Online Minimax Regret Bounded by Sequential Rademacher Complexity

Yuantong Li

Purdue University

li3551@purdue.edu

Sep 26, 2019

Online Learning Model

1. *Online Learning Model*:
   Let $\mathcal{F}$ be a class of functions and $\mathcal{X}$ some set. The Online Learning Model is defined as the following T-round interaction between the learner and the adversary: On round $t = 1, ..., T$, the learner chooses $f_t \in \mathcal{F}$, the adversary picks $x_t \in \mathcal{X}$, and the learner suffers loss $f_t(x_t)$. At the end of T rounds we define *regret*

$$\mathbf{R}\left(f_{1:T}, x_{1:T}\right) = \sum_{t=1}^{T} f_t\left(x_t\right) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f\left(x_t\right)$$

   as the difference between the cumulative loss of the player as compared to the cumulative loss of the best fixed comparator.

2. *Online learnable*:
   For the given pair $(\mathcal{F}, \mathcal{X})$, the problem is said to be online learnable if there exists an algorithm for the learner such that regret grows sublinearly.

Value of the Game

Theorem 1

Let $\mathcal{F}$ and $\mathcal{X}$ be the sets of moves for the two players, satisfying the necessary conditions for the minimax theorem to hold. Denote by $\mathcal{Q}$ and $\mathcal{P}$ the sets of probability distributions (mixed strategies) on $\mathcal{F}$ and $\mathcal{X}$, respectively. Then

$$\begin{aligned}\mathcal{V}_T(\mathcal{F},\mathcal{X}) &= \inf_{q_1\in\mathcal{Q}} \sup_{x_1\in\mathcal{X}} \mathbb{E}_{f_1\sim q_1} \inf_{q_T\in\mathcal{Q}} \sup_{x_T\in\mathcal{X}} \mathbb{E}_{f_T\sim q_T} \left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t)\right] \\ &= \sup_{p_1} \mathbb{E}_{x_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T\sim p_T} \left[\sum_{t=1}^{T} \inf_{f_t\in\mathcal{F}} \mathbb{E}_{x_t\sim p_t}[f_t(x_t)] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t)\right]\end{aligned}$$

- $\mathcal{F}$: is a subset of a separable metric space.
- $\mathcal{Q}$: the set of probability distributions on $\mathcal{F}$.
- p_t: the distribution on x_t

Some definitions

Definition 1. Online Learnable (Formal)

A class $\mathcal{F}$ is said to be online learnable with respect to the given $\mathcal{X}$ if

$$\limsup_{T\to\infty} \frac{\mathcal{V}_T(\mathcal{F},\mathcal{X})}{T} = 0$$

Since complexity of $\mathcal{F}$ is the focus, we shall often write $\mathcal{V}(\mathcal{F})$, and the dependence on $\mathcal{X}$ will be implicit.

Definition 2. Sequential Rademacher Complexity (SRC)

The Sequential Rademacher Complexity of a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is defined as

$$\mathfrak{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t f\left(\mathbf{x}_t(\epsilon)\right) \right]$$

where the outer supremum is taken over all $\mathcal{X}$-valued trees of depth T and $\epsilon = (\epsilon_1, \dots, \epsilon_T)$ is a sequence of i.i.d. Rademacher random variables.

Upper bound by SRC

Theorem 2

The minimax value of a randomized game is bounded as

$$\mathcal{V}_T(\mathcal{F}) \leq 2\mathfrak{R}_T(\mathcal{F}) \tag{1}$$

Proof:

From Theorem (1),

$$\begin{aligned}
\mathcal{V}_T(\mathcal{F}) &= \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[\sum_{t=1}^{T} \inf_{f_t \in \mathcal{F}} \mathbb{E}_{x_t \sim p_t} \left[f_t\left(x_t\right) \right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f\left(x_t\right) \right] \\
&= \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} \inf_{f_t \in \mathcal{F}} \mathbb{E}_{x_t \sim p_t} \left[f_t\left(x_t\right) \right] - \sum_{t=1}^{T} f\left(x_t\right) \right\} \right] \\
&\leq \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{x_t \sim p_t} \left[f\left(x_t\right) \right] - \sum_{t=1}^{T} f\left(x_t\right) \right\} \right]
\end{aligned} \tag{2}$$

The upper bound is obtained by replacing each infimum by a particular choice f.

Proof:

Renaming variables,

$$\mathcal{V}_T(\mathcal{F}) = \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{x_i' \sim p_t} \left[f\left(x_t'\right) \right] - \sum_{t=1}^{T} f\left(x_t\right) \right\} \right]$$

$$\leq_1 \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[\mathbb{E}_{x_1' \sim p_1} \dots \mathbb{E}_{x_T' \sim p_T} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} f\left(x_t'\right) - \sum_{t=1}^{T} f\left(x_t\right) \right\} \right]$$

$$\leq_2 \sup_{p_1} \mathbb{E}_{x_1, x_1' \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T, x_T' \sim p_T} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} f\left(x_t'\right) - \sum_{t=1}^{T} f\left(x_t\right) \right\} \right]$$

where the last two steps are using jensen inequality for the supremum.

Quick Lemma: sup is a convex function

Proof: Let $(g_i)_{i \in I}$ be a family of convex functions on a convex compact set Ω. Let $g := \sup_{i \in I} g_i$. Take $x, y \in \Omega$ and $t \in [0, 1]$. Fix $i \in I$, since g_i is convex and bounded by g, we have

$$g_i(tx + (1-t)y) \leq t g_i(x) + (1-t) g_i(y) \leq t g(x) + (1-t) g(y)$$

holds for all g_i, so g is convex.

Proof of (1)

Proof: We take $T = 2$ as an illustration,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left[\mathbb{E}_{x_1'\sim p_1}[f(x_1')] + \mathbb{E}_{x_2'\sim p_2}[f(x_2')]\right] \\
&= \sup_{f\in\mathcal{F}}\left[\mathbb{E}_{x_1'\sim p_1}\mathbb{E}_{x_2'\sim p_2}[f(x_1') + f(x_2')]\right] \\
&\leq \mathbb{E}_{x_1'\sim p_1}\sup_{f\in\mathcal{F}}\left[\mathbb{E}_{x_2'\sim p_2}[f(x_1') + f(x_2')]\right] \\
&\leq \mathbb{E}_{x_1'\sim p_1}\mathbb{E}_{x_2'\sim p_2}\sup_{f\in\mathcal{F}}\left[\sum_{t=1}^{T} f(x_i')\right]
\end{aligned}$$

Proof of (2)

$$\begin{aligned}
&\sup_{p_1}\mathbb{E}_{x_1\sim p_1}\sup_{p_2}\mathbb{E}_{x_2\sim p_2}\left[\mathbb{E}_{x_1'\sim p_1}\mathbb{E}_{x_2'\sim p_2}\sup_{f\in\mathcal{F}} f(x_1, x_1', x_2, x_2')\right] \\
&= \sup_{p_1}\mathbb{E}_{x_1\sim p_1}\sup_{p_2}\mathbb{E}_{x_1'\sim p_1}\left[\mathbb{E}_{x_2\sim p_2}\mathbb{E}_{x_2'\sim p_2}\sup_{f\in\mathcal{F}} f(x_1, x_1', x_2, x_2')\right] \\
&\leq \sup_{p_1}\mathbb{E}_{x_1\sim p_1}\mathbb{E}_{x_1'\sim p_1}\sup_{p_2}\mathbb{E}_{x_2\sim p_2}\mathbb{E}_{x_2'\sim p_2}\left[\sup_{f\in\mathcal{F}} f(x_1, x_1', x_2, x_2')\right]
\end{aligned}$$

Continue proof of Theorem 2

By the Key Technical Lemma (See Lemma 1 below) with $\phi(u) = u$ and $\Delta_f(x_t, x_t') = f(x_t') - f(x_t)$,

$$\sup_{p_1} \mathbb{E}_{x_1, x_1' \sim p_1} \ldots \sup_{p_T} \mathbb{E}_{x_T, x_T' \sim p_T} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} f(x_t') - f(x_t) \right\} \right]$$

$$\leq \sup_{x_1, x_1'} \left\{ \mathbb{E}_{\epsilon_1} \left[\ldots \sup_{x_T, x_T'} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t \left(f(x_t') - f(x_t) \right) \right] \right\} \cdots \right] \right\}$$

Thus,

$$\mathcal{V}_T(\mathcal{F}) \leq \sup_{x_1, x_1'} \left\{ \mathbb{E}_{\epsilon_1} \left[\cdots \sup_{x_T, x_T'} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t \left(f(x_t') - f(x_t) \right) \right] \right\} \cdots \right] \right\}$$

$$\leq \sup_{x_1, x_1'} \left\{ \mathbb{E}_{\epsilon_1} \left[\cdots \sup_{x_T, x_T'} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} \epsilon_t f(x_t') \right\} + \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} -\epsilon_t f(x_t) \right\} \right] \right\} \cdots \right] \right\}$$

$$= 2 \sup_{x_1} \left\{ \mathbb{E}_{\epsilon_1} \left[\ldots \sup_{x_T} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{T} \epsilon_t f(x_t) \right\} \right] \right\} \cdots \right] \right\}$$

Lemma 1

Key Technical Lemma

Let $(x_1, \ldots, x_T) \in \mathcal{X}^T$ be a sequence distributed according to **D** and let $(x'_1, \ldots, x'_T) \in \mathcal{X}^T$ be a tangent sequence. Let $\Delta_f(x_t, x'_t)$ be a functional $\mathcal{F} \mapsto \mathbb{R}$ such that

$$\Delta_f(x_t, x'_t) = -\Delta_f(x'_t, x_t)$$

Let $\Phi(\Omega) = \phi\left(\sup_{f\in\mathcal{F}} \Omega(f)\right)$ or $\Phi(\Omega) = \phi\left(\sup_{f\in\mathcal{F}} |\Omega(f)|\right)$, where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is some measurable real valued function and $\Omega : \mathcal{F} \mapsto \mathbb{R}$. Then

$$\sup_{p_1} \mathbb{E}_{x_1,x'_1\sim p_1} \ldots \sup_{p_T} \mathbb{E}_{x_T,x'_T\sim p_T} \left[\Phi\left(\sum_{t=1}^{T} \Delta_f(x_t, x'_t)\right)\right]$$
$$\leq \sup_{x_1,x'_1} \left\{\mathbb{E}_{\epsilon_1}\left[\ldots \sup_{x_T,x'_T} \left\{\mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T} \epsilon_t \Delta_f(x_t, x'_t)\right)\right]\right\}\ldots\right]\right\}$$

Continue proof of Theorem 2

Now, we need to move the suprema over x_t's outside. This is achieved via an idea similar to skolemization in logic. We basically exploit the identity

$$\mathbb{E}_{\epsilon_{1:t-1}}\left[\sup_{x_t} G\left(\epsilon_{1:t-1}, x_t\right)\right] = \sup_{\mathbf{x}_t} \mathbb{E}_{\epsilon_{1:t-1}}\left[G\left(\epsilon_{1:t-1}, \mathbf{x}_t\left(\epsilon_{1:t-1}\right)\right)\right] \tag{3}$$

[*Proof of (3) on next page*] Use this identity once, we get,

$$\mathcal{V}_T(\mathcal{F}) \leq 2 \sup_{x_1, \mathbf{x}_2}\left\{\mathbb{E}_{\epsilon_1,\epsilon_2}\left[\sup_{x_3} \ldots \sup_{x_T}\left\{\mathbb{E}_{\epsilon_T}\left[\sup_{f \in \mathcal{F}}\left\{\epsilon_1 f\left(x_1\right) + \epsilon_2 f\left(\mathbf{x}_2\left(\epsilon_1\right)\right) + \sum_{t=3}^{T} \epsilon_t f\left(x_t\right)\right\}\right]\right\}\right]\right\}$$

Now, we apply this identity $T-2$ more times to successively move the supremums over $x_3, \ldots, x_T$ out side, to get

$$\begin{aligned}\mathcal{V}_T(\mathcal{F}) &\leq 2 \sup_{x_1, \mathbf{x}_2, \ldots, \mathbf{x}_T} \mathbb{E}_{\epsilon_1, \ldots, \epsilon_T}\left[\sup_{f \in \mathcal{F}}\left\{\epsilon_1 f\left(x_1\right) + \sum_{t=2}^{T} \epsilon_t f\left(\mathbf{x}_t\left(\epsilon_{1:t-1}\right)\right)\right\}\right] \\ &= 2 \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_1, \ldots, \epsilon T}\left[\sup_{f \in \mathcal{F}}\left\{\sum_{t=1}^{T} \epsilon_t f\left(\mathbf{x}_t(\epsilon)\right)\right\}\right]\end{aligned}$$

Proof of equation (3)

We take $t = 2$ as an illustration

$$\mathbb{E}_{\epsilon_1} \sup_{x_2} G(\epsilon_1, x_2) = \sup_{\mathbf{x}_2(\epsilon_1)} \mathbb{E}_{\epsilon_1} G(\epsilon_1, \mathbf{x}_2(\epsilon_1))$$

Proof of the Key Technical Lemma (1)

We start by noting that since x_T, x'_T are both drawn from p_T,

$$\begin{aligned}
\mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T}\Delta_f\left(x_t,x'_t\right)\right)\right] &= \mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f\left(x_t,x'_t\right)+\Delta_f\left(x_T,x'_T\right)\right)\right]\\
&= \mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f\left(x_t,x'_t\right)+\Delta_f\left(x'_T,x_T\right)\right)\right]\\
&= \mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f\left(x_t,x'_t\right)-\Delta_f\left(x_T,x'_T\right)\right)\right]
\end{aligned}$$

Since the first line and last lines are equal, they are both equal to their average and hence

$$\mathbb{E}_{x_T,x'_T}\sim p_T\left[\Phi\left(\sum_{t=1}^{T}\Delta_f\left(x_t,x'_t\right)\right)\right]=$$

$$\mathbb{E}_{x_T,x'_T\sim p_T}\left[\mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f\left(x_t,x'_t\right)+\epsilon_T\Delta_f\left(x_T,x'_T\right)\right)\right]\right]$$

Proof of the Key Technical Lemma (2)

Hence we conclude

$$\begin{aligned}
\sup_{p_T} \mathbb{E}_{x_T, x'_T} \sim p_T & \left[\Phi \left(\sum_{t=1}^{T} \Delta_f \left(x_t, x'_t \right) \right) \right] \\
&= \sup_{p_T} \mathbb{E}_{x_T, x'_T \sim p_T} \left[\mathbb{E}_{\epsilon_T} \left[\Phi \left(\sum_{t=1}^{T-1} \Delta_f \left(x_t, x'_t \right) + \epsilon_T \Delta_f \left(x_T, x'_T \right) \right) \right] \right] \\
&\leq \sup_{x_T, x'_T} \mathbb{E}_{\epsilon_T} \left[\Phi \left(\sum_{t=1}^{T-1} \Delta_f \left(x_t, x'_t \right) + \epsilon_T \Delta_f \left(x_T, x'_T \right) \right) \right]
\end{aligned}$$

Using the above and noting that x_{T-1}, x'_{T-1} are both from p_{T-1} and hence similar to previous step introducing Rademacher variable ϵ_{T-1} we get that

$$\begin{aligned}
& \sup_{p_{T-1}} \mathbb{E}_{x_{T-1}, x'_{T-1} \sim p_{T-1}} \sup_{p_T} \mathbb{E}_{x_T, x'_T \sim p_T} \left[\Phi \left(\sum_{t=1}^{T} \Delta_f \left(x_t, x'_t \right) \right) \right] \\
&\leq \sup_{x_{T-1}, x'_{T-1}} \mathbb{E}_{\epsilon_{T-1}} \left[\sup_{x_T, x'_T} \mathbb{E}_{\epsilon_T} \left[\Phi \left(\sum_{t=1}^{T-2} \Delta_f \left(x_t, x'_t \right) + \sum_{t=T-1}^{T} \epsilon_t \Delta_f (x_t, x'_t) \right) \right] \right]
\end{aligned}$$