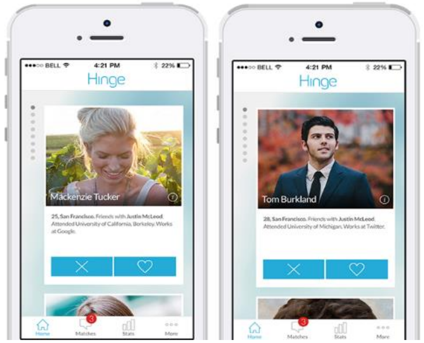
Adaptivity and Confounding in Multi-Armed Bandit Experiments

Daniel Russo and Chao Qin

Columbia University

Efficient interactive learning



Academic literature on bandit algorithms

- Pricing [Ferria et al, 2018] [Javanmard et. al, 2019]
- Recommendations [Li et al, 2010]…
- Personalized medicine [Bastani & Bayati, 2020], Susan Murhy's lab…
- Clinical trials [Villar, 2015][Chick et. al, 2020] [Aziz, 2021]
- A/B/n Testing [Scott, 2010]…
- Public policy experiments [Athey and Wager, 2021] [Kasy, 2021]
- Advertising [Schwartz et al, 2017]

Documented industry applications…

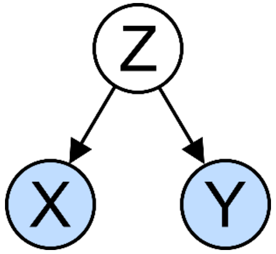
- Adobe, Amazon, Facebook, Google, Linkedin, Netflix, Twitter…

But classical randomized controlled trials (RCTs) are still the standard

This work

Modeling

- Take seriously (some of the) concerns underlying classical RCTs.
- A new twist on models of bandit experiments requiring robustness to delay and nonstationary confounders.



Algorithm design

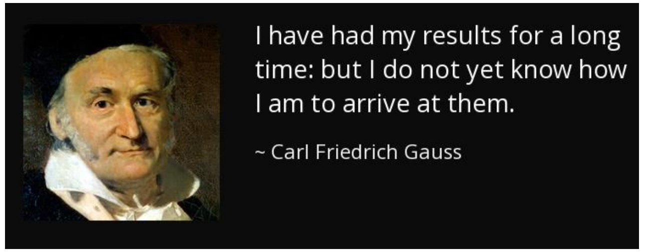
- Propose *deconfounded Thompson sampling*.
- Build on a foundational algorithm, rather than create a solution restricted to our narrow problem formulation.

Algorithm 1: DTS allocation rule in Gaussian best-arm learning

Input prior parameters $(\mu_{1,i}, \Sigma_{1,i})_{i\in[k]}$, population weights X_{pop} and noise variance σ^2.
for $t = 1, 2, \cdots$ **do**
  Sample $\nu_i \sim N(m_{t,i}, s^2_{t,i})$ for $i \in [k]$ and set $I_t^{(1)} = \arg\max_{i\in[k]} \nu_i$;
  do
    Sample $\nu_i \sim N(m_{t,i}, s^2_{t,i})$ for $i \in [k]$ and set $I_t^{(2)} = \arg\max_{i\in[k]} \nu_i$;
  while $I_t^{(1)} = I_t^{(2)}$;
  Flip coin $C_t \in \{0, 1\}$ with bias $\mathbb{P}(C_t = 1) = \beta_t$;
  Play arm $I_t = I_t^{(1)} C_t + I_t^{(2)}(1 - C_t)$;
  Gather delayed observation $o = (I_{t-L}, X_{t-L}, R_{t-L})$.;
  Calculate posterior parameters $m_{t+1,i}$, $s^2_{t+1,i}$ for $i \in [k]$ according to (6) to reflect o;
  Calculate new tuning parameter β_{t+1} if using adaptive tuning;
end

Theory

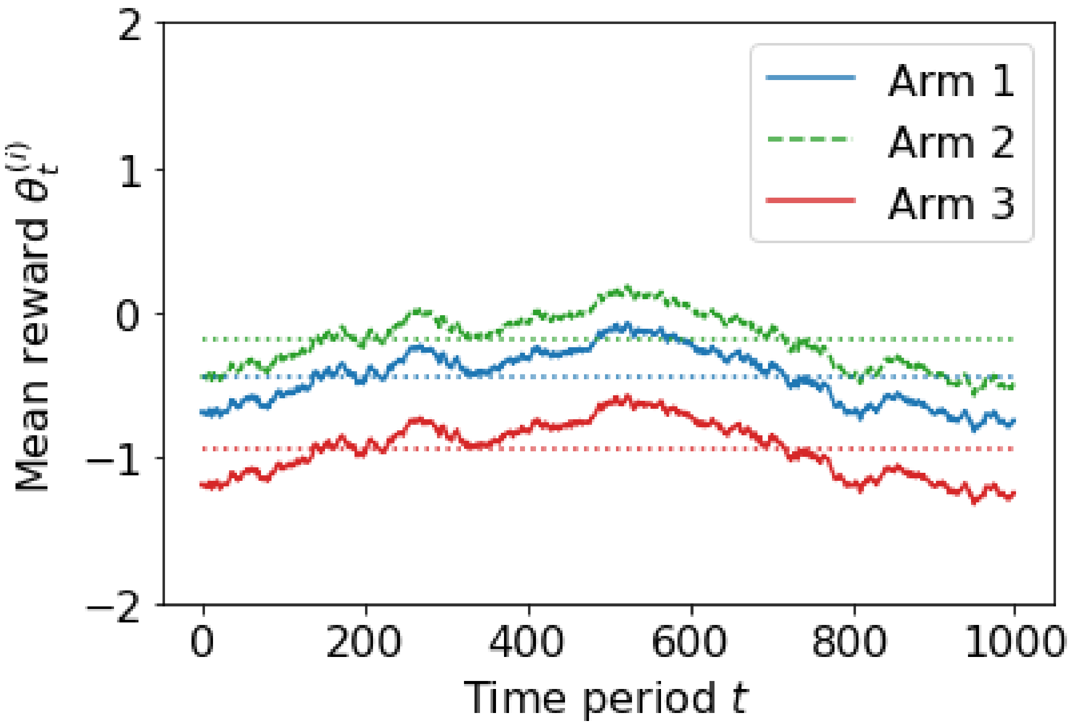
- Robustness in ‘hard’ nonstationary instances
- (Asymptotically optimal) efficiency in ‘easy’ stationary instances.

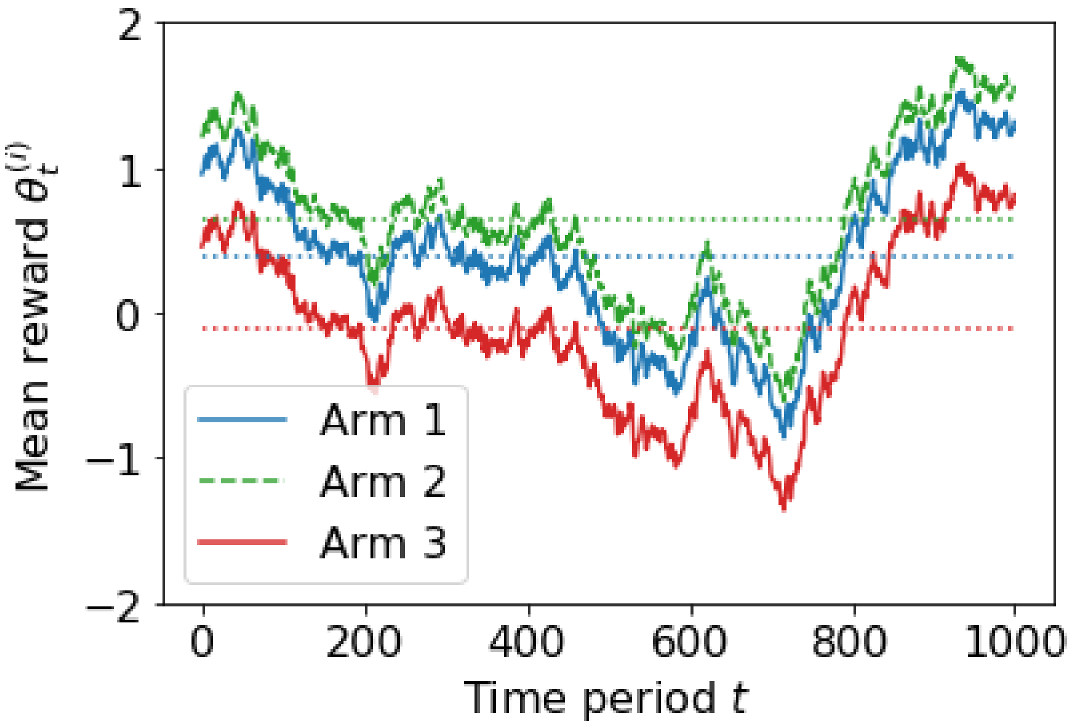


A core tension

Efficiency: Quickly zero-in on the competitive part of the decision-space; Focus most measurement effort on arms 1&2, less on arm 3.

Robustness: Guard against nonstationary confounders by fixing the probability of measuring each arm in each period.

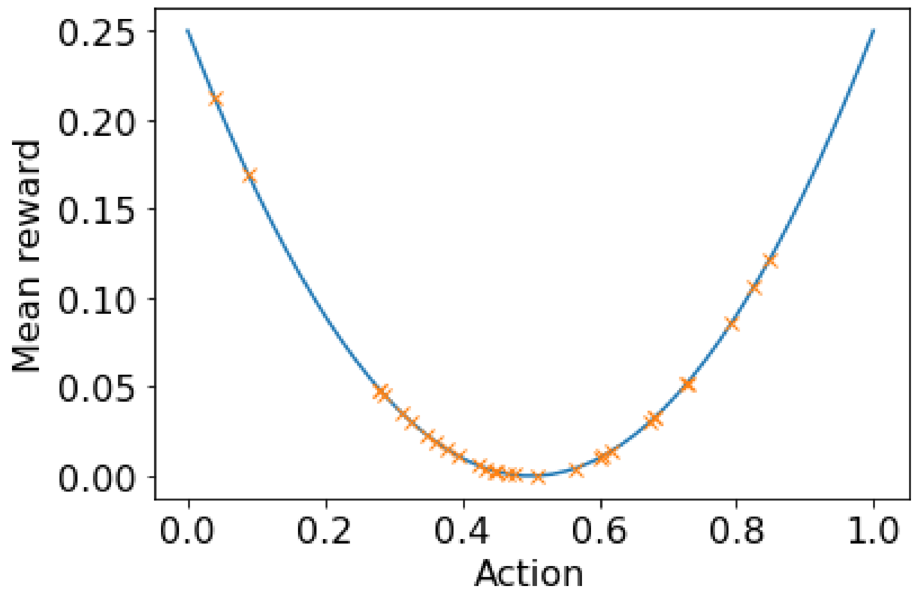




Just give up on adaptivity?

The literature (on bandits, RL, online optimization) offers algorithms that efficiently learn by experimentation to make effective decisions in complex problems.

- Non-adaptive algorithms often require $n \geq \exp\{\text{problem dimension}\}$ samples.



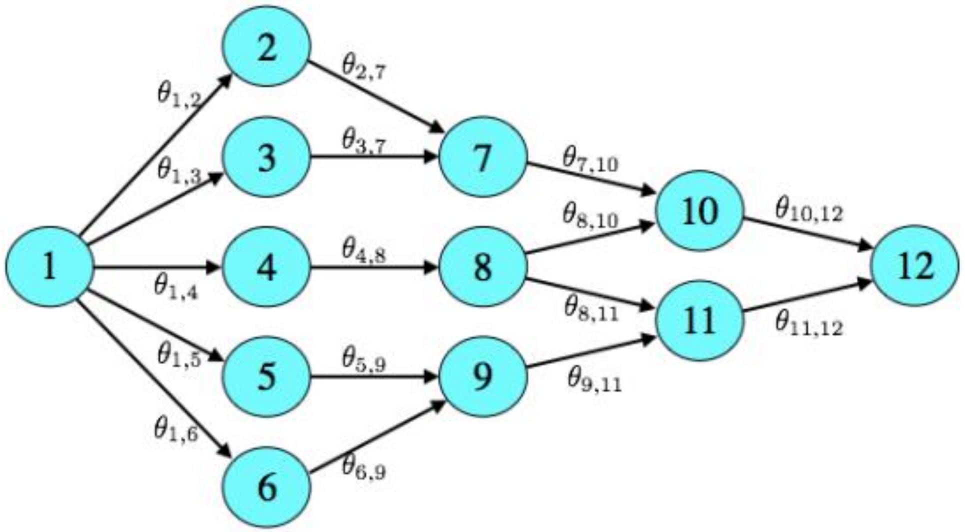


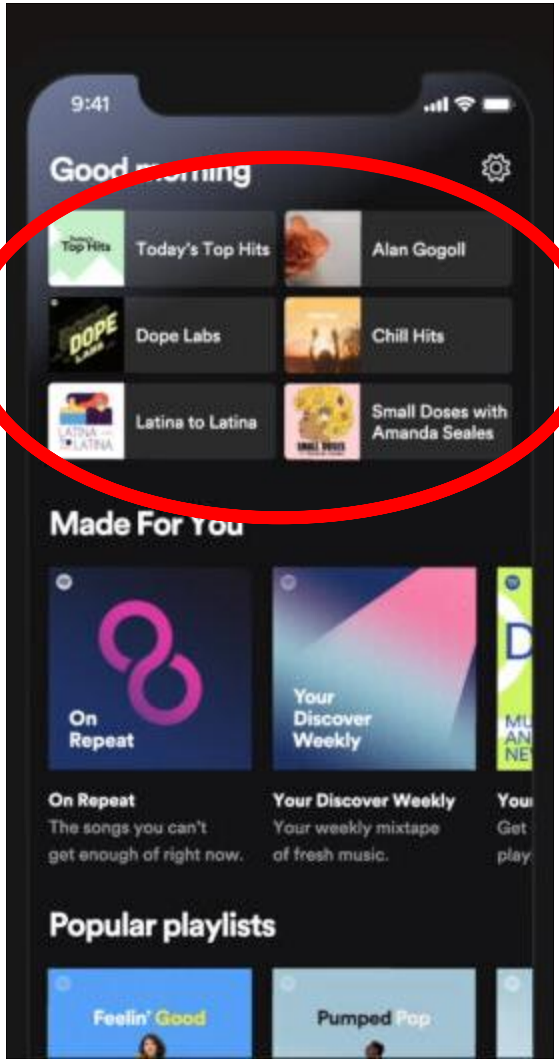
Figure 1.1: Shortest path problem.

Searching for minimum of a convex function. Here shown with 1 dimensional action space.

Searching over paths through a graph or over policies in a control problem.

Part I: Modeling

Example: product testing



Goal of the experiment:

- Should we use 4,6, or 8 icons in the future?
- Should we productionize ML model variant A,B,C or D?

Decisions are standardized across the population.

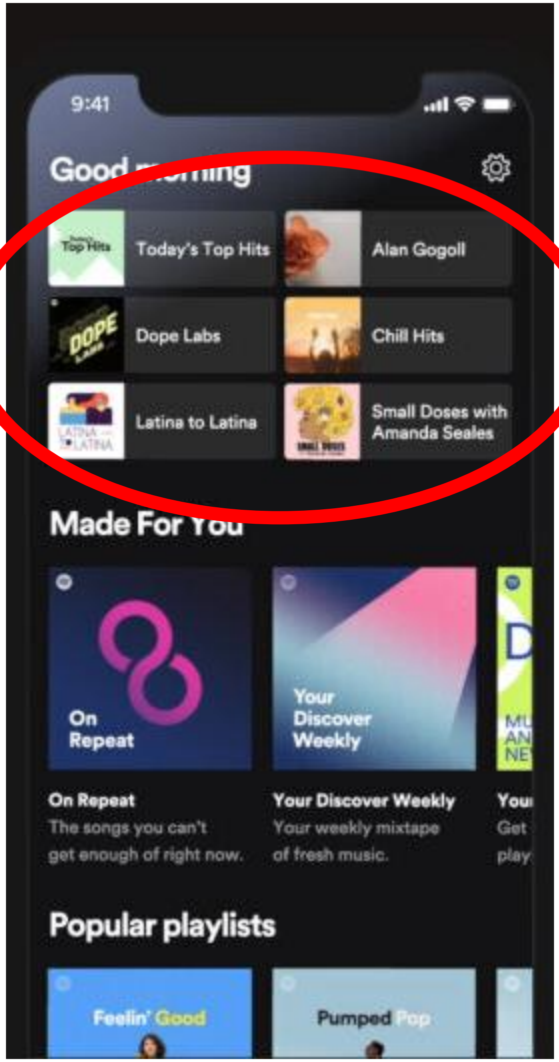
Context

- Day, time of day, promo running?
- Age, gender, location, device.
- Taste, app usage.
- Usage in previous 10 minutes

Reward measure

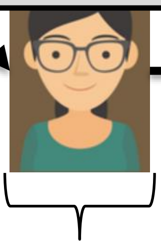
- Overall app usage?
- % of streams from home-page?
- Minutes of streaming from home-page?

Example: product testing



Comments on the role of context

- Context explains much more of the variability in user responses than the treatment decision. ‘Controlling for context’ reduces sample complexity.
- The company knows about the distribution of contexts.
  - This is logged passively before the experiment and during the experiment is logged for users held back from the test.

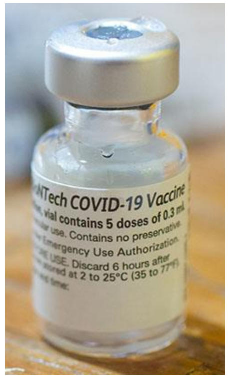


- % of streams from home-page?
- Minutes of streaming from home-page?

Context

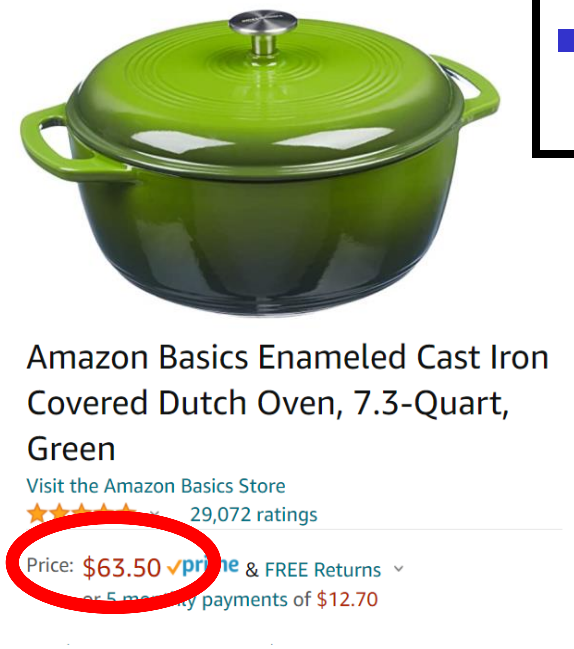
- Day, time of day, promo running?
- Age, gender, location, device.
- Taste, app usage.
- Usage in previous 10 minutes

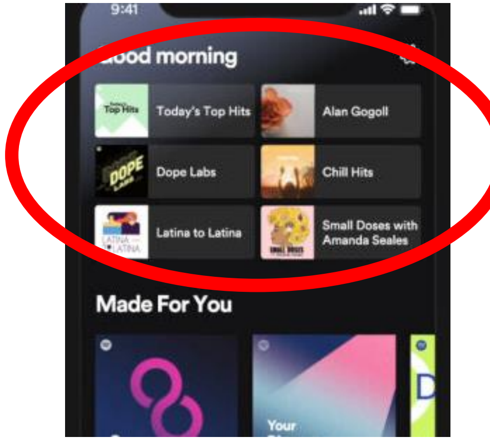
Why standardize not personalize?

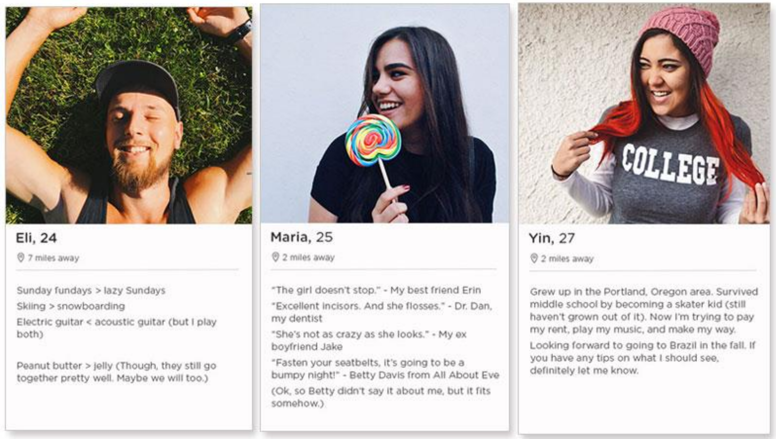


There are many reasons to want decisions to be invariant to (aspects of) the context

- Operational benefits
- Fairness, ethical, or legal constraints
- Incentive compatibility constraints
- Sample complexity benefits
- Social benefits
- Consistency benefits







Model: decision goal and prior knowledge

- The objective in running the experiment is to learn which arm (a.k.a. treatment or action) to employ throughout a population and across future contexts.

- With perfect knowledge, we would make the utilitarian choice of the arm with the highest average treatment effect:

$$I^* = \text{argmax}_{i \in [k]} \left\{ \mu(\theta, i, w) := \sum_x w(x) \langle x, \theta^i \rangle \right\}$$

Population context distribution

Context's feature vector

Uncertain parameter determining arm i's expected performance in each context.

<u>Prior knowledge</u>

1. The population distribution w is known.

2. The experimenter begins with a prior $\theta = \left(\theta^1, \ldots, \theta^k\right) \sim N(\mu, \Sigma)$.

Model: information gathering

Adaptive experimentation

For t=1,2,…

- Observe context $X_t \in \mathbb{R}^d$
- Play $I_t \in [k]$
- Observe reward $R_t = \langle \theta^i, X_t \rangle + N(0, \sigma^2)$

We allow for delay that limits feasible adaptivity:

→ I_t chosen based on $(R_1, .. R_{t-L})$.

Post-experiment decision

Experimentation yields information

$$H_T^+ = (X_1, I_1, R_1, \ldots, X_T, I_T, R_T)$$

The price of unresolved uncertainty is:

$$\Delta_T = \mu(\theta, I^*, w) - \mu(\theta, \hat{I}_T, w)$$

Unknown best arm

$$I^* = \operatorname*{argmax}_{i \in [k]} \mu(\theta, i, w)$$

Bayes selection

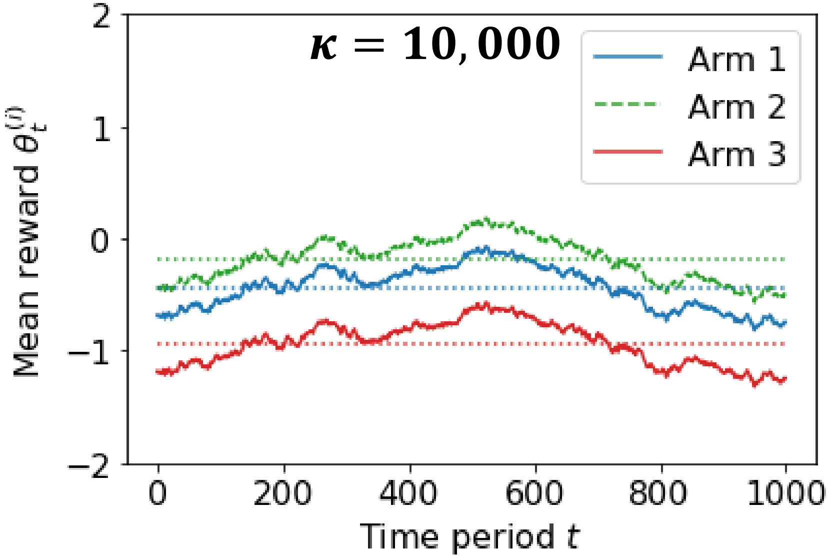
$$\hat{I}_T = \operatorname*{argmax}_{i \in [k]} \mathbb{E}[\mu(\theta, i, w) \mid H_T^+]$$

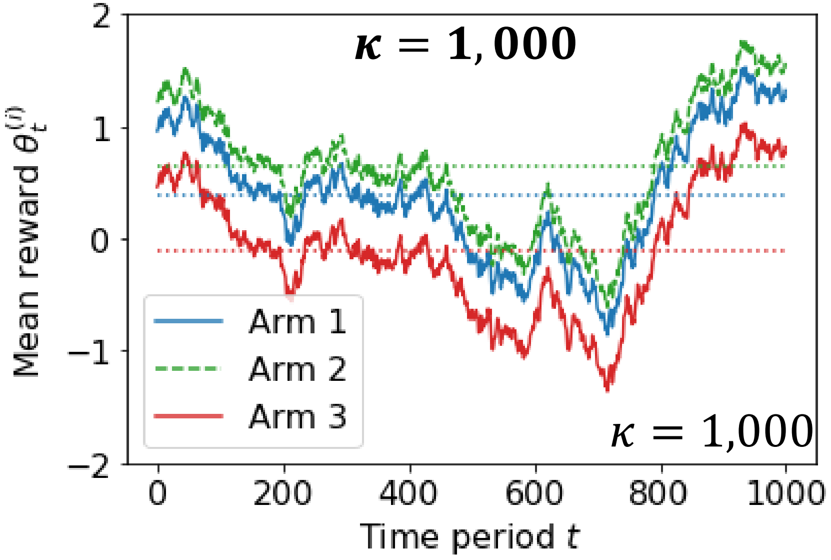
Contexts in the experiment might be i.i.d or might follow a nonstationary pattern.

Part IB: ‘Hard’ nonstationary examples

Ex: Bayesian model of latent confounders

- Almost surely $X_t = e_t \in \mathbb{R}^T$, the t-th standard basis vector.
- Ideal choice is $I^* = \text{argmax}_{i \in [k]} \left\{ \frac{\theta_1^i + \cdots + \theta_T^i}{T} \right\}$ when $w = \text{Uniform}[e_1, \ldots, e_T]$
  - Plots below use: $\theta_t^i = \underbrace{Z^i}_{\text{arm-effect}} + \underbrace{Z_t}_{\text{time-effect}}$ and $\text{Cor}(Z_t, Z_{t'}) = e^{-|t-t'|/\kappa}$.





Example: day-of-week effects

Run a weeklong experiment to decide which **fixed arm** to employ in future weeks.

- Ideal choice is $I^* = \text{argmax}_{i \in [k]} \left\{ \mu(\theta, i, w) := \frac{\theta_1^i + \cdots + \theta_7^i}{7} \right\}$

- Latent variable model induces structured prior covariance:
  - $\theta_x^i = \underbrace{Z^i}_{\text{arm-effect}} + \underbrace{Z_x}_{\text{day-effect}} + \underbrace{Z_{x,i}}_{\text{interaction-effect}}$

Monday

Period $t = 1$
Context $X_t = e_1$
Action $I_t \in [k]$
Reward $R_t = \theta_1^{I_t} + W_t$

Period $t = T/7$
Context $X_t = e_1$
Action $I_t \in [k]$
Reward $R_t = \theta_1^{I_t} + W_t$

Sunday

Period $t = 6T/7+1$
Context $X_t = e_7$
Action $I_t \in [k]$
Reward $R_t = \theta_7^{I_t} + W_t$

Period $t = T$
Context $X_t = e_7$
Action $I_t \in [k]$
Reward $R_t = \theta_7^{I_t} + W_t$

Example: day-of-week effects

Two challenges

1. **Distribution shift:** Day-of-week effects will confound inferences if unmodeled.
2. **Information Delays:** If they're modeled, uncertainty does not fully resolve until Sunday.
   - …even if an arm is played repeatedly on earlier days.

Monday

Period $t = 1$
Context $X_t = e_1$
Action $I_t \in [k]$
Reward $R_t = \theta_1^{I_t} + W_t$

…

Period $t = T/7$
Context $X_t = e_1$
Action $I_t \in [k]$
Reward $R_t = \theta_1^{I_t} + W_t$

…

Sunday

Period $t = 6T/7+1$
Context $X_t = e_7$
Action $I_t \in [k]$
Reward $R_t = \theta_7^{I_t} + W_t$

…

Period $t = T$
Context $X_t = e_7$
Action $I_t \in [k]$
Reward $R_t = \theta_7^{I_t} + W_t$

Failure of context-unaware algorithms

1. **Distribution shift:** Day-of-week effects will confound inferences if unmodeled.

2. **Information Delays:** If they're modeled, uncertainty does not fully resolve until Sunday.
   - …even if an arm is played repeatedly on earlier days.

Context unaware algorithms fail due to distribution shift:

Ignore contexts and apply TS/UCB pretending each arm generates i.i.d rewards.

- Can get stuck only sampling whichever arm is best on Mondays.
- Failure to gather adequate information means $\inf_T \mathbb{E}[\Delta_T] > 0$.

Failure of deconfounded UCB

1. **Distribution shift:** Day-of-week effects will confound inferences if unmodeled.
2. **Information Delays:** If they're modeled, uncertainty does not fully resolve until Sunday.
    - …even if an arm is played repeatedly on earlier days.

Deconfounded UCB fails due to information delays:

Correct adaptation of UCB: $I_t \in \underset{i \in [k]}{\operatorname{argmax}} \mathbb{E}[\mu(\theta, i, w) | H_t] + z\sqrt{\mathrm{Var}(\mu(\theta, i, w) | \mathrm{H}_t)}$

- Can get stuck only sampling one uncertain arm on Monday, Tuesday etc.
    - This does not resolve uncertainty about the weeklong average $\mu(\theta, i, w)$.
- Failure to gather adequate information means $\inf_T \mathbb{E}[\Delta_T] > 0$.

Part II: Deconfounded TS

Proper inference

As observations are gathered, algorithms can track beliefs about:

1. **The uncertain parameters:** $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k)}\right)$
   - $\theta | H_t \sim N(\mu_t, \Sigma_t)$
   - E.g if beliefs are independent across arms: $\Sigma_{t,i} = \left(\Sigma_{1,i} + \sum_1^t 1(I_t = i) X_t X_t^\top\right)^{-1}$.

2. **Marginals, like the population avg reward** $\boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{i}, \boldsymbol{w}) = \langle \boldsymbol{\theta}^{(i)}, \boldsymbol{X}_{\mathbf{pop}} \rangle$
   - $\mu(\theta, i, w) | H_t \sim N\left(\langle \mu_{t,i}, X_{\text{pop}} \rangle, X_{\text{pop}}^\top \Sigma_{t,i} X_{\text{pop}}\right)$

$= \mathbb{E}_{x \sim w}[x]$.

Form beliefs about population performance…

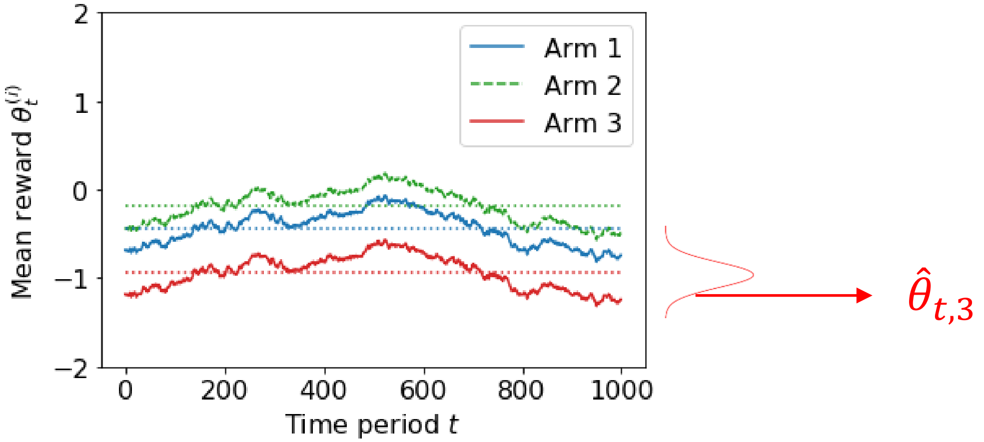
…while accounting for exogenous variation driven by contexts.

Deconfounded Thompson Sampling (DTS)

Deconfounded TS makes two modifications to standard TS

1) **Change the learning target:** Sample an arm according to the posterior probability it maximizes *the population average reward*.

- Intellectual def: $\mathbb{P}(I_t = i|H_t) = \mathbb{P}(I^* = i|H_t)$
- Algorithmic def: $I_t \in \mathrm{argmax}_{i\in[k]}\, \hat{\theta}_{t,i}$ where $\hat{\theta}_{t,i}|H_t \sim N\big(\langle\mu_{t,i}, X_{\mathrm{pop}}\rangle\,, X_{\mathrm{pop}}^{\top}\,\Sigma_{t,i}\, X_{\mathrm{pop}}\big)$



Deconfounded Thompson Sampling (DTS)

Deconfounded TS makes two modifications to standard TS

1) **Change the learning target:** Sample an arm according to the posterior probability it maximizes *the population average reward*.

- Intellectual def: $\mathbb{P}(I_t = i | H_t) = \mathbb{P}(I^* = i | H_t)$
- Algorithmic def: $I_t \in \operatorname{argmax}_{i \in [k]} \hat{\theta}_{t,i}$ where $\hat{\theta}_{t,i} | H_t \sim N\left(\langle \mu_{t,i}, X_{\text{pop}} \rangle, X_{\text{pop}}^\top \Sigma_{t,i} X_{\text{pop}}\right)$

Why should this work when deconfounded UCB does not?

- *Randomizing in the face of uncertainty lets it cope with information delays.*

Deconfounded Thompson Sampling (DTS)

Deconfounded TS makes two modifications to standard TS

1) **Change the learning target:** Sample an arm according to the posterior probability it maximizes *the population average reward*.

- Intellectual def: $\mathbb{P}(I_t = i | H_t) = \mathbb{P}(I^* = i | H_t)$
- Algorithmic def: $I_t \in \mathrm{argmax}_{i\in[k]}\, \hat{\theta}_{t,i}$ where $\hat{\theta}_{t,i} | H_t \sim N\big(\langle \mu_{t,i}, X_{\mathrm{pop}} \rangle\,, X_{\mathrm{pop}}^{\top}\, \Sigma_{t,i}\, X_{\mathrm{pop}}\big)$

---

2) **Top-two sampling:** (A modification to make focus on post-experiment performance)

- Continue sampling arms according to $\mathbb{P}(I^* = \cdot\, | H_t)$ until two distinct choices are drawn.
- Flip a coin to select among those top two.

Simple Bayesian Algorithms for Best-Arm Identification, Russo, Operations Research, 2019.

Part III: Theory

DTS strikes a delicate balance between

- ❑ Aggressive adaptivity
- ❑ Robustness to nonstationary confounders

Robustness / Efficiency

Result 1: Robustness

With arbitrary delay in observing rewards, arbitrary context sequence,

$$\mathbb{E}[\,\Delta_T \mid X_1, \ldots, X_T\,] = \tilde{O}\left(\sigma\sqrt{\frac{k \cdot X_{\text{pop}}^{\top}\left(T^{-1}\sum X_t X_t^{\top}\right)^{-1} X_{\text{pop}}}{T}}\right)$$

Result 2: Asymptotic efficiency

Assume contexts are i.i.d with $\mathbb{E}[X_1 X_1^{\top}] \succ 0$, and no delay. Then, with some stopping rule $\tau = \tau(c)$,

$$\mathbb{E}[\,c\tau + \Delta_\tau \mid \theta\,] = c \cdot \log\left(\frac{1}{c}\right) \cdot \left(\Gamma_\theta + o(1)\right) \;\text{ as }\; c \to 0.$$

Robustness / Efficiency

Result 1: Robustness

With arbitrary delay in observing rewards, arbitrary context sequence,

$$\mathbb{E}[\, \Delta_T \mid X_1, \dots, X_T \,] = \tilde{O}\left(\sigma\sqrt{\frac{k \cdot X_{\text{pop}}^{\top} \left(T^{-1} \sum X_t X_t^{\top} \right)^{-1} X_{\text{pop}}}{T}}\right)$$

Where RCTs shine

Result 2: Asymptotic efficiency

Assume contexts are i.i.d with $\mathbb{E}[X_1 X_1^{\top}] \succ 0$, and no delay. Then, with some stopping rule $\tau = \tau(c)$,

$$\mathbb{E}[\, c\tau + \Delta_\tau \mid \theta \,] = c \cdot \log\left(\frac{1}{c}\right) \cdot \left(\Gamma_\theta + o(1)\right) \text{ as } c \to 0.$$

Where bandit algos shine

Robustness / Efficiency

Result 1: Robustness

With arbitrary delay in observing rewards, arbitrary context sequence,

$$\mathbb{E}[\,\Delta_T \mid X_1, \dots, X_T\,] = \tilde{O}\left(\sigma\sqrt{\frac{k \cdot X_{\text{pop}}^{\top}\left(T^{-1}\sum X_t X_t^{\top}\right)^{-1} X_{\text{pop}}}{T}}\right)$$

Punchline: For a hard instances, where there is severe delay, severe nonstationary, and arms may not be well separated….

…DTS gets roughly the same bound as non-adaptive uniform sampling

Result 2: Asymptotic efficiency

Assume contexts are i.i.d with $\mathbb{E}[X_1 X_1^{\top}] \succ 0$, and no delay. Then, with some stopping rule $\tau = \tau(c)$,

$$\mathbb{E}[\,c\tau + \Delta_\tau \mid \theta\,] = c \cdot \log\left(\frac{1}{c}\right) \cdot \left(\Gamma_\theta + o(1)\right) \text{ as } c \to 0.$$

Punchline: In 'easy' instances where contexts are i.i.d, and large sample sizes let the algorithm focus on competitive arms…

…DTS incurs minimal cost up to first-order asymptotically.

Robustness / Efficiency

Result 1: Robustness

With arbitrary delay in observing rewards, arbitrary context sequence,

$$\mathbb{E}[\,\Delta_T \mid X_1, \dots, X_T\,] = \tilde{O}\left(\sigma\sqrt{\frac{k \cdot X_{\text{pop}}^{\top}\left(T^{-1}\sum X_t X_t^{\top}\right)^{-1} X_{\text{pop}}}{T}}\right)$$

How to achieve this?

Don't over-react to rewards earned in a limited set of contexts. Continue to gather enough information about all arms.

Result 2: Asymptotic efficiency

Assume contexts are i.i.d with $\mathbb{E}[X_1 X_1^{\top}] \succ 0$, and no delay. Then, with some stopping rule $\tau = \tau(c)$,

$$\mathbb{E}[\,c\tau + \Delta_\tau \mid \theta\,] = c \cdot \log\left(\frac{1}{c}\right) \cdot \left(\Gamma_\theta + o(1)\right) \ \text{as} \ c \to 0.$$

How to achieve this?

Quickly zero-in on the competitive arms. Play inferior ones *just enough*.

Robustness / Efficiency

Result 1: Robustness

With arbitrary delay in observing rewards, arbitrary context sequence,

$$\mathbb{E}[\,\Delta_T \mid X_1, \dots, X_T\,] = \tilde{O}\left(\sigma\sqrt{\frac{k \cdot X_{\mathrm{pop}}^\top \left(T^{-1}\sum X_t X_t^\top\right)^{-1} X_{\mathrm{pop}}}{T}}\right)$$

Conditions on contexts

Integrates over the prior <--> "Bayesian"

Fixed experimentation horizon

Result 2: Asymptotic efficiency

Assume contexts are i.i.d with $\mathbb{E}[X_1 X_1^\top] \succ 0$, and no delay. Then, with some stopping rule $\tau = \tau(c)$,

$$\mathbb{E}[\,c\tau + \Delta_\tau \mid \theta\,] = c \cdot \log\left(\frac{1}{c}\right) \cdot \left(\Gamma_\theta + o(1)\right) \text{ as } c \to 0.$$

Integrates over the draw of contexts

Conditions on θ <--> "Frequentist"

Allows for adaptive stopping to sidestep open theoretical questions

Result 1: Robustness (A)

Posterior variance of $\mu(\theta, i, w)$ if you observed arm i's reward in each context:

$$V(X_{1:T}) = X_{\text{pop}}^\top \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^{T} X_t X_t^\top \right)^{-1} X_{\text{pop}}$$

Proposition 1. *Suppose that* $\|X_t\|_2 \leqslant 1$ *almost surely for* $t \in \mathbb{N}$. *If DTS applied with tuning parameters satisfying* $\inf_{t \in \mathbb{N}} \beta_t \geqslant 1/2$ *almost surely and with the Bayes optimal selection rule in (7), then for any* $T \in \mathbb{N}$,

$$\mathbb{E}\left[\Delta_T \mid X_{1:T}\right] \leqslant \sqrt{2\iota \cdot k \cdot \mathbb{H}(I^* \mid H_T^+) \cdot V(X_{1:T})}$$

where $\iota = \max\left\{ 9 \log\left(d\lambda_{\max}(\Sigma_1) \left[\lambda_{\max}\left(\Sigma_1^{-1}\right) + T \right] \right) \cdot \lambda_{\max}(\Sigma_1), 9 \right\}$.

As if you saw each arm in every context, but with k times the noise.

Result 1: Robustness (B)

This corollary applies in a problem like the day of week example:

- The empirical context distribution is the same as the population distribution.
- It's the order which is challenging.

Corollary 1. *Under the conditions of Proposition 1, for any sequence* $x_{1:T} \in \mathcal{X}^T$, *with* $\frac{1}{T}\sum_{t=1}^{T} x_t x_t^\top \succeq X_{\text{pop}} X_{\text{pop}}^\top$,

$$\mathbb{E}\left[\Delta_T \mid X_{1:T} = x_{1:T}\right] \leqslant \sigma \sqrt{\frac{2\iota \cdot k \cdot \mathbb{H}(I^* \mid H_T^+)}{T}} \leqslant \sigma \sqrt{\frac{2\iota \cdot k \cdot \log(k)}{T}}$$

where ι *is given in Proposition 1.*

Bound has no dependence on the dimension of the context space.
Similar to common 'gap-independent' lower bounds for k-armed bandits.

Result 1: Robustness (C)

- Proof uses inverse propensity weights implicitly to analyze the posterior
  - Special care is required because 'overlap' condition is violated,

Step 1: Simple regret is small if you can estimate the quality of I^*

$$\mathbb{E}\left[\Delta_T\right] \lessapprox \sqrt{O(\log(kT)) X_{\text{pop}}^\top \mathbb{E}\left[\tilde{S}_{T,I^*}\right] X_{\text{pop}}} \quad \textit{where} \quad \tilde{S}_{T,i} \equiv \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{T} \mathbb{P}(I_\ell = i \mid H_\ell) X_\ell X_\ell^\top\right)^{-1}$$

Step 2: Posterior variance is less than the sampling variance of a propensity score estimator.

Lemma (Propensity matching type variance bound). *For any $i \in [k]$, with probability one,*

$$\tilde{S}_{T,i} \preceq S_{\text{full}} \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^{T} \frac{X_t X_t^\top}{\mathbb{P}(I_t = i \mid H_t)}\right) S_{\text{full}}.$$

Step 3: DTS can neglect bad actions, but it's expected to assign large propensity to I^*

Lemma (Inverse propensity of the optimal action). *Define $\alpha_{t,i} = \mathbb{P}(I^* = i | H_t)$. Then,*

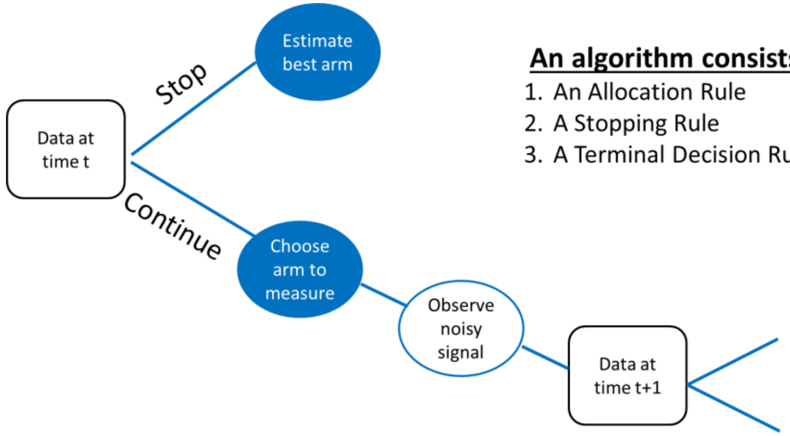
$$\mathbb{E}\left[1/\alpha_{t,I^*}\right] = k$$

Result 2: Asymptotic Efficiency (A)

Very rough view: DTS minimizes total cost

$$\mathbb{E}[c\tau + \Delta_\tau \mid \theta]$$

as $c \to 0$ among all admissible procedures.



An algorithm consists of:
1. An Allocation Rule
2. A Stopping Rule
3. A Terminal Decision Rule

Proposition 3. *Suppose Assumption 1 holds and $L = 1$ (no delay). If DTS is applied with β_t set by Algorithm 2 and stopping time τ defined in (26) with parameter $\delta = c$, and the Bayes optimal selection rule in (7), then*

$$\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_0\right] \leqslant \Gamma_{\theta_0}[c + o_{\theta_0}(1)]\log(1/c) \qquad \text{for all } \theta_0 \in \Theta.$$

Under any admissible sampling rule, selection rule, and stopping rule $\tau = \tau(c)$, if

$$\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_0\right] < \Gamma_{\theta_0}[c + o_{\theta_0}(1)]\log(1/c) \qquad \text{for some } \theta_0 \in \Theta$$

as $c \to 0$, then

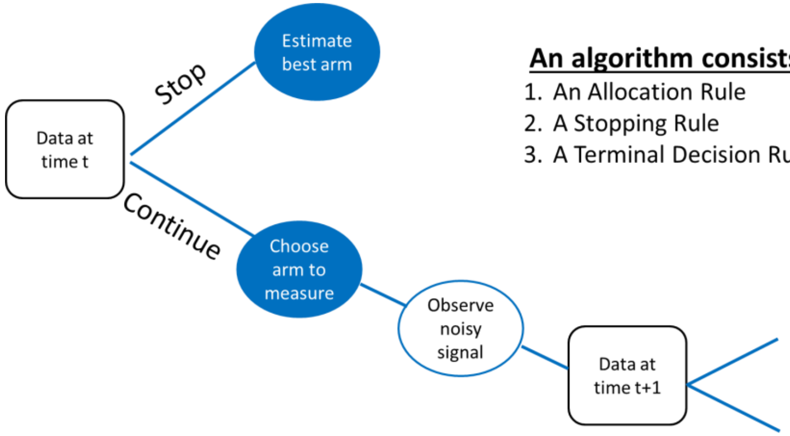
$$\lim_{c \to \infty} \frac{\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_1\right]}{c\log(1/c)} = \infty \quad \text{for some } \theta_1 \in \Theta. \tag{27}$$

Result 2: Asymptotic Efficiency (A)

Very rough view: DTS minimizes total cost

$$\mathbb{E}[c\tau + \Delta_\tau \mid \theta]$$

as $c \to 0$ among all admissible procedures.



An algorithm consists of:

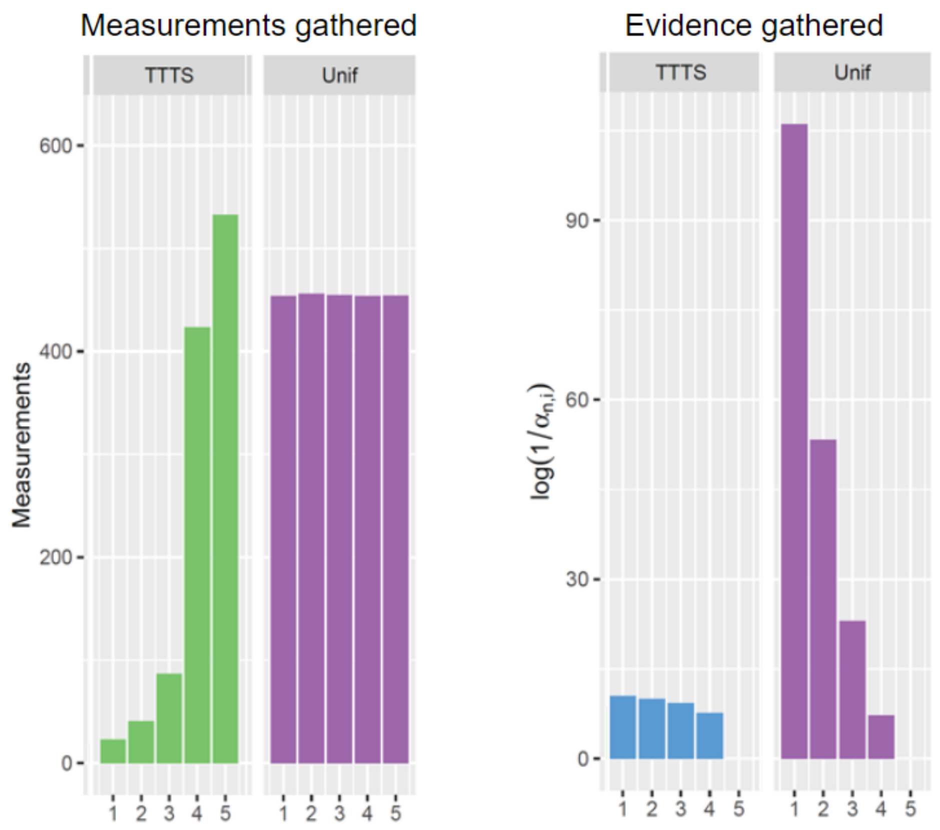
1. An Allocation Rule
2. A Stopping Rule
3. A Terminal Decision Rule

Extends known optimality of top-two Sampling to contextual problems

- Past work due to [Russo, 2016], [Qin et. al 2017], [Shang et. al 2020]
- Total cost objective is similar, but not identical, to past work.

Result 2: Asymptotic Efficiency (B)

Information balance property.



From an experiment w/o contexts,
arm_means = $(.5, .4, .3, .2, .1)$

Asymptotic behavior of DTS:

Under DTS, almost surely

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} 1(I_t = i)X_tX_t^\top = p_i^*(\theta)\mathbb{E}[X_1X_1^\top]$$

where $p^* = p^*(\theta)$ satisfies the scalar information balance constraint:

$$\frac{\left(\theta^{(I^*)} - \theta^{(i)}\right)^\top X_{\text{pop}}}{\sqrt{(p_{I^*}^*)^{-1} + (p_i^*)^{-1}}} = \frac{\left(\theta^{(I^*)} - \theta^{(j)}\right)^\top X_{\text{pop}}}{\sqrt{(p_{I^*}^*)^{-1} + (p_j^*)^{-1}}} \qquad \forall i, j \neq I^*$$

- *Coin's bias is separate from information balance and determines tradeoff between regret and speed of learning*

Result 2: Asymptotic Efficiency (C)

- Asymptotic sample complexity largely worked out in abstract form by Chernoff, [1959]
  - Specific limits for best- arm identification worked out by Jennison et. al [1982], Chan and Lai [2006] and Garivier and Kaufmann [2016]…
- Sample complexity is determined by the equilibrium value:

$$\Gamma_\theta^{-1} = \max_{M_{1:k} \in \mathbb{M}} \min_{\theta_a \in \mathrm{Alt}(\theta)} \frac{1}{2\sigma^2} \sum_{i \in [k]} \left(\theta^{(i)} - \theta_a^{(i)}\right)^\top M_i \left(\theta^{(i)} - \theta_a^{(i)}\right)$$

Experimenter picks 'design matrices'

Skeptic picks most confusing instance w/ $I^*(\theta_a) \neq I^*(\theta)$

Measures accumulated KL divergence

As a step toward showing the optimality of DTS, we show the experimenters optimal strategy uses 'context-independent' sampling $\longleftrightarrow$ $M_i^* = p_i^*(\theta)\mathbb{E}\left[X_1 X_1^\top\right]$

Part IV: Key Related Work

(Some) key related work

- Decision-theoretic approximations — [Frazier et. al, 2008/9] [Chick et al, 2018] ……
- Best-of both worlds in Nonstochastic best-arm identification — [Cong Shen, 2018] [Jamieson & Talwalkar, 2015] [Abbasi-Yadkori et. al, 2018]
- Asymptotic limits of best-arm identification problems — [Chernoff, 1959], [Glynn & Juneja, 2004], [Chang & Lai, 2006] [Garivier & Kaufmann, 2016], [Russo, 2016] [Qin et. al 2017] [Shang et al, 2020]
- Causal estimation techniques & semiparametric efficiency in contextual bandits — [Dudík etl. Al, 2011] [Bareinboim et. al, 2015] [Dimakopoulou et. al, 2017] [Kallus, 2018] [Athey & Wager, 2021] ….


NOW I WILL TAKE
YOUR QUESTIONS.