Reinforcement Learning II

Tengyuan Liang[1]

[1] The University of Chicago
Booth School of Business

DLA Lecture 6: Explore vs. Exploit

Bandit algorithms and details about exploration vs. exploitation. Readings: Hardt and Recht [2] Chapters 11 and 12, Lattimore and Szepesvari [3] Chapters 6, 7 and 11.

[2] Moritz Hardt and Benjamin Recht. *Patterns, Predictions, and Actions: A Story about Machine Learning*. Princeton University Press, 2022

[3] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020

Contents

1 Problem Setup

Multi-arm bandit [4] is a special class of sequential decision-making (SDM) problems we have considered before with no dynamic systems for how the state evolves: one simply decides based on the history the current action to take from a finite dictionary of arms, and then observes the reward. In other words, there is no states X, the reward can be represented by a vector of length k, and the action space is $\mathcal{A}_k = \{1, 2, \ldots, k\}$.

[4] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4): 285–294, 1933; and Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(6):527–535, 1952

Definition 1 (Stochastic Bandit)**.** *Let the action space be $\mathcal{A}_k = \{1, 2, \ldots, k\}$. Let the reward vectors $\mathbf{r}_t \in \mathbb{R}^k, t = 1, 2 \ldots, T$ be i.i.d. samples from an unknown rewards distribution. We denote the average reward for arm i as $\mu(i) := \mathbb{E}[\mathbf{r}_t(i)]$, $i = 1, 2, \ldots k$.*

At each time, the player takes an action $u_t \in \mathcal{A}_k$ (based on the past information $\{u_s, \mathbf{r}_s(u_s)\}_{s<t}$), then only observes the reward $\mathbf{r}_t(u_t) \in \mathbb{R}$. The player tries to maximize the cumulative reward

$$\max_{u_t} \ \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t(u_t)\right] = \max_{u_t} \ \mathbb{E}\left[\sum_{t=1}^{T} \mu(u_t)\right] .$$

Define the cumulative regret for an algorithm/policy π, where the actions $u_t \sim \pi$

$$\mathcal{R}_T^{\text{sto}}(\pi) := T \cdot \max_{i \in \mathcal{A}_k} \mu(i) - \mathbb{E}\left[\sum_{t=1}^{T} \mu(u_t)\right]$$

Definition 2 (Adversarial Bandit). *Let the action space be $\mathcal{A}_k = \{1, 2, \ldots, k\}$. The reward vectors $\mathbf{r}_t \in \mathbb{R}^k, t = 1, 2 \ldots, T$ now can be arbitrary vectors.*

At each time, the player takes an action $u_t \in \mathcal{A}_k$ (based on the past information $\{u_s, \mathbf{r}_s(u_s)\}_{s<t}$), then observes the reward $\mathbf{r}_t(u_t) \in \mathbb{R}$. The player tries to maximize the cumulative reward

$$\max_{u_t} \sum_{t=1}^{T} \mathbf{r}_t(u_t) \ .$$

Define the cumulative regret for an algorithm/policy π, where the actions $u_t \sim \pi$

$$\mathcal{R}_T^{\text{adv}}(\pi) := \max_{i \in \mathcal{A}_k} \sum_{t=1}^{T} \mathbf{r}_t(i) - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t(u_t)\right]$$

2 *Stochastic Bandits*

2.1 *Explore-Then-Commit (ETC) Algorithm*

Definition 3 (ETC Algorithm). ***Explore-Then-Commit*** *(ETC) Algorithm specifies an exploration budget of time $m \times k$.*

Define

$$N^t(i) := \sum_{s=1}^{t} 1_{u_s = i} \tag{2.1}$$

$$\hat{\mu}^t(i) := \frac{1}{N^t(i)} \sum_{s=1}^{t} 1_{u_s = i} \cdot \mathbf{r}_s(i) \tag{2.2}$$

which correspond to the number of times action i is taken up till time t, and the empirical estimate of the average reward for arm i.

The ETC algorithm π^{ETC} implements the following

1. *Input an integer m*

2. *In round t, choose action*

$$u_t = \begin{cases} (t-1 \mod k) + 1 & \text{if } t \le mk \\ \arg\max_i \ \hat{\mu}^{mk}(i) & \text{if } t > mk \end{cases}$$

One more notation we need is the gap of the problem

$$\Delta_i := \max_{i'} \mu(i') - \mu_i \ , \ i \in [k] \ . \tag{2.3}$$

Theorem 1 (Regret for ETC Algorithm). *Consider stochastic bandits with bounded rewards $\mathbf{r}(i) \in [-1/\sqrt{2}, 1/\sqrt{2}]$.*

TL: The constants are for convenience only.

$$\mathcal{R}_T^{\text{sto}}(\pi^{ETC}) \le \sum_{i=1}^{k} m\Delta_i + (T - mk)\Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right) \tag{2.4}$$

A few remarks follow regarding the **exploration and exploitation tradeoff**. Consider $k = 2$: $\Delta_1 = 0$, and $\Delta_2 = \Delta > 0$. Then the above regret bound reads

$$\mathcal{R}_T^{\text{sto}}(\pi^{ETC}) \le m\Delta + T\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \tag{2.5}$$

Let us denote

$$m_0 = \lceil \frac{4}{\Delta^2} \log\left(\frac{T\Delta^2}{4}\right) \rceil \tag{2.6}$$

- **Gap dependent regret**.
  - If $m_0 \ge 1$, then set $m = m_0$

    $$\mathcal{R}_T^{\text{sto}}(\pi^{ETC}) \le \Delta + \frac{4}{\Delta}\left(\log\left(\frac{T\Delta^2}{4}\right) + 1\right)$$

    TL: $\log(T)$ regret
  - If $m_0 < 1$, then $\frac{T\Delta^2}{4} < 1$ which implies that the gap is small $\Delta < \frac{2}{\sqrt{T}}$, in such case, set $m = T/2$ will result in

    $$\mathcal{R}_T^{\text{sto}}(\pi^{ETC}) = \frac{T}{2}\Delta \le \sqrt{T}$$

- **Gap independent regret**. If $m_0 \ge 1$, then set $m = m_0$ and note that

  $$\frac{4}{\Delta}\left(\log\left(\frac{T\Delta^2}{4}\right) + 1\right) \le 2\sqrt{T} \sup_{x \ge 0} \frac{2\log x + 1}{x} \le 4\sqrt{e}\sqrt{T}$$

  and thus

  $$\mathcal{R}_T^{\text{sto}}(\pi^{ETC}) \le \Delta + 4\sqrt{e}\sqrt{T}$$

  This shows the worst case $\sqrt{T}$ regret.

  TL: $\sqrt{T}$ regret

- **Gap independent policy**. So far the m depends on the knowledge of Δ. Without knowing it and simply set $m = T^{2/3}$, we have

  $$\begin{aligned}\mathcal{R}_T^{\text{sto}}(\pi^{ETC}) &\le T^{2/3}\Delta + T^{2/3} \cdot T^{1/3}\Delta \exp\left(-\frac{T^{2/3}\Delta^2}{4}\right) \\ &\le T^{2/3}\Delta + T^{2/3} \cdot 2 \sup_{x \ge 0} x \exp\left(-x^2\right) \asymp T^{2/3}\end{aligned}$$

  TL: $T^{2/3}$ regret

Proof. Observe the simple identity

$$\begin{aligned}\mathcal{R}_T^{\text{sto}}(\pi) &:= T \cdot \max_{i \in \mathcal{A}_k} \mu(i) - \mathbb{E}\left[\sum_{t=1}^{T} \mu(u_t)\right] \\ &= \sum_{i=1}^{k} \Delta_i \, \mathbb{E}[N^T(i)]\end{aligned}$$

WLOG, assume $\Delta_1 = 0$, and $\Delta_i > 0$ for $i \geq 2$. Let's control

$$\begin{aligned}
\mathbb{E}[N^T(i)] &= \sum_{t=1}^{T} \mathbb{P}[u_t = i] \\
&= m + \sum_{t=mk+1}^{T} \mathbb{P}[u_t = i] \\
&= m + \sum_{t=mk+1}^{T} \mathbb{P}\left[\widehat{\mu}^{mk}(i) > \widehat{\mu}^{mk}(j),\ \forall j\right] \\
&\leq m + \sum_{t=mk+1}^{T} \mathbb{P}\left[\widehat{\mu}^{mk}(i) > \widehat{\mu}^{mk}(1)\right] \\
&\leq m + (T - mk)\, \mathbb{P}\left[\frac{1}{m}\sum_{z=0}^{m-1}\big(\mathbf{r}_{zk+i}(i) - \mathbf{r}_{zk+1}(1) - \mu(i) + \mu(1)\big) > \Delta_i\right] \\
&\leq m + (T - mk)\exp\left(-\frac{m\Delta_i^2}{4}\right)
\end{aligned}$$

Clearly $B_z := \mathbf{r}_{zk+i}(i) - \mathbf{r}_{zk+1}(1) \in [-\sqrt{2}, \sqrt{2}]$ and that

$$\mathbb{P}\left[\frac{1}{m}\sum_{z=0}^{m-1} B_z - \mathbb{E}[B_z] > t\right] \leq \exp\left(-\frac{mt^2}{4}\right) \tag{2.7}$$

and the last line follows from Azuma-Hoeffding's inequality.

□

2.2 *Upper Confidence Bound (UCB) Algorithm*

The next algorithm, UCB, is derived based on the *optimism principle*. The name **upper confidence bound** came from the Azuma Hoffding's inequality

$$\mathbb{P}\left[\mu > \widehat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}}\right] \leq \delta\ . \tag{2.8}$$

It shows an optimistic estimate of the true μ based on n-empirical samples, with accuracy parameter δ.

Definition 4. ***Upper Confidence Bound*** *(UCB) Algorithm Define the*

$$\mathrm{UCB}^{t,\delta}(i) = \begin{cases} \infty & \text{if } N^t(i) = 0 \\ \widehat{\mu}^t(i) + \sqrt{\frac{2\log(1/\delta)}{N^t(i)}} & \text{otherwise} \end{cases} \tag{2.9}$$

The UCB algorithm π^{UCB} *implements the following*

1. *Input a small accuracy parameter* $\delta \in (0, 1)$
2. *In round* t*, choose action*

$$u_t := \arg\max_i\ \mathrm{UCB}^{t,\delta}(i)$$

3. *Observe the new reward and update the upper confidence bounds*

Theorem 2 (Regret for UCB Algorithm)**.** *Consider stochastic bandits with bounded rewards* $\mathbf{r}(i) \in [-1, 1]$ *and* $\delta = \frac{1}{T(T+1)}$

$$\mathcal{R}_T^{\text{sto}}(\pi^{UCB}) \leq 2\sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(T+1)}{\Delta_i} \tag{2.10}$$

Proof. First, let us introduce one new notation.

$$\mathbf{R} = [r_{it}]_{i\in[k],t\in[T]} \in \mathbb{R}^{k\times T} \tag{2.11}$$

be the random reward matrix, where $r_{i,t}, t \in [T]$ are i.i.d. draws from the distribution $\mathcal{L}(\mathbf{r}(i))$. We introduce the empirical mean as

$$\widehat{\mu}_m(i) := \frac{1}{m}\sum_{t=1}^{m} r_{i,t}\ . \tag{2.12}$$

The Stochastic bandit model is stochastically equivalent to the following model: at each time t, if an arm i is pulled, then reveal the reward $\mathbf{r}_t(i) = r_{i,N^t(i)}$.

TL: the distribution of $\mathbf{r}_t(i)$ is the same as $r_{i,1}$ though $N^t(i)$ is a random variable.

Again recall the basic identity,

$$\mathcal{R}_T^{\text{sto}}(\pi) := \sum_{i=1}^{k} \Delta_i\, \mathbb{E}[N^T(i)]$$

WLOG, assume $\Delta_1 = 0$, and $\Delta_i > 0$ for $i \geq 2$. Let's control

$$\mathbb{E}[N^T(i)]$$

Choose an integer m_i

$$m_i = \lceil \frac{8\log(1/\delta)}{\Delta_i^2} \rceil < T$$

so to satisfy

TL: m_i conceptually is the right level of number of arms one need to pull to figure out in order to eliminate the bad arm.

$$\sqrt{\frac{2\log(1/\delta)}{m_i}} < \frac{\Delta_i}{2} \tag{2.13}$$

Define two events

$$E_i = \left\{\widehat{\mu}_{m_i}(i) + \sqrt{\frac{2\log(1/\delta)}{m_i}} < \mu(1)\right\} \tag{2.14}$$

$$F = \left\{\forall t \in [T],\ \text{UCB}^{t,\delta}(1) > \mu(1)\right\} \tag{2.15}$$

Let's bound the probability of each event

$$\begin{aligned}
\mathbb{P}[E_i^c] &= \mathbb{P}\left[\widehat{\mu}_{m_i}(i) + \sqrt{\frac{2\log(1/\delta)}{m_i}} \geq \mu(1)\right] \\
&= \mathbb{P}\left[\widehat{\mu}_{m_i}(i) - \mu(i) \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{m_i}}\right] \\
&\leq \mathbb{P}\left[\widehat{\mu}_{m_i}(i) - \mu(i) \geq \tfrac{\Delta_i}{2}\right] \\
&\leq \exp\left(-\tfrac{m_i \Delta_i^2}{8}\right) \leq \delta
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}[F^c] &= \mathbb{P}\left[\exists t \in [T],\ \mathrm{UCB}^{t,\delta}(1) \leq \mu(1)\right] \\
&\leq \mathbb{P}\left[\exists m \in [T],\ \widehat{\mu}_m(1) + \sqrt{\frac{2\log(1/\delta)}{m}} \leq \mu(1)\right] \\
&\leq T\delta
\end{aligned}$$

Thus

$$\mathbb{P}[(E_i \cap F)^c] \leq \mathbb{P}[E_i^c] + \mathbb{P}[F^c] = (T+1)\delta\ . \tag{2.16}$$

On the event $E_i \cap F$, we claim that

$$N^T(i) \leq m_i\ . \tag{2.17}$$

If not, define

$$\tau := \inf\{t \in [T]\ :\ N^t(i) = m_i\} \leq T - 1 \tag{2.18}$$

and at time τ, arm i is pulled again. Then one must have

$$\mathrm{UCB}^{\tau,\delta}(i) > \mathrm{UCB}^{\tau,\delta}(1) \tag{2.19}$$

Recall that on the event E_i, we have

$$\mathrm{UCB}^{\tau,\delta}(i) = \widehat{\mu}^{\tau}(i) + \sqrt{\frac{2\log(1/\delta)}{N^{\tau}(i)}} = \widehat{\mu}_{m_i}(i) + \sqrt{\frac{2\log(1/\delta)}{m_i}} < \mu(1) \tag{2.20}$$

yet on the event F we know

$$\mathrm{UCB}^{\tau,\delta}(1) > \mu(1)\ . \tag{2.21}$$

Therefore on the event $E_i \cap F$, (2.20) and (2.21) contradicts with (2.19). Therefore, we have shown on the event $E_i \cap F$, $N^T(i) \leq m_i$.

Now we are ready to bound

$$\begin{aligned}
\mathbb{E}[N^T(i)] &= \mathbb{E}[N^T(i) \cdot 1_{E_i \cap F}] + \mathbb{E}[N^T(i) \cdot 1_{(E_i \cap F)^c}] \\
&\leq m_i \cdot \mathbb{P}[E_i \cap F] + T \cdot \mathbb{P}[(E_i \cap F)^c] \\
&\leq \frac{8\log(1/\delta)}{\Delta_i^2} + 1 + T(T+1)\delta
\end{aligned}$$

Plug in $\delta = \frac{1}{T(T+1)}$, we know

$$\mathbb{E}[N^T(i)] \le 2 + \frac{16\log(T+1)}{\Delta_i^2} \tag{2.22}$$

and thus reach the final bound. □

3 Adversarial Bandits

Now we relieve the i.i.d. assumptions, and see that won't change the regret in the worst case $\sqrt{T}$. However, note in the best case, the regret in the stochastic bandit setting grows at a rate of $\log(T)$.

3.1 Exponential Weight for Exploration and Exploitation (EXP3) Algorithm

Definition 5 (**EXP3** Algorithm). *Define the inverse propensity score estimate of the reward vector*

$$\widehat{\mathbf{r}}_t(i) = \begin{cases} \frac{\mathbf{r}_t(i)}{\mathbb{P}[u_t=i]} & \text{if } i = u_t \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

At each round t, let $u_t \sim \mathbf{p}_t$ where $\mathbf{p}_t \in \mathbb{R}^k$ is a probability vector that sums up to 1.

1. *Input a learning rate η*
2. *In round t, choose action $u_t \sim \mathbf{p}_t$ where $\mathbf{p}_t \in \mathbb{R}^k$ is a probability distribution over arms*
3. *Observe the reward $\mathbf{r}_t(u_t)$ scalar, and update*

$$\mathbf{p}_{t+1}(i) = \frac{\mathbf{p}_t(i)\exp(\eta\widehat{\mathbf{r}}_t(i))}{\sum_j \mathbf{p}_t(j)\exp(\eta\widehat{\mathbf{r}}_t(j))} \tag{3.2}$$

Theorem 3 (Regret for EXP3 Algorithm). *Consider the adversarial bandits with bounded rewards $\mathbf{r}(i) \in [-1,1]$ and $\eta \in (0,1)$*

$$\mathcal{R}_T^{\text{adv}}(\pi^{EXP3}) \le \frac{\log(k)}{\eta} + \eta \cdot Tk \tag{3.3}$$

Remark. *Note this bound is optimized when $\eta = \sqrt{\frac{\log(k)}{Tk}}$, in this case*

$$\mathcal{R}_T^{\text{adv}}(\pi^{EXP3}) \le 2\sqrt{T \cdot k\log k}$$

Proof. We first introduce an inequality based on KL divergence $D(\cdot \,\|\, \cdot)$, define

$$\mathbf{p}_1(i) = \frac{\mathbf{p}_0(i)\exp(\eta\mathbf{r}(i))}{\sum_j \mathbf{p}_0(j)\exp(\eta\mathbf{r}(j))} \tag{3.4}$$

We then claim that $\forall \mathbf{q}$

$$\langle \mathbf{r}, \mathbf{q} \rangle - \langle \mathbf{r}, \mathbf{p}_0 \rangle = \frac{D(\mathbf{q} \parallel \mathbf{p}_0) - D(\mathbf{q} \parallel \mathbf{p}_1)}{\eta} + \frac{D(\mathbf{p}_0 \parallel \mathbf{p}_1)}{\eta} . \tag{3.5}$$

To derive this, notice

$$D(\mathbf{q} \parallel \mathbf{p}) = \langle \mathbf{q}, \log \mathbf{q} - \log \mathbf{p} \rangle \tag{3.6}$$

and thus the RHS equals

$$\begin{aligned}
&\frac{\langle \mathbf{q}, \log \mathbf{q} - \log \mathbf{p}_0 \rangle - \langle \mathbf{q}, \log \mathbf{q} - \log \mathbf{p}_1 \rangle}{\eta} + \frac{\langle \mathbf{p}_0, \log \mathbf{p}_0 - \log \mathbf{p}_1 \rangle}{\eta} \\
&= \frac{1}{\eta} \left(\langle \mathbf{q}, \log \mathbf{p}_1 - \log \mathbf{p}_0 \rangle - \langle \mathbf{p}_0, \log \mathbf{p}_1 - \log \mathbf{p}_0 \rangle \right) \\
&= \frac{1}{\eta} \left(\langle \mathbf{q}, \mathbf{r} \rangle - \langle \mathbf{p}_0, \mathbf{r} \rangle + \text{const.} \cdot \langle \mathbf{q} - \mathbf{p}_0, \mathbf{1} \rangle \right) \\
&(\text{notice } \log \mathbf{p}_1 - \log \mathbf{p}_0 = \eta \mathbf{r} + \text{const.} \cdot \mathbf{1})
\end{aligned}$$

The claim is thus proved.

A second fact we will use is a bound for the KL divergence through the local norm: suppose $z \in [-1, 1]$, then we have

$$\exp(z) - 1 - z \leq z^2 \tag{3.7}$$

and thus

$$\frac{D(\mathbf{p}_0 \parallel \mathbf{p}_1)}{\eta} \leq \eta \sum_{i=1}^{k} \mathbf{p}_0(i) \mathbf{r}^2(i) . \tag{3.8}$$

The proof is due to the fact

$$\begin{aligned}
D(\mathbf{p}_0 \parallel \mathbf{p}_1) &= \log(\langle \mathbf{p}_0, \exp(\eta \mathbf{r}) \rangle) - \langle \mathbf{p}_0, \eta \mathbf{r} \rangle \\
&\leq \langle \mathbf{p}_0, \exp(\eta \mathbf{r}) - 1 \rangle - \langle \mathbf{p}_0, \eta \mathbf{r} \rangle \quad \text{notice } \log(1+z) \leq z, \forall z \geq -1. \\
&= \sum_{i=1}^{k} \mathbf{p}_0(i) \big(\exp(\eta \mathbf{r}(i)) - 1 - \eta \mathbf{r}(i) \big) \quad \text{notice } \exp(z) - 1 - z \leq z^2, \forall z \in [-1, 1] \\
&\leq \eta^2 \sum_{i=1}^{k} \mathbf{p}_0(i) \mathbf{r}^2(i) .
\end{aligned}$$

So far we have derived

$$\langle \mathbf{r}, \mathbf{q} \rangle - \langle \mathbf{r}, \mathbf{p}_0 \rangle \leq \frac{D(\mathbf{q} \parallel \mathbf{p}_0) - D(\mathbf{q} \parallel \mathbf{p}_1)}{\eta} + \eta \sum_{i=1}^{k} \mathbf{p}_0(i) \mathbf{r}^2(i) . \tag{3.9}$$

Recursively using the above (3.9), we can obtain the following telescoping inequality for the EXP3 algorithm

$$\langle \widehat{\mathbf{r}}_t, \mathbf{q} \rangle - \langle \widehat{\mathbf{r}}_t, \mathbf{p}_t \rangle \leq \frac{D(\mathbf{q} \parallel \mathbf{p}_t) - D(\mathbf{q} \parallel \mathbf{p}_{t+1})}{\eta} + \eta \sum_{i=1}^{k} \mathbf{p}_t(i) \widehat{\mathbf{r}}_t^2(i) . \tag{3.10}$$

Notice that due to the inverse propensity rule being unbiased, we have

$$\mathop{\mathbb{E}}_{u_t}[\langle \widehat{\mathbf{r}}_t, \mathbf{q} \rangle] = \langle \mathop{\mathbb{E}}_{u_t}[\widehat{\mathbf{r}}_t], \mathbf{q} \rangle = \langle \mathbf{r}_t, \mathbf{q} \rangle \tag{3.11}$$

$$\mathop{\mathbb{E}}_{u_t}[\langle \widehat{\mathbf{r}}_t, \mathbf{p}_t \rangle] = \langle \mathop{\mathbb{E}}_{u_t}[\widehat{\mathbf{r}}_t], \mathbf{p}_t \rangle = \langle \mathbf{r}_t, \mathbf{p}_t \rangle = \mathop{\mathbb{E}}_{u_t}[\mathbf{r}_t(u_t)] \tag{3.12}$$

and thus

$$\langle \mathbf{r}_t, \mathbf{q} \rangle - \mathop{\mathbb{E}}_{u_t}[\mathbf{r}_t(u_t)] \leq \mathop{\mathbb{E}}_{u_t}[\frac{D(\mathbf{q} \parallel \mathbf{p}_t) - D(\mathbf{q} \parallel \mathbf{p}_{t+1})}{\eta}] + \eta \sum_{i=1}^{k} \mathbf{p}_t(i) \mathop{\mathbb{E}}_{u_t}[\widehat{\mathbf{r}}^2(i)] \tag{3.13}$$

where

$$\mathop{\mathbb{E}}_{u_t}[\widehat{\mathbf{r}}^2(i)] = \frac{\mathbf{r}_t^2(i)}{\mathbf{p}_t(i)} \tag{3.14}$$

and thus

$$\langle \mathbf{r}_t, \mathbf{q} \rangle - \mathop{\mathbb{E}}_{u_t}[\mathbf{r}_t(u_t)] \leq \mathop{\mathbb{E}}_{u_t}[\frac{D(\mathbf{q} \parallel \mathbf{p}_t) - D(\mathbf{q} \parallel \mathbf{p}_{t+1})}{\eta}] + \eta \sum_{i=1}^{k} \mathbf{p}_t(i) \frac{\mathbf{r}_t^2(i)}{\mathbf{p}_t(i)} \tag{3.15}$$

$$\leq \mathop{\mathbb{E}}_{u_t}[\frac{D(\mathbf{q} \parallel \mathbf{p}_t) - D(\mathbf{q} \parallel \mathbf{p}_{t+1})}{\eta}] + \eta k \tag{3.16}$$

and thus marginally we have

$$\langle \mathbf{r}_t, \mathbf{q} \rangle - \mathbb{E}[\mathbf{r}_t(u_t)] \leq \mathbb{E}[\frac{D(\mathbf{q} \parallel \mathbf{p}_t) - D(\mathbf{q} \parallel \mathbf{p}_{t+1})}{\eta}] + \eta \sum_{i=1}^{k} \mathbf{p}_t(i) \frac{\mathbf{r}_t^2(i)}{\mathbf{p}_t(i)} \tag{3.17}$$

$$\leq \mathbb{E}[\frac{D(\mathbf{q} \parallel \mathbf{p}_t) - D(\mathbf{q} \parallel \mathbf{p}_{t+1})}{\eta}] + \eta k \tag{3.18}$$

summing over $t = 1, 2, \ldots, T$

$$\begin{aligned} \mathcal{R}_T^{\text{adv}}(\pi^{EXP3}) &\leq \frac{D(\mathbf{q} \parallel \mathbf{p}_1) - \mathbb{E}[D(\mathbf{q} \parallel \mathbf{p}_{T+1})]}{\eta} + T \cdot \eta k \\ &\leq \frac{\log(k)}{\eta} + \eta \cdot Tk \end{aligned}$$

□

References

Moritz Hardt and Benjamin Recht. *Patterns, Predictions, and Actions: A Story about Machine Learning*. Princeton University Press, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(6):527–535, 1952.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.