MKTG 596: Adaptive Experimental Design Methods

Lalit Jain

April 19, 2023

1 Introduction

This course is about the algorithms and techniques that drive experimentation in online platforms. At many firms, thousands of experiments are run daily to gather the vast amount of data needed to enable data-driven decision-making. For any one of these experiments, an experimenter faces the challenge of deciding what data to collect, how long to run the experiment, and how to glean insights from the data collected. As a result, there is an increased demand by practitioners for methods and algorithms that deliver statistically sound results faster and with less opportunity cost. This course develops a modern toolbox of experimentation, with a focus on adaptive experimental design. Rooted in classical statistical and machine learning techniques, AED decides what future data to collect based on past measurements in a closed loop. Due to both theoretical gains and empirical success, AED has quickly become one of the most commonly employed algorithmic paradigms in practice with a promise of cutting experimentation time by up to half. However, practitioners who employ AED blindly can easily bias their results, or potentially not collect the data needed to make any useful inferences.

In practice experimental systems need to

- Return results rapidly (minimize sample complexity)
- Reduce opportunity cost (minimize regret)
- Provide valid inference (valid confidence intervals)
- Be robust to time variation
- Effectively incorporate customer heterogeneity

In addition, As we will see in this course, unfortunately there is no master algorithm and practitioners often have to trade-off several competing objectives.

The course is organized as follows:

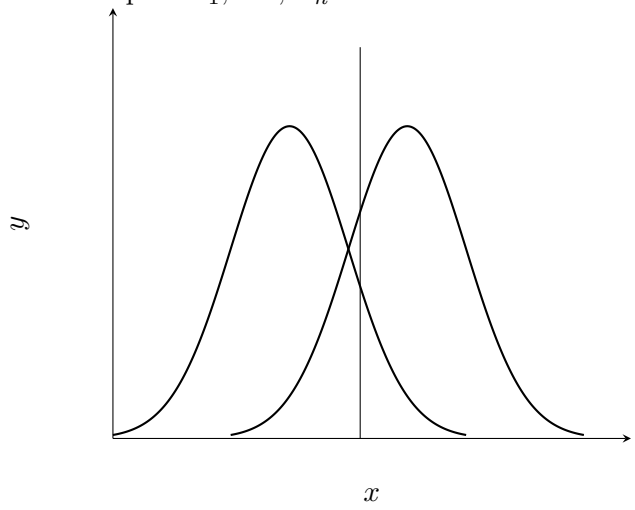
- We begin with a short overview and discussion of the three pillars of experimentation, A/B Testing, Multi-Armed Bandits, and Multiple Hypothesis Testing
- Building upon our experience with SPRT and MAB we will next discuss Anytime-Valid-Inference
- This will be followed by a more in-depth discussion of MAB with a focus on Optimistic Strategies, such as MAB and Thompson Sampling.
- Contextual Bandits

- Interference in Experiments
- Off Policy Evaluation
- Experimental Design
- Orthogonal ML
- Reinforcement Learning

2 Three Pillars

2.1 Pillar 1: A/B Testing

The following captures a common setting in online A/B testing and statistics.
Example. We have a single distribution that is assumed to be Gaussian $N(\mu, 1)$, and we have collected samples $X_1, \cdots, X_n$ from this distribution.



Consider the following hypothesis test:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_0 + \Delta$$

where μ_0 and Δ are known. In practice, μ_0 could be known because the control variant may have been running for many months. The "gap" Δ is known as the *minimum detectable effect* and captures the smallest deviation from the control that we are interested in capturing.

We consider the following test. Choose a threshold τ, if $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i > \tau$ then we will declare for H_1 otherwise we declare for H_0. We want this test to have a type-1 error bounded by α, and a type-2 error bounded by β. That is, the probability we accept the alternative given that the null is true is at most α, and the probability we accept the null given the alternative is at most β.

A natural question is how to choose τ, n to guarantee this result. This is a common problem considered in most introductory statistics courses, and we quickly review it. Define $z_\alpha = \Phi^{-1}(1-\alpha)$. To guarantee that the type-1 error is indeed bounded by α we need

$$\frac{\tau - \mu_0}{\sqrt{1/n}} \geq z_\alpha$$

where Φ is the CDF function for a $N(0,1)$. Similarly for the type-2 error,

$$\frac{\tau - (\mu_0 + \Delta)}{\sqrt{1/n}} \leq -z_\beta$$

See Figure **??**.

Adding these together, we see that it suffices to take

$$n \geq \frac{(z_\alpha + z_\beta)^2}{\Delta^2}$$

Now consider the specific case where $\beta = \alpha = \delta$. Using the fact that $1 - \Phi(t) \leq e^{-t^2/2}$ and that $\Phi(t)$ is monotonically increasing, we can upper bound $z_\delta \leq \sqrt{2\log(1/\delta)}$, so we see that with probability greater than $1 - 2\delta$, if $n \geq 8\log(1/\delta)/\Delta^2$ we will return the correct hypothesis. □

Lower Bounds. The natural question to ask is, how tight is this? Can we do better? In fact, we can prove a lower bound. Recall that the KL-Divergence between two distributions p_0, p_1 (supported on $\mathbb{R}$) is

$$KL(p_0, p_1) = \int \log\left(\frac{p_0(x)}{p_1(x)}\right) p_0(x)dx$$

Theorem 1. *Any hypothesis test Ψ that can distinguish $H_0 : X_1, \cdots, X_n \sim p_0$ and $H_1 : X_1, \cdots, X_n \sim p_1$ has a probability of error lower bounded by*

$$\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\psi = 0)) \geq \frac{1}{4}e^{-nKL(p_0,p_1)}$$

Proof. Without loss of generality, we refer to the

$$\begin{aligned}
\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\psi = 0)) &\geq \frac{1}{2}\left(\mathbb{P}_0(\Psi = 1) + \mathbb{P}_1(\psi = 0)\right)) \\
&= \frac{1}{2}\left(\int_{\psi=1} dp_0 + \int_{\psi=0} dp_1\right) \\
&\geq \frac{1}{2}\int \min(dP_0, dP_1)
\end{aligned}$$

Now note that a) $\int \max(dp_0, dp_1) \leq 2$ (since the max is less than the sum), and b)

$$\int \min(dp_0, dp_1) \int \max(dp_0, dp_1) \geq (\int \sqrt{\min(dp_0, dp_1)\max(dp_0, dp_1)})^2$$

. Thus

$$\begin{aligned}
\max(\mathbb{P}_0(\Psi=1),\mathbb{P}_1(\psi=0)) &\geq \frac{1}{4}(\int \sqrt{dp_0 dp_1})^2 \\
&= \frac{1}{4}(\int dp_0 \sqrt{\frac{dp_1}{dp_0}})^2 \\
&= \frac{1}{4}\exp[2\log\left(\int dp_0 \sqrt{\frac{dp_1}{dp_0}}\right)] \\
&\geq \frac{1}{4}\exp[2\int \log(\sqrt{\frac{dp_1}{dp_0}})dp_0] \qquad \text{(Jensen's inequality)} \\
&= \frac{1}{4}\exp[-\int \log(\sqrt{\frac{dp_0}{dp_1}})dp_0] \\
&= \frac{1}{4}\exp[-\int \log(\frac{dp_0}{dp_1})dp_0] \\
&= \frac{1}{4}\exp(-KL(p_0^n,p_1^n))
\end{aligned}$$

Exercise: Show that $KL(p_0^n,p_1^n) = nKL(p_0,p_1)$ □

For our setting above, $KL(N(\mu,1),N(\mu+\Delta,1)) = \Delta^2/2$ so we see that

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

Thus, our probability of error will be at least δ unless $n \geq 2\log(1/\delta)/\Delta^2$. You will notice that this is a factor of 4 tighter than our upper bound. How can we do better? Answer: Adaptive Experimentation, namely the Sequential Probability Ratio Test.

2.2 The Sequential Probability Ratio Test

Let's return to our hypothesis testing setting.

$$\begin{aligned}
H_0 &: X_1,\cdots,X_n \sim p_0 \\
H_1 &: X_1,\cdots,X_n \sim p_1
\end{aligned}$$

Indeed, denote the expectation and probability measure with respect to $p_i, i=0,1$ as $\mathbb{E}_i$.

Define the *likelihood ratio* $\Lambda_t = \prod_{s=1}^t \frac{p_1(X_s)}{p_0(X_s)}$, $t \leq n$. The SPRT choose two thresholds, γ_0,γ_1 with $\gamma_1 > \gamma_0$. The SPRT stops at a time τ, and outputs H_1 if $\Lambda_\tau > \gamma_1$ and otherwise outputs H_0. Our goal is to set the thresholds to guarantee that Type-1,2 error are bounded by α,β respectively.

To make the following argument precise, we will need to employ the fact that Λ_t forms a Martingale sequence under $\mathbb{P}_{\nvdash}$. Then

$$\mathbb{E}[\Lambda_t|X_1,\cdots,X_{t-1}] = \Lambda_{t-1}\mathbb{E}_0\left[\frac{p_1(X_t)}{p_0(X_t)}\right] = \Lambda_{t-1}$$

Similarly, Λ_t^{-1} forms a martingale sequence under P_1.

Let $x = (X_1, \cdots, X_n)$, and abusing notation, we let $p_0(x) = p(X_1) \cdots p(X_n)$. Again letting α denote our Type 1 error and β to denote our Type 2 error, we compute: we now compute

$$\begin{aligned}
1 - \beta &= \mathbb{P}_1(\Lambda_\tau > \gamma_1) \\
&= \int_{\Lambda_\tau > \gamma_1} p_1(x)dx \\
&= \int_{\Lambda_\tau > \gamma_1} \frac{p_1(x)}{p_0(x)} p_0(x)dx && \text{(Wald's Ratio Identity)} \\
&\geq \gamma_1 \int_{\Lambda_\tau > \gamma_1} p_0(x)dx \\
&= \gamma_1 \alpha
\end{aligned}$$

and similarly

$$\begin{aligned}
1 - \alpha &= 1 - \mathbb{P}_0(\Lambda_\tau > \gamma_1) \\
&= \int_{\Lambda_\tau < \gamma_0} p_0(x)dx \\
&= \int_{\Lambda_\tau < \gamma_0} \frac{p_0(x)}{p_1(x)} p_1(x)dx && \text{(Wald's Ratio Identity)} \\
&\geq \gamma_0^{-1} \int_{\Lambda_\tau < \gamma_0} p_1(x)dx \\
&\geq \gamma_0^{-1} \beta
\end{aligned}$$

The first series of inequalities implies that we should choose

$$\gamma_1 \leq \frac{1-\beta}{\alpha}$$

and similarly

$$\gamma_0 \geq \frac{\beta}{1-\alpha}$$

In general we set γ, γ_1 to be equal to these values.

This calculation demonstrates the trade-off between $\gamma_0, \gamma_1, \alpha, \beta$. Fixing γ_0, we could increase γ_1, which would diminish α however would cause β to increase.

Exercise. Fix α, β and set $\gamma_1 = (1-\beta)/\alpha$ and $\gamma_0 = \beta/(1-\alpha)$. Now imagine a threshold $\gamma_0' = \beta'/(1-\alpha') < \gamma_0$ and $\gamma_1' = (1-\beta')/\alpha' > \gamma_1$. Show that $\alpha' + \beta' < \alpha + \beta$. What does this mean in practice for using a threshold that is slightly smaller than γ_0 or slightly larger than γ_1?

It remains to bound the expected stopping time of this procedure. At the stopping time τ, we can now use Wald's Theorem and the fact that the data are drawn i.i.d. to see,

$$\mathbb{E}_0[\log(\Lambda_\tau)] = \mathbb{E}_0[\tau]\mathbb{E}_0[\Lambda_1] = \mathbb{E}_0[\tau]\mathbb{E}_0[\log(p_1(X)/p_0(X))] = -\mathbb{E}_0[\tau]KL(p_0, p_1)$$

and similarly

$$\mathbb{E}_1[\log(\Lambda_\tau)] = \mathbb{E}_1[\tau]\mathbb{E}_1[\Lambda_1] = \mathbb{E}_1[\tau]\mathbb{E}_1[p_1(X)/p_0(X)] = \mathbb{E}_1[\tau]KL(p_1, p_0)$$

We now compute these expected stopping times in a slightly different way. Ignoring the "overshoot" of the path.

$$\mathbb{E}_0[\log(\Lambda_\tau)] = \mathbb{E}_0[\log(\Lambda_\tau)\mathbf{1}\{\Lambda_\tau \leq \gamma_0\}] + \mathbb{E}_0[\log(\Lambda_\tau)\mathbf{1}\{\Lambda_\tau > \gamma_1\}] \quad \approx \log(\gamma_0)(1-\alpha) + \log(\gamma_1)\alpha$$

and similarly

$$\mathbb{E}_1[\log(\Lambda_\tau)] = \mathbb{E}_1[\log(\Lambda_\tau)\mathbf{1}\{\Lambda_\tau \le \gamma_0\}] + \mathbb{E}_0[\log(\Lambda_\tau)\mathbf{1}\{\Lambda_\tau > \gamma_1\}] \quad \approx \log(\gamma_0)\beta + \log(\gamma_1)(1-\beta)$$

Combining the last four displays, we see

$$\mathbb{E}_0[\tau] \approx \frac{\alpha \log\left(\frac{\alpha}{1-\beta}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right)}{KL(p_0, p_1)}$$

and

$$\mathbb{E}_1[\tau] \approx \frac{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) + \beta \log\left(\frac{\beta}{1-\alpha}\right)}{KL(p_1, p_0)}$$

Example. Now we instantiate for our running A/B testing example. Consider $\alpha = \beta = \delta < .5$, $p_0 = N(\mu, 1)$ and $p_1 = N(\mu + \Delta, 1)$. Then

$$\mathbb{E}_0[\tau] \approx \frac{2\log(1/\delta)}{\Delta^2} \text{ and } \mathbb{E}_1[\tau] \approx \frac{2\log(1/\delta)}{\Delta^2} \tag{1}$$

which matches our previous lower bound!

It's worth thinking about what the test is explicitly is in this case. To make calculations slightly easier, let's set $\mu = 0$. By definition,

$$\begin{aligned}
\prod_{i=1}^n \frac{p_1(X_i)}{p_0(X_i)} &= \prod_{i=1}^n \frac{e^{-(x-\Delta)^2/2}}{e^{-x^2/2}} \\
&= \prod_{i=1}^n e^{\Delta(2x_i - \Delta)/2} \\
&= e^{\Delta S_n - \Delta^2 t/2} && (S_n = \textstyle\sum_{t=1}^n)
\end{aligned}$$

So $\log(\Lambda_n) = \Delta S_n - \Delta^2 t/2$. In particular (assuming $\alpha = \beta = \delta$), the SPRT turns into the following

$$\begin{aligned}
S_\tau &\ge \frac{n\Delta}{2} + \frac{\log((1-\delta)/\delta)}{\Delta} \to \text{return } H_1 \\
S_\tau &\le \frac{n\Delta}{2} - \frac{\log((1-\delta)/\delta)}{\Delta} \to \text{return } H_0
\end{aligned}$$

In particular the optimality of the SPRT shows us that a linear boundary optimally decides between two *known* means. Later on we will see the SPRT as a special case of a more general maximal inequality.

□

There are a couple of key details missing in this argument. Firstly, we need to handle the overshoot. Secondly, we can ask how tight this bound is. For details see Chapter 3 of [TNB14].

Remark.

Algorithm 1 An algorithm with caption

$[K]$
for $t = 1, 2, \cdots$ **do**
  Choose $I_t \in K$
  Observe $r_t = X_{I_t,t}$ where $X_{I_t,t} \sim \nu_{I_t}$
end for

2.3 Pillar 2: Multi-Armed Bandits

In the multi-armed bandit we have K distributions (referred to as *arms*), $\nu_1, \cdots, \nu_K$, and for each arm we can choose a distribution to receive a reward from (*pull*).

We assume that $\mu_i = \mathbb{E}_{X \sim \nu_i}[X]$ is the expectation of the i-th arm. We consider two different goals.

1. **Best-Arm Identification.** Let $i_* = \arg\max_{i \in [K]} \mu_i$. Identify i_* with probability greater than $1 - \delta$ in the fewest number of samples.

2. **Regret Minimization.** The (expected) regret at time n is defined as

$$R_n = \max_{i \le [K]} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right] \qquad = \max_{i \le [K]} \mu_i T - \mathbb{E}[\sum_{t=1}^{n} X_{I_t,t}]$$

   Our goal is to design a procedure that minimizes the regret. Ideally the regret is sub-linear in n. Here we should be careful about what we mean by expectation. The expectation is being taken over all randomness in the rewards *and* the randomness of the algorithm.

   [1]

Example. A natural strategy to try to minimize regret is to pull each arm once, maintain an estimate $\hat{\mu}_{i,t}$ for each arm at each time, and then set $I_t = \arg\max_{i \in [K]} \hat{\mu}_{i,t-1}$. This is often known as a *Greedy* heuristic.

Here is a simple example showing that this can incur linear regret. Imagine a simple example where we have three arms each of which is a Bernoulli distribution with means set to $\mu_1 > \mu_2 > \mu_3$. Imagine a setting where on the pull of arms 1 and 2 we get a reward of 0 and for arm 3 we get a reward of 3. Then, the empirical mean of arm 3 is 1 and for the other arms is 0. Thus in each round after, the empirical mean of arm 3 will be greater than 0, so we will pull it in each round and we will never pull arm 1 or 2 again. This happens with probability, $\mathbb{P}(\bar{\mu}_{3,1} \ge \max_{i=1,2} \bar{\mu}_{i,t-1}) = \mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 1) = (1 - \mu_1)(1 - \mu_2)\mu_3$. Thus in this setting, with some finite probability, we will incur linear regret! We have totally failed to balance exploration and exploitation. □

Let $\mu_* = \max_{i \in [K]} \mu_i$.

Lemma 1. *Define* $\Delta_i = \mu_* - \mu_i$. *Then*

$$R_n = \sum_{t=1}^{n} \Delta_i \mathbb{E}[T_i]$$

where $T_i = \sum_{t=1}^{n} \mathbf{1}\{I_t = i\}$

[1] To be precise, define $\mathcal{F}_t$ as the sigma-algebra induced by the random variables. See [LS20] for details.

Proof.

$$\begin{aligned}
R_n &= \mu_* T - \mathbb{E}\left[\sum_{t=1}^{n} X_{I_t,t}\right] \\
&= \mu_* T - \mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{n} \mathbf{1}\left\{I_t = i\right\} X_{I_t,t}\right] \\
&= \mu_* T - \sum_{i=1}^{K} \mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\left\{I_t = i\right\} X_{I_t,t}\right] \\
&= \mu_* T - \sum_{i=1}^{K} \mu_i \mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\left\{I_t = i\right\}\right] \\
&= \mu_* T - \sum_{i=1}^{K} \mu_i \mathbb{E}[T_i] \\
&= \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i]
\end{aligned}$$

□

The above characterization of regret characterizes the fundamental balance between exploration vs explotiation. We need to pull each arm sufficiently many times to conclude that it is the best, or not, but we incur far too much regret if we give the arm too many pulls.

2.3.1 A quick introduction to concentration.

Given i.i.d random variables $X_1, \cdots, X_t$, we would like to understand how quickly their empirical mean $\bar{X} = \frac{1}{t}\sum_{s=1}^{t} X_i$ converges to the true mean $\mu = \mathbb{E}[X]$. By the Central Limit Theorem (and various moment conditions), defining $Z_t = \sum_{i=1}^{t}(X_i - \mu)$ and denoting $\sigma^2 = \mathrm{var}(X)$

$$\frac{\frac{1}{\sqrt{n}} Z_t}{\sigma} \to N(0,1)$$

Thus we may believe that

$$\mathbb{P}(\bar{X} - \mu > \epsilon) = \mathbb{P}(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{\epsilon}{\sigma/\sqrt{n}}) \leq 1 - \Phi(\frac{\epsilon}{\sigma/\sqrt{n}}) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}$$

However, unfortunately, this is far from true for any finite number of samples.

Example. Consider $X_1, \cdots, X_n \sim \mathrm{Ber}(p)$ so that $\sum_{i=1}^{t} X_i \sim \mathrm{Bin}(n,p)$ Asymptotic theory suggests the following α-confidence intervals on the mean

$$\bar{X} - z_{\alpha/2}\frac{\bar{X}(1-\bar{X})}{n} \leq p \leq \bar{X} + z_{\alpha/2}\frac{\bar{X}(1-\bar{X})}{n}$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

But it's easy to see that fails.

□

In practical applications we are making *decisions based on uncertainty quantification*, so we need something much tighter that holds in any finite time horizon.

Theorem 2 (Markov's Inequality)**.** *Let X be a positive random variable and $\gamma > 0$, then $\mathbb{P}(X > \gamma) \leq \frac{\mathbf{E}[X]}{\gamma}$.*

Proof.

$$\begin{aligned}
\mathbb{P}(X > \gamma) &= \int_{\gamma}^{\infty} dp(x) \\
&= \frac{1}{\gamma} \int_{\gamma}^{\infty} \gamma dp(x) \\
&\leq \frac{1}{\gamma} \int_{\gamma}^{\infty} x dp(x) \\
&\leq \frac{1}{\gamma} \int_{0}^{\infty} x dp(x) \\
&= \frac{\mathbb{E}[X]}{\gamma}
\end{aligned}$$

□

Definition 1. *Given a random variable, X, let $\psi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}]$ for all $\lambda \geq 0$ and define $\psi^*(t) = \sup_{\lambda \geq 0} \lambda t - \psi_X(\lambda)$.*

Lemma 2 (Cramer-Chernoff Trick)**.** *Let X be a random variable. Then*

$$\mathbb{P}(X \geq \gamma) \leq e^{-\psi_Z^*(t)}$$

Proof.

$$\begin{aligned}
\mathbb{P}(X \geq \gamma) &= \mathbb{P}(e^{\lambda X} \geq e^{\lambda \gamma}) \\
&\leq e^{-\lambda\gamma} \mathbb{E}[e^{\lambda X}] && \text{(Markov's Inequality)} \\
&= e^{-\lambda\gamma} e^{\log \mathbb{E}[e^{\lambda X}]} \\
&= e^{-(\lambda\gamma - \psi(\lambda)} \\
&\leq \sup_{\lambda \geq 0} e^{-(\lambda\gamma - \psi(\lambda)} \\
&= e^{-\inf_{\lambda \geq 0}(\lambda\gamma - \psi(\lambda))} = e^{-\psi_*(\gamma)}
\end{aligned}$$

□

Definition 2. *We say that a random variable X with is σ^2-subGaussian, if $\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2\sigma^2/2}$.*

- If $X \sim N(0, \sigma^2)$ then $\mathbb{E}[e^{\lambda X}] = e^{\lambda^2\sigma^2/2}$
- If X is a random variable bounded in the interval $[a, b]$ then $\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\lambda^2(b-a)^2/8}$. In particular, a $\text{Ber}(p)$ random variable is 1-subGaussian.

  Note that this does not account for the variance! To do so, we have to be a bit more careful and consider *sub-exponential* random variables. More on this later.

Let $Z_n = \sum_{t=1}^{n}(X_n - \mu)$ where $X_1, \cdots, X_n$ are i.i.d. samples, $\mathbb{E}[X_t] = \mu$, and X_t is σ^2-subGaussian. Then

$$\mathbb{E}[e^{\lambda Z}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda X_i}] \leq e^{n\lambda^2\sigma^2/2}.$$

And

$$\begin{aligned}\psi_Z^*(t) &= \inf_{\lambda \geq 0} \lambda\gamma - \psi_Z(t) \\ &\geq \inf_{\lambda \geq 0} \lambda\gamma - n\lambda^2\sigma^2/2 && \text{(Set } \lambda = \gamma/n\sigma^2\text{)} \\ &\geq \frac{\gamma^2}{2n\sigma^2}\end{aligned}$$

In particular, this immediately implies that w.p. $\geq 1-\delta$,

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \gamma) = \mathbb{P}(\sum_{i=1}^n (X_i - \mu) \geq n\gamma) \leq e^{\frac{-n\gamma^2}{2\sigma^2}}$$

Setting the left-hand side less than some failure probability δ, we see that with probability $\geq 1-\delta$

$$\frac{1}{n}\sum_{i=1}^n X_i - \mu \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}$$

Using an identical argument (check!) on $\mathbb{P}(Z < -\gamma) = \mathbb{P}(-Z > \gamma)$, we also have that with probability greater than $1-\delta$,

$$\mu - \frac{1}{n}\sum_{i=1}^n X_i \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}$$

Thus we can conclude the following two-sided inequality (which we state as a theorem).

Theorem 3. *Let $X_1, \cdots, X_n$ be i.i.d. σ^2-subGaussian random variables with mean μ. Then with probability greater than $1-\delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} \tag{2}$$

Proof. By the previous, the upper and lower bounds each fail with probability at most $1\delta/2$. So the probability that either fail is at most $\delta/2 + \delta/2 = \delta$.

Let's make this more formal. Let $\mathcal{E}_1 = \mathbf{1}\left\{\bar{X}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}\right\}$ and $\mathcal{E}_2 = \mathbf{1}\left\{\bar{X}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}\right\}$. Then

$$\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2) \leq \frac{\delta}{2} + \frac{\delta}{2} \leq \delta$$

□

2.3.2 Explore Than Commit

Let's exercise some of our concentration knowledge. We assume that each ν_i is 1-subGaussian and we consider the following strategy.

Define $\hat{\mu}_{i,n} = \frac{1}{T_i}\sum_{t=1}^n \mathbf{1}\{I_t = i\} X_{i,t}$ (i.e. the empirical mean of the i-th arm).

Theorem 4.

$$R_n \leq \tau \sum_{i=1}^K \Delta_i + (n - K\tau)\sum_{i=1}^K \Delta_i e^{-\tau\Delta_i^2/4}$$

Algorithm 2 An algorithm with caption

$[K]$, τ, n
for $i = 1, \cdots, K$ **do**
  Pull Arm i τ times.
end for
Define $\hat{\mu}_{i,n} = \frac{1}{T_i} \sum_{t=1}^{n} \mathbf{1}\{I_t = i\} X_{i,t}$
Pull arm $\hat{i} = \arg\max \hat{\mu}_{i,n}$ for the rest of time, $t \in [K\tau, n]$.

Proof. From the above

$$\begin{aligned} R_n &= \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i] \\ &= \tau \sum_{i=1}^{K} \Delta_i + (n - K\tau)\mathbb{E}\left[\sum_{i=1}^{K} \Delta_i \mathbf{1}\left\{\hat{i} = i\right\}\right] \\ &= \tau \sum_{i=1}^{K} \Delta_i + (n - K\tau) \sum_{i=1}^{K} \Delta_i \mathbb{P}(\mathbf{1}\left\{\hat{i} = i\right\}) \end{aligned}$$

Now

$$\begin{aligned} \mathbb{P}(\hat{i} = i) &\leq \mathbb{P}(\hat{\mu}_i \geq \hat{\mu}_*) \\ &\leq \mathbb{P}(\hat{\mu}_i - \hat{\mu}_* \geq 0) \\ &\leq \mathbb{P}(\hat{\mu}_i - \mu_i - (\hat{\mu}_* - \mu_*) \geq \mu_* - \mu_i) \\ &\leq \mathbb{P}((\hat{\mu}_i - \mu_i) - (\hat{\mu}_* - \mu_*) \geq \mu_* - \mu_i) \qquad \text{(This is } 2/\tau\text{-subGaussian)} \\ &\leq e^{-\tau\Delta_i^2/4} \end{aligned}$$

from which the result follows. □

Let's consider the case when $K = 2$, and assume $\mu_* = \mu_1 > \mu_2$. We can further bound this as follows:

$$R_n \leq \tau\Delta + n\Delta e^{-\tau\Delta^2/4}$$

This expression kind of tells us how to choose τ. By taking $\tau = \lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \rceil$, we see that the regret is bounded as

$$R_n \leq \min\{n\Delta, \Delta + \frac{4}{\Delta} \log\left(\frac{n\Delta^2}{4}\right)\}$$

Exercise. Show this choice of τ minimizes the regret. How do you interpret τ from our previous perspective of A/B testing?

This is an *instance-dependent* bound for this algorithm - it scales logarithmically in T! Certainly, sub-linear regret. Setting $\Delta = \sqrt{4\log(4n)/n}$ we actually see that

$$R_n \leq O(\sqrt{n \log(n)}).$$

This is a minimax or worst-case bound.

These bounds scale sub-linearly with T, but depend on knowledge of Δ. What if we don't know Δ? Let's go back to the case of K arms, then

$$\begin{aligned} R_n &\leq \tau K \Delta_{\max} + nK \exp(-\tau \Delta_{\min}^2/4) \\ &\leq \tau K \Delta_{\max} + \frac{nK}{\sqrt{\tau \Delta_{\min}^2/4}} \\ &\leq \tau K \Delta_{\max} + \frac{2nK}{\Delta_{\min}\sqrt{\tau}} \end{aligned}$$

If $\tau = n^{2/3}$, we have that

$$R_n \leq K\Delta_{\max} n^{2/3} + 2K\frac{n^{2/3}}{\Delta_{\min}} = O(n^{2/3}).$$

2.3.3 Elimination

Algorithm 3 An algorithm with caption

$[K]$, τ, n
for $i = 1, \cdots, K$ **do**
  Pull Arm i τ times.
end for
Define $\hat{\mu}_{i,n} = \frac{1}{T_i}\sum_{t=1}^{n} \mathbf{1}\left\{I_t = i\right\} X_{i,t}$
Pull arm $\hat{i} = \arg\max \hat{\mu}_{i,n}$ for the rest of time, $t \in [K\tau, n]$.

Without loss of generality we assume that arm 1 has the highest mean. Define $\Delta_i = \mu_1 - \mu_i, 2 \leq i \leq K$ and $\Delta_1 = \Delta_2$. Define

$$\mathcal{E}_{\rangle,\sqcup} = \{|\hat{\mu}_{i,t} - \mu_i| < \sqrt{\frac{2\log(2Kt^2/\delta)}{t}}\}$$

$$\mathcal{E}_i = \bigcap_{t=1}^{\infty} \mathcal{E}_{i,t}$$

$$\mathcal{E} = \bigcap_{i=1}^{k} \mathcal{E}_i$$

Lemma 3. $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$

Proof. It's easiest to consider the complement. Firstly by Hoeffding's inequality, $\mathbb{P}(\mathcal{E}_{i,t}^c) \leq \frac{\delta}{Kt^2}$. Then the union bound ($\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$) tells us

$$\mathbb{P}(\mathcal{E})_i \leq \sum_{t=1}^{\infty}\sum \mathbb{P}(\mathcal{E}_{i,t}) \leq \sum_{t=1}^{\infty} \frac{\delta}{Kt^2} \leq \frac{2}{K}$$

Then finally

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{i=1}^{K} \frac{\delta}{K} = \delta$$

□

*Remark.*This is known as an *anytime* confidence interval since it bounds the deviation at all times uniformly with high probability. More on this in the next chapter.

Lemma 4. *On $\mathcal{E}$, arm 1 is never eliminated*

Proof. Assume $\mathcal{E}$.

$$\begin{aligned}\text{arm 1 is eliminated} &\iff \exists j, \hat{\mu}_{j,t} - \sqrt{\frac{2\log(2Kt^2/\delta)}{t}} \geq \hat{\mu}_{1,t} + \sqrt{\frac{2\log(2Kt^2/\delta)}{t}} \\ &\iff \mu_j \geq \mu_1\end{aligned}$$

□

Lemma 5. *Define $T_i = \sum_{t=1}^{\infty} \mathbf{1}\{i \in S_t\}$, on $\mathcal{E}$, $T_i \leq \frac{c\log(K\Delta_i^{-1})}{\Delta_i^2}$ for some absolute constant $c > 0$.*

Proof. Assume $\mathcal{E}$. Then for any arm i,

$$\hat{\mu}_{i,t} + \sqrt{\frac{2\log(2Kt^2/\delta)}{t}} \leq \mu_i + 2\sqrt{\frac{2\log(2Kt^2/\delta)}{t}}$$

and

$$\hat{\mu}_{1,t} - \sqrt{\frac{2\log(2Kt^2/\delta)}{t}} \geq \mu_1 - 2\sqrt{\frac{2\log(2Kt^2/\delta)}{t}}$$

Now, we will have that arm j is eliminated by arm 1 if

$$\hat{\mu}_{1,t} - \sqrt{\frac{2\log(2Kt^2/\delta)}{t}} > \hat{\mu}_{i,t} + \sqrt{\frac{2\log(2Kt^2/\delta)}{t}}$$

and the latter will happen at any time when

$$\mu_1 - 2\sqrt{\frac{2\log(2Kt^2/\delta)}{t}} > \mu_i + 2\sqrt{\frac{2\log(2Kt^2/\delta)}{t}}$$

Solving for t, we see that there exists an absolute constant c such that the previous is satisfied when $t > \frac{c\log(K\Delta_i^{-1})}{\Delta_i^2}$. □

Exercise. Show that if $t > 2\log(1/a)/a$, then $at > \log(t)$ for $a > 0$

Theorem 5. *With probability greater than $1-\delta$, Arm 1 is returned at a time τ satisfying*

$$\tau \leq c\sum_{i=1}^{K} \frac{\log(K\Delta_i^{-1}/\delta)}{\Delta_i^2}$$

Proof. On event $\mathcal{E}$

$$\begin{aligned}\tau &= \sum_{t=1}^{\infty}\sum_{i=1}^{K} \mathbf{1}\{i \in S_t\} \\ &= \sum_{i=1}^{K}\sum_{t=1}^{\infty} \mathbf{1}\{i \in S_t\} \\ &= \sum_{i=1}^{K} T_i \\ &\leq c\sum_{i=1}^{K} \frac{\log(K\Delta_i^{-1}/\delta)}{\Delta_i^2}\end{aligned}$$

□

*Remark.*The result of 5 can be improved by a $\log(K)$ [JMNB14]. Matching lower bounds are given in [KCG16].

Let's take a step back, up to constant and log-factors, the above result tells us that the bandit is effectively eliminating each sub-optimal time in a time that is proportional to what would happen in an A/B test if we knew the effect size! It's worth thinking about what a *passive* algorithm which knew each effect size would do. Such a passive algorithm has to choose it's number of samples ahead of time.

It still remans to ask what the opportunity cost of this procedure. The following theorem bounds the regret.

Theorem 6. *After any number of pulls T, the regret is bounded as $O(\sqrt{KT\log(T/\delta)})$.*

Proof. Say we have had T total pulls. For any $\gamma > 0$,

$$\begin{aligned}
R_t &= \sum_{i=2}^{n} \Delta_i T_i(T) && (T_i(T) \text{ denotes the number of pulls up to time } T)\\
&= \sum_{i:\Delta_i \le \gamma}^{n} \Delta_i T_i(T) + \sum_{i:\Delta_i > \gamma} \Delta_i T_i(T)\\
&= \gamma T + \sum_{i:\Delta_i > \gamma} \Delta_i T_i(T)\\
&\le \gamma T + c \sum_{i:\Delta_i > \gamma} \Delta_i \frac{\log(K\Delta_i^{-1}/\delta)}{\Delta_i^2}\\
&\le \gamma T + c \sum_{i:\Delta_i > \gamma} \frac{\log(K\Delta_i^{-1}/\delta)}{\Delta_i}
\end{aligned}$$

Now let's consider a few cases.

- $\gamma = 0$: $R_t \le O\left(\sum_{i=2}^{K} \frac{\log(K\Delta_i^{-1}/\delta)}{\Delta_i}\right)$
- $\gamma = \sqrt{K\log(T/\Delta)/T}$: $R_t \le O\left(\sqrt{KT\log T}\right)$

□

It remains to bound the expected regret, note that $\mathbb{P}(E^c) \le \delta$. So,

$$\begin{aligned}
\mathbb{E}[R_T] &= \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i]\\
&= \sum_{i=1}^{K} \Delta_i \frac{\log(K\Delta_i^{-1}/\delta)}{\Delta_i} + \Delta_i T\mathbb{P}(\mathcal{E})\\
&\le KT\delta + \sum_{i=1}^{K} \Delta_i \frac{\log(K\Delta_i^{-1}/\delta)}{\Delta_i}
\end{aligned}$$

Taking $\delta = 1/T$ gives us a regret of $O(\sum_{i=2}^{K} \frac{\log(KT\Delta_i^{-1})}{\Delta_i})$

2.4 Martingale Digression

For a moment we will be formal. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a measure space. That is Ω is a fixed set, and $\mathcal{A}$ forms a σ-algebra of subsets of Ω.
[2]

Definition 3. *We say that $\mathcal{H} = \{\mathcal{H}_i\}_{i=0}^n$ is a filtration if*

- $\mathcal{H}_0 = \emptyset$
- *Each* $\mathcal{H}_t \subset \mathcal{A}$
- $\mathcal{H}_t \subset \mathcal{H}_{t+1}$

A sequence of random variables $\{X_t\}_{t=1}^\infty$ is adapted to to $\mathcal{H}$ if X_t is $\mathcal{H}_t$ measurable.

Intuitively, the filtration captures all information available to us at each time t. Given a collection of random variables $X_1, \ldots, X_t, \cdots$, i.e. measurable with respect to $\mathcal{A}$. The most interesting filtrations that we will consider in this class are $\mathcal{H}_t = \sigma\{X_s\}_{s=1}^t$ - that is the filtration determined by the random variables. Intuitively, the decisions we make at time t will be dependent on what information we have up to that time.

Definition 4. *Fix some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a filtration $\mathcal{H} = \{\mathcal{H}_t\}_{t=1}^\infty$. We say that an adapted sequence of random variables $\{X_t\}_{t=1}^\infty$ is*

- *A martingale of* $\mathbb{E}[X_t|\mathcal{H}_{t-1}] = X_{t-1}$
- *A super-martingale of* $\mathbb{E}[X_t|\mathcal{H}_{t-1}] \leq X_{t-1}$
- *A sub-martingale of* $\mathbb{E}[X_t|\mathcal{H}_{t-1}] \geq X_{t-1}$

Example 1. Again consider a sequence of random variables $\{X_t\}_{t\geq 1}$ where each $\mathbb{E}[X_{t+1}|X_t] = 0$ (eg iid mean zero random variables). Then define $S_t = X_1 + \cdots + X_t$.

$$\mathbb{E}[S_{t+1}|\{X_t\}_{t=1}^n] = \mathbb{E}[S_t + X_{t+1}|X_t] = S_t$$

so $\{S_t\}_{t\geq 1}$ forms a martingal sequence.
Example 2. Assume that $X_1, \cdots, X_t \sim p_0(X)$ and let p_1 be a different distribution. Then Λ_t forms a martingale sequence under p_0.

Definition 5. *We say that a random variable $\tau \in \mathbb{N} \cup \{\infty\}$ is a stopping time with respect to $\mathcal{H} = \{\mathcal{H}_t\}_{t=1}^\infty$ if $\{\tau = t\} \in \mathcal{H}_t$.*

Intuitively, a stopping time is a decision on whether to stop or proceed dependent on all the information up to and including time t.

INCLUDE INTUITION ON STOPPING TIME BASED ON RANDOM WALK.

Theorem 7 (Optional Stopping)**.** *If*

[2] Recall that a σ-algebra if

- $\Omega, \emptyset \in \mathcal{A}$
- $A \in \mathcal{A} \to A^c \in \mathcal{A}$
- For any countable sequence of sets $\{F_i\}_{i=1}^n \subset 2^\Omega$, $\cup_{i=1}^n F_i \in \mathcal{A}$.

- $\mathbb{P}(\tau < c) = 1$ *for some* c
- *There exists a constant* c *so that* $\mathbb{E}[|X_{t+1} - X_t||\mathcal{F}_t] \leq c$ *and* $\mathbb{E}[\tau] \leq \infty$.
- *If* X_t *forms a martingale,* $\mathbb{E}[X_t] = \mathbb{E}[X_0]$
- *If* X_t *forms a super - martingale,* $\mathbb{E}[X_t] \leq \mathbb{E}[X_0]$
- *If* X_t *forms a sub - martingale,* $\mathbb{E}[X_t] \geq \mathbb{E}[X_0]$

We will use Doob's optimal stopping several times. However, our first application is an important result in it's own right that we will need for the SPRT, Wald's theorem.

Theorem 8 (Wald's Theorem)**.** *Let* $\{X_t\}_{t=1}^{\infty}$ *be i.i.d. random variables with* $\mathbb{E}[X_t] = \mu$*. Assume that* τ *with* $\mathbb{E}[\tau] \leq \infty$ *is a stopping time wrt the filtration* $\mathcal{F}_t = \{X_s\}_{s=1}^{t}$*. Then* $\mathbb{E}[\sum_{s=1}^{\tau} X_s] = \mu\mathbb{E}[\tau]$

Proof. Define the martingale $M_n = \sum_{i=1}^{n} X_i - \mathbb{E}[X_i]$. By the optional stopping theorem

$$\mathbb{E}[M_\tau] = \mathbb{E}[S_\tau - T_\tau] = \mathbb{E}[M_0] = 0$$

Exercise. Show that it is valid to apply the second condition of the optional stopping theorem.

Subtracting $\mathbb{E}[T_\tau] = \mu\mathbb{E}[\tau]$ now gives the result. □

Theorem 9 (Ville's Inequality)**.** *Let* $\{X_t\}_{t=1}^{\infty}$ *be a family of random variables adapted to the filtration* $\mathcal{F}_t$ *with* $X_t \geq 0$ *almost surely. If* X_t *forms a super-martingale,* $\mathbb{P}(\max_{t\geq 1} X_t \geq \epsilon) \leq \frac{\mathbb{E}[X_0]}{\epsilon}$

Proof. Define the stopping time $\tau(t) = \min\{t+1, \min\{s \leq t : X_t \geq \epsilon\}\}$ and the event $A_t = \{\sup_{s\leq t} X_t \geq \epsilon\}$. Clearly $\tau(t)$ is a bounded stopping time. Then by Doob's optimal stopping

$$\begin{aligned}\mathbb{E}[X_0] &\geq \mathbb{E}[X_{\tau(t)}] \\ &\geq \mathbb{E}[X_{\tau(t)}\mathbf{1}\{\{\}\tau \leq t\}] \\ &\geq \epsilon\mathbb{P}(\tau \leq t).\end{aligned}$$

Rearranging shows that $\mathbb{P}(\sup_{s\leq t} X_t \geq \epsilon) \leq \frac{\mathbb{E}[X_0]}{\epsilon}$ for all $t \geq 1$. However note that $A_1 \subset A_2 \subset A_3 \subset \cdots$ so $\mathbb{P}(\sup_{t\geq 1} X_t > \epsilon) \leq \frac{\mathbb{E}[X_0]}{\epsilon}$. □

2.5 Always-Valid Confidence Intervals

Given a sequence of mean-zero independent random variables $X_1, X_2, \cdots$, an α-*always-valid* confidence interval is a sequence of bounds $c(t, \alpha)$ such that

$$\mathbb{P}\left(\left|\frac{1}{t}\sum_{s=1}^{t} X_i\right| > c(t,\alpha)\right) < \delta.$$

As we saw above, for 1-subGaussian random variables, $c(t,\alpha) = \sqrt{2\log(2t^2/\delta)/t}$ forms an anytime-confidence bound. The key to developing anytime confidence intervals is given by the following maximal inequailities for martingales.

Line-Cross Inequalities. Assume that $X_1, X_2, \cdots$, is a sequence of i.i.d. centered 1-subGaussian random variables. Define $S_t = \sum_{s=1}^{t} X_s$ and $M_t = \exp\left(\lambda S_t - \frac{t\lambda^2}{2}\right), t \geq 1, M_0 = 1$ Then,

$$\mathbb{E}[M_t|M_{t-1}] = M_{t-1}\mathbb{E}\left[\exp\left(\lambda X_t - \frac{\lambda^2 t}{2}\right)|M_{t-1}\right] \leq M_{t-1}.$$

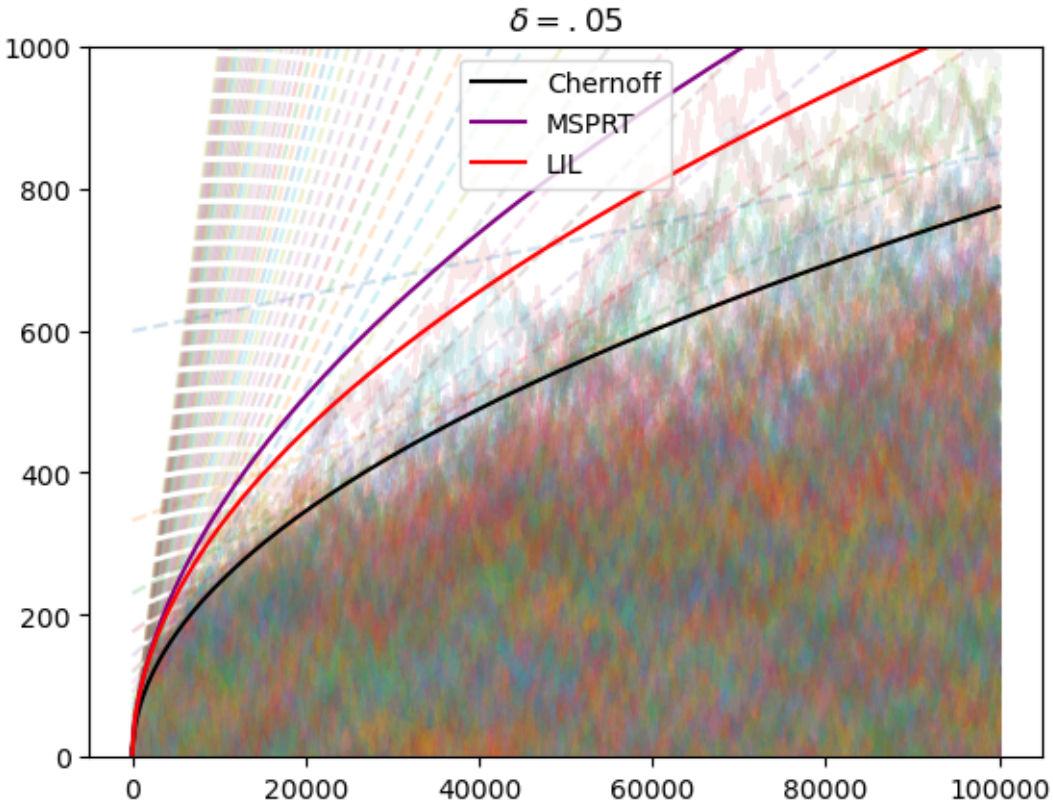


Figure 1: Chernoff, MSPRT, and LIL boundaries. Note that the Chernoff boundary is the pointwise minimum at each time of the set of linear boundaries.

Then by Ville's inequality,

$$\mathbb{P}(\max_{t\in\mathbb{N}} M_t \geq \delta) \leq 1/\delta$$
$$\iff \mathbb{P}\left(\max_{t\in\mathbb{N}} S_t \geq \frac{\lambda t}{2} + \frac{\log 1/\delta}{\lambda}\right) \leq \delta.$$

Mixture Sequential Probability Ratio Test

Let's connect this back to the SPRT. The (one-sided) SPRT (at least in the case of a mean-0 Gaussian) of $H_0 : \mu = 0$ vs $H_1 : \mu = \Delta$, guaranteed that with probability greater than $1 - \delta$, if there exists a time t such that

$$S_t \geq \frac{\lambda\Delta}{2} + \frac{\log 1/\delta}{\Delta}$$

we should return H_1. We also saw this test was optimal in the sense that if H_1 is true the expected hitting time of this boundary is minimized.

There is a nice geometric connection between the family of linear-crossing boundaries given above and the fixed-time Chernoff bound $\sqrt{2\log(1/\delta)/t}$. Namely, the line $\Delta t/2 + \log(1/\delta)/t$ is tangent to the Chernoff at precisely the expected stopping time $t = 2\log(1/\delta)/\Delta^2$!

The previous example established a linear boundary that is always-valid. As we discusses, the fixed-time Chernoff-bound we saw above is the minimum of these linear bounds at each time t - but unfortunately, it doesn't form an always-valid confidence interval.

If we are testing a null-hypothesis versus an alternative, and we suspect that the alternative is at Δ, then we would use the linear boundary $\ell(\Delta, \delta)$. If instead, we believed it was in a set of fixed effects $\Delta_1, \cdots, \Delta_n$, we could consider using multiple lines to define a boundary. More general, we may take a prior distribution over the various lines. The MSPRT makes this concrete.

Let $p(\lambda)$ be a distribution over λ and define

$$M_t = \int M_t(\lambda) dp(\lambda).$$

Then $\{M_t\}_{t=0}^{\infty}$ defines a super-martingale.

$$\begin{aligned}
\mathbb{E}[M_t|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\int M_t(\lambda) dp(\lambda)|\mathcal{F}_{t-1}\right] \\
&= \int \mathbb{E}[M_t(\lambda)|\mathcal{F}_{t-1}] dp(\lambda) \\
&\leq \int M_{t-1}(\lambda) dp(\lambda) \\
&= M_{t-1}.
\end{aligned}$$

A reasonable choice of $h(\lambda) = \frac{1}{\sqrt{2\pi\rho^2}} e^{-\lambda^2/2\rho^2}$. Then

$$\begin{aligned}
M_t &= \frac{1}{\sqrt{2\pi\rho^2}} \int e^{\lambda S_t - \frac{t\lambda^2}{2} - \frac{\lambda^2}{2\rho^2}} d\lambda \\
&= \frac{1}{\sqrt{2\pi\rho^2}} \int e^{\lambda S_t - \frac{\lambda^2}{2}(t+1/\rho^2)} d\lambda \\
&= \frac{1}{\sqrt{2\pi\rho^2}} \int e^{\frac{S_t^2}{2(t+\rho^{-2})} - \frac{(S_t(t+\rho^{-2})^{-1} - \lambda)^2}{2(t+\rho^{-2})}} d\lambda \\
&= \frac{1}{\sqrt{2\pi\rho^2}} \int e^{\frac{S_t^2}{2(t+\rho^{-2})} - \frac{(S_t(t+\rho^{-2})^{-1} - \lambda)^2 (t+\rho^{-2})}{2}} d\lambda \\
&= \frac{\sqrt{(t+\rho^{-2})^{-1}}}{\sqrt{2\pi(t+\rho^{-2})^{-1}\rho^2}} e^{\frac{S_t^2}{2(t+\rho^{-2})}} \int e^{-\frac{(S_t(t+\rho^{-2})^{-1} - \lambda)^2 (t+\rho^{-2})}{2}} d\lambda \\
&= \sqrt{\frac{\rho^{-2}}{t+\rho^{-2}}} e^{\frac{S_t^2}{2(t+\rho^{-2})}}
\end{aligned}$$

Now by Ville's Inequality,

$$\mathbb{P}(\exists t, \log(M_t) > \log(1/\delta)) \leq \delta$$

which is equivalent to the statement that with probability greater than $1 - \delta$

$$|S_t| \leq \sqrt{\frac{t+\rho^{-2}}{\rho^{-2}} \log\left(\frac{t+\rho^{-2}}{\rho^{-2}\delta^2}\right)}$$

Taking $\rho = 1$ we see that

$$\mathbb{P}\left(\exists t : |S_t| \geq \sqrt{(t+1)\log\left(\frac{t+1}{\delta^2}\right)}\right) \leq \delta$$

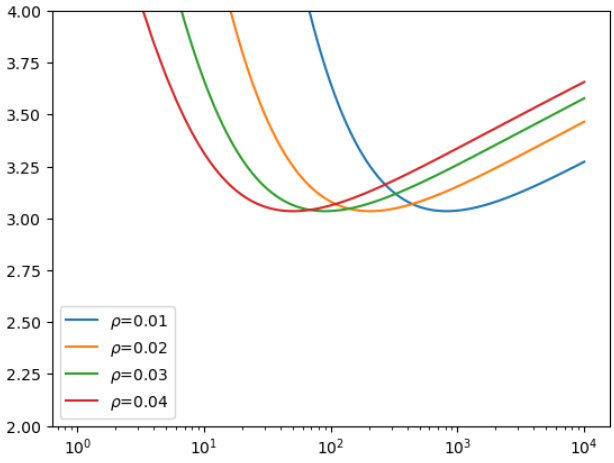


Figure 2: $u(t)/\sqrt{t}$ for various values of ρ

How do we choose ρ ? Essentially the idea is given by Figure **??**. Setting $u(t) = \sqrt{\frac{t+\rho^{-2}}{\rho^{-2}} \log\left(\frac{t+\rho^{-2}}{\rho^{-2}\delta^2}\right)}$, we have plotted $u(t)/\sqrt{t}$. Effectively, considering the normalized process $S_t/\sqrt{t}$, we want to choose ρ so that the resulting boundary $u(t)/\sqrt{t}$ is minimized at an appropriate time t_0. In practice, this means that if you have the budget for a two week experiment, choose set t_0 appropriately and then solve for the value of ρ which results in a curve pinching at t_0. See [HRMS21] for more details.

Law of Iterated Logarithm.

A natural question to ask is how to choose the tightest possible anytime confidence interval. As the discussion on the SPRT (and the MSPRT above) demonstrates, this question is not necessarily that meaningful - in practice we want to choose anytime intervals that are tight when we need them to be. A bound may be tight at a period of two weeks, but then be extremely loose after.

Nevertheless, it's still reasonable to ask for the tightest possible anytime confidence interval *asymptotically*. The answer to this is given by the law of the iterated logarithm.

Theorem 10. *Let $X_1, X_2, \cdots$, be i.i.d. mean zero 1-subGaussian random variables. Then*

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2t\log\log t}} = 1$$

almost surely.

In general the LIL does not hold in finite time with a constant of 2. However, several works have provided finite time versions with slightly worse constants. The best one that I know of is from [HRMS21] which shows,

$$\mathbb{P}(\exists t \leq 10^20 : S_t \geq 1.7\sqrt{t(\log\log(et) + 3.46)}) \leq 0.025$$

Anytime Inference in Linear Regression.

Imagine that we have collected a dataset $(x_1, y_1), \cdots, (x_n, y_n) \subset \mathbb{R}^d \times \mathbb{R}$ where $y_s = x_s^\top \theta + \epsilon_s$ where we assume ϵ_s is mean-zero 1-sub-Gaussian noise. We also assume that $\{x_s\}_{s=1}^t$ form

a *fixed-design*, that is, we are assuming that the choice of x_s is independent of the choice of $(x_1, y_1), \cdots, (x_{s-1}, y_s)$.

For a minute we pretend that ϵ_s is exactly Gaussian noise with variance 1, and we assume that $V_t = \sum_{s=1}^{t} x_s x_s$ is invertible. Then recall the least-squares estimator,

$$\hat{\theta}_\lambda = \arg\min_{\theta' \in \mathbb{R}^d} \|Y - X\theta'\|_2^2, \hat{\theta} = V_t^{-1} X^\top y$$

where $X = [x_1, \cdots, x_n]^\top \in \mathbb{R}^{n \times d}$ and $Y = [y_1, \cdots, y_n] \in \mathbb{R}^n$.

We have that

$$\begin{aligned}
\hat{\theta} &= V_t^{-1} X^\top y \\
&= V_t^{-1} X^\top (X\theta + \epsilon) \\
&= \theta + V_t^{-1} X^\top \epsilon \\
&= \theta + N(0, V_t^{-1})
\end{aligned}$$

Thus $V_t^{1/2}(\hat{\theta} - \theta) \sim N(0, I)$ so that $\|\hat{\theta} - \theta\|_V \sim \chi_d^2$. We can now appeal to the Cramer-Chernoff method, note that for a χ_d^2 random-variable,

Let's try this again. Define the filtration $\mathcal{F}_t = \sigma(\{x_1, y_1, \cdots, x_{t-1}, y_{t-1}, x_t\})$ and assume that ϵ_t is 1-subGaussian conditional on $\mathcal{F}_{t-1}$, that is, $\mathbb{E}[e^{\lambda \epsilon_s} | \mathcal{F}_{t-1}] \leq e^{\lambda^2/2}$.

As above, let the least squares estimator by

$$\begin{aligned}
\hat{\theta} &= (V_t + \gamma I)^{-1} X^\top y \\
&= (V_t + \gamma I)^{-1} V_t \theta + (V_t + \gamma I)^{-1} \sum_{s=1}^{t} x_s \epsilon_s \\
&= (V_t + \gamma I)^{-1} V_t \theta + (V_t + \gamma I)^{-1} S_t
\end{aligned}$$

Now at the same time

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

So the main game in town is S_t - if we could indeed conclude that the x_s's were independent, standard concentration bounds would kick in and we could easily bound it as above. However, we are allowing our sequences of x_s's to be adaptive so we need a much more careful analysis. We will employ the ideas we used for the MSPRT.

To proceed, firstly note that for any $\lambda \in \mathbb{R}^d$ that $M_t(\lambda) = \exp\left(\langle \lambda, S_t \rangle - \frac{\|\lambda\|_{V_t}^2}{2}\right)$, where $V_t = \sum_{s=1}^{t} x_s x_s^\top$, is a super-Martingale. Then Ville's inequality automatically implies a linear boundary in this case, guaranteeing that $\mathbb{P}(\sup_{t \geq} \log(\bar{M}_t) \geq \log(1/\delta)) \leq \delta$

Exercise. Check that M_t forms a super-martingale.

Now we place a distribution $N(0, \gamma^{-1}I)$ over our choice of λ. As we did with the MSPRT, we will now consider the Martingale $\bar{M}_t = \frac{1}{\sqrt{2\pi\gamma^{-d}}} \int M_t(\lambda) e^{-\frac{\|\lambda\|^2}{2\gamma^{-1}}} d\lambda$. We compute,

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

Plugging this into our confidence inequality above, gives that with probability greater than $1-\delta$, for all $t \geq 0$

$$\|S_t\|_{(V_t+\gamma I)^{-1}} \leq \sqrt{2\log(1/\delta) + \log \frac{|V_t + \gamma I|}{\gamma^d}}$$

It remains to bound the second term under the square root.

$$\begin{aligned}
\frac{|V_t + \gamma I|}{\gamma^d} &= \frac{1}{\gamma^d} \prod_{i=1}^{d} (\lambda_i + \gamma) \\
&\leq \left(\frac{\sum_{i=1}^{d} \lambda_i + \gamma d}{\gamma} \right)^d \\
&= \left(\frac{\mathrm{trace}(V_t) + \gamma d}{\gamma} \right)^d \\
&= \left(\frac{tL + \gamma d}{\gamma} \right)^d
\end{aligned}$$

We summarize the above in a Theorem.

Theorem 11. *Fix $\delta \in (0,1)$. Assume that we have a sequence of data $x_1, y_1, \cdots, x_t, y_t, \cdots$. Let $\mathcal{F}_{t-1} = \sigma(\{x_1, y_1, \cdots, x_{t-1}, y_{t-1}, x_t\})$ and assume for all $t \geq 1$*

- *x_t is $\mathcal{F}_{t-1}$ measurable and $\|x_t\|2 \leq 2$*
- *$y_t = \theta_*^\top x_t + \epsilon_t$*
- *η_t is conditionally subGaussian, $\mathbb{E}[e^{\lambda \epsilon_t} | \mathcal{F}_{t-1}] \leq e^{\lambda^2/2}$.*

If $\hat{\theta}_t = (V_t + \gamma I)^{-1} S_t$ with $V_t = \sum_{t=1}^{n} x_t x_t^\top$ and $S_t = \sum_{t=1}^{n} x_t \epsilon_t$ then

$$\|\hat{\theta} - \theta_*\|_{V_t+\gamma I} \leq \sqrt{\gamma}\|\theta\|_2 + \sqrt{2\log(1/\delta) + d\log\left(\frac{tL + \gamma d}{\gamma}\right)}$$

3 Linear Bandits

This section is based on Chapter 19 of [LS20]. Consider the following setting. At each time t we receive a set of arms $\mathcal{A}_t \subset \mathbb{R}^d$ at which point we choose an arm $a_t \in \mathcal{A}_t$ and receive a reward $r_t = \langle \theta_*, a_t \rangle + \epsilon_t$ where ϵ_t is assumed to be 1-subGaussian noise. Our goal is to minimize the regret,

$$R_n \leq \sum_{t=1}^{n} \max_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle - \langle a_t, \theta_* \rangle$$

This linear bandit setting is extremely general.

Example. If $\mathcal{A}_t = \{e_1, \cdots, e_d\}$ at each time, then we recover the multi-armed bandit setting of before.

Example. We can also model more general e-commerce settings. Assume at each time a customer arrives to our platform and we we receive a context vector $c_t \in \mathcal{C}$ representing them. We also have a set of items $\mathcal{A}_t \subset \mathcal{A}$ which could represent a set of products or ads that we need to choose between them to show. The reward can represent a click, conversion or purchase. In particular, we also assume that we have a featurization map $\psi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$ so that the reward at time t is given by $r_t = \langle \psi(c_t, a_t) \rangle + \epsilon_t$. Clearly choosing a set of products to minimize regret is equivalent to the linear bandit model above.

Example.

We now describe our algorithms. Given the data up to time t, $(A_1, r_1, \cdots, a_{t-1}, r_{t-1})$, a regularization parameter $\lambda > 0$ and an upper bound B on $\|\theta_*\|_2$, construct the confidence interval based on Theorem **??**,

$$\beta_t = \sqrt{\lambda} B + \sqrt{2 \log(1/\delta) + d \log(\frac{d\lambda + t}{d\lambda})}$$

Algorithm 4 Optimism if the Face of Uncertainty

Define $V_0 = \lambda I$
for $t = 1, \cdots,$ **do**
  Receive $\mathcal{A}_t$
  Compute the confidence set

$$C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{V_{t-1}} \leq \beta_t\}$$

  Let $a_t = \arg\max_{a \in \mathcal{A}_t, \theta \in C_t} \langle a, \theta \rangle$
  Receive reward r_t.
  Update $V_t = V_{t-1} + a_t a_t^\top$ and $\hat{theta} a_{t+1} = V_t^{-1} \sum_{s=1}^{t} r_t a_t$.
end for

Theorem 12. *Assume that* $\max_{a \in \mathcal{A}_t} |\langle \theta_*, a \rangle| \leq 2, \forall t \geq 1$. *With probability greater than* $1 - \delta$, *the regret of OFUL over* n *steps is bounded by*

$$\begin{aligned} R_n &\leq 2\sqrt{8n\beta_n \log\left(\frac{nL + \gamma d}{\gamma}\right)} \\ &\leq O(d\sqrt{n} \log(nL/\delta)) \end{aligned}$$

Proof. We begin by defining the event

$$\mathcal{E} = \bigcup_{t=1}^{\infty} \{\theta_* \in \mathcal{C}_t\}$$

By Theorem 11, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Throughout the following abbreviate $V_t(\gamma) = V_t + \gamma I$.

Let $a_t^* = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle$ be the true best arm in round t. Then in any round t

$$\begin{aligned}
\langle \theta_*, a_t^* - a_t \rangle &= \langle \theta_*, a_t^* \rangle - \langle \theta_*, a_t \rangle \\
&\leq \max_{\theta \in C_t} \langle \theta, a_t \rangle - \langle \theta_*, a_t \rangle && \text{(Optimism)} \\
&= \max_{\theta \in C_t} \langle \theta - \theta_*, a_t \rangle \\
&= \max_{\theta \in C_t} \|\theta - \theta_*\|_{V_{t-1}(\gamma)} \|a_t\|_{V_{t-1}^{-1}(\gamma)} \\
&\leq 2\beta_t \|a_t\|_{V_{t-1}^{-1}} && (\|\theta - \theta_*\|_{V_{t-1}(\gamma)} \leq \|\theta - \theta_*\|_{V_{t-1}(\gamma)} + \|\theta - \theta_*\|_{V_{t-1}(\gamma)})
\end{aligned}$$

Note that we also have (by assumption) $\langle \theta_*, a_t^* - a_t \rangle \leq 2$. We can now bound the regret

$$\begin{aligned}
R_n &= \sum_{t=1}^{n} \langle \theta_*, a_t^* - a_t \rangle \\
&\leq \sqrt{n \sum_{t=1}^{n} \langle \theta_*, a_t^* - a_t \rangle^2} \\
&\leq \sqrt{2\beta_t^2 n \sum_{t=1}^{n} 1 \wedge \|a_t\|^2_{V_{t-1}(\gamma)^{-1}}}
\end{aligned}$$

We now need to bound the second term. Firstly note that

$$\begin{aligned}
|V_t(\gamma)| &= |V_{t-1}(\gamma) + a_t a_t^\top| \\
&= |V_{t-1}^{1/2}(\gamma)(I + V_{t-1}^{-1/2}(\gamma) a_t a_t^\top V_{t-1}^{-1/2}(\gamma)) V_{t-1}^{1/2}(\gamma)| \\
&= |V_{t-1}(\gamma)|(1 + \|a\|^2_{V_{t-1}^{-1}(\gamma)})
\end{aligned}$$

Using the fact that $x \leq 2\log(1 + x)$

$$\begin{aligned}
\sum_{t=1}^{n} 1 \wedge \|a_t\|^2_{V_{t-1}^{-1}} &\leq 2 \sum_{t=1}^{n} \log(1 + \|a_t\|^2_{V_{t-1}^{-1}}) \\
&\leq 2 \sum_{t=1}^{n} \log(\frac{|V_t|}{V_{t-1}}) \\
&\leq 2 \log(\frac{|V_n|}{|\gamma I|}) \\
&\leq 2d \log\left(\frac{nL + \gamma d}{\gamma}\right)
\end{aligned}$$

Thus we see that the regret of the algorithm is roughly $O(d\sqrt{n}\log(n))$ □

4 Thompson Sampling

We will continue discussing Linear Bandits and the setting of above, but we will now switch algorithms to Thompson Sampling. Developed in 1933, Thompson Sampling is one of the oldest multi-armed bandit algorithms and it is widely used in production systems due to it's ease of implementation, versatility, and good regret guarantees. Indeed one of the main downsides of our OFUL framework is that explicitly computing the best action can be very computationally challenging.

Unlike our previous methods which forced exploration or relied on exploration from confidence bounds, Thompson Sampling relies on exploration from a Bayesian posterior. Intuitively, if we are very uncertain about some action, the posterior will be large and Thompson sampling will explore more.

Let's fix notation - we will try to be more general to exploit the power of TS. We refer to a Bayesian bandit environment as a tuple $(\Theta, \Pi, \mathcal{A}, \mathcal{F}, P)$, where

- Θ is an underlying parameter or index set (implicitly with an underlying σ-algebra)[3]
- Π, the prior, is a measure on Θ
- $\mathcal{A}$ is a set of actions
- $\mathcal{F} = \{f_\theta : \mathcal{A} \to \mathbb{R} | \theta \in \Theta\}$
- $P(\theta, a), \theta \in \Theta, a \in \mathcal{A}$ is a distribution.

At each time the agent selects an action $A_t \in \mathcal{A}_t$ and observes a reward R_t. More precisely,

- At the start of the game, we assume $\theta \sim \Theta$
- We assume that the learner plays probabilistically according to a policy $\pi = \{\pi_t\}_{t=1}^{\infty}$ where given the history $\mathcal{H}_t = (A_1, R_1, \cdots, A_{t-1}, R_{t-1})$, $\pi_t(\mathcal{H})_t$ is a distribution over $\mathcal{A}$. I.e., $\mathbb{P}(A_t = a|\mathcal{H}_t) = \pi_t(a)$.
- Finally, $\mathbb{P}(R_t = \cdot|\mathcal{H}_{t-1}, A_t, \theta) = P(\theta, A_t)(\cdot)$ and $\mathbb{E}[R_t|\mathcal{H}_{t-1}] = f_\theta(A_t)$

The goal of the learner is to minimize their Bayesian Regret, that is

$$BR(T, \pi, \Pi) = \mathbb{E}_{\theta \sim \Pi}[\sum_{t=1}^{T} \max a \in \mathcal{A} f_\theta(a) - f_\theta(A_t)]$$

Remark. **Comparison between Normal Regret and Bayesian Regret.**

Example. Bernoulli Bandits.

Example. Linear Bandits. In that case, we assume that $\mathcal{A} \subset \mathbb{R}^d$ and that $f_\theta(a) = \theta^\top a$.

As we will see, Thompson Sampling enjoys excellent regret bounds in many settings.

[3] I will not be too careful here. For a very precise study of the measure theoretic issues see, Chapter 34 of [LS20].

Algorithm 5 Thompson Sampling

Input: $\mathcal{F}, \mathcal{A}, \Pi$
$\Pi_0 \to \Pi$
for $t = 1, \cdots,$ **do**
  Sample $\theta_t \sim \Pi_{t-1}$
  Set $A_t = \arg\max_{a \in \mathcal{A}} f_{\theta_t}(a)$
  Receive reward R_t
  Update $\Pi_{t+1} = \mathbb{P}(\theta \in \cdot | \mathcal{H}_t)$
end for

4.1 Linear Bandits Revisited

Let's now specialize in the case of Linear Bandits. Recall from Theorem **??**, that on some event $\mathcal{E}$, with $\mathbb{P}(\mathcal{E}) \geq 1 - 1/n$, we have that

$$\|\hat{\theta}_t - \theta\| \leq \beta$$

where $\beta = 1 + \sqrt{s \log(n) + d \log\left(\frac{d + nS^2L^2}{d}\right)}$ (we have taken $\delta = 1/n$ and $\gamma = 1/S$.

Based on this confidence interval, we can define a UCB on each arm at time t,

$$\beta_t(a) = \langle \hat{\theta}_{t-1}, a \rangle + \beta \|a\|_{V_{t-1}^{-1}}$$

By the guarantees of the confidence bound,

$$\begin{aligned}
BR_n &= \mathbb{E}[\sum_{t=1}^{n} \langle a_* - a_t, \theta \rangle] \\
&= \mathbb{E}[\mathbf{1}\{\mathcal{E}\} \sum_{t=1}^{n} \langle a_* - a_t, \theta \rangle] + \mathbb{E}[\mathbf{1}\{\mathcal{E}^c\} \sum_{t=1}^{n} \langle a_* - a_t, \theta \rangle] \\
&= \mathbb{E}[\mathbf{1}\{\mathcal{E}\} \sum_{t=1}^{n} \langle a_* - a_t, \theta \rangle] + 2\mathbb{E}[\mathbf{1}\{\mathcal{E}\} \sum_{t=1}^{n} \langle a_* - a_t, \theta \rangle] + 2
\end{aligned}$$

Let's go after this first term by using a trick. Firstly note, that under the posterior, A_* and A_t are identically distributed! Indeed recall that A_* is a random variable given by $A_* = \arg\max \langle A, \theta_* \rangle$ where θ_* is drawn from Π. Conditioned on the history, the distribution of θ_* at time t is precisely the posterior, so conditioned on the history, A_t is distributed identically as A_*.

Thus, we have that $\mathbb{E}[\beta_t(A_t) | \mathcal{H}_{t-1}] = \mathbb{E}[\beta_t(A_*) | \mathcal{H}_{t-1}]$. Continuing, we have,

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{\mathcal{E}\} \langle a_* - a_t, \theta \rangle | \mathcal{H}_{t-1}] &= \mathbb{E}[\mathbf{1}\{\mathcal{E}\} (\langle a_*, \theta \rangle - \beta_t(a_*) + \beta_t(a_*) - \langle a_t, \theta \rangle) | \mathcal{H}_{t-1}] \\
&= \mathbb{E}[\mathbf{1}\{\mathcal{E}\} (\langle a_*, \theta \rangle - \beta_t(a_*) + \beta_t(a_t) - \langle a_t, \theta \rangle) | \mathcal{H}_{t-1}] \\
&\leq \mathbb{E}[\mathbf{1}\{\mathcal{E}\} (\beta_t(a_t) - \langle a_t, \theta \rangle) | \mathcal{H}_{t-1}] \\
&\leq \mathbb{E}[\mathbf{1}\{\mathcal{E}\} (\langle a_t, \hat{\theta}_{t-1} - \theta \rangle + \beta \|A_t\|_{V_{t-1}^{-1}}) | \mathcal{H}_{t-1}] \\
&\leq \mathbb{E}[\mathbf{1}\{\mathcal{E}\} (\langle a_t, \hat{\theta}_{t-1} - \theta \rangle + \beta \|A_t\|_{V_{t-1}^{-1}}) | \mathcal{H}_{t-1}]
\end{aligned}$$

References

[HRMS21] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. 2021.

[JMNB14] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[KCG16] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[TNB14] Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.