CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Zhaolin Ren

Lecture 17
November 26, 2018

1 Review and Overview

In the last few lectures, we have been studying the problem of online convex optimization (OCO). In particular, we have been working under the assumption of full access to the loss function at every iteration. In this lecture, we introduce the partial observation/bandit setting, where we do not have full access to the loss function.

To recap, recall the set-up of (OCO).

In each iteration $t = 1 \dots T$:

- Player picks an action $\omega_t \in \Omega$, where Ω is a convex set
- Environment picks a convex function $f_t : \Omega \mapsto [0, 1]$
- Player observes some information about f_t

Our goal is to choose ω_t to minimize regret:

$$\text{regret} \triangleq \mathbb{E}\left[\sum_{t=1}^{T} f_t(\omega_t) - \min_{\omega\in\Omega} \sum_{t=1}^{T} f_t(\omega)\right],$$

where the expectation is taken (potentially) over the randomness of both the algorithm and the loss functions $\{f_t\}_{t=1}^{T}$. The OCO framework can vary in its settings. One way in which it can differ is in the power of **observations**.

Power of observations. Player can observe different amounts of information on f_t.

- Full-information setting: player observes the entire loss function f_t
- Bandit/partial observation setting: player only observes $f_t(\omega_t)$

In this lecture, we will introduce examples of bandit OCO, and show how one might solve it by reducing to the full-information setting. Before we continue, note that the OCO framework can also vary in the power of **environment**. Below are two possible environment settings.

Power of environment. The environment can choose the f_t with different levels of adaptivity and adversarialism:

- Stochastic setting: $f_t, \dots, f_T$ are i.i.d samples from some distribution P. Under this setting, the envirommment is not adversarial.
- Oblivious setting: $f_1, \dots, f_T$ are chosen before the game starts. Under this setting, the environment can be adversarial, but not adaptive.

The lack of adversarialism means that the stochastic setting is easier to study than the oblivious setting. As such, in analyzing bandit problems, we will begin with the harder problem of working under the oblivious setting.

2 Multi-armed bandit problem (oblivious setting)

We begin by studying the multi-armed bandit problem, which is essentially the expert problem with partial feedback. This is a special case of the partial observation/bandit OCO.

2.1 Setup

The setup is as follows. Suppose there are N experts, and that $\ell_t \in [0,1]^N$.

In each iteration $t = 1, \ldots, T$:

- player plays action $a_t \in [N]$ based on some probability distribution p_t, such that $a_t = a$ with probability $p_t(a)$
- player receives loss $\ell_t(a_t) \in \mathbb{R}$

The difference with the expert problem is that here we only have access to the loss $\ell_t(a_t)$ for the expert we chose at time t, instead of access to the entire loss function ℓ_t. We assume also that we are in the oblivious setting, where $\ell_1, \ldots, \ell_T$ are chosen before the game starts.

The regret in this problem is as follows:

$$\begin{aligned}\text{regret}_{\text{bandit}} &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \arg\min_{a\in[N]} \sum_{t=1}^{T} \ell_t(a) \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, \ell_t\rangle\right] - \text{argmin}_{a\in[N]} \sum_{t=1}^{T} \ell_t(a),\end{aligned}$$

where the expectation is taken over the randomness of our algorithm.

2.2 Exploration vs Exploitation Tradeoff

At a high level, in contrast with the full-information setting, in the partial observation setting, there is a tradeoff between exploitation and exploration.

For example, suppose we pick expert $a_t = 1$ for the first few iterations, and the loss $\ell_t(1)$ appears small. How do we know if expert 1 has really done well so far in comparison with the other experts? In the full-information setting, we can compare the performance of expert 1 against the performance of the other experts so far. In the partial observation setting, we do not have this luxury, and hence there is a need to explore more experts to see if better experts exist. This creates a need for **exploration**.

However, if other experts are worse-off, exploration might lead to more losses. Therefore, at some point, we need to **exploit** the best expert(s) we have seen so far.

2.3 Reduction to expert problem

We will show that we can in fact reduce the multi-armed bandit problem to the expert problem. One natural idea is to estimate the functions ℓ_t from the information we have, namely $\ell_t(a_t)$.

2.3.1 Estimation of ℓ_t

In each iteration $t = 1, \ldots, T$:

- Player picks expert a_t from some distribution p_t
- Based on p_t, consider the following estimator, $\hat{\ell}_t$, where

$$\hat{\ell}(a) = \frac{\ell_t(a)}{p_t(a)} \mathbb{1}(a = a_t) = \begin{cases} \dfrac{\ell_t(a_t)}{p_t(a_t)} & \text{if } a = a_t \\ 0 & \text{if } a \neq a_t \end{cases}$$

As an aside, the above estimator $\hat{\ell}_t$ is essentially a form of importance sampling. We will next show that conditional on knowing p_t, the estimator $\hat{\ell}_t$ is unbiased.

Lemma 2.1. *Suppose $p_t(a) > 0$ for every $a \in [N]$. Then,*

$$\mathbb{E}\left[\hat{\ell}_t \mid p_t\right] = \ell_t.$$

As a natural consequence, taking expectation over the randomness of the algorithm,

$$\mathbb{E}\left[\hat{\ell}_t\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{\ell}_t \mid p_t\right]\right] = \ell_t.$$

Proof. For any $a \in [N]$,

$$\begin{aligned} \mathbb{E}[\hat{\ell}_t(a) \mid p_t] &= \mathbb{P}\left[a_t = a\right] \mathbb{E}\left[\hat{\ell}_t(a) \mid a_t = a\right] + \mathbb{P}\left[a_t \neq a\right] \mathbb{E}\left[\hat{\ell}_t(a) \mid a_t \neq a\right] \\ &= p_t(a) \cdot \frac{\ell_t(a)}{p_t(a)} + (1 - p_t(a)) \cdot 0 \\ &= \ell_t(a) \end{aligned}$$

□

Building on Lemma 2.1, we next introduce a result that allows us to interchange ℓ_t and $\hat{\ell}_t$ when analyzing the regret. This will be useful later.

Lemma 2.2. *Taking expectation over the randomness of the algorithm, we have*

$$\mathbb{E}\left[\langle p_t, \hat{\ell}_t \rangle\right] = \mathbb{E}\left[\langle p_t, \ell_t \rangle\right]$$

Proof. The result can be proven using a combination of the law of iterated expectations and Lemma 2.1. Observe that

$$\begin{aligned} \mathbb{E}\left[\langle p_t, \hat{\ell}_t \rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle p_t, \hat{\ell}_t \rangle \mid p_t\right]\right] \\ &= \mathbb{E}\left[\langle p_t, \mathbb{E}\left[\hat{\ell}_t \mid p_t\right] \rangle\right] \\ &= \mathbb{E}\left[\langle p_t, \ell_t \rangle\right], \end{aligned}$$

where the first step is by law of iterated expectations. Note that the outer expectation is with respect to the first $t - 1$ iterations of the algorithm, which determine p_t. The second step uses the fact that conditional on p_t, we can take p_t out of the conditional expectation term, and the last step is by Lemma 2.1. □

2.3.2 Passing $\hat{\ell}_t$ to the expert problem algorithm

Recall that in the original expert problem, there was an assumption that the losses ℓ_t was bounded in the range $[0,1]^N$. Suppose now that $\hat{\ell}_t(a)$ is bounded in the range $[0,1]^N$. This allows us to call the algorithm $\mathcal{A}$ for the expert problem with the estimator $\hat{\ell}_t$.

Let $\text{regret}_{\text{bandit}}(a)$ denote the regret, compared to action a, in the bandit problem, and let $\text{regret}_{\text{expert}}(a)$ denote the regret, compared to action a, in the expert problem. Then, if we plug in the estimator $\hat{\ell}_t$ to the algorithm for the expert problem, we have that

$$\begin{aligned}
\text{regret}_{\text{expert}}(a) &\triangleq \mathbb{E}\left[\sum_{t=1}^{T}\langle\hat{\ell}_t, p_t\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_t(a)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle\ell_t, p_t\rangle\right] - \sum_{t=1}^{T}\ell_t(a) \\
&\triangleq \text{regret}_{\text{bandit}}(a),
\end{aligned}$$

where in moving from the first to the second line we used Lemma 2.1 and Lemma 2.2.

Hence, if we pass in the estimator $\hat{\ell}_t$ to the algorithm $\mathcal{A}$ for the expert problem, minimizing regret for the bandit problem is equivalent to minimizing regret for the expert problem.

Caveat: There is however one caveat to the above analysis - it is possible that $\hat{\ell}_t \notin [0,1]^N$.

Fix: To fix this issue, we need to make $\hat{\ell}_t$ bounded, and then rescale it such that it stays in the range $[0,1]^N$. To this end, we mix in the uniform distribution to form the probability distribution p_t.

In each iteration $t = 1, \ldots, T$:

- Let $\hat{p}_t$ be the distribution of actions suggested by the expert problem algorithm $\mathcal{A}$.
- Set $p_t = (1-\alpha)\hat{p}_t + \alpha\dfrac{\mathbf{1}}{N}$, for some $0 < \alpha < 1$ that we choose later. Note that here $\mathbf{1}$ denotes the N-dimensional vector with all ones, so $\mathbf{1}/N$ denotes the uniform distribution.
- As a consequence, $p_t(a) \geq \dfrac{\alpha}{N}$.
- Since $\ell_t \in [0,1]^N$, it follows that

$$\begin{aligned}
\hat{\ell}_t(a) &= \frac{\ell_t(a)}{p_t(a)}\mathbb{1}(a = a_t) \\
&\leq \frac{1}{p_t(a)} \\
&\leq \frac{N}{\alpha}, \quad \text{since } p_t(a) \geq \frac{\alpha}{N}.
\end{aligned}$$

- This implies then that

$$\frac{\alpha}{N}\hat{\ell}_t \in [0,1]^N$$

We have resolved the issue of $\hat{\ell}_t \notin [0,1]^N$ at a cost: now, for any $a \in [N]$, we only know that $\hat{\ell}_t(a) \leq N/\alpha$, where $N/\alpha > 1$, while the true loss ℓ_t satisfies $\ell_t \in [0,1]^N$.

2.3.3 Analysis of regret after reduction

Let $\mathcal{A}$ be the algorithm used by the expert problem to output the optimal ω_t at every iteration t, which we detailed in the previous lecture. By the reduction outlined earlier, we have the following algorithm for the bandit problem:

In each iteration $t = 1, \ldots, T$:

- Call $\mathcal{A}$ to get a probability distribution $\hat{p}_t$
- Set $p_t = (1-\alpha) \cdot \hat{p}_t + \alpha \cdot \dfrac{\mathbf{1}}{N}$
- Compute $\hat{\ell}_t$ using p_t
- Feed $(\alpha/N)\hat{\ell}_t$ to $\mathcal{A}$, where we note that $(\alpha/N)\hat{\ell}_t \in [0,1]^N$.

Next, similar to in Lemma 2.2 earlier, we can show another result allowing us to interchange ℓ_t and $\hat{\ell}_t$ that is useful.

Lemma 2.3. *Taking expectation over randomness of the algorithm,*

$$\mathbb{E}\left[\langle \hat{p}_t, \ell_t \rangle\right] = \mathbb{E}\left[\langle \hat{p}_t, \hat{\ell}_t \rangle\right]$$

Proof.

$$\begin{aligned}
\mathbb{E}\langle \hat{p}_t, \hat{\ell}_t \rangle &= \mathbb{E}\left[\mathbb{E}\left[\langle \hat{p}_t, \hat{\ell}_t \rangle \mid \hat{p}_t\right]\right] \\
&= \mathbb{E}\left[\langle \hat{p}_t, \mathbb{E}\left[\hat{\ell}_t \mid \hat{p}_t\right] \rangle\right] \\
&= \mathbb{E}\left[\langle \hat{p}_t, \mathbb{E}\left[\hat{\ell}_t \mid p_t\right] \rangle\right] \\
&= \mathbb{E}\left[\langle \hat{p}_t, \ell_t \rangle\right].
\end{aligned}$$

In moving from the second to the third line we used the fact that knowing $\hat{p}_t$ is equivalent to knowing p_t. For the last step, we used Lemma 2.1. □

By the guarantee of the expert problem shown in the last lecture, where we used the entropic regularizer, we have that

$$\forall a \in [N], \quad \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \hat{p}_t, \frac{\alpha}{N}\hat{\ell}_t \right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha}{N}\hat{\ell}_t(a)\right] \leq O(\sqrt{T \log N}) \tag{1}$$

We proceed to upper bound the regret of the bandit problem via reduction. Observe that

$$\begin{aligned}
\text{regret}_{\text{bandit}}(a) &= \mathbb{E}\left[\sum_{t=1}^{T}\langle p_t, \ell_t\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_t, \hat{\ell}_t\right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle (1-\alpha)\hat{p}_t + \frac{\alpha}{N}\mathbf{1}, \hat{\ell}_t\right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_t(a)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle \hat{p}_t, \hat{\ell}_t\right\rangle\right] + \alpha\cdot\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \frac{\mathbf{1}}{N} - \hat{p}_t, \hat{\ell}_t\right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_t(a)\right] \\
&\leq \frac{N}{\alpha}O(\sqrt{T\log N}) + \alpha\cdot\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \frac{\mathbf{1}}{N} - \hat{p}_t, \ell_t\right\rangle\right] \\
&\leq \frac{N}{\alpha}O(\sqrt{T\log N}) + 2\alpha T. \qquad (2)
\end{aligned}$$

Going from the first to the second line, we used Lemma 2.2 to exchange ℓ_t and $\hat{\ell}_t$. Going from the second to the third line, we used Lemma 2.1 to exchange $\ell_t(a)$ and $\hat{\ell}_t(a)$. Going from the third to the fourth line, we used the fact that we can write

$$(1-\alpha)\hat{p}_t + \alpha\cdot\mathbf{1}/N = \hat{p}_t + \alpha(\mathbf{1}/N - \hat{p}_t).$$

To move from the fourth to the fifth line, we used the result for the regret of the expert problem shown above in (1), and Lemma 2.3 to exchange ℓ_t and $\hat{\ell}_t$. For the fifth to the final line, we used duality of the L_1 and L_∞ norms, and the bounds

$$\left\|\frac{\mathbf{1}}{N} - \hat{p}_t\right\|_1 \leq 2, \quad \|\ell_t\|_\infty \leq 1.$$

To choose the optimal α in order to minimize the bound in (2), we can differentiate the quantity

$$Q(\alpha) = (N/\alpha)O(\sqrt{T\log N}) + 2\alpha T$$

with respect to α, giving us

$$\alpha^* = O(T^{-1/4}\sqrt{N}\log^{1/4} N).$$

Then, by (2) above,

$$\begin{aligned}
\text{regret}_{\text{bandit}}(a) &\leq \frac{N}{\alpha^*}O(\sqrt{T\log N}) + 2\alpha^* T \\
&\leq O(T^{3/4}\sqrt{N}\log^{1/4} N),
\end{aligned}$$

which implies that

$$\text{regret}_{\text{bandit}} \leq O(T^{3/4}\sqrt{N}\log^{1/4} N).$$

While the dependence on T here is not as good as the $\sqrt{T}$ for the expert problem, it is better than the trivial bound $O(T)$. In fact, the regret for the multi-armed bandit problem can be further reduced to $O(\sqrt{T})$, but the proof is quite complicated, and we will not do it in class. Those interested can refer to Lecture 16 in Percy Liang's notes for more details.

3 General OCO with partial observation

The reduction approach we saw earlier also works for more general OCO problems with partial observation. As described in section 5 of the previous lecture (Lecture 16), we can reduce the more general problem of online convex optimization to online linear optimization. In particular, recall that in this setup, the full loss function is the linear function

$$g_t(\omega) = \langle \nabla f_t(\omega_t), \omega \rangle.$$

In the case of general OCO with partial information, we only observe $f_t(\omega_t)$. We can then reduce the problem to OCO with full information if we can find an unbiased estimate of $\nabla f_t(\omega_t)$ using $f_t(\omega_t)$.

3.1 Estimating $\nabla f_t(\omega_t)$ from $f_t(\omega_t)$.

We note that the following result is useful more generally, since it allows us to estimate the gradient of a function using only information about function values.

Lemma 3.1. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, with bounded Hessian. Then,*

$$\nabla f(w) = \lim_{\delta \to 0} \mathbb{E}_{\xi \sim \mathbb{S}^{d-1}} \left[\frac{d\xi}{\delta} \cdot f(w + \xi\delta) \right],$$

where $\mathbb{S}^{d-1}$ refers to the unit sphere $\left\{ x \in \mathbb{R}^d : \|x\|_2 = 1 \right\}$ in $\mathbb{R}^d$, and ξ is drawn uniformly from the sphere $\mathbb{S}^{d-1}$.

Proof. The key idea behind the proof is to use Taylor expansion. We have that

$$\begin{aligned}
& f(w + \delta\xi) = f(w) + \langle \nabla f(w), \xi\delta \rangle + O(\delta^2) \\
\implies & \frac{f(w + \delta\xi)\xi}{\delta} = \frac{f(w)\xi}{\delta} + \langle \nabla f(w), \xi \rangle \xi + O(\delta) \\
\implies & \mathbb{E}\left[\frac{f(w + \delta\xi)\xi}{\delta} \right] = 0 + \mathbb{E}\left[\xi\xi^T \nabla f(w) \right] + O(\delta) \\
\implies & \mathbb{E}\left[\frac{f(w + \delta\xi)\xi}{\delta} \right] = \begin{bmatrix} 1/d & & \\ & \ddots & \\ & & 1/d \end{bmatrix} \nabla f(w) + O(\delta) \\
\implies & \mathbb{E}\left[\frac{d\xi}{\delta} \cdot f(w + \xi\delta) \right] = \nabla f(w) + O(\delta).
\end{aligned}$$

Taking limit $\delta \to 0$ then completes the proof. The first step uses the assumption of a bounded Hessian to control the second-order term. To get from the third to the fourth line, we relied on a calculation of $\mathbb{E}[\xi\xi^T]$. Since the ξ's are uniformly distributed on, it follows that

- $\mathbb{E}[\xi_i \xi_j] = 0$ for any $1 \le i \ne j \le d$, by symmetry.
- $\mathbb{E}[\xi_i^2] = 1/d$ for any $i \in [d]$, since $\|\xi\|_2 = 1$, and ξ has d coordinates, each of which should behave similarly given that ξ is uniformly distributed on $\mathbb{S}^{d-1}$.

This completes the proof. □

4 Stochastic Multi-Armed Bandit Problem

For the remaining time in this lecture, we will begin analyzing the multi-armed bandit problem in the stochastic setting.

The setup is as follows:

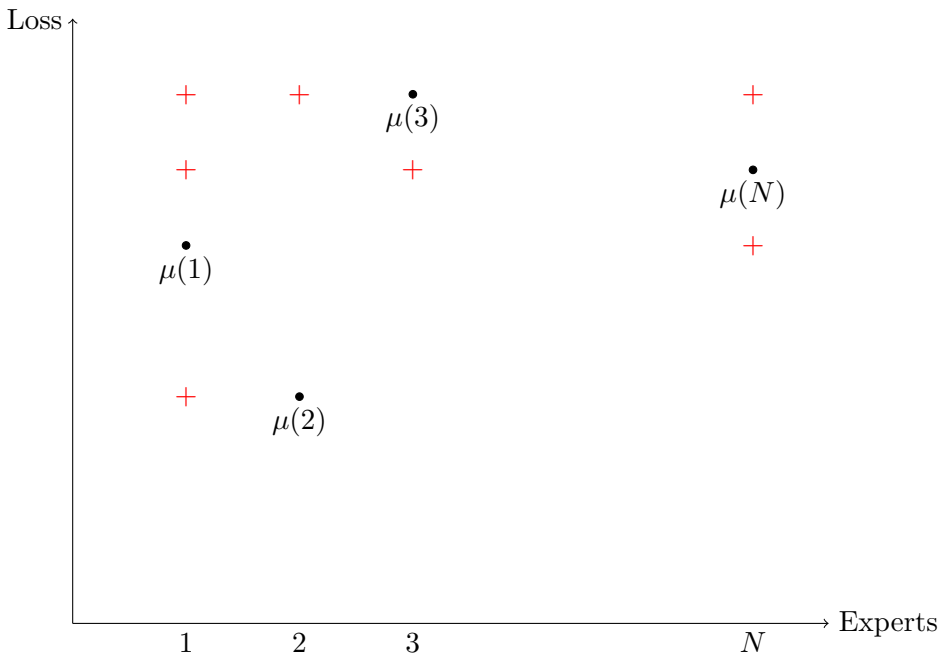
- Again, assume that there are N experts.
- Suppose that $\ell_1, \ldots, \ell_T \overset{i.i.d}{\sim} D$ for some probability distribution D
- Let D_a denote the distribution of $\ell(a)$ when $\ell \sim D$
- Let $\mu = \mathbb{E}_{\ell \sim D}\mathbb{E}[\ell]$ denote the expected loss vector, where $\mu \in \mathbb{R}^N$.
- Let $a^* = \arg\min_{a \in [N]} \mu(a)$
- We only see $\ell_t(a_t)$ at every iteration t, where a_t is the expert we choose.

The regret is the following:

$$\text{regret} \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(a^*)\right],$$

where the expectation is taken over the randomness of our algorithm, and the stochastic loss functions $\ell_1, \ldots, \ell_T$.

A challenge in the stochastic setting with partial feedback is that we never see the true mean loss function μ. Consider the following scenario.



In the above plot, the red plus signs denote the (random) losses we see during the course of the algorithm for each expert. While the expected loss for expert 2, $\mu(2)$, is the lowest in the plot above, the one time we did see a loss for expert 2, the loss was significantly higher than the true mean $\mu(2)$. As such, on the basis of the available evidence, we might erroneously conclude that expert 2 is not a good expert, when the truth was that we simply got unlucky with our random draws.

The above scenario highlights the tradeoff between exploitation and exploration in a stochastic multi-armed bandit problem. We will continue the discussion in the next lecture.