Some Notes on Multi-armed Bandits

Kevin Jamieson, University of Washington

March 18, 2021

These notes were written for myself to refer to while lecturing. They are not a replacement for the course textbook [Lattimore and Szepesvári, 2020]. There may contain errors, there will contain typos. I have posted them by request.

Contents

Part I

Stochastic Bandits

Chapter 1

Multi-armed Bandits

1.1 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 $t = 1, 2, 3, \ldots$

- Algorithm chooses an action $I_t \in \{1, \ldots, n\}$

- Observes a reward $X_{I_t, t} \sim P_{I_t}$ where $P_1, \ldots, P_n$ 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 $\Delta_i = \max_{j=1,\ldots,n} \theta_j^* - \theta_i^*$. 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.

1.1.1 Regret Minimization

After T time steps, define the *regret* as

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

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

The goal is to have $R(T) = o(T)$ to achieve sub-linear regret (e.g., $R(T) \leq \sqrt{T}$).

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.

1.1.2   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 $\epsilon = 0$.

1.1.3   Warm-up: A/B testing

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 $\arg\max_i \widehat{\theta}_i$

Without loss of generality, assume $\theta_1^* > \theta_2^*$. If $\widehat{\theta}_i$ is the empirical mean of arm i after pulling it τ times, it is a random variable that intuitively should be "close" to θ_i^*. Suppose we could guarantee that $\widehat{\theta}_1 > \widehat{\theta}_2$ with probability $1 - \delta$. If this were true then we have an algorithm for identifying the best arm with probability at least $1 - \delta$ using at most 2τ pulls. Moreover, with probability at least $1 - \delta$ the sub-optimal arm is pulled at most τ times incurring a regret of at most $\tau\Delta$ where $\Delta := \theta_1^* - \theta_2^*$. To make this argument rigorous, we need to be able to build a confidence interval on each $\widehat{\theta}_i - \theta_i^*$ with high probability. By the central limit theorem (CLT) we know that $\widehat{\theta}_i - \theta_i^* \sim \mathcal{N}(0, \frac{\mathbb{V}\mathrm{ar}(Z)}{\tau})$ where $\mathbb{V}\mathrm{ar}(Z)$ denotes the variance of each individual observation (assumed the same for each arm). This suggests that $\frac{\widehat{\theta}_i - \theta_i^*}{\sqrt{\mathbb{V}\mathrm{ar}(Z)}} \in [-1.96, 1.96]$ with probability at least .95 using a standard Normal distribution look up. But this is asymptotic, can we get non-asymptotic and mathematically convenient quantities?

1.1.4 Finite-sample confidence intervals

Proposition 1 (Chernoff Bounding technique). *Fix ϵ, δ. If $Z_1, Z_2, \ldots$ are independent mean-zero random variables with $\psi_Z(\lambda) := \log(\mathbb{E}[\exp(\lambda Z_i)])$ then $\mathbb{P}(\frac{1}{\tau} \sum_{t=1}^{\tau} Z_t > \epsilon) \leq \inf_\lambda \exp(-\tau \epsilon \lambda + \tau \psi_Z(\lambda))$.*

Proof.

$$\mathbb{P}(\frac{1}{\tau} \sum_{t=1}^{\tau} Z_t > \epsilon) = \mathbb{P}(\exp\left(\lambda \sum_{t=1}^{\tau} Z_t\right) > \exp(\lambda \tau \epsilon))$$

$$\leq e^{-\lambda \tau \epsilon} \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{\tau} Z_t\right)\right] \qquad \text{(Markov's)}$$

$$= e^{-\lambda \tau \epsilon} \prod_{t=1}^{\tau} \mathbb{E}\left[\exp\left(\lambda Z_t\right)\right] \qquad \text{(Independence)}$$

$$= \exp\left(-\lambda \tau \epsilon + \tau \psi_Z(\lambda)\right)$$

$\square$

Corollary 1. *Let $Z_1, Z_2, \ldots$ be independent mean-zero σ^2-sub-Gaussian random variables so that $\psi_Z(\lambda) := \log(\mathbb{E}[\exp(\lambda Z_t)]) \leq \exp(\lambda^2 \sigma^2 / 2)$, then for $\tau = \lceil 2\sigma^2 \epsilon^{-2} \log(1/\delta) \rceil$ we have $\mathbb{P}(\frac{1}{\tau} \sum_{t=1}^{\tau} Z_t \leq \epsilon) \geq 1 - \delta$.*

Lemma 1 (Hoeffding's Lemma). *Let X be an independent random variable with support in $[a, b]$ almost surely and $\mathbb{E}[X] = 0$. Then $\log(\mathbb{E}[\exp(\lambda X)]) \leq (b-a)^2 \lambda^2 / 8$.*

Proof. This proof is adapted from [Boucheron et al., 2013]. Let P_X denote the distribution of X so that for any function $g : \mathbb{R} \to \mathbb{R}$ we have $\mathbb{E}_X[g(X)] = \int_x g(x) dP(x)$. Define a new random variable Z with distribution P_Z defined as $dP_Z(x) = \frac{1}{\mathbb{E}_X[\exp(\lambda X)]} e^{\lambda x} dP_X(x)$. Note that P_Z is a valid distribution as $dP_Z(x) \geq 0$ for all x and $\int_x dP_Z(x) = \frac{1}{\mathbb{E}_X[\exp(\lambda X)]} \int_x e^{\lambda x} dP_X(x) = \frac{1}{\mathbb{E}_X[\exp(\lambda X)]} \mathbb{E}_X[\exp(\lambda X)] = 1$. The key observation is to notice that

$$\psi_X(\lambda) := \log(\mathbb{E}_X[\exp(\lambda X)])$$

$$\psi_X'(\lambda) = \frac{1}{\mathbb{E}_X[\exp(\lambda X)]} \mathbb{E}_X[X \exp(\lambda X)]$$

$$\psi_X''(\lambda) = \frac{1}{\mathbb{E}_X[\exp(\lambda X)]} \mathbb{E}_X[X^2 \exp(\lambda X)] - \left(\frac{1}{\mathbb{E}_X[\exp(\lambda X)]} \mathbb{E}_X[X \exp(\lambda X)]\right)^2$$

$$= \mathbb{E}_Z[Z^2] - \mathbb{E}_Z[Z]^2$$

$$= \mathbb{V}ar(Z)$$

$$\leq (b-a)^2 / 4$$

where the last line follows from the fact that the support of P_Z is contained in $[a, b]$ so that

$$\mathbb{V}ar(Z) = \mathbb{E}_Z[(Z - \mathbb{E}_Z[Z]^2)^2] \leq \mathbb{E}_Z[(Z - \tfrac{a+b}{2})^2] \leq (b-a)^2 / 4.$$

By Taylor's remainder theorem, for some $\theta \in [0, \lambda]$ we have

$$\psi_X(\lambda) = \psi_X(0) + \psi_X'(0)\lambda + \psi_X''(\theta)\lambda^2 / 2$$

$$= \psi_X''(\theta)\lambda^2 / 2$$

$$\leq (b-a)^2 \lambda^2 / 8$$

which completes the proof.   □

1.1.5 A/B testing solution

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

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

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

$$
\begin{aligned}
\widehat{\theta}_1 &> \theta_1^* - \sqrt{\frac{2\log(4/\delta)}{\tau}} \\
&> \theta_1^* - \Delta/2 \\
&\geq \theta_2^* + \Delta/2 \\
&\geq \widehat{\theta}_2 - \sqrt{\frac{2\log(4/\delta)}{\tau}} + \Delta/2 \\
&> \widehat{\theta}_2
\end{aligned}
$$

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

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

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

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

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.

1.2 Elimination Algorithm for Pure exploration

Input: n arms $\mathcal{X} = \{1, \ldots, n\}$, confidence level $\delta \in (0, 1)$.
Let $\widehat{\mathcal{X}}_1 \leftarrow \mathcal{X}, \ell \leftarrow 1$
while $|\mathcal{X}l| > 1$ **do**
    $\epsilon_\ell = 2^{-\ell}$
    Pull each arm in $\mathcal{X}l$ 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 $i \in \mathcal{X}l$
    $\mathcal{X}lp \leftarrow \mathcal{X}l \setminus \{i \in \mathcal{X}l : \max_{j \in \mathcal{X}l} \widehat{\theta}_{j,\ell} - \widehat{\theta}_{i,\ell} > 2\epsilon_\ell\}$
    $\ell \leftarrow \ell + 1$
Output: $\widehat{\mathcal{X}}_{\ell+1}$ (or play the last arm forever in the regret setting)

Lemma 2. *Assume that* $\max_{i \in \mathcal{X}} \Delta_i \le 4$. *With probability at least* $1 - \delta$, *we have* $1 \in \mathcal{X}l$ *and* $\max_{i \in \mathcal{X}l} \Delta_i \le 8\epsilon_\ell$ *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 $1 \in \mathcal{X}l$ (note $1 \in \widehat{\mathcal{X}}_1$). Then for any $j \in \mathcal{X}l$ we have

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

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

$$\max_{j \in \mathcal{X}l} \widehat{\theta}_{j,\ell} - \widehat{\theta}_{i,\ell} \ge \widehat{\theta}_{1,\ell} - \widehat{\theta}_{i,\ell}$$
$$= (\widehat{\theta}_{1,\ell} - \theta_1) - (\widehat{\theta}_{i,\ell} - \theta_i) + \Delta_i$$
$$> -2\epsilon_\ell + 4\epsilon_\ell = 2\epsilon_\ell$$

which implies this $\max_{j \in \mathcal{X}lp} \theta_j^* \ge \theta_1^* - 4\epsilon_\ell = \theta_1^* - 8\epsilon_{\ell+1}$. $\qquad \square$

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

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

Proof. Assume $\mathcal{E}$ holds, as it does with probability at least $1 - \delta$. If $\Delta = \min_{i \neq 1} \Delta_i$ then $\mathcal{X}l = \{1\}$ for $t \geq \lceil \log_2(8\Delta^{-1}) \rceil$ since all other arms would have been removed. Note that

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

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

$$
\begin{aligned}
\sum_{i=1}^n T_i &\leq T_1 + \sum_{i=1}^n c\Delta_i^{-2} \log(\tfrac{4 \log_2^2(16\Delta_i^{-2})|\mathcal{X}|}{\delta}) \\
&\leq 2 \sum_{i=1}^n c\Delta_i^{-2} \log(\tfrac{4 \log_2^2(16\Delta_i^{-2})|\mathcal{X}|}{\delta})
\end{aligned}
$$

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

1.3   Elimination Algorithm for Regret minimization

We will use the same algorithm and Lemma as above, but now analyze the regret of the algorithm.

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. For any $\nu \geq 0$ the regret is bounded by:

$$\sum_{i=2}^{n} \Delta_i T_i = \sum_{i:\Delta_i \leq \nu} \Delta_i T_i + \sum_{i:\Delta_i > \nu} \Delta_i T_i$$

$$\leq \nu T + \sum_{i:\Delta_i > \nu} \Delta_i T_i$$

$$= \nu T + \sum_{i:\Delta_i > \nu} \sum_{\ell=1}^{\infty} \Delta_i \tau_\ell \mathbf{1}\{i \in \mathcal{X}l\}$$

$$\leq \nu T + \sum_{i:\Delta_i > \nu} \sum_{\ell=1}^{\infty} \Delta_i \tau_\ell \mathbf{1}\{\Delta_i \leq 8\epsilon_\ell\}$$

$$\leq \nu T + \sum_{i=2}^{n} \sum_{\ell=1}^{\infty} \Delta_i \tau_\ell \mathbf{1}\{\Delta_i \vee \nu \leq 8\epsilon_\ell\}$$

$$\leq \nu T + \sum_{i=2}^{n} \sum_{\ell=1}^{\lceil \log_2(8(\Delta_i \vee \nu)^{-1})\rceil} 8\epsilon_\ell \tau_\ell$$

$$= \nu T + \sum_{i=2}^{n} \sum_{\ell=1}^{\lceil \log_2(8(\Delta_i \vee \nu)^{-1})\rceil} 8\epsilon_\ell \lceil 2\epsilon_\ell^{-2} \log(\tfrac{4\ell^2|\mathcal{X}|}{\delta})\rceil$$

$$\leq \nu T + \sum_{i=2}^{n} c\log(\tfrac{4\log_2^2(8(\Delta_i \vee \nu)^{-2})|\mathcal{X}|}{\delta}) \sum_{\ell=1}^{\lceil \log_2(8(\Delta_i \vee \nu)^{-1})\rceil} 2^\ell$$

$$\leq \nu T + \sum_{i=2}^{n} c(\Delta_i \vee \nu)^{-1} \log(\tfrac{\log((\Delta_i \vee \nu)^{-1})|\mathcal{X}|}{\delta})$$

where the second inequality follows from Lemma 2. 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

$$\sum_{i=2}^{n} \Delta_i \mathbb{E}[T_i] = \mathbb{E}\left[\sum_{i=2}^{n} \Delta_i T_i\right]$$

$$\leq \sum_{i=2}^{n} \Delta_i^{-1} \log(n \log(\Delta_i^{-1})/\delta) + T\mathbb{P}(\mathcal{E}^c)$$

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. One can avoid knowing this by using a double trick: guess a value of T, then when you this value double T and restart using this value of T.

1.4   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?

1.4.1   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] You'll show this on your homework.

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

$$\inf_\phi \max\{\mathbb{P}_0(\phi = 1), \mathbb{P}_1(\phi = 0)\} \geq \inf_\phi \frac{1}{2}(\mathbb{P}_0(\phi = 1), \mathbb{P}_1(\phi = 0))$$

$$= \inf_\phi \frac{1}{2}\left(\int_{x\in\mathbb{R}^n} \mathbf{1}\{\phi(x) = 1\}p_0(x)dx + \int_{x\in\mathbb{R}^n} \mathbf{1}\{\phi(x) = 0\}p_1(x)dx\right)$$

$$= \frac{1}{2}\int_{x\in\mathbb{R}^n} \min\{p_0(x), p_1(x)\}dx$$

$$\geq \frac{1}{4}\left(\int_{x\in\mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx\right)^2 \qquad \text{(Cauchy-Schwartz)}$$

$$\geq \frac{1}{4}\exp\left(-\int_{x\in\mathbb{R}^n} \log(\tfrac{p_1(x)}{p_0(x)})p_1(x)dx\right) \qquad \text{(Jensen's)}$$

where

$$\left(\int_{x\in\mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx\right)^2 = \left(\int_{x\in\mathbb{R}^n} \sqrt{\min\{p_0(x), p_1(x)\}\max\{p_0(x), p_1(x)\}}dx\right)^2$$

$$\leq \int_{x\in\mathbb{R}^n} \min\{p_0(x), p_1(x)\}dx \int_{x\in\mathbb{R}^n} \max\{p_0(x), p_1(x)\}dx \qquad \text{(Cauchy-Schwartz)}$$

$$\leq 2\int_{x\in\mathbb{R}^n} \min\{p_0(x), p_1(x)\}dx$$

and (integrating only over support of pq)

$$\left(\int_{x\in\mathbb{R}^n} \sqrt{p_0(x)p_1(x)}dx\right)^2 = \exp\left(2\log(\int_{x\in\mathbb{R}^n} p_0(x)\sqrt{p_1(x)/p_0(x)}dx)\right)$$

$$\geq \exp\left(2\int_{x\in\mathbb{R}^n} p_0(x)\log(\sqrt{p_1(x)/p_0(x)})dx\right)$$

$$= \exp\left(-\int_{x\in\mathbb{R}^n} \log(\tfrac{p_1(x)}{p_0(x)})p_1(x)dx\right)$$

Note that

$$KL(\mathbb{P}_1|\mathbb{P}_0) = \int_x \log\left(\prod_{i=1}^{n} \frac{p_1(x_i)}{p_0(x_i)}\right) \prod_{i=1}^{n} p_1(x_i)dx$$

$$= nKL(p_1|p_0) = n\Delta^2/2$$

---

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

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 $n \geq 2\Delta^{-2} \log(1/4\delta)$.

1.4.2 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.

1.4.3 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.

1.4.4   Gap-dependent regret

Lemma 3. *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.

1.4.5   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 $\sqrt{nT}$.

Chapter 2

Linear bandits

2.1 Problem statement

Now suppose each arm $i = 1, \ldots, n$ 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.

2.2 Review of 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 $1 - \delta$

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

19

Aside: Gaussian to sub-Gaussian

For an arbitrary constant μ,

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

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 $1 - \delta$,

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

2.3   Experimental design and Kiefer-Wolfowitz

Note that if I take measurements $(x_1, \ldots, x_n) \in \mathcal{X}$ and observe their corresponding observations $y_i = \langle x_i, \theta^* \rangle + \eta_i$ where $\eta_i \in \prime, \infty$, then $\mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\top] = \sigma^2(X^TX)^{-1}$ and also, $\hat{\theta} - \theta^* \sim \mathcal{N}(0, \sigma^2(X^\top X)^{-1})$. 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

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

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

- **A-optimality**: minimize $f_A(\lambda) = \text{Tr}(A_\lambda^{-1})$ minimizes $\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2]$

- **E-optimality**: minimize $f_E(\lambda) = \max_{u:\|u\|\leq 1} u^\top A_\lambda^{-1} u$ minimizes $\max_{u:\|u\|\leq 1} \mathbb{E}[(\langle u, \widehat{\theta} - \theta\rangle)^2]$

- **D-optimality**: maximize $g_D(\lambda) = \log(|A_\lambda|)$ maximizes the entropy of distribution. Also, if $\mathcal{E}_\lambda = \{x : x^\top A_\lambda^{-1} x \leq d\}$ 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 4 (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 $1 - \delta$*

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

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

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, 201

2.3.1   Frank-Wolfe for D/G-optimal design

Define $g(\lambda) = \text{logdet}(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)$. Recall that for any $\lambda \in \triangle_{\mathcal{X}}$ we have by Kiefer-Wolfowitz that
$[\nabla g(\lambda)]_{x'} = \|x'\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}} \geq d$.

Input: Finite set $(x_1, \ldots, x_n) \subset \mathbb{R}^d$, $\lambda^1 \in \triangle_n$.
for $k = 1, 2, \ldots$
  $I_k = \arg\max_i \|x_i\|^2_{(\sum_{j=1}^n \lambda_j^k x_j x_j^\top)^{-1}}$
  $\lambda^{k+1} = (1 - \gamma_k)\lambda^k + \gamma_k [\nabla g(\lambda)]_{I_k}$ where

$$\gamma_k = \arg\max_\gamma g\left((1 - \gamma)\lambda^k + \gamma \mathbf{e}_{I_k}\right)$$

If $\max_i \|x_i\|^2_{(\sum_{j=1}^n \lambda_j^{k+1} x_j x_j^\top)^{-1}} \leq 2d$ **Terminate**

The analysis of the algorithm critically leverages the step size selection. Let $A(\lambda) = \sum_{i=1}^n \lambda_i x_i x_i^\top$ so that $g(\lambda) = \text{logdet}(A(\lambda))$. Note that for some step size γ and $i \in [n]$ we have

$$
\begin{aligned}
g\left((1 - \gamma)\lambda + \gamma \mathbf{e}_i\right) &= g\left((1 - \gamma)[\lambda + \tfrac{\gamma}{1-\gamma}\mathbf{e}_i]\right) \\
&= d\log(1 - \gamma) + \text{logdet}(A(\lambda) + \tfrac{\gamma}{1-\gamma} x_i x_i^\top) \\
&= d\log(1 - \gamma) + \text{logdet}(A(\lambda)(I + \tfrac{\gamma}{1-\gamma} A(\lambda)^{-1} x_i x_i^\top)) \\
&= d\log(1 - \gamma) + \text{logdet}(A(\lambda)) + \log(1 + \tfrac{\gamma}{1-\gamma}\|x_i\|^2_{A(\lambda)^{-1}})
\end{aligned}
$$

which means, plugging in the definition of γ_k and I_k,

$$
\begin{aligned}
g(\lambda^{k+1}) - g(\lambda^k) &= \max_\gamma d\log(1 - \gamma) + \log(1 + \tfrac{\gamma}{1-\gamma}\|x_{I_k}\|^2_{A(\lambda^k)^{-1}}) \\
&= \max_\tau -d\log(1 + \tau) + \log(1 + \tau\|x_{I_k}\|^2_{A(\lambda^k)^{-1}}) \\
&\geq \max_\tau -d\tau + \tau\|x_{I_k}\|^2_{A(\lambda)^{-1}} - \tau^2\|x_{I_k}\|^4_{A(\lambda^k)^{-1}}/2 \\
&= \frac{(\|x_{I_k}\|^2_{A(\lambda)^{-1}} - d)^2}{2\|x_{I_k}\|^4_{A(\lambda^k)^{-1}}}
\end{aligned}
$$

where we have taken the reparameterization $\tau = \frac{\gamma}{1-\gamma}$ which implies $\gamma = \tau/(1 + \tau)$, and the fact $\log(1 + x) \geq x - x^2/2$ for $x \geq 0$.

Let K be the final iterate of the algorithm when it terminates. By the definition of the termination condition and I_k, for all $k \leq K$ we have that $\|x_{I_k}\|^2_{A(\lambda^k)^{-1}} > 2d$. Thus,

$$
\begin{aligned}
g(\lambda^{K+1}) &\geq g(\lambda^K) + \frac{(\|x_{I_K}\|^2_{A(\lambda^K)^{-1}} - d)^2}{2\|x_{I_K}\|^4_{A(\lambda^K)^{-1}}} \\
&\geq g(\lambda^1) + \sum_{k=1}^K \frac{(\|x_{I_k}\|^2_{A(\lambda)^{-1}} - d)^2}{2\|x_{I_k}\|^4_{A(\lambda^k)^{-1}}} \\
&= g(\lambda^1) + \sum_{k=1}^K (1 - \frac{d}{\|x_{I_K}\|^2_{A(\lambda)^{-1}}})^2/2 \\
&\geq g(\lambda^1) + K/8
\end{aligned}
$$

which implies $K \leq 8(\max_\lambda g(\lambda) - g(\lambda^1))$.

There exists an algorithm to produce an initiliazation with $|\text{support}(\lambda^1)| = 2d$ and $\max_\lambda g(\lambda) - g(\lambda^1) \leq 5d \log(d)$ [**?**]. This implies that $|\text{support}(\lambda^{(K+1)})| \leq 40d \log(d)$. But our above analysis is actually quite weak, one can improve this result to $O(d \log \log(d))$ with a more careful Taylor series upper bound.

2.4 Elimination algortihm for 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 $|\mathcal{X}l| > 1$ **do**
  Let $\widehat{\lambda}_\ell \in \triangle_{\mathcal{X}l}$ be a $\frac{d(d+1)}{2}$-sparse minimizer of $f(\lambda) = \max\limits_{x \in \mathcal{X}l} \|x\|^2_{(\sum_{x \in \mathcal{X}l} \lambda_x x x^\top)^{-1}}$
  $\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = 2d\epsilon_\ell^{-2} \log(4\ell^2 |\mathcal{X}|/\delta)$
  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
  $\mathcal{X}lp \leftarrow \mathcal{X}l \setminus \{x \in \mathcal{X}l : \max_{x' \in \mathcal{X}l} \langle x' - x, \widehat{\theta}_\ell \rangle > 2\epsilon_\ell\}$
  $\ell \leftarrow \ell + 1$
Output: $\mathcal{X}l$

After T time steps, define the *regret* as

$$R_T = \langle x^\star, \theta^* \rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t, \theta^* \rangle\right]$$

$$= \mathbb{E}\left[\sum_{x \neq x^\star} T_x \Delta_x\right]$$

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

Lemma 5. *Assume that $\max_{x \in \mathcal{X}} \langle x^\star - x, \theta^* \rangle \leq 4$. With probability at least $1 - \delta$, we have $x^\star \in \mathcal{X}l$ and $\max_{x \in \mathcal{X}l} \langle x^\star - x, \theta^* \rangle \leq 8\epsilon_\ell$ 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 $\mathcal{X}l = \mathcal{V}$. Note that this is precisely the analogous events of multi-armed bandits. The

key piece of the analysis is that

$$\mathbb{P}\left(\bigcup_{\ell=1}^{\infty}\bigcup_{x\in\mathcal{X}l}\{\mathcal{E}_{x,\ell}^{c}(\widehat{\mathcal{X}}_{\ell})\}\right) \le \sum_{\ell=1}^{\infty}\mathbb{P}\left(\bigcup_{x\in\mathcal{X}l}\{\mathcal{E}_{x,\ell}^{c}(\mathcal{X}l)\}\right)$$

$$= \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{X}}\mathbb{P}\left(\bigcup_{x\in\mathcal{V}}\{\mathcal{E}_{x,\ell}^{c}(\mathcal{V})\},\widehat{\mathcal{X}}_{\ell}=\mathcal{V}\right)$$

$$= \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{X}}\mathbb{P}\left(\bigcup_{x\in\mathcal{V}}\{\mathcal{E}_{x,\ell}^{c}(\mathcal{V})\}\right)\mathbb{P}(\widehat{\mathcal{X}}_{\ell}=\mathcal{V})$$

$$\le \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{X}}\frac{\delta|\mathcal{V}|}{2\ell^{2}|\mathcal{X}|}\mathbb{P}(\widehat{\mathcal{X}}_{\ell}=\mathcal{V}) \le \delta$$

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 $x^{\star}\in\mathcal{X}l$ (note $x^{\star}\in\widehat{\mathcal{X}}_{1}$). Then for any $x\in\mathcal{X}l$ we have

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

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

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

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

For any $\ell\ge\lceil\log_{2}(8\Delta^{-1})\rceil$ we have that $\mathcal{X}l = \{x^{\star}\}$. Suppose you run for T timesteps. Then for any $\nu\ge 0$ the regret is bounded by:

$$\sum_{x\in\mathcal{X}\backslash x^{\star}}\Delta_{x}T_{x} = \sum_{x\in\mathcal{X}\backslash x^{\star}:\Delta_{x}\le\nu}\Delta_{x}T_{x} + \sum_{x\in\mathcal{X}\backslash x^{\star}:\Delta_{x}>\nu}\Delta_{x}T_{x}$$

$$\le \nu T + \sum_{\ell=1}^{\infty}\sum_{x\in\mathcal{X}\backslash x^{\star}:\Delta_{x}>\nu}\Delta_{x}\lceil\tau_{\ell}\widehat{\lambda}_{\ell}\rceil$$

$$\le T\nu + \sum_{\ell=1}^{\lceil\log_{2}(8(\Delta\vee\nu)^{-1})\rceil}8\epsilon_{\ell}(|\text{support}(\widehat{\lambda}_{\ell})|+\tau_{\ell})$$

$$= T\nu + \sum_{\ell=1}^{\lceil\log_{2}(8(\Delta\vee\nu)^{-1})\rceil}8\epsilon_{\ell}(\tfrac{(d+1)d}{2}+2d\epsilon_{\ell}^{-2}\log(4\ell^{2}|\mathcal{X}|/\delta))$$

$$\le T\nu + 4(d+1)d\lceil\log_{2}(8(\Delta\vee\nu)^{-1})\rceil + \sum_{\ell=1}^{\lceil\log_{2}(8(\Delta\vee\nu)^{-1})\rceil}16d\epsilon_{\ell}^{-1}\log(4\ell^{2}|\mathcal{X}|/\delta)$$

$$\le T\nu + 4(d+1)d\lceil\log_{2}(8(\Delta\vee\nu)^{-1})\rceil + 16d\log(4\log_{2}^{2}(16(\Delta\vee\nu)^{-1})|\mathcal{X}|/\delta)\sum_{\ell=1}^{\lceil\log_{2}(8(\Delta\vee\nu)^{-1})\rceil}2^{\ell}$$

$$\le T\nu + 4(d+1)d\lceil\log_{2}(8(\Delta\vee\nu)^{-1})\rceil + 512d(\Delta\vee\nu)^{-1}\log(4\log_{2}^{2}(16(\Delta\vee\nu)^{-1})|\mathcal{X}|/\delta)$$

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

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**

2.5 Elimination algorithm for 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^*.

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

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

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

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

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

Lemma 6. *Assume that* $\max_{x \in \mathcal{X}} \langle x^\star - x, \theta^* \rangle \leq 2$. *With probability at least* $1 - \delta$, *we have* $x^\star \in \mathcal{X}l$ *and* $\max_{x \in \mathcal{X}l} \langle x^\star - x, \theta^* \rangle \leq 4\epsilon_\ell$ *for all* $\ell \in \mathbb{N}$.

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

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

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

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

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

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

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

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

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

Theorem 5. *Assume that $\max_{x \in \mathcal{X}} \langle x^\star - x, \theta^\star \rangle \leq 2$. 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 $S_\ell = \{x \in \mathcal{X} : \langle x^\star - x, \theta^\star \rangle \leq 4\epsilon_\ell\}$. 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 $\bigcap_{\ell=1}^{\infty} \{\mathcal{X}l \subseteq S_\ell\}$. This implies that

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

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

$$= f(S_\ell)$$

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

$$\sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \sum_{x \in \mathcal{X}} \lceil \tau_\ell \widehat{\lambda}_{\ell,x} \rceil = \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \left(\frac{(d+1)d}{2} + \tau_\ell \right)$$

$$= \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \left(\frac{(d+1)d}{2} + 2\epsilon_\ell^{-2} f(\mathcal{X}l) \log(4\ell^2 |\mathcal{X}|/\delta) \right)$$

$$\leq \frac{(d+1)d}{2} \lceil \log_2(4\Delta^{-1}) \rceil + \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2\epsilon_\ell^{-2} f(S_\ell) \log(4\ell^2 |\mathcal{X}|/\delta)$$

$$\leq \frac{(d+1)d}{2} \lceil \log_2(4\Delta^{-1}) \rceil + 4 \log\left(\frac{4 \log_2^2(8\Delta^{-1}) |\mathcal{X}|}{\delta} \right) \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} f(S_\ell).$$

We now note that

$$\rho^\star = \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X}} \frac{\|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}}{(\langle x - x^\star, \theta^\star \rangle)^2}$$

$$= \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{\ell \leq \lceil \log_2(4\Delta^{-1}) \rceil} \max_{x \in S_\ell} \frac{\|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}}{(\langle x - x^\star, \theta^\star \rangle)^2}$$

$$\geq \frac{1}{\lceil \log_2(4\Delta^{-1}) \rceil} \inf_{\lambda \in \triangle_{\mathcal{X}}} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \max_{x \in S_\ell} \frac{\|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}}{(\langle x - x^\star, \theta^\star \rangle)^2}$$

$$\geq \frac{1}{16 \lceil \log_2(4\Delta^{-1}) \rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in S_\ell} \|x - x^\star\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}$$

$$\geq \frac{1}{64 \lceil \log_2(4\Delta^{-1}) \rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x,x' \in S_\ell} \|x - x'\|^2_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}$$

$$\geq \frac{1}{64 \lceil \log_2(4\Delta^{-1}) \rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 2^{2\ell} f(S_\ell)$$

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. $\square$

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

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*

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

Note that

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

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.

Chapter 3

Sequential statistics and Martingales

3.1 Preliminaries, Optional stopping, Wald's identity

Additional material on this section can be found in [Lattimore and Szepesvári, 2020] and [Howard et al., 2018]. Let $X_1, X_2, \ldots$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathbb{F} = \{\mathcal{F}_t\}_{t=1}^n$ is a filtration of $\mathcal{F}$. We say the sequence $\{X_t\}_{t=1}^n$ is $\mathbb{F}$-adapted if X_t is $\mathcal{F}_t$ measurable for all $1 \leq t \leq n$.

Definition 2. *An $\mathbb{F}$-adapted sequence of random variables is an $\mathbb{F}$-adapted martingale if $\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t$ for all t and $\mathbb{E}[|X_t|] < \infty$. Furthermore, if*

- *X_t is a super-martingale if $\mathbb{E}[X_{t+1}|\mathcal{F}_t] \leq X_t$*

- *X_t is a sub-martingale if $\mathbb{E}[X_{t+1}|\mathcal{F}_t] \geq X_t$*

Definition 3. *Let $\mathbb{F} = \{\mathcal{F}_t\}_{t \in \mathbb{N}}$ be a filtration. A random variable $\tau \in \mathbb{N}$ is a stopping time with respect to $\mathbb{F}$ with values in $\mathbb{N} \cup \{\infty\}$ if $\mathbf{1}\{\tau \leq t\}$ is $\mathcal{F}_t$ measurable for all $t \in \mathbb{N}$.*

Example 1. *Let $Z_1, Z_2, \ldots$ be an $\mathbb{F}$ adapted sequence and define $S_t = \sum_{i=1}^t$. A valid stopping time may be $\tau = \min\{t \in \mathbb{N} : S_t \geq \epsilon\}$ because τ is $\mathcal{F}_t$ measurable: given S_t we can determine whether it is greater than or equal to ϵ or not. An example of a time that is not a stopping time is $\tau' = \{t \in \mathbb{N} : S_t \geq \epsilon\}$ because given only $\mathcal{F}_t$, the information up to time t, we do not know whether $S_{t'}$ will exceed ϵ again at some future time $t' > t$. Thus, $\mathbf{1}\{\tau' = t\}$ is not meaurable with respect to $\mathcal{F}_t$ and thus, is not a stopping time.*

Lemma 7 (Doob's optional stopping). *Let $\mathbb{F} = \{\mathcal{F}_t\}_{t \in \mathbb{N}}$ be a filtration and $\{X_t\}_t$ be an $\mathbb{F}$-adapted martingale and τ be an $\mathbb{F}$-stopping time. If either of the following two events holds*

- *$\exists N \in \mathbb{N}$ such that $\mathbb{P}(\tau \leq N) = 1$, or*

- *$\mathbb{E}[\tau] < \infty$ and $\mathbb{E}[|X_{t+1} - X_t| \,|\mathcal{F}_t] < c$ for all $t < \tau$ for some $c > 0$,*

then X_τ is well-defined and $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$. Furthermore, if

- *X_t is a super-martingale then $\mathbb{E}[X_\tau] \leq \mathbb{E}[X_0]$*

- *X_t is a sub-martingale then $\mathbb{E}[X_\tau] \geq \mathbb{E}[X_0]$*

Lemma 8 (Wald's identity). *Let Z_t be IID random variables with $\mathbb{E}[Z_t] = \mu$. If τ is a stopping time with $\mathbb{E}[\tau] < \infty$ then $\mathbb{E}[\sum_{t=1}^\tau Z_t] = \mu \mathbb{E}[\tau]$.*

Proof. Note that $X_n = \sum_{t=1}^n Z_t - \mu n$ is a martingale. If $\mu \in \{-\infty, \infty\}$ the result is trivial so assume otherwise. Note

$$\mu = \mathbb{E}[Z_t] = \int z dP_Z(z) = \int \max\{0, z\} dP_Z(z) - \int \max\{0, -z\} dP_Z(z)$$

implies at most one of these summands could be infinite in magnitude (since $\infty - \infty$ is not defined). But if μ is finite, neither piece can be infinite in magnitude. Thus, there exists some $c > 0$ such that $\mathbb{E}[|X_{t+1} - X_t| \, |\mathcal{F}_t] = \mathbb{E}[|Z_{t+1}| \, |\mathcal{F}_t] < c$ and we can apply Doob's optional stopping to conclude

$$\mathbb{E}[\sum_{t=1}^{\tau} Z_t - \mu\tau] = \mathbb{E}[X_\tau] = \mathbb{E}[X_0] = 0$$

which implies the result by subtracting $\mu\mathbb{E}[\tau]$ from both sides.   □

3.2   Maximal inequalities

Lemma 9 (Maximal inequality). *Let $\{X_t\}_t$ be an $\mathbb{F}$-adapted sequence of random variables with $X_t \geq 0$ almost surely. Then for any $\epsilon > 0$, if*

- *X_t is a super-martingale then $\mathbb{P}(\max_{t \in \mathbb{N}} X_t \geq \epsilon) \leq \mathbb{E}[X_0]/\epsilon$*

- *X_t is a sub-martingale then $\mathbb{P}(\max_{t \in \{1, \ldots, n\}} X_t \geq \epsilon) \leq \mathbb{E}[X_n]/\epsilon$*

Proof. We first present the proof of the super-martingale case, which is thanks to [Lattimore and Szepesvári, 2020]. Fix $n \in \mathbb{N}$ and let $\tau = \min\{n + 1, \min\{t : X_t \geq \epsilon\}\}$. Clearly, τ is finite (less than $n + 1$) and thus we can apply Doob's optional stopping. Note that $\tau \leq n \iff \exists t \leq n : X_t \geq \epsilon$ Then by optional stopping,

$$\mathbb{E}[X_0] \geq \mathbb{E}[X_\tau] \geq \mathbb{E}[X_\tau \mathbf{1}\{\tau \leq n\}] \geq \epsilon\mathbb{P}(\tau \leq n) = \epsilon\mathbb{P}(\exists t \leq n : X_t \geq \epsilon).$$

Now we note that $\mathbb{P}(\exists t \in \mathbb{N} : X_t \geq \epsilon) = \lim_{n \to \infty} \mathbb{P}(\exists t \leq n : X_t \geq \epsilon) = \frac{\mathbb{E}[X_0]}{\epsilon}$ by...

The following proof of the sub-martingale case is thanks to [Lawler, 2006]. Let $\tau = \min\{t \leq n : X_t \geq \epsilon\}$. Again, τ is finite so we can apply Doob's optional stopping. Using the tower rule of

expectations we have

$$\mathbb{E}[X_n] \geq \mathbb{E}[X_n \mathbf{1}\{\tau \leq n\}]$$

$$= \sum_{j=1}^{n} \mathbb{E}[X_n \mathbf{1}\{\tau = j\}]$$

$$= \sum_{j=1}^{n} \mathbb{E}[\mathbb{E}[X_n \mathbf{1}\{\tau = j\}|\mathcal{F}_j]]$$

$$= \sum_{j=1}^{n} \mathbb{E}[\mathbb{E}[X_n|\mathcal{F}_j]\mathbf{1}\{\tau = j\}]$$

$$= \sum_{j=1}^{n} \mathbb{E}[X_j \mathbf{1}\{\tau = j\}]$$

$$\geq \epsilon \sum_{j=1}^{n} \mathbb{E}[\mathbf{1}\{\tau = j\}]$$

$$= \epsilon \mathbb{P}(\max_{j=1,\ldots,n} X_j \geq \epsilon)$$

$\square$

Example: Maximal inequality Let $Z_1, Z_2, \ldots$ be Bernoulli(1/2) random variables in $\{-1, 1\}$. Verify that $S_t = \sum_{i=1}^{t} Z_i$ is a martingale. Also note that for any $\lambda > 0$ we have by Jensen's inequality that $\mathbb{E}[\exp(\lambda S_t)|\mathcal{F}_{t-1}] = \mathbb{E}[\exp(\lambda Z_t)|\mathcal{F}_{t-1}]\exp(\lambda S_{t-1}) \geq \exp(\lambda \mathbb{E}[Z_t|\mathcal{F}_{t-1}])\exp(\lambda S_{t-1}) = \exp(\lambda S_{t-1})$. Thus, $\exp(\lambda S_t)$ is a *sub-martingale*. Applying the maximal inequality for sub-martingales we have for any $N \in \mathbb{N}$ that

$$\mathbb{P}(\max_{t\in\{1,\ldots,N\}} S_t \geq \sqrt{2N\log(1/\delta)}) = \mathbb{P}(\max_{t\in\{1,\ldots,N\}} \exp(\lambda S_t) \geq \exp(\lambda\sqrt{2N\log(1/\delta)}))$$

$$\leq \exp(-\lambda\sqrt{2N\log(1/\delta)})\mathbb{E}[\exp(\lambda S_N)]$$

$$\leq \exp(-\lambda\sqrt{2N\log(1/\delta)})\exp(\lambda^2 N/2)$$

where the last inequality follows from the fact that S_N is a sum of N IID random variables, so $\mathbb{E}[\exp(\lambda S_N)] \leq \exp(\lambda^2 N/2)$. By setting $\lambda = \sqrt{2\log(1/\delta)/N}$ we obtain $\mathbb{P}(\max_{t\in\{1,\ldots,N\}} S_t \geq \sqrt{2N\log(1/\delta)}) \leq \delta$. Since all we used is that $\mathbb{E}[\exp(\lambda S_N)] \leq \exp(\lambda^2 N/2)$, we could have also applied a standard Chernoff bound at time N to obtain $\mathbb{P}(S_N \geq \sqrt{2N\log(1/\delta)}) \leq \delta$. This above example seems to be getting a guarantee on $t \in \{1, \ldots, N-1\}$ for free! It turns out we can do *even better*.

3.3 Anytime concentration inequalities

3.3.1 Linear boundaries

Let $Z_1, Z_2, \ldots$ be Bernoulli(1/2) random variables in $\{-1, 1\}$. Define the random walk $S_t = \sum_{i=1}^{t} Z_i$. If $M_t(\lambda) = \exp(\lambda S_t - t\lambda^2/2)$ then M_t is a super-martingale since

$$\mathbb{E}[M_{t+1}(\lambda)|\mathcal{F}_t] = \mathbb{E}[\exp(\lambda S_{t+1} - (t+1)\lambda^2/2)|\mathcal{F}_t] = \exp(\lambda S_t - t\lambda^2/2)\mathbb{E}[\exp(\lambda Z_{t+1} - \lambda^2/2)|\mathcal{F}_t] \leq M_t(\lambda) \cdot 1$$

Applying the maximal inequality for super-martingales we have

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t \geq t\lambda/2 + \log(1/\delta)/\lambda) = \mathbb{P}(\exists t \in \mathbb{N} : M_t(\lambda) \geq 1/\delta) \leq \delta$$

since $\mathbb{E}_[M_0(\lambda)] = 1$. The above holds for any λ and says the random walk S_t, with probability at least $1 - \delta$ does not go above the line $t\lambda/2 + \log(1/\delta)/\lambda$ for all $t \in \mathbb{N}$. But if we take $\lambda = \sqrt{2\log(1/\delta)/N}$ then we have that

$$\mathbb{P}(\max_{t\in\{1,\dots,N\}} S_t \geq (t/\sqrt{N} + \sqrt{N})\sqrt{\log(1/\delta)/2}) \leq \delta,$$

a strict improvement over the maximal inequality!

3.3.2   Curved boundaries with a mixing distribution

Let $Z_1, Z_2, \dots$ be Bernoulli(1/2) random variables in $\{-1, 1\}$. If $S_t = \sum_{i=1}^{t} Z_i$ then $M_t(\lambda) = \exp(\lambda S_t - t\lambda^2/2)$ is a super-martingale for any $\lambda \in \mathbb{R}$. Let h be any probability distribution over $\mathbb{R}$. Define $\bar{M}_t = \int_\lambda M_t(\lambda)dh(\lambda)$. Then $\bar{M}_t$ is a super-martingale since

$$\mathbb{E}[\bar{M}_{t+1}|\mathcal{F}_t] = \mathbb{E}\left[\int_\lambda M_{t+1}(\lambda)dh(\lambda)|\mathcal{F}_t\right]$$
$$= \int_\lambda \mathbb{E}\left[M_{t+1}(\lambda)|\mathcal{F}_t\right] dh(\lambda)$$
$$\leq \int_\lambda M_t(\lambda)dh(\lambda)$$
$$= \bar{M}_t.$$

Suppose we take $h(\lambda) = \frac{1}{\sqrt{2\pi\nu^2}}e^{-\lambda^2/2\nu^2}$. Then

$$\bar{M}_t = \int_\lambda M_t(\lambda)dh(\lambda) = \frac{1}{\sqrt{2\pi\nu^2}} \int \exp(\lambda S_t - t\lambda^2/2 - \lambda^2/2\nu^2)d\lambda$$
$$= \frac{1}{\sqrt{2\pi\nu^2}} \int \exp(\lambda S_t - \lambda^2(t + \nu^{-2})/2)d\lambda$$
$$= \frac{1}{\sqrt{2\pi\nu^2}} \int \exp(S_t^2(t + \nu^{-2})^{-1}/2 - (S_t(t + \nu^{-2})^{-1} - \lambda)^2/2(t + \nu^{-2})^{-1})d\lambda$$
$$= \sqrt{\frac{(t + \nu^{-2})^{-1}}{\nu^2}} \exp(S_t^2(t + \nu^{-2})^{-1}/2)$$
$$= \sqrt{\frac{\nu^{-2}}{t + \nu^{-2}}} \exp(S_t^2(t + \nu^{-2})^{-1}/2).$$

Applying the maximal inequality for super-martingales we have

$$\mathbb{P}(\exists t : |S_t| \geq \sqrt{2(t + \nu^{-2})\left(\log(1/\delta) + \tfrac{1}{2}\log(\frac{t + \nu^{-2}}{\nu^{-2}})\right)}) = \mathbb{P}(\exists t : \bar{M}_t \geq 1/\delta) \leq \delta.$$

A particularly convenient choice for $\nu = 1$, which implies

$$\mathbb{P}(\exists t : |S_t| \geq \sqrt{(t + 1)\log(\frac{t+1}{\delta^2})}) = \leq \delta.$$
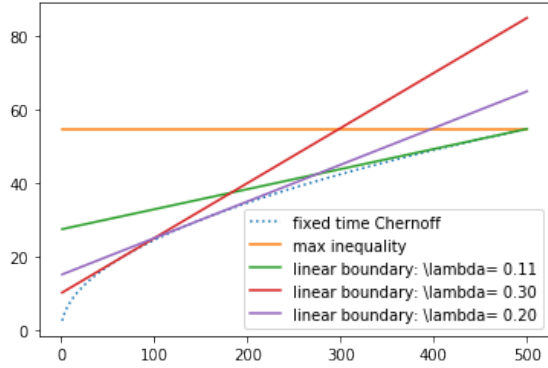
(a) Fix $\delta = 0.05$. The 'fixed time Chernoff' represents $\sqrt{2t \log(1/\delta)}$ which holds at each t but not all $t \le 500$ simultaneously (which is why it is dotted). The 'max inequality' holds for all $t \le 500$, and the linear boundaries hold for all $t \in \mathbb{N}$ simultaneously.
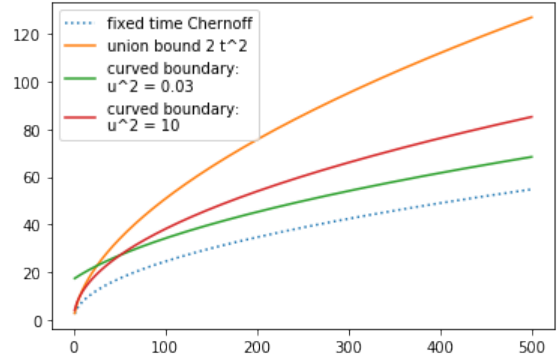
(b) Fix $\delta = 0.05$. The 'fixed time Chernoff' represents $\sqrt{2t \log(1/\delta)}$ which holds at each t but not all $t \in \mathbb{N}$ simultaneously (which is why it is dotted). All other curves do hold for all $t \in \mathbb{N}$ simultaneously. "union bound $2t^2$" plots $\sqrt{2 \log(2t^2/\delta)}$.

Intuitively, $h(\lambda)$ is a probability distribution over linear boundaries parameterized by λ.

The above Figures compares these linear and curved boundaries. We see that the curved boundary just derived appears much tighter than our naive union bound used in the proofs of the early days of this course. Let us consider a few more interesting examples.

3.3.3 Predictable sequences

Let $Z_1, Z_2, \ldots$ be an $\mathcal{F}_t$-adapted sequence and assume σ_t is predictable in the sense that σ_t is $\mathcal{F}_{t-1}$-measurable. Furthermore, assume that for any $\lambda > 0$ we have $\mathbb{E}[\exp(\lambda Z_t)|\mathcal{F}_{t-1}] \le \exp(\lambda^2 \sigma_t^2 / 2)$. Define $S_t = \sum_{i=1}^{t} Z_i$ and $V_t = \sum_{i=1}^{t} \sigma_i^2$. Then $M_t(\lambda) = \exp(\lambda S_t - \lambda^2 V_t / 2)$ is a super-martingale. Thus,

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t \ge \lambda V_t / 2 + \log(1/\delta)/\lambda) = \mathbb{P}(\exists t : M_t(\lambda) \ge 1/\delta) \le \delta.$$

Likewise, computing $\bar{M}_t$ with $h(\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\lambda^2/2}$ yields

$$\mathbb{P}(\exists t \in \mathbb{N} : |S_t| \ge \sqrt{(V_t + 1) \log(\tfrac{V_t+1}{\delta^2})}) = \mathbb{P}(\exists t : \bar{M}_t \ge 1/\delta) \le \delta.$$

Note that "time" t does not appear anywhere in these bounds explicitly, and has been replaced by V_t.

Now suppose for $\lambda \in (0, 1/c)$ we have $\mathbb{E}[\exp(\lambda Z_t)|\mathcal{F}_{t-1}] \le \exp(\frac{\lambda^2 \sigma_t^2}{2(1-c\lambda)})$. Then $M_t(\lambda) = \exp(\lambda S_t - \frac{\lambda^2 V_t}{2(1-c\lambda)})$ is a supermartingale. Thus,

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t \ge \frac{\lambda V_t}{2(1 - c\lambda)} + \log(1/\delta)/\lambda) = \mathbb{P}(\exists t : M_t(\lambda) \ge 1/\delta) \le \delta. \tag{3.1}$$

3.3.4 Vector-valued martingales

Now suppose $Z_1, Z_2, \cdots \in \mathbb{R}^d$ is a $\mathcal{F}_t$-adapted random sequence that satisfies $\mathbb{E}[\exp(\langle \lambda, Z_t \rangle)|\mathcal{F}_{t-1}] \le \exp(\|\lambda\|_{\Sigma_t}^2 / 2)$ for any $\lambda \in \mathbb{R}^d$ for a Σ_t predictable sequence. Define $S_t = \sum_{i=1}^{t} Z_i$ and $V_t = \sum_{i=1}^{t} \Sigma_i$.

Then $M_t(\lambda) = \exp(\langle \lambda, S_t \rangle - \|\lambda\|_{V_t}^2/2)$ is a super-martingale. If $h(\lambda) = \frac{1}{(2\pi/\gamma)^{d/2}} \exp(-\|\lambda\|^2 \gamma/2)$ be a mean-zero Gaussian distribution with covariance $\gamma^{-1}I$. If $\bar{M}_t = \int_\lambda M_t(\lambda) dh(\lambda)$ then

$$
\begin{aligned}
\bar{M}_t &= \int_\lambda M_t(\lambda) dh(\lambda) \\
&= \frac{1}{(2\pi/\gamma)^{d/2}} \int_\lambda \exp(\langle \lambda, S_t \rangle - \|\lambda\|_{V_t}^2/2 - \|\lambda\|^2\gamma/2) dh(\lambda) \\
&= \frac{1}{(2\pi/\gamma)^{d/2}} \int_\lambda \exp(\langle \lambda, S_t \rangle - \|\lambda\|_{V_t+\gamma I}^2/2) dh(\lambda) \\
&= \frac{1}{(2\pi/\gamma)^{d/2}} \int_\lambda \exp(\tfrac{1}{2}\|S_t\|_{(V_t+\gamma I)^{-1}}^2 - \tfrac{1}{2}\|(V_t+\gamma I)^{-1}S_t - \lambda\|_{(V_t+\gamma I)}^2) dh(\lambda) \\
&= \frac{|V_t + \gamma I|^{-1/2}}{\gamma^{-d/2}} \exp(\tfrac{1}{2}\|S_t\|_{(V_t+\gamma I)^{-1}}^2)
\end{aligned}
$$

then repeating the same steps as above we conclude that

$$
\mathbb{P}(\exists t : \|S_t\|_{(V_t+\gamma I)^{-1}} \geq \sqrt{2\log(1/\delta) + \log(\frac{|V_t + \gamma I|}{\gamma^d})}) \leq \delta. \tag{3.2}
$$

3.3.5   Application: Online linear regression

Let $x_1, x_2, \cdots \in \mathbb{R}^d$ be an $\mathcal{F}_{t-1}$-measurable sequence, and for each $t \in \mathbb{N}$ let $y_t \in \mathbb{R}$ be $\mathcal{F}_t$-measurable. We assume there exists $\theta_* \in \mathbb{R}^d$ such that each $y_t = \langle \theta_*, x_t \rangle + \eta_t$ where η_t is mean-zero, independent of x_t, and $\mathbb{E}[\exp(s\eta_t)|\mathcal{F}_{t-1}] \leq \exp(s^2/2)$ for any $s \in \mathbb{R}$. In the previous example let $Z_i = x_i\eta_i$ so that $S_t = \sum_{i=1}^t x_t\eta_t$ and $V_t = \sum_{i=1}^t x_t x_t^\top$ since

$$
\begin{aligned}
\mathbb{E}[\exp(\langle \lambda, x_t\eta_t \rangle)|\mathcal{F}_{t-1}] &= \mathbb{E}[\exp(\langle \lambda, x_t \rangle \eta_t)|\mathcal{F}_{t-1}] \\
&\leq \exp(\langle \lambda, x_t \rangle^2/2) \\
&= \exp(\|\lambda\|_{x_t x_t^\top}^2/2).
\end{aligned}
$$

Thus, Equation 3.2 holds for any $\gamma > 0$. Fix some $\gamma > 0$ and define

$$
\begin{aligned}
\widehat{\theta}_t &= \arg\min_\theta \sum_{i=1}^t (y_i - \langle x_i, \theta \rangle)^2 + \gamma\|\theta\|_2^2 \\
&= (\sum_{i=1}^t x_i x_i^\top + \gamma I)^{-1} \sum_{i=1}^t x_i y_i \\
&= (V_t + \gamma I)^{-1} V_t \theta_* + (V_t + \gamma I)^{-1} S_t
\end{aligned}
$$

Now notice

$$
\begin{aligned}
\|\widehat{\theta}_t - \theta_*\|_{(V_t+\gamma I)} &= \|\widehat{\theta}_t - (V_t+\gamma I)^{-1}(V_t+\gamma I)\theta_*\|_{(V_t+\gamma I)} \\
&= \|(V_t+\gamma I)^{-1}S_t - \gamma(V_t+\gamma I)^{-1}\theta_*\|_{(V_t+\gamma I)} \\
&= \|S_t - \gamma\theta_*\|_{(V_t+\gamma^{-1}I)^{-1}} \\
&\leq \|S_t\|_{(V_t+\gamma I)^{-1}} + \gamma\|\theta_*\|_{(V_t+\gamma I)^{-1}} \\
&\leq \|S_t\|_{(V_t+\gamma I)^{-1}} + \sqrt{\gamma}\|\theta_*\|_2.
\end{aligned}
$$

We conclude that

$$\mathbb{P}\Big(\exists t : \|\widehat{\theta}_t - \theta_*\|_{(V_t+\gamma I)} \geq \sqrt{\gamma}\|\theta_*\|_2^2 + \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t + \gamma I|)}\Big) \tag{3.3}$$

$$\leq \mathbb{P}\Big(\exists t : \|S_t\|_{(V_t+\gamma I)^{-1}} \geq \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t + \gamma I|)}\Big) \leq \delta.$$

If we assume $\max_t \|x_t\|_2^2 \leq L$ then we also have by Jensen's inequality

$$\log\left(|V_t + \gamma I|^{1/d}\right) = \log\left(\prod_{i=1}^{d} \lambda_i\right)^{1/d}$$

$$= \sum_{i=1}^{d} \frac{1}{d}\log(\lambda_i)$$

$$\leq \log\left(\sum_{i=1}^{d} \frac{1}{d}\lambda_i\right)$$

$$= \log\left(\frac{1}{d}\text{Trace}(V_t + \gamma I)\right)$$

$$= \log\left(tL/d + \gamma\right)$$

so that for all $t \in \mathbb{N}$ we have

$$\|\widehat{\theta}_t - \theta_*\|_{(V_t+\gamma I)} \leq \sqrt{\gamma}\|\theta_*\|_2^2 + \sqrt{2\log(1/\delta) + d\log(\tfrac{tL}{d\gamma} + 1)}$$

Due to its usefulness, we summarize the above discussion in a proposition.

Proposition 3. *Fix $\delta \in (0,1)$, $\gamma \geq 0$, and $\theta_* \in \mathbb{R}^d$. Assume for all $t \geq 1$ that $y_t = \langle \theta_*, x_t \rangle + \eta_t$ and $\mathbb{E}[\exp(s\eta_t)|\mathcal{F}_{t-1}] \leq \exp(s^2/2)$ for any $s \in \mathbb{R}$ where $\mathcal{F}_t$ is such that $x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t$ are $\mathcal{F}_{t-1}$ measurable. If $S_t = \sum_{i=1}^{t} x_t\eta_t$, $V_t = \sum_{i=1}^{t} x_t x_t^\top$, and $\widehat{\theta}_t = (V_t + \gamma I)^{-1}S_t$, then*

$$\|\widehat{\theta}_t - \theta_*\|_{(V_t+\gamma I)} \leq \sqrt{\gamma}\|\theta_*\|_2^2 + \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t + \gamma I|)}$$

for all $t \geq 1$ simultaneously with probability at least $1 - \delta$. Moreover, if $\max_t \|x_t\|_2^2 \leq L$ then $\log(\gamma^{-d}|V_t + \gamma I|) \leq d\log(\tfrac{tL}{d\gamma} + 1)$.

3.4   Hypothesis testing and Likelihood ratios

Let $X_1, X_2, \ldots$ be an $\mathcal{F}_t$-adapted sequence of random variables. Consider the hypothesis test

$$\mathbf{H}_0 : X_t \sim p_0 \; \forall t$$
$$\mathbf{H}_1 : X_t \sim p_1 \; \forall t.$$

Define the likelihood ratio $L_t = \prod_{s=1}^{t} \frac{p_1(X_s)}{p_0(X_s)}$. Let $\mathbb{E}_i[\cdot], \mathbb{P}_i[\cdot]$ denote expectation and probability under $\mathbf{H}_i$. Note that under $\mathbf{H}_0$ we have that L_t is a martingale since

$$\mathbb{E}_0[L_{t+1}|\mathcal{F}_t] = L_t \int_x \frac{p_1(x)}{p_0(x)}p_0(x)dx = L_t \int_x p_1(x)dx = L_t.$$

Similarly, we have that L_t^{-1} is a martingale under $\mathbf{H}_1$. This allows us to apply the maximal inequality to conclude that

$$\max\{\mathbb{P}_0(\exists t \in \mathbb{N} : L_t \geq 1/\delta), \mathbb{P}_1(\exists t \in \mathbb{N} : L_t \leq \delta)\} \leq \delta.$$

We will show that if $\tau := \min\{t \in \mathbb{N} : L_t \notin (\delta, 1/\delta)\}$ then

$$\mathbb{E}_0[\tau] \leq \frac{\log(e/\delta)}{KL(p_0|p_1)} + 1 \qquad \text{and} \qquad \mathbb{E}_1[\tau] \leq \frac{\log(e/\delta)}{KL(p_1|p_0)} + 1.$$

Compare this with our minimax lower bound of above. Since they nearly match, we conclude that this method known as the sequential probability ratio test (SPRT) is optimal.

By Wald's inequality we have

$$\mathbb{E}_0[\log(L_\tau)] = \mathbb{E}_0[\tau]\mathbb{E}_0[\log(\tfrac{p_1(X_1)}{p_0(X_1)})] = -\mathbb{E}_0[\tau]KL(p_0|p_1).$$

But on the other hand, we also have

$$\begin{aligned}
\mathbb{E}_0[\log(L_\tau)] &= \mathbb{E}_0[\log(L_\tau)\mathbf{1}\{L_\tau > 1/\delta\}] + \mathbb{E}_0[\log(L_\tau)\mathbf{1}\{L_\tau < \delta\}] \\
&\geq \mathbb{E}_0[\log(L_\tau)\mathbf{1}\{L_\tau < \delta\}] \\
&\geq \mathbb{E}_0[(\log(L_{\tau-1}) + \log(\tfrac{p_1(X_\tau)}{p_0(X_\tau)}))\mathbf{1}\{L_\tau < \delta\}] \\
&\geq \log(\delta) + \mathbb{E}_0[\log(\tfrac{p_1(X_\tau)}{p_0(X_\tau)})\mathbf{1}\{L_\tau < \delta\}] \\
&= \log(\delta) + \mathbb{E}_0[\log(\tfrac{p_1(X_\tau)}{p_0(X_\tau)})\mathbf{1}\{\log(\tfrac{p_1(X_\tau)}{p_0(X_\tau)}) < \log(\delta) - \log(L_{\tau-1})\}] \\
&= \log(\delta) - \mathbb{E}_0[\log(\tfrac{p_0(X_1)}{p_1(X_1)})\mathbf{1}\{p_1(X_1) < p_0(X_1)\}] \\
&\geq \log(\delta) - KL(p_0|p_1) - 1
\end{aligned}$$

where the last line follows from

$$\begin{aligned}
\mathbb{E}_0[\log(\tfrac{p_0(X_1)}{p_1(X_1)})\mathbf{1}\{p_1(X_1) < p_0(X_1)\}] &= \int_x p_0(x)\log(\tfrac{p_0(x)}{p_1(x)})\mathbf{1}\{p_1(x) < p_0(x)\}dx \\
&= \int_{x:p_1(x)<p_0(x)} p_0(x)\log(\tfrac{p_0(x)}{p_1(x)})dx \\
&= \int_x p_0(x)\log(\tfrac{p_0(x)}{p_1(x)})dx - \int_{x:p_1(x)>p_0(x)} p_0(x)\log(\tfrac{p_0(x)}{p_1(x)})dx \\
&= KL(p_0|p_1) + \int_{x:p_1(x)>p_0(x)} p_0(x)\log(\tfrac{p_1(x)}{p_0(x)})dx \\
&= KL(p_0|p_1) + \int_{x:p_1(x)>p_0(x)} p_0(x)\log(1 + \tfrac{p_1(x)-p_0(x)}{p_0(x)})dx \\
&\leq KL(p_0|p_1) + \int_{x:p_1(x)>p_0(x)} (p_1(x) - p_0(x))dx \\
&\leq KL(p_0|p_1) + 1
\end{aligned}$$

where the first inequality follows from $\log(1 + x) \leq x$. Putting the pieces together, we conclude that $\mathbb{E}_0[\tau] \leq \frac{\log(e/\delta)}{KL(p_0|p_1)} + 1$. Repeating the process for $\mathbf{H}_1$ produces an analogous result.

Binary hypothesis test for Gaussians with known variance Let $p_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $p_1(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\Delta)^2/2}$ so that we are deciding two Gaussian distributions, each with variance 1 and seperated by Δ. Note that

$$
\begin{aligned}
L_t &= \prod_{s=1}^{t} \frac{p_1(X_s)}{p_0(X_s)} \\
&= \prod_{s=1}^{t} \exp(-(X_s - \Delta)^2/2 + X_s^2/2) \\
&= \exp((\sum_{s=1}^{t} X_s)\Delta - t\Delta^2/2) \\
&= \exp(\Delta(S_t - t\Delta/2)
\end{aligned}
$$

where $S_t = \sum_{s=1}^{t} X_s$. Applying the maximal inequality of above and rearranging, we have

$$
\mathbb{P}_0(\exists t \in \mathbb{N} : S_t \geq t\Delta/2 + \log(1/\delta)/\Delta) = \mathbb{P}_0(\exists t \in \mathbb{N} : L_t \geq 1/\delta) \leq \delta.
$$

Compare this to the line-crossing super-martingale bound of above. They are equivalent with $\lambda = \Delta$. Because of the optimality of the SPRT, we conclude that a linear boundary is optimal for deciding between two means. Unfortunately, the precise parameterization of λ requires knowledge of the unknown parameter.

Chapter 4

Contextual Bandits

4.1 Introduction

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

For $t = 1, 2, \ldots$

- 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.

4.1.1 Finite contexts

Suppose the space of contexts, denoted $\mathcal{C}$, is finite. Then a natural algorithm would be to run an individual multi-armed bandit algorithm (action elimination, UCB, etc.) per context. We know that after T time steps, such a strategy applied to context $c \in \mathcal{C}$ would satisfy

$$\max_{x \in \mathcal{X}} \sum_{t=1}^{T} \mathbf{1}\{c_t = c\} v(c, x) - \sum_{t=1}^{T} \mathbf{1}\{c_t = c\} v(c, x_t) \lesssim \sqrt{T_c |\mathcal{X}| \log(T_c |\mathcal{X}|/\delta)} \leq \sqrt{T_c |\mathcal{X}| \log(T |\mathcal{X}|/\delta)}.$$

where $T_c := \sum_{t=1}^{T} \mathbf{1}\{c_t = c\}$. Summing over contexts were have

$$\sum_{c \in \mathcal{C}} \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \mathbf{1}\{c_t = c\} \left(v(c, x) - v(c, x_t) \right) \lesssim \sum_{c \in \mathcal{C}} \sqrt{T_c |\mathcal{X}| \log(T |\mathcal{X}|/\delta)}$$

$$\leq \sqrt{|\mathcal{C}| \, |\mathcal{X}| \log(T |\mathcal{X}|/\delta)} \tag{4.1}$$

by Cauchy-Schwartz.

The clear problem with this strategy is that this regret is trivial if $|\mathcal{C}|$ is very large. Indeed, if we had just *ignored* the context altogether at each time t, we could just play a multi-armed bandit algorithm to achieve

$$\max_{x \in \mathcal{X}} \sum_{t=1}^{T} \left(v(c_t, x) - v(c_t, x_t) \right) \lesssim \sqrt{|\mathcal{X}| \log(T |\mathcal{X}|/\delta)}. \tag{4.2}$$

While the right hand side of (4.2) appears much smaller than the right hand side of (4.1) (by a factor of $\sqrt{|\mathcal{C}|}$) we point out that the left had sides are not equivalent – they're using a different benchmark of regret! The first compares to the best single action *per context* whereas the latter compares itself to the single best action with respect to *all contexts*. The first is a much higher standard. The next section formalizes a unifying benchmark, which we denote as policy regret.

4.1.2   Policy Regret

A policy $\pi : \mathcal{C} \to \mathcal{X}$ maps contexts to actions. The value of a policy π is defiend as

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

We assume at the start of the game the learner has access to a set of policies $\diamond$ which may be infinite, but for simplicity we will assume it is finite. At each time, we assume the action taken is according to some policy $\pi_t \in \Pi$ so that the regret is defined as

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

and for convenience, we will fix a $\pi^\star := \arg\max_{\pi \in \Pi} V(\pi)$.

Connecting to the previous section, $|\Pi| = |\mathcal{X}|^{|\mathcal{C}|}$ and $|\Pi| = |\mathcal{X}|$ respectively.

4.2   Policy evaluation

Suppose for each policy $\pi \in \Pi$ we wished to estimate $V(\pi)$ up to tolerance $\epsilon > 0$ with probability at least $1 - \delta$. A naive strategy would be to simple play policy π for some number of trials to estimate its value. That is, for some $\tau \in \mathbb{N}$, play each $\pi \in \Pi$ for τ trials in response to the IID contexts. For each π this would result in a set of rewards $\{r_t^\pi\}_{t=1}^{\tau}$ each in $[0,1]$ which we can use to define $\widehat{V}(\pi) = \frac{1}{\tau}\sum_{t=1}^{\tau} r_t^\pi$. By Hoeffding's inequality and a union bound, we have that

$$\mathbb{P}\left(\bigcup_{\pi \in \Pi} \{|\widehat{V}(\pi) - V(\pi)| \geq \sqrt{\log(2|\Pi|/\delta)/2\tau}\right) \leq \delta.$$

Thus, if $\tau \geq \epsilon^{-2}\log(2|\Pi|/\delta)/2$ samples were taken for each policy $\pi \in \Pi$ we can estimate each $V(\pi)$ up to tolerance ϵ with probability at least $1 - \delta$ using $\epsilon^{-2}|\Pi|\log(2|\Pi|/\delta)/2$. But this linear scaling in $|\Pi|$ is awful if $|\Pi|$ is large! Can we do better?

4.2.1   Logging policy

The core difficulty of model evaluation in 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 $j \neq i$. But if every context appears very rarely so that I cannot rely on seeing the same context multiple times, how can I predict what I should have done? We will employ the use of a randomized logging policy to help us solve this riddle.

For any context c fix an exploration distribution $\mu(x|c) \in \triangle_{\mathcal{X}}$ such that $\mu(x|c) > 0$ for all x, c. The distribution $\mu(x|c)$ will act as our randomized logging policy to collect data in aid of estimating each $V(\pi)$ efficiently. We can define $\mu(x|c)$ independently of Π, or we can perform "proper learning"

so that $\mu(x|c)$ is actually playing a random policy $\pi \in \Pi$ at each time. To do this, fix some distribution over policies $\lambda \in \triangle_{\mathcal{X}}$, and then at each time t, draw $\pi_t \sim \lambda$ and play $x_t = \pi_t(c_t)$. Here, $\mu(x|c) = \sum_{\pi \in \Pi} \lambda_\pi \mathbf{1}\{\pi(c) = x\}$.

If we play the logging policy for $\tau \in \mathbb{N}$ rounds, at each time $t = 1, \ldots, \tau$ nature reveals a context $c_t \sim \mathcal{D}$, the logging policy plays $x_t \sim \mu(\cdot|c)$, and receives reward $r_t = v(c_t, x_t) + \epsilon_t \in [0, 1]$ where $\mathbb{E}[\epsilon_t] = 0$, by assumption. This results in a dataset $\{(c_t, x_t, r_t, p_t)\}_{t=1}^\tau$ where $p_t := \mu(x_t|c_t)$. Given this dataset, we wish to estimate each $V(\pi)$. We describe two ways to do so described as *model the bias* and *model the world*. We will also consider a hybrid of the two.

4.2.2 Model the bias

If we just naively estimated $V(\pi)$ with $\frac{1}{\tau} \sum_{t=1}^\tau r_t \mathbf{1}\{\pi(c_t) = x_t\}$ this would be a biased estimator: its expectation may not converge to $V(\pi)$ no matter how large τ is. We now define an unbiased estimator for $V(\pi)$. Define the *inverse propensity scoring* estimator as

$$\widehat{v}(c_t, x) := r_t \frac{\mathbf{1}\{x_t = x\}}{p_t}, \qquad \text{and} \qquad \widehat{V}(\pi) = \frac{1}{\tau} \sum_{t=1}^\tau \widehat{v}(c_t, \pi(c_t)).$$

Note that $\widehat{V}(\pi)$ is unbiased since

$$\mathbb{E}[\widehat{V}(\pi)] = \frac{1}{\tau} \sum_{t=1}^\tau \mathbb{E}[\widehat{v}(c_t, \pi(c_t))] = \frac{1}{\tau} \sum_{t=1}^\tau \mathbb{E}[\mathbb{E}[\widehat{v}(c_t, \pi(c_t))|c_t]] = \mathbb{E}_{C \sim \mathcal{D}}[v(C, \pi(C))] = V(\pi)$$

and

$$\mathbb{E}[\widehat{v}(c_t, x)|c_t] = \mathbb{E}\left[r_t \frac{\mathbf{1}\{x_t = x\}}{p_t}|c_t\right]$$

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

$$= \sum_{x' \in \mathcal{X}} \mu(x'|c_t)v(c_t, x')\frac{\mathbf{1}\{x' = x\}}{\mu(x'|c_t)}$$

$$= v(c_t, x).$$

The variance of $\widehat{V}(\pi)$ is

$$\mathbb{E}[(\widehat{V}(\pi) - V(\pi))^2] = \frac{1}{\tau^2} \sum_{t=1}^\tau \mathbb{E}[\mathbb{E}[(\widehat{v}(c_t, \pi(x)) - v(c_t, \pi(x)))^2|c_t]] \leq \frac{1}{\tau}\mathbb{E}_{C \sim \mathcal{D}}\left[\frac{1}{\mu(\pi(C)|C)}\right]$$

due to

$$\mathbb{E}[(\widehat{v}(c_t, x) - v(c_t, x))^2|c_t] \leq \mathbb{E}[(\widehat{v}(c_t, x))^2|c_t]$$

$$\leq \mathbb{E}\left[\frac{\mathbf{1}\{x_t = x\}}{\mu(x_t|c_t)^2}|c_t\right] \qquad (r_t \in [0, 1])$$

$$\leq \sum_{x' \in \mathcal{X}} \mu(x'|c_t)\frac{\mathbf{1}\{x' = x\}}{\mu(x'|c_t)^2}$$

$$= \frac{1}{\mu(x|c_t)}.$$

If $v_{\max} := \max_c \max_{\pi \in \Pi} \frac{1}{\mu(\pi(c)|c)}$ then by Bernstein's inequality we have

$$\mathbb{P}\left(\bigcup_{\pi \in \Pi} \{ |V(\pi) - \widehat{V}(\pi)| \leq \sqrt{\frac{2\log(2|\Pi|/\delta)}{\tau} \mathbb{E}_{C \sim \mathcal{D}} \left[\frac{1}{\mu(\pi(C)|C)} \right]} + \frac{2\widehat{v}_{\max}\log(2|\Pi|/\delta)}{3\tau} \} \right) \leq \delta$$

In particular, if we set $\mu(x|c) = \frac{1}{|\mathcal{X}|}$ for all x, c then for any $\tau \geq 2|\mathcal{X}|\log(2|\Pi|/\delta)$ we have

$$|V(\pi) - \widehat{V}(\pi)| \leq \sqrt{\frac{2|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau}} + \frac{2|\mathcal{X}|\log(2|\Pi|/\delta)}{3\tau} \leq \sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau}}$$

for all $\pi \in \Pi$ with probability at least $1 - \delta$. Thus, it suffices to take $\tau = 4\epsilon^{-2}|\mathcal{X}|\log(2|\Pi|/\delta)$ samples to estimate every $V(\pi)$ up to tolerance ϵ with probability at least $1 - \delta$.

Lemma 10 (Bernstein's inequality). *Let $X_1, \ldots, X_m$ be independent random variables such that $\frac{1}{m}\sum_{i=1}^m \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq \sigma^2$ and $|X_i| \leq B$. Then*

$$\left| \frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}[X_i] \right| \leq \sqrt{\frac{2\sigma^2\log(2/\delta)}{m}} + \frac{2B\log(2/\delta)}{3m}$$

with probability at least $1 - \delta$.

4.2.3   Model the world

Consider a function class $\mathcal{F}$ such that for each $f \in \mathcal{F}$ we have $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$. The idea is that some function $f_* \in \mathcal{F}$ is close enough to $v(c, x)$ to be useful, and that we can identify f_* or some function close to it using our collected dataset $\{(c_t, x_t, r_t, p_t)\}_{t=1}^\tau$. Let

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^\tau (r_t - f(c_t, x_t))^2.$$

We can then estimate $V(\pi)$ with $\widehat{V}(\pi) = \frac{1}{\tau}\sum_{t=1}^\tau \widehat{f}(c_t, \pi(c_t))$. Note that for any $f \in \mathcal{F}$ we have that

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^\tau (r_t - f(c_t, x_t))^2 \right] = \mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^\tau (r_t - v(c_t, x_t) + v(c_t, x_t) - f(c_t, x_t))^2 \right]$$

$$= \frac{1}{\tau}\sum_{t=1}^\tau \mathbb{E}\left[(r_t - v(c_t, x_t))^2 \right] + \frac{1}{\tau}\sum_{t=1}^\tau \mathbb{E}\left[(v(c_t, x_t) - f(c_t, x_t))^2 \right]$$

$$= \mathbb{E}\left[(r_1 - v(c_1, x_1))^2 \right] + \mathbb{E}\left[\sum_{x \in \mathcal{X}} \mathbf{1}\{x = x_1\}(v(c_1, x) - f(c_1, x))^2 \right]$$

$$= \mathbb{E}\left[(r_1 - v(c_1, x_1))^2 \right] + \mathbb{E}\left[\sum_{x \in \mathcal{X}} \mu(x|c_1)(v(c_1, x) - f(c_1, x))^2 \right]$$

$$\leq 1/4 + \mathbb{E}\left[\sum_{x \in \mathcal{X}} \mu(x|C)(v(C, x) - f(C, x))^2 \right]$$

Also note that by a similar calculation, we could consider an importance sampled version that satisfies

$$
\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{(r_t - f(c_t, x_t))^2}{p_t}\right] \leq \frac{1}{4}\mathbb{E}\left[\sum_{x\in\mathcal{X}}\frac{\mathbf{1}\{x_1 = x\}}{\mu(x|c_1)}\right] + \mathbb{E}\left[\sum_{x\in\mathcal{X}}\mathbf{1}\{x_1 = x\}\frac{(v(c_1, x) - f(c_1, x))^2}{\mu(x|c_1)}\right]
$$

$$
\leq \frac{|\mathcal{X}|}{4} + \mathbb{E}\left[\sum_{x\in\mathcal{X}}(v(c_1, x) - f(c_1, x))^2\right]
$$

but note that its not clear that this is a much better metric, but it will certainly have higher variance. Importantly, note that if $v \in \mathcal{F}$ and $\mu(x|c) > 0$ for all x, c then the minimum of both of these objectives are identical and equal to v, and as $\tau \to \infty$ we have that $\widehat{f} \to v$. We say the contextual bandit instance is *realizable* if $v \in \mathcal{F}$, and a great number of works have taken advantage of this fact. If $v \notin \mathcal{F}$ then this technique is biased and no matter how much data you collect, you may never get accurate estimates for the true value of a policy $\pi \in \Pi$. Nevertheless, this method is extremely popular in practice because its so easy to solve a least squares problem like the above for arbitrary function classes, like neural networks. Then people will just use $x_t = \arg\max_{x\in\mathcal{X}} f(c_t, x)$, totally bypassing the definition of the policy class Π.

4.2.4 Doubly robust estimators

The *model the bias* approach potentially has high variance and the *model the world* approach potentially has high bias. Doubly robust methods get the best of both worlds: unbiased but if the model $\widehat{f}$ is close to the true v then the variance is reduced [Dudík et al., 2011]. Technically, $\widehat{f}$ should be trained using data independent of our dataset $\{(c_t, x_t, r_t, p_t)\}_{t=1}^{\tau}$, like a hold-out set, but in practice people will often just reuse the dataset. Define

$$
\widehat{v}_{DR}(c_t, x) = \widehat{f}(c_t, x) + (r_t - \widehat{f}(c_t, x))\frac{\mathbf{1}\{x_t = x\}}{p_t}.
$$

It is easy to check this is unbiased with expectation $v(c_t, x)$. In the variance calculation of IPS we simply used the fact that r_t was bounded in magnitude by 1. Here, we will take advantage of the possibility that r_t may be close to $\widehat{f}(c_t, x)$. If r_t still has lots of intrinsic variance, this method won't help much, but if $|r_t - \widehat{f}(c_t, x)|$ is small, it can help a lot.

4.3 Stochastic Linear model

Consider a very special case of *model the world* where we assume that $v(c, x) = \langle \phi(c, x), \theta_* \rangle$ for some $\theta_* \in \mathbb{R}^d$ and Π is induced by all possible $\theta \in \mathbb{R}^d$ with $\pi(c) = \arg\max_{x\in\mathcal{X}}\langle \phi(c, x), \theta_* \rangle$ 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 $\widehat{\mathcal{X}}_t = \mathcal{X} \setminus \{x : \max_{x' \in \mathcal{X}} \langle x' - x, \theta \rangle < 0 \; \forall \theta \in C_t\}$ and play uniformly in this set. 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

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

where $\widetilde{\theta} = \arg\max_{\theta \in C_t} \langle x_t, \theta \rangle$. 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 7. *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*

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

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

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

4.4   Stochastic Contextual Bandits for General policy classes

Let's return to the general setting of trying to minimize policy regret without assuming a parametric structure on $v(c, x)$. Fix some policy set Π.

4.4.1 τ-greedy

Consider running the uniform exploration logging policy of above for τ steps. If $\epsilon_\tau := \sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau}}$ and $\widehat{\pi} = \arg\max_{\pi\in\Pi}\widehat{V}(\pi)$ then

$$V(\widehat{\pi}) = \underbrace{V(\widehat{\pi}) - \widehat{V}(\widehat{\pi})}_{\geq -\epsilon_\tau} + \underbrace{\widehat{V}(\widehat{\pi}) - \widehat{V}_t(\pi^\star)}_{\geq 0} + \underbrace{\widehat{V}(\pi^\star) - V(\pi^\star)}_{\geq -\epsilon_\tau} + V(\pi^\star)$$

$$\geq V(\pi^\star) - 2\epsilon_\tau$$

If we explore uniformly for τ rounds according to our logging policy and than exploit or $T - \tau$ rounds then we achieve a regret of at most

$$1 \cdot \tau + 2\epsilon_\tau(T - \tau) \leq \tau + T\sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau}}$$

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

4.4.2 Reduction to cost-sensitive classification

The above τ-greedy procedure requires a solution to the optimization problem $\widehat{\pi} = \arg\max_{\pi\in\Pi}\widehat{V}(\pi)$. Note that

$$\arg\max_{\pi\in\Pi}\widehat{V}(\pi) = \arg\max_{\pi\in\Pi}\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{\mathbf{1}\{x_t = \pi(c_t)\}}{p_t}r_t$$

$$= \arg\max_{\pi\in\Pi}\frac{1}{\tau}\sum_{t=1}^{\tau}(1 - \mathbf{1}\{x_t \neq \pi(c_t)\})\frac{r_t}{p_t}$$

$$= \arg\min_{\pi\in\Pi}\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbf{1}\{x_t \neq \pi(c_t)\}\frac{r_t}{p_t}.$$

where the last line is empirical risk minimization of the 0/1-loss with example-label pairs (c_t, x_t) weighted by $\frac{r_t}{p_t}$.

Example 2. *Let $\phi : \mathcal{C} \times \mathcal{X} \to \mathbb{R}^d$ be a feature map and assume Π is parameterized by $\mathbb{R}^d$ so that for every $\theta \in \mathbb{R}^d$ there exists a $\pi \in \Pi$ such that $\pi(c_t) = \arg\max_{x\in\mathcal{X}}\langle\theta, \phi(c_t, x)\rangle$. Note, unlike Section 4.3, we are not assuming anything about the relationship between $\langle\theta, \phi(c, x)\rangle$ and $v(c, x)$. A natural convex relaxation of $\mathbf{1}\{x_t \neq \pi(c_t)\}$ is cross-entropy loss $-\log(\frac{\exp(\langle\theta,\phi(c_t,x_t)\rangle)}{\sum_{x\in\mathcal{X}}\exp(\langle\theta,\phi(c_t,x)\rangle)})$. We can approximate $\widehat{\pi} = \arg\max_{\pi\in\Pi}\widehat{V}(\pi)$ with an iterative algorithm where $\theta_{k+1} = \theta_k + \eta_k\sum_{t=1}^{\tau}\frac{r_t}{p_t}\nabla_\theta\log(\frac{\exp(\langle\theta,\phi(c_t,x_t)\rangle)}{\sum_{x\in\mathcal{X}}\exp(\langle\theta,\phi(c_t,x)\rangle)})$ for some step size sequence η_k.*

4.4.3 Elimination algorithm

We will make the strong assumption that the distribution of contexts $\mathcal{D}$ is known a priori. This is not so implausible due to historical data, and often one can proceed in stages where the previous stage's data can be used to approximate the context distribution. The algorithm and analysis is inspired by [Dudik et al., 2011].

Recall the value of a policy $\pi \in \Pi$ as $V(\pi) = \mathbb{E}_{C \sim \mathcal{D}}[v(C, \pi(C))]$. Taking our G-optimality approach, we aim to sequentially define a distribution λ over policies Π and at each time play according to $\pi_t \sim \lambda$. Given an active set of policies still under consideration, we wish to identify the distribution that minimizes

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

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 $\widehat{\Pi}_1 \leftarrow \Pi, \ell \leftarrow 1, T_0 \leftarrow 0$
while $|\widehat{\Pi}_\ell| > 1$ **do**

$\quad \epsilon_\ell = 2^{-\ell}$, $\tau_\ell = \lceil 16n\epsilon_\ell^{-2} \log(2|\Pi|T/\delta) \rceil$, $\gamma_\ell = \min\{\frac{1}{2n}, \sqrt{\frac{\log(2|\Pi|T/\delta)}{9n\tau_\ell}}\}$, $T_\ell = T_{\ell-1} + \tau_\ell$

$\quad Q_\ell = \arg\min_{Q \in \triangle_{\widehat{\Pi}_\ell}} \max_{\pi \in \widehat{\Pi}_\ell} \mathbb{E}_C \left[\frac{1}{Q^{\gamma_\ell}(\pi(C)|C)} \right]$
$\quad\quad\quad$ s.t. $Q^\gamma(x|c) = \gamma + (1-\gamma n) \sum_{\pi \in \widehat{\Pi}_\ell : \pi(c) = x} Q(\pi)$

$\quad$ **for** $t = T_{\ell-1} + 1, \ldots, T_\ell$
$\quad\quad$ Observe context c_t
$\quad\quad$ Play $x_t \sim Q^\gamma(\cdot|c_t)$, set $p_t = Q^\gamma(x_t|c_t)$ and observe reward $r_t = v(c_t, x_t) + \eta_t$
$\quad$ Set $\widehat{V}_\ell(\pi) = \frac{1}{T_\ell - T_{\ell-1}} \sum_{t \in (T_{\ell-1}, T_\ell]} r_t \frac{\mathbf{1}\{\pi(c_t) = c_t\}}{p_t}$
$\quad \widehat{\Pi}_{\ell+1} \leftarrow \widehat{\Pi}_\ell \setminus \left\{ \pi \in \widehat{\Pi}_\ell | \max_{\pi' \in \widehat{\Pi}_\ell} \widehat{V}_\ell(\pi') - \widehat{V}_\ell(\pi) \geq 2\epsilon_\ell \right\}$
$\quad t \leftarrow t + 1$
Output: Π_{t+1}

The following lemma is somewhat of a generalization of Kiefer-Wolfowitz.

Lemma 11. *Let $\xi \in \Xi$ be a random variable and let $\phi : \mathcal{X} \times \Xi \to \mathbb{R}^d$. Then*

$$\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X}} \mathbb{E}_\xi \left[\phi(x, \xi)^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'} \phi(x', \xi) \phi(x', \xi)^\top \right)^\dagger \phi(x, \xi) \right] \leq d,$$

with equality if dimspan$(\{\phi(x, \xi) : x \in \mathcal{X}\}) = d$ *for all* $\mathcal{X} \in \mathcal{X}$.

Proof. Define $f(\lambda) = \mathbb{E}_\xi [f(\lambda; \xi)]$ and $f(\lambda; \eta) = \text{logdet} \left(\sum_{x \in \mathcal{X}} \lambda_x V_\eta \phi(x, \eta) \phi(x, \eta)^\top V_\eta^\top \right)$ where $V_\eta \in \mathbb{R}^{k \times d}$ satisfies $V_\eta^\top V_\eta \phi(x; \eta) = \phi(x; \eta)$ and $k = \text{dimspan}(\{\phi(x, \eta) : x \in \mathcal{X}\})$. First note that for any $A : \mathbb{R} \to \mathbb{R}^{k \times k}$ we have $\frac{d}{dt}\text{logdet}(A(t))|_{t=t_0} = \text{Trace}(A(t_0)^{-1} \frac{dA(t)}{dt}|_{t=t_0})$. Thus

$$\frac{\partial f(\lambda; \xi)}{\partial \lambda_x} = \text{Trace} \left(\left(\sum_{x' \in \mathcal{X}} \lambda_{x'} V_\xi \phi(x', \xi) \phi(x', \xi)^\top V_{\mathcal{X}}^\top \right)^{-1} V_\xi \phi(x, \xi) \phi(x, \xi)^\top V_\xi^\top \right)$$

$$= \phi(x, \xi)^\top V_\xi^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'} V_\xi \phi(x', \xi) \phi(x', \xi)^\top V_\xi^\top \right)^{-1} V_\xi \phi(x, \xi)$$

$$= \phi(x, \xi)^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'} \phi(x', \xi) \phi(x', \xi)^\top \right)^\dagger \phi(x, \xi)$$

Note that for any λ we have $\langle \nabla f(\lambda; \xi), \lambda \rangle = \text{dimspan}(\phi(x', \xi) : x' \in \mathcal{X})$. Let $\lambda^* = \arg\min_{\lambda \in \mathcal{X}} f(\lambda)$

and fix any $x \in \mathcal{X}$. Then by first order conditions,

$$
\begin{aligned}
0 &\geq \langle \nabla f(\lambda^*), \mathbf{e}_x - \lambda^* \rangle \\
&= \mathbb{E}_\xi \left[\langle \nabla f(\lambda^*; \xi), \mathbf{e}_x \rangle - \text{dimspan}(\phi(x', \xi) : x' \in \mathcal{X}) \right] \\
&\geq \mathbb{E}_\xi \left[\phi(x, \xi)^\top \left(\sum_{x' \in \mathcal{X}} \lambda^*_{x'} \phi(x', \xi) \phi(x', \xi)^\top \right)^\dagger \phi(x, \xi) \right] - \max_{\eta \in \mathcal{X}} \text{dimspan}(\phi(x', \eta) : x' \in \mathcal{X})
\end{aligned}
$$

Because $x \in \mathcal{X}$ was arbitrary, this completes the first part of the proof. Now suppose $d = \max_{\eta \in \mathcal{X}} \text{dimspan}(\phi(x', \eta) : x' \in \mathcal{X})$. Then by the previous display we have

$$
\begin{aligned}
d &\geq \max_{x \in \mathcal{X}} \mathbb{E}_\xi \left[\phi(x, \xi)^\top \left(\sum_{x' \in \mathcal{X}} \lambda^*_{x'} \phi(x', \xi) \phi(x', \xi)^\top \right)^\dagger \phi(x, \xi) \right] \\
&\geq \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X}} \mathbb{E}_\xi \left[\phi(x, \xi)^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'} \phi(x', \xi) \phi(x', \xi)^\top \right)^\dagger \phi(x, \xi) \right] \\
&\geq \min_{\lambda \in \triangle_{\mathcal{X}}} \sum_{x \in \mathcal{X}} \lambda_x \mathbb{E}_\xi \left[\phi(x, \xi)^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'} \phi(x', \xi) \phi(x', \xi)^\top \right)^\dagger \phi(x, \xi) \right] \\
&= d
\end{aligned}
$$

which completes the proof. $\square$

However, the original proof in [Dudik et al., 2011] proves this result in a very different way, appealing to Sion's minimax theorem.

Lemma 12. *For any finite policy set* Π *we have*

$$
\min_{Q \in \triangle_\Pi} \max_{\pi \in \Pi} \mathbb{E}_C \left[\frac{1}{Q(\pi(C)|C)} \right] \leq |\mathcal{X}|,
$$

with equality if $| \cup_{\pi \in \Pi} \pi(c)| = |\mathcal{X}|$ *for all* c. *Moreover, for any finite policy set* Π *and* $\gamma \leq \frac{1}{2|\mathcal{X}|}$ *we have*

$$
\min_{Q \in \triangle_\Pi} \max_{\pi \in \Pi} \mathbb{E}_C \left[\frac{1}{Q^\gamma(\pi(C)|C)} \right] = \min_{Q \in \triangle_\Pi} \max_{\pi \in \Pi} \mathbb{E}_C \left[\frac{1}{\gamma + (1 - \gamma|\mathcal{X}|)Q(\pi(C)|C)} \right] \leq 2|\mathcal{X}|.
$$

Proof. Consider $i = 1, \ldots, |\mathcal{X}|$ actions and for each $c \in \mathcal{C}$ define $\pi_c := \mathbf{e}_{\pi(c)} \in \{0, 1\}^{|\mathcal{X}|}$. For any $Q \in \triangle_\Pi$

$$
\pi_c^\top \left(\sum_{\pi' \in \Pi} Q(\pi') \pi'_c \pi'^\top_c \right)^\dagger \pi_c = \frac{1}{\sum_{\pi' \in \Pi : \pi'(c) = \pi(c)} Q(\pi')} = \frac{1}{Q(\pi(c)|c)}
$$

Applying the above lemma we have

$$
\min_{q \in \triangle_\Pi} \max_{\pi \in \Pi} \mathbb{E}_C \left[\pi_C^\top \left(\sum_{\pi' \in \Pi} q_{\pi'} \pi'_C \pi'^\top_C \right)^\dagger \pi_C \right] \leq |\mathcal{X}|.
$$

$\square$

Remark 1. *In light of the connection to Kiefer-Wolfowitz, where one appealed to Caratheodory's theorem to find a sparse solution that grew only quadratically in the dimension, one may wonder if Q can also be sparse in this setting. If $|\mathcal{C}| < \infty$ then by constructing a sparse solution via Caratheodory for each $c \in \mathcal{C}$ we can always find a solution Q that is $|\mathcal{C}||\mathcal{X}|$ sparse. Can we do better? Unfortunately, for some absolute constant $\alpha \gg 2$, [Agarwal et al., 2014] prove that for sufficiently small $\gamma > 0$ there exists a contextual bandit instance with $|\mathcal{C}| = \frac{1}{2\sqrt{2}\alpha|\mathcal{X}|\gamma}$ such that if a $Q \in \triangle_\Pi$ satisfies $\max_{\pi \in \Pi} \mathbb{E}_C \left[\frac{1}{\gamma + (1-\gamma n)Q(\pi(C)|C)} \right] \leq \alpha|\mathcal{X}|$, then $|support(Q)| \geq (|\mathcal{C}|-1)|\mathcal{X}| = \frac{1}{4\sqrt{2}\alpha\gamma}$. They also show how to obtain a $O(1/\gamma)$-sparse solution for a very similar optimization problem for any contextual bandit instance.*

Lemma 13. *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

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

$$\leq \sqrt{\frac{16n \log(2|\Pi|T/\delta)}{\tau_\ell}}$$

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

$$\widehat{V}_\ell(\pi) - \widehat{V}_\ell(\pi^\star) = \widehat{V}_\ell(\pi) - V(\pi) + \underbrace{V(\pi) - V(\pi^\star)}_{\leq 0} + V(\pi^\star) - \widehat{V}_\ell(\pi^\star)$$

$$\leq 2\epsilon_\ell.$$

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

$$\max_{\pi' \in \widehat{\Pi}_\ell} \widehat{V}_\ell(\pi') - \widehat{V}_\ell(\pi) \geq \widehat{V}_\ell(\pi^\star) - \widehat{V}_\ell(\pi)$$

$$= \widehat{V}_\ell(\pi^\star) - V(\pi^\star) + \underbrace{V(\pi^\star) - V(\pi)}_{>4\epsilon_\ell} + V(\pi) - \widehat{V}_\ell(\pi)$$

$$> 2\epsilon_\ell$$

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. $\square$

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:

$$T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1}) \rceil} (\gamma_\ell n + 8\epsilon_\ell(1 - \gamma_\ell n))\tau_\ell$$

$$= T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1}) \rceil} (n\sqrt{\tfrac{\log(2|\Pi|T/\delta)}{9n\tau_\ell}} + 8\epsilon_\ell)\tau_\ell$$

$$= T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1}) \rceil} n\sqrt{\lceil 16n\epsilon_\ell^{-2} \log(2|\Pi|T/\delta)\rceil \log(2|\Pi|T/\delta)/9n} + 8\epsilon_\ell\lceil 16n\epsilon_\ell^{-2}\log(2|\Pi|T/\delta)\rceil$$

$$\leq T\nu + \sum_{\ell=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1}) \rceil} 2n\epsilon_\ell^{-1}\log(4|\Pi|T/\delta) + 128n\epsilon_\ell^{-1}\log(4|\Pi|T/\delta)$$

$$\leq T\nu + 8 + 130n\log(2|\Pi|T/\delta) \sum_{t=1}^{\lceil \log_2(4(\Delta \vee \nu)^{-1}) \rceil} 2^t$$

$$\leq T\nu + 8 + 2080n(\Delta \vee \nu)^{-1}\log(2|\Pi|T/\delta).$$

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

Notes: [Dudik et al., 2011] compute a different Q_t every time a new context c_t arrives instead of our algorithm, above, which computes it only once per stage. Also, when estimating the value of a policy, they use all the observed data up to the current time, whereas our algorithm only uses the data from that round. Reusing data introduces dependencies that are easily handled by the martingale-based bounds of above since we can define our filtration as $\mathcal{F}_{t-1} = (c_1, x_1, r_1, \ldots, c_{t-1}, x_{t-1}, r_{t-1}, c_t)$ so that $c_t \in \mathcal{F}_{t-1}$ which makes p_t a predictable sequence.

While the τ-greedy algorithm is computationally efficient via a reduction to cost-sensitive classification, it is unclear how to make the above elimination algorithm computationally efficient. Fortunately, [Agarwal et al., 2014] did precisely that by approximately solving the optimization problem over $\triangle_\Pi$ using an iterative algorithm which results in a finite cover over Π.

4.4.4 A $\sqrt{T}$ computationally efficient algorithm

In this section we will propose an algorithm analogous to [Agarwal et al., 2014], but presented a bit differently.

Input: Policy set Π such that $\pi : \mathcal{C} \to [n]$ for all $\pi \in \Pi$, confidence level $\delta \in (0,1)$.
Let $\widehat{\Pi}_1 \leftarrow \Pi, \ell \leftarrow 1, T_0 \leftarrow 0$
for $\ell = 1, 2, \dots$

$\quad \epsilon_\ell = 2^{-\ell}, \tau_\ell = \lceil 16n\epsilon_\ell^{-2} \log(2|\Pi|T/\delta) \rceil, \gamma_\ell = \min\{\frac{1}{2n}, \sqrt{\frac{\log(2|\Pi|T/\delta)}{9n\tau_\ell}}\}, T_\ell = T_{\ell-1} + \tau_\ell$
$\quad$ Let Q_ℓ be any $Q \in \triangle_\Pi$ that satisfies both

$$\sum_{\pi \in \Pi} \widehat{\Delta}_{\ell-1}(\pi) Q(\pi) \le c_0 \epsilon_\ell \tag{4.3}$$

$$\mathbb{E}_C \left[\frac{1}{Q^{\gamma_\ell}(\pi(C)|C)} \right] \le n + \frac{\widehat{\Delta}_{\ell-1}(\pi)}{\gamma_\ell} \qquad \forall \pi \in \Pi \tag{4.4}$$

$$\text{where } Q^{\gamma}(x|c) = \gamma + (1 - \gamma n) \sum_{\pi \in \Pi : \pi(c) = x} Q(\pi)$$

$\quad$ **for** T_ℓ steps
$\quad\quad$ Observe context c_t
$\quad\quad$ Play $x_t \sim Q_\ell^{\gamma_\ell}(\cdot|c_t)$, set $p_t = Q_\ell^{\gamma_\ell}(x_t|c_t)$ and observe reward $r_t = v(c_t, x_t) + \eta_t$
$\quad$ Set $\widehat{V}_\ell(\pi) = \frac{1}{T_\ell - T_{\ell-1}} \sum_{t \in (T_{\ell-1}, T_\ell]} r_t \frac{\mathbf{1}\{\pi(c_t)=c_t\}}{p_t}$ and $\widehat{\Delta}_\ell(\pi) = \max_\pi \widehat{V}_\ell(\pi') - \widehat{V}_\ell(\pi)$
$\quad \ell \leftarrow \ell + 1$
Output: Π_{t+1}

On stage ℓ, for any $\lambda < \gamma_\ell$ and $\pi \in \Pi$, we have by Equation 3.1 that with probability at least $1 - \delta$ that $S \le \frac{\lambda V}{2(1-c\lambda)} + \log(|\Pi|/\delta)/\lambda$ where $S = \sum_{t \in (T_{\ell-1}, T_\ell]} r_t \frac{\mathbf{1}\{\pi(c_t)=c_t\}}{p_t}$, $V = \tau_\ell \mathbb{E}_C \left[\frac{1}{Q^{\gamma_\ell}(\pi(C)|C)} \right]$, and $c = 1/3\gamma_\ell$. Thus, taking $\lambda = \gamma_\ell/2$ we have with probability at least $1 - \delta$ that for all $\pi \in \Pi$

$$|\widehat{V}_\ell(\pi) - V(\pi)| \le \lambda \mathbb{E}_C \left[\frac{1}{Q^{\gamma_\ell}(\pi(C)|C)} \right] + \frac{\log(2|\Pi|T/\delta)}{\tau_\ell \lambda}$$

$$= \frac{\gamma_\ell}{2} \mathbb{E}_C \left[\frac{1}{Q^{\gamma_\ell}(\pi(C)|C)} \right] + \frac{\log(2|\Pi|T/\delta)}{\tau_\ell \gamma_\ell/2}.$$

Since $\gamma_\ell \approx \frac{\epsilon_\ell}{n}$, $\frac{\log(2|\Pi|T/\delta)}{\tau_\ell \gamma_\ell/2} \lesssim \epsilon_\ell$, and there always exists a Q to with $\mathbb{E}_C \left[\frac{1}{Q^{\gamma_\ell}(\pi(C)|C)} \right] \le 2n$, we have that we will have that $|\widehat{V}_1(\pi) - V(\pi)| \le \epsilon_1$. By an inductive argument, one can show that Equation 4.4 guarantees that $\widehat{V}_\ell(\pi) - V(\pi) \le c(\epsilon_\ell + \Delta(\pi))$ for some small $c < 1$. In particular, it implies that

$$\widehat{\Delta}_\ell(\pi) = \max_\pi \widehat{V}_\ell(\pi') - \widehat{V}_\ell(\pi)$$

$$\ge \widehat{V}_\ell(\pi^*) - \widehat{V}_\ell(\pi)$$

$$= \Delta(\pi) + \widehat{V}_\ell(\pi^*) - V(\pi^*) + V(\pi) - \widehat{V}_\ell(\pi)$$

$$= \Delta(\pi) - c(2\epsilon_\ell + \Delta(\pi))$$

which is at least $\Delta(\pi)/2$ when $\epsilon_\ell \le \Delta(\pi)/2$. This, in turn, implies that Equation 4.3 is using accurate estimates of $\widehat{\Delta}_\ell(\pi)$. Thus, approximating $\widehat{\Delta}_\ell(\pi) \approx \max\{\Delta(\pi), \epsilon_\ell$, if a feasible Q_ℓ is identified at each round, then during round ℓ one will incur an average regret of at most $c_0(\epsilon_\ell + \gamma_\ell n)$ since $\mathbb{P}(x_t = x) = \gamma_\ell + (1 - \gamma_\ell n)\mathbf{1}\{\pi_t(c_t) = x\}$ and $\pi_t \sim Q_\ell$. Then the regret analysis of above follows identically. Now all that remains is to show that (i) there exists a feasible Q_ℓ at each round with high probability, and (ii) such a Q_ℓ can be identified using a computationally efficient procedure.

To solve (i) we will explicitly construct a feasible Q_ℓ. For $j \le \ell$ define $\Pi_j = \{\pi \in \Pi : \Delta(\pi) \le \epsilon_\ell\}$ and

$$P_j = \arg \min_{P \in \triangle_\Pi : \text{support}(P) \subset \Pi_j} \max_{\pi \in \Pi_j} \mathbb{E}_C \left[\frac{1}{P^{\gamma_j}(\pi(C)|C)} \right]$$

where $P^\gamma(x|c) = \gamma + (1 - \gamma n) \sum_{\pi \in \Pi_j : \pi(c) = x} P(\pi)$. With τ_j and γ_j defined as above, let $\bar{P}_\ell = \frac{1}{T_\ell} \sum_{j=1}^\ell P_j \tau_j$ where we recall that $T_\ell = \sum_{j=1}^\ell \tau_j$. Note that P_j is essentially the distribution identified in the computationally inefficient algorithm at each round j. Thus, Equation 4.4 will be satisfied almost immediately, since for any π such that $\Delta(\pi) \leq \epsilon_j$ we have $\bar{P}_\ell^{\gamma_\ell}(\pi(c)|c) T_\ell \geq P_j^{\gamma_j}(\pi(c)|c)\tau_j$ which makes the LHS less than $O(\epsilon_j)$. To show that Equation 4.3 is satisfied, we observe that

$$\sum_{\pi \in \Pi} \hat{\Delta}_{\ell-1}(\pi) \bar{P}_\ell(\pi) \approx \sum_{\pi \in \Pi} \Delta(\pi) \bar{P}_\ell(\pi)$$

$$= \sum_{j=0}^\ell \sum_{\pi \in \Pi : \Delta(\pi) \leq \epsilon_j} \Delta(\pi) \bar{P}_\ell(\pi)$$

$$\leq \frac{1}{T_\ell} \sum_{j=0}^\ell \sum_{\pi \in \Pi : \Delta(\pi) \leq \epsilon_j} \epsilon_j P_j(\pi) \tau_j$$

$$\lesssim \frac{1}{T_\ell} \sum_{j=0}^\ell \sum_{\pi \in \Pi : \Delta(\pi) \leq \epsilon_j} \epsilon_j P_j(\pi) n \epsilon_j^{-2} \log(2|\Pi|T/\delta)$$

$$= \frac{1}{T_\ell} \sum_{j=0}^\ell n \epsilon_j^{-1} \log(2|\Pi|T/\delta)$$

$$\lesssim \frac{1}{T_\ell} n \epsilon_\ell^{-1} \log(2|\Pi|T/\delta)$$

$$\lesssim \frac{1}{\tau_\ell} n \epsilon_\ell^{-1} \log(2|\Pi|T/\delta)$$

$$\lesssim \epsilon_\ell$$

where we have used the inequality $\sum_{j=0}^\ell a^j \leq 2a^\ell$ for $a \geq 2$.

To solve (ii) we employ the use of a Frank-Wolfe style algorithm, as used in [Agarwal et al., 2014]. In particular, at each step we find some π that invalidates Equation 4.4 (which can be cast as a cost-sensitive classification problem) increase $Q(\pi)$, and then renormalize Q so that Equation 4.3 is satisfied. The process will eventually terminate with a Q that doesn't necessarily sum to one (but does not exceed it). To make the policy sum to 1, use the empirical best policy.

4.4.5 Frank-Wolfe

Frank-Wolfe for contextual bandits

Input: $\phi : [n] \times \Xi \to \mathbb{R}^d$, $\gamma > 0$, $\alpha \in \mathbb{R}_+^n$, $\lambda^{(0)} \in \triangle_n$

Set: $t = 1$, $\omega_i(\lambda) = \mathbb{E}\left[\phi(j,\xi)^\top (\sum_{i=1}^n \lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)^{-1} \phi(j,\xi)\right]$

do

$\quad k = \arg\max_{i=1,\ldots,n} \omega_i(\lambda^{(t-1)}) - \alpha_i$

$\quad \eta = \frac{\omega_i(\lambda^{(t-1)}) - \alpha_i}{\omega_i(\lambda^{(t-1)})^2}$

$\quad \lambda^{(t)} = \lambda^{(t-1)} + \eta \mathbf{e}_k$

$\quad$ **if** $\sum_{j=1}^n \alpha_j \lambda_j^{(t)} > d$

$\quad\quad \lambda^{(t)} = \lambda^{(t)} \cdot \frac{d}{\sum_{j=1}^n \alpha_j \lambda_j^{(t)}}$

$\quad t = t + 1$

until $\max_{i=1,\ldots,n} \omega_i - 2\alpha_i \leq 0$

Output: Single element in V

$$f(\lambda) = -\mathbb{E}_\xi[\text{logdet}(\sum_{i=1}^{n} \lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)] + \sum_{i=1}^{n} \lambda_i \alpha_i$$

$$\frac{\partial f(\lambda)}{\partial \lambda_j} = -\mathbb{E}\left[\text{Trace}\left((\sum_{i=1}^{n} \lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)^{-1}\phi(j,\xi)\phi(j,\xi)^\top\right)\right] + \alpha_j$$

$$= -\mathbb{E}\left[\phi(j,\xi)^\top(\sum_{i=1}^{n} \lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)^{-1}\phi(j,\xi)\right] + \alpha_j$$

$$=: -\omega_j + \alpha_j$$

For some λ and $k \in [n]$ such that $\omega_k \geq 2\alpha_k$ consider

$$f(\lambda + \eta \mathbf{e}_k) = -\mathbb{E}_\xi[\text{logdet}(\sum_{i=1}^{n} \lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I + \eta\phi(k,\xi)\phi(k,\xi)^\top)] + \sum_{i=1}^{n} \lambda_i \alpha_i + \eta\alpha_k$$

$$= f(\lambda) - \log(1 + \eta\omega_k) + \eta\alpha_k$$

$$\leq f(\lambda) - \eta\omega_k + \eta^2\omega_k^2/2 + \eta\alpha_k$$

$$= f(\lambda) - \frac{(\omega_k - \alpha_k)^2}{2\omega_k^2}$$

$$\leq f(\lambda) - 1/2$$

using the fact that $\log(1 + x) \geq x - x^2/2$ and the prescribed step size.

Fix λ and let $g(c) = f(c\lambda)$ so that

$$g(c) = -\mathbb{E}_\xi[\text{logdet}(\sum_{i=1}^{n} c\lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)] + \sum_{i=1}^{n} c\lambda_i \alpha_i$$

$$g'(c) = -\mathbb{E}\left[\text{Trace}\left((\sum_{i=1}^{n} c\lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)^{-1}\sum_{j=1}^{n} \lambda_j\phi(j,\xi)\phi(j,\xi)^\top\right)\right] + \sum_{j=1}^{n} \alpha_j\lambda_j$$

$$= -\frac{1}{c}\mathbb{E}\left[\text{Trace}\left((\sum_{i=1}^{n} c\lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)^{-1}(c\sum_{j=1}^{n} \lambda_j\phi(j,\xi)\phi(j,\xi)^\top + \gamma I - \gamma I)\right)\right] + \sum_{j=1}^{n} \alpha_j\lambda_j$$

$$= -\frac{1}{c}\mathbb{E}\left[\text{Trace}\left(I - \gamma(\sum_{i=1}^{n} c\lambda_i \phi(i,\xi)\phi(i,\xi)^\top + \gamma I)^{-1}\right)\right] + \sum_{j=1}^{n} \alpha_j\lambda_j$$

$$\geq -\frac{d}{c} + \sum_{j=1}^{n} \alpha_j\lambda_j.$$

Thus, $g(1) \geq g(c)$ if $c = \frac{d}{\sum_{j=1}^{n} \alpha_j\lambda_j}$. So if at some point in the algorithm, if $\sum_{j=1}^{n} \alpha_j\lambda_j > d$ then $\lambda \mapsto c\lambda$. Thus, the projection step does not increase the objective.

If $\lambda^{(0)} = 0$ then the algorithm terminates after $O(n\log(1/\gamma_L))$ steps.

Chapter 5

Other topics in bandits

5.1 Non-parametric bandits

5.1.1 Bandits in an RKHS, Gaussian Process Bandits, Bayesian Optimization

5.1.2 Bandit Convex Optimization

5.1.3 Lipschitz Bandits

5.2 Infinite-armed bandits

5.3 Alternative kinds of feedback

5.3.1 Dueling bandits

5.3.2 Slates

Chapter 6

Active Learning for Classification

This chapter is concerned learning a binary classifier while requesting as few labels as possible. Specifically, for an example space $\mathcal{X}$ and label space $\{0,1\}$ let $\mathcal{H}$ be a *hypothesis class* such that for each $h \in \mathcal{H}$ we have $h : \mathcal{X} \to \{0,1\}$. For example, $\mathcal{X}$ could be a set of images in some database, and each image $x \in \mathcal{X}$ either contains a man-made object $y = 1$ or not $y = 1$. After labels for examples in a subset of $\mathcal{X}$ are observed, one can construct $\widehat{h} \in \mathcal{H}$ which can then be applied to every image in the database $\mathcal{X}$ to predict their unknown labels.

In the *pool-based* setting, at any point we can choose any $x \in \mathcal{X}$ and then request its label $y \in \{0,1\}$. We evaluate an algorithm by the number of labels requested and the number of errors the learned classifier $\widehat{h}$ makes over the entire pool $\mathcal{X}$, whether they were queried or not. In the *streaming* setting, there exists a distribution $\mathcal{D}$ over $\mathcal{X}$ and we can only obtain samples $x \sim \mathcal{D}$. In this setting, we evaluate an algorithm based on how many total examples are drawn from $\mathcal{D}$ (amount of unlabeled data), the number of labels that are requested, and the error of the classifier over $\mathcal{X}$ with respect to $\mathcal{D}$. Note, by taking $\mathcal{D}$ to be a uniform distribution over $\mathcal{X}$, one can always apply an algorithm for the streaming setting to the pool-based setting. On the other hand, if we had no restriction on the amount of unlabeled data, one could keep sampling from $\mathcal{D}$ until any desired $x \in \text{support}(\mathcal{D}) \subset \mathcal{X}$ eventually is returned (though this could take a very long time).

6.1 Separable, pool-based setting

We say a problem is *separable* if there exists an $h^* \in \mathcal{H}$ such that for every $x \in \mathcal{X}$ and its corresponding label $y \in \{0,1\}$, $h^*(x) = y$. The goal of *exact learning* in the pool-based setting is to identify h^* using as few queries as possible. In this section, we propose different strategies for accomplishing this. For a deterministic algorithm $\mathcal{A}$, let $\mathfrak{S}(\mathcal{X}, \mathcal{H}, h, \mathcal{A})$ be the number of labels that algorithm $\mathcal{A}$ requests before identifying the true hypothesis $h^* = h$, and let $\mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A}) = \max_{h \in \mathcal{H}} \mathfrak{S}(\mathcal{X}, \mathcal{H}, h, \mathcal{A})$. In the pool-based setting, $|\mathcal{X}| < \infty$ which implies $|\mathcal{H}| < \infty$. Thus, any deterministic algorithm can be thought of as a binary tree where each node specifies which $x \in \mathcal{X}$ to request the label for, and each child node to move to depends on the label $y \in \{0,1\}$. The leaves of this tree correspond to a unique $h \in \mathcal{H}$. Because a binary tree with $|\mathcal{H}|$ leaves has depth at least $\lceil \log_2(|\mathcal{H}|) \rceil$ we have the immediate proposition.

Proposition 4. *For any hypothesis space $\mathcal{H}$ defined over $\mathcal{X}$, any algorithm $\mathcal{A}$ satisfies $\mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A}) \geq \lceil \log_2 \mathcal{H} \rceil$.*

Note that for some hypothesis classes, this lower bound is achievable:

Example 3 (Binary search). *Let $n \in \mathbb{N}$ be a power of two, $\mathcal{X} = \{1, \ldots, n\}$, and $\mathcal{H}_{thresholds} = \{h(x) = \mathbf{1}\{x \in \{1, \ldots, k\}\} : k \in \{1, \ldots, n\}\}$. Each $h \in \mathcal{H}$ corresponds to a unique index $k \in [n]$ and the label for each query $i \in \mathcal{X} = [n]$ is equivalent to asking whether $i \leq k$ or $i > k$. Performing bisection/binary search identifies the correct hypothesis after exactly $\log_2(n) = \log_2(|\mathcal{H}|)$. Thus, $\min_{\mathcal{A}} \mathfrak{S}(\mathcal{X}, \mathcal{H}_{thresholds}, \mathcal{A}) = \lceil \log_2 |\mathcal{H}_{thresholds}| \rceil$.*

However, clearly $\log_2(|\mathcal{H}|)$ is not always possible:

Example 4 (Needle in a haystack). *Fix $n \in \mathbb{N}$, $\mathcal{X} = \{1, \ldots, n\}$, and $\mathcal{H}_{needle} = \{h(x) = \mathbf{1}\{x = k\} : k \in \{1, \ldots, n\}\}$. Exhaustive search is clearly the best one could hope for here, which takes $n - 1 = |\mathcal{H}| - 1$ queries. Thus, $\min_{\mathcal{A}} \mathfrak{S}(\mathcal{X}, \mathcal{H}_{needle}, \mathcal{A}) = |\mathcal{H}_{needle}| - 1$*

Moreover, its easy to see that $\min_{\mathcal{A}} \mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A}) \leq |\mathcal{H}| - 1$ in general since without loss of generality we have $|\mathcal{X}| \leq |\mathcal{H}|$, otherwise queries in $\mathcal{X}$ are equivalent. The above two examples show that $\min_{\mathcal{A}} \mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A})$ lies somewhere between $\lceil \log_2 |\mathcal{H}| \rceil$ and $|\mathcal{H}| - 1$. Can we get a bit tighter?

6.1.1   Extended teaching dimension and the Halving algorithm

Here we introduce a combinatorial quantity called the extended teaching dimension that describes the complexity of exact learning [Hegedus, 1995]. First, some definitions. Let $n = |\mathcal{X}|$.

Definition 4. *We say $S \subset \mathcal{X}$ is a* specifying set *for $b \in \{0,1\}^n$ (with respect to $\mathcal{H}$) if $|\{h \in \mathcal{H} : h(x) = b(x) \ \forall x \in S\}| \leq 1$.*

Definition 5. *For any $\mathcal{X}$ and hypothesis class $\mathcal{H}$ over $\mathcal{X}$, define extended teaching dimension $\mathrm{EXT\text{-}TD}(\mathcal{H})$ as $\mathrm{EXT\text{-}TD}(\mathcal{H}) = \min\{k : \forall b \in \{0,1\}^{|\mathcal{X}|}, \exists$ specifying set S for b with $|S| \leq k\}$.*

Examples make these definitions clearer.

Example 5 (Thresholds). *If $b = \mathbf{0}$ then $S = \{1\}$ suffices since $h(1) = 1$ for all $h \in \mathcal{H}$. If $b \in \{0,1\}^n$ such that $b \notin \mathcal{H}_{thresholds}$ and $b \neq \mathbf{0}$ then there exists some $1 \leq i < j \leq n$ such that $b(i) = 0$ and $b(j) = 1$ so $S = \{i, j\}$ suffices. Finally, if $b \in \mathcal{H}$ then there exists an index i such that $b(i) = 1$ and $b(i+1) = 0$ and so $S = \{i, i+1\}$ suffices. Thus $\mathrm{EXT\text{-}TD}(\mathcal{H}_{thresholds}) = 2$.*

Example 6 (Needle in a Haystack). *Take S to be any subset of $[n]$ with $|S| = n - 1$. Fix any $b \in \{0,1\}^n$. If $b(i) = 0$ for all $i \in S$ then the remaining index either uniquely specifies an $h \in \mathcal{H}$ or is equal to 0, in which no $h \in \mathcal{H}$ is consistent. If $|\{b(i) = 1 : i \in S\}| \geq 1$ then either no $h \in \mathcal{H}$ is consistent or the correct h is specified (since b cannot contain multiple 1s). Finally, if $b = \mathbf{0}$ and $|S| < n - 2$ then there are two hypotheses in $\mathcal{H}$ that are consistent with any $S \subset \mathcal{X}$. Together, these cases imply that $\mathrm{EXT\text{-}TD}(\mathcal{H}_{needle}) = |S| - 1$.*

Theorem 8. *For any $\mathcal{H}$ we have $\mathrm{EXT\text{-}TD}(\mathcal{H}) \leq \min_{\mathcal{A}} \mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A}) \leq \mathrm{EXT\text{-}TD}(\mathcal{H}) \lceil \log_2(|\mathcal{H}|) \rceil$. Moreover, the upper bound is achieved by the* Halving *algorithm.*

Proof. Suppose $\mathcal{A}$ is an algorithm that takes at most $\mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A})$ queries when run on any instance $h \in \mathcal{H}$. Without loss of generality, we may assume that after receiving a label in each round, $\mathcal{A}$ checks whether there exists more than one hypothesis consistent with the observations. If so it continues, if not it stops. Moreover, assume that if $\mathcal{A}$ is played on some $b \notin \mathcal{H}$ and it encounters a query that is inconsistent with any $h \in \mathcal{H}$ it also stops and outputs FAIL. Note that this algorithm takes no more than $\mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A})$ queries when it is run on any $b \in \{0,1\}^n$. Thus, when run on any

$b \in \{0, 1\}^n$, the set of measured example, label pairs is a specifying set for b with respect to $\mathcal{H}$. This proves the lower bound.

The upper bound follows from the HALVING algorithm. At each round, if $b(x) = h^*(x)$ for all $x \in S$ then $V = \{h^*\}$ by the definition of the specifying set. If there exists an $x \in S$ such that $b(x) \neq h^*(x)$ then at least half of the hypotheses in V at the start of the round are removed, due to majority vote. Because this can occur at most $\lceil \log_2(|\mathcal{H}|) \rceil$ times and the size of each specifying set is at most EXT-TD($\mathcal{H}$), the result follows. $\square$

HALVING Algorithm for exact learning
Input: Finite hypothesis set $\mathcal{H}$ such that each $h : \mathcal{X} \rightarrow \{0, 1\}$.
Initialize: version space $V = \mathcal{H}$
while $|V| > 1$
    Set $b(x) = $ MAJORITYVOTE($h(x) : h \in V$) for all $x \in \mathcal{X}$
    Let $S \subset \mathcal{X}$ be a minimal specifying set for $b \in \{0, 1\}^{|\mathcal{X}|}$ with respect to $\mathcal{H}$ and request $h^*(x)$ for all $x \in S$
    Update $V = \{h \in V : h(x) = h^*(x)\}$
Output: Single element in V

The result of Theorem 8 sheds light on the sample complexity of exact learning. However, the HALVING algorithm and its analysis has many downsides. First, its not clear how to compute a minimal specifying set. Second, even if one could, the computational complexity of the algorithm scales like $|\mathcal{X}| |\mathcal{H}|$ which is almost always intractable. Finally, even as a theoretical result alone, the definition of EXT-TD($\mathcal{H}$) is very combinatorial and can be difficult to bound.

6.1.2 Generalized binary search

There have been many variants of greedy information gain algorithms that have been proposed and analyzed. They go by names like query by committee, splitting algorithm, or generalized binary search [Freund et al., 1997, Dasgupta, 2005a, Nowak, 2011, Golovin and Krause, 2011]. These generalized binary search algorithms define a probability distribution p over $\mathcal{H}$ and take queries to remove as much mass as possible.

Generalized Binary Search (GBS) for exact learning
Input: Finite hypothesis set $\mathcal{H}$ such that each $h : \mathcal{X} \rightarrow \{0, 1\}$. Probability distribution $p \in \triangle_{\mathcal{H}}$
Initialize: Version space $V = \mathcal{H}$
while $|V| > 1$
    Set $x' \leftarrow \arg\min_{x \in \mathcal{X}} \left| \frac{1}{2} - \sum_{h \in V} p(h) h(x) \right|$
    Request $h^*(x')$ and update $V = \{h \in V : h(x') = h^*(x')\}$
Output: Single element in V

Theorem 9 ([Dasgupta, 2005a]). *Fix any finite hypothesis class $\mathcal{H}$ and $p \in \triangle_{\mathcal{H}}$. Let* OPT $= \min_{\mathcal{A}} \mathbb{E}_{h \sim p} [\mathfrak{S}(\mathcal{X}, \mathcal{H}, h, \mathcal{A})]$. *The GBS algorithm satisfies* $\mathbb{E}_{h \sim p} [\mathfrak{S}(\mathcal{X}, \mathcal{H}, h, \mathcal{A}_{GBS})] \leq 4$OPT $\max_{h \in \mathcal{H}} \log(1/p(h))$.

When p is taken to be the uniform distribution over $\mathcal{H}$ this amounts to a $\log(|\mathcal{H}|)$ approximation ratio. Note that while the previous section has been considering $\mathfrak{S}(\mathcal{X}, \mathcal{H}, \mathcal{A}) \max_{h \in \mathcal{H}} \mathfrak{S}(\mathcal{X}, \mathcal{H}, h, \mathcal{A})$, this theorem is concerned with the average case sample complexity $\mathbb{E}_{h \sim p} [\mathfrak{S}(\mathcal{X}, \mathcal{H}, h, \mathcal{A})]$.

For some special classes of $\mathcal{H}$, or those that possess certain geometrical properties, this average-case approximation ratio can be shown to be a constant [Nowak, 2011]. A related greedy approach attempts to remove as many pairs of hypotheses that cannot be distinguished between. Its sample complexity is given in terms of the *splitting index* [Dasgupta, 2005b].

6.1.3   Open problems

All of the above algorithms are computationally infeasible since they scale like $|\mathcal{X}||\mathcal{H}|$. If $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{H}$ is the space of linear half-spaces then $O(|\mathcal{X}|^d)$. Ideally, an algorithm for this setting would only require calls to an oracle of the form $\arg\min_{h\in\mathcal{H}} \sum_{x\in S} \mathbf{1}\{h(x) \neq h^*(x)\}$ for any $S \in \mathcal{X}$. In special cases, such as the space of half-space classifiers, one can efficiently approximate the volume of the version space thereby making algorithms like GBS with a uniform prior potentially runnable.

For this pool-based setting, the extended teaching dimension is the only result that I am aware of that provides a nearly matching upper and lower bound that quantifies the sample complexity (i.e., not a non-constructive OPT). [Dasgupta, 2005b] proves an upper bound in terms of the splitting index. However, the lower bound relies on either the number of unlabeled data being large *or* the number of labels needed being large–this is vacuous in the pool-based setting. Reducing the gap between the upper and lower bound of Theorem 8 is a great open problem. Especially if the new bounds are in terms of efficiently computable quantities.

The extended teaching dimension bounds are worst-case over the class of $\mathcal{H}$. Which means that if I embed a set of hard instances in $\mathcal{H}$, the teaching dimension and thus the sample complexity of this appended class is at least as hard as learning the hard instances. Is there some notion of segmenting $\mathcal{H}$ into equivalence classes of instances of difficulty such that some instances are learnable with very queries while others may require a lot, and this is reflected in the sample complexity?

6.2   Separable, streaming setting

While in the pool-based setting we assumed that $|\mathcal{X}| < \infty$ and can be enumerated over, when we move to the streaming setting we make no such restriction. Here $\mathcal{X}$ can be uncountable and we assume we have access to a sampling oracle such that we can query an $x_t \sim \mathcal{D}_X$ over $\mathcal{X}$. If we request a corresponding label $y_t \in \{0, 1\}$ we assume that the pair $(x_t, y_t) \sim \mathcal{D}$. For simplicity we will still assume $\mathcal{H}$ is finite.

Define the *risk* of any $h \in \mathcal{H}$ with respect to $\mathcal{D}$ as $R(h) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathbf{1}\{h(X) \neq Y\}]$. We do not assume that we know $\mathcal{D}$ directly, but we can collect an iid dataset $\{(x_i, y_i)\}_{i=1}^n$ from drawn from $\mathcal{D}$. Hence, define the *empirical risk* as $\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$. For i.i.d. draws $\{x_i, y_i\}_{i=1}^n$, the empirical risk $\widehat{R}(h)$ is an unbiased estimator of the true risk $R(h)$ for any hypothesis $h \in \mathcal{H}$. Let $h^* = \arg\min_{h\in\mathcal{H}} R(h)$ be a hypothesis of minimum true risk in $\mathcal{H}$. Then, the goal of the learner is to return a hypothesis $h \in \mathcal{H}$ with true risk $R(h)$ as close as possible to the minimum true risk $R(h^*)$.

6.2.1   Review of passive learning

We start with the problem of passive learning for binary classification. Herein, we have a set of data points and labels $\{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from distribution $\mathcal{D}$. Our goal is to find how fast the true risk goes down as a function of n. Intuitively, the lower the risk we want, the more data points we need to sample. Further, there must be a positive relationship between the number of hypotheses we have in the hypothesis space and the number of data points we need to identify the hypothesis with the minimum true risk almost surely. The following theorem characterizes these relationships:

Theorem 10. *Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ and a finite set of hypotheses $\mathcal{H}$ (i.e., $|\mathcal{H}| < \infty$). We assume that the data is separable so that* $\min_{h\in\mathcal{H}} \mathbb{P}_{(X,Y)\sim\mathcal{D}}(h(X) \neq Y) = 0$. *Given n IID*

draws from $\mathcal{D}$, $\{(x_i, y_i)\}_{i=1}^{n}$, let $\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h)$ be the empirical risk minimizer. For any $\epsilon, \delta \in (0, 1)$, we have $\mathbb{P}\big(R(\hat{h}_n) > \epsilon\big) \leq \delta$ whenever $n \geq \epsilon^{-1} \log(\frac{|\mathcal{H}|}{\delta})$. In other words, for any $\epsilon, \delta \in (0, 1)$, with probability $1 - \delta$, we have $R(\hat{h}_n) \leq \frac{\log(\frac{|\mathcal{H}|}{\delta})}{n}$

Before going over the proof, it is worth noting that the theorem assumes a finite set of hypotheses in the hypothesis class $\mathcal{H}$. Such assumption, besides being practical from an optimization point of view, also simplifies the theory. Given that an infinite hypothesis class applied to a finite number of points observations yields a finite number of labellings, intuitively, assuming a finite hypothesis should not be too restrictive. We refer the interested reader to Boucheron et al. [Boucheron, 2005] in order to better understand how to make this argument rigorous by appealing to quantities like the Vapnik-Chervonenkis (VC) dimension and Rademacher complexity.

Proof. First note that $\hat{R}_n(\hat{h}_n) = 0$ since $\hat{R}_n(\hat{h}_n) = \min_{h \in \mathcal{H}} \hat{R}_n(h) \leq \hat{R}_n(h^*) = 0$. Now we can write:

$$\Pr\big(R(\hat{h}_n) > \epsilon\big) = \Pr\big(\bigcup_{h \in \mathcal{H}} \{R(h) > \epsilon \wedge \hat{R}_n(h) = 0\}\big) \leq \sum_{h \in \mathcal{H}} \Pr\big(\{R(h) > \epsilon \wedge \hat{R}_n(h) = 0\}\big), \quad (6.1)$$

where we performed a union bound to obtain the inequality. We can now find a bound for each element—$\Pr\big(\{R(h) > \epsilon, \hat{R}_n(h) = 0\}\big)$. This is the probability that a hypothesis with true risk greater than ϵ shows zero empirical risk in n points drawn i.i.d. from $\mathcal{D}$. Since the true risk for this hypothesis is greater than ϵ, the probability that this hypothesis correctly identifies a random point is lower than $1 - \epsilon$. Thus, we can write:

$$\Pr\big(R(\hat{h}_n) > \epsilon\big) \leq \sum_{h \in \mathcal{H}} \Pr\big(\{R(h) > \epsilon \wedge \hat{R}_n(h) = 0\}\big) \leq \sum_{h \in \mathcal{H}} (1 - \epsilon)^n \leq |\mathcal{H}| e^{-n\epsilon}$$

using the approximation $1 - x \leq e^{-x}$ for $x \geq 0$. Thus, $\Pr\big(R(\hat{h}_n) > \epsilon\big) \leq |\mathcal{H}| e^{-\epsilon n} = \delta$. Solving for n, we find $n \geq \epsilon^{-1} \log(\frac{|\mathcal{H}|}{\delta})$, and this completes the proof. $\square$

Example 7. *As a concrete example, let us assume x being uniform on $[0, 1]$, and that the hypothesis class is defined as $\mathcal{H} = \{\mathbf{1}\{x \leq \frac{i-1}{m-1}\} : i = 1, ..., m\}$; that is, there are m classifiers uniformly spaced in our hypothesis class (i.e., $|\mathcal{H}| = m$). If $y = h^*(x)$ for some $h^* \in \mathcal{H}$ (i.e., the perfect classifier exists in the hypothesis class), then we can apply Theorem ?? and say that after n observations with probability at least $1 - \delta$ we have $R(\hat{h}_n) \leq \frac{\log(m/\delta)}{n}$.*

6.2.2 CAL, Disagreement-based learning

Now we will attempt to reduce the number of labels that are requested while achieving the same performance as passive learning. The style of algorithm we consider here is known as a disagreement-based learner. [Hanneke et al., 2014] provides an excellent survey of these methods.

Definition 6. *For some hypothesis class $\mathcal{H}$ and subset $V \subset \mathcal{H}$ where for each $h \in \mathcal{H}, h : \mathcal{X} \to \{0, 1\}$, the* region of disagreement *is defined as*

$$DIS(V) = \{x \in \mathcal{X} : \exists\, h, h' \in \mathcal{H} \text{ s.t. } h(x) \neq h'(x)\}$$

which is the set of unlabeled examples x for which there are hypotheses in V that disagree on how to label x.

The CAL[1] algorithm [Cohn et al., 1994] (see Algorithm 3) at time t, in response to Nature revealing $x_t \sim \mathcal{D}_X$, decides to request y_t if and only if $x_t \in DIS(V_{t-1})$ where V_t represents the subset of hypotheses in $\mathcal{H}$ that is consistent with all data requested up to time t. Note that if $x_t \notin DIS(V_{t-1})$ then *all* $h \in V_{t-1}$ agree on its label. Because $h^* \in V_t$ for all t on the assumption that $y_t = h^*(x_t)$, if $x_t \notin DIS(V_{t-1})$ then we can conclude that $y_t = h(x_t)$ for all $h \in V_{t-1}$. Thus, after n unlabeled examples, CAL has the same performance guarantee as the passive learning algorithm that observes all n labels. All that is left to do is bound how many samples CAL takes.

Computational efficiency Note, the check "$x_t \in DIS(V_{t-1})$" can often be computed efficiently. To see how, note that $x_t \in DIS(V_{t-1})$ if and only if there exists an $h \in V_{t-1}$ such that $h(x_t) \neq h^*(x_t)$. Now, $h \in V_{t-1}$ if and only if $h(x_s) = h^*(x_s)$ for all $s \in Z_{t-1}$. Thus, an equivalent check to $x_t \in DIS(V_{t-1})$ is a check if there exist an $h_1 \in \mathcal{H}$ such that $h_1(x_s) = y_s$ for all $(x_s, y_s) \in Z_{t-1} \cup (x_t, 1)$ *and* and an $h_0 \in \mathcal{H}$ such that $h_0(x_s) = y_s$ for all $(x_s, y_s) \in Z_{t-1} \cup (x_t, 0)$. This inspires the algorithm efficient CAL. This idea was first introduced in [Dasgupta et al., 2008]. If $\mathcal{H}$ is the set of linear halfspaces, such a check is a linear program.

Algorithm 1 CAL	**Algorithm 2** Efficient CAL
1: Initialize: $Z_0 = \emptyset$, $V_0 = \mathcal{H}$	1: Initialize: $Z_0 = \emptyset$
2: **for** $t = 1, 2...n$ **do**	2: **for** $t = 1, 2...n$ **do**
3: Nature reveals unlabeled data point x_t	3: Nature reveals unlabeled data point x_t
4: **if** $x_t \in DIS(V_{t-1})$ **then**	4: **if** for $\widehat{y} \in \{0,1\}$ $\exists h_{\widehat{y}} \in \mathcal{H}$: $h_{\widehat{y}}(x_s) = y_s$, $\forall (x_s, y_s) \in Z_{t-1} \cup (x_t, \widehat{y})$ **then**
5: Query y_t, and set $Z_t = Z_{t-1} \cup (x_t, y_t)$	5: Query y_t, and set $Z_t = Z_{t-1} \cup (x_t, y_t)$
6: **else**	6: **else**
7: $Z_t = Z_{t-1}$	7: $Z_t = Z_{t-1}$
8: **end if**	8: **end if**
9: $V_t = \{h \in \mathcal{H} : h(x_i) = y_i \ \forall (x_i, y_i) \in Z_t\}$	9: **end for**
10: **end for**	10: **return** $\arg\min_{h \in \mathcal{H}} \sum_{(x,y) \in Z_t} \mathbf{1}\{h(x) \neq y\}$.
11: **return** any $h \in V_n$	

CAL (and other disagreement-based methods) use a concept called the *disagreement coefficient* for analyzing the label complexity. Define the disagreement (pseudo) metric ρ on $\mathcal{H}$ as $\rho(h, h') := \mathbb{P}_{X \sim \mathcal{D}_X}(h(X) \neq h'(X))$. Let $B(h, r) := \{h' \in \mathcal{H} : \rho(h, h') \leq r\}$ denote the closed ball centered at $h \in \mathcal{H}$ with radius r.

Definition 7. *The* disagreement coefficient *of $h \in \mathcal{H}$ with respect to a hypothesis class $\mathcal{H}$ and distribution $\mathcal{D}_X$ is defined as*

$$\theta_h = \sup_r \frac{\mathbb{P}_{X \sim \mathcal{D}_X}(X \in DIS(B(h, r)))}{r}$$

As we will see, a small disagreement coefficient is a sufficient condition for efficient active learning algorithms. Essentially, it says that as the version space collapses around h^*, there is always sufficient mass between some sub-optimal h and h^* to rule out h in a bounded amount of time.

As an example, consider our example of before with $\mathcal{D}_X$ uniform on $[0, 1]$ and $\mathcal{H}$ as thresholds. Then, interpreting h^* as a number in $[0, 1]$ denoting the threshold location, $DIS(B(h^*, r)) = [h^* - r, h^* + r]$ and so $\mathbb{P}_{X \sim \mathcal{D}_X}(X \in DIS(B(h^*, r))) = 2r$. Therefore, the disagreement coefficient is equal to $\theta_{h^*} = \sup_r \frac{Pr(DIS(B(h^*, r)))}{r} = \frac{2r}{r} = 2$. With the exception of very nice situations (uniform distribution, symmetric geometry, etc.) the disagreement coefficient is often very difficult to calculate. Some

---

[1]Named for its inventors Cohn, Atlas, and Ladner

"nice" classes include i) homogeneous hyperplanes in $\mathbb{R}^d$ with data uniformly distributed on a sphere: $\theta \leq \sqrt{d}$, ii) general hyperplanes in $\mathbb{R}^d$ with the data density bounded below: $\theta = O(d)$, and iii) intervals $[a, b]$ on $\mathbb{R}$: $\theta = \infty$.

Theorem 11. *Let $h^* = \arg\min_{h \in \mathcal{H}} R(h)$ and assume $R(h^*) = 0$. Suppose n iid labeled examples $\{(x_i, y_i)\}_{i=1}^n$ are drawn from $\mathcal{D}$ and $V_n = \{h \in \mathcal{H} : h(x_i) = y_i \ \forall i \in [n]\}$. If we request λ additional labels only when the samples lie in the disagreement region $DIS(V_n)$, where $\lambda = 2\theta_{h^*} \log(|\mathcal{H}|/\delta)$, then, with probability greater than $1 - \delta$ we have $\sup_{h \in V_{n+\lambda}} R(h) \leq \sup_{h \in V_n} \frac{1}{2} R(h)$.*

Proof. The disagreement coefficient allows for a bound that relates the region of disagreement to the true risk of any $h \in V_n$. First, observe that:

$$\frac{\mathbb{P}_{X \sim \mathcal{D}_X}(X \in DIS(V_n))}{\sup_{h \in V_n} R(h)} \leq \frac{\mathbb{P}_{X \sim \mathcal{D}_X}(X \in DIS(B(h^*, \sup_{h \in V_n} R(h))))}{\sup_{h \in V_n} R(h)}$$

$$\leq \theta_{h^*}$$

where the first inequality follows from the fact that in the RHS we replace V_n with a bigger set. To see this, for any $h \in V_n$ we have that $\rho(h, h^*) = \mathbb{P}_{X \sim \mathcal{D}_X}(h(X) \neq h^*(X)) = \mathbb{P}_{X \sim \mathcal{D}_X}(h(X) \neq Y) = R(h) \leq \max_{h \in V_n} R(h)$, which implies $h \in B(h^*, \sup_{h \in V_n} R(h))$. The second inequality follows by the definition of the disagreement coefficient.
By the definition of risk we have:

$$\sup_{h \in V_{n+\lambda}} R(h) = \sup_{h \in V_{n+\lambda}} \mathbb{P}(h(X) \neq Y)$$

$$= \sup_{h \in V_{n+\lambda}} \mathbb{P}(h(X) \neq Y | X \in DIS(V_n)) \mathbb{P}(X \in DIS(V_n))$$

$$\leq \sup_{h \in V_{n+\lambda}} \mathbb{P}(h(X) \neq Y | X \in DIS(V_n)) \theta_{h^*} \sup_{h \in V_n} R(h)$$

where the second equality exploits the fact that $\mathbb{P}(h(X) \neq Y | X \notin DIS(V_n)) = 0$ since $h^*, h \in V_n$ and $Y = h^*(X)$. To bound, $\mathbb{P}(h(X) \neq Y | X \in DIS(V_n))$ note that it only looks at λ points that land in the disagreement region of V_n. This is like a brand new problem, where we can do passive learning only with points that land in $DIS(V_n)$. If we apply Theorem **??** and condition on the new version space $V_{n+\lambda}$, then the risk of any classifier in the new version space if we see only λ samples satisfies:

$$Pr(h(X) \neq Y | X \in DIS(V_n)) \leq \frac{\log(|V_{n+\lambda}|/\delta)}{\lambda} \leq \frac{\log(|\mathcal{H}|/\delta)}{\lambda}$$

Thus, for our specified value of λ, we conclude that $\sup_{h \in V_{n+\lambda}} R(h) \leq \frac{1}{2} \sup_{h \in V_n} R(h)$. $\qquad\square$

Finally, we need to do the previous procedure $\log_2(1/\epsilon)$ times in order to achieve ϵ-error, meaning that the bound holds simultaneously for all epochs. By taking a union bound over $\lambda, 2\lambda, ..., \lceil \log_2(1/\epsilon) \rceil / \lambda$, we have that after $n \geq \lambda \lceil \log_2(1/\epsilon) \rceil$ labels, the true risk of any classifier satisfies $R(h) \leq \epsilon$, and the total number of requested labels is bounded by $2\theta_{h^*} \log(|\mathcal{H}|/\delta) \log(1/\epsilon)$ with probability at least $1 - \delta$. Compare this to passive learning which requires $O(\log(|\mathcal{H}|)/\epsilon)$ labels to reach an ϵ risk.

6.2.3 Splitting index

The CAL algorithm is very mellow in the sense that it will request the label of any example that will remove at least one hypothesis from the current version space. It is intuitive that it may be

more advantageous for the learner to pass on some examples in favor of waiting for future examples to cut off a more substantial chunk of the version space. If we only have access to a sample oracle $x \sim \mathcal{D}$, this naturally leads to the question of what is the fundamental trade off between labeled data and unlabeled data?

A hard instance [Dasgupta, 2005b]. For some $\tau \in (0,1)$ let the distribution over $\mathcal{X}$ be denoted as $(1-\tau)\Gamma + \tau\Gamma'$ where Γ is the uniform distribution on the set $\{x \in \mathbb{R}^d : x_1^2 + x_2^2 = 1, x_3 = 0\}$ and Γ' is the uniform distribution on the set $\{x \in \mathbb{R}^d : x_1^2 + x_2^2 = 1, x_3 = 1\}$. Consider the set of tilted half-spaces that go through the origin so that they bisect Γ', and capture just a fraction ϵ of Γ. Assume $h^* \in \mathcal{H}$. Recall at each time Nature reveals $x \sim (1-\tau)\Gamma + \tau\Gamma'$, assume τ is very tiny. Note that by construction, if we only request labels that land on Γ', CAL would find an ϵ-close hypothesis to h^* using just $\log(1/\epsilon)$ queries. But it would only get a sample from Γ' every $1/\tau$ unlabeled examples. If τ is tiny and one seeks to minimize the amount of unlabeled data, it is better to run CAL on the full stream where the vast amount of data comes from Γ. This will find an ϵ-good classifier after $1/\epsilon$ labeled and unlabeled data. Can we formalize this trade off?

Consider a finite hypothesis space $\mathcal{H}$ and consider any $Q \subset \binom{\mathcal{H}}{2}$ where $(h, h') \in Q$ can be considered an edge connecting any two hypotheses. For any $\widehat{y} \in \{0,1\}$ define $\mathcal{H}_{(x,\widehat{y})} = \{h \in \mathcal{H} : h(x) = \widehat{y}\}$. We say an example x ρ-splits Q if requesting its label reduces the number of edges by at least a fraction $\rho \in (0,1)$:

$$\max\{|Q \cap \mathcal{H}_{(x,0)}|, |Q \cap \mathcal{H}_{(x,1)}|\} \leq (1-\rho)|Q|.$$

We are now ready to introduce the splitting index.

Definition 8. *Fix any subset $S \subset \mathcal{H}$ and $Q \subset \binom{S}{2}$ such that $\mathbb{P}(h(X) \neq h'(X)) \geq \epsilon, \forall(h, h') \in Q$. Then we say S is (ρ, ϵ, τ)-splittable if $\mathbb{P}(X \text{ splits } Q) \geq \tau$.*

Basically, the definition is saying that to reduce the number of pairs of hypotheses that differ by at least ϵ by a fraction at least ρ, requires $1/\tau$ unlabeled data. If $\mathcal{H}$ is finite, and $\mathcal{H}$ is (ρ, ϵ, τ)-splittable, then it is almost immediate that there exists an algorithm that requires $1/(\tau\rho)$ unlabeled data and $1/\rho$ labels to identify an ϵ-good classifier ([Dasgupta, 2005b] suggests one, though it is computationally intractable). What is more important is the reproduced lower bound:

Theorem 12 ([Dasgupta, 2005b]). *Fix any hypothesis space $\mathcal{H}$ and distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$. Suppose that for some $\rho \in (0,1)$, $\epsilon \in (0,1)$ and some $\tau \in (0,1/2)$, the set $S \subset \mathcal{H}$ is not (ρ, ϵ, τ)-splittable. Then any active learning strategy that achieves an accuracy of $\epsilon/2$ on all target hypotheses in S must, with probability at least $3/4$ (taken over the random sampling of data), either draw $\geq 1/\tau$ unlabeled samples, or must request $\geq 1/\rho$ labels.*

The above theorem characterizes the trade off between the streaming sampling oracle from $\mathcal{X}$ and the pool-based setting. In the streaming setting, one may be able to use as few labels as the learner in the pool-based setting, but they may have to wade through an arbitrarily large amount of unlabeled data first.

6.2.4   Lower bounds

The somewhat trivial covering lower bound of Proposition 4 for the separable, pool-based setting can be extended to the separable, streaming setting by replacing finite cardinality classes with ϵ-covers with respect to the underlying distribution [Kulkarni et al., 1993]. Just like in the pool-based setting, this style of lower bound is unlikely to be achieved by an algorithm in all cases because it is

not considering how little information each query may provide. The splitting-index lower bound of Theorem 12 is one of the only results I am aware of that has a corresponding upper bound for a fixed instance of $\mathcal{D}_X$ and $\mathcal{H}$, indicating that it may be fundamental. Hanneke developed a framework in the streaming setting inspired by Hegedus's extended teaching dimension. Essentially, the extended teaching dimension is now a random quantity based on data, and is bounded with high probability. The sample complexity is upper and lower bounded by quantities that match in limiting cases, but is not as strong as the same quantity appearing in both the upper and lower bounds like in the pool-based case. There are *minimax* lower bounds known for the separable setting which considers a worst-case choice of $\mathcal{D}_X$ for each choice of $\mathcal{H}$. As described in the lower bounds section below, these can be misleading.

6.2.5   Open problems

The algorithm in [Dasgupta, 2005b] that achieves the lower bound is computationally intractable. Recent work has demonstrated that if one has access to a sampling distribution over $\mathcal{H}$, one can use rejection sampling to write down an algorithm that can be run, though it may still have unbounded computation [Tosh and Dasgupta, 2017]. A significant advancement would be an algorithm that achieves the same performance but only uses empirical risk oracles, like those used in the efficient version of CAL. It is much more natural to efficiently minimize a loss than define a favorable distribution over hypotheses.

6.3   Agnostic, sampling-oracle setting

In contrast to the separable setting, in the agnostic setting we now make minimal assumptions on how labels related to our hypothesis class $\mathcal{H}$. Specifically, in this setting we assume that when we request the label of some $x \in \mathcal{X}$ we observe a Bernoulli random variable $Y \in \{0,1\}$ with $\mathbb{P}(Y = 1|X = x) = \eta(x)$ where $\eta : \mathcal{X} \to [0,1]$ is arbitrary. Define $R(h) = \mathbb{E}_{X \sim \nu, Y \sim \eta(X)}[\mathbf{1}\{h(X) \neq Y\}]$.

6.3.1   Passive learning

First, let us establish the baseline of a passive learning algorithm that observes the label of every $x \sim \nu$. Let $\{(x_t, y_t)\}_{t=1}^{n}$ be a dataset such that $x_t \sim \nu$ and $y_t \sim \eta(x_t)$. Let $\widehat{R}_n(h) = \frac{1}{n}\sum_{t=1}^{n}\mathbf{1}\{h(x_t) \neq y_t\}$ and $\widehat{h}_n = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$. Note that for any $h \in \mathcal{H}$ we have

$$
\begin{aligned}
\mathbb{E}_{X \sim \nu}[(\mathbf{1}\{h(X) \neq Y\} - \mathbf{1}\{h^*(X) \neq Y\})^2] &= \mathbb{E}_{X \sim \nu}[\mathbf{1}\{h(X) \neq h^*(X)\}] \\
&= \mathbb{E}_{X \sim \nu}[\mathbf{1}\{h(X) = Y, h^*(X) \neq Y\} + \mathbf{1}\{h(X) \neq Y, h^*(X) = Y\}] \\
&\leq R(h) + R(h^*) \\
&\leq 2\max\{R(h) - R(h^*), 2R(h^*)\}
\end{aligned}
$$

By Bernstein's inequality, we have with probability at least $1 - \delta$

$$R(\widehat{h}_n) - R(h^*) \leq \widehat{R}_n(\widehat{h}_n) - \widehat{R}_n(h^*) + \sqrt{\mathbb{E}_{X \sim \nu}[\mathbf{1}\{\widehat{h}_n(X) \neq h^*(X)\}] \frac{2 \log(|\mathcal{H}|/\delta)}{n}} + \frac{\log(|\mathcal{H}|/\delta)}{n}$$

$$\leq \sqrt{\max\{R(\widehat{h}_n) - R(h^*), 2R(h^*)\} \frac{4 \log(|\mathcal{H}|/\delta)}{n}} + \frac{\log(|\mathcal{H}|/\delta)}{n}$$

$$\leq \max\{\frac{8 \log(|\mathcal{H}|/\delta)}{n}, \sqrt{\frac{4R(h^*) \log(|\mathcal{H}|/\delta)}{n}} + \frac{\log(|\mathcal{H}|/\delta)}{n}\}$$

$$\leq \frac{8 \log(|\mathcal{H}|/\delta)}{n} + \sqrt{\frac{4R(h^*) \log(|\mathcal{H}|/\delta)}{n}}$$

where the third inequality follows from the quadratic equation. We summarize our findings in the following theorem.

Theorem 13. *Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ and a finite set of hypotheses $\mathcal{H}$ (i.e., $|\mathcal{H}| < \infty$). Given n IID draws from $\mathcal{D}$, $\{(x_i, y_i)\}_{i=1}^n$, let $\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h)$ be the empirical risk minimizer. Let $h^* \in \arg\min_{h \in \mathcal{H}} R(h)$ be the true risk minimizer. For any $\epsilon, \delta \in (0, 1)$, we have $\mathbb{P}(R(\hat{h}_n) - R(h^*) > \epsilon) \leq \delta$ whenever $n \geq \left(\frac{R(h^*)}{\epsilon^2} + \frac{1}{\epsilon}\right) 8 \log(|\mathcal{H}|/\delta)$. In other words, for any $\delta \in (0, 1)$, with probability $1 - \delta$, we have $R(\hat{h}_n) \leq R(h^*) + \sqrt{\frac{4R(h^*) \log(|\mathcal{H}|/\delta)}{n}} + \frac{8 \log(|\mathcal{H}|/\delta)}{n}$.*

6.3.2 Robust CAL

Our strategy will be to use disagreement-based learning. This algorithm and its analysis is based on [Dasgupta et al., 2008]. While the argument here is somewhat coherent, I encourage you to go and read the paper because like most works Sanjoy Dasgupta is involved in, it is an exemplar of technical writing.

Algorithm 3 Robust CAL

1: Initialize: $Z_0 = \emptyset$, $V_1 = \mathcal{H}$
2: **for** $t = 1, 2 ... n$ **do**
3:     Nature reveals unlabeled data point x_t
4:     **if** $x_t \in DIS(V_t)$ **then**
5:       Query y_t, and set $Z_t = Z_{t-1} \cup (x_t, y_t)$
6:     **else**
7:       $Z_t = Z_{t-1}$
8:     **end if**
9:     **if** $\log_2(t) \in \mathbb{N}$ **then**
10:       $\widehat{L}_t(h) = \frac{1}{t} \sum_{(x_s, y_s) \in Z_t} \mathbf{1}\{h(x_s) \neq y_s\}$ for all $h \in \mathcal{H}$, $\widehat{h}_t = \arg\min_{h \in V_t} \widehat{L}_t(h)$
11:       $\widehat{\rho}_t(h, h') = \frac{1}{t} \sum_{s=1}^{t} \mathbf{1}\{h(x_s) \neq h'(x_s)\}$, $\beta_t = \sqrt{\frac{2 \log(2 \log_2(t)^2 |\mathcal{H}|^2/\delta)}{t}}$
12:       $V_{t+1} = \{h \in V_t : \widehat{L}_t(h) - \widehat{L}_t(\widehat{h}_t) \leq \beta_t \sqrt{\widehat{\rho}_t(h, \widehat{h}_t)} + \beta_t^2/2\}$
13:     **else**
14:       $V_{t+1} = V_t$, $\beta_{t+1} = \beta_t$, $\widehat{h}_{t+1} = \widehat{h}_t$
15:     **end if**
16: **end for**
17: **return** any $h \in V$

Define the empirical risk of all the data up to time t as $\widehat{R}_t(h) = \frac{1}{t}\sum_{s=1}^{t}\mathbf{1}\{h(x_s) \neq y_s\}$. The first thing to note about this algorithm is that for any $h, h' \in V_t$ we have

$$\widehat{R}_t(h) - \widehat{R}_t(h') = \frac{1}{t}\sum_{s=1}^{t}(\mathbf{1}\{h(x_s) \neq y_s\} - \mathbf{1}\{h'(x_s) \neq y_s\})$$

$$= \underbrace{\frac{1}{t}\sum_{(x_s,y_s)\in Z_t}(\mathbf{1}\{h(x_s) \neq y_s\} - \mathbf{1}\{h'(x_s) \neq y_s\})}_{=L_t(h)-L_t(h')} + \underbrace{\frac{1}{t}\sum_{(x_s,y_s)\notin Z_t}(\mathbf{1}\{h(x_s) \neq y_s\} - \mathbf{1}\{h'(x_s) \neq y_s\})}_{=0}$$

because if for some $s \leq t$ we have $(x_s, y_s) \notin Z_t$, then $x_s \notin DIS(V_s)$, and thus all $h, h' \in V_s$ satisfy $(\mathbf{1}\{h(x_s) \neq y_s\} - \mathbf{1}\{h'(x_s) \neq y_s\}) = 0$. Since $V_t \subset V_{t-1} \subset \ldots V_s$ we also have that all $h, h' \in V_t$ have this difference equal to 0. However, it is important to note that while $\widehat{R}_t(h) - \widehat{R}_t(h') = \widehat{L}_t(h) - \widehat{L}_t(h')$, we can only guarantee that $\widehat{L}_t(h) \leq \widehat{R}_t(h)$.

Define the events

$$\mathcal{E}_0 := \bigcap_{\ell=1}^{\infty}\bigcap_{h\in\mathcal{H}}\{|\widehat{R}_{2^\ell}(h) - R(h)| \leq \beta_{2^\ell}\sqrt{R(h)} + \beta_{2^\ell}^2/2\},$$

$$\mathcal{E}_1 := \bigcap_{\ell=1}^{\infty}\bigcap_{h\in\mathcal{H}}\{\widehat{R}_{2^\ell}(h) - \widehat{R}_{2^\ell}(h') \leq R(h) - R(h') + \beta_{2^\ell}\sqrt{\widehat{\rho}_{2^\ell}(h,h')} + \beta_{2^\ell}^2/2\}.$$

To show $\mathbb{P}(\mathcal{E}_0) \geq 1 - \delta/2$ one simply applies Bernstein's inequality and a union bound, exploiting the fact that $\mathrm{Variance}(\mathbf{1}\{h(X) \neq Y\}) \leq \mathbb{E}[(\mathbf{1}\{h(X) \neq Y\})^2] = R(h)$ (note: the chosen value of β_t is chosen to be larger than necessary for convenience). To show $\mathbb{P}(\mathcal{E}_1) \geq 1 - \delta/2$ we apply empirical Bernstein's inequality [Maurer and Pontil, 2009] similar to as in Section 6.3.1 plus a union bound. This implies that with probability at least $1 - \delta/2$ for any t that is a power of 2 and any $h, h' \in \mathcal{H}$ we have

$$\widehat{R}_t(h) - \widehat{R}_t(h') \leq R(h) - R(h') + \sqrt{\frac{2\widehat{\rho}_t(h,h)\log(2\log_2(t)^2|\mathcal{H}|/\delta)}{t}} + \frac{\log(2\log_2(t)^2|\mathcal{H}|/\delta)}{t}$$

$$= R(h) - R(h') + \beta_t\sqrt{\widehat{\rho}_t(h,h')} + \beta_t^2/2$$

In particular, using the observations above, this implies that

$$\widehat{L}_t(h^*) - \widehat{L}_t(\widehat{h}_t) = \widehat{R}_t(h^*) - \widehat{R}_t(\widehat{h}_t)$$

$$\leq R(h^*) - R(\widehat{h}_t) + \beta_t\sqrt{\widehat{\rho}_t(h^*, \widehat{h}_t)} + \beta_t^2/2$$

$$\leq \beta_t\sqrt{\widehat{\rho}_t(h^*, \widehat{h}_t)} + \beta_t^2/2$$

which implies $h^* \in V_t$ for all t. This proves correctness, we next prove the sample complexity.

Sample complexity analysis For any $h \in V_{t+1}$ we have

$$
\widehat{R}_t(h) - \widehat{R}_t(h^*) = \widehat{L}_t(h) - \widehat{L}_t(\widehat{h}_t)
$$

$$
\leq \sqrt{\widehat{\rho}_t(h, \widehat{h}_t)} \beta_t + \beta_t^2/2
$$

$$
= \sqrt{\widehat{L}_t(h)} \beta_t + \sqrt{\widehat{L}_t(\widehat{h}_t)} \beta_t + \beta_t^2/2
$$

$$
\leq \sqrt{\widehat{L}_t(h)} \beta_t + \sqrt{\widehat{L}_t(h^*)} \beta_t + \beta_t^2/2
$$

$$
\leq \sqrt{\widehat{R}_t(h)} \beta_t + \sqrt{\widehat{R}_t(h^*)} \beta_t + \beta_t^2/2.
$$

where we have exploited that for any $h, h' \in V_t$

$$
\widehat{\rho}_t(h, h') = \frac{1}{t} \sum_{s=1}^{t} \mathbf{1}\{h(x_s) \neq h'(x_s)\}
$$

$$
= \frac{1}{t} \sum_{(x_s, y_s) \in Z_t} \mathbf{1}\{h(x_s) \neq h'(x_s)\}
$$

$$
\leq \frac{1}{t} \sum_{s=1}^{t} \mathbf{1}\{h(x_s) \neq y_s\} + \mathbf{1}\{h'(x_s) \neq y_s\} \leq \widehat{L}_t(h) + \widehat{L}_t(h')
$$

and the facts that (i) $\widehat{R}_t(h) - \widehat{R}_t(h') = \widehat{L}_t(h) - \widehat{L}_t(h')$ for all $h, h' \in V_t$, (ii) $\widehat{L}_t(\widehat{h}_t) \leq \widehat{L}_t(h^*)$, and (iii) $\widehat{L}_t(h) \leq \widehat{R}_t(h)$ for all $h \in V_t$. We now use the crazy useful fact that $A \leq B + C\sqrt{A} \implies A \leq B + C^2 + C\sqrt{B}$ for $A, B, C \geq 0$ we have

$$
R(h) - \beta_t \sqrt{R(h)} - \beta_t^2/2 \leq \widehat{R}_t(h)
$$

$$
\leq \widehat{R}_t(h^*) + \sqrt{\widehat{R}_t(h)} \beta_t + \sqrt{\widehat{R}_t(h^*)} \beta_t + \beta_t^2/2
$$

$$
\leq \widehat{R}_t(h^*) + \sqrt{\widehat{R}_t(h^*)} \beta_t + \beta_t^2/2 + \beta_t^2 + \beta_t \sqrt{\widehat{R}_t(h^*) + \sqrt{\widehat{R}_t(h^*)} \beta_t + \beta_t^2/2}
$$

$$
\leq R(h^*) + \beta_t \sqrt{R(h^*)} + \beta_t^2 + \beta_t \sqrt{R(h^*) + \beta_t \sqrt{R(h^*)} + \beta_t^2} + (3/2)\beta_t^2
$$

$$
+ \beta_t \sqrt{R(h^*) + \beta_t \sqrt{R(h^*)} + \beta_t^2 + \beta_t \sqrt{R(h^*) + \beta_t \sqrt{R(h^*)} + \beta_t^2} + \beta_t^2/2}
$$

$$
\leq R(h^*) + c_1' \beta_t \sqrt{R(h^*)} + c_2' \beta_t^2
$$

for some constants c_1, c_2'. Applying the crazy useful fact again to $A = R(h)$ we get that

$$
R(h) \leq c_0 R(h^*) + c_1 \beta_t \sqrt{R(h^*)} + c_2 \beta_t^2
$$

for some constants c_0, c_1, c_2. Noting that $\rho_\nu(h, h^*) = \mathbb{E}_{X \sim \nu}[\mathbf{1}\{h(x) \neq h^*(X)\}] \leq R(h) + R(h^*)$ we have that

$$
\mathbb{P}(x_{t+1} \in DIS(V_{t+1})) = \mathbb{P}(\exists h, h' \in V_{t+1} : h(x_t) \neq h'(x_t))
$$

$$
= \mathbb{P}(\exists h \in V_{t+1} : h(x_t) \neq h^*(x_t))
$$

$$
\leq \mathbb{P}(\exists h \in \mathcal{H} : h(x_t) \neq h^*(x_t), \ R(h) \leq c_0 R(h^*) + c_1 \beta_t \sqrt{R(h^*)} + c_2 \beta_t^2)
$$

$$
\leq \mathbb{P}(\exists h \in \mathcal{H} : h(x_t) \neq h^*(x_t), \ \rho(h, h^*) \leq (1 + c_0) R(h^*) + c_1 \beta_t \sqrt{R(h^*)} + c_2 \beta_t^2)
$$

$$
\leq \theta^*(R(h^*) + c_1 \beta_t \sqrt{R(h^*)} + c_2 \beta_t^2)\big((1 + c_0) R(h^*) + c_1 \beta_t \sqrt{R(h^*)} + c_2 \beta_t^2\big)
$$

where we have used the definition of the disagreement coefficient. We summarize our findings in the following theorem.

Theorem 14. *Fix $\epsilon > 0$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$ if $t \gtrsim (\frac{R(h^*)}{\epsilon^2} + \frac{1}{\epsilon}) \log(|\mathcal{H}| \log(1/\epsilon)/\delta)$ then $R(\widehat{h}_t) - R(h^*) \le \epsilon$ and*

$$\# \text{ labels requested} \lesssim \theta^*(\nu + \epsilon)\Big(R(h^*)t + \sqrt{t \log(|\mathcal{H}| \log(t)/\delta)} + \log(t) \log(|\mathcal{H}| \log(t)/\delta)\Big)$$

$$\lesssim \theta^*(\nu + \epsilon)\Big(\frac{R(h^*)^2}{\epsilon^2} + \log(\frac{1}{\epsilon})\Big) \log(|\mathcal{H}| \log(1/\epsilon)/\delta)$$

Note that we always have $\theta^*(\nu + \epsilon) \le 1/(\nu + \epsilon)$. If $\theta^*(\nu + \epsilon)$ is a constant independent of ϵ then this theorem says that it requires just $O(\log(1/\epsilon))$ labels to obtain an ϵ-good classifier when $\epsilon \ge R(h^*)$. On the other hand, if $\epsilon \ll R(h^*)$ then one requires $O(R(h^*)^2/\epsilon^2)$ labels. Note, this is not much better than the passive guarantee of $O(R(h^*)/\epsilon^2)$! While this ladder observation is somewhat discouraging, note that under favorable noise distributions such as Massart noise (separable, but each label is flipped with a constant probability bounded away from $1/2$) or Tsybakov noise then the number of labels required by active and passive can be quite large.

6.3.3 Computationally efficient algorithms

Researchers approach computationally efficiency in different ways. The first approach simply assumes access to an empirical risk oracle that returns a classifier from $\mathcal{H}$ that minimizes the (weighted) empirical 0/1-loss on any set of examples. The second replaces the objective of 0/1-loss with a convex loss and performs active learning to minimizes this convex loss (c.f., [Beygelzimer et al., 2009]). The third approach uses a convex surrogate loss in place of a empirical 0/1-loss but still remains the objective of 0/1-loss (c.f., [Hanneke et al., 2014]). This third approach relies on the function class being sufficiently rich and the loss function being *classification calibrated* which allows one to relate the quality of the solution of the convex loss to the 0/1-loss [Bartlett et al., 2006]. [Hanneke et al., 2014] argues against the second approach as there are examples where the second approach requires exponentially more labels than the first or third. The third approach requires unverifiable assumptions, such as the Bayes classifier being in $\mathcal{H}$. The first approach rests on naive hope that the minimizer of the convex loss will be close to the minimizer of 0/1-loss without justification. But it is also the most practical solution, and tends to work well if not too much is demanded from the oracle (e.g., oracle is constrained in some way).

Just like with CAL, we can derive an exactly equivalent version of Robust CAL that does not explicitly maintain a version space. However, this reduction requires a minimization oracle that minimizes empirical risk on one set of examples subject to making no mistakes on a different set of examples. In practice constructing such an oracle is very awkward so we do not discuss it further, see the original paper [Dasgupta et al., 2008] for details. [Beygelzimer et al., 2010] requires far less from its procedure by requiring the oracle to minimize a weighted empirical subject to classifying just a single data point as a specific label, so a single constraint. There are natural hypothesis classes like halfspaces and trees where this is easy satisfy, and in any case, one can simply take the Lagrange multiplier approach and place a large weight on this one loss to satisfy a constraint. Ideally, we would like our method to rely on performing empirical risk minimization on a set of examples, without any constraints. [Huang et al., 2015] is one of the first algorithms to achieve this feat.

6.3.4   Minimax lower bounds

To the best of my knowledge, only *minimax* lower bounds are known for active learning outside of a few specific settings [Kääriäinen, 2006, Castro and Nowak, 2008, Raginsky and Rakhlin, 2011, Hanneke and Yang, 2015]. An instance is defined as a joint probability distribution $\mathcal{D}$ defined over $\mathcal{X} \times \{0, 1\}$. That is, if we interpreted $\mathcal{D}$ as a density then $\mathcal{D}(x, y) = \nu(x)(\eta(x)\mathbf{1}\{y = 1\} + (1 - \eta(x))\mathbf{1}\{y = 0\})$. Fix a collection of instances $\mathbb{D}$. For any $\epsilon \in (0, \epsilon_0)$ and $\delta \in (0, \delta_0)$, let $(S)(\mathcal{D}, \mathcal{H}, \epsilon, \delta, \mathcal{A})$ denote the number of queries taken by algorithm $\mathcal{A}$ on instance $\mathcal{D}$ with hypothesis class $\mathcal{D}$ to output a ϵ-good classifier with probability at least $1 - \delta$. Minimax lower bounds are stated as follows: for any $\epsilon \in (0, \epsilon_0)$ and $\delta \in (0, \delta_0)$ we have that $\min_{\mathcal{A}} \max_{\mathcal{D} \in \mathbb{D}} \mathfrak{S}(\mathcal{D}, \mathcal{H}, \epsilon, \delta, \mathcal{A}) \geq \Lambda(\mathbb{D}, \mathcal{H}, \epsilon, \delta)$. We say an algorithm $\mathcal{A}$ is minimax optimal if $\mathfrak{S}(\mathcal{D}, \mathcal{H}, \epsilon, \delta, \mathcal{A}) \lesssim \Lambda(\mathbb{D}, \mathcal{H}, \epsilon, \delta)$, but typically we will allow the upper and lower bounds to differ by small amounts. An excellent survey of known minimax lower bounds is available in [Hanneke and Yang, 2015] for a variety of classes $\mathbb{D}$ with assumed access to a sampling oracle with no cost attributed to sampling unbounded amounts of unlabeled data. Some important classes of $\mathbb{D}$ include:

- **Realizable/separable**: There exists an $h^* \in \mathcal{H}$ such that $h^*(x) = 2\eta(x) - 1 \in \{0, 1\}$ for all $x \in \mathcal{X}$. ν is arbitrary.

- **Massart noise**: There exists an $h^* \in \mathcal{H}$ such that $h^*(x) = \text{sign}(2\eta(x) - 1)$, and there exists a constant $\alpha > 0$ such that $|\eta(x) - 1/2| \geq \alpha$. ν is arbitrary.

- **Tsybakov noise**: There exists an $h^* \in \mathcal{H}$ such that $h^*(x) = \text{sign}(2\eta(x) - 1)$, and there exists constants $a \geq 1$ and $\alpha \in (0, 1)$ such that for all $\gamma > 0$ we have $\mathbb{P}_{X \sim \nu}(|\eta(X) - 1/2| \leq \gamma) \leq a'\gamma^{a/(1-a)}$ where $a' = (1 - \alpha)(2\alpha)^{\alpha/(1-\alpha)}a^{1/(1-\alpha)}$.

- **Agnostic noise**: For $\alpha \geq 0$ there exists an $h^* \in \mathcal{H}$ such that $R(h^*) \leq \alpha$. ν is arbitrary.

Note that each of the above classes is a subset of the class that follows it. Algorithms are typically designed for either the realizable setting or agnostic setting, and then only algorithms for the ladder are analyzed under the easier settings.

Minimax lower bounds are sometimes **very weak** and can even be misleading. For example, consider the "hard instance" $\mathcal{D}$ of Section 6.2.3. The splitting index algorithm can identify an ϵ-good classifier with just $\log(1/\epsilon)$ queries using binary search on Γ'. On the other hand, CAL will essentially only sample from Γ if τ is very small, and require $1/\epsilon$ labels to obtain an ϵ-good classifier. Thus, the splitting algorithm requires exponentially fewer samples than CAL on this instance, but nevertheless, CAL is nearly minimax optimal for the realizable case [Hanneke and Yang, 2015]. This is because the hard instance is a particular choice of ν, not the worst-case distribution ν.

Ideally we would want instance-dependent lower bounds. That is, how many labels does an algorithm require for the *particular* instance $\mathcal{D}$ you care about? We will expand on this discussion and describe what is known for this case in the next section.

6.4   Agnostic, pool-based setting

We are again in the agnostic setting and therefore, when we request the label of some $x \in \mathcal{X}$ we observe a Bernoulli random variable $Y \in \{0, 1\}$ with $\mathbb{P}(Y = 1 | X = x) = \eta(x)$ where $\eta : \mathcal{X} \to [0, 1]$ is arbitrary. In this pool-based setting we are going to assume that $\mathcal{X}$ is finite with $|\mathcal{X}| = n$ and that there exists a *known* probability density ν defined over $\mathcal{X}$ that we wish to evaluate risk with

respect to. Define the risk of any $h \in \mathcal{H}$ as $R(h) := \mathbb{E}_{X \sim \nu, Y \sim \eta(X)}[\mathbf{1}\{Y \neq h(X)\}]$. For any $\epsilon > 0$ and $\delta \in (0,1)$ we want an algorithm that identifies an $\widehat{h} \in \mathcal{H}$ such that $R(\widehat{h}) - \min_{h \in \mathcal{H}} R(h) \leq \epsilon$ with probability at least $1 - \delta$ using as few total requested labels as possible. While we could apply a disagreement-based algorithm from the streaming setting to accomplish this, instead we will reduce the problem to an instance of linear bandits. We will see that this algorithm obtains a sample complexity that is provably superior to disagreement-based learning, sometimes requiring exponentially fewer labels.

6.4.1 Reduction to linear bandits

Note that

$$R(h) := \mathbb{E}_{X \sim \nu, Y \sim \eta(X)}[\mathbf{1}\{Y \neq h(X)\}] = \sum_{x \in \mathcal{X}} \nu(x)(\eta(x)\mathbf{1}\{h(x) \neq 1\} + (1 - \eta(x))\mathbf{1}\{h(x) \neq 0\})$$

$$= \sum_{x \in \mathcal{X}} \nu(x)\eta(x) + \sum_{x \in \mathcal{X}} \nu(x)(1 - 2\eta(x))h(x).$$

Binary classification is intimately related with transductive linear bandits as we now show. Define $\theta^* = [(2\eta(x) - 1)]_{x \in \mathcal{X}} \in \mathbb{R}^n$ and for each $h \in \mathcal{H}$ we define a vector $z_h = [\nu(x)h(x)]_{x \in \mathcal{X}} \in \mathbb{R}^n$ and set $\mathcal{Z} = \{z_h : h \in \mathcal{H}\}$. When requesting the label of example $x \in \mathcal{X}$, we observe $Y \in \{0,1\}$ and convert this into a "pull" of arm $\mathbf{e}_x \in \{0,1\}^n$ by feeding the bandit algorithm $2Y - 1$ so that $\mathbb{E}[2Y - 1 | X = x] = 2\eta(x) - 1 = \langle \mathbf{e}_x, \theta* \rangle$. Thus, identifying a $\widehat{h}$ such that $R(\widehat{h}) - \min_{h \in \mathcal{H}} R(h) \leq \epsilon$ is equivalent to identifying a $\widehat{z} \in \mathcal{Z}$ such that $\max_{z \in \mathcal{Z}} \langle z, \theta^* \rangle - \langle \widehat{z}, \theta^* \rangle \leq \epsilon$. We can now apply the pure-exploration algorithm for linear bandits! Note that in that algorithm, at each stage we solved $\arg\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{z, z' \in \mathcal{Z}_\ell} \|z - z'\|_{A(\lambda)^{-1}}^2$. We observe that

$$\|z - z'\|_{A(\lambda)^{-1}}^2 = \sum_{x \in \mathcal{X}} \frac{|h(x) - h'(x)|^2 \nu(x)^2}{\lambda(x)} = \sum_{x \in \mathcal{X}} \nu(x) \frac{\mathbf{1}\{h(x) \neq h'(x)\}}{\lambda(x)/\nu(x)} = \mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h'(X)\}}{\lambda(X)/\nu(X)}\right].$$

It appears we can run the linear bandit algorithm precisely as before. But there is a problem: that algorithm used the least squares estimator for θ^*, which assumes the number of samples exceeds that dimension, which is $|\mathcal{X}| = n$ here–this is equivalent to labeling every example, trivial. Fortunately, we can solve this problem by using a regularized estimator.

6.4.2 Regularized empirical risk minimization

First, we need a deviation result. Define $\rho_\lambda(h, h') = \mathbb{E}_{X \sim \nu}[\frac{\mathbf{1}\{h(X) \neq h'(X)\}}{\lambda(X)/\nu(X)}]$.

Lemma 14. *Let λ and ν be two probability densities defined over $\mathcal{X}$ with $\lambda \ll \nu^2$. Consider a dataset $\{(x_t, y_t, w_t)\}_{t=1}^\tau$ where $x_t \sim \lambda$, $y_t \sim Bernoulli(\eta(x_t))$, and $w_t = \frac{\lambda(x_t)}{\nu(x_t)}$. For any $\gamma > 0$ define*

$$\widetilde{R}^\gamma(h) = \frac{1}{\tau} \sum_{t=1}^\tau \frac{1}{w_t + \gamma} \mathbf{1}\{h(x_t) \neq y_t\}.$$

Then for any $h, h' \in \mathcal{H}$ we have with probability at least $1 - \delta$

$$\widetilde{R}^\gamma(h) - \widetilde{R}^\gamma(h') \leq R(h) - R(h') + \gamma \rho_\lambda(h, h') + \sqrt{\rho_\lambda(h, h') \frac{2 \log(1/\delta)}{\tau}} + \frac{\log(1/\delta)}{\tau(\gamma + \min_{x \in \mathcal{X}} \lambda(x)/\nu(x))}.$$

---

[2]We say density p dominates q, or $p \gg q$, if support$(q) \subseteq$ support(p)

Proof. Since γ is fixed, to simplify notation in the remainder of this proof let $\widetilde{R}(h) = \widetilde{R}^\gamma(h)$. First note that for any $h, h' \in \mathcal{H}_\ell$ we have $\mathbb{E}[\frac{1}{\tau}\sum_{t=1}^\tau \frac{1}{w_t}(\mathbf{1}\{h(x_t) \neq y_t\} - \mathbf{1}\{h'(x_t) \neq y_t\})] = R(h) - R(h')$. Thus,

$$
\begin{aligned}
\mathbb{E}[\widetilde{R}(h) - \widetilde{R}(h')] - [R(h) - R(h')] &= \mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^\tau (\frac{1}{w_t+\gamma} - \frac{1}{w_t})(\mathbf{1}\{h(x_t) \neq y_t\} - \mathbf{1}\{h'(x_t) \neq y_t\})\right] \\
&= \mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^\tau \frac{-\gamma}{w_t(w_t+\gamma)}(\mathbf{1}\{h(x_t) \neq y_t\} - \mathbf{1}\{h'(x_t) \neq y_t\})\right] \\
&\leq \mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^\tau \frac{\gamma}{w_t(w_t+\gamma)}\mathbf{1}\{h(x_t) \neq h'(x_t)\}\right] \\
&= \gamma\mathbb{E}_{X\sim\nu}\left[\frac{\mathbf{1}\{h(X) \neq h'(X)\}}{\lambda(X)/\nu(X) + \gamma}\right] \\
&\leq \gamma\mathbb{E}_{X\sim\nu}\left[\frac{\mathbf{1}\{h(X) \neq h'(X)\}}{\lambda(X)/\nu(X)}\right] = \gamma\rho_\lambda(h, h')
\end{aligned}
$$

If $\alpha_t := \frac{1}{w_t+\gamma}(\mathbf{1}\{h(x_t) \neq y_t\} - \mathbf{1}\{h'(x_t) \neq y_t\})$ then $\widetilde{R}(h) - \widetilde{R}(h') = \frac{1}{\tau}\sum_{t=1}^\tau \alpha_t$ where each α_t is IID. Note that $\alpha_t \leq 1/(\gamma + \min_{x\in\mathcal{X}} \lambda(x)/\nu(x))$ and

$$
\begin{aligned}
\mathbb{E}[\alpha_t^2] &\leq \mathbb{E}[\left(\frac{1}{w_t+\gamma}(\mathbf{1}\{h(x_t) \neq y_t\} - \mathbf{1}\{h'(x_t) \neq y_t\})\right)^2] \\
&\leq \mathbb{E}[\left(\frac{1}{w_t+\gamma}\right)^2 \mathbf{1}\{h(x_t) \neq h'(x_t)\}] \\
&= \mathbb{E}_{X\sim\nu}\left[\frac{\lambda(X)/\nu(X)}{(\lambda(X)/\nu(X) + \gamma)^2}\mathbf{1}\{h(X) \neq h'(X)\}\right] \\
&= \mathbb{E}_{X\sim\nu}\left[\frac{\mathbf{1}\{h(X) \neq h'(X)\}}{\lambda(X)/\nu(X)}\right] = \rho_\lambda(h, h').
\end{aligned}
$$

By Bernstein's inequality we have with probability at least $1 - \delta$

$$
\begin{aligned}
\widetilde{R}(h) - \widetilde{R}(h') &\leq \mathbb{E}[\widetilde{R}(h) - \widetilde{R}(h')] + \sqrt{\rho_\lambda(h, h')\frac{2\log(1/\delta)}{\tau}} + \frac{\log(1/\delta)}{\gamma\tau} \\
&\leq R(h) - R(h') + \gamma\rho_\lambda(h, h') + \sqrt{\rho_\lambda(h, h')\frac{2\log(1/\delta)}{\tau}} + \frac{\log(1/\delta)}{\gamma\tau}.
\end{aligned}
$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Note that the regularized estimator $\widetilde{R}^\gamma(h)$ is only an unbiased estimator of $R(h)$ when $\gamma = 0$. Also, because $a + \sqrt{2ab} + b \leq (\sqrt{a} + \sqrt{b})^2 \leq 2(a + b)$ we always have that for each $h, h' \in \mathcal{H}$

$$
\widetilde{R}^\gamma(h) - \widetilde{R}^\gamma(h') \leq R(h) - R(h') + 2\gamma\rho_\lambda(h, h') + \frac{2\log(1/\delta)}{\tau\gamma}
$$

with probability $1 - \delta$ which may be more convenient.

Algorithm 4 An efficient estimator with importance-sampled data

---

1: **Input:** $\gamma_0 > 0$, $\epsilon > 0$, $\{(x_t, y_t, w_t)\}_{t=1}^\tau$ s.t. $w_t = \lambda(x_t)/\nu(x_t)$, $\mathcal{H}$, $\delta \in (0,1)$
2: Pick $h_0 \in \mathcal{H}$ arbitrarily, set $\Gamma = \{\gamma' : 1/\epsilon > \gamma' > \gamma_0 : \log_2(\gamma) \in \mathbb{Z}\}$
3: **for** $k = 0, 1, 2, \ldots$ **do**
4:    For every $\gamma \in \Gamma$ let $h_{k+1}^\gamma = \arg\min_{h \in \mathcal{H}} f(h, h_k; \gamma)$ where

$$f(h, h'; \gamma) = \widetilde{R}^\gamma(h) - \widetilde{R}^\gamma(h') + 2\gamma\rho_\lambda(h, h') + \frac{2\log(|\mathcal{H}|^2 \log_2(8/\gamma_0\epsilon)/\delta)}{\tau\gamma}$$

5:    **if** $\min_{\gamma \in \Gamma} f(h_{k+1}^\gamma, h_k; \gamma) \geq -\epsilon$ **then**
6:       **Terminate** and **output** h_k
7:    **else**
8:       Set $h_{k+1} = h_{k+1}^{\gamma_{k+1}}$ where $\gamma_{k+1} = \arg\min_{\gamma \in \Gamma} f(h_{k+1}^\gamma, h_k; \gamma)$
9:    **end if**
10: **end for**

Lemma 15. *Consider the setting of Lemma 14. If one runs Algorithm 4 with $\gamma_0 > 0$ and $\epsilon > 0$ then after $k \leq 1/\gamma_0\epsilon$ iterations the procedure returns an $h_k \in \mathcal{H}$ such that with probability at least $1 - \delta$*

$$R(h_k) - R(h^*) \leq \epsilon + \sqrt{\frac{\rho_\lambda(h^*, h_k)\ 256\log(|\mathcal{H}|^2 \log_2(8/\gamma_0\epsilon)/\delta)}{\tau}}.$$

Moreover, without loss of generality one can take $\gamma_0 = 1/\tau$.

Proof. Fix a finite subset $\Gamma \subset \{2^k : k \in \mathbb{Z}\}$ that will be determined later. By Lemma 14 we have with probability at least $1 - \delta$ that

$$|\widetilde{R}^\gamma(h) - \widetilde{R}^\gamma(h') - R(h) + R(h')| \leq 2\gamma\rho_\lambda(h, h') + \frac{2\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau\gamma}$$

for all $\gamma \in \Gamma$ and $h, h' \in \mathcal{H}$.

Note that since $R(h) \leq 1$ and we want a non-trivial guarantee, without loss of generality we can take $\gamma_0 \geq \frac{1}{\tau}$. This means the number of iterations before the stopping criteria is met is at most $1/\gamma_0\epsilon$ since before the stopping criteria is met, the objective is reduced by at least ϵ each iteration, and $0 \leq \widetilde{R}^\gamma(h) \leq \tau$. We also would never need $\gamma \geq 1/\epsilon$. Thus, it suffices to take $\Gamma = \{2^k : -\lceil\log_2(1/\gamma_0)\rceil k \leq \lceil\log_2(1/\epsilon)\rceil\}$ which means $|\Gamma| \leq \log_2(8/\gamma_0\epsilon)$.

To see the performance guarantee, at the final iterate,

$$
\begin{aligned}
-\epsilon &\leq \widetilde{R}^{\gamma_{k+1}}(h_{k+1}) - \widetilde{R}^{\gamma_{k+1}}(h_k) + 2\gamma_{k+1}\rho_\lambda(h_{k+1}, h_k) + \frac{2\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau\gamma_{k+1}} \\
&= \min_{\gamma\in\Gamma} \min_{h\in\mathcal{H}} \widetilde{R}^\gamma(h) - \widetilde{R}^\gamma(h_k) + 2\gamma\rho_\lambda(h, h_k) + \frac{2\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau\gamma} \\
&\leq \min_{\gamma\in\Gamma} \min_{h\in\mathcal{H}} R(h) - R(h_k) + 4\gamma\rho_\lambda(h, h_k) + \frac{4\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau\gamma} \\
&\leq \min_{\gamma\in\Gamma} R(h^*) - R(h_k) + 4\gamma\rho_\lambda(h^*, h_k) + \frac{4\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau\gamma} \\
&\leq \min_{\gamma\geq 0} R(h^*) - R(h_k) + 8\gamma\rho_\lambda(h^*, h_k) + \frac{8\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau\gamma} \\
&= R(h^*) - R(h_k) + \sqrt{\frac{\rho_\lambda(h^*, h_k)\ 256\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau}}.
\end{aligned}
$$

Thus, if we output h_k then $R(h_k) - R(h^*) \leq \epsilon + \sqrt{\frac{\rho_\lambda(h^*,h_k)\ 256\log(|\mathcal{H}|^2|\Gamma|/\delta)}{\tau}}$. $\square$

6.4.3   A Version-space Elimination Algorithm

Consider Algorithm 5. It leverages the estimator of the previous section.

Algorithm 5 An Elimination-style algorithm for Binary Classification
---
1: **Input**: Policy set $\mathcal{H}$ such that $h : \mathcal{X} \to \{0,1\}$ for all $h \in \mathcal{H}$, confidence level $\delta \in (0,1)$.
2: Let $\widehat{\mathcal{H}}_1 \leftarrow \mathcal{H}$
3: **for** $\ell = 1, 2, \dots$ **do**
4:     Let $\rho_\ell = \min_{\lambda\in\triangle_\mathcal{X}} \max_{h,h'\in\widehat{\mathcal{H}}_\ell} \mathbb{E}_{X\sim\nu}\left[\frac{\mathbf{1}\{h(X) \neq h'(X)\}}{\lambda(X)/\nu(X)}\right]$ and λ_ℓ be its minimizer
5:     Set $\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = \lceil 12\rho_\ell\epsilon_\ell^{-2}\log(2\ell^2|\mathcal{H}|/\delta)\rceil$, $\gamma_\ell = \sqrt{\frac{\log(2\ell^2|\mathcal{H}|/\delta)}{\tau_\ell\rho_\ell}}$
6:     Draw $x_1, \dots, x_{\tau_\ell} \sim \lambda_\ell$, let $w_t = \lambda(x_t)/\nu(x_t)$, and request their labels to obtain $\{(x_t, y_t, w_t)\}_{t=1}^{\tau_\ell}$

7:     Set $\widetilde{R}_\ell(h) = \frac{1}{\tau_\ell}\sum_{t=1}^{\tau_\ell}\frac{1}{w_t+\gamma_\ell}\mathbf{1}\{h(x_t) \neq y_t\}$ for all $h \in \mathcal{H}_\ell$
8:     $\widehat{\mathcal{H}}_{\ell+1} \leftarrow \widehat{\mathcal{H}}_\ell \setminus \{h \in \widehat{\mathcal{H}}_\ell | \widetilde{R}_\ell(h) - \min_{h'\in\widehat{\mathcal{H}}_\ell}\widetilde{R}_\ell(h') \geq \epsilon_\ell\}$
9:     $\ell \leftarrow \ell + 1$
10: **end for**
11: **Output**:

For some $h^* \in \arg\min_{h\in\mathcal{H}}$ define the event

$$
\mathcal{E} := \bigcap_{\ell=1}^\infty \bigcap_{h\in\mathcal{H}} \{\widetilde{R}_\ell(h^*) - \widetilde{R}_\ell(h) \leq R(h^*) - R(h) \leq \epsilon_\ell\}
$$

Noting that

$$\gamma_\ell \rho_\ell + \sqrt{\frac{2\rho_\ell \log(2\ell^2|\mathcal{H}|/\delta)}{\tau_\ell}} + \frac{\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma_\ell \tau_\ell} = (1+\sqrt{2})\sqrt{\frac{2\rho_\ell \log(2\ell^2|\mathcal{H}|/\delta)}{\tau_\ell}}$$

$$< \sqrt{\frac{12\rho_\ell \log(2\ell^2|\mathcal{H}|/\delta)}{\tau_\ell}} \leq \epsilon_\ell$$

showing that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ is nearly identical to the linear bandit case. Thus, in what follows assume $\mathcal{E}$ holds.

If $h^* \in \widehat{\mathcal{H}}_\ell$ then for *any* $h' \in \mathcal{H}$ we have

$$\widetilde{R}_\ell(h^*) - \widetilde{R}_\ell(h') \leq R(h^*) - R(h') + \epsilon_\ell \leq \epsilon_\ell.$$

Since $h^* \in \widehat{\mathcal{H}}_1$, we have by induction that $h^* \in \widehat{\mathcal{H}}_\ell$ for all ℓ, showing correctness. Now suppose for some $h \in \mathcal{H}$ we have that $R(h) - R(h^*) > 2\epsilon_\ell$. Then

$$\widetilde{R}_\ell(h) - \min_{h' \in \widehat{\mathcal{H}}_\ell} \widetilde{R}_\ell(h') \geq \widetilde{R}_\ell(h) - \widetilde{R}_\ell(h^*)$$

$$\geq R(h) - R(h^*) - \epsilon_\ell$$

$$> \epsilon_\ell$$

which implies $h \notin \widehat{\mathcal{H}}_\ell$. Moreover, it implies that $\max_{h \in \widehat{\mathcal{H}}_{\ell+1}} R(h) - R(h^*) \leq 2\epsilon_\ell$ for all ℓ, or equivalently $\max_{h \in \widehat{\mathcal{H}}_\ell} R(h) - R(h^*) \leq 4\epsilon_\ell$ for all ℓ. Thus, the number of samples taken before some round ℓ satisfies $\max_{h \in \widehat{\mathcal{H}}_\ell} R(h) - R(h^*) \leq \epsilon$ is bounded by

$$\sum_{\ell=1}^{\lceil \log_2(4/\epsilon) \rceil} \tau_\ell = \sum_{\ell=1}^{\lceil \log_2(4/\epsilon) \rceil} \lceil 12\rho_\ell \epsilon_\ell^{-2} \log(2\ell^2|\mathcal{H}|/\delta) \rceil$$

$$\lesssim \log(\log(1/\epsilon)|\mathcal{H}|/\delta) \sum_{\ell=1}^{\lceil \log_2(4/\epsilon) \rceil} \rho_\ell \epsilon_\ell^{-2}$$

$$= \log(\log(1/\epsilon)|\mathcal{H}|/\delta) \sum_{\ell=1}^{\lceil \log_2(4/\epsilon) \rceil} \epsilon_\ell^{-2} \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h,h' \in \widehat{\mathcal{H}}_\ell} \mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h'(X)\}}{\lambda(X)/\nu(X)}\right]$$

$$\lesssim \log(\log(1/\epsilon)|\mathcal{H}|/\delta) \sum_{\ell=1}^{\lceil \log_2(4/\epsilon) \rceil} \epsilon_\ell^{-2} \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \widehat{\mathcal{H}}_\ell} \mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda(X)/\nu(X)}\right]$$

$$\lesssim \log(\log(1/\epsilon)|\mathcal{H}|/\delta) \sum_{\ell=1}^{\lceil \log_2(4/\epsilon) \rceil} \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq 4\epsilon_\ell} \frac{\mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda(X)/\nu(X)}\right]}{4\epsilon_\ell^2}$$

$$\lesssim \log(\log(1/\epsilon)|\mathcal{H}|/\delta) \lceil \log_2(4/\epsilon) \rceil \max_{\ell=1,\dots,\lceil \log_2(4/\epsilon) \rceil} \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq 4\epsilon_\ell} \frac{\mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda(X)/\nu(X)}\right]}{4\epsilon_\ell^2}$$

$$\lesssim \log(\log(1/\epsilon)|\mathcal{H}|/\delta) \log(1/\epsilon) \max_{\xi \geq \epsilon} \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq \xi} \frac{\mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda(X)/\nu(X)}\right]}{\xi^2}$$

where we have upper bounded the sum over ℓ by the max. We immediately obtain the following theorem.

Theorem 15. *Fix any $\epsilon > 0$ and $\delta \in (0,1)$. Define*

$$\rho^*(\epsilon) := \sup_{\xi \geq \epsilon} \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq \xi} \frac{\mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda(X)/\nu(X)}\right]}{\xi^2} \leq \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}} \frac{\mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda(X)/\nu(X)}\right]}{\epsilon^2 \vee (R(h)-R(h^*))^2}$$

Any $\widehat{h}$ in the version space of Algorithm 5 satisfies $R(\widehat{h}) - R(h^*) \leq \epsilon$ once $\gtrsim \rho^*(\epsilon)\log(\log(1/\epsilon)|\mathcal{H}|/\delta)\log(1/\epsilon)$ labels have been requested.

Note that $\rho^*(\epsilon)$ is balancing the variance (numerator) with the sub-optimality gap squared (denominator) for every $h \in \mathcal{H}$. By inspecting the proof of the following proposition, disagreement based learning corresponds to a very particular choice of λ_ℓ at each round, namely, uniform distribution over the disagreement region. Our optimized choice of λ is never worse than this particular choice.

Proposition 5. *Define the disagreement coefficient as*

$$\theta^*(u) = \sup_{\xi \geq u} \frac{\mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H} : h(X) \neq h^*(X), \rho_\nu(h, h^*) \leq \xi\}]}{\xi}.$$

Then $\rho^*(\epsilon) \leq 4(\frac{R(h^*)^2}{\epsilon^2} + 1)\theta^*(R(h^*) + \epsilon)$.

Proof. Now, for any each ξ, if we take $\lambda^\xi(x) = \frac{\nu(x)\mathbf{1}\{\exists h \in \mathcal{H}: h(x) \neq h^*(x), R(h)-R(h^*) \leq \xi\}}{\mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H}: h(X) \neq h^*(X), R(h)-R(h^*) \leq \xi\}]}$ then

$$\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq \xi} \frac{\mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda(X)/\nu(X)}\right]}{\xi^2} \leq \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq \xi} \frac{\mathbb{E}_{X \sim \nu}\left[\frac{\mathbf{1}\{h(X) \neq h^*(X)\}}{\lambda^\xi(X)/\nu(X)}\right]}{\xi^2}$$

$$= \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq \xi} \frac{\mathbb{E}_{X \sim \nu}\left[\mathbf{1}\{h(X) \neq h^*(X)\}\right] \mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H} : h(X) \neq h^*(X), R(h)-R(h^*) \leq \xi\}]}{\xi^2}$$

$$= \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq \xi} \frac{\rho_\nu(h, h^*) \mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H} : h(X) \neq h^*(X), R(h)-R(h^*) \leq \xi\}]}{\xi^2}$$

$$\leq \max_{h \in \mathcal{H}: R(h)-R(h^*) \leq \xi} \frac{\rho_\nu(h, h^*) \mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H} : h(X) \neq h^*(X), \rho_\nu(h, h^*) \leq 2R(h^*)+\xi\}]}{\xi^2}$$

$$\leq \frac{(2R(h^*)+\xi)\mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H} : h(X) \neq h^*(X), \rho_\nu(h, h^*) \leq 2R(h^*)+\xi\}]}{\xi^2}$$

$$\leq \begin{cases} \frac{9R(h^*)^2}{\xi^2} \frac{\mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H}: h(X) \neq h^*(X), \rho_\nu(h,h^*) \leq 2R(h^*)+\xi\}]}{2R(h^*)+\xi} & \text{if } \xi \leq R(h^*) \\ \frac{9\mathbb{E}_{X \sim \nu}[\mathbf{1}\{\exists h \in \mathcal{H}: h(X) \neq h^*(X), \rho_\nu(h,h^*) \leq 2R(h^*)+\xi\}]}{2R(h^*)+\xi} & \text{if } \xi > R(h^*) \end{cases}$$

$$\leq 9(\frac{R(h^*)^2}{\xi^2} + 1)\theta^*(2R(h^*)+\xi)$$

where we have used the fact that

$$\rho_\nu(h, h^*) = \mathbb{E}_{X \sim \nu}\left[\mathbf{1}\{h(X) \neq h^*(X)\}\right] \leq \mathbb{E}_{X \sim \nu}\left[\mathbf{1}\{h(X) \neq Y\} + \mathbf{1}\{Y \neq h^*(X)\}\right] = R(h) + R(h^*)$$

so that $\rho_\nu(h, h^*) \leq \xi + 2R(h^*)$ whenever $R(h) - R(h^*) \leq \xi$. $\qquad\square$

Claim 1. *For the "hard instance" of Section 6.2.3, $R(h^*) = 0$, $\theta^*(\epsilon) \geq \frac{1}{\epsilon}$, and $\rho^*(\epsilon) \leq \log(1/\epsilon)$. Thus, Algorithm 5 requires exponentially fewer labels than the (nearly) minimax optimal Robust CAL algorithm of Algorithm 3. Algorithm 5 is also (nearly) minimax optimal.*

6.4.4   A Computationally efficient Algorithm

Consider Algorithm 6.

Algorithm 6 Computationally efficient Algorithm for Binary Classification

---

1: **Input**: Policy set $\mathcal{H}$ such that $h : \mathcal{X} \to \{0, 1\}$ for all $h \in \mathcal{H}$, confidence level $\delta \in (0, 1)$.
2: Choose $\widehat{h}_{-1} \in \mathcal{H}$ arbitrarily, set $\lambda_0 = \arg\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}} \rho_\lambda(h, \widehat{h}_{-1})$ and $\tau_0 \approx \epsilon_0^{-2} \max_{h \in \mathcal{H}} \rho_\lambda(h, \widehat{h}_0) \log(|\mathcal{H}|/\delta)$.
3: Draw $x_1, \ldots, x_{\tau_0} \sim \lambda_0$, set $w_t = \lambda_\ell(x_t)/\nu(x_t)$, and request their labels to obtain $\{(x_t, y_t, w_t)\}_{t=1}^{\tau_0}$. Set $\widetilde{R}_0^\gamma(h) = \frac{1}{\tau_\ell} \sum_{t=1}^{\tau_0} \frac{1}{w_t + \gamma} \mathbf{1}\{h(x_t) \neq y_t\}$ for all $h \in \mathcal{H}$ and $\gamma > 0$. Find $\widetilde{h}_0$
4: **for** $\ell = 1, 2, \ldots$ **do**
5:   Let τ_ℓ be a minimal value of τ such that the objective, achieved by λ_ℓ, is no greater than ϵ_ℓ:

$$\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{h \in \mathcal{H}} \min_{\gamma' \in \Gamma} -\frac{8}{9}\widetilde{\Delta}_{\ell-1}(h) + 384\gamma'\rho_\lambda(h, \widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma'\tau}$$

6:   Draw $x_1, \ldots, x_{\tau_\ell} \sim \lambda_\ell$, set $w_t = \lambda_\ell(x_t)/\nu(x_t)$, and request their labels to obtain $\{(x_t, y_t, w_t)\}_{t=1}^{\tau_\ell}$. Set $\widetilde{R}_\ell^\gamma(h) = \frac{1}{\tau_\ell} \sum_{t=1}^{\tau_\ell} \frac{1}{w_t + \gamma} \mathbf{1}\{h(x_t) \neq y_t\}$ for all $h \in \mathcal{H}$ and $\gamma > 0$.
7:   Set

$$\widehat{h}_\ell = \arg\min_{h \in \mathcal{H}} \min_\gamma \widetilde{R}_\ell^\gamma(h) - \widetilde{R}_\ell^\gamma(\widehat{h}_{\ell-1}) + 6\gamma\rho_{\lambda_\ell}(h, \widehat{h}_{\ell-1}) + \frac{6\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma\tau_\ell}$$

8:   Set

$$\widetilde{\Delta}_\ell(h) = \min_\gamma \widetilde{R}_\ell^\gamma(h) - \widetilde{R}_\ell^\gamma(\widetilde{h}_\ell) + 2\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_\ell) + \frac{4\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}$$

9: **end for**
10: **Output**:

Let $\epsilon_\ell = 2^{-\ell}$ and $S_\ell = \{h \in \mathcal{H} : R(h) - R(h^*) \leq \epsilon_\ell\}$ for all $\ell \in \mathbb{N}$. Define the events

$$\mathcal{E}_\ell = \bigcap_{h,h' \in \mathcal{H}} \{\widetilde{R}_\ell^\gamma(h) - \widetilde{R}_\ell^\gamma(h') - R(h) + R(h') \leq 2\gamma\rho_{\lambda_\ell}(h, h') + \frac{2\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}\}$$

and $\mathcal{E} = \cap_{\ell=0}^\infty \mathcal{E}_\ell$. Define $\Delta(h, h') = R(h) - R(h')$ and for each ℓ define

$$\widetilde{\Delta}_\ell(h, h') = \min_\gamma \widetilde{R}_\ell^\gamma(h) - \widetilde{R}_\ell^\gamma(h') + 2\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_\ell) + 2\gamma\rho_{\lambda_\ell}(h', \widetilde{h}_\ell) + \frac{4\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}$$

Note that $\widetilde{\Delta}_\ell(h, h')$ is a *pessimistic* estimate of the true gap $\Delta(h, h')$ in the sense that on $\mathcal{E}$ we have $\widetilde{\Delta}_\ell(h, h') \geq \Delta(h, h')$ for all h, h'.

We need to show that we are estimating the gaps well. We will prove a helper lemma first.

Lemma 16. *Fix $\ell \in \mathbb{N}$. Then, on the event $\{\widetilde{\Delta}_{\ell-1}(h^*) \leq \epsilon_{\ell-1}/8\} \cap \mathcal{E}$ we have that $R(\widehat{h}_\ell) - R(h^*) \leq \epsilon_\ell/8$.*

Proof. On $\mathcal{E}$ and the event that $\widetilde{\Delta}_{\ell-1}(h^*) \leq \epsilon_{\ell-1}/8 = \epsilon_\ell/4$

$$
\begin{aligned}
\epsilon_\ell &\geq \max_{h \in \mathcal{H}} \min_{\gamma' \in \Gamma} -\frac{8}{9}\widetilde{\Delta}_{\ell-1}(h) + 384\gamma' \rho_{\lambda_\ell}(h, \widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma'\tau_\ell} \\
&\geq \min_{\gamma' \in \Gamma} -\frac{8}{9}\widetilde{\Delta}_{\ell-1}(h^*) + 384\gamma' \rho_{\lambda_\ell}(h^*, \widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma'\tau_\ell} \\
&\geq \min_{\gamma' \in \Gamma} -\frac{2}{9}\epsilon_\ell + 384\gamma' \rho_{\lambda_\ell}(h^*, \widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma'\tau_\ell} \\
&\geq 64\sqrt{\frac{\rho_{\lambda_\ell}(h^*, \widehat{h}_{\ell-1})\log(2\ell^2|\mathcal{H}|/\delta)}{\tau_\ell}}.
\end{aligned}
$$

Now,

$$
\begin{aligned}
R(\widehat{h}_\ell) - R(\widehat{h}_{\ell-1}) &\leq \min_\gamma \widetilde{R}_\ell^\gamma(\widehat{h}_\ell) - \widetilde{R}_\ell^\gamma(\widehat{h}_{\ell-1}) + 2\gamma\rho_{\lambda_\ell}(\widehat{h}_\ell, \widehat{h}_{\ell-1}) + \frac{2\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma\tau_\ell} \\
&= \min_{h \in \mathcal{H}} \min_\gamma \widetilde{R}_\ell^\gamma(h) - \widetilde{R}_\ell^\gamma(\widehat{h}_{\ell-1}) + 2\gamma\rho_{\lambda_\ell}(h, \widehat{h}_{\ell-1}) + \frac{2\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma\tau_\ell} \\
&\leq \min_\gamma \widetilde{R}_\ell^\gamma(h^*) - \widetilde{R}_\ell^\gamma(\widehat{h}_{\ell-1}) + 2\gamma\rho_{\lambda_\ell}(h^*, \widehat{h}_{\ell-1}) + \frac{2\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma\tau_\ell} \\
&\leq \min_\gamma R(h^*) - R(\widehat{h}_{\ell-1}) + 4\gamma\rho_{\lambda_\ell}(h^*, \widehat{h}_{\ell-1}) + \frac{4\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma\tau_\ell} \\
&= R(h^*) - R(\widehat{h}_{\ell-1}) + 8\sqrt{\frac{\rho_{\lambda_\ell}(h^*, \widehat{h}_{\ell-1})\log(2\ell^2|\mathcal{H}|/\delta)}{\tau_\ell}} \\
&= R(h^*) - R(\widehat{h}_{\ell-1}) + \epsilon_\ell/8
\end{aligned}
$$

where the last line follows from the above display. Rearranging, we conclude the proof. $\qquad \square$

Lemma 17. *On event $\mathcal{E}$ we have for all $\ell \in \mathbb{N}$*

$$
0 \leq \widetilde{\Delta}_\ell(h, h^*) - \Delta(h, h^*) \leq \begin{cases} \frac{\Delta(h, h^*)}{8} & \textit{if } h \notin S_\ell \\ \epsilon_\ell/8 & \textit{if } h \in S_\ell, \end{cases}
$$

.

Proof. We will proceed inductively assuming it holds for $\ell - 1$ and then show this implies the result

for round ℓ. Note that for $\ell = 0$ we have for any $h \in S_0 = \mathcal{H}$

$$\max_{h \in S_0} \widetilde{\Delta}_\ell(h, h^*) - \Delta(h, h^*)$$

$$= \max_{h \in S_0} \min_\gamma \widetilde{R}_0^\gamma(h) - \widetilde{R}_0^\gamma(h^*) + 2\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_0) + 2\gamma\rho_{\lambda_\ell}(h^*, \widetilde{h}_0) + \frac{4\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_0} - R(h) + R(h^*)$$

$$\leq \max_{h,h' \in S_0} \min_\gamma \widetilde{R}_0^\gamma(h) - \widetilde{R}_0^\gamma(h') + 2\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_0) + 2\gamma\rho_{\lambda_\ell}(h', \widetilde{h}_0) + \frac{4\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_0} - R(h) + R(h')$$

$$\leq \max_{h,h' \in S_0} \min_\gamma 4\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_0) + 4\gamma\rho_{\lambda_\ell}(h', \widetilde{h}_0) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_0}$$

$$\leq \max_{h \in S_0} \min_\gamma 8\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_0) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_0}$$

$$= 16\sqrt{\frac{\max_{h \in \mathcal{H}} \rho_{\lambda_0}(h, \widetilde{h}_0) \log(2\ell^2|\mathcal{H}|^2/\delta)}{\tau_0}}$$

$$\leq \epsilon_0/8$$

by choosing τ_0 sufficiently large. Thus $\widetilde{\Delta}_0(h, h^*) - \Delta(h, h^*) \leq \epsilon_0/8$ for all $h \in S_0 = \mathcal{H}$.
Now we use induction to prove the rest. For any $\widetilde{h} \in \mathcal{H}$ we have

$$\widetilde{\Delta}_\ell(\widetilde{h}, h^*) - \Delta(\widetilde{h}, h^*)$$

$$= \min_\gamma \widetilde{R}_\ell^\gamma(\widetilde{h}) - \widetilde{R}_\ell^\gamma(h^*) + 2\gamma\rho_{\lambda_\ell}(\widetilde{h}, \widetilde{h}_\ell) + 2\gamma\rho_{\lambda_\ell}(h^*, \widetilde{h}_\ell) + \frac{4\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell} - R(\widetilde{h}) + R(h^*)$$

$$\leq \min_\gamma 4\gamma\rho_{\lambda_\ell}(\widetilde{h}, \widetilde{h}_\ell) + 4\gamma\rho_{\lambda_\ell}(h^*, \widetilde{h}_\ell) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}$$

$$\leq \min_\gamma 4\gamma\rho_{\lambda_\ell}(\widetilde{h}, \widetilde{h}_{\ell-1}) + 4\gamma\rho_{\lambda_\ell}(\widetilde{h}_{\ell-1}, h^*) + 8\gamma\rho_{\lambda_\ell}(\widetilde{h}_\ell, \widetilde{h}_{\ell-1}) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}.$$

This last line will be used as the starting point for the two cases.
Case 1: $\widetilde{h} \in S_\ell$
Now, on the base case and Lemma 17 we have that $\widetilde{h}_{\ell-1} \in S_{\ell+2} \subset S_\ell$ and $\widetilde{h}_\ell \in S_{\ell+3} \subset S_\ell$. Thus, for any $\widetilde{h} \in S_\ell$ we have

$$\widetilde{\Delta}_\ell(\widetilde{h}, h^*) - \Delta(\widetilde{h}, h^*) \leq \min_\gamma 4\gamma\rho_{\lambda_\ell}(\widetilde{h}, \widetilde{h}_{\ell-1}) + 4\gamma\rho_{\lambda_\ell}(\widetilde{h}_{\ell-1}, h^*) + 8\gamma\rho_{\lambda_\ell}(\widetilde{h}_\ell, \widetilde{h}_{\ell-1}) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}$$

$$\leq \max_{h \in S_\ell} \min_\gamma 16\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_{\ell-1}) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}$$

$$\leq \frac{1}{20} \max_{h \in S_\ell} \min_\gamma -\frac{8}{9}\widetilde{\Delta}_{\ell-1}(h, \widetilde{h}_{\ell-1}) + \frac{8}{9}\widetilde{\Delta}_{\ell-1}(h, \widetilde{h}_{\ell-1}) + 384\gamma\rho_{\lambda_\ell}(h, \widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma\tau_\ell}$$

$$\leq \frac{1}{20} \max_{h \in S_\ell} \min_\gamma -\frac{8}{9}\widetilde{\Delta}_{\ell-1}(h, \widetilde{h}_{\ell-1}) + \epsilon_\ell + 384\gamma\rho_{\lambda_\ell}(h, \widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma\tau_\ell}$$

$$\leq 2\epsilon_\ell/20 \leq \epsilon_\ell/8$$

where we have used the inductive hypothesis that $\frac{8}{9}\widetilde{\Delta}_{\ell-1}(h, \widetilde{h}_{\ell-1}) \leq \Delta(h, \widetilde{h}_{\ell-1}) \leq \epsilon_\ell$ since $\widetilde{h}_{\ell-1} \in S_\ell$.
This completes the first case.
Case 2: $\widetilde{h} \notin S_\ell$
If $\widetilde{h} \notin S_\ell$ then there exists some $j \leq \ell$ such that $2^{-j} < R(h) - R(h^*) \leq 2^{-j+1}$. On Lemma 17, we

have again that on the inductive hypothesis, $\widetilde{h}_\ell \in S_{\ell+3} \subset S_j$ so

$$\frac{\widetilde{\Delta}_\ell(\widetilde{h}, h^*) - \Delta(\widetilde{h}, h^*)}{\Delta(\widetilde{h}, h^*)} \leq \frac{\min_\gamma 4\gamma\rho_{\lambda_\ell}(\widetilde{h}, \widetilde{h}_{\ell-1}) + 4\gamma\rho_{\lambda_\ell}(\widetilde{h}_{\ell-1}, h^*) + 8\gamma\rho_{\lambda_\ell}(\widetilde{h}_\ell, \widetilde{h}_{\ell-1}) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}}{\Delta(\widetilde{h}, h^*)}$$

$$\leq \max_{h \in S_j} \frac{\min_\gamma 16\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_{\ell-1}) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}}{R(\widetilde{h}) - R(h^*)}$$

$$\leq 3\max_{h \in S_j} \frac{\min_\gamma 16\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_{\ell-1}) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}}{2^{-\ell} + 2[R(\widetilde{h}) - R(h^*)]}$$

$$\leq 3\max_{h \in \mathcal{H}} \frac{\min_\gamma 16\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_{\ell-1}) + \frac{8\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}}{2^{-\ell} + R(h) - R(h^*)}$$

$$= \max_{h \in \mathcal{H}} \frac{\min_\gamma 48\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_{\ell-1}) + \frac{24\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}}{\epsilon_\ell + \Delta(h)}$$

Now, on the inductive hypothesis we have for any $h \in \mathcal{H}$ that

$$\widetilde{\Delta}_{\ell-1}(h, h^*) - \Delta(h) \leq \max\{\epsilon_{\ell-1}, \Delta(h)\}/8 \leq \epsilon_\ell/4 + \Delta(h)/8.$$

Rearranging, we have that

$$\frac{9}{8}\Delta(h) + \epsilon_\ell/4 \geq \widetilde{\Delta}_{\ell-1}(h, h^*)$$

$$= \min_\gamma \widetilde{R}^\gamma_{\ell-1}(h) - \widetilde{R}^\gamma_{\ell-1}(h^*) + 2\gamma\rho_{\lambda_{\ell-1}}(h, \widetilde{h}_{\ell-1}) + 2\gamma\rho_{\lambda_{\ell-1}}(\widetilde{h}_{\ell-1}, h^*) + \frac{4\log(\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_{\ell-1}}$$

$$\geq \min_\gamma \widetilde{R}^\gamma_{\ell-1}(h) - \widetilde{R}^\gamma_{\ell-1}(\widetilde{h}_{\ell-1}) + \widetilde{R}^\gamma_{\ell-1}(\widetilde{h}_{\ell-1}) - \widetilde{R}^\gamma_{\ell-1}(h^*) + 2\gamma\rho_{\lambda_{\ell-1}}(h, \widetilde{h}_{\ell-1}) + 2\gamma\rho_{\lambda_{\ell-1}}(\widetilde{h}_{\ell-1}, h^*) + \frac{4\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_{\ell-1}}$$

$$\geq \min_\gamma \widetilde{R}^\gamma_{\ell-1}(h) - \widetilde{R}^\gamma_{\ell-1}(\widetilde{h}_{\ell-1}) + 2\gamma\rho_{\lambda_{\ell-1}}(h, \widetilde{h}_{\ell-1}) + \frac{2\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_{\ell-1}} + R(\widetilde{h}_{\ell-1}) - R(h^*)$$

$$\geq \widetilde{\Delta}_{\ell-1}(h, \widetilde{h}_{\ell-1})/2 = \widetilde{\Delta}_{\ell-1}(h)/2.$$

Thus, rearranging once more we have $\Delta(h) \geq \frac{4}{9}\widetilde{\Delta}_{\ell-1}(h) - \epsilon_\ell/4$ and

$$\frac{\widetilde{\Delta}_\ell(\widetilde{h}, h^*) - \Delta(\widetilde{h}, h^*)}{\Delta(\widetilde{h}, h^*)} \leq \max_{h \in \mathcal{H}} \frac{\min_\gamma 48\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_{\ell-1}) + \frac{24\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}}{\epsilon_\ell + \Delta(h)}$$

$$\leq \max_{h \in \mathcal{H}} \frac{\min_\gamma 48\gamma\rho_{\lambda_\ell}(h, \widetilde{h}_{\ell-1}) + \frac{24\log(2\ell^2|\mathcal{H}|^2/\delta)}{\gamma\tau_\ell}}{\epsilon_\ell/2 + \frac{4}{9}\widetilde{\Delta}_{\ell-1}(h)} \leq 1/8$$

where the last inequality follows from the solution of the optimization problem.   □

The above lemmas imply that

$$\Delta(h) = \Delta(h, h^*)$$
$$= \Delta(h, \widetilde{h}_{\ell-1}) + \Delta(\widetilde{h}_{\ell-1}, h^*)$$
$$\leq \widetilde{\Delta}_{\ell-1}(h, \widetilde{h}_{\ell-1}) + \epsilon_{\ell-1}/8$$
$$= \widetilde{\Delta}_{\ell-1}(h) + \epsilon_{\ell-1}/8.$$

We leverage this to compute the sample complexity as follows:

$$\min_{\lambda\in\triangle_\mathcal{X}} \max_{h\in\mathcal{H}} \min_{\gamma'\in\Gamma} -\frac{8}{9}\widetilde{\Delta}_{\ell-1}(h) + 384\gamma'\rho_\lambda(h,\widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma'\tau}$$

$$\leq \min_{\lambda\in\triangle_\mathcal{X}} \max_{h\in\mathcal{H}} \min_{\gamma'\in\Gamma} -\frac{8}{9}\Delta(h) + \epsilon_\ell/8 + 384\gamma'\rho_\lambda(h,\widehat{h}_{\ell-1}) + \frac{192\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma'\tau}$$

$$\leq \min_{\lambda\in\triangle_\mathcal{X}} \max_{h\in\mathcal{H}} -\frac{8}{9}\Delta(h) + \epsilon_\ell/8 + 544\sqrt{\rho_\lambda(h,\widehat{h}_{\ell-1})\frac{\log(2\ell^2|\mathcal{H}|/\delta)}{\gamma'\tau}}$$

$$\leq \min_{\lambda\in\triangle_\mathcal{X}} \max_{h\in\mathcal{H}} -\frac{8}{9}\Delta(h) + \epsilon_\ell/8 + 544\sqrt{\rho_\lambda(h,h^*)\frac{\log(2\ell^2|\mathcal{H}|/\delta)}{\tau}} + 544\sqrt{\rho_\lambda(h^*,\widehat{h}_{\ell-1})\frac{\log(2\ell^2|\mathcal{H}|/\delta)}{\tau}}$$

$$\leq \min_{\lambda\in\triangle_\mathcal{X}} \max_{h\in\mathcal{H}} -\frac{8}{9}\Delta(h) + \epsilon_\ell/8 + 544\sqrt{\rho_\lambda(h,h^*)\frac{\log(2\ell^2|\mathcal{H}|/\delta)}{\tau}} + 544\sqrt{\max_{h'\in S_\ell}\rho_\lambda(h^*,h')\frac{\log(2\ell^2|\mathcal{H}|/\delta)}{\tau}}$$

which is less than ϵ_ℓ whenever

$$\tau \gtrsim \min_{\lambda\in\triangle_\mathcal{X}} \max_{h\in\mathcal{H}} \frac{\rho_\lambda(h,h^*)}{\epsilon_\ell^2 + \Delta(h)^2}\log(|\mathcal{H}|/\delta).$$

We summarize the conclusions in the following theorem.

Theorem 16. *Fix $\delta\in(0,1)$. Then on the ℓth round, $R(\widetilde{h}_\ell) - R(h^*) \leq \epsilon_\ell/8$ and the total number of samples is bounded by $\min_{\lambda\in\triangle_\mathcal{X}} \max_{h\in\mathcal{H}} \frac{\rho_\lambda(h,h^*)}{\epsilon_\ell^2+\Delta(h)^2}\log(|\mathcal{H}|/\delta)\log(1/\epsilon_\ell)$.*

6.4.5 Instance-dependent Lower bounds

Combinatorial bandits sheds some light on instance-dependent lower-bounds for the pool-based setting.

6.5 Heuristics of note

6.5.1 Uncertainty sampling

6.5.2 Covering algorithms

6.5.3 Hypothesis-class agnostic algorithms

Part II

Adversarial bandits

6.6 Stochastic online mirror descent

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 $\ell(A_t, z_t) = A_t^\top z_t$

6.6.1 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

$$\widetilde{a}_{t+1} = \nabla F^*(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t)) \tag{6.2}$$

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} D_F(a, \widetilde{a}_{t+1}) \tag{6.3}$$

where we have assumed the iterates exist.

Theorem 17 (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 $\nabla F(a) - \eta z \in \text{dom}(\nabla F(\mathcal{D}))$ 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 2 (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:

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

Corollary 3 (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*

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

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

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

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

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

$\square$

6.7 Simplex games 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

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

6.7.1 Full information game, simplex action set

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 4 (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 $a \in \mathcal{A}$

$$\sum_{t=1}^{T} \ell(a_t, z_t) - \ell(a, z_t) = \sum_{t=1}^{T} z_t^\top (a_t - a)$$

$$\leq \frac{\log(d)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} \sum_{i=1}^{d} a_{t,i}(\exp(-\eta z_{t,i}) - 1 + \eta z_{t,i})$$

$$\leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} a_{t,i} z_{t,i}^2$$

$$\leq \frac{\log(d)}{\eta} + \frac{\eta T}{2}$$

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. $\square$

6.7.2 Full information game, 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

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

$$= \mathbb{E}\left[\sum_{t=1}^{T} z_t^\top(a_t - a)\right]$$

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

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 $a_1 = (1/d, \ldots, 1/d)$. Adversary chooses $\{z_t\}_{t=1}^{T} \subset [0,1]^d$.

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

    Player chooses $a_t \in \mathcal{A}$

    Player 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 5 (Exponential weights over finite actions). *Under the conditions of Example 1 where the player can only play $\mathbf{e}_i$ for $i \in \{1, \ldots, d\}$ with let $\ell(a, z) = a^\top z$, the exponential weights over finite actions algorithm satisfies*

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

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] $\qquad\qquad\square$

6.7.3 Bandit feedback, finite action set

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

$$\mathbb{E}[\widehat{z}_{t,i}|a_1, I_1, \ldots, a_{t-1}, I_{t-1}, a_t] = \mathbb{E}[\frac{\mathbf{1}\{I_t=i\}z_{t,i}}{a_{t,i}}|a_1, I_1, \ldots, a_{t-1}, I_{t-1}, a_t]$$

$$= \sum_{j=1}^{d} a_{t,j}\frac{\mathbf{1}\{j=i\}z_{t,i}}{a_{t,i}}$$

$$= z_{t,i}.$$

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 6 (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*

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

$$\leq \frac{\log(d)}{\eta} + \frac{\eta T d}{2} \leq \sqrt{2dT\log(d)}$$

Proof. We can directly apply Corollary 2:

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

$$= \frac{\log(d)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d}a_{t,i}(\exp(-\eta\widehat{z}_{t,i}) - 1 + \eta\widehat{z}_{t,i})\right]$$

$$\leq \frac{\log(d)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d}a_{t,i}\widehat{z}_{t,i}^2\right]$$

$$= \frac{\log(d)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d}a_{t,i}\frac{\mathbf{1}\{I_t=i\}z_{t,i}^2}{a_{t,i}^2}\right]$$

$$= \frac{\log(d)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d}z_{t,i}^2$$

$$\leq \frac{\log(d)}{\eta} + \frac{\eta dT}{2}$$

□

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

6.8   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:

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

$$\leq \frac{1}{\eta} + \frac{\eta}{2}\sum_{t=1}^T \mathbb{E}\left[\|\widehat{z}_t\|_2^2\right]$$

$$= \frac{1}{\eta} + \frac{\eta}{2}\sum_{t=1}^T \mathbb{E}\left[\sum_{i=1}^d \widehat{z}_{t,i}^2\right]$$

$$= \frac{1}{\eta} + \frac{\eta}{2}\sum_{t=1}^T \mathbb{E}\left[\sum_{i=1}^d \frac{z_{t,i}^2}{a_{t,i}}\right].$$

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 6. *If F is twice continuously differntiable, and if its Hessian $\nabla^2 F(x)$ 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*

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

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

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

For the choice of $F(x) = \frac{1}{2}\|x\|_2^2$ we have $\nabla^2 F(\zeta_t) = I$ 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 $\nabla^2 F(x) = \operatorname{diag}(1/x_1, \ldots, 1/x_d)$ which blows up as a component of x approaches 0. But this is

perfect, since then

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

$$= \mathbb{E}\left[\frac{\eta^2}{2}\sum_{i=1}^d \widehat{z}_{t,i}^2 \zeta_{t,i}\right]$$

$$= \mathbb{E}\left[\frac{\eta^2}{2}\sum_{i=1}^d \frac{z_{t,i}^2}{a_{t,i}} \zeta_{t,i}\right]$$

$$\approx \frac{\eta^2}{2}\sum_{i=1}^d z_{t,i}^2$$

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.

6.8.1 Bandit feedback, unit ball action set

Here we address the action set $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. 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:

- $\mathcal{X}_t \sim \text{Bernoulli}(\|a_t\|_2)$, $I_t \sim \text{uniform}([d])$, $\epsilon_t \in \{-1, 1\}$ with equal probability

- $A_t = (1 - \mathcal{X}_t)\epsilon_t \mathbf{e}_{I_t} + \mathcal{X}_t \frac{a_t}{\|a_t\|_2}$

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

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

$$\mathbb{E}\left[\|\widehat{z}_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}\right] = \mathbb{E}\left[\|(1 - \mathcal{X}_t)\frac{d}{1 - \|a_t\|_2} A_t A_t^\top z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}\right]$$

$$= \mathbb{E}\left[\frac{d^2}{1 - \|a_t\|_2}\|A_t A_t^\top z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}|\mathcal{X}_t = 0\right]$$

$$= \sum_{i=1}^d \frac{1}{d}\frac{d^2}{1 - \|a_t\|_2}\|e_i e_i^\top z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}$$

$$= \frac{d}{1 - \|a_t\|_2}\|z_t\|^2_{(\nabla^2 F(\zeta_t))^{-1}}$$

$$\leq \frac{d}{1 - \|a_t\|_2}(1 - \|\zeta_t\|_2)\|z_t\|^2_2$$

$$\leq 2d$$

where we have used the fact that $\nabla^2 F(x) \succeq I/(1 - \|x\|_2)$ and $\frac{1 - \|z_t\|_2}{1 - \|a_t\|_2} \leq 2$ (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, 20

6.8.2 Bandit feedback, finite action sets; Linear bandits

Here we study the case when $|\mathcal{A}| = n$ and $\mathcal{A} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and $\max_{x \in \mathcal{A}} \max_t |x^\top z_t| \leq 1$. 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 7 (EXP3(γ)). *The regret of EXP3(γ) algorithm satisfies*

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

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

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

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

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

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

so that

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

Finally,

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

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

6.9   Contextual bandits, EXP4

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 $Q_1 = (1/m, \ldots, 1/m) \in [0,1]^{1 \times m}$. 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 $p_{t,i} = \mathbb{E}_{M \sim Q_t}[E_{M,i}^{(t)}] = \sum_{j=1}^m Q_{t,j} E_{j,i}^{(t)}$

Player draws $M_t \sim Q_t$ and $I_t \sim E_{M_t}^{(t)} \in \triangle_n$ (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 $\widehat{y}_{t,j} = \sum_{i=1}^n E_{j,i}^{(t)} \widehat{\ell}_{t,i}$ for all $j = 1, \ldots, m$

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:

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

and

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

The expert regret is defined as

$$\max_{k=1,\ldots,m} \mathbb{E}\left[\sum_{t=1}^T \ell_{t,I_t} - \sum_{i=1}^n E_{k,i}^{(t)} \ell_{t,i}\right] = \max_{k=1,\ldots,m} \mathbb{E}\left[\sum_{t=1}^T \sum_{j=1}^m Q_{t,j} \sum_{i=1}^n E_{j,i}^{(t)} \ell_{t,i} - \sum_{i=1}^n E_{k,i}^{(t)} \ell_{t,i}\right]$$

$$= \max_{k \in [m]} \mathbb{E}\left[\sum_{t=1}^T \sum_{j=1}^m Q_{t,j} \widehat{y}_{t,j} - \widehat{y}_{t,k}\right]$$

$$= \max_{k \in [m]} \mathbb{E}\left[\sum_{t=1}^T \widehat{y}_{t,M_t} - \widehat{y}_{t,k}\right]$$

$$\leq \frac{\log(m)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}\left[\sum_{j=1}^m Q_{t,j} \widehat{y}_{t,j}^2\right]$$

$$\leq \frac{\log(m)}{\eta} + \frac{\eta n T}{2}$$

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

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

so

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

Bibliography

[Agarwal et al., 2014] Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR.

[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.

[Bartlett et al., 2006] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[Beygelzimer et al., 2009] Beygelzimer, A., Dasgupta, S., and Langford, J. (2009). Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56.

[Beygelzimer et al., 2010] Beygelzimer, A., Hsu, D., Langford, J., and Zhang, T. (2010). Agnostic active learning without constraints. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1*, pages 199–207.

[Boucheron, 2005] Boucheron, Stéphane, B. O. L. G. (2005). Theory of classification : a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[Bubeck et al., 2012] Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and non-stochastic 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.

[Castro and Nowak, 2008] Castro, R. M. and Nowak, R. D. (2008). Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353.

[Cohn et al., 1994] Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. *Mach. Learn.*, 15(2):201–221.

[Dasgupta, 2005a] Dasgupta, S. (2005a). Analysis of a greedy active learning strategy. *Advances in neural information processing systems*, 17:337–344.

[Dasgupta, 2005b] Dasgupta, S. (2005b). Coarse sample complexity bounds for active learning. In *NIPS*, volume 18, pages 235–242.

[Dasgupta et al., 2008] Dasgupta, S., Hsu, D. J., and Monteleoni, C. (2008). A general agnostic active learning algorithm. In *Advances in neural information processing systems*, pages 353–360.

[Dudik et al., 2011] Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178.

[Dudík et al., 2011] Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104.

[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.

[Freund et al., 1997] Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168.

[Golovin and Krause, 2011] Golovin, D. and Krause, A. (2011). Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486.

[Hanneke et al., 2014] Hanneke, S. et al. (2014). Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309.

[Hanneke and Yang, 2015] Hanneke, S. and Yang, L. (2015). Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(109):3487–3602.

[Hegedus, 1995] Hegedus, T. (1995). Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 108–117.

[Howard et al., 2018] Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2018). Time-uniform, nonparametric, nonasymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*.

[Huang et al., 2015] Huang, T.-K., Agarwal, A., Hsu, D. J., Langford, J., and Schapire, R. E. (2015). Efficient and parsimonious agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 2755–2763.

[Kääriäinen, 2006] Kääriäinen, M. (2006). Active learning in the non-realizable case. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

[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.

[Kulkarni et al., 1993] Kulkarni, S. R., Mitter, S. K., and Tsitsiklis, J. N. (1993). Active learning using arbitrary binary valued queries. *Machine Learning*, 11(1):23–35.

[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*.

[Lawler, 2006] Lawler, G. F. (2006). *Introduction to stochastic processes*. CRC Press.

[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.

[Maurer and Pontil, 2009] Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

[Nowak, 2011] Nowak, R. D. (2011). The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906.

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

[Raginsky and Rakhlin, 2011] Raginsky, M. and Rakhlin, A. (2011). Lower bounds for passive and active learning.

[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.

[Tosh and Dasgupta, 2017] Tosh, C. and Dasgupta, S. (2017). Diameter-based active learning. In *International Conference on Machine Learning*, pages 3444–3452. PMLR.

[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.