MKTG 596: Adaptive Experimental Design Methods

Lalit Jain

March 29, 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:

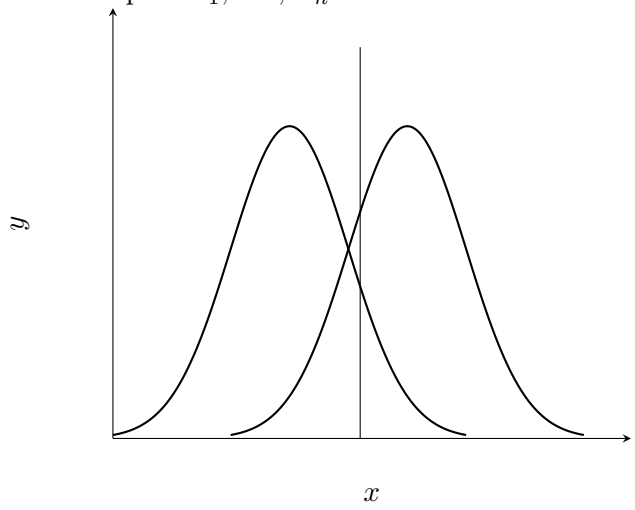
- We begin with a short overview and discussion of the three pillars of experimentation, A/B Testing, Multi-Armed Bandits, and Multiple Hypothesis Testing
- Building upon our experience with SPRT and MAB we will next discuss Anytime-Valid-Inference
- This will be followed by a more in-depth discussion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

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

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

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

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.

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

□

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 $\mathcal{E}_1 = \mathbf{1}\{\bar{X}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}\}$ and $\mathcal{E}_2 = \mathbf{1}\{\bar{X}_n - \mu \geq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}\}$. 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

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

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

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

References

[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.