Some Notes on Multi-armed Bandits

Kevin Jamieson, University of Washington
February 7, 2021

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

1 Introduction

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

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

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

- 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 Regret Minimization

After T time steps, define the *regret* as

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

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.

1.2 Best-arm identification

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

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

1.3 Warm-up: 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?

1.4 Finite-sample confidence intervals

Proposition 1 (Chernoff Bounding technique)**.** *Fix ϵ, δ. If $Z_1, Z_2, \dots$ 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.

$$\begin{aligned}
\mathbb{P}(\frac{1}{\tau}\sum_{t=1}^{\tau} Z_t > \epsilon) &= \mathbb{P}(\exp\left(\lambda\sum_{t=1}^{\tau} Z_t\right) > \exp(\lambda\tau\epsilon)) \\
&\leq e^{-\lambda\tau\epsilon}\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{\tau} Z_t\right)\right] && \text{(Markov's)} \\
&= e^{-\lambda\tau\epsilon}\prod_{t=1}^{\tau}\mathbb{E}\left[\exp\left(\lambda Z_t\right)\right] && \text{(Independence)} \\
&= \exp\left(-\lambda\tau\epsilon + \tau\psi_Z(\lambda)\right)
\end{aligned}$$

□

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

$$\begin{aligned}
\psi_X(\lambda) :=& \log(\mathbb{E}_X[\exp(\lambda X)]) \\
\psi'_X(\lambda) =& \frac{1}{\mathbb{E}_X[\exp(\lambda X)]}\mathbb{E}_X[X\exp(\lambda X)] \\
\psi''_X(\lambda) =& \frac{1}{\mathbb{E}_X[\exp(\lambda X)]}\mathbb{E}_X[X^2\exp(\lambda X)] - \left(\frac{1}{\mathbb{E}_X[\exp(\lambda X)]}\mathbb{E}_X[X\exp(\lambda X)]\right)^2 \\
=& \mathbb{E}_Z[Z^2] - \mathbb{E}_Z[Z]^2 \\
=& \mathbb{V}ar(Z) \\
\leq& (b-a)^2/4
\end{aligned}$$

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

$$\begin{aligned}
\psi_X(\lambda) &= \psi_X(0) + \psi'_X(0)\lambda + \psi''_X(\theta)\lambda^2/2 \\
&= \psi''_X(\theta)\lambda^2/2 \\
&\leq (b-a)^2\lambda^2/8
\end{aligned}$$

which completes the proof. □

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

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

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

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

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

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

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

2 Action Elimination Algorithm for Multi-armed Bandits

Input: n arms $\mathcal{X} = \{1, \dots, n\}$, confidence level $\delta \in (0,1)$.
Let $\widehat{\mathcal{X}}_1 \leftarrow \mathcal{X}, \ell \leftarrow 1$
while $|\widehat{\mathcal{X}}_\ell| > 1$ **do**
  $\epsilon_\ell = 2^{-\ell}$
  Pull each arm in $\widehat{\mathcal{X}}_\ell$ exactly $\tau_\ell = \lceil 2\epsilon_\ell^{-2}\log(\frac{4\ell^2|\mathcal{X}|}{\delta})\rceil$ times
  Compute the empirical mean of these rewards $\widehat{\theta}_{i,\ell}$ for all $i \in \widehat{\mathcal{X}}_\ell$
  $\widehat{\mathcal{X}}_{\ell+1} \leftarrow \widehat{\mathcal{X}}_\ell \setminus \left\{i \in \widehat{\mathcal{X}}_\ell : \max_{j\in\widehat{\mathcal{X}}_\ell} \widehat{\theta}_{j,\ell} - \widehat{\theta}_{i,\ell} > 2\epsilon_\ell\right\}$
  $\ell \leftarrow \ell + 1$
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* $1 \in \widehat{\mathcal{X}}_\ell$ *and* $\max_{i\in\widehat{\mathcal{X}}_\ell} \Delta_i \le 8\epsilon_\ell$ *for all* $\ell \in \mathbb{N}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thus, the total number of samples taken before $\widehat{\mathcal{X}}_\ell = \{1\}$ 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 $\sum_{i=2}^{n}\Delta_i^{-2}\log(n\log(\Delta_i^{-2})/\delta)$. □

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

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

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

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

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

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 $\Delta_i \vee \nu \ge \nu$ and minimizing over ν yields a regret of $\sqrt{nT\log(n\log(T)/\delta)}$. The expected regret, of course, is then bounded by

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

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

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

3 Lower bounds for Multi-armed Bandits

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

3.1 Mean of a Gaussian

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

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

where

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

and (integrating only over support of pq)

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

Note that

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

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

We conclude that

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

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

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

3.2 Indentification

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

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

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

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

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

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

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

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

3.3 Regret, minimax

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

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

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

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

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

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

3.4 Gap-dependent regret

Lemma 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.

3.5 Revisiting MAB with Optimism

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

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

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

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

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

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

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

4 Linear Bandits Intro

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

4.1 Least Squares

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

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

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

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

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

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

so that

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

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

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

4.1.1 Aside: Gaussian to sub-Gaussian

For an arbitrary constant μ,

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

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

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

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

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

5 Experimental design

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

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

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

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

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

Lemma 4 (Kiefer-Wolfowitz (1960))**.** *For any $\mathcal{X}$ with $d = dim(span(\mathcal{X}))$, there exists a $\lambda^* \in \triangle_{\mathcal{X}}$ that*

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

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

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

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

Notes:

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

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

6 Linear Bandits: Regret Minimization

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

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

After T time steps, define the *regret* as

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

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

Lemma 5. *Assume that* $\max_{x\in\mathcal{X}}\langle x^\star - x, \theta^*\rangle \le 4$. *With probability at least* $1-\delta$, *we have* $x^\star \in \widehat{\mathcal{X}}_\ell$ *and* $\max_{x\in\widehat{\mathcal{X}}_\ell}\langle x^\star - x, \theta^*\rangle \le 8\epsilon_\ell$ *for all* $\ell \in \mathbb{N}$.

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

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

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

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

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

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

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

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

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

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

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

the regret is bounded by:

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

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

Remarks:

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

7 Linear Bandits: Pure exploration

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We now note that

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

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

8 Linear bandits: regret minimization revisited

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

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

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

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

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

Note that

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

Notes

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

R.V. $Z_1, Z_2 \ldots \overset{iid}{\sim} Y$ where $\mathbb{E}[Z_i] = \mu$

$$\mathbb{E}[\exp(\lambda(Z_i - \mu))] \le \exp(\lambda^2/2)$$

Chernoff technique: For a fixed n, $\delta > 0$

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} Z_i > \mu + \sqrt{\frac{2\log(1/\delta)}{n}} \right) \le \delta$$

Fix $N \in \mathbb{N}$. Suppose we want a bound to hold for all $1 \le n \le N$ simultaneously? Union bound!

$$\mathbb{P}\left(\bigcup_{n=1}^{N} \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i > \mu + \sqrt{\frac{2\log(N/\delta)}{n}} \right\} \right)$$

$$\le \sum_{n=1}^{N} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} Z_i > \mu + \sqrt{\frac{2\log(N/\delta)}{n}} \right)$$

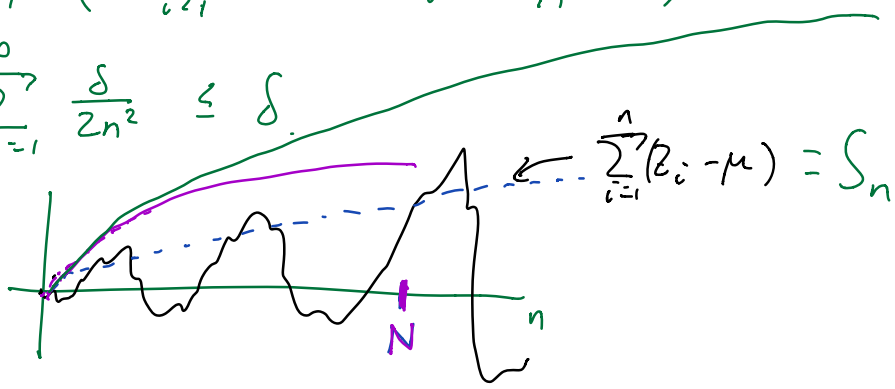
$$\le \sum_{n=1}^{N} \frac{\delta}{N} = \delta.$$

And if we wanted a bound to hold for all $n \in \mathbb{N}$

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i > \mu + \sqrt{\frac{2\log(2n^2/\delta)}{n}} \right\} \right)$$

$$\le \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} Z_i > \mu + \sqrt{\frac{2\log(2n^2/\delta)}{n}} \right)$$

$$\le \sum_{n=1}^{\infty} \frac{\delta}{2n^2} \le \delta.$$



$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$

9 Sequential statistics and Martingales

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 $\{X_t\}_{t=1}^{\mathbb{N}}$ is $\mathbb{F}$-adapted if X_t is $\mathcal{F}_t$ measurable for all $1 \le t \le n$.

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

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 $\mathbf{1}\{\tau \le t\}$ is $\mathcal{F}_t$ measurable for all $t \in \mathbb{N}$.*

$\therefore$

Martingale example $\quad Z_i \overset{iid}{\sim} N(0,1)$

$X_t = \sum_{i=1}^{t} Z_i \qquad$ claim X_t is a martingale.

$\mathbb{E}[|X_t|] \le c\sqrt{t} < \infty \quad \forall t \in \mathbb{N}$

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = \mathbb{E}\left[\sum_{i=1}^{t+1} Z_i \mid \mathcal{F}_t\right]$$

$$= \mathbb{E}[Z_{t+1} + X_t \mid \mathcal{F}_t]$$

$$= \mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] + X_t$$

$$= X_t$$

Consider $X_t = \sum_{i=1}^{t} Z_i$ as above

$\tau = \min\{t : X_t > \varepsilon\}$ is a stopping time.

$X_t \in \mathcal{F}_t \Rightarrow \mathbb{1}\{t \le \tau\} \in \mathcal{F}_t$

Counter-example 1: Fix N.

Consider $\tau' = \left\{ \min t : \sum_{s=t+1}^{N} X_s < \sum_{s=1}^{t} X_s \right\}$

τ' is <u>not</u> a stopping time b/c $\sum_{s=t+1}^{N} X_s \notin \mathcal{F}_t$

Counter example 2:

$\tau'' = \max\{t \le N : X_t > \varepsilon\}$

Given $\mathcal{F}_t$, I don't know if $X_s > \varepsilon$ for some $s > t$. Not meas. wrt $\mathcal{F}_t$.

τ'' is <u>not</u> a stopping time

9 Sequential statistics and Martingales

Additional material on this section can be found in [Lattimore and Szepesvári, 2020] and [Howard et al., 2018].

Let $X_1, X_2, \dots$ 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 $1 \le t \le n$.

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

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 $\mathbf{1}\{\tau \le t\}$ is $\mathcal{F}_t$ measurable for all $t \in \mathbb{N}$.*

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. Then if $\mathbb{E}[\tau] < \infty$ 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 $\mathbb{E}[X_\tau] \le \mathbb{E}[X_0]$*
- *X_t is a sub-martingale then $\mathbb{E}[X_\tau] \ge \mathbb{E}[X_0]$*

Lemma 8 (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 $\mathbb{P}(\max_{t\in\{1,\ldots,n\}} X_t \geq \epsilon) \leq \mathbb{E}[X_n]/\epsilon$*

Fix n.

$A_n = \{ \sup_{t\leq n} X_t \geq \varepsilon \}$ assume X_t super-martingale.

$\tau = \min\{ (n+1), \min\{ t\leq n : X_t \geq \varepsilon \} \}$

Note: $\tau \leq n \Rightarrow X_\tau \geq \varepsilon \Rightarrow A_n$

$$\mathbb{E}[X_0] \geq \mathbb{E}[X_\tau] \geq \mathbb{E}[X_\tau \mathbb{1}\{\tau\leq n\}]$$
$$\geq \mathbb{E}[\varepsilon \mathbb{1}\{\tau\leq n\}]$$
$$= \varepsilon \mathbb{P}(A_n)$$
$$= \varepsilon \mathbb{P}(\sup_{t\leq n} X_t \geq \varepsilon)$$

$$\Rightarrow \mathbb{P}(\sup_{t\leq n} X_t \geq \varepsilon) \leq \frac{\mathbb{E}[X_0]}{\varepsilon} \quad \forall n \in \mathbb{N}$$

$$\Rightarrow \mathbb{P}(\sup_{t\in\mathbb{N}} X_t \geq \varepsilon) \leq \frac{\mathbb{E}[X_0]}{\varepsilon}$$

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 ineqaulity we have for any $N \in \mathbb{N}$ that

Claim: M_t is a sub-martingale.

$M_t = \exp(\lambda S_t)$

$\mathbb{E}[M_{t+1} | \mathcal{F}_t] = \mathbb{E}[\exp(\lambda S_{t+1}) | \mathcal{F}_t]$

$= \mathbb{E}[\exp(\lambda Z_{t+1}) \exp(\lambda S_t) | \mathcal{F}_t]$

$= \exp(\lambda S_t) \mathbb{E}[\exp(\lambda Z_{t+1}) | \mathcal{F}_t]$

exp is convex so we apply Jensen's

$\geq \exp(\lambda S_t) \exp(\mathbb{E}[\lambda Z_{t+1}])$

$\geq \exp(\lambda S_t)$

$\mathbb{E}[Z_i] = 0.$

M_t is sub-martingale, thus by maximal ineq.

$$\mathbb{P}\left(\max_{t=1,\dots,n} M_t > \varepsilon\right) \le \frac{\mathbb{E}[M_n]}{\varepsilon}$$

$$M_n = \exp(\lambda S_n) = \exp\left(\lambda \sum_{i=1}^{n} Z_i\right)$$

$$\mathbb{E}[M_n] = \mathbb{E}[\exp(\lambda Z_1)]^n \quad \text{(Independence)}$$

$$\le \exp(\lambda^2 n/2) \quad \text{(Hoeffding)}$$

-

$$\mathbb{P}\left(\max_{t=1,\dots,n} S_t \ge \sqrt{2n\log(1/\delta)}\right)$$

$$= \mathbb{P}\left(\max_{t=1,\dots,n} \underbrace{\exp(\lambda S_t)}_{M_t} \ge \underbrace{\exp\left(\lambda\sqrt{2n\log(1/\delta)}\right)}_{\varepsilon}\right)$$

$$\le \frac{\mathbb{E}[\exp(\lambda n)]}{\exp\left(\lambda\sqrt{2n\log(1/\delta)}\right)} = \exp\left(\lambda^2 n\tfrac{1}{2} - \lambda\sqrt{\;}\right)$$

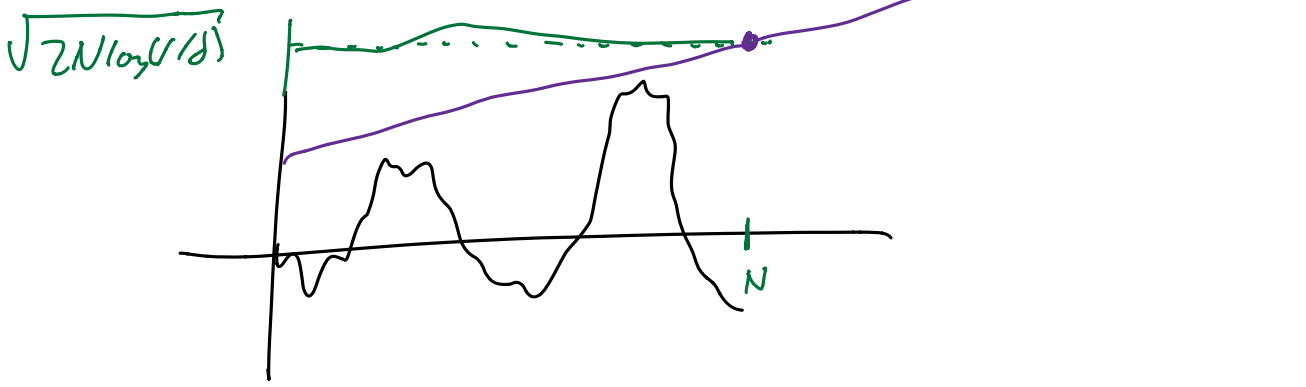
$$\le \delta$$

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 ineqaulity we have for any $N \in \mathbb{N}$ that

$$\begin{aligned}\mathbb{P}(\max_{t\in\{1,\ldots,N\}} S_t \geq \sqrt{2N\log(1/\delta)}) &= \mathbb{P}(\max_{t\in\{1,\ldots,N\}} \exp(\lambda S_t) \geq \exp(\lambda\sqrt{2N\log(1/\delta)})) \\ &\leq \exp(-\lambda\sqrt{2N\log(1/\delta)})\mathbb{E}[\exp(\lambda S_N)] \\ &\leq \exp(-\lambda\sqrt{2N\log(1/\delta)})\exp(\lambda^2 N/2)\end{aligned}$$

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*.

$$\Rightarrow \quad \mathbb{P}\left(\bigcup_{t=1}^{N}\left\{\sum_{i=1}^{t} Z_i > \sqrt{2N\log(1/\delta)}\right\}\right) \leq \delta$$



Recall: from 1st slide $\mathbb{P}\left(\sum_{i=1}^{N} Z_i > \sqrt{2N\log(1/\delta)}\right) \leq \delta$

Accomplishing a bound that holds $t = 1, \ldots, N$

but not paying for a union bound.

Example: Linear boundary crossing 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

$$\mathbb{E}[M_{t+1}(\lambda) \mid \mathcal{F}_t] = \mathbb{E}[\exp(\lambda S_{t+1} - (t+1)\lambda^2/2) \mid \mathcal{F}_t]$$

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

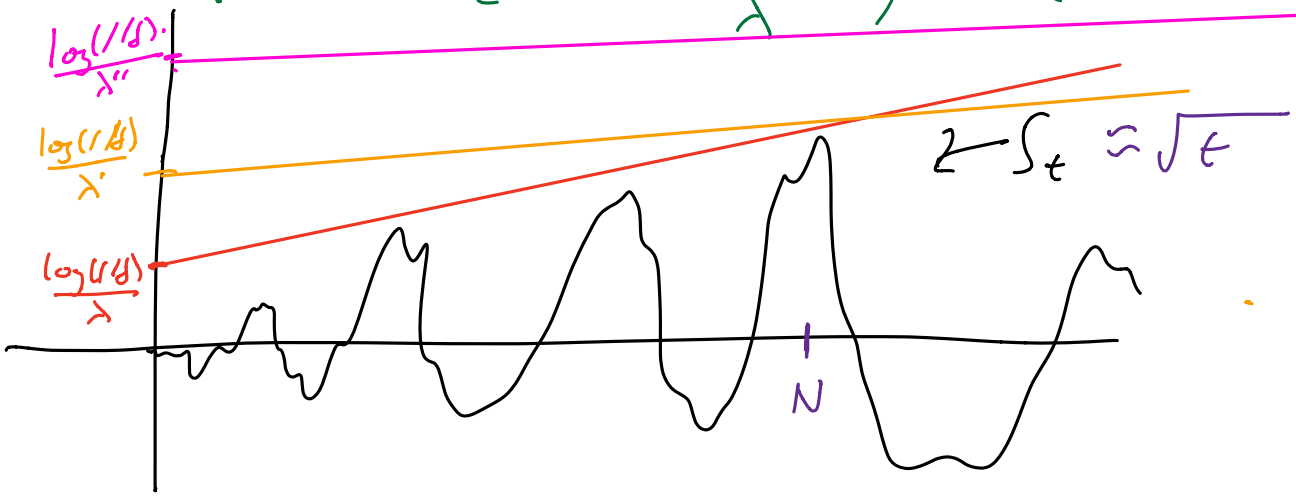
$$\leq 1 \cdot M_t(\lambda)$$

$\Rightarrow$ $M_t(\lambda)$ is supermartingale.

Maximal ineq

$$\mathbb{P}(\exists t : M_t(\lambda) > 1/\delta) \leq \frac{\mathbb{E}[M_0(\lambda)]}{1/\delta} = \delta$$

$$\mathbb{P}\left(\exists t : S_t \geq t\lambda/2 + \frac{\log(1/\delta)}{\lambda}\right) = \mathbb{P}(\exists t : M_t(\lambda) > \tfrac{1}{\delta})$$



Each λ defines a linear boundary in which S_t crosses w.p. $\leq \delta$.

Example: Linear boundary crossing 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

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

~~Let $\tau = \mathbf{1}\{M_t(\lambda) \geq 1/\delta\}$~~. Then by Doob's optional stopping theorem

$$\begin{aligned}\mathbb{P}(\exists t \in \mathbb{N} : S_t \geq t\lambda/2 + \log(1/\delta)/\lambda) &= \mathbb{P}(\exists t \in \mathbb{N} : M_t(\lambda) \geq 1/\delta) \\ &= \mathbb{P}(M_\tau(\lambda) \geq 1/\delta) \\ &\leq \delta \mathbb{E}[M_\tau(\lambda)] \\ &\leq \delta\end{aligned}$$

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

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

a strict improvement over the maximal inequality!

Example: Curved boundaries with a mixing distribution Let $Z_1, Z_2, \ldots$ 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

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

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

Q: If I just looked at $\bar{M}_t$ would I know it's a supermartingale?

If X_t is a martingale and ϕ is convex (concave) then $\phi(X_t)$ is a submartingale (super)

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

← prove supermartingale directly?

~~Using the same logic as above, if $\tau = \mathbb{1}\{\min_t : \bar{M}_t \geq 1/\delta\}$ then~~ Applying maximal inequality for supermartingales

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

~~$= \mathbb{P}(\bar{M}_\tau \geq 1/\delta)$~~

$\leq \delta.$

$\mathbb{E}[\bar{M}_0] = 1$

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

Compare to naive union bound $\sqrt{2t\left(\log(1/\delta) + \log(2t^2)\right)}$

Example: Curved boundaries with a mixing distribution Let $Z_1, Z_2, \ldots$ 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

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

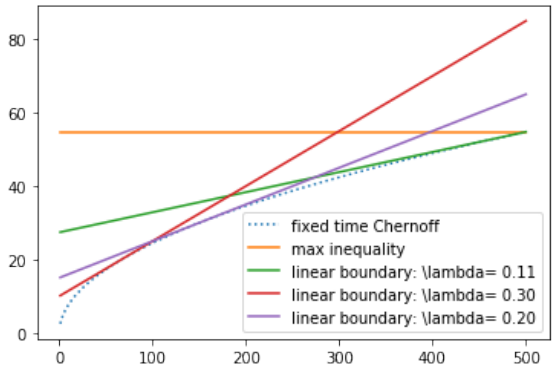
Suppose we take $h(\lambda) = \frac{1}{\sqrt{2\pi\nu^2}} e^{-\lambda^2/2\nu^2}$. Then

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

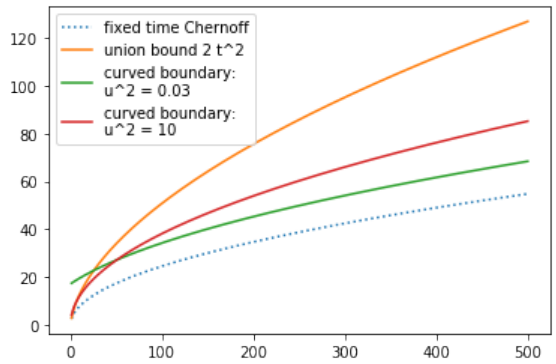
Using the same logic as above, if $\tau = \mathbf{1}\{\min t : \bar{M}_t \geq 1/\delta\}$ then

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

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



(a) Fix $\delta = 0.05$. The 'fixed time Chernoff' represents $\sqrt{2t\log(1/\delta)}$ which holds at each t but not all $t \leq 500$ simultaneously (which is why it is dotted). The 'max inequality' holds for all $t \leq 500$, 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 $\sqrt{2\log(2t^2/\delta)}$.

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 exampes.

If you choose $\{\lambda_i\}_{i=1}^{\infty}$ "smartly" so that each applies to a region $[t_i, 2t_i)$

union bound $\mathbb{P}\left(\bigcup_i \left\{ M_t(\lambda) > \frac{2i^2}{\delta} \right\}\right) \leq \delta$ } stitching

Can show that this implies

$$|S_t| \leq \sqrt{c\, t \left(\log(1/\delta) + \log\log(t)\right)}$$

Law of the iterated Logarithm. (LIL)

(LIL)

Theorem Let Z_i be iid mean-0 w/ variance σ^2 then

$$\limsup_t \frac{S_t}{\sqrt{2\sigma^2 t \log\log t}} = 1 \quad \text{where} \quad S_t := \sum_{i=1}^{t} Z_i$$

A sequence H_t is predictable wrt $\mathcal{F}_t$ if $H_t \in \mathcal{F}_{t-1}$.

Example: 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 $\mathbb{E}[\exp(\lambda Z_t)|\mathcal{F}_{t-1}] \leq \exp(\lambda\sigma_t^2/2)$. Define $S_t = \sum_{i=1}^t Z_i$ and $V_t = \sum_{i=1}^t \sigma_i^2$. Then $M_t = \exp(\lambda S_t - \lambda^2 V_t/2)$ is a super-martingale. Thus, $\mathbb{P}(\exists t \in \mathbb{N} : S_t \geq \lambda V_t/2 + \log(1/\delta)/\lambda) \leq \delta$. Note that "time" t does not appear anywhere in this bound explicitly, and has been replaced by V_t.

Example: Vector-valued martingales Now suppose $Z_1, Z_2, \cdots \in \mathbb{R}^d$ is a $\mathcal{F}_t$-adapted random sequence that satisfies $\mathbb{E}[\exp(\langle \lambda, Z_t \rangle) | \mathcal{F}_{t-1}] \leq \exp(\|\lambda\|^2_{\Sigma_t}/2)$ 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 $M_t(\lambda) = \exp(\langle \lambda, S_t \rangle - \|\lambda\|^2_{V_t}/2)$ 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

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

then repeating the same steps as above we conclude that

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

Example: Vector-valued martingales Now suppose $Z_1, Z_2, \cdots \in \mathbb{R}^d$ is a $\mathcal{F}_t$-adapted random sequence that satisfies $\mathbb{E}[\exp(\langle \lambda, Z_t \rangle)|\mathcal{F}_{t-1}] \leq \exp(\|\lambda\|^2_{\Sigma_t}/2)$ 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 $M_t(\lambda) = \exp(\langle \lambda, S_t \rangle - \|\lambda\|^2_{V_t}/2)$ 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

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

then repeating the same steps as above we conclude that

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

Example: 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 $\mathbb{E}[\exp(s\eta_t)|\mathcal{F}_{t-1}] \le \exp(s^2/2)$ 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

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

$\sum_t = x_t x_t^\top$

Thus, Equation 1 holds for any $\gamma > 0$. Fix some $\gamma > 0$ and define

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

$\|a+b\|_A = \|A^{1/2}a + A^{1/2}b\|_2$
$\le \|A^{1/2}a\|_2 + \|A^{1/2}b\|_2$
$= \|a\|_A + \|b\|_A$

Now notice

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

(triangle ineq)
$V_t \succ 0$

We conclude that

$$\begin{aligned}\mathbb{P}\Big(\exists t : \|\widehat{\theta}_t - \theta_*\|_{(V_t+\gamma I)} \ge \sqrt{\gamma}\|\theta_*\|_2^2 + \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t + \gamma I|)}\Big) \\ \le \mathbb{P}\Big(\exists t : \|S_t\|_{(V_t+\gamma I)^{-1}} \ge \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t + \gamma I|)}\Big) \le \delta.\end{aligned} \tag{2}$$

Let λ_i be ith eigenvector of $V_t + \gamma I$ $\le d \log\left(\frac{Lt}{d\gamma} + 1\right)$

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

$\frac{1}{d}\sum_i \log(\lambda_i) \le \log\left(\frac{1}{d}\sum_i \lambda_i\right) =$
$= \log\left(\frac{1}{d}\mathrm{Tr}(V_t + \gamma I)\right)$
$\le \log\left(\frac{1}{d}(Lt + d\gamma)\right)$

$\|x\|^2_{V+\gamma I} = x^\top (V + \gamma I) x$
$= x^\top V x + \gamma x^\top x$
$\ge \gamma \|x\|_2^2$

if $\|x_t\|_2^2 \le L$

Example: 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

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

Thus, Equation 1 holds for any $\gamma > 0$. Fix some $\gamma > 0$ and define

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

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

$$\begin{aligned}\mathbb{P}\Big(\exists t : \|\widehat{\theta}_t - \theta_*\|_{(V_t+\gamma I)} \geq &\sqrt{\gamma}\|\theta_*\|_2^2 + \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t+\gamma I|)}\Big) \qquad (2) \\ &\leq \mathbb{P}\Big(\exists t : \|S_t\|_{(V_t+\gamma I)^{-1}} \geq \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t+\gamma I|)}\Big) \leq \delta.\end{aligned}$$

How to choose γ?

$R_T \lesssim \gamma\|\theta_*\|_2 + d\sqrt{T\log(\frac{T}{\delta})}$

Typically w/o of knowledge of $\|\theta_*\|_2$, just pick $\gamma = 1$

9.1 Hypothesis testing and Likelihood ratios

The previous section show-cased the *method of mixtures* to generate curved boundaries that random walks will not pass. However, the method seems mysterious and unmotivated. In this section we present an alternative perspective on the same derivations that, at least to me, is quite illuminating.

References

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

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

[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 nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

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

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

[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*.

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

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

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

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

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

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

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

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

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

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