Some Notes on Multi-armed Bandits

Kevin Jamieson, University of Washington

These notes were written for myself to refer to while lecturing. They are not a replacement for the course textbooks [Lattimore and Szepesvári, 2020, Bubeck et al., 2012] and may contain errors! I have posted them by request.

1 Introduction

Machine learning, and in particular, supervised learning, is the study of making statistical inferences from previously collected data. Multi-armed bandits is more about an interaction between an agent (algorithm) and an environment where one simultaneously collects data and makes inferences in a closed-loop.

You have n "arms" or actions, representing distributions. "Pulling" an arm represents requesting a sample from that arm.

At each time $t = 1, 2, 3, \dots$

- Algorithm chooses an action $I_t \in \{1, \dots, n\}$
- Observes a reward $X_{I_t,t} \sim P_{I_t}$ where $P_1, \dots, P_n$ are unknown distributions

That is, playing arm i and time s results in a reward $X_{i,s}$ from the ith distribution. In these lectures, all distributions will be Gaussian (or sub-Gaussian) with variance 1 unless otherwise specified. Example of sub-Gaussian distribution is bounded distributions on $[-1, 1]$ or Gaussian $\mathcal{N}(0, 1)$. Formally, a distribution of X is 1-sub-Gaussian if $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2/2)$.

We will find that the means of the distribution are the most pertinent parameters of these distributions. Let $\theta_i^* = \mathbb{E}_{X \sim P_i}[X]$ be the mean of the ith distribution. Define $\Delta_i = \max_{j=1,\dots,n} \theta_j^* - \theta_i^*$. We measure performance of an algorithm in two ways: 1) how much total reward is accumulated, and 2) how many total pulls are required to identify the best mean.

1.1 Regret Minimization

After T time steps, define the *regret* as

$$R_T = \max_{j=1,\dots,n} \theta_j^* T - \mathbb{E}\left[\sum_{t=1}^{T} X_{I_t,t}\right]$$

The goal is to have $R(T) = o(T)$ to achieve sub-linear regret.

If at time T the ith arm has been played T_i times, then

$$\begin{aligned}
R_T &= \max_{j=1,\dots,n} \theta_j^* T - \mathbb{E}\left[\sum_{t=1}^{T} X_{I_t,t}\right] \\
&= \max_{j=1,\dots,n} \theta_j^* T - \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{n} X_{i,t} \mathbf{1}\{I_t = i\}\right] \\
&= \max_{j=1,\dots,n} \theta_j^* T - \sum_{i=1}^{n} \theta_i^* \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{I_t = i\}\right] \\
&= \max_{j=1,\dots,n} \theta_j^* T - \sum_{i=1}^{n} \theta_i^* \mathbb{E}\left[T_i\right] \\
&= \sum_{i=1}^{n} \Delta_i \mathbb{E}\left[T_i\right]
\end{aligned}$$

Thus, we want to minimize the number of times we play sub-optimal arms.

1.2 Best-arm identification

Given a $\delta \in (0,1)$ identify the best arm with probability at least $1-\delta$ using as few total pulls as possible.

While related, these objectives are at odds with one another. Sometimes called the (ϵ, δ)-PAC setting, but for simplicity we'll take $\epsilon = 0$.

1.3 Warm-up

Suppose $n = 2$. How long would it take to decide one arm was better than another using sub-gaussian bounds? Consider the trivial algorithm:

> **Input**: 2 arms, time $\tau \in \mathbb{N}$.
> Pull each arm $i \in \{1,2\}$ exactly τ times and compute empirical mean $\widehat{\theta}_i$.
> For all $t > 2\tau$ play arm $\arg\max_i \widehat{\theta}_i$

Proposition 1. *Fix ϵ, δ. If $Z_1, Z_2, \ldots$ are independent mean-zero σ^2-sub-Gaussian random variables so that $\mathbb{E}[\exp(\lambda Z_t)] \le \exp(\lambda^2\sigma^2/2)$, then for $\tau = \lceil 2\sigma^2\epsilon^{-2}\log(1/\delta)\rceil$ we have $\mathbb{P}(\frac{1}{\tau}\sum_{t=1}^{\tau} Z_t \le \epsilon) \ge 1-\delta$.*

Proof.

$$\begin{aligned}
\mathbb{P}(\frac{1}{\tau}\sum_{t=1}^{\tau} Z_t > \epsilon) &= \mathbb{P}(\exp\left(\lambda\sum_{t=1}^{\tau} Z_t\right) > \exp(\lambda\tau\epsilon)) \\
&\le e^{-\lambda\tau\epsilon}\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{\tau} Z_t\right)\right] \\
&= e^{-\lambda\tau\epsilon}\prod_{t=1}^{\tau}\mathbb{E}\left[\exp\left(\lambda Z_t\right)\right] \\
&= e^{-\lambda\tau\epsilon}\exp\left(\lambda^2\sigma^2\tau/2\right) \\
&= \exp\left(-\lambda\tau\epsilon + \lambda^2\sigma^2\tau/2\right) \\
&\le \exp\left(-\tau\epsilon^2/2\sigma^2\right) \le \delta
\end{aligned}$$

for the chosen τ. □

Set $\tau = \lceil 8\Delta^{-2}\log(4/\delta)\rceil$ and let $\widehat{\theta}_i = \frac{1}{\tau}\sum_{s=1}^{\tau} X_{i,s}$ for $i = 1, 2$. Define the event

$$\mathcal{E}_i := \left\{|\widehat{\theta}_i - \theta_i^*| \le \sqrt{\frac{2\log(4/\delta)}{\tau}}\right\}.$$

Then $\mathbb{P}(\mathcal{E}_1^c \cup \mathcal{E}_2^c) \le \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_1^c) \le \delta$. Thus, if we pull each arm τ times then on $\mathcal{E}_1 \cap \mathcal{E}_2$ we have

$$\begin{aligned}
\widehat{\theta}_1 &> \theta_1^* - \Delta/2 \\
&\ge \theta_2^* + \Delta/2 \\
&> \widehat{\theta}_2
\end{aligned}$$

so that we have determined the best-arm. And we can play it forever.

After any T total plays such that arm i has been played T_i times and $T = T_1 + T_2$, the expected regret is at most

$$\begin{aligned}
\theta_1^* T - \mathbb{E}\left[\sum_{s=1}^{T} X_{I_s,s}\right] &= \theta_1^* T - \mathbb{E}\left[(T_1\theta_1^* + T_2\theta_2^*)\right] \\
&= \mathbb{E}\left[T_2\Delta\right] \\
&= \mathbb{E}\left[T_2\Delta\mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} + T_2\Delta\mathbf{1}\{\mathcal{E}_1^c \cup \mathcal{E}_2^c\}\right] \\
&\le \mathbb{E}\left[\tau\Delta\mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} + T\Delta\mathbf{1}\{\mathcal{E}_1^c \cup \mathcal{E}_2^c\}\right] \\
&\le 8\Delta^{-1}\log(4/\delta) + \Delta T\mathbb{P}(\mathcal{E}_1^c \cup \mathcal{E}_2^c) \\
&\le 8\Delta^{-1}\log(4/\delta) + \Delta T\delta.
\end{aligned}$$

If we take $\delta = 1/T$ then the expected regret is less than $\Delta + 8\Delta^{-1}\log(4T)$. On the other hand, the regret can't possibly be greater than ΔT, thus the total regret is bounded by

$$\begin{aligned}\theta_1^* T - \mathbb{E}\left[\sum_{s=1}^{T} X_{I_s,s}\right] &= \min\{T\Delta, \Delta + 8\Delta^{-1}\log(4T)\} \\ &\le 1 + 2\sqrt{8T\log(4T)}\end{aligned}$$

where the last step takes the worst case $\Delta = \sqrt{8\log(4T)/T}$.
Takeaway: For very small Δ we lose almost nothing, for very large Δ its easy to distinguish, its maximized at around $1/\sqrt{T}$. We'll see this again.

1.4 Disclaimers and goals of these lectures

To goal of these lectures is to provide an overview of the kinds of problems multi-armed bandits can solve, and basic algorithmic tropes. By the end of these lectures I hope you will be able to model a real-world problem as an appropriate bandit problem and propose an algorithm that is a strong baseline.

As a consequence, I will not be stressing important practicalities of algorithms that muck up the analysis. That is, I will pay for a slightly weaker performing algorithm in exchange for a dramatically simpler analysis. This means that I will i) assume the horizon times T for regret analyses is known in advance, ii) I will take liberties with constants and some log factors, making notes of the best known results when pertinent, iii) will always assume there is a uniquely optimal arm for best-arm identification and aim to find it, not merely an ϵ-good arm, iv) I will only study best-arm identification in the fixed confidence setting (exists a whole literature on fixed budget setting), v) will assume sub-Gaussian distributions everywhere and nothing more.

2 Action Elimination Algorithm for Multi-armed Bandits

Input: n arms $\mathcal{X} = \{1, \dots, n\}$, confidence level $\delta \in (0, 1)$.
Let $\widehat{\mathcal{X}}_1 \leftarrow \mathcal{X}, t \leftarrow 1$
while $|\widehat{\mathcal{X}}_\ell| > 1$ **do**
  $\epsilon_t = 2^{-t}$
  Pull each arm in $\widehat{\mathcal{X}}_\ell$ exactly $\tau_\ell = \lceil 2\epsilon_\ell^{-2}\log(\frac{4\ell^2|\mathcal{X}|}{\delta})\rceil$ times
  Compute the empirical mean of these rewards $\widehat{\theta}_{i,\ell}$ for all $i \in \widehat{\mathcal{X}}_\ell$
  $\widehat{\mathcal{X}}_{\ell+1} \leftarrow \widehat{\mathcal{X}}_\ell \setminus \left\{i \in \widehat{\mathcal{X}}_\ell : \max_{j\in\widehat{\mathcal{X}}_\ell} \widehat{\theta}_{j,\ell} - \widehat{\theta}_{i,\ell} > 2\epsilon_\ell\right\}$
  $t \leftarrow t + 1$
Output: $\widehat{\mathcal{X}}_{t+1}$

Lemma 1. *Assume that* $\max_{i\in\mathcal{X}} \Delta_i \le 4$. *With probability at least* $1-\delta$, *we have* $1 \in \widehat{\mathcal{X}}_\ell$ *and* $\max_{i\in\widehat{\mathcal{X}}_\ell} \Delta_i \le 8\epsilon_\ell$ *for all* $\ell \in \mathbb{N}$.

Proof. For any $\ell \in \mathbb{N}$ and $i \in [n]$ define

$$\mathcal{E}_{i,\ell} = \left\{|\widehat{\theta}_{i,\ell} - \theta_i^*| \le \epsilon_\ell\right\}$$

and $\mathcal{E} = \bigcap_{i=1}^n \bigcap_{\ell=1}^\infty \mathcal{E}_{i,\ell}$. Noting that $\epsilon_\ell = \sqrt{\frac{2\log(4n\ell^2/\delta)}{\tau_\ell}}$ we have

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\left(\bigcup_{i=1}^n \bigcup_{\ell=1}^\infty \mathcal{E}_{i,\ell}^c\right) \le \sum_{i=1}^n \sum_{\ell=1}^\infty \frac{\delta}{2n\ell^2} \le \delta.$$

In what follows assume $\mathcal{E}$ holds.

Fix any ℓ for which $1 \in \widehat{\mathcal{X}}_\ell$ (note $1 \in \widehat{\mathcal{X}}_1$). Then for any $j \in \widehat{\mathcal{X}}_\ell$ we have

$$\begin{aligned}\widehat{\theta}_{j,\ell} - \widehat{\theta}_{1,\ell} &= (\widehat{\theta}_{j,\ell} - \theta_j^*) - (\widehat{\theta}_{1,\ell} - \theta_1^*) - \Delta_\ell \\ &\overset{\mathcal{E}}{\le} 2\epsilon_\ell\end{aligned}$$

which implies $1 \in \widehat{\mathcal{X}}_{\ell+1}$. Thus, $1 \in \widehat{\mathcal{X}}_\ell$ for all ℓ. On the other hand, any i for which $\Delta_i = \theta_1^* - \theta_i^* > 4\epsilon_\ell$ we have

$$\begin{aligned}\max_{j \in \widehat{\mathcal{X}}_\ell} \widehat{\theta}_{j,\ell} - \widehat{\theta}_{i,\ell} &\geq \widehat{\theta}_{1,\ell} - \widehat{\theta}_{i,\ell} \\ &= (\widehat{\theta}_{j,\ell} - \theta_j) - (\widehat{\theta}_{i,\ell} - \theta_i) - \Delta_i \\ &> 2\epsilon_\ell - 4\epsilon_\ell = 2\epsilon_\ell\end{aligned}$$

which implies this $\max_{j \in \widehat{\mathcal{X}}_{\ell+1}} \theta_j^* \geq \theta_1^* - 4\epsilon_\ell = \theta_1^* - 8\epsilon_{\ell+1}$. □

Theorem 1. *Assume that* $\max_{i \in \mathcal{X}} \Delta_i \leq 4$. *Then with probability at least* $1 - \delta$, 1 *is returned from the algorithm at a time* τ *that satisfies*

$$\tau \leq c \sum_{i=2}^n \Delta_i^{-2} \log(n \log(\Delta_i^{-2})/\delta)$$

Proof. If $\Delta = \min_{i \neq 1} \Delta_i$ then $\widehat{\mathcal{X}}_\ell = \{1\}$ for $t \geq \lceil \log_2(8\Delta) \rceil$. Note that

$$\begin{aligned}T_i &= \sum_{\ell=1}^{\lceil \log_2(8\Delta) \rceil} \tau_\ell \mathbf{1}\{i \in \widehat{\mathcal{X}}_\ell\} \\ &\leq \sum_{\ell=1}^{\lceil \log_2(8\Delta) \rceil} \tau_\ell \mathbf{1}\{\Delta_i \leq 8\epsilon_\ell\} \\ &= \sum_{\ell=1}^{\lceil \log_2(8\Delta_i^{-1}) \rceil} \tau_\ell \\ &= \sum_{\ell=1}^{\lceil \log_2(8\Delta_i^{-1}) \rceil} \lceil 2\epsilon_\ell^{-2} \log(\tfrac{4\ell^2|\mathcal{X}|}{\delta}) \rceil \\ &\leq \lceil 2 \log(\tfrac{4 \log_2^2(16\Delta_i^{-2})|\mathcal{X}|}{\delta}) \rceil \sum_{\ell=1}^{\lceil \log_2(8\Delta_i^{-1}) \rceil} 4^\ell \\ &\leq c\Delta_i^{-2} \log(\tfrac{4 \log_2^2(16\Delta_i^{-2})|\mathcal{X}|}{\delta})\end{aligned}$$

which implies

Thus, the total number of samples taken before $\widehat{\mathcal{X}}_\ell = \{1\}$ is equal to

$$\sum_{i=1}^n T_i \leq \sum_{i=1}^n c\Delta_i^{-2} \log(\tfrac{4 \log_2^2(16\Delta_i^{-2})|\mathcal{X}|}{\delta})$$

which implies that one can identify the best arm after no more than $\sum_{i=2}^n \Delta_i^{-2} \log(n \log(\Delta_i^{-2})/\delta)$. □

Theorem 2. *Assume that* $\max_{i \in \mathcal{X}} \Delta_i \leq 4$. *For any* $T \in \mathbb{N}$, *with probability at least* $1 - \delta$

$$\sum_{i:\Delta_i > 0} T_i \Delta_i \leq \inf_{\nu \geq 0} \nu T + \sum_{i=1}^n c(\Delta_i \vee \nu)^{-1} \log(\tfrac{\log((\Delta_i \vee \nu)^{-1})|\mathcal{X}|}{\delta}).$$

Moreover, if the algorithm is run with $\delta = 1/T$ *then* $R_T \leq c \sum_{i=2}^n \Delta_i^{-1} \log(T)$ *and* $R_T \leq c\sqrt{nT \log(T)}$.

Suppose you run for T timesteps. Then for any $\nu \geq 0$ the regret is bounded by:

$$\begin{aligned}
\sum_{i=2}^{n} \Delta_i T_i &\leq \nu T + \sum_{\ell : 8\epsilon_\ell > \nu} 8\epsilon_\ell \tau_\ell |\widehat{\mathcal{X}}_\ell| \\
&\leq \nu T + \sum_{\ell=1}^{\lceil \log_2(8(\Delta \vee \nu)^{-1}) \rceil} 8\epsilon_\ell \tau_\ell |\widehat{\mathcal{X}}_\ell| \\
&\leq \nu T + \sum_{i=1}^{n} \sum_{\ell=1}^{\lceil \log_2(8(\Delta \vee \nu)^{-1}) \rceil} 8\epsilon_\ell \tau_\ell \mathbf{1}\{\Delta_i \leq 8\epsilon_\ell\} \\
&= \nu T + \sum_{i=1}^{n} \sum_{\ell=1}^{\lceil \log_2(8(\Delta_i \vee \nu)^{-1}) \rceil} 8\epsilon_\ell \tau_\ell \\
&= \nu T + \sum_{i=1}^{n} \sum_{\ell=1}^{\lceil \log_2(8(\Delta_i \vee \nu)^{-1}) \rceil} 8\epsilon_\ell \lceil 2\epsilon_\ell^{-2} \log(\tfrac{4\ell^2 |\mathcal{X}|}{\delta}) \rceil \\
&\leq \nu T + \sum_{i=1}^{n} c \log(\tfrac{4 \log_2^2(8(\Delta_i \vee \nu)^{-2}) |\mathcal{X}|}{\delta}) \sum_{\ell=1}^{\lceil \log_2(8(\Delta_i \vee \nu)^{-1}) \rceil} 2^\ell \\
&\leq \nu T + \sum_{i=1}^{n} c(\Delta_i \vee \nu)^{-1} \log(\tfrac{\log((\Delta_i \vee \nu)^{-1}) |\mathcal{X}|}{\delta})
\end{aligned}$$

Setting $\nu = 0$ yields a regret of $\sum_{i=2}^{n} \Delta_i^{-1} \log(n \log(\Delta_i^{-1})/\delta)$. On the other hand, using $\Delta_i \vee \nu \geq \nu$ and minimizing over ν yields a regret of $\sqrt{nT \log(n \log(T)/\delta)}$. The expected regret, of course, is then bounded by

$$\begin{aligned}
\sum_{i=2}^{n} \Delta_i \mathbb{E}[T_i] &= \mathbb{E}\left[\sum_{i=2}^{n} \Delta_i T_i\right] \\
&\leq \sum_{i=2}^{n} \Delta_i^{-1} \log(n \log(\Delta_i^{-1})/\delta) + T\mathbb{P}(\mathcal{E}^c)
\end{aligned}$$

Setting $\delta = 1/T$ implies the regret is less than $\sum_{i=2}^{n} c\Delta_i^{-1} \log(T)$.
Some remarks:

- This analysis doesn't reuse samples from previous rounds, it is easy to make this change.
- Regret bound requires knowledge of T a priori.

3 Lower bounds for Multi-armed Bandits

Let us briefly pause to consider how far off from optimal we are, and then think about an algorithm that could get us to optimality. How do we know we're doing okay?

3.1 Mean of a Gaussian

Suppose I get n samples from a Gaussian distribution $\mathcal{N}(\mu, 1)$. You compute the empirical mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We know that $|\widehat{\mu} - \mu| \leq \sqrt{2\log(2/\delta)/n}$. How tight is this? If $\mu \in \{0, \Delta\}$ then we just need $n = 8\Delta^{-2}\log(2/\delta)$[1]

[1] Using the SPRT, as $\delta \to 0$ one needs just an expected number of samples equal to $2\Delta^{-2}\log(2/\delta)$.

Let $p_\mu(x) = \frac{1}{2\pi} e^{-(x-\mu)^2/2\sigma^2}$ be the Gaussian distribution with mean μ. Under H_0, $X_i \sim p_0$ and under H_1, $X_i \sim p_\Delta$. Let $\phi : \mathbb{R}^n \to \{0, \Delta\}$. Then the minimax probability of error is equal to

$$\begin{aligned}
\inf_\phi \max\{\mathbb{P}_0(\phi = 1), \mathbb{P}_1(\phi = 0)\} &\geq \inf_\phi \frac{1}{2}(\mathbb{P}_0(\phi = 1), \mathbb{P}_1(\phi = 0)) \\
&= \inf_\phi \frac{1}{2}\left(\int_{x\in\mathbb{R}^n} \mathbf{1}\{\phi(x) = 1\} p_0(x)dx + \int_{x\in\mathbb{R}^n} \mathbf{1}\{\phi(x) = 0\} p_1(x)dx\right) \\
&= \frac{1}{2}\int_{x\in\mathbb{R}^n} \min\{p_0(x), p_1(x)\}dx \\
&\geq \frac{1}{4}\left(\int_{x\in\mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx\right)^2 \qquad \text{(Cauchy-Schwartz)} \\
&\geq \frac{1}{4}\exp\left(-\int_{x\in\mathbb{R}^n} \log(\tfrac{p_1(x)}{p_0(x)}) p_1(x)dx\right) \qquad \text{(Jensen's)}
\end{aligned}$$

where

$$\begin{aligned}
\left(\int_{x\in\mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx\right)^2 &= \left(\int_{x\in\mathbb{R}^n} \sqrt{\min\{p_0(x), p_1(x)\}\max\{p_0(x), p_1(x)\}}dx\right)^2 \\
&\leq \int_{x\in\mathbb{R}^n} \min\{p_0(x), p_1(x)\}dx \int_{x\in\mathbb{R}^n} \max\{p_0(x), p_1(x)\}dx \qquad \text{(Cauchy-Schwartz)} \\
&\leq 2\int_{x\in\mathbb{R}^n} \min\{p_0(x), p_1(x)\}dx
\end{aligned}$$

and (integrating only over support of pq)

$$\begin{aligned}
\left(\int_{x\in\mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx\right)^2 &= \exp\left(2\log(\int_{x\in\mathbb{R}^n} p_0(x)\sqrt{p_1(x)/p_0(x)}dx)\right) \\
&\geq \exp\left(2\int_{x\in\mathbb{R}^n} p_0(x)\log(\sqrt{p_1(x)/p_0(x)})dx\right) \\
&= \exp\left(-\int_{x\in\mathbb{R}^n} \log(\tfrac{p_1(x)}{p_0(x)}) p_1(x)dx\right)
\end{aligned}$$

Note that

$$\begin{aligned}
KL(\mathbb{P}_1|\mathbb{P}_0) &= \int_x \log\left(\prod_{i=1}^n \frac{p_1(x_i)}{p_0(x_i)}\right)\prod_{i=1}^n p_1(x_i)dx \\
&= nKL(p_1|p_0) = n\Delta^2/2
\end{aligned}$$

and that $KL(\mathcal{N}(0,1)|\mathcal{N}(\Delta,1)) = \Delta^2/2$.

We conclude that

$$\inf_\phi \max\{\mathbb{P}_0(\phi = 1), \mathbb{P}_1(\phi = 0)\} \geq \frac{1}{4}\exp\left(-n\Delta^2/2\right)$$

Thus, to determine whether or not n samples are from a Gaussian with mean 0 or Δ with probability of failure less than δ, one needs $n \geq 2\Delta^{-2}\log(1/4\delta)$.

3.2 Indentification

An algorithm for best-arm identification at time t is described by given a history $(I_s, X_s)_{s<t}$ for each time t is described by a

- **selection rule** $I_t \in [n]$ is $\mathcal{F}_{t-1}$ measurable where $\mathcal{F}_t = \sigma(I_1, X_1, I_2, X_2, \ldots, I_{t-1}, X_{t-1})$
- **stopping time** τ is $\mathcal{F}_t$ measurable, and

- **recommendation rule** $\widehat{i} \in [n]$ invoked at time τ which is $\mathcal{F}_\tau$-measurable.

Definition 1. *We say that an algorithm for best-arm identification is δ-PAC if for all $\theta^* \in \mathbb{R}^n$ we have $\mathbb{P}_{\theta^*}(\widehat{i} = \arg\max_{i\in[n]} \theta_i^*) \geq 1 - \delta$.*

The following is due to [Kaufmann et al., 2016], a strengthening of the first time it appeared in [Mannor and Tsitsiklis, 2004

Theorem 3 (Best-arm identification lower bound)**.** *Any algorithm that is δ-PAC on $\{P : P_i = \mathcal{N}(\theta_i, 1), \theta_1 > \max_{i\neq 1} \theta_i, \theta \in [0,1]^n\}$ for $\delta < 0.15$ satisfies $\mathbb{E}_{\theta^*}[\tau] \geq 2\log(\frac{1}{2.4\delta}) \sum_{i=1}^n \Delta_i^{-2}$.*

Proof sketch: The original instance has $P_i = \mathcal{N}(\theta_i^*, 1)$. Pick some $j \in [n]$ and define an alternative mean vector $\theta^{(j)} \in [0,1]^n$ such that $\theta_i^{(j)} = \theta_i^*$ if $i \neq j$ and $\theta_i^{(j)} = \theta_1 + \epsilon$ for $j = i$ for some arbitrarily small number ϵ. Note that under $\theta^{(j)}$, arm j is the best arm.

Because the algorithm claims to be δ-PAC, it has to output arm 1 under θ^* and arm j under $\theta^{(j)}$. But these two bandit games only differ on arm j so to tell the difference between them its only natural to sample arm j until one can figure out which instance is being played (i.e., is its mean θ_j or $\theta_1 + \epsilon$?) The discussion above suggests that to make this distinction with probability at least $1 - \delta$, it is necessary to sample arm j at least $2(\theta_1 - \theta_j + \epsilon)^{-2}\log(1/4\delta)$ times. Taking ϵ to zero and noticing that j was arbitrary completes the sketch.

This is *not* a proof, however, because the number of times the algorithm samples arm j is random whereas in the above argument it was fixed. The proof of [Kaufmann et al., 2016] provides convenient tools to prove general lower bounds for δ-PAC settings.

3.3 Regret, minimax

Theorem 4 (Minimax regret lower bound)**.** *For every $T \geq n$ there exists an instance $P = \mathcal{N}(\theta^*, I)$ such that $R_T \geq \sqrt{(n-1)T}/27$.*

Proof sketch: Let $\theta^* = \theta = (\Delta, 0, \ldots, 0)$. For any algorithm, by the pigeon hole principle, there exists an arm $\widehat{i} \in [n]$ such that $\mathbb{E}[T_{\widehat{i}}] \leq T/n$.

Define an alternative Gaussian instance with mean vector θ' that is identical to θ other than $\theta_{\widehat{i}} = 2\Delta$.

If $\Delta \approx \sqrt{n/T}$ then $\widehat{i}$ will not be given enough samples to distinguish between the two instances, which means $\mathbb{E}[T_1]$ will be about the same under both models.

Under θ, if $\mathbb{E}[T_1] \leq T/2$ then the regret incurred is at least $\Delta T/2 \approx \sqrt{nT}$. On the other hand, under θ', if $\mathbb{E}[T_1] > T/2$ then the regret again is at least $\Delta T/2 \approx \sqrt{nT}$.

This is *not* a proof because again the number of times an arm is pulled is random, but as before, these arguments can be made precise.

3.4 Gap-dependent regret

Lemma 2. *Any strategy that satisfies $\mathbb{E}[T_i(t)] = o(t^a)$ for any arm i with $\Delta_i > 0$ and $a \in (0,1)$, we have that $\lim_{T\to\infty} \inf \frac{\bar{R}_T}{\log(T)} = \sum_{i=2}^n \frac{2}{\Delta_i}$.*

Takeaway: This is what his field does: prove an initial upper, then lower, then chase it.

3.5 Revisiting MAB with Optimism

Why go beyond action elimination algorithms? Because they will never hit the asymptotic lower bound, for one thing, since if we look at when the second to last arm exits, the lowerbounds are the same.

α-UCB which is $\arg\max_i \widehat{\theta}_{i,T_i(t)} + \sqrt{\frac{2\alpha\log(t)}{T_i(t)}}$ as $\alpha \to 1$ achieves the lower bound.

Any sub-linear regret algorithm plays arm 1 an infinite number of times, so assume $\widehat{\mu}_1 \approx \mu_1$. Minimizing the maximum upper bound. Thus, we expect the number of times the ith arm is pulled is $2\Delta_i^{-2}\log(T)$, which is optimal.

UCB1 in its most popular form was developed by [Auer et al., 2002].

MOSS first achieved $\sqrt{nT}$ regret [Audibert and Bubeck, 2009].

KL-UCB is finite-time analysis with optimal constants for asymptotic regret [Cappé et al., 2013].

The recent work of [Lattimore, 2018] defined a UCB-based algorithm that achieves asymptotic optimal constants, and finite regret bounds of $\sum_i \frac{\log(T)}{\Delta_i^{-1}}$ and $\sqrt{nT}$.

4 Linear Bandits Intro

Now suppose each arm $i = 1, \dots, n$ has a feature vectors $x_i \in \mathbb{R}^d$. And more over, there exists some $\theta^* \in \mathbb{R}^d$ such that a pull of arm $I_t \in [n]$ results in a reward $y_t = \langle x_{I_t}, \theta^* \rangle + \eta_t$ where $\eta_t \sim \mathcal{N}(0, 1)$.

Applications: Drug-discovery, Spotify, Netflix, ads

In the previous setup, pulling arm i provided no information about arm j, but now suddenly it does.

4.1 Least Squares

Given a sequence of arm choices and observed rewards let $\{x_t, y_t, \eta_t\}_{t=1}^{\tau}$ we denote the stacked sequences of each as $X \in \mathbb{R}^{\tau \times d}, Y \in \mathbb{R}^{\tau}$, and $\eta \in \mathbb{R}^{\tau}$ respectively where $Y = X\theta^* + \eta$. Using this information we can derive a least-squares estimate of θ_* given as follows

$$\hat{\theta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\theta_* + \eta) = \theta_* + (X^T X)^{-1} X^T \eta.$$

Fix any $z \in \mathbb{R}^d$, then Thus

$$z^\top (\hat{\theta} - \theta_*) = z^\top (X^\top X)^{-1} X^\top \eta.$$

Note that $\eta \sim \mathcal{N}(0, I)$. For any $W \sim \mathcal{N}(\mu, \Sigma)$ we have $AW + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$. Thus

$$z^\top (\hat{\theta} - \theta_*) \sim \mathcal{N}(0, z^\top (X^\top X)^{-1} z).$$

so that

$$\mathbb{P}\left(z^\top (\hat{\theta} - \theta_*) \geq \sqrt{2 z^\top (X^\top X)^{-1} z \log(1/\delta)} \right) \leq \delta.$$

We will use the notation $\|z\|_A^2 = z^\top A z$ so that with probability at least $1 - \delta$

$$z^\top (\hat{\theta} - \theta_*) \leq \|z\|_{(X^\top X)^{-1}} \sqrt{2 \log(1/\delta)}$$

4.1.1 Aside: Gaussian to sub-Gaussian

For an arbitrary constant μ,

$$\begin{aligned}
P(x^T(\hat{\theta} - \theta_*) > \mu) &= P(w^T \eta > \mu) \\
&\leq \exp(-\lambda\mu)\mathbb{E}[\exp(\lambda w^T \eta)], \quad \text{let } \lambda > 0 && \text{Chernoff Bound} \\
&= \exp(-\lambda\mu)\mathbb{E}[\exp(\lambda \sum_{i=1}^{t} w_i \eta_i)] \\
&= \exp(-\lambda\mu) \prod_{i=1}^{t} \mathbb{E}[\exp(\lambda w_i \eta_i)] && \text{independence of } w_i\eta_i \\
&\leq \exp(-\lambda\mu) \prod_{i=1}^{t} \exp(\lambda^2 w_i^2/2) && \text{sub-Gaussian assumption} \\
&= \exp(-\lambda\mu) \exp(\frac{\lambda^2}{2} ||w||_2^2) \\
&\leq \exp(-\frac{\mu^2}{2||w||_2^2}) && \lambda = \frac{\mu}{||w||_2^2} \\
&= \exp(-\frac{\mu^2}{2x^T(X^TX)^{-1}x}) = \delta,
\end{aligned}$$

where in the final step we made use of the following equality

$$||w||_2^2 = x^T(X^TX)^{-1}X^TX(X^TX)^{-1}x = x^T(X^TX)^{-1}x.$$

Thus with probability at least $1-\delta$,

$$\begin{aligned} x^T(\hat{\theta}-\theta_*) &\le \sqrt{2x^T(X^\top X)^{-1}x\log(\frac{1}{\delta})} \\ &=: \|x\|_{(X^\top X)^{-1}}\sqrt{2\log(1/\delta)} \end{aligned}$$

5 Experimental design

Note that if I take measurements $(x_1,\dots,x_n) \in \mathcal{X}$ and observe their corresponding observations $y_i = \langle x_i, \theta^*\rangle + \eta_i$ where $\eta_i \in \prime, \infty$, then $\mathbb{E}[(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^\top] = \sigma^2(X^TX)^{-1}$ and also, $\widehat{\theta}-\theta^* \sim \mathcal{N}(0,\sigma^2(X^\top X)^{-1})$. We can visualize this as a confidence ellipsoid for each choice of X. And we can even think of optimizing the choice. Recall that the PDF of a Gaussian is $\phi(x) = \frac{1}{(2\pi|\Sigma|)^{d/2}}e^{-x^\top\Sigma^{-1}x/2}$. With entropy $\frac{1}{2}\log(2\pi e|\Sigma|)$.

When the number of selected points is large, its more convenient to think of sampling n points from a distribution placed over $\mathcal{X}$. Define

$$A_\lambda = \sum_{x\in\mathcal{X}} \lambda_x xx^\top$$

so that for every $X \in \mathbb{R}^{\tau\times d}$ there exists some $\lambda \in \triangle_\mathcal{X}$ such that $X^\top X = \sum_{x\in\mathcal{X}}\lceil\lambda_x\tau\rceil xx^\top = A_\lambda$. This A_λ can then be used to shape the covariance $\widehat{\theta}$:

- **A-optimality**: minimize $f_A(\lambda) = \mathrm{Tr}(A_\lambda^{-1})$ minimizes $\mathbb{E}[\|\widehat{\theta}-\theta\|_2^2]$
- **E-optimality**: minimize $f_E(\lambda) = \max_{u:\|u\|\le 1} u^\top A_\lambda^{-1}u$ minimizes $\max_{u:\|u\|\le 1}\mathbb{E}[(\langle u,\widehat{\theta}-\theta\rangle)^2]$
- **D-optimality**: maximize $g_D(\lambda) = \log(|A_\lambda|)$ maximizes the entropy of distribution. Also, if $\mathcal{E}_\lambda = \{x : x^\top A_\lambda^{-1}x \le d\}$ then D-optimality is the minimum volume ellipsoid that contains $\mathcal{X}$.
- **G-optimality**: minimize $f_G(\lambda) = \max_{x\in\mathcal{X}} x^\top A_\lambda^{-1}x$ minimizes $\max_{x\in\mathcal{X}}\mathbb{E}[(\langle x,\widehat{\theta}-\theta^*\rangle)^2]$

Lemma 3 (Kiefer-Wolfowitz (1960))**.** *For any $\mathcal{X}$ with $d = dim(span(\mathcal{X}))$, there exists a $\lambda^* \in \triangle_\mathcal{X}$ that*

- $\max_\lambda g_D(\lambda) = g_D(\lambda^*)$
- $\min_\lambda f_G(\lambda) = f_G(\lambda^*)$
- $f_G(\lambda^*) = g_D(\lambda^*) = d$
- $support(\lambda^*) = (d+1)d/2$

Proposition 2. *If λ^* is the G-optimal design for $\mathcal{X}$ then if we pull arm $x\in\mathcal{X}$ exactly $\lceil\tau\lambda_x^*\rceil$ times for some $\tau > 0$ and compute the least squares estimator $\widehat{\theta}$. Then for each $x\in\mathcal{X}$ we have with probability at least $1-\delta$*

$$\begin{aligned} \langle x,\widehat{\theta}-\theta^*\rangle &\le \|x\|_{(\sum_{x\in\mathcal{X}}\lceil\tau\lambda_x^*\rceil xx^\top)^{-1}}\sqrt{2\log(1/\delta)} \\ &\le \frac{1}{\sqrt{\tau}}\|x\|_{(\sum_{x\in\mathcal{X}}\lambda_x^* xx^\top)^{-1}}\sqrt{2\log(1/\delta)} \\ &\le \sqrt{\frac{2d\log(1/\delta)}{\tau}} \end{aligned}$$

and we have taken at most $\tau + \frac{d(d+1)}{2}$ pulls. Thus, for any $\delta' \in (0,1)$ we have $\mathbb{P}(\bigcup_{x\in\mathcal{X}}\{|\langle x,\widehat{\theta}-\theta^*\rangle| > \sqrt{\frac{2d\log(2|\mathcal{X}|/\delta')}{\tau}}\}) \le \delta'$.

Notes:

- The support size of $(d+1)d/2$ is trivial application of Caratheodory's theorem. Many algorithms to find this efficiently.
- Note that one can find a λ^* with a constant approximation with just support $O(d)$.
- Leverage scores if V-optimality
- John's ellipsoid is equivalent to G/D-optimality

[Pukelsheim, 2006, Yu et al., 2006]. [Yu et al., 2006, Soare et al., 2014, Soare, 2015, Lattimore and Szepesvari, 2017],

6 Linear Bandits: Regret Minimization

This section is inspired by [Lattimore and Szepesvári, 2020].

Input: Finite set $\mathcal{X} \subset \mathbb{R}^d$, confidence level $\delta \in (0,1)$.
Let $\widehat{\mathcal{X}}_1 \leftarrow \mathcal{X}, \ell \leftarrow 1$
while $|\widehat{\mathcal{X}}_\ell| > 1$ **do**
  Let $\widehat{\lambda}_\ell \in \triangle_{\widehat{\mathcal{X}}_\ell}$ be a $\frac{d(d+1)}{2}$-sparse minimizer of $f(\lambda) = \max_{x \in \widehat{\mathcal{X}}_\ell} \|x\|^2_{(\sum_{x \in \widehat{\mathcal{X}}_\ell} \lambda_x x x^\top)^{-1}}$
  $\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = 2d\epsilon_\ell^{-2} \log(4\ell^2|\mathcal{X}|/\delta)$
  Pull arm $x \in \mathcal{X}$ exactly $\lceil \widehat{\lambda}_{\ell,x} \tau_\ell \rceil$ times and construct the least squares estimator $\widehat{\theta}_\ell$ using only the observations of this round
  $\widehat{\mathcal{X}}_{\ell+1} \leftarrow \widehat{\mathcal{X}}_\ell \setminus \left\{ x \in \widehat{\mathcal{X}}_\ell : \max_{x' \in \widehat{\mathcal{X}}_\ell} \langle x' - x, \widehat{\theta}_\ell \rangle > 2\epsilon_\ell \right\}$
  $\ell \leftarrow \ell + 1$
Output: $\widehat{\mathcal{X}}_\ell$

After T time steps, define the *regret* as

$$\begin{aligned} R_T &= \langle x^\star, \theta^* \rangle - \mathbb{E}\left[\sum_{t=1}^T \langle x_t, \theta^* \rangle \right] \\ &= \mathbb{E}\left[\sum_{x \neq x^\star} T_x \Delta_x \right] \end{aligned}$$

where $\Delta_x = \langle x^\star - x, \theta^* \rangle$.

Lemma 4. *Assume that* $\max_{x \in \mathcal{X}} \langle x^\star - x, \theta^* \rangle \leq 4$. *With probability at least* $1-\delta$, *we have* $x^\star \in \widehat{\mathcal{X}}_\ell$ *and* $\max_{x \in \widehat{\mathcal{X}}_\ell} \langle x^\star - x, \theta^* \rangle \leq 8\epsilon_\ell$ *for all* $\ell \in \mathbb{N}$.

Proof. For any $\mathcal{V} \subseteq \mathcal{X}$ and $x \in \mathcal{V}$ define

$$\mathcal{E}_{x,\ell}(\mathcal{V}) = \{ |\langle x, \widehat{\theta}_\ell - \theta^* \rangle| \leq \epsilon_\ell \}$$

where it is implicit that $\widehat{\theta}_\ell$ is the G-optimal design constructed in the algorithm at stage ℓ with respect to $\widehat{\mathcal{X}}_\ell = \mathcal{V}$. Note that this is precisely the analogous events of multi-armed bandits. The key piece of the

analysis is that

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\ell=1}^{\infty}\bigcup_{x\in\widehat{\mathcal{X}}_\ell}\{\mathcal{E}^c_{x,\ell}(\widehat{\mathcal{X}}_\ell)\}\right) &\le \sum_{\ell=1}^{\infty}\mathbb{P}\left(\bigcup_{x\in\widehat{\mathcal{X}}_\ell}\{\mathcal{E}^c_{x,\ell}(\widehat{\mathcal{X}}_\ell)\}\right) \\
&= \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{X}}\mathbb{P}\left(\bigcup_{x\in\mathcal{V}}\{\mathcal{E}^c_{x,\ell}(\mathcal{V})\}, \widehat{\mathcal{X}}_\ell=\mathcal{V}\right) \\
&= \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{X}}\mathbb{P}\left(\bigcup_{x\in\mathcal{V}}\{\mathcal{E}^c_{x,\ell}(\mathcal{V})\}\right)\mathbb{P}(\widehat{\mathcal{X}}_\ell=\mathcal{V}) \\
&\le \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{X}}\tfrac{\delta|\mathcal{V}|}{2\ell^2|\mathcal{X}|}\mathbb{P}(\widehat{\mathcal{X}}_\ell=\mathcal{V}) \le \delta
\end{aligned}$$

Thus, in what follows, assume $\mathcal{E} := \bigcap_{x\in\mathcal{X}}\bigcap_{\ell=1}^{\infty}\{\mathcal{E}_{x,\ell}(\widehat{\mathcal{X}}_\ell)\}$ holds.

Fix any ℓ for which $x^\star \in \widehat{\mathcal{X}}_\ell$ (note $x^\star \in \widehat{\mathcal{X}}_1$). Then for any $x \in \widehat{\mathcal{X}}_\ell$ we have

$$\begin{aligned}
\langle x - x^\star, \widehat{\theta}_\ell\rangle &= \langle x, \widehat{\theta}_\ell - \theta^*\rangle - \langle x^\star, \widehat{\theta}_\ell - \theta^*\rangle + \langle x - x^\star, \theta^*\rangle \\
&\le 2\epsilon_\ell
\end{aligned}$$

which implies $x^\star \in \widehat{\mathcal{X}}_{\ell+1}$. Thus, $x^\star \in \widehat{\mathcal{X}}_\ell$ for all ℓ. On the other hand, any x for which $\langle x^\star - x, \theta^*\rangle > 4\epsilon_\ell$ we have

$$\begin{aligned}
\max_{x'\in\widehat{\mathcal{X}}_\ell}\langle x' - x, \widehat{\theta}_\ell\rangle &\ge \langle x^\star - x, \widehat{\theta}_\ell\rangle \\
&= \langle x^\star, \widehat{\theta}_\ell - \theta^*\rangle - \langle x, \widehat{\theta}_\ell - \theta^*\rangle + \langle x^\star - x, \theta^*\rangle \\
&> 2\epsilon_\ell
\end{aligned}$$

which implies $\max_{x\in\widehat{\mathcal{X}}_{\ell+1}}\langle x, \theta^*\rangle \ge \langle x^\star, \theta^*\rangle - 4\epsilon_\ell = \langle x^\star, \theta^*\rangle - 8\epsilon_{\ell+1}$. □

For any $\ell \ge \lceil \log_2(8\Delta^{-1})\rceil$ we have that $\widehat{\mathcal{X}}_\ell = \{x^\star\}$. Suppose you run for T timesteps. Then for any $\nu \ge 0$ the regret is bounded by:

$$\begin{aligned}
\sum_{x\in\mathcal{X}\setminus x^\star}\Delta_x T_x &= T\nu + \sum_{\ell=1}^{\lceil\log_2(8(\Delta\vee\nu)^{-1})\rceil} 8\epsilon_\ell(|\text{support}(\widehat{\lambda}_\ell)| + \tau_\ell) \\
&= T\nu + \sum_{\ell=1}^{\lceil\log_2(8(\Delta\vee\nu)^{-1})\rceil} 8\epsilon_\ell(\tfrac{(d+1)d}{2} + 2d\epsilon_\ell^{-2}\log(4\ell^2|\mathcal{X}|/\delta)) \\
&\le T\nu + 4(d+1)d\lceil\log_2(8(\Delta\vee\nu)^{-1})\rceil + \sum_{\ell=1}^{\lceil\log_2(8(\Delta\vee\nu)^{-1})\rceil} 16d\epsilon_\ell^{-1}\log(4\ell^2|\mathcal{X}|/\delta) \\
&\le T\nu + 4(d+1)d\lceil\log_2(8(\Delta\vee\nu)^{-1})\rceil + 16d\log(4\log_2^2(16(\Delta\vee\nu)^{-1})|\mathcal{X}|/\delta)\sum_{\ell=1}^{\lceil\log_2(8(\Delta\vee\nu)^{-1})\rceil} 2^\ell \\
&\le T\nu + 4(d+1)d\lceil\log_2(8(\Delta\vee\nu)^{-1})\rceil + 512d(\Delta\vee\nu)^{-1}\log(4\log_2^2(16(\Delta\vee\nu)^{-1})|\mathcal{X}|/\delta)
\end{aligned}$$

Setting $\nu = 0$ yields a regret bound of $O(d\Delta^{-1}\log(|\mathcal{X}|\log(\Delta^{-1})/\delta))$ which implies $R_T \le c\frac{d}{\Delta}\log(|\mathcal{X}|T)$. Minimizing over $\nu > 0$ yields a regret bound of $O(\sqrt{dT\log(\log(T/d)|\mathcal{X}|/\delta)})$ which implies $R_T \le c\sqrt{dT\log(|\mathcal{X}|T)}$.

Remarks:

- Let $\mathcal{X} = \{\mathbf{e}_i : i \in [d]\}$. Then for this action set, this bound is nearly minimax according to our lower bounds!

- However, this is also concerning: we know that in the bandit setting the regret scales like $\sum_{i=2}^{d} \Delta_i^{-1} \log(T)$ but this scales $d\Delta^{-1}\log(T)$, which is significantly worse. Can we achieve this?
- For **pure-exploration**, an analogous analysis shows that one can identify the best-arm in $\frac{d}{\Delta^2}\log(1/\delta)$ pulls. But this is exactly the same rate we would have gotten if we did G-optimal *once* in the beginning and sample according to that!
- **Optimism won't help here**

7 Linear Bandits: Pure exploration

This section is inspired by [Fiez et al., 2019].

Showing that $x^\star$ is the best arm is equivalent to showing that $\langle x^\star - x, \theta^* \rangle > 0$ for all $x \in \mathcal{X} \setminus x^\star$. Given a finite number of observations, we have an estimate $\widehat{\theta}$ and a confidence set for θ^*.

$$\begin{aligned}\langle x^\star - x, \widehat{\theta} \rangle &= \langle x^\star - x, \widehat{\theta} - \theta^* \rangle + \langle x^\star - x, \theta^* \rangle \\ &= \langle x^\star - x, \widehat{\theta} - \theta^* \rangle + \Delta_x\end{aligned}$$

Recalling above, we have for any vector $z \in \mathbb{R}^d$ that $|\langle z, \widehat{\theta} - \theta^* \rangle| \le \|z\|_{(X^\top X)^{-1}} \sqrt{2\log(1/\delta)}$ w.p. $\ge 1-\delta$.

We need to show that this confidence set is completely inside the $x^\star$ region. Where we need to decrease uncertainty is in the directions $x - x^\star$, clearly, which is not the G-optimal design. The most realistic optimization program

$$\begin{aligned}\rho^\star := &\inf_{\lambda \in \triangle_{\mathcal{X}}, \tau \in \mathbb{N}} \tau \\ &\text{subject to} \quad \max_{x \in \mathcal{X}} \frac{\|x^\star - x\|^2_{(\sum_{x \in \mathcal{X}} \tau \lambda_x x x^\top)^{-1}}}{\Delta_x^2} \le \frac{1}{2} \\ &= \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X}} \frac{\|x^\star - x\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}}{\Delta_x^2}\end{aligned}$$

Once can prove a lower bound of $\log(1/2.4\delta)\rho^\star$.

Input: Finite set $\mathcal{X} \subset \mathbb{R}^d$, confidence level $\delta \in (0,1)$.
Let $\widehat{\mathcal{X}}_1 \leftarrow \mathcal{X}, t \leftarrow 1$
while $|\widehat{\mathcal{X}}_\ell| > 1$ **do**
  Let $\widehat{\lambda}_\ell \in \triangle_{\mathcal{X}}$ be a $\frac{d(d+1)}{2}$-sparse minimizer of $f(\lambda; \widehat{\mathcal{X}}_\ell)$ where
    $f(\mathcal{V}) = \inf_{\lambda \in \mathcal{X}} f(\lambda; \mathcal{V}) = \inf_{\lambda \in \mathcal{X}} \max_{x, x' \in \mathcal{V}} \|x - x'\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}$
  Set $\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = 2\epsilon_\ell^{-2} f(\widehat{\mathcal{X}}_\ell) \log(4\ell^2 |\mathcal{X}|/\delta)$
  Pull arm $x \in \mathcal{X}$ exactly $\lceil \tau_\ell \widehat{\lambda}_{\ell,x} \rceil$ times and construct $\widehat{\theta}_\ell$
  $\widehat{\mathcal{X}}_{\ell+1} \leftarrow \widehat{\mathcal{X}}_\ell \setminus \left\{ x \in \widehat{\mathcal{X}}_\ell : \max_{x' \in \widehat{\mathcal{X}}_\ell} \langle x' - x, \widehat{\theta}_t \rangle > \epsilon_\ell \right\}$
  $t \leftarrow t + 1$
Output: $\widehat{\mathcal{X}}_{t+1}$

Lemma 5. *Assume that* $\max_{x \in \mathcal{X}} \langle x^\star - x, \theta^* \rangle \le 2$. *With probability at least* $1-\delta$, *we have* $x^\star \in \widehat{\mathcal{X}}_\ell$ *and* $\max_{x \in \widehat{\mathcal{X}}_\ell} \langle x^\star - x, \theta^* \rangle \le 4\epsilon_\ell$ *for all* $\ell \in \mathbb{N}$.

Proof. For any $\mathcal{V} \subseteq \mathcal{X}$ and $x \in \mathcal{V}$ define

$$\mathcal{E}_{x,\ell}(\mathcal{V}) = \{ |\langle x - x^\star, \widehat{\theta}_\ell - \theta^* \rangle| \le \epsilon_\ell \}$$

where it is implicit that $\widehat{\theta}_\ell$ is the design constructed in the algorithm at stage ℓ with respect to $\widehat{\mathcal{X}}_\ell = \mathcal{V}$. Given $\widehat{\mathcal{X}}_\ell$, with probability at least $1 - \frac{\delta}{2\ell^2|\mathcal{X}|}$

$$\begin{aligned}
|\langle x - x^\star, \widehat{\theta}_\ell - \theta^* \rangle| &\leq \|x - x^\star\|_{(\sum_{x \in \mathcal{V}} \lceil \tau_\ell \lambda_{\ell,x}(\mathcal{V}) \rceil x x^\top)^{-1}} \sqrt{2 \log(4\ell^2|\mathcal{X}|/\delta)} \\
&\leq \frac{\|x - x^\star\|_{(\sum_{x \in \mathcal{V}} \lambda_{\ell,x}(\mathcal{V}) x x^\top)^{-1}}}{\sqrt{\tau_\ell}} \sqrt{2 \log(4\ell^2|\mathcal{X}|/\delta)} \\
&\leq \sqrt{\frac{\|x - x^\star\|^2_{(\sum_{x \in \mathcal{V}} \lambda_{\ell,x}(\mathcal{V}) x x^\top)^{-1}}}{2\epsilon_\ell^{-2} f(\mathcal{V}) \log(4\ell^2|\mathcal{X}|/\delta)}} \sqrt{2 \log(4\ell^2|\mathcal{X}|/\delta)} \\
&= \epsilon_\ell
\end{aligned}$$

By exactly the same sequence of steps as above, we have $\mathbb{P}(\bigcap_{\ell=1}^\infty \bigcap_{x \in \widehat{\mathcal{X}}_\ell} \{|\langle x - x^\star, \widehat{\theta}_t - \theta^* \rangle| > \epsilon_t\}) = \mathbb{P}\left(\bigcap_{x \in \mathcal{X}} \bigcap_{\ell=1}^\infty \mathcal{E}_{x,\ell}(\widehat{\mathcal{X}}_\ell)\right) \geq 1 - \delta$, so assume these events hold. Consequently, for any $x' \in \widehat{\mathcal{X}}_\ell$

$$\begin{aligned}
\langle x' - x^\star, \widehat{\theta}_\ell \rangle &= \langle x' - x^\star, \widehat{\theta}_\ell - \theta^* \rangle + \langle x' - x^\star, \theta^* \rangle \\
&\leq \langle x' - x^\star, \widehat{\theta}_\ell - \theta^* \rangle \\
&\leq \epsilon_\ell
\end{aligned}$$

so that $x^\star$ would survive to round $\ell + 1$. And for any $x \in \widehat{\mathcal{X}}_\ell$ such that $\langle x^\star - x, \theta^* \rangle > 2\epsilon_\ell$ we have

$$\begin{aligned}
\max_{x' \in \widehat{\mathcal{X}}_\ell} \langle x' - x, \widehat{\theta}_\ell \rangle &\geq \langle x^\star - x, \widehat{\theta}_\ell \rangle \\
&= \langle x^\star - x, \widehat{\theta}_\ell - \theta^* \rangle + \langle x^\star - x, \theta^* \rangle \\
&> -\epsilon_\ell + 2\epsilon_\ell \\
&= \epsilon_\ell
\end{aligned}$$

which implies this x would be kicked out. Note that this implies that $\max_{x \in \widehat{\mathcal{X}}_{\ell+1}} \langle x^\star - x, \theta^* \rangle \leq 2\epsilon_\ell = 4\epsilon_{\ell+1}$. □

Theorem 5. *Assume that* $\max_{x \in \mathcal{X}} \langle x^\star - x, \theta^* \rangle \leq 2$. *Then with probability at least* $1 - \delta$, $x^\star$ *is returned from the algorithm at a time* τ *that satisfies*

$$\tau \leq c\rho^\star \log(\Delta^{-1})[\log(1/\delta) + \log(\log(\Delta^{-1})) + \log(|\mathcal{X}|)].$$

Proof. Define $S_\ell = \{x \in \mathcal{X} : \langle x^\star - x, \theta^* \rangle \leq 4\epsilon_\ell\}$. Note that by assumption $\mathcal{X} = \widehat{\mathcal{X}}_1 = S_1$. The above lemma implies that with probability at least $1 - \delta$ we have $\bigcap_{\ell=1}^\infty \{\widehat{\mathcal{X}}_\ell \subseteq S_\ell\}$. This implies that

$$\begin{aligned}
f(\widehat{\mathcal{X}}_\ell) &= \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x, x' \in \widehat{\mathcal{X}}_\ell} \|x - x'\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}} \\
&\leq \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x, x' \in S_\ell} \|x - x'\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}} \\
&= f(S_\ell)
\end{aligned}$$

For $\ell \geq \lceil \log_2(4\Delta^{-1}) \rceil$ we have that $S_\ell = \{x^\star\}$, thus, the sample complexity to identify $x^\star$ is equal to

$$\begin{aligned}
\sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \sum_{x \in \mathcal{X}} \lceil \tau_\ell \widehat{\lambda}_{\ell,x} \rceil &= \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \left(\tfrac{(d+1)d}{2} + \tau_\ell \right) \\
&= \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \left(\tfrac{(d+1)d}{2} + 2\epsilon_\ell^{-2} f(\widehat{\mathcal{X}}_\ell) \log(4\ell^2 |\mathcal{X}|/\delta) \right) \\
&\leq \tfrac{(d+1)d}{2} \lceil \log_2(4\Delta^{-1}) \rceil + \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2\epsilon_\ell^{-2} f(S_\ell) \log(4\ell^2 |\mathcal{X}|/\delta) \\
&\leq \tfrac{(d+1)d}{2} \lceil \log_2(4\Delta^{-1}) \rceil + 4 \log(\tfrac{4 \log_2^2(8\Delta^{-1})|\mathcal{X}|}{\delta}) \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} f(S_\ell).
\end{aligned}$$

We now note that

$$\begin{aligned}
\rho^\star &= \inf_{\lambda \in \triangle_\mathcal{X}} \max_{x \in \mathcal{X}} \frac{\|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}}{(\langle x - x^\star, \theta^* \rangle)^2} \\
&= \inf_{\lambda \in \triangle_\mathcal{X}} \max_{\ell \leq \lceil \log_2(4\Delta^{-1}) \rceil} \max_{x \in S_\ell} \frac{\|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}}{(\langle x - x^\star, \theta^* \rangle)^2} \\
&\geq \frac{1}{\lceil \log_2(4\Delta^{-1}) \rceil} \inf_{\lambda \in \triangle_\mathcal{X}} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \max_{x \in S_\ell} \frac{\|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}}{(\langle x - x^\star, \theta^* \rangle)^2} \\
&\geq \frac{1}{16 \lceil \log_2(4\Delta^{-1}) \rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} \inf_{\lambda \in \triangle_\mathcal{X}} \max_{x \in S_\ell} \|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}} \\
&\geq \frac{1}{64 \lceil \log_2(4\Delta^{-1}) \rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} \inf_{\lambda \in \triangle_\mathcal{X}} \max_{x, x' \in S_\ell} \|x - x'\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}} \\
&\geq \frac{1}{64 \lceil \log_2(4\Delta^{-1}) \rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} f(S_\ell)
\end{aligned}$$

where we have used the fact that $\max_{x,x' \in S_t} \|x - x'\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}} \leq 4 \max_{x \in S_t} \|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}$ by the triangle inequality. □

8 Linear bandits: regret minimization revisited

Okay, now that we know how to do optimal pure exploration, how do we turn this into an algorithm that is optimal?

Let $R_T(\mathcal{X}, \theta) = \mathbb{E}_\theta[\sum_{t=1}^T \Delta_{X_t}]$, $\Delta_x = \max_{x' \in \mathcal{X}} \langle x' - x, \theta \rangle$

The next theorem is from [Lattimore and Szepesvári, 2020].

Theorem 6. *Fix any $\mathcal{X} \subset \mathbb{R}^d$ that spans $\mathbb{R}^d$ and $\theta^* \in \mathbb{R}^d$ such that $\arg\max_{x \in \mathcal{X}} \langle x, \theta^* \rangle$ is unique. Any policy for which $R_T(\mathcal{X}, \theta^*) = o(T^a)$ for any $a > 0$ also satisfies $\liminf_{T \to \infty} \frac{R_T(\mathcal{X}, \theta^*)}{\log(T)} \geq r^\star$ where*

$$\begin{aligned}
r^\star := \inf_{\alpha \in [0,\infty)^\mathcal{X}} & \sum_{x \in \mathcal{X}} \alpha_x \Delta_x \\
\text{subject to} \quad & \max_{x \in \mathcal{X}} \frac{\|x^\star - x\|^2_{(\sum_{x \in \mathcal{X}} \alpha_x x x^\top)^{-1}}}{\Delta_x^2} \leq \frac{1}{2}
\end{aligned}$$

Note that

$$\rho^\star := \inf_{\alpha \in [0,\infty)^{\mathcal{X}}} \frac{1}{2} \sum_{x \in \mathcal{X}} \alpha_x$$
$$\text{subject to} \quad \max_{x \in \mathcal{X}} \frac{\|x^\star - x\|^2_{(\sum_{x \in \mathcal{X}} \alpha_x x x^\top)^{-1}}}{\Delta_x^2} \leq \frac{1}{2}$$

Notes

- There exists an asymptotic algorithm [Lattimore and Szepesvari, 2016], but no satisfying finite-time algorithm as of yet.
- Information directed sampling may be near-optimal and very high performance.

9 Adversarial Linear Bandits

Protocal for Linear Bandits
Input: Time horizon T, action set $\mathcal{A} \subset \mathbb{R}^d$.
Initialize: Adversary chooses $\{z_t\}_{t=1}^T \subset \mathbb{R}^d$.
for: $t = 1, \cdots, T$
Player chooses action $A_t \in \mathcal{A}$
Player suffers (and observes) loss $\ell(A_t, z_t) = A_t^\top z_t$

9.1 Preliminaries

This presentation of mirror descent follows [Bubeck et al., 2012, Ch. 5].

For any open convex set $\mathcal{D} \subset \mathbb{R}^d$ and its closure denoted $\bar{\mathcal{D}}$, for any Legendre F on $\bar{\mathcal{D}}$ define $F^*(x) := \sup_{y \in \bar{\mathcal{D}}} x^\top y - F(y)$.

Define $D_F(x, y) = F(x) - F(y) - (x - y)^\top \nabla F(y)$.

Let the Mirror Descent iterations satisfy, $a_1 = \arg\min_{a \in \mathcal{A}} F(a)$ then

$$\widetilde{a}_{t+1} = \nabla F^*(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t)) \tag{1}$$
$$a_{t+1} = \arg\min_{a \in \mathcal{A}} D_F(a, \widetilde{a}_{t+1}) \tag{2}$$

where we have assumed the iterates exist.

Theorem 7 (Online Mirror Descent)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ be a closed convex action set, ℓ a subdifferentiable loss, and F a Legendre function defined on $\mathcal{A} \subset \bar{\mathcal{D}}$, such that $\nabla F(a) - \eta z \in dom(\nabla F(\mathcal{D}))$ for all $(a, z) \in \mathcal{A} \times \mathcal{Z}$ is satisfied. Then OMD satisfies*

$$\sup_{a \in \mathcal{A}} \sum_{t=1}^T \ell(a_t, z_t) - \ell(a, z_t) \leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^T D_{F^*}\left(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t), \nabla F(a_t)\right).$$

Online Mirror Descent with Linear Losses
Input: Time horizon T, convex action set $\mathcal{A} \subset \mathbb{R}^d$.
Initialize: Player sets $a_1 = \arg\min_{a \in \mathcal{A}} F(a)$. Adversary chooses $\{z_t\}_{t=1}^T \subset [0, 1]^d$.
for: $t = 1, \cdots, T$
Player suffers (and observes) loss $\ell(a_t, z_t) = a_t^\top z_t$
Player observes z_t
Update mirror descent iterates:

$$\widetilde{a}_{t+1} = \nabla F^*(\nabla F(a_t) - \eta z_t)$$
$$a_{t+1} = \arg\min_{a \in \mathcal{A}} D_F(a, \widetilde{a}_{t+1})$$

Corollary 1 (Online Mirror Descent with Linear Losses)**.** *Let $\mathcal{A} \subset \mathcal{D} \subset \mathbb{R}^d$ be a closed convex action set, $\{z_t\}_{t=1}^T \subset \mathcal{Z}$, $\ell(a,z) = a^\top z$, and F a Legendre function defined on $\mathcal{A} \subset \bar{\mathcal{D}}$, such that $\nabla F(a) - \eta z \in \nabla F(\mathcal{D})$ for all $(a,z) \in \mathcal{A} \times \mathcal{Z}$ is satisfied. Then OMD satisfies*

$$\sup_{a\in\mathcal{A}} \sum_{t=1}^T (a_t - a)^\top z_t \leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^T D_{F^*}\left(\nabla F(a_t) - \eta z_t, \nabla F(a_t)\right).$$

Stochastic Online Mirror Descent with Linear Losses
Input: Time horizon T, action set $\mathcal{A} \subset \mathbb{R}^d$.
Initialize: Player sets $a_1 = \arg\min_{a\in\mathcal{A}} F(a)$. Adversary chooses $\{z_t\}_{t=1}^T \subset [0,1]^d$.
for: $t = 1, \cdots, T$
Player chooses distribution P_t over $\mathcal{A}$ with $a_t = \mathbb{E}[A_t|P_t] = \sum_{a\in\mathcal{A}} a P_t(a)$
Player samples A_t from P_t and suffers (and observes) loss $\ell(A_t, z_t) = A_t^\top z_t$
Player computes estimate $\widehat{z}_t$ with $\mathbb{E}[\widehat{z}_t|P_t] = z_t$
Update mirror descent iterates:

$$\widetilde{a}_{t+1} = \nabla F^*(\nabla F(a_t) - \eta \widehat{z}_t)$$
$$a_{t+1} = \arg\min_{a\in\text{convhull}(\mathcal{A})} D_F(a, \widetilde{a}_{t+1})$$

Corollary 2 (Stochastic Online Mirror Descent with Linear Losses)**.** *Let $\mathcal{A} \subset \mathcal{D} \subset \mathbb{R}^d$ be a finite action set, $\{\widehat{z}_t\}_{t=1}^T \subset \mathcal{Z}$, $\ell(a,z) = a^\top z$, and F a Legendre function defined on $\mathcal{A} \subset \bar{\mathcal{D}}$, such that $\nabla F(a) - \eta z \in \nabla F(\mathcal{D})$ for all $(a,z) \in \mathcal{A} \times \mathcal{Z}$ is satisfied. Then OMD satisfies*

$$\sup_{a\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^T (A_t - a)^\top z_t\right] \leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^T \mathbb{E}\left[D_{F^*}\left(\nabla F(a_t) - \eta\widehat{z}_t, \nabla F(a_t)\right)\right].$$

Proof. Applying Corollary 1 with $\widehat{z}_t$ we have

$$\sup_{a\in\mathcal{A}} \sum_{t=1}^T (a_t - a)^\top \widehat{z}_t \leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^T D_{F^*}\left(\nabla F(a_t) - \eta\widehat{z}_t, \nabla F(a_t)\right).$$

Taking the expectation on both sides yields the result by noting that

$$\mathbb{E}\left[\sum_{t=1}^T z_t^\top (A_t - a)\right] = \mathbb{E}\left[\sum_{t=1}^T z_t^\top (a_t - a)\right] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[z_t^\top (a_t - a)|P_t]\right] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[\widehat{z}_t^\top (a_t - a)|P_t]\right] = \mathbb{E}\left[\sum_{t=1}^T \widehat{z}_t^\top (a_t - a)\right].$$

□

9.2 Simplex with unnormalized negative entropy

Example 1 Let $\mathcal{A} = \{x \in \mathbb{R}^d : x_i \geq 0, \sum_{i=1}^d = 1\}$, $F(x) = \sum_{i=1}^n x_i \log(x_i) - x_i$ with $\mathcal{D} = (0,\infty)^d$. F is Legendre and

$$[\nabla F(x)]_i = \log(x_i)$$
$$D_F(x,y) = \sum_{i=1}^d x_i \log(\tfrac{x_i}{y_i}) - \sum_{i=1}^d (x_i - y_i)$$
$$F^*(x) = \sum_{i=1}^d \exp(x_i)$$
$$[\nabla F^*(x)]_i = \exp(x_i)$$
$$D_{F^*}(x,y) = \sum_{i=1}^d \exp(y_i)(\exp(x_i - y_i) - 1 - (x_i - y_i))$$

9.2.1 Full information game

Assume the setting of Example 1. The following algorithm implements the updates of Mirror Descent above for the particular loss $\ell(a, z) = a^\top z$.

Exponential Weights
Input: Time horizon T.
Initialize: Player sets $a_1 = (1/d, \ldots, 1/d)$. Adversary chooses $\{z_t\}_{t=1}^T \subset [0,1]^d$.
for: $t = 1, \cdots, T$
Player chooses $a_t \in \mathcal{A}$
Player suffers (and observes) loss $\ell(a_t, z_t) = a_t^\top z_t$
Player observes z_t
Update mirror descent iterates:

$$\widetilde{a}_{t+1,i} = \exp(-\eta \sum_{s=1}^{t} z_{s,i}) \qquad a_{t,i} = \widetilde{a}_{t+1,i} / \sum_{j=1}^{d} \widetilde{a}_{t+1,j}.$$

Corollary 3 (Exponential weights)**.** *Under the conditions of Example 1 with let $\ell(a, z) = a^\top z$, the exponential weights algorithm satisfies*

$$\sup_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a_t, z_t) - \ell(a, z_t) \leq \frac{\log(d)}{\eta} + \frac{\eta T}{2} \leq \sqrt{2T \log(d)}$$

Proof. Note $\nabla_a \ell(a, z) = z$. Plug in quantities of the example to obtain for any $a \in \mathcal{A}$

$$\begin{aligned}
\sum_{t=1}^{T} \ell(a_t, z_t) - \ell(a, z_t) &= \sum_{t=1}^{T} z_t^\top (a_t - a) \\
&\leq \frac{\log(d)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} \sum_{i=1}^{d} a_{t,i} (\exp(-\eta z_{t,i}) - 1 + \eta z_{t,i}) \\
&\leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} a_{t,i} z_{t,i}^2 \\
&\leq \frac{\log(d)}{\eta} + \frac{\eta T}{2}
\end{aligned}$$

where the second line uses $F(x) \leq 0$ and $F(a_1) = \log(d)$, the third line uses $\exp(-x) \leq 1 - x + \frac{1}{2}x^2$ for $x \geq 0$, and the last line follows from $z_{t,i} \in [0,1]$ and a_t is a probability distribution. □

9.2.2 Full information, finite action set

Analogous to the categorial weather prediction problem in class, we now consider the case where the player can only play from a distinct set $\{1, \ldots, d\}$ (i.e., predict rain, snow, sunny). As discussed in class, any deterministic algorithm will suffer linear regret, so instead, at time t we choose a probability distribution $a_t \in \mathcal{A}$ (in the setting of Example 1), choose distinct action I_t drawn according to a_t so that $A_t := \mathbf{e}_{I_t}$, and then suffer loss $\ell(A_t, z_t) = A_t^\top z_t = z_{t,I_t}$. Note that $\mathbb{E}[A_t] = \mathbb{E}[\mathbf{e}_{I_t}] = \sum_{i=1}^{d} \mathbf{e}_i a_{t,i} = a_t$ so that $\mathbb{E}[\ell(A_t, z_t)] = \mathbb{E}[A_t^\top z_t] = a_t^\top z_t$. Thus, the expected regret relative to any probability distribution $a \in \mathcal{A}$ over distinct items

in hindsight is

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\ell(A_t,z_t)-\ell(a,z_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}z_t^\top(A_t-a)\right]\\ &= \mathbb{E}\left[\sum_{t=1}^{T}z_t^\top(a_t-a)\right]\\ &= \mathbb{E}\left[\sum_{t=1}^{T}\ell(a_t,z_t)-\ell(a,z_t)\right]\end{aligned}$$

where the expectation is with to the random selection of each I_t from a_t. Alternatively, we can directly apply Corollary 2 with $\widehat{z}_t = z_t$ since we are in the full information setting.

Exponential Weights over finite actions

Input: Time horizon T.
Initialize: Player sets $a_1 = (1/d, \dots, 1/d)$. Adversary chooses $\{z_t\}_{t=1}^T \subset [0,1]^d$.
for: $t = 1, \cdots, T$
Player chooses $a_t \in \mathcal{A}$
Player draws $I_t \sim a_t$, sets $A_t = \mathbf{e}_{I_t}$ and suffers (and observes) loss $\ell(A_t, z_t) = A_t^\top z_t = z_{t,I_t}$
Player observes z_t
Update mirror descent iterates:

$$\widetilde{a}_{t+1,i} = \exp(-\eta\sum_{s=1}^{t} z_{s,i}) \qquad a_{t,i} = \widetilde{a}_{t+1,i}/\sum_{j=1}^{d}\widetilde{a}_{t+1,j}.$$

Corollary 4 (Exponential weights over finite actions)**.** *Under the conditions of Example 1 where the player can only play* $\mathbf{e}_i$ *for* $i \in \{1, \dots, d\}$ *with let* $\ell(a,z) = a^\top z$*, the exponential weights over finite actions algorithm satisfies*

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell(\mathbf{e}_{I_t},z_t)-\ell(a,z_t)\right] \le \frac{\log(d)}{\eta}+\frac{\eta T}{2} \le \sqrt{2T\log(d)}$$

Proof. Immediate from reduction described above and the previous corollary since the iterates are identical in the full information game. Due to the oblivious adversary we have

$$\sup_{a\in\mathcal{A}}\mathbb{E}\left[\sum_{t=1}^{T}\ell(a_t,z_t)-\ell(a,z_t)\right] = \mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell(a_t,z_t)-\ell(a,z_t)\right]$$

Note, as we did in class, one can also prove a high probability bound that would apply to a general reactive adversary [**?**, Ch. 2.7] □

9.2.3 Bandit feedback over finite action sets

This setting is identical to the previous setting, but now we do not observe the entire vector z_t at each time t, we only observe the element we played z_{t,I_t}. Using this single value, the player constructs an unbiased estimate of z_t with $\widehat{z}_{t,i} = \frac{\mathbf{1}\{I_t=i\}z_{t,i}}{a_{t,i}}$ for all i. Note that

$$\begin{aligned}\mathbb{E}[\widehat{z}_{t,i}|a_1,I_1,\dots,a_{t-1},I_{t-1},a_t] &= \mathbb{E}[\frac{\mathbf{1}\{I_t=i\}z_{t,i}}{a_{t,i}}|a_1,I_1,\dots,a_{t-1},I_{t-1},a_t]\\ &= \sum_{j=1}^{d}a_{t,j}\frac{\mathbf{1}\{j=i\}z_{t,i}}{a_{t,i}}\\ &= z_{t,i}.\end{aligned}$$

Also note that $\mathbb{E}[A_t] = \mathbb{E}[\mathbf{e}_{I_t}] = \sum_{i=1}^d \mathbf{e}_i a_{t,i} = a_t$.

EXP3: Exponential Weights for Exploration Exploitation

Input: Time horizon T, $\mathcal{A} = \{x \in \mathbb{R}^d : x_i \geq 0, \sum_{i=1}^d = 1\}$.
Initialize: Player sets $a_1 = (1/d, \ldots, 1/d)$. Adversary chooses $\{z_t\}_{t=1}^T \subset [0,1]^d$.
for: $t = 1, \cdots, T$
Player draws $I_t \sim a_t$ and suffers (and observes) loss $\ell(\mathbf{e}_{I_t}, z_t) = z_{t,I_t}$
Player sets $\widehat{z}_{t,i} = \frac{\mathbf{1}\{I_t = i\} z_{t,i}}{a_{t,i}}$
Update mirror descent iterates:

$$\widetilde{a}_{t+1,i} = \exp(-\eta \sum_{s=1}^t \widehat{z}_{s,i}) \qquad a_{t,i} = \widetilde{a}_{t+1,i} / \sum_{j=1}^d \widetilde{a}_{t+1,j}.$$

Corollary 5 (EXP3)**.** *Under the conditions of Example 1 where the player can only play $\mathbf{e}_i$ for $i \in \{1, \ldots, d\}$ with $\ell(a, z) = a^\top z$ and only observe bandit feedback, the EXP3 algorithm satisfies*

$$\begin{aligned}\sup_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^T \ell(A_t, z_t) - \ell(a, z_t)\right] &\leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}\left[\sum_{i=1}^d a_{t,i} \widehat{z}_{t,i}^2\right] \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta T d}{2} \leq \sqrt{2dT \log(d)}\end{aligned}$$

Proof. We can directly apply Corollary 2:

$$\begin{aligned}\mathbb{E}\left[\sup_{a \in \mathcal{A}} \sum_{t=1}^T (A_t - a)^\top z_t\right] &\leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^T \mathbb{E}\left[D_{F^*}\left(\nabla F(a_t) - \eta \widehat{z}_t, \nabla F(a_t)\right)\right] \\ &= \frac{\log(d)}{\eta} + \frac{1}{\eta} \sum_{t=1}^T \mathbb{E}\left[\sum_{i=1}^d a_{t,i}(\exp(-\eta \widehat{z}_{t,i}) - 1 + \eta \widehat{z}_{t,i})\right] \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}\left[\sum_{i=1}^d a_{t,i} \widehat{z}_{t,i}^2\right] \\ &= \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}\left[\sum_{i=1}^d a_{t,i} \frac{\mathbf{1}\{I_t = i\} z_{t,i}^2}{a_{t,i}^2}\right] \\ &= \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{i=1}^d z_{t,i}^2 \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta d T}{2}\end{aligned}$$

□

For an alternative proof of EXP3, see [Lattimore and Szepesvári, 2020, Ch. 11]

9.3 Other action sets

The previous section addressed the case of the action set being equal to the simplex: $\mathcal{A} = \{x \in \mathbb{R}^d : x_i \geq 0, \sum_{i=1}^d = 1\}$. As our Legendre potential we chose unnormalize negative entropy $F(x) = \sum_{i=1}^d x_i \log(x_i) - x_i$. Consider what the guarantee would be from Corollary 2 if we chose a different function, say, $F(x) = \frac{1}{2}\|x\|_2^2$

then:

$$\begin{aligned}\mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{t=1}^{T}(A_t-a)^\top z_t\right] &\le \frac{\sup_{a\in\mathcal{A}}F(a)-F(a_1)}{\eta}+\frac{1}{\eta}\sum_{t=1}^{T}\mathbb{E}\left[D_{F^*}\left(\nabla F(a_t)-\eta\widehat{z}_t,\nabla F(a_t)\right)\right]\\ &\le \frac{1}{\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\widehat{z}_t\|_2^2\right]\\ &= \frac{1}{\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d}\widehat{z}_{t,i}^2\right]\\ &= \frac{1}{\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d}\frac{z_{t,i}^2}{a_{t,i}}\right].\end{aligned}$$

The issue here is that $a_{t,i}$ can become arbitrarily close to 0 and blow up the bound. If we mix in unfirom exploration at each round, one can show that the regret bound is $O((dT)^{2/3})$ which is significantly worse than $O(\sqrt{dT\log(d)})$ of EXP3 above. So given an action set $\mathcal{A}$ how do we choose F? The next proposition sheds some light on this question.

Proposition 3. *If F is twice continuously differntiable, and if its Hessian $\nabla^2F(x)$ is invertible $\forall x\in\mathcal{D}$, then $\forall x,y\in\mathcal{D}$, there exists $\zeta\in\mathcal{D}$ such that $\nabla F(\zeta)\in[\nabla F(x),\nabla F(y)]$ and*

$$D_{F^*}(\nabla F(x),\nabla F(y))=\frac{1}{2}\|\nabla F(x)-\nabla F(y)\|^2_{(\nabla^2F(\zeta))^{-1}}.$$

The implication of the above proposition is that $\exists\nabla F(\zeta_t)\in[\nabla F(a_t)-\eta\widehat{z}_t,\nabla F(a_t)]$ and

$$D_{F^*}\left(\nabla F(a_t)-\eta\widehat{z}_t,\nabla F(a_t)\right)=\frac{\eta^2}{2}\|\widehat{z}_t\|^2_{(\nabla^2F(\zeta_t))^{-1}}$$

For the choice of $F(x)=\frac{1}{2}\|x\|_2^2$ we have $\nabla^2F(\zeta_t)=I$ for any ζ_t so that the Hessian is flat across the action set. On the other hand, with the choice $F(x)=\sum_{i=1}^{d}x_i\log(x_i)-x_i$, we have $\nabla^2F(x)=\mathrm{diag}(1/x_1,\ldots,1/x_d)$ which blows up as a component of x approaches 0. But this is perfect, since then

$$\begin{aligned}\mathbb{E}\left[D_{F^*}\left(\nabla F(a_t)-\eta\widehat{z}_t,\nabla F(a_t)\right)\right] &= \mathbb{E}\left[\frac{\eta^2}{2}\|\widehat{z}_t\|^2_{(\nabla^2F(\zeta_t))^{-1}}\right]\\ &= \mathbb{E}\left[\frac{\eta^2}{2}\sum_{i=1}^{d}\widehat{z}_{t,i}^2\zeta_{t,i}\right]\\ &= \mathbb{E}\left[\frac{\eta^2}{2}\sum_{i=1}^{d}\frac{z_{t,i}^2}{a_{t,i}}\zeta_{t,i}\right]\\ &\approx \frac{\eta^2}{2}\sum_{i=1}^{d}z_{t,i}^2\end{aligned}$$

where we have used the approximation that $\zeta_t\approx a_t$. The hessian of F is blowing up precisely at the locations where $\widehat{z}_t$ blows up, essentially cancelling each other! We'll see another example of this in the next subsection.

9.3.1 Unit ball action set

Here we address the action set $\mathcal{A}=\{x\in\mathbb{R}^d:\|x\|_2\le 1\}$. To use Corollary 2, we need to define P_t (or equivalently, A_t) to make sure that $\mathbb{E}[A_t]=a_t$ and we need to define $\widehat{z}_t$ with $\mathbb{E}[\widehat{z}_t]=z_t$. Consider the following choices:

- $\xi_t\sim\text{Bernoulli}(\|a_t\|_2)$, $I_t\sim\text{uniform}([d])$, $\epsilon_t\in\{-1,1\}$ with equal probability
- $A_t=(1-\xi_t)\epsilon_t\mathbf{e}_{I_t}+\xi_t\frac{a_t}{\|a_t\|_2}$

- $\widehat{z}_t = (1-\xi_t)\frac{d}{1-\|a_t\|_2} A_t A_t^\top z_t$

It is straightforward to verify that $\mathbb{E}[A_t|a_t] = a_t$ and $\mathbb{E}[\widehat{z}_t|a_t] = z_t$. Following the intuition of the previous section, we need to choose F to make $\mathbb{E}\left[\frac{\eta^2}{2}\|\widehat{z}_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}\right]$ small. Let $F(x) = -\log(1-\|x\|_2) - \|x\|_2$. Note that

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{z}_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}\right] &= \mathbb{E}\left[\|(1-\xi_t)\frac{d}{1-\|a_t\|_2} A_t A_t^\top z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}\right] \\
&= \mathbb{E}\left[\frac{d^2}{1-\|a_t\|_2}\|A_t A_t^\top z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}} | \xi_t = 0\right] \\
&= \sum_{i=1}^d \frac{1}{d}\frac{d^2}{1-\|a_t\|_2}\|e_i e_i^\top z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}} \\
&= \frac{d}{1-\|a_t\|_2}\|z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}} \\
&\le \frac{d}{1-\|a_t\|_2}(1-\|\zeta_t\|_2)\|z_t\|_2^2 \\
&\le 2d
\end{aligned}$$

where we have used the fact that $\nabla^2 F(x) \succeq I/(1-\|x\|_2)$ and $\frac{1-\|z_t\|_2}{1-\|a_t\|_2} \le 2$ (see [Lattimore and Szepesvári, 2020] for second fact).

Using a shrunken action set, one can prove that the regret is bounded by $O(\sqrt{dT})$ (see [Lattimore and Szepesvári, 2020]).

9.3.2 Finite action sets

Here we study the case when $|\mathcal{A}| = n$ and $\mathcal{A} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and $\max_{x\in\mathcal{A}} \max_t |x^\top z_t| \le 1$. Our approach will rely on a slight modification of the EXP3 algorithm:

EXP3(γ): Exponential Weights for Exploration Exploitation

Input: Time horizon T, n arms, $\eta > 0$, $\gamma \in [0,1]$, $\lambda \in \triangle_n$.
Initialize: Player sets $p_1 = (1/n, \ldots, 1/n) \in \triangle_n$. Adversary chooses $\{y_t\}_{t=1}^T \subset [-1,1]^n$.
for: $t = 1, \cdots, T$
Player draws $I_t \sim q_t := (1-\gamma)p_t + \gamma\lambda$ and suffers (and observes) loss $\ell(I_t, y_t) = y_{t,I_t}$
Player computes $\widehat{y}_{t,i}$ where $\mathbb{E}[\widehat{y}_{t,i}|p_t] = y_{t,i}$
Update iterates:

$$\widetilde{p}_{t+1,i} = \exp(-\eta\sum_{s=1}^t \widehat{y}_{s,i}) \qquad p_{t,i} = \widetilde{p}_{t+1,i}/\sum_{j=1}^n \widetilde{p}_{t+1,j}.$$

A straightforward modification for the stochastic mirror descent proof accounts for the forced exploration:

Proposition 4 (EXP3(γ))**.** *The regret of EXP3(γ) algorithm satisfies*

$$\max_{i\in[n]} \mathbb{E}\left[\sum_{t=1}^T y_{t,I_t} - y_{t,i}\right] \le \frac{\log(n)}{\eta} + 2\gamma T + \frac{1}{\eta}\sum_{t=1}^T \mathbb{E}\left[\sum_{i=1}^n q_{t,i}\phi(-\eta\widehat{y}_{t,i})\right]$$

where $\phi(x) = e^x - 1 - x \le x^2$ for $|x| \le 1$.

We will use the EXP3(γ) algorithm as follows:

- Set $A_t = x_{I_t}$, $Q_t = \sum_{i=1}^n q_{t,i} x_i x_i^\top$, $\widehat{z}_t = Q_t^{-1} A_t A_t^\top z_t$
- $y_{t,i} = x_i^\top z_t$, $\widehat{y}_{t,i} = x_i^\top \widehat{z}_t$

to obtain the following theorem:

Theorem 8. *Let $|\mathcal{A}| = n$ and $\mathcal{A} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ and $\max_{x\in\mathcal{A}} \max_t |x^\top z_t| \le 1$. Using the reduction to EXP3(γ) with $\gamma = \eta d$ we have*

$$\max_{a\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} (A_t - a)^\top z_t\right] \le \frac{\log(n)}{\eta} + 3\eta dT \le \sqrt{3dT\log(n)}$$

First note that

$$\begin{aligned}
\mathbb{E}[\widehat{y}_{t,i}|p_t] &= \mathbb{E}[x_i^\top \widehat{z}_t|p_t] \\
&= \mathbb{E}[x_i^\top Q_t^{-1} A_t A_t^\top z_t|p_t] \\
&= x_i^\top Q_t^{-1} \mathbb{E}[A_t A_t^\top|p_t] z_t \\
&= x_i^\top z_t = y_{t,i}
\end{aligned}$$

We also have

$$\begin{aligned}
\mathbb{E}[\widehat{y}_{t,i}^2] &= \mathbb{E}[(x_i^\top \widehat{z}_t)^2|p_t] \\
&= \mathbb{E}[x_i^\top Q_t^{-1} A_t A_t^\top Q_t^{-1} x_i (A_t^\top z_t)^2|p_t] \\
&\le x_i^\top Q_t^{-1} x_i \cdot h^2
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} q_{t,i}\widehat{y}_{t,i}^2\right] &\le h^2 \sum_{i=1}^{n} q_{t,i} x_i^\top Q_t^{-1} x_i \\
&\le h^2 \sum_{i=1}^{n} \mathrm{Trace}(q_{t,i} x_i x_i^\top Q_t^{-1}) \\
&\le h^2 \mathrm{Trace}(Q_t Q_t^{-1}) = dL^2
\end{aligned}$$

Finally,

$$\begin{aligned}
|\eta\widehat{y}_{t,i}| &= |\eta x_i^\top \widehat{z}_t| \\
&= |\eta x_i^\top Q_t^{-1} A_t A_t^\top z_t| \\
&\le \eta h |x_i^\top Q_t^{-1} A_t| \\
&= \eta h |x_i^\top Q_t^{-1} x_{I_t}| \\
&\le \eta h \|x_i\|_{Q_t^{-1}} \|x_{I_t}\|_{Q_t^{-1}} \\
&\le \eta h \|x_i\|_{(\gamma \sum_{i=1}^n \lambda_i x_i x_i^\top)^{-1}} \|x_{I_t}\|_{(\gamma \sum_{i=1}^n \lambda_i x_i x_i^\top)^{-1}} \\
&\le \frac{\eta h d}{\gamma}
\end{aligned}$$

so it suffices to take $\gamma = \eta h d$.

10 Contextual Bandits

This section is inspired by [Lattimore and Szepesvári, 2020]

For $t = 1, 2, \dots$

- Nature reveals $c_t \overset{iid}{\sim} \mathcal{D}$
- Player chooses $x_t \in \mathcal{X}$ and observes $y_t = v(c_t, x_t) + \epsilon_t$

which models users showing up to websites, or patients showing up to the doctor with different symptoms. A policy π maps contexts to actions in $[n]$ and the value of a policy π is defiend as

$$V(\pi) = \mathbb{E}_{C,\epsilon}[v(C, \pi(C)) + \epsilon] = \mathbb{E}_C[v(C, \pi(C))]$$

At each time, we assume the action taken is according to some policy $\pi_t \in \Pi$ so that the regret is defined as

$$R_T = T \cdot V(\pi^\star) - \mathbb{E}[\sum_{t=1}^{T} V(\pi_t)]$$

Notes

- This is not the only way we could have defined regret. Perhaps an even more natural approach would be to define the optimal policy wrt the actual rewards, now the expected rewards.

One naive way to handle this is to run a different linear bandit for each c_t, but this is impractical for infinite sets of c_t.

10.1 Stochastic Linear model

Suppose $v(c, x) = \langle \phi(c, x), \theta_* \rangle$. This can be learned through historical data (e.g., a deep network), or through polynomials of inputs. We can restate the above models as For $t = 1, 2, \ldots$

- Nature reveals $(x_{t,1}, \ldots, x_{t,n}) = \mathcal{X}_t \subset \mathbb{R}^d$
- Player chooses $I_t \in [n]$ and observes $y_t = \langle x_{t,I_t}, \theta^* \rangle + \epsilon_t$

When we had a fixed action set, we built confidence intervals on $\langle x_i, \widehat{\theta} - \theta^* \rangle$. Now that we don't know what action sets to expect, a natural to assume $\max_{i,t} \|x_{i,t}\| \leq 1$ and build confidence intervals on $\sup_{u:\|u\|_2 \leq 1} \langle u, \widehat{\theta} - \theta^* \rangle = \|\widehat{\theta} - \theta^*\|_2$, or equivalently, define a set C_t with the guarantee that $\theta^* \in C_t$ for all t. When an action set $\mathcal{X}_t$ shows up, we could eliminate all provably sub-optimal arms by setting $\widehat{\mathcal{X}}_t = \mathcal{X} \setminus \{x : \max_{x' \in \mathcal{X}} \langle x' - x, \theta \rangle < 0 \ \forall \theta \in C_t\}$. An alternative is to run UCB, defining:

$$UCB_t(x) = \max_{\theta \in C_t} \langle x, \theta \rangle$$

and play $x_t = \arg\max_{x \in \mathcal{X}_t} UCB_t(x)$. If $x_t^\star = \arg\max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle$ then

$$\langle x_t^\star, \theta^* \rangle \leq UCB_t(x_t^\star) \leq UCB_t(x_t) = \langle x_t, \widetilde{\theta} \rangle.$$

Thus, the instantaneous regret at time t satisfies

$$\begin{aligned} r_t &= \langle x_t^\star - x_t, \theta^* \rangle \\ &\leq \langle x_t, \widetilde{\theta} - \theta^* \rangle \\ &\leq \|x_t\|_{A_{t-1}^{-1}} \|\widetilde{\theta} - \theta^*\|_{A_{t-1}} \\ &\leq 2\|x_t\|_{A_{t-1}^{-1}} \sqrt{\beta_{t-1}} \end{aligned}$$

Thus, the random regret satisfies

$$\widehat{R}_T = \sum_{t=1}^{T} r_t \leq \sqrt{T \sum_{t=1}^{T} r_t^2} \approx \sqrt{2T\beta_T \sum_{t=1}^{T} \|x_t\|_{A_{t-1}^{-1}}^2}$$

Let $\hat{\theta}_t$ be the ℓ^2-regularized least-squares estimate of θ_* with regularization parameter $\lambda > 0$ given by

$$\hat{\theta}_t = \arg\min_\theta \|\mathbf{X}_{1:t}\theta - \mathbf{Y}_{1:t}\| + \lambda\|\theta\|_2^2 = (\mathbf{X}_{1:t}^T \mathbf{X}_{1:t} + \lambda I)^{-1} \mathbf{X}_{1:t}^T \mathbf{Y}_{1:t}$$

where we are denoting $\mathbf{X}_{1:t}$ as a matrix with rows $X_1^T, X_2^T, \ldots, X_t^T$ and $\mathbf{Y}_{1:t}$ as the vector $(Y_1, \ldots, Y_t)^T$. The following theorem says that with high probability θ_* lies with high probability in an ellipsoid with center at $\hat{\theta}_t$.

Theorem 9. *Confidence Ellipsoid. Assume the same as in Theorem* **??**, *let* $V = I\lambda, \lambda > 0$, *define* $Y_t = \langle X_t, \theta_t \rangle + \eta_t$ *and assume that* $\|\theta_*\| \leq S$. *Then for any* $\delta > 0$, *with probability at least* $1-\delta$, *for all* $t \geq 0, \theta_*$ *lies in the set*

$$C_t = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta} - \theta\|_{\overline{V}_t} \leq R\sqrt{2\log(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta})} + \lambda^{1/2} S \right\}.$$

Furthermore, if for all $t \geq 1, \|X_t\|_2 \leq L$ *then with probability at least* $1-\delta$, *for all* $t \geq 0, \theta_*$ *lies in the set*

$$C'_t = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta} - \theta\|_{\overline{V}_t} \leq R\sqrt{d\log(\frac{1+tL^2/\lambda}{\delta})} + \lambda^{1/2} S \right\}.$$

10.2 Stochastic Contextual Bandits for General policy classes

This section is inspired by Alekh's Agarwal's excellent notes on contextual bandits. See `https://courses.cs.washington.edu/courses/cse599m/19sp/`.

Let's return to the general setting of trying to minimize $V(\pi^\star) - V(\pi)$. The core diffuclty of contextual bandits is that if I take action i and receive a reward with mean $v(c_t, i)$, I don't observe $v(c_t, j)$ for some $j \neq i$. If I don't see that context ever again, how can I know what I should have done?

10.2.1 τ-greedy

Here we explore a simple strategy that explores for τ steps then exploits or the last $T - \tau$ steps. For any context c fix an exploration distribution $\mu(x|c) \in \triangle_n$ such that $\mu(x|c) > 0$ for all x, c.

Then suppose at time t, I observed some context c_t and played $x_t \sim \mu(\cdot|c)$ to receive reward $r_t = v(c_t, x_t) + \epsilon_t$. Then define the **inverse propensity scoring** estimator as

$$\widehat{v}(c_t, x) = r_t \frac{\mathbf{1}\{x_t = x\}}{\mu(x|c_t)}$$

Note that

$$\begin{aligned}
\mathbb{E}[\widehat{v}(c_t, x)|c_t] &= \mathbb{E}\left[r_t \frac{\mathbf{1}\{x_t = x\}}{\mu(x|c_t)} | c_t \right] \\
&= \sum_{x' \in \mathcal{X}} \mu(x'|c_t) \mathbb{E}\left[r_t \frac{\mathbf{1}\{x_t = x\}}{\mu(x|c_t)} | c_t, x_t = x' \right] \\
&= \sum_{x' \in \mathcal{X}} \mu(x'|c_t) v(c_t, x') \frac{\mathbf{1}\{x' = x\}}{\mu(x|c_t)} \\
&= v(c_t, x)
\end{aligned}$$

with variance

$$\begin{aligned}
\mathbb{E}[(\widehat{v}(c_t, x) - v(c_t, x))^2 | c_t] &\leq \mathbb{E}[(\widehat{v}(c_t, x))^2 | c_t] \\
&\leq \mathbb{E}\left[\frac{\mathbf{1}\{x_t = x\}}{\mu(x|c_t)^2} | c_t \right] \qquad (r_t \leq 1) \\
&\leq \sum_{x' \in \mathcal{X}} \mu(x'|c_t) \frac{\mathbf{1}\{x' = x\}}{\mu(x|c_t)^2} \\
&= \frac{1}{\mu(x|c_t)}
\end{aligned}$$

and trivially, $\widehat{v}(c, x) \leq \widehat{v}_{\max} := \max_{c,x} \frac{1}{\mu(x|c)}$. Finally, if $\widehat{V}_t(\pi) = \frac{1}{s}\sum_{s=1}^t \widehat{v}(c_t, \pi(c_t))$ then $\mathbb{E}[\widehat{V}_t(\pi)] = V(\pi)$ for any $\pi \in \Pi$. By Bernstein's inequality we have with probability at least $1-\delta$

$$|V(\pi) - \widehat{V}_t(\pi)| \leq \sqrt{\frac{2\log(2/\delta)}{t} \mathbb{E}_{c,x\sim\mu}\left[\frac{1}{\mu(\pi(c)|c)} \right]} + \frac{2\widehat{v}_{\max}\log(2/\delta)}{3t}$$

and in particular, if $\mu(x|c) = \frac{1}{n}$ for all x, c then

$$|V(\pi) - \widehat{V}_t(\pi)| \le \sqrt{\frac{2n\log(2/\delta)}{t}} + \frac{2n\log(2/\delta)}{3t} \le \sqrt{\frac{4n\log(2/\delta)}{t}}$$

for some $t \ge 2n\log(2/\delta)$. Define $\epsilon_t := \sqrt{\frac{4n\log(2|\Pi|/\delta)}{t}}$. If $\widehat{\pi} = \arg\max_{\pi\in\Pi} \widehat{V}_t(\pi)$

$$\begin{aligned} V(\widehat{\pi}) &= \underbrace{V(\widehat{\pi}) - \widehat{V}_t(\widehat{\pi})}_{\ge -\epsilon_t} + \underbrace{\widehat{V}_t(\widehat{\pi}) - \widehat{V}_t(\pi^\star)}_{\ge 0} + \underbrace{\widehat{V}_t(\pi^\star) - V(\pi^\star)}_{\ge -\epsilon_t} + V(\pi^\star) \\ &\ge V(\pi^\star) - 2\epsilon_t \end{aligned}$$

If we explore uniformly for τ rounds than exploit or $T - \tau$ rounds then we achieve a regret of at most

$$1 \cdot t + 2\epsilon_\tau (T - \tau) \le \tau + T\sqrt{\frac{4n\log(2|\Pi|/\delta)}{\tau}}$$

which is minimized at $\tau = (nT^2\log(2|\Pi|/\delta))^{1/3}$ which yields a regret of $O(T^{2/3}(n\log(2|\Pi|/\delta))^{1/3})$ regret.

Lemma 6 (Bernstein's inequality)**.** *Let $X_1, \dots, X_m$ be independent random variables such that $\frac{1}{m}\sum_{i=1}^m \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \le \sigma^2$ and $|X_i| \le B$. Then*

$$\left|\frac{1}{m}\sum_{i=1}^m X_i\right| \le \sqrt{\frac{2\sigma^2\log(2/\delta)}{m}} + \frac{2B\log(2/\delta)}{3m}$$

with probability at least $1 - \delta$.

10.2.2 Action Elimination

We will make the strong assumption that the distribution of contexts is known a priori. Not so implausible due to historical data.

Recall the value of a policy $\pi \in \Pi$ as $V(\pi) = \mathbb{E}_c[v(c, \pi(c))]$. Taking our G-optimality approach, we wish to sequentially define probability vectors at each time and play probabilistically following

$$\min_{\lambda\in\Delta_{\widehat{\Pi}}} \max_{\pi\in\widehat{\Pi}} \mathbb{E}[(\widehat{V}_t(\pi) - V(\pi))^2]$$

for some active set $\widehat{\Pi} \subset \Pi$.

Input: Policy set Π such that $\pi : \mathcal{X} \to [n]$ for all $\pi \in \Pi$, confidence level $\delta \in (0, 1)$.
Let $\widehat{\Pi}_1 \leftarrow \Pi, \ell \leftarrow 1$
while $|\widehat{\Pi}_\ell| > 1$ **do**
  $\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = \lceil 16n\epsilon_\ell^{-2}\log(2|\Pi|T/\delta)\rceil$, $\gamma_\ell = \min\{\frac{1}{2n}, \sqrt{\frac{\log(2|\Pi|T/\delta)}{9n\tau_\ell}}\}$
  $Q_\ell = \arg\min_{Q\in\triangle_{\widehat{\Pi}_\ell}} \max_{\pi\in\widehat{\Pi}_\ell} \mathbb{E}_C\left[\frac{1}{Q^\gamma(\pi(C)|C)}\right]$
    s.t. $Q^\gamma(x|c) = \gamma + (1 - \gamma n)\sum_{\pi\in\widehat{\Pi}_\ell : \pi(c) = x} Q(\pi)$
  for $t = T_{\ell-1} + 1, \dots, T_\ell$
    Observe context c_t
    Play $x_t \sim Q^\gamma(\cdot|c_t)$, set $p_t = Q^\gamma(x_t|c_t)$ and observe reward $r_t = v(c_t, x_t) + \eta_t$
  Set $\widehat{V}_\ell(\pi) = \frac{1}{T_\ell - T_{\ell-1}}\sum_{t\in(T_{\ell-1}, T_\ell]} r_t \frac{\mathbf{1}\{\pi(c_t) = c_t\}}{p_t}$
  $\widehat{\Pi}_{\ell+1} \leftarrow \widehat{\Pi}_\ell \setminus \left\{\pi \in \widehat{\Pi}_\ell \,|\, \max_{\pi'\in\widehat{\Pi}_\ell} \widehat{V}_\ell(\pi') - \widehat{V}_\ell(\pi) \ge 2\epsilon_\ell\right\}$
  $t \leftarrow t + 1$
Output: Π_{t+1}

We state without proof:

Lemma 7. *For any finite policy set Π and $\gamma \le \frac{1}{2n}$ we have*

$$\min_{Q\in\triangle_\Pi} \max_{\pi\in\Pi} \mathbb{E}_C\left[\frac{1}{Q^\gamma(\pi(C)|C)}\right] \le 2n$$

Lemma 8. *For all $\ell = 1, 2, \ldots$ we have $\pi^\star \in \widehat{\Pi}_\ell$ and $\max_{\pi \in \widehat{\Pi}_\ell} V(\pi) \geq V(\pi^\star) - 8\epsilon_\ell$.*

Proof. Let $\tau_\ell = T_\ell - T_{\ell-1}$. Noting that the variance of $\widehat{V}_\ell(\pi)$ is bounded by $\max_{\pi \in \widehat{\Pi}_\ell} \mathbb{E}_C \left[\frac{1}{Q_\ell^\gamma(\pi(C)|C)}\right]$, we apply Bernstein's inequality at each stage ℓ to find

$$\begin{aligned}|V(\pi) - \widehat{V}_\ell(\pi)| &\leq \sqrt{\frac{4n\log(2|\Pi|T/\delta)}{\tau_\ell}} + \frac{2\log(2|\Pi|T/\delta)}{3\gamma_\ell\tau_\ell} \\ &\leq \sqrt{\frac{16n\log(2|\Pi|T/\delta)}{\tau_\ell}}\end{aligned}$$

for the choice of $\gamma_\ell = \min\{\frac{1}{2n}, \sqrt{\frac{\log(2|\Pi|T/\delta)}{9n\tau_\ell}}\}$ to equalize the terms for large τ_ℓ. The last inequality holds if $\tau_\ell \geq n\log(2|\Pi|T/\delta)$. To make the right hand side less than ϵ_ℓ, it suffices to take $\tau_\ell = \lceil 16n\epsilon_\ell^{-2}\log(2|\Pi|T/\delta)\rceil$.

For any fixed $\widehat{\Pi}_\ell$ with $\pi^\star \in \widehat{\Pi}_\ell$, we have that any $\pi \in \widehat{\Pi}_\ell$ satisfies

$$\begin{aligned}\widehat{V}_\ell(\pi) - \widehat{V}_\ell(\pi^\star) &= \widehat{V}_\ell(\pi) - V(\pi) + \underbrace{V(\pi) - V(\pi^\star)}_{\leq 0} + V(\pi^\star) - \widehat{V}_\ell(\pi^\star) \\ &\leq 2\epsilon_\ell.\end{aligned}$$

On the other hand, for any π such that $V(\pi^\star) - V(\pi) > 4\epsilon_\ell$

$$\begin{aligned}\max_{\pi' \in \widehat{\Pi}_\ell} \widehat{V}_\ell(\pi') - \widehat{V}_\ell(\pi) &\geq \widehat{V}_\ell(\pi^\star) - \widehat{V}_\ell(\pi) \\ &= \widehat{V}_\ell(\pi^\star) - V(\pi^\star) + \underbrace{V(\pi^\star) - V(\pi)}_{>4\epsilon_\ell} + V(\pi) - \widehat{V}_\ell(\pi) \\ &> 2\epsilon_\ell\end{aligned}$$

which implies this π will be kicked out. This means that $\max_{\pi \in \widehat{\Pi}_{\ell+1}} V(\pi) \geq V(\pi^\star) - 4\epsilon_\ell \geq V(\pi^\star) - 8\epsilon_{\ell+1}$.

Extending the proof to all ℓ and random $\widehat{\Pi}_\ell$ is identical to above for linear bandits. □

Suppose you run for T timesteps. Let $\Delta = \min_{\pi \neq \pi^\star} V(\pi^\star) - V(\pi)$. Then for any $\nu \geq 0$ the regret is bounded by:

$$\begin{aligned}&T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1})\rceil} (\gamma_\ell n + 8\epsilon_\ell(1 - \gamma_\ell n))\tau_\ell \\ &= T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1})\rceil} (n\sqrt{\tfrac{\log(2|\Pi|T/\delta)}{9n\tau_\ell}} + 8\epsilon_\ell)\tau_\ell \\ &= T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1})\rceil} n\sqrt{\lceil 16n\epsilon_\ell^{-2}\log(2|\Pi|T/\delta)\rceil \log(2|\Pi|T/\delta)/9n} + 8\epsilon_\ell\lceil 16n\epsilon_\ell^{-2}\log(2|\Pi|T/\delta)\rceil \\ &\leq T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1})\rceil} 2n\epsilon_\ell^{-1}\log(4|\Pi|T/\delta) + 128n\epsilon_\ell^{-1}\log(4|\Pi|T/\delta) \\ &\leq T\nu + 8 + 130n\log(2|\Pi|T/\delta) \sum_{t=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1})\rceil} 2^t \\ &\leq T\nu + 8 + 2080n(\Delta \vee \nu)^{-1}\log(2|\Pi|T/\delta).\end{aligned}$$

As before, using the upper bound $(\Delta \vee \nu) \leq \nu$ and optimizing over ν we have that the regret is no great than $O(\sqrt{nT\log(|\Pi|T/\delta)})$.

10.3 EXP4 for oblivious adversaries

The following algorithm and proof are very standard. However, the textbooks [Lattimore and Szepesvári, 2020, Bubeck et al., 2012] have some unfortunate typos and/or notation that I found confusing so I have reproduced EXP4 here.

EXP4: Exponential Weights for Exploration Exploitation

Input: Time horizon T, n arms, m expers, $\eta > 0$, $\gamma \in [0,1]$, $\lambda \in \triangle_n$.

Initialize: Player sets $Q_1 = (1/m, \dots, 1/m) \in [0,1]^{1\times m}$. Adversary chooses $\{\ell_t\}_{t=1}^T \subset [0,1]^n$.

for: $t = 1, \cdots, T$

Nature reveals expert advice $E^{(t)} \in [0,1]^{m\times n}$

Set $p_{t,i} = \mathbb{E}_{M\sim Q_t}[E^{(t)}_{M,i}] = \sum_{j=1}^m Q_{t,j} E^{(t)}_{j,i}$

Player draws $M_t \sim Q_t$ and $I_t \sim E^{(t)}_{M_t} \in \triangle_n$ (equivalent to $I_t \sim p_t \in \triangle_n$)

Player suffers (and observes) loss ℓ_{t,I_t}

Estimate arm losses $\widehat{\ell}_{t,i} = \frac{\mathbf{1}\{I_t = i\}\ell_{t,i}}{p_{t,i}}$

Estimate expert losses $\widehat{y}_{t,j} = \sum_{i=1}^n E^{(t)}_{j,i}\widehat{\ell}_{t,i}$ for all $j = 1, \dots, m$

Update iterates:

$$\widetilde{Q}_{t+1,i} = \exp(-\eta \sum_{s=1}^t \widehat{y}_{s,i}) \qquad Q_{t,i} = \widetilde{Q}_{t+1,i} / \sum_{j=1}^m \widetilde{Q}_{t+1,j}.$$

First we will prove some simple identities:

$$\mathbb{E}[\widehat{y}_{t,j}] = \mathbb{E}[\sum_{i=1}^n E^{(t)}_{j,i}\widehat{\ell}_{t,i}] = \sum_{i=1}^n E^{(t)}_{j,i}\ell_{t,i}$$

and

$$\mathbb{E}[\ell_{t,I_t}] = \mathbb{E}[\sum_{i=1}^n p_{t,i}\ell_{t,i}] = \sum_{i=1}^n \sum_{j=1}^m Q_{t,j}E^{(t)}_{j,i}\ell_{t,i} = \sum_{j=1}^m Q_{t,j}\sum_{i=1}^n E^{(t)}_{j,i}\ell_{t,i} = \mathbb{E}[\sum_{j=1}^m Q_{t,j}\sum_{i=1}^n \widehat{y}_{t,j}]$$

The expert regret is defined as

$$\begin{aligned}
\max_{k=1,\dots,m} \mathbb{E}\left[\sum_{t=1}^T \ell_{t,I_t} - \sum_{i=1}^n E^{(t)}_{k,i}\ell_{t,i}\right] &= \max_{k=1,\dots,m} \mathbb{E}\left[\sum_{t=1}^T\sum_{j=1}^m Q_{t,j}\sum_{i=1}^n E^{(t)}_{j,i}\ell_{t,i} - \sum_{i=1}^n E^{(t)}_{k,i}\ell_{t,i}\right] \\
&= \max_{k\in[m]} \mathbb{E}\left[\sum_{t=1}^T\sum_{j=1}^m Q_{t,j}\widehat{y}_{t,j} - \widehat{y}_{t,k}\right] \\
&= \max_{k\in[m]} \mathbb{E}\left[\sum_{t=1}^T \widehat{y}_{t,M_t} - \widehat{y}_{t,k}\right] \\
&\le \frac{\log(m)}{\eta} + \frac{\eta}{2}\sum_{t=1}^T \mathbb{E}\left[\sum_{j=1}^m Q_{t,j}\widehat{y}^2_{t,j}\right] \\
&\le \frac{\log(m)}{\eta} + \frac{\eta n T}{2}
\end{aligned}$$

where the first two lines follow from plugging in the identities of above, the third from the definition of M_t,

and the first inequality follows from the guarantee of EXP3. The last inequality follows from

$$\begin{aligned}
\mathbb{E}[\widehat{y}_{t,j}^2] &= \mathbb{E}\left[(\sum_{i=1}^{n} E_{j,i}^{(t)} \widehat{\ell}_{t,i})^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} E_{j,i}^{(t)} \frac{\mathbf{1}\{I_t = i\}\ell_{t,i}}{p_{t,i}}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{E_{j,I_t}^{(t)} \ell_{t,I_t}}{p_{t,I_t}}\right)^2\right] \\
&= \sum_{i=1}^{n} p_{t,i} \left(\frac{E_{j,i}^{(t)} \ell_{t,i}}{p_{t,i}}\right)^2 \\
&\leq \sum_{i=1}^{n} \frac{E_{j,i}^{(t)}}{p_{t,i}}
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^{m} Q_{t,j} \widehat{y}_{t,j}^2\right] &= \sum_{j=1}^{m} Q_{t,j} \sum_{i=1}^{n} \frac{E_{j,i}^{(t)}}{p_{t,i}} \\
&= \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} Q_{t,j} E_{j,i}^{(t)}}{p_{t,i}} = n
\end{aligned}$$

References

[Audibert and Bubeck, 2009] Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

[Bubeck et al., 2012] Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

[Cappé et al., 2013] Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., Stoltz, G., et al. (2013). Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541.

[Fiez et al., 2019] Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. (2019). Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, pages 10666–10676.

[Kaufmann et al., 2016] Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42.

[Lattimore, 2018] Lattimore, T. (2018). Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796.

[Lattimore and Szepesvari, 2016] Lattimore, T. and Szepesvari, C. (2016). The end of optimism? an asymptotic analysis of finite-armed linear bandits. *arXiv preprint arXiv:1610.04491*.

[Lattimore and Szepesvari, 2017] Lattimore, T. and Szepesvari, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737.

[Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). Bandit algorithms. *https://tor-lattimore.com/downloads/book/book.pdf*.

[Mannor and Tsitsiklis, 2004] Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648.

[Pukelsheim, 2006] Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

[Soare, 2015] Soare, M. (2015). *Sequential resource allocation in linear stochastic bandits*. PhD thesis, Université Lille 1-Sciences et Technologies.

[Soare et al., 2014] Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pages 828–836.

[Yu et al., 2006] Yu, K., Bi, J., and Tresp, V. (2006). Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088. ACM.