**EE 290 Theory of Multi-armed Bandits and Reinforcement Learning** Lecture 13 - 3/02/2021

# Lecture 13: Online Mirror Descent

*Lecturer: Jiantao Jiao* *Scribe: Alyssa Li Dayan, Han Feng, Sandy Tanwisuth*

In this lecture we cover an algorithm based on Online Mirror Descent that achieves the optimal (within constants) regret for both the adversarial and stochastic multi-armed bandits settings. The major references are [1] and its extended version [2].

## 1 Review of Stochastic and Adversarial Bandits Regret Bounds

Bandit has a huge literature. Except for some recent work related to Bayesian and frequency arguments, we have covered the major ideas: successive elimination, UCB, Thompson Sampling, EXP3.

Our aim is to minimize the expected regret of our algorithm at time $T$, $E[R(T)]$. In previous lectures we introduced different algorithms to minimize this quantity in the stochastic and adversarial settings. In the stochastic case we saw that the UCB1 algorithm achieves instance-dependent and instance-independent expected regret bounds of $\tilde{\Theta}(\sqrt{kT})$ and $\Theta(\log(T)\sum_{i\neq i^*}\frac{1}{\Delta_i})$, respectively. Meanwhile in the adversarial case, the EXP3 algorithm achieves $E[R(T)] \lesssim \sqrt{kT\log(k)}$.

However, applying either algorithm in the other setting is not guaranteed to perform well. Our goal is to find a single efficient algorithm that maintains these upper bounds in both stochastic and adversarial settings. Prior to online mirror descent, this can be achieved by initially assuming the stochastic setting, but change strategy upon evidence of adversarial behavior [3].

## 2 Preliminaries

### 2.1 Convex Conjugate

The convex conjugate of function $f\colon \mathbb{R}^k \to \mathbb{R}$ is denoted $f^*(y)$ and defined as

$$f^*(y) = \max_{x\in\mathbb{R}^k}\{\langle x, y\rangle - f(x)\}. \tag{1}$$

This expression has the convenient property that if $f(x)$ is convex, then $\langle x, y\rangle - f(x)$ is concave in $x$, so the maximum value over $x$ is easy to compute for any $y$. This is because the negative of a convex function $-f(x)$ is concave, and adding the linear term $\langle x, y\rangle$ does not affect concavity.

### 2.2 Indicator Function

The indicator function of a set $A$ is defined as

$$\mathcal{I}_A(x) = \begin{cases} 0 \text{ if } x \in A \\ \infty \text{ otherwise .} \end{cases} \tag{2}$$

This is also called the characteristic function, and is commonly used in optimization contexts. Note that this is different from the indicator function used in probability theory, which we denote by $\mathcal{I}$.

When $I_A(x)$ is added to another function $f(x)$, the combined result is $f(x)$ when $x \in A$ and $\infty$ everywhere else. Thus, when taking the convex conjugate we can restrict ourselves to $x \in A$ and simplify the expression to:

$$(f + \mathcal{I}_A)^*(y) = \max_{x\in A}\{\langle x, y\rangle - f(x)\}. \tag{3}$$

Suppose $f$ is convex and $(\nabla f)^{-1}$ is invertible, the gradient of the convex conjugate [1] is simply the value of $x$ that attains the maximum:

$$\nabla(f + \mathcal{I}_A)^*(y) = \arg\max_{x \in A}\{\langle x, y\rangle - f(x)\}. \tag{4}$$

# 3 Online Mirror Descent

Online Mirror Descent chooses a sequence of regularizers $\Psi_t$. Like EXP3, we would like to obtain estimates of the "loss", run "gradient descent", and then translate that to the sampling of the next arm. To represent the space of categorical probability distributions over $k$ alternatives, we define the simplex over $k$ variables $\mathbf{w} = (w_1, w_2, \ldots, w_k)$ as

$$\Delta^k = \{\mathbf{w} | w_i \geq 0, \sum_i w_i = 1\}. \tag{5}$$

Note that $\Delta^k$ has dimension $k-1$, and the paper [2] used $k-1$ instead of $k$ in the superscript of $\Delta$. Algorithm 1 describes online mirror descent for bandits [2].

**Algorithm 1** Online Mirror Descent for Bandits

**Input**:$(\Psi_t)_{t=1,2,\ldots}$.
**Initialize**: $\hat{L}_0 = \mathbf{0}_k$ (vector of $k$ zeros).
**for** $i \leftarrow 1, \ldots,$ **do**
  choose $\mathbf{w}_t = \nabla(\Psi_t + \mathcal{I}_{\Delta^k})^*(-\hat{L}_{t-1}) = \arg\min_{\mathbf{w} \in \Delta^k} \langle \mathbf{w}, \hat{L}_{t-1}\rangle + \Psi_t(\mathbf{w})$;
  sample $I_t \sim \mathbf{w}_t$ and observe $l_{t,I_t}$;
  construct $\hat{l}_t \in \mathbb{R}^k$: $\hat{l}_{t,i} = \mathcal{I}(I_t = i)\frac{l_{t,i}}{w_{t,i}}$, set $\hat{L}_t = \hat{L}_{t-1} + \hat{l}_t$.
**end for**

Note that the indicator function $\mathcal{I}(I_t = i)$ in the penultimate line is the probability theory indicator function, which is 1 when $I_t = i$ and 0 otherwise.

As before, $\hat{L}_t$ is an unbiased estimate of the cumulative loss for each arm at time $t$. The regularization term $\Psi_t$ in $\arg\min_{\mathbf{w} \in \Delta^k} \langle \mathbf{w}, \hat{L}_{t-1}\rangle + \Psi_t(\mathbf{w})$ pushes the optimal distribution of $\mathbf{w}$ away from the boundary of $\Delta^k$, where the expected loss $\langle \mathbf{w}, \hat{L}_{t-1}\rangle$ is minimized. This prevents the algorithm from being biased too heavily towards one arm to defend against the adversarial setting.

# 4 $\alpha$-Tsallis Entropy

We introduce a very important family of potential functions called $\alpha$-Tsallis Entropy. Unlike the Shannon entropy used in information theory, $\alpha$-Tsallis Entropy is negative. It is defined as:

$$H_\alpha(x) \triangleq \frac{1}{1-\alpha}(1 - \sum_{i \in [k]} x_i^\alpha).$$

Online mirror descent uses $\alpha$-Tsallis Entropy with $\alpha = \frac{1}{2}$ as its regularizer, with an additional time-dependent factor $\eta_{t,i}$ which can be interpreted as the learning rate (and commonly only depends on time $t$):

$$\Psi_{t,\alpha}(\mathbf{w}) \triangleq -\sum_{i \in [k]} \frac{w_i^\alpha}{\alpha \eta_{t,i}},$$

[1] To see how to take the gradient over the maximum, see the Envelop Theorem, or in the case of convex functions, Danskin's Theorem.

Here is some intuition on how OMD is connected to EXP3 algorithm. As $\alpha$ approaches 1, both the numerator and the denominator of $\alpha$-Tsallis Entropy approach 0. Computing the limit using L'Hôpital's rule we obtain:

$$\lim_{\alpha \to 1} H_\alpha(x) = \sum_{i \in [k]} x_i \log(x_i) \leq 0.$$

This quantity is the negative of Shannon entropy $H_p(x) = \sum_{i \in [k]} p_i \log(\frac{1}{p_i})$. With this entropy, OMD reduces exactly to the EXP3 algorithm, which you will prove in homework 2. To see how this regularization pushes the optimal distribution of $w$ away from the boundary of $\Delta^k$, set $\alpha = \frac{1}{2}$. Then

$$\Psi_{t,\frac{1}{2}}(x) \propto -\sum \sqrt{w_i}$$

The term has interesting properties:

$$\max_{w \in \Delta^k} \sum \sqrt{w_i} = \sqrt{k}$$
$$\min_{w \in \Delta^k} \sum \sqrt{w_i} = 1.$$

The max is attained for uniform $w$ at the center of $\Delta^k$, while the min is attained at a corner of the simplex where one $w_i = 1$ and all the rest are 0 (the singleton distribution). Since we choose $w_t$ to minimize the sum of the linear and regularization terms, we are much more inclined to set $w_t$ away from the boundary. This allows us to reserve some probability mass for trying other actions, knowing that our loss estimates may be inaccurate (which is especially the case in the adversarial setting).

# 5 Theorem by Zimmert & Seldin

This theorem was introduced in the reference [2]. We choose $\alpha = \frac{1}{2}$, because other values of $\alpha$ necessitate the learning rate to depend on the unknown gap $\Delta_i$, which is impractical. They choose the learning rate $\eta_{t,i} = \eta_t = \frac{1}{\sqrt{t}}$, so that it decreases over time. Previously we always fixed a particular value of $\eta$ (sometimes calculated from the horizon, when known). OMD is appealing because it does not need to know the horizon in advance, while maintaining optimal regret.

In the adversarial environment, OMD achieves

$$\mathbb{E}[R(T)] \leq 4\sqrt{kT} + 1.$$

In the stochastic case, OMD achieves

$$\mathbb{E}[R(T)] \leq (\sum_{i \neq i^*} \frac{4\log(T) + 68}{\Delta_i}) + 4\sqrt{k}.$$

The comparison of the two terms suggests that initially you might have to wait for some number of time-steps depending on $k$, before focusing on the optimal arm.

Note that the adversarial bound is stronger than the stochastic bound. This result is impressive because it achieves the information-theoretic limit (the instance-independent bound) within a universal constant for both the stochastic and adversarial settings.

# 6 Self-bounding Inequality

We wish to upper bound the regret, which by Lemma 3 in [1] satisfies the following inequality:

$$l_{t,I_t} \leq \sum_{i \neq I_t} \frac{4\eta w_{i,t}}{w_{t,I_t}} + \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t).$$

In order to further upper bound the regret, we use the *self-bounding inequality*, one of whose manifestations states that the adversarial regime satisfies

$$\mathbb{E}[R(T)] \geq \mathbb{E} \sum_{i \neq i^*} \sum_{t=1}^{T} w_{t,i} \Delta_i. \tag{6}$$

This is called the self-bounding inequality because applying it to the bound of $\mathbb{E}[R(T)]$ introduces a small multiple of $\mathbb{E}[R(T)]$, which can be moved to the left-hand side. For details, see Lemma 4 in [1]. Setting $\Phi_t \triangleq (\Psi_t + I_{\Delta^k})^*$ , $\alpha = \frac{1}{2}$ and $\eta_t = \frac{1}{\sqrt{t}}$, it can be shown that

$$\mathbb{E}[\sum_{t=1}^{T} \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)] \leq \mathbb{E}[\sum_{t=1}^{T} \sum_{i \neq i^*} \sqrt{\frac{w_{t,i}}{t}}] + \mathbb{E}[L_{i^*}] + \sqrt{2k}.$$

Where $L_{i^*}$ is the oracle performance. Rearranging and combining the inequalities allows us to upper bound $l_{t,I_t}$ by $\mathbb{E}[\sum_{t=1}^{T} \sum_{i \neq i^*} \sqrt{\frac{w_{t,i}}{t}}]$ (omitting constant terms), from which the final result can be derived.

# References


[1] J. Zimmert and Y. Seldin, "An Optimal Algorithm for Stochastic and Adversarial Bandits," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2019, pp. 467–475.

[2] ——, "Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits," *Journal of Machine Learning Research*, vol. 22, no. 28, pp. 1–49, 2021.

[3] S. Bubeck and A. Slivkins, "The Best of Both Worlds: Stochastic and Adversarial Bandits," in *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, Jun. 2012, pp. 42.1–42.23.