}} := A$ as the prediction space and $\mathcal{Y} := B$ as the label space.

The set A'_ϵ is therefore a hypothesis class with constant-valued functions (i.e., with no features), and $f(a,b)$ is a loss function.

Let $\Phi : \mathcal{Y}^* \to \hat{\mathcal{Y}}$ be the (generalized) EWA algorithm with no feature inputs.

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ for any f (why?).

For converse, by compactness of A, there exists a finite ϵ-net $A'_\epsilon \subset A$ of size N.

We now view $\hat{\mathcal{Y}} := A$ as the prediction space and $\mathcal{Y} := B$ as the label space.

The set A'_ϵ is therefore a hypothesis class with constant-valued functions (i.e., with no features), and $f(a,b)$ is a loss function.

Let $\Phi : \mathcal{Y}^* \to \hat{\mathcal{Y}}$ be the (generalized) EWA algorithm with no feature inputs.

Consider the following strategy for Nature: $\forall t \leq T$, choose $y_t \in \mathcal{Y}$ such that

$$f(\hat{y}_{t-1}, y_t) \geq \sup_{y \in \mathcal{Y}} f(\hat{y}_{t-1}, y) - \frac{1}{T},$$

where $\hat{y}_{t-1} = \Phi(y^{t-1})$ is learner's prediction using EWA.

Proof of Minimax Theorem via the EWA algorithm

It is obvious that $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ for any f (why?).

For converse, by compactness of A, there exists a finite ϵ-net $A'_\epsilon \subset A$ of size N.

We now view $\hat{\mathcal{Y}} := A$ as the prediction space and $\mathcal{Y} := B$ as the label space.

The set A'_ϵ is therefore a hypothesis class with constant-valued functions (i.e., with no features), and $f(a,b)$ is a loss function.

Let $\Phi : \mathcal{Y}^* \to \hat{\mathcal{Y}}$ be the (generalized) EWA algorithm with no feature inputs.

Consider the following strategy for Nature: $\forall t \leq T$, choose $y_t \in \mathcal{Y}$ such that

$$f(\hat{y}_{t-1}, y_t) \geq \sup_{y \in \mathcal{Y}} f(\hat{y}_{t-1}, y) - \frac{1}{T},$$

where $\hat{y}_{t-1} = \Phi(y^{t-1})$ is learner's prediction using EWA.

By the regret guarantee for EWA (Theorem 1), we have:

$$\frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y_t) \leq \inf_{\hat{y} \in A'_\epsilon} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}, y_t) + O(\sqrt{\frac{\log N}{T}}).$$

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) \leq \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t, y\right)$$

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$\begin{aligned}
\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) &\leq \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t, y\right) \\
&\leq \sup_{y} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y), \text{ by convexity of } f(\cdot, y)
\end{aligned}$$

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$\begin{aligned}
\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) &\leq \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t, y\right) \\
&\leq \sup_{y} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y), \text{ by convexity of } f(\cdot, y) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y_t) + \frac{1}{T}, \text{ by definition of } y_t
\end{aligned}$$

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$\begin{aligned}
\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) &\le \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t, y\right) \\
&\le \sup_{y} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y), \text{ by convexity of } f(\cdot, y) \\
&\le \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y_t) + \frac{1}{T}, \text{ by definition of } y_t \\
&\le \inf_{\hat{y}\in A'_\epsilon} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}, y_t) + O\left(\sqrt{\frac{\log N}{T}}\right), \text{ by regret bound of EWA}
\end{aligned}$$

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$
\begin{aligned}
\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) &\leq \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t, y\right) \\
&\leq \sup_{y} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y), \text{ by convexity of } f(\cdot, y) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y_t) + \frac{1}{T}, \text{ by definition of } y_t \\
&\leq \inf_{\hat{y}\in A'_\epsilon} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}, y_t) + O\left(\sqrt{\frac{\log N}{T}}\right), \text{ by regret bound of EWA} \\
&\leq \inf_{\hat{y}\in A'_\epsilon} f\left(\hat{y}, \frac{1}{T}\sum_{t=1}^{T} y_t\right) + O\left(\sqrt{\frac{\log N}{T}}\right) \text{ by concavity of } f(\hat{y}, \cdot)
\end{aligned}
$$

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$
\begin{aligned}
\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) &\leq \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t, y\right) \\
&\leq \sup_{y} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y), \text{ by convexity of } f(\cdot, y) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y_t) + \frac{1}{T}, \text{ by definition of } y_t \\
&\leq \inf_{\hat{y}\in A'_\epsilon} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}, y_t) + O\left(\sqrt{\frac{\log N}{T}}\right), \text{ by regret bound of EWA} \\
&\leq \inf_{\hat{y}\in A'_\epsilon} f\left(\hat{y}, \frac{1}{T}\sum_{t=1}^{T} y_t\right) + O\left(\sqrt{\frac{\log N}{T}}\right) \text{ by concavity of } f(\hat{y}, \cdot) \\
&\leq \sup_{y} \inf_{\hat{y}\in A'_\epsilon} f(\hat{y}, y) + O(\sqrt{\log N/T}).
\end{aligned}
$$

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$\begin{aligned}
\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) &\leq \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t, y\right) \\
&\leq \sup_{y} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y), \text{ by convexity of } f(\cdot, y) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y_t) + \frac{1}{T}, \text{ by definition of } y_t \\
&\leq \inf_{\hat{y}\in A'_\epsilon} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}, y_t) + O\left(\sqrt{\frac{\log N}{T}}\right), \text{ by regret bound of EWA} \\
&\leq \inf_{\hat{y}\in A'_\epsilon} f\left(\hat{y}, \frac{1}{T}\sum_{t=1}^{T} y_t\right) + O\left(\sqrt{\frac{\log N}{T}}\right) \text{ by concavity of } f(\hat{y}, \cdot) \\
&\leq \sup_{y} \inf_{\hat{y}\in A'_\epsilon} f(\hat{y}, y) + O(\sqrt{\log N/T}).
\end{aligned}$$

Sending $T \to \infty$, we have $\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) \leq \sup_{y} \inf_{\hat{y}\in A'_\epsilon} f(\hat{y}, y)$.

Proof of Minimax Theorem via the EWA algorithm

Observe that

$$
\begin{aligned}
\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) &\le \sup_{y} f\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t, y\right) \\
&\le \sup_{y} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y), \text{ by convexity of } f(\cdot, y) \\
&\le \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}_t, y_t) + \frac{1}{T}, \text{ by definition of } y_t \\
&\le \inf_{\hat{y}\in A'_\epsilon} \frac{1}{T}\sum_{t=1}^{T} f(\hat{y}, y_t) + O\left(\sqrt{\frac{\log N}{T}}\right), \text{ by regret bound of EWA} \\
&\le \inf_{\hat{y}\in A'_\epsilon} f\left(\hat{y}, \frac{1}{T}\sum_{t=1}^{T} y_t\right) + O\left(\sqrt{\frac{\log N}{T}}\right) \text{ by concavity of } f(\hat{y}, \cdot) \\
&\le \sup_{y} \inf_{\hat{y}\in A'_\epsilon} f(\hat{y}, y) + O(\sqrt{\log N / T}).
\end{aligned}
$$

Sending $T \to \infty$, we have $\inf_{\hat{y}} \sup_{y} f(\hat{y}, y) \le \sup_{y} \inf_{\hat{y}\in A'_\epsilon} f(\hat{y}, y)$. The theorem follows by sending $\epsilon \to 0$ and continuity of $f(\cdot, y)$, since $A'_\epsilon \subset A$ is an ϵ-net.

Concluding Remarks

- In this lecture, we discussed the minimax regret of online learning games by focusing on the structure of the hypothesis class.
- We demonstrate that the Littlestone dimension tightly characterizes the minimax regret for binary-valued classes.
- Most of the techniques can be extended to real-valued classes, but need more care to get it right. This will be discussed in the upcoming lecture.