A/B Tests Under a Safety Budget: A Simulation-Optimization Point of View

Nian Si

Joint work with Jose Blanchet, Ramesh Johari, and Zeyu Zheng

INFORMS 2022

October 18, 2022

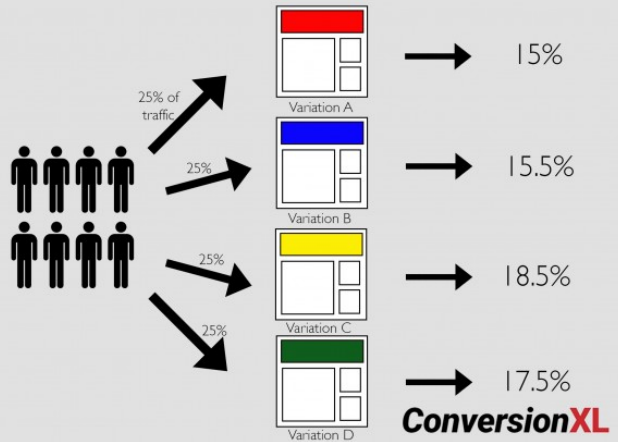
Road map of this talk

1. Motivation: unknown risk in online experiments
2. Formulation and main results by large deviation principles
   - A special case on equal variances
3. Numerical illustrations

A/B tests

Stanford University

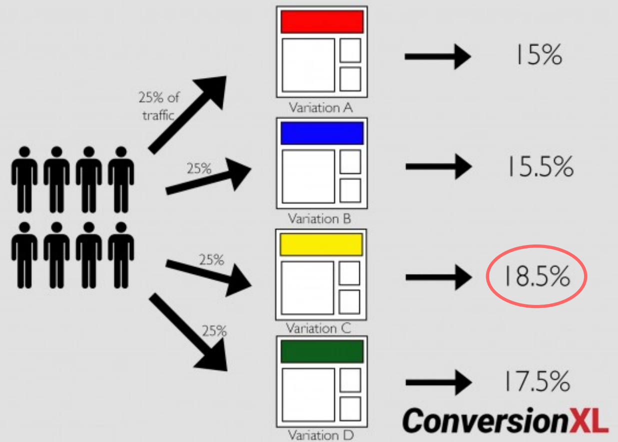
- A/B tests: effectively identify the best from a pool of different designs.



A/B tests

Stanford University

- A/B tests: effectively identify the best from a pool of different designs.



Safety and risky new design

Stanford University

- New designs could be risky: incur large costs and a simple mistake may threaten the whole system.

POSTED ON OCTOBER 5, 2021 TO NETWORKING & TRAFFIC

More details about the October 4 outage

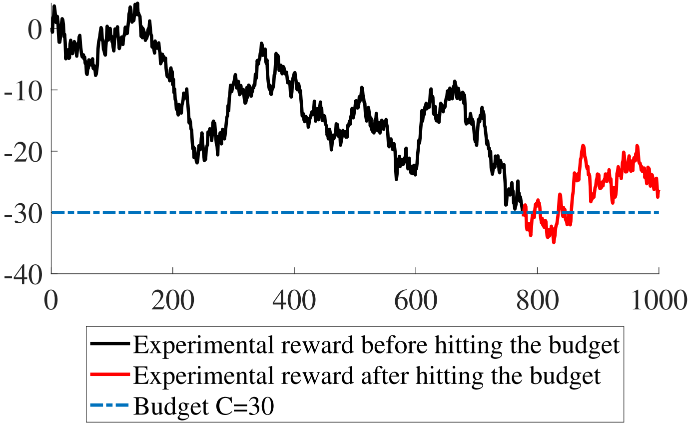
This was the source of yesterday's outage. During one of these routine maintenance jobs, a command was issued with the intention to assess the availability of global backbone capacity, which unintentionally took down all the connections in our backbone network, effectively disconnecting Facebook data centers globally. Our systems are designed to audit commands like these to prevent mistakes like this, but a bug in that audit tool prevented it from properly stopping the command. [1]

[1] https://engineering.fb.com/2021/10/05/networking-traffic/outage-details/

Unknown risk in online experiments

Stanford University

- A safety budget is set to regulate the total cost that can be tolerated in the experiment.



Formulation: ranking and selection

Stanford University

- One control action with known mean reward μ_0; K treatment actions with unknown mean rewards $\mu_1, \ldots, \mu_K$ and follow Gaussian distributions $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Formulation: ranking and selection

Stanford University

- One control action with known mean reward μ_0; K treatment actions with unknown mean rewards $\mu_1, \ldots, \mu_K$ and follow Gaussian distributions $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.
- At time t, a treatment action $I_t \in \{1, 2, \ldots, K\}$ is chosen and a random reward $X_{I_t,t}$ is revealed.

Formulation: ranking and selection

Stanford University

- One control action with known mean reward μ_0; K treatment actions with unknown mean rewards $\mu_1, \ldots, \mu_K$ and follow Gaussian distributions $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.
- At time t, a treatment action $I_t \in \{1, 2, \ldots, K\}$ is chosen and a random reward $X_{I_t,t}$ is revealed.
- The experiment horizon is T and the safety budget is C. Define the stopping time

$$\tau_{C,T} = T \wedge \inf \left\{ t \,\middle|\, \sum_{s=1}^{t} X_{I_s,s} \leq \mu_0 t - C \right\}.$$

Formulation: ranking and selection

Stanford University

- One control action with known mean reward μ_0; K treatment actions with unknown mean rewards $\mu_1, \ldots, \mu_K$ and follow Gaussian distributions $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.
- At time t, a treatment action $I_t \in \{1, 2, \ldots, K\}$ is chosen and a random reward $X_{I_t,t}$ is revealed.
- The experiment horizon is T and the safety budget is C. Define the stopping time

$$\tau_{C,T} = T \wedge \inf \left\{ t \,\middle|\, \sum_{s=1}^{t} X_{I_s,s} \le \mu_0 t - C \right\}.$$

- $I_{\tau_{C,T}+1}$ is the experimenter's decision of the treatment action with the highest mean upon stopping. Goal: minimize the probability of false selection:

$$\mathbb{P}\left\{ I_{\tau_{C,T}+1} \notin \arg\max_{1 \le i \le K} \mu_i \right\}.$$

Literature review

Stanford University

- Safety concerns in standard practice of A/B tests in industry [Xu et al., 2018, Kohavi et al., 2020];

Literature review

Stanford University

- Safety concerns in standard practice of A/B tests in industry [Xu et al., 2018, Kohavi et al., 2020];
- Ranking and selection [Chen et al., 2000, Glynn and Juneja, 2004, Batur and Kim, 2005, Morrice and Butler, 2006, Kim and Nelson, 2006, Hong and Nelson, 2007, Chick and Gans, 2009, Frazier et al., 2009, Andradóttir and Kim, 2010, Waeber et al., 2010, Lee et al., 2012, Chick and Frazier, 2012, Healey et al., 2013, Hunter and Pasupathy, 2013, Pasupathy et al., 2014, Song et al., 2015, Hunter and Nelson, 2017, Gao et al., 2018, Lam and Li, 2018, Wu and Zhou, 2018, Chen and Ryzhov, 2019, Hong et al., 2021, Kim et al., 2022];
- Best arm identification [Even-Dar et al., 2002, Mannor and Tsitsiklis, 2004, Audibert et al., 2010, Gabillon et al., 2012, Karnin et al., 2013, Jamieson and Nowak, 2014, Chen and Li, 2015, Garivier and Kaufmann, 2016, Kaufmann et al., 2014, 2016, Russo, 2020, Agrawal et al., 2020];

Literature review: continue

Stanford University

- Feasible arm identification: Find the best treatment for one key metric provided that other metrics are not bad [Katz-Samuels and Scott, 2018, 2019];

Literature review: continue

Stanford University

- Feasible arm identification: Find the best treatment for one key metric provided that other metrics are not bad [Katz-Samuels and Scott, 2018, 2019];
- Safe and conservative contextual bandit and reinforcement learning [Driessens and Džeroski, 2004, Koppejan and Whiteson, 2009, Taylor and Stone, 2007, Garcıa and Fernández, 2015, Wu et al., 2016, Kazerouni et al., 2017, Amani et al., 2019, Xu et al., 2021];

Literature review: continue

Stanford University

- Feasible arm identification: Find the best treatment for one key metric provided that other metrics are not bad [Katz-Samuels and Scott, 2018, 2019];
- Safe and conservative contextual bandit and reinforcement learning [Driessens and Džeroski, 2004, Koppejan and Whiteson, 2009, Taylor and Stone, 2007, Garcıa and Fernández, 2015, Wu et al., 2016, Kazerouni et al., 2017, Amani et al., 2019, Xu et al., 2021];
- Connections between regret minimization and best arm identification. [Degenne et al, 2019, Zhong et al, 2021].

Our results: setup

Stanford University

- Limiting regime: $C, T \to +\infty$ with $T/C \to \beta$, where β represents the safety level.
  - $\beta \uparrow$ means relatively small $C \Rightarrow$ safer.
- Gaussian setting: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ with $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$.

Our results: setup

Stanford University

- Limiting regime: $C, T \to +\infty$ with $T/C \to \beta$, where β represents the safety level.
    - $\beta \uparrow$ means relatively small $C \Rightarrow$ safer.
- Gaussian setting: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ with $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$.
- Static allocation rule $\sum_{i=1}^{K} p_i = 1$ stationary over time: up to time t, we collect $p_i t$ samples from the treatment action i for every $t \leq \tau_{C,T}$.
- Decision rule: $I_{\tau_{C,T}+1} \in \arg\max_{1 \leq i \leq K} \bar{X}_i(\tau_{C,T})$.

The main theorem

Stanford University

Theorem

For $C, T \to +\infty$ with $T/C \to \beta$, $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ with $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$, any allocation rules uniform over time and the empirical-maximizer decision rule, we have

$$\lim_{C,T\to\infty} -\frac{1}{C}\log\left(\mathbb{P}\left(I_{\tau_{C,T}+1} \neq 1\right)\right) = \min_{j\geq 2}\left\{\min\left\{H_j(p), \beta G_j(p)\right\}\right\},$$

where

$$G_j(p) = \frac{\left(\mu_1 - \mu_j\right)^2}{2\left(\sigma_1^2/p_1 + \sigma_j^2/p_j\right)},$$

- $H_j(p)$ corresponds to the event of early stopping, i.e., $\tau_{C,T} < T$, and wrong selection of the j-th action, i.e., $\bar{X}_j(\tau_{C,T}) > \bar{X}_1(\tau_{C,T})$, for $j = 1, 2, \ldots, K$.
- $G_j(p)$ corresponds to the event of stopping at time T, i.e., $\tau_{C,T} = T$, and wrong selection of the j-th action, i.e., $\bar{X}_j(\tau_{C,T}) > \bar{X}_1(\tau_{C,T})$, for $j = 1, 2, \ldots, K$.

Comparision with the vanilla case without safety constraints

Stanford University

	W/ safety		W/o safety ($C=+\infty$) [2]
$T/C=\beta$ range	$[0,+\infty)$		$\beta=0$
Stopping time	$\tau_{C,T}$	$\leq$	T
$\lim_{T\to\infty} -\frac{1}{T}\log(PFS)$	$\min_{j\geq 2}\left\{\min\left\{\frac{1}{\beta}H_j(p), G_j(p)\right\}\right\}$		$\min_{j\geq 2}\{G_j(p)\}$

[2] Glynn and Juneja [2004]

Equal variances: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2 = \mathcal{V}$

Stanford University

Proposition

We assume $X_i \sim \mathcal{N}(\mu_i, \mathcal{V})$ for $i = 1, 2, \ldots, K$ and $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$. For any allocations $p_1, p_2, \ldots, p_K$ satisfying $\sum_{i=1}^K p_i = 1$. For $C, T \to +\infty$ and $T/C \to \beta$

$$\begin{aligned}
&\lim_{C,T\to\infty} -\frac{1}{C}\log\left(\mathbb{P}\left(I_{\tau_{C,T}+1} \neq 1\right)\right) \\
= \;& \frac{1}{\mathcal{V}}\min\left\{\mathcal{D} + \sqrt{\mathcal{D}^2 + \min_{j\geq 2}\frac{(\mu_1-\mu_j)^2}{1/p_j + 1/p_1}}, \beta\min_{j\geq 2}\left\{\frac{(\mu_1-\mu_j)^2}{2\left(1/p_1 + 1/p_j\right)}\right\}\right\},
\end{aligned}$$

where $\mathcal{D}$ is the mean extra reward per unit:

$$\mathcal{D} = \sum_{i=1}^K p_i\left(\mu_i - \mu_0\right).$$

Equal variances: optimal allocation

Stanford University

- The optimal allocation $p^* = [p_1^*, p_2^*, \ldots, p_K^*]^\top$ is defined as

$$\{p_1^*, p_2^*, \ldots, p_K^*\} = \operatorname*{arg\,max}_{p \geq 0, \sum_{i=1}^K p_i = 1} \min_{j \geq 2} \{\min\{H_j(p), \beta G_j(p)\}\}.$$

Theorem

For the equal variance case, we have the optimal allocation rule satisfies

$$\frac{(\mu_1 - \mu_i)^2}{1/p_1^* + 1/p_i^*} = \frac{(\mu_1 - \mu_j)^2}{1/p_1^* + 1/p_j^*} \text{ for } i \neq j \neq 1. \tag{1}$$

Comparison with the vanilla case without safety constraints

Stanford University

- $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2 = \mathcal{V}$.
- The optimal allocation without safety constraints $\left\{p_1^{0,*}, p_2^{0,*}, \ldots, p_K^{0,*}\right\}$ is defined as

$$\left\{p_1^{0,*}, p_2^{0,*}, \ldots, p_K^{0,*}\right\} = \underset{p\geq 0, \sum_{i=1}^K p_i = 1}{\arg\max} \min_{j\geq 2} \left\{G_j(p)\right\}. \quad (2)$$

	W/ safety		W/o safety ($C = +\infty$) [3]
$T/C = \beta$ range	$[0, +\infty)$		$\beta = 0$
Stopping time	$\tau_{C,T}$	$\leq$	T
$\lim_{T\to\infty} -\frac{1}{T}\log(PFS)$	$\min_{j\geq 2}\left\{\min\left\{\frac{1}{\beta}H_j(p), G_j(p)\right\}\right\}$		$\min_{j\geq 2}\left\{G_j(p)\right\}$
Optimal allocation	$\frac{(\mu_1-\mu_i)^2}{1/p_1^*+1/p_i^*} = \frac{\left(\mu_1-\mu_j\right)^2}{1/p_1^*+1/p_j^*}$		$\frac{(\mu_1-\mu_i)^2}{1/p_1^{0,*}+1/p_i^{0,*}} = \frac{\left(\mu_1-\mu_j\right)^2}{1/p_1^{0,*}+1/p_j^{0,*}}$
Leading probability	p_1^*	$\geq$	$p_1^{0,*}$

[3] Glynn and Juneja [2004]

Comparison with the vanilla case without safety constraints

Stanford University

- $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2 = \mathcal{V}$.
- The optimal allocation without safety constraints $\left\{p_1^{0,*}, p_2^{0,*}, \ldots, p_K^{0,*}\right\}$ is defined as

$$\left\{p_1^{0,*}, p_2^{0,*}, \ldots, p_K^{0,*}\right\} = \underset{p \geq 0, \sum_{i=1}^K p_i = 1}{\arg\max} \min_{j \geq 2} \left\{G_j(p)\right\}. \tag{2}$$

	W/ safety		W/o safety ($C = +\infty$) [3]
$T/C = \beta$ range	$[0, +\infty)$		$\beta = 0$
Stopping time	$\tau_{C,T}$	$\leq$	T
$\lim_{T\to\infty} -\frac{1}{T}\log(PFS)$	$\min_{j\geq 2}\left\{\min\left\{\frac{1}{\beta}H_j(p), G_j(p)\right\}\right\}$		$\min_{j\geq 2}\left\{G_j(p)\right\}$
Optimal allocation	$\frac{(\mu_1-\mu_i)^2}{1/p_1^*+1/p_i^*} = \frac{\left(\mu_1-\mu_j\right)^2}{1/p_1^*+1/p_j^*}$		$\frac{(\mu_1-\mu_i)^2}{1/p_1^{0,*}+1/p_i^{0,*}} = \frac{\left(\mu_1-\mu_j\right)^2}{1/p_1^{0,*}+1/p_j^{0,*}}$
Leading probability	p_1^*	$\geq$	$p_1^{0,*}$

[3] Glynn and Juneja [2004]

More structural insights

Stanford University

$$\text{Stop at T:} \qquad \{p_1^{0,*}, p_2^{0,*}, \ldots, p_K^{0,*}\} = \underset{p\geq 0, \sum_{i=1}^K p_i = 1}{\arg\max} \min_{j\geq 2} \{G_j(p)\};$$

$$\text{Early stop:} \qquad \{p_1^{\infty,*}, p_2^{\infty,*}, \ldots, p_K^{\infty,*}\} = \underset{p\geq 0, \sum_{i=1}^K p_i = 1}{\arg\max} \min_{j\geq 2} \{H_j(p)\}.$$

Theorem

For the equal variance case, we have $G_j(p), H_j(p), j = 1, 2, \ldots K$ *are all quasi-concave. Therefore,* p_1^* *is monotonic with respect to* β*, and there exists* $0 \leq \underline{\beta} \leq \overline{\beta} \leq +\infty$ *(*$\underline{\beta}, \overline{\beta}$ *could possibly be zero or* $+\infty$*) such that*

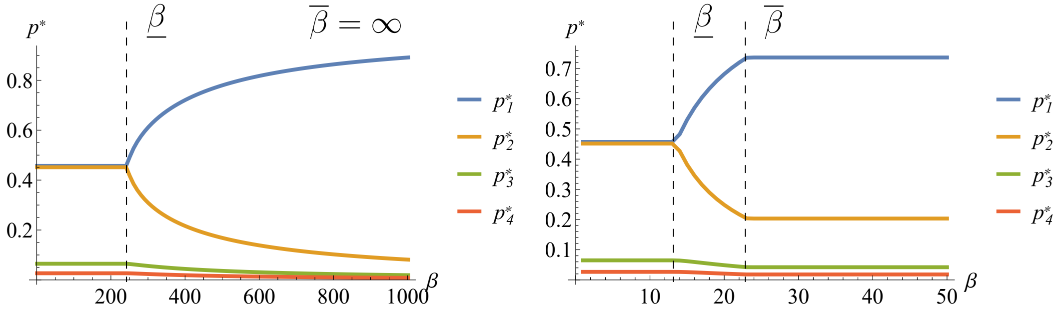
$$p^* = \begin{cases} p^{0,*} & \text{for } \beta < \underline{\beta} \\ p^{\infty,*} & \text{for } \beta \geq \overline{\beta} \end{cases},$$

and if $\beta \in [\underline{\beta}, \overline{\beta})$, p^* *satisfies* $\min_{j\geq 2}\{H_j(p^*)\} = \beta \min_{j\geq 2}\{G_j(p^*)\}$.

Numerical Illustrations for the Equal-Variance Case

Optimal allocation

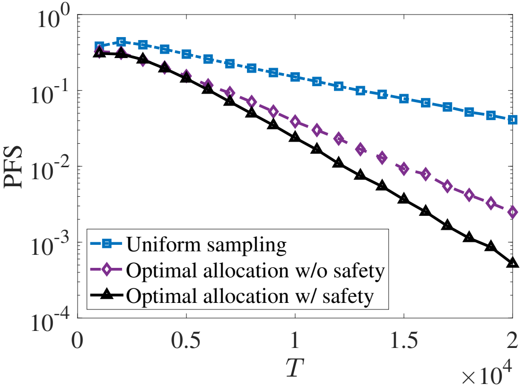
Stanford University



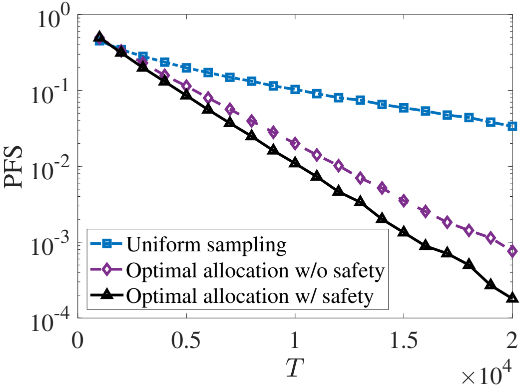
(a) $\mu_1 = 0.2, \mu_2 = 0.1, \mu_3 = 0.0, \mu_4 = -0.1$ (b) $\mu_1 = 0.0, \mu_2 = -0.1, \mu_3 = -0.2, \mu_4 = -0.3$

Figure 1: The optimal allocation rules with respect to different $\beta = T/C$ for $K = 4$ actions with equal variances and $\mu_0 = 0$

Probability of false selection



(a) $\mu_1 = 0.2, \mu_2 = 0.1, \mu_3 = 0.0, \mu_4 = -0.1$, $T/C = \beta = 1000$



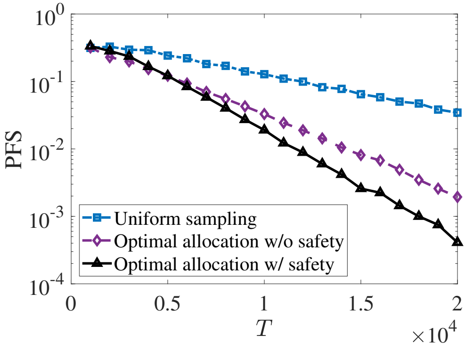
(b) $\mu_1 = 0, \mu_2 = -0.1, \mu_3 = -0.2, \mu_4 = -0.3$, $T/C = \beta = 50$

Figure 2: The probability of false selection with respect to different horizons T for $K = 4$ actions with equal variances and $\mu_0 = 0$

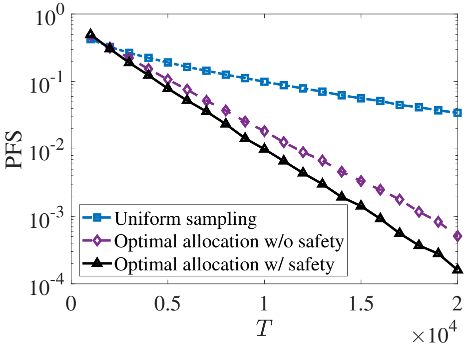
Probability of false selection: model misspecification

Stanford University

- Two-point distributions supported on $\{-1, 1\}$: for $i = 1, 2, 3, 4$
$$\mathbb{P}\{X_i = 1\} = 1/2 + \mu_i/2 \text{ and } \mathbb{P}\{X_i = -1\} = 1/2 - \mu_i/2;$$
- Consider optimal allocation rules derived under Gaussian assumptions.



(a) $\mu_1 = 0.2, \mu_2 = 0.1, \mu_3 = 0.0, \mu_4 = -0.1$, $T/C = \beta = 1000$



(b) $\mu_1 = 0, \mu_2 = -0.1, \mu_3 = -0.2, \mu_4 = -0.3$, $T/C = \beta = 50$

Conclusion

Stanford University

- We emphasize the importance of safety in online A/B tests and we propose a framework to study this issue based on ranking and selection.
- We provide a large deviation theory for the probability of false selection.
- We explicitly solve the optimal sampling budget allocation problem that minimizes the probability of false selection under safety constraints for the equal-variance case.
- The optimal allocation rule exhibits similar structures with the vanilla rule without safety considerations but has a systematical shift.

The paper

Stanford University

Nian Si, Jose Blanchet, Ramesh Johari, and Zeyu Zheng. "A/B Tests under a Safety Budget: A Simulation-Optimization Point of View." *Available soon*, 2022+.

Thanks!

References I

Stanford University

Shubhada Agrawal, Sandeep Juneja, and Peter Glynn. Optimal δ -correct best-arm selection for heavy-tailed distributions. In *Algorithmic Learning Theory*, pages 61–110. PMLR, 2020.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Sigrún Andradóttir and Seong-Hee Kim. Fully sequential procedures for comparing constrained systems via simulation. *Naval Research Logistics (NRL)*, 57(5):403–421, 2010.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer, 2010.

D Batur and S Kim. Finding the best in the presence of multiple constraints. In *Proceedings of the 2005 Winter Simulation Conference, IEEE: Piscataway, NJ*, pages 732–738, 2005.

Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen E Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3): 251–270, 2000.

Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015.

References II

Stanford University

Ye Chen and Ilya O Ryzhov. Complete expected improvement converges to an optimal budget allocation. *Advances in Applied Probability*, 51(1):209–235, 2019.

Stephen E Chick and Peter Frazier. Sequential sampling with economics of selection procedures. *Management Science*, 58(3):550–569, 2012.

Stephen E Chick and Noah Gans. Economic analysis of simulation selection problems. *Management Science*, 55(3):421–437, 2009.

Kurt Driessens and Sašo Džeroski. Integrating guidance into relational reinforcement learning. *Machine Learning*, 57(3):271–304, 2004.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

References III

Stanford University

Fei Gao, Siyang Gao, Hui Xiao, and Zhongshun Shi. Advancing constrained ranking and selection with regression in partitioned domains. *IEEE Transactions on Automation Science and Engineering*, 16 (1):382–391, 2018.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.

Christopher M Healey, Sigrún Andradóttir, and Seong-Hee Kim. Efficient comparison of constrained systems using dormancy. *European journal of operational research*, 224(2):340–352, 2013.

L Jeff Hong and Barry L Nelson. Selecting the best system when systems are revealed sequentially. *Iie Transactions*, 39(7):723–734, 2007.

L Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021.

References IV

Stanford University

Susan R Hunter and Barry L Nelson. Parallel ranking and selection. In *Advances in Modeling and Simulation*, pages 249–275. Springer, 2017.

Susan R Hunter and Raghu Pasupathy. Optimal sampling laws for stochastically constrained simulation optimization on finite sets. *INFORMS Journal on Computing*, 25(3):527–542, 2013.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of a/b testing. In *Conference on Learning Theory*, pages 461–481. PMLR, 2014.

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

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

References V

Stanford University

Seong-Hee Kim and Barry L Nelson. Selecting the best system. *Handbooks in operations research and management science*, 13:501–534, 2006.

Taeho Kim, Kyoung-kuk Kim, and Eunhye Song. Selection of the most probable best. *arXiv preprint arXiv:2207.07533*, 2022.

Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press, 2020.

Rogier Koppejan and Shimon Whiteson. Neuroevolutionary reinforcement learning for generalized helicopter control. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 145–152, 2009.

Henry Lam and Fengpei Li. Sampling uncertain constraints under parametric distributions. In *2018 Winter Simulation Conference (WSC)*, pages 2072–2083. IEEE, 2018.

Loo Hay Lee, Nugroho Artadi Pujowidianto, Ling-Wei Li, Chun-Hung Chen, and Chee Meng Yap. Approximate simulation budget allocation for selecting the best design in the presence of stochastic constraints. *IEEE Transactions on Automatic Control*, 57(11):2940–2945, 2012.

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

References VI

Stanford University

Douglas J Morrice and John C Butler. Ranking and selection with multiple" targets". In *Proceedings of the 2006 winter simulation conference*, pages 222–230. IEEE, 2006.

Raghu Pasupathy, Susan R Hunter, Nugroho A Pujowidianto, Loo Hay Lee, and Chun-Hung Chen. Stochastically constrained ranking and selection via score. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 25(1):1–26, 2014.

Daniel Russo. Simple bayesian algorithms for best arm identification. *Operations Research*, pages 1625–1647, 2020.

Eunhye Song, Barry L Nelson, and L Jeff Hong. Input uncertainty and indifference-zone ranking & selection. In *2015 Winter Simulation Conference (WSC)*, pages 414–424. IEEE, 2015.

Matthew E Taylor and Peter Stone. Representation transfer for reinforcement learning. In *AAAI Fall Symposium: Computational Approaches to Representation Change during Learning and Development*, pages 78–85, 2007.

Rolf Waeber, Peter I Frazier, and Shane G Henderson. Performance measures for ranking and selection procedures. In *Proceedings of the 2010 Winter Simulation Conference*, pages 1235–1245. IEEE, 2010.

References VII

Stanford University

Di Wu and Enlu Zhou. Analyzing and provably improving fixed budget ranking and selection algorithms. *arXiv preprint arXiv:1811.12183*, 2018.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.

Wanqiao Xu, Kan Xu, Hamsa Bastani, and Osbert Bastani. Safely bridging offline and online reinforcement learning. *arXiv preprint arXiv:2110.13060*, 2021.

Ya Xu, Weitao Duan, and Shaochen Huang. Sqr: Balancing speed, quality and risk in online experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 895–904, 2018.

Results without the budget constraint[4]

Stanford University

- No budget constraint and stop at time T. Gaussian setting: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Assume $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$.
- For the allocation rule $\sum_{i=1}^K p_i = 1$ and the decision rule $I_{T+1} \in \arg\max_{1\leq i\leq K} \bar{X}_i(T)$:

$$\lim_{T\to\infty} -\frac{1}{T}\log\left(\mathbb{P}\left(I_{T+1} \neq 1\right)\right) = \min_{j\geq 2}\left\{\frac{(\mu_1-\mu_j)^2}{2\left(\sigma_1^2/p_1+\sigma_j^2/p_j\right)}\right\}. \tag{*}$$

- Optimal decision rule satisfies

$$\frac{(\mu_1-\mu_i)^2}{\sigma_1^2/p_1^*+\sigma_i^2/p_i^*} = \frac{(\mu_1-\mu_j)^2}{\sigma_1^2/p_1^*+\sigma_j^2/p_j^*} \text{ for } i \neq j. \tag{**}$$

[4] Glynn and Juneja [2004]

Our results: details

Stanford University

- $G_j(p)$ is the same as the large deviation rate function in the vanilla case without safety constraints (*).

Our results: details

Stanford University

- $G_j(p)$ is the same as the large deviation rate function in the vanilla case without safety constraints (*).
- $H_j(p)$ satisfies

$$H_j(p) = H_j^{(1)}(p) := 2\mathcal{D}/\mathcal{V} \text{ if } \mu_1 - \mu_j < 2\left(\sigma_1^2 - \sigma_j^2\right)\left(\mathcal{D}/\mathcal{V}\right) \text{ with } \mathcal{D} > 0,$$

and otherwise, $H_j(p) = H_j^{(2)}(p) :=$

$$\frac{\left(\frac{(\mu_1-\mu_j)(\sigma_j^2-\sigma_1^2)}{\sigma_j^2/p_j+\sigma_1^2/p_1} + \mathcal{D}\right) + \sqrt{\left(\mathcal{D} + \frac{(\sigma_j^2-\sigma_1^2)(\mu_1-\mu_j)}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)^2 + \frac{(\mu_1-\mu_j)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\left(\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)}}{\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}},$$

where $\mathcal{V}$ is the variance per unit and $\mathcal{D}$ is the mean extra reward per unit:

$$\mathcal{V} = \sum_{i=1}^{K} p_i\sigma_i^2, \text{ and } \mathcal{D} = \sum_{i=1}^{K} p_i\left(\mu_i - \mu_0\right).$$

Our results: details

Stanford University

- $G_j(p)$ is the same as the large deviation rate function in the vanilla case without safety constraints (*).
- $H_j(p)$ satisfies

$$H_j(p) = H_j^{(1)}(p) := 2\mathcal{D}/\mathcal{V} \text{ if } \mu_1 - \mu_j < 2\left(\sigma_1^2 - \sigma_j^2\right)\left(\mathcal{D}/\mathcal{V}\right) \text{ with } \mathcal{D} > 0,$$

and otherwise, $H_j(p) = H_j^{(2)}(p) :=$

$$\frac{\left(\frac{(\mu_1-\mu_j)(\sigma_j^2-\sigma_1^2)}{\sigma_j^2/p_j+\sigma_1^2/p_1} + \mathcal{D}\right) + \sqrt{\left(\mathcal{D} + \frac{(\sigma_j^2-\sigma_1^2)(\mu_1-\mu_j)}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)^2 + \frac{(\mu_1-\mu_j)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\left(\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)}}{\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}},$$

where $\mathcal{V}$ is the variance per unit and $\mathcal{D}$ is the mean extra reward per unit:

$$\mathcal{V} = \sum_{i=1}^{K} p_i\sigma_i^2, \text{ and } \mathcal{D} = \sum_{i=1}^{K} p_i\left(\mu_i - \mu_0\right).$$

Our results: details

Stanford University

- $G_j(p)$ is the same as the large deviation rate function in the vanilla case without safety constraints (*).
- $H_j(p)$ satisfies

$$H_j(p) = H_j^{(1)}(p) := 2\mathcal{D}/\mathcal{V} \text{ if } \mu_1 - \mu_j < 2\left(\sigma_1^2 - \sigma_j^2\right)(\mathcal{D}/\mathcal{V}) \ \text{ with } \mathcal{D} > 0,$$

and otherwise, $H_j(p) = H_j^{(2)}(p) :=$

$$\frac{\left(\frac{(\mu_1-\mu_j)(\sigma_j^2-\sigma_1^2)}{\sigma_j^2/p_j+\sigma_1^2/p_1} + \mathcal{D}\right) + \sqrt{\left(\mathcal{D} + \frac{(\sigma_j^2-\sigma_1^2)(\mu_1-\mu_j)}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)^2 + \frac{(\mu_1-\mu_j)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\left(\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)}}{\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}},$$

where $\mathcal{V}$ is the variance per unit and $\mathcal{D}$ is the mean extra reward per unit:

$$\mathcal{V} = \sum_{i=1}^{K} p_i \sigma_i^2, \text{ and } \mathcal{D} = \sum_{i=1}^{K} p_i\left(\mu_i - \mu_0\right).$$

Our results: details

Stanford University

- $G_j(p)$ is the same as the large deviation rate function in the vanilla case without safety constraints (*).
- $H_j(p)$ satisfies

$$H_j(p) = H_j^{(1)}(p) := 2\mathcal{D}/\mathcal{V} \text{ if } \mu_1 - \mu_j < 2\left(\sigma_1^2 - \sigma_j^2\right)\left(\mathcal{D}/\mathcal{V}\right) \text{ with } \mathcal{D} > 0,$$

and otherwise, $H_j(p) = H_j^{(2)}(p) :=$

$$\frac{\left(\frac{(\mu_1-\mu_j)(\sigma_j^2-\sigma_1^2)}{\sigma_j^2/p_j+\sigma_1^2/p_1} + \mathcal{D}\right) + \sqrt{\left(\mathcal{D} + \frac{(\sigma_j^2-\sigma_1^2)(\mu_1-\mu_j)}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)^2 + \frac{(\mu_1-\mu_j)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\left(\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}\right)}}{\mathcal{V} - \frac{(\sigma_j^2-\sigma_1^2)^2}{\sigma_j^2/p_j+\sigma_1^2/p_1}},$$

where $\mathcal{V}$ is the variance per unit and $\mathcal{D}$ is the mean extra reward per unit:

$$\mathcal{V} = \sum_{i=1}^{K} p_i\sigma_i^2, \text{ and } \mathcal{D} = \sum_{i=1}^{K} p_i\left(\mu_i - \mu_0\right).$$

Comparison with the control action

Our theorem also holds when some of σ_i's is zero and as long as $\sum_{i=1}^{K} p_i\sigma_i^2 > 0$. Therefore, we allow some treatment action i to be a control action, i.e., $\mu_i = \mu_0$ and $\sigma_i = 0$. In particular, if $\mu_1 = \mu_0$ and $\sigma_1 = 0$, we have for $j = 2, \ldots, K$

$$H_j(p) = H_j^{(2)}(p)$$

$$= \frac{\sum_{i=2,i\neq j}^{K} p_i\left(\mu_i - \mu_0\right) + \sqrt{\left(\sum_{i=2,i\neq j}^{K} p_i\left(\mu_i - \mu_0\right)\right)^2 + \frac{p_j}{\sigma_j^2}\left(\mu_1 - \mu_j\right)^2\left(\sum_{i=2,i\neq j}^{K} p_i\sigma_i^2\right)}}{\sum_{i=2,i\neq j}^{K} p_i\sigma_i^2}.$$

Otherwise, if $\mu_j = \mu_0$ and $\sigma_j = 0$ for $j \neq 1$, we have

$$H_j(p) = H_j^{(1)}(p) = 2\mathcal{D}/\mathcal{V} = \frac{2\sum_{i=1,i\neq j}^{K} p_i\left(\mu_i - \mu_0\right)}{\sum_{i=1,i\neq j}^{K} p_i\sigma_i^2}.$$

Special case 2: two treatment actions

Stanford University

- $K = 2$.

Proposition

We assume $K = 2$, $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ *for* $i = 1, 2$, *and* $\mu_1 > \mu_2$. *For any allocations* p_1, p_2 *satisfying* $p_1 + p_2 = 1$, *we have*

$$H_j^{(2)}(p) = \sqrt{\left(\frac{p_1}{\sigma_1^2} + \frac{p_2}{\sigma_2^2}\right)\left(\frac{p_1}{\sigma_1^2}(\mu_1 - \mu_0)^2 + \frac{p_2}{\sigma_2^2}(\mu_2 - \mu_0)^2\right)} + \left(\frac{p_1}{\sigma_1^2}(\mu_1 - \mu_0) + \frac{p_2}{\sigma_2^2}(\mu_2 - \mu_0)\right).$$

Numerical algorithms

Stanford University

Top-two Thompson sampling (TTTS) method proposed in Russo (2020).

- With probability $\hat{\alpha}_t$, sample from the posterior distribution and select the largest.
- With probability $1 - \hat{\alpha}_t$, continue sampling until an action different from the first sampling action is selected.
- Consistently tuning $\hat{\alpha}_t$.

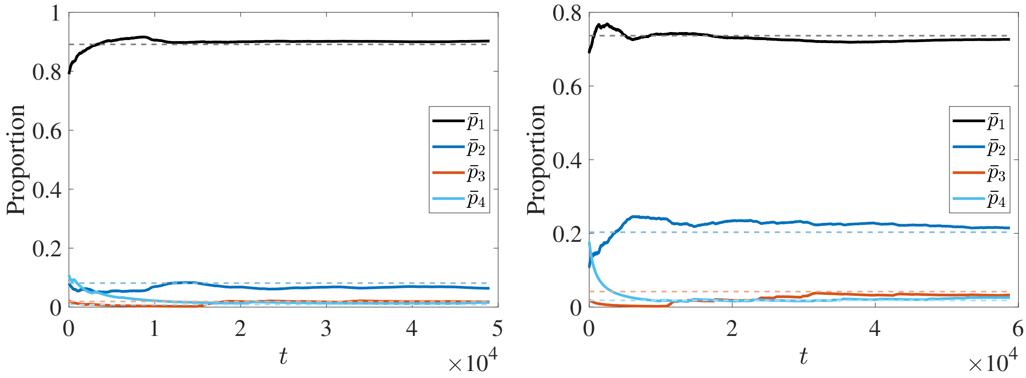
Proposition (Consistency)

We assume $X_i \sim \mathcal{N}(\mu_i, \mathcal{V})$ for $i = 1, 2, \ldots, K$ and $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$. Under the algorithm, for C_n, $T_n \to +\infty$ with $T_n / C_n \to \beta$, we have

$$\lim_{n \to +\infty} \frac{N_i(\tau_n)}{\tau_n} = p_i^* \text{ almost surely for } i = 1, 2, \ldots, K.$$

Numerical performance: consistency

Stanford University

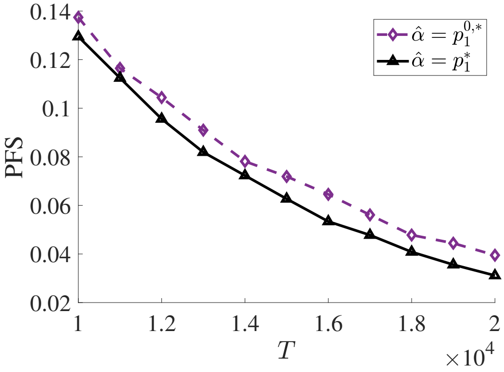


(c) $\mu_1 = 0.2, \mu_2 = 0.1, \mu_3 = 0.0, \mu_4 = -0.1$ (d) $\mu_1 = 0, \mu_2 = -0.1, \mu_3 = -0.2, \mu_4 = -0.3$

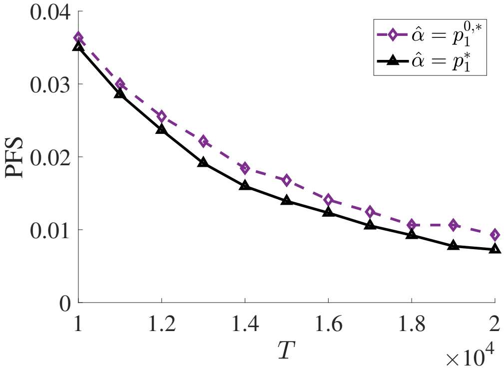
Figure 3: The convergence of the sampling proportions to the optimal allocations with $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$

Numerical performance: probability of false selection

Stanford University



(a) $\mu_1 = 0.2, \mu_2 = 0.1, \mu_3 = 0.0, \mu_4 = -0.1, \beta = 1000$



(b) $\mu_1 = 0.0, \mu_2 = -0.1, \mu_3 = -0.2, \mu_4 = -0.3, \beta = 50$

Figure 4: The probability of false selection with respect to different horizons T for TTTS/T3C with fixed $\hat{\alpha} = p_1^*$ and $\hat{\alpha} = p_1^{0,*}$