From Convex Analysis to Learning, Prediction, and Elicitation[*]
Lecture 4: No-Regret Online Learning

Lunjia Hu

In this lecture, we apply what we have learned about the minimax theorem to obtain a no-regret algorithm for a famous online learning problem: *experts problem*, as well as its generalization to *Online Linear Optimization (OLO)*. The experts problem and OLO are a fundamental building block for solving many other online learning problems that we will discuss later in the course (e.g., online calibration and Blackwell's approachability).

1 The Classic Experts Problem

Here is the classic setup of the experts problem. As the learner, our goal is to learn from d experts, labeled $1, \ldots, d$. In each round $t = 1, \ldots, T$, our task is to choose a probability distribution $x_t \in \Delta_d$ over the experts. After that, the loss incurred by each expert is revealed as a vector $y_t \in \{-1, 1\}^d$, where the i-th coordinate of y_t is the loss incurred by expert i in round t. The loss we incur in round t is $\langle x_t, y_t \rangle \in [-1, 1]$.

Our goal is to minimize the regret R compared to the best distribution $x^* \in \Delta_d$ in hindsight:

$$R(x_{1,\ldots,T}, y_{1,\ldots,T}) := \sum_{t=1}^{T} \langle x_t, y_t \rangle - \min_{x^* \in \Delta_d} \sum_{t=1}^{T} \langle x^*, y_t \rangle. \tag{1}$$

Clearly, the regret for T rounds is at most T. Can we do substantially better than that and develop a strategy that guarantees $o(T)$ regret? Is that even possible?

The challenge in the experts problem is that the learner needs to choose the distribution $x_t \in \Delta_d$ *before* y_t is revealed. Moreover, we do not make any assumptions on how the y_t's are chosen—we assume that an *adversary* chooses $y_1, \ldots, y_T$, and we even assume that the adversary knows the learner's strategy. This is a common challenge in many online learning problems. At every round t, the learner cannot see the outcome y_t before taking an action or making a prediction. Therefore, in order to achieve a low regret, the learner needs to *learn* from the previous outcomes $y_1, \ldots, y_{t-1}$.

We will show, somewhat surprisingly, that it is possible to ensure $O(\sqrt{T \log d})$ regret in T rounds. The idea will be to use the minimax theorem to switch the order of play. The intuition is that if the adversary plays first and reveals y_t or the distribution of y_t before the learner has to choose x_t, then it is straightforward to find the best response x_t that minimizes the expected regret.

To formally state the result, we need to formalize the strategy space of the learner and the adversary. The learner's strategy L consists of functions $\ell_1, \ldots, \ell_T$, where each ℓ_t maps the history $h_{t-1} := (y_1, \ldots, y_{t-1})$ to some $x_t = \ell_t(h_{t-1}) \in \Delta_d$. The adversary's strategy is simply $A =$

[*] `https://lunjiahu.com/convex-analysis/`

$(y_1, \dots, y_T) \in (\{-1,1\}^d)^T$. The entire sequence $x_1, y_1, \dots, x_T, y_T$ is determined by L and A, where $y_1, \dots, y_T$ are immediately given by A, and each x_t is given by $x_t = \ell(h_{t-1}) = \ell((y_1, \dots, y_{t-1}))$. We define

$$R(L, A) := R(x_{1,\dots,T}, y_{1,\dots,T}).$$

We use $\mathcal{L}$ to denote the set of all possible learner's strategies $L = (\ell_1, \dots, \ell_T)$, and use $\mathcal{A} = (\{-1,1\}^d)^T$ to denote the set of all possible adversary's strategies.

Our main theorem shows that the learner has a strategy that guarantees $O(\sqrt{T \log d})$ regret against any adversary:

Theorem 1. *There exists a learner's strategy $L \in \mathcal{L}$ such that*

$$\max_{A \in \mathcal{A}} R(L, A) = O(\sqrt{T \log d}).$$

As mentioned earlier, we prove Theorem 1 by considering an order-reversed game where the adversary chooses a (randomized) strategy first, in which case it becomes easy for the learner to achieve low expected regret:

Lemma 2. *Let Π be an arbitrary distribution over $\mathcal{A}$. There exists a learner's strategy $L \in \mathcal{L}$ such that*

$$\mathbb{E}_{A \sim \Pi} R(L, A) = O(\sqrt{T \log d}).$$

Proof. Consider $A = (y_1, \dots, y_T) \in (\{-1,1\}^d)^T = \mathcal{A}$ drawn from the distribution Π. For $t = 1, \dots, T$, we define μ_t as the conditional expectation of y_t given the history $y_1, \dots, y_{t-1}$:

$$\mu_t := \mathbb{E}_\Pi[y_t | y_1, \dots, y_{t-1}].$$

Note that μ_t is a (deterministic) function of the history $h_{t-1} := (y_1, \dots, y_{t-1})$. We construct our learner's strategy $L = (\ell_1, \dots, \ell_T)$ simply by defining $\ell_t(h_{t-1}) := \arg\min_{x \in \Delta_d} \langle x, \mu_t \rangle$ for every $t = 1, \dots, T$, where we break ties arbitrarily.

Given $A = (y_1, \dots, y_T) \sim \Pi$, our learner's strategy L gives the corresponding $x_1, \dots, x_T \in \Delta_d$, where

$$x_t = \ell(h_{t-1}) = \arg\min_{x \in \Delta_d} \langle x, \mu_t \rangle \quad \text{for every } t = 1, \dots, T. \tag{2}$$

The lemma is proved by the following calculation:

$$\begin{aligned}
\mathbb{E}_{A \sim \Pi} R(L, A) &= \mathbb{E}_\Pi \left[\sum_{t=1}^T \langle x_t, y_t \rangle - \min_{x^* \in \Delta_d} \sum_{t=1}^T \langle x^*, y_t \rangle \right] \\
&= \mathbb{E} \left[\max_{x^* \in \Delta_d} \left(\sum_{t=1}^T \langle x_t, \mu_t \rangle - \sum_{t=1}^T \langle x^*, y_t \rangle \right) \right] \\
&\le \mathbb{E} \left[\max_{x^* \in \Delta_d} \left(\sum_{t=1}^T \langle x^*, \mu_t \rangle - \sum_{t=1}^T \langle x^*, y_t \rangle \right) \right] && \text{(by (2))} \\
&\le O(\sqrt{T \log d}). && \text{(by Lemma 3)}
\end{aligned}$$

□

The above proof of Lemma 2 uses the following concentration inequality.

Lemma 3. *Let* Π *be an arbitrary distribution of* $(y_1, \ldots, y_T) \in ([-1,1]^d)^T$. *Define*

$$\mu_t := \mathbb{E}_\Pi[y_t | y_1, \ldots, y_{t-1}].$$

Then

$$\mathbb{E}_{Y\sim\Pi}\left[\max_{x\in\Delta_d}\sum_{t=1}^{T}\langle x, \mu_t - y_t\rangle\right] \leq O(\sqrt{T\log d}).$$

This lemma can be proved based on the fact that $z_s := \sum_{t=1}^{s}(\mu_t - y_t)$ is a martingale. We omit the detailed proof of this lemma since concentration inequalities are not the focus of this course.

Proof of Theorem 1. By Lemma 2, we have

$$\max_{\Pi\in\Delta_{\mathcal{A}}}\min_{L\in\mathcal{L}}\mathbb{E}_{A\sim\Pi}[R(L,A)] = O(\sqrt{T\log d}).$$

Note that $\mathcal{L}$ and $\Delta_{\mathcal{A}}$ are both compact convex sets, and $\mathbb{E}_{A\sim\Pi}[R(L,A)]$ is an affine function of $L \in \mathcal{L}$ and $\Pi \in \Delta_{\mathcal{A}}$, so by the minimax theorem, we have

$$\min_{L\in\mathcal{L}}\max_{\Pi\in\Delta_{\mathcal{A}}}\mathbb{E}_{A\sim\Pi}[R(L,A)] = O(\sqrt{T\log d}).$$

This means that there exists $L \in \mathcal{L}$ such that

$$\max_{A\in\mathcal{A}} R(L,A) = O(\sqrt{T\log d}). \qquad \square$$

Remark 1. *While Theorem 1 proves the existence of a learner's strategy L that achieves a good regret guarantee, it does not explicitly tell us how to construct such a strategy. Later in the course we will see a concrete strategy achieving the same* $O(\sqrt{T\log d})$ *regret guarantee using the* multiplicative weights *algorithm, which is a special case of* mirror descent*.*

2 Generalizations of the Experts Problem: Online Linear Optimization

In the classic experts problem shown in the previous section, in each round, the learner chooses an action $x_t \in \Delta_d$, whereas the adversary chooses an action $y_t \in \{-1,1\}^d$. It is natural to consider other choices of the two action spaces Δ_d and $\{-1,1\}^d$. Such generalized experts problems obtained from general action spaces are termed Online Linear Optimization (OLO) in the literature.

Specifically, an OLO problem is the following generalized experts problem defined by the learner's action set $X \subseteq \mathbb{R}^d$ and the adversary's action set $Y \subseteq \mathbb{R}^d$: in each round $t = 1, \ldots, T$,

1. learner chooses $x_t \in X$;

2. adversary reveals $y_t \in Y$;

3. learner incurs loss $\langle x_t, y_t\rangle$.

The regret of the learner after T rounds is defined analogously to (1):

$$R(x_{1,\dots,T}, y_{1,\dots,T}) := \sum_{t=1}^{T} \langle x_t, y_t \rangle - \inf_{x^* \in X} \sum_{t=1}^{T} \langle x^*, y_t \rangle. \tag{3}$$

Similarly to Theorem 1, we can prove regret bounds for OLOs. We summarize some important examples in Table 1. These examples will become useful for our discussions about online calibration and Blackwell's approachability later in the course. In Table 1, we use $\bar{B}_{\ell_2}(\mathbf{0}, 1)$ to denote the closed unit ℓ_2-ball $\{v \in \mathbb{R}^d : \|v\|_2 \le 1\}$ and use $\bar{B}_{\ell_1}(\mathbf{0}, 1)$ to denote the closed unit ℓ_1-ball $\{v \in \mathbb{R}^d : \|v\|_1 \le 1\}$.

Regret bounds for OLOs (including the ones in Table 1) can be proved in the same way as we prove Theorem 1 in Section 1. In the following, we give a high-level explanation for how these general regret bounds are proved.

Learner's action space	Adversary's action space	Regret bound
$x_t \in \Delta_d$	$y_t \in [-1, 1]^d$	$O(\sqrt{T \log d})$
$x_t \in \bar{B}_{\ell_2}(\mathbf{0}, 1)$	$y_t \in \bar{B}_{\ell_2}(\mathbf{0}, 1)$	$O(\sqrt{T})$
$x_t \in [-1, 1]^d$	$y_t \in \bar{B}_{\ell_1}(\mathbf{0}, 1)$	$O(\sqrt{Td})$

Table 1: Regret bounds for Online Linear Optimization problems.

First, we note that our entire proof of Theorem 1 in Section 1 is still valid if we replace $y_t \in \{-1, 1\}^d$ with $y_t \in Y$ for an arbitrary finite subset $Y \subseteq [-1, 1]^d$. Our analysis requires $S \subseteq [-1, 1]^d$ because Lemma 3 requires $y_t \in [-1, 1]^d$. We need Y to be finite because in our proof of Theorem 1, we apply the minimax theorem to distributions Π on Y^T, and we would like Π to be a compact convex set in a linear space with finite dimension. Our analysis does not need any additional properties of Y beyond these two requirements.

Second, the finiteness assumption on Y can be removed, and the same $O\sqrt{T \log d})$ regret bound in Theorem 1 holds even if we allow each y_t to be chosen arbitrarily from $[-1, 1]^d$ (first line of Table 1), despite that $[-1, 1]^d$ is an infinite set. This is achieved by applying Theorem 1 to a finite ε-cover Y of $[-1, 1]^d$ for a sufficiently small $\varepsilon > 0$.

Third, we can apply the same proof strategy to get regret bounds when the learner's action space $X \in \mathbb{R}^d$ and adversary's action space $Y \in \mathbb{R}^d$ are general, as long as X is a compact convex set and Y is bounded. The different regret bounds simply come from different variants of the concentration inequality Lemma 3. For instance, the second regret bound in Table 1 comes from the following concentration inequality:

Lemma 4. *Let Π be an arbitrary distribution of $(y_1, \dots, y_T) \in (\bar{B}_{\ell_2}(\mathbf{0}, 1))^T$. Define*

$$\mu_t := \mathbb{E}_{\Pi}[y_t | y_1, \dots, y_{t-1}].$$

Then

$$\mathbb{E}_{Y \sim \Pi}\left[\max_{x \in \bar{B}_{\ell_2}(\mathbf{0}, 1)} \sum_{t=1}^{T} \langle x, \mu_t - y_t \rangle\right] \le O(\sqrt{T}).$$