On multi-armed bandit algorithms for (early stage) clinical trials

Emilie Kaufmann,

based on a joint work with
Maryam Aziz (Spotify) and Marie-Karelle Riviere (Sanofi)

Statistics and Biopharmacy Conference
Paris, September 2022

The multi-armed bandit model

K arms $= K$ probability distributions:


ν_1 ν_2 ν_3 ν_4 ν_5

At round t, an agent:

- chooses an arm A_t
- gets an observation $X_t \sim \nu_{A_t}$ (reward)

Adaptive sampling strategy (**bandit algorithm**):

$$A_{t+1} = F_t(A_1, X_1, \ldots, A_t, X_t).$$

Possible objectives:

- maximize rewards, $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$
- identify the best arm $a_\star = \text{argmax}_{a \in [K]} \mathbb{E}_{X \sim \nu_a}[X]$

The multi-armed bandit model

K arms $= K$ probability distributions:





$\mathcal{B}(p_1)$ $\mathcal{B}(p_2)$ $\mathcal{B}(p_3)$ $\mathcal{B}(p_4)$ $\mathcal{B}(p_5)$

At round t, a doctor:

- chooses a treatment A_t
- observes the outcome (success/failure) : $\mathbb{P}(X_t = 1) = p_{A_t}$

Adaptive sampling strategy (**design**):

$$A_{t+1} = F_t(A_1, X_1, \ldots, A_t, X_t).$$

Possible objectives:

- maximize rewards, $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$
- identify the best arm $a_\star = \text{argmax}_{a \in [K]} p_a$

The multi-armed bandit model

K arms $= K$ probability distributions:





$\mathcal{B}(p_1)$ $\mathcal{B}(p_2)$ $\mathcal{B}(p_3)$ $\mathcal{B}(p_4)$ $\mathcal{B}(p_5)$

At round t, a doctor:

- chooses a treatment A_t
- observes the outcome (success/failure) : $\mathbb{P}(X_t = 1) = p_{A_t}$

Adaptive sampling strategy (**design**):

$$A_{t+1} = F_t(A_1, X_1, \ldots, A_t, X_t).$$

Possible objectives:

- maximize the number of cured patients (*treatment*)
- identify the best arm $a_\star = \text{argmax}_{a \in [K]} p_a$

The multi-armed bandit model

K arms $= K$ probability distributions:





$\mathcal{B}(p_1)$ $\mathcal{B}(p_2)$ $\mathcal{B}(p_3)$ $\mathcal{B}(p_4)$ $\mathcal{B}(p_5)$

At round t, a doctor:

- chooses a treatment A_t
- observes the outcome (success/failure) : $\mathbb{P}(X_t = 1) = p_{A_t}$

Adaptive sampling strategy (**design**):

$$A_{t+1} = F_t(A_1, X_1, \ldots, A_t, X_t).$$

Possible objectives:

- maximize the number of cured patients (*treatment*)
- identify the best treatment (*identification*)

Bandits and clinical trials

Difficulties with bandit algorithms:

- efficacy takes a long time to be observed: a fully-sequential design would lead to a very long trial
  (+ the DSMB cannot meet after every patient)
- standard statistical methodology (e.g. tests) is not always compatible with data collected adaptively

This talk:

- learn about bandit algorithms for *treatment* and *identification*
- ... who come up with quite strong theoretical guarantees

Outline

1. Bandit algorithms
   - Treatment: Maximizing rewards
   - Identification: Best Arm Identification

2. A bandit perspective on early-stage trials
   - MTD identification
   - A phase I/II example

Outline

1. Bandit algorithms
   - Treatment: Maximizing rewards
   - Identification: Best Arm Identification

2. A bandit perspective on early-stage trials
   - MTD identification
   - A phase I/II example

Objective





$\mathcal{B}(p_1)$ $\mathcal{B}(p_2)$ $\mathcal{B}(p_3)$ $\mathcal{B}(p_4)$ $\mathcal{B}(p_5)$

Maximize rewards $\leftrightarrow$ select arm $a_\star$ as much as possible

$\leftrightarrow$ few allocations to sub-optimal arms $a \neq a_\star$

$$a_\star = \arg\max_{a \in [K]} p_a \qquad p_\star = \max_{a \in [K]} p_a$$

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} p_{A_t}\right] = T \times p_\star - \mathbb{E}\left[\sum_{t=1}^{T} (p_\star - p_{A_t})\right]$$

$$= \underbrace{T \times p_\star}_{\substack{\text{reward of the} \\ \text{oracle playing } a_\star}} - \sum_{a \neq a_\star} (p_\star - p_a) \times \underbrace{\mathbb{E}[N_a(T)]}_{\substack{\text{expected number of} \\ \text{selections of arm } a \neq a_\star}}$$

A first idea: (Don't) Follow the Leader

Select each arm once, then exploit our current knowledge:

$$A_{t+1} = \arg\max_{a \in [K]} \hat{p}_a(t)$$

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{p}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the MLE estimate of p_a

A first idea: (Don't) Follow the Leader

Select each arm once, then exploit our current knowledge:

$$A_{t+1} = \arg\max_{a \in [K]} \hat{p}_a(t)$$

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{p}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the MLE estimate of p_a

Follow the leader can fail! $K = 2$, $p_1 > p_2$

$$\mathbb{E}[N_2(T)] \geq (1 - p_1)p_2 \times (T - 1)$$

(*linear* number of sub-optimal allocations)

Exploitation is not enough, we need to add some exploration

➜ **Exploration/Exploitation trade-off**

Thompson Sampling

[Thompson, 1933] suggests the very first bandit algorithm

A Bayesian algorithm: $K = 2$, uniform prior distribution

$$(p_1, p_2) \sim \mathcal{U}([0,1]) \otimes \mathcal{U}([0,1])$$

At time $t+1$, choose

- arm 1 with probability $f(P_t)$
- arm 2 with probability $1 - f(P_t)$

where f is some non-decreasing function and P_t is the posterior probability that arm 1 is optimal

$$P_t = \mathbb{P}_{\substack{\tilde{p}_1 \sim \text{Beta}(S_1(t)+1, F_1(t)+1) \\ \tilde{p}_2 \sim \text{Beta}(S_2(t)+1, F_2(t)+1)}} \left(\tilde{p}_1 \geq \tilde{p}_2 \right)$$

$S_a(t)$ / $F_a(t)$: number of successes/failures observed on arm a up to time t

Thompson Sampling

[Thompson, 1933] suggests the very first bandit algorithm

A Bayesian algorithm: $K = 2$, uniform prior distribution

$$(p_1, p_2) \sim \mathcal{U}([0,1]) \otimes \mathcal{U}([0,1])$$

At time $t+1$, choose

- arm 1 with probability $f(P_t)$
- arm 2 with probability $1 - f(P_t)$

where f is some non-decreasing function and P_t is the posterior probability that arm 1 is optimal

$$P_t = \mathbb{P}_{\substack{\tilde{p}_1 \sim \text{Beta}(S_1(t)+1, F_1(t)+1) \\ \tilde{p}_2 \sim \text{Beta}(S_2(t)+1, F_2(t)+1)}} \Big(\tilde{p}_1 \geq \tilde{p}_2 \Big)$$

$S_a(t)$ / $F_a(t)$: number of successes/failures observed on arm a up to time t

Remark: computing P_t can be costly (especially in 1933)

Thompson Sampling

For $f(x) = x$, we get a simpler implementation, which can be further extended to any prior/posterior distributions.

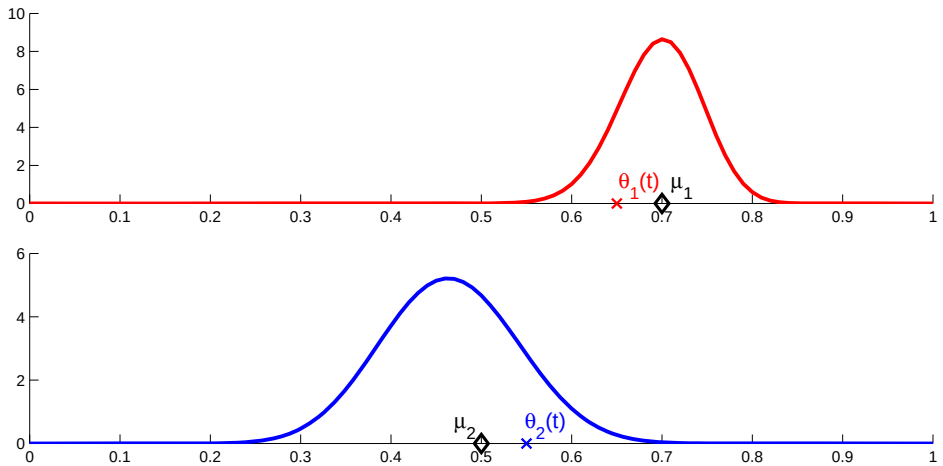
Thompson Sampling

Π_t: posterior distribution on $(p_1, \ldots, p_K)$ after t rounds

At round $t+1$:

- draw a posterior sample $(\tilde{p}_1(t), \ldots, \tilde{p}_K(t)) \sim \Pi_t$
- select arm $A_{t+1} = \arg\max_{a \in [K]} \tilde{p}_a(t)$

"act optimally in a possible model sampled from the posterior"



Thompson Sampling

Thompson Sampling was rediscovered in the 2010s for its very good empirical performance, but little was known in theory.

[Scott, 2010, Chapelle and Li, 2011]

Theorem [Kaufmann et al., 2012, Agrawal and Goyal, 2013]

For every $a \neq a_\star$, for every $\varepsilon > 0$,

$$\mathbb{E}[N_a(T)] \leq (1+\varepsilon)\frac{\log(T)}{\mathrm{kl}(p_a, p_\star)} + o(\log(T))$$

with $\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ the binary relative entropy.

- TS achieves the optimal allocation! (at least asymptotically)

$$\text{[Lai and Robbins, 1985]} : \liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(p_a, p_\star)}$$

Outline

Thompson Sampling for Identification?

What if we want to select arms with Thompson Sampling and propose a guess for the best arm, B_T at time T?

Possible ideas:

1. $B_T = \arg\max_a \hat{p}_a(T)$ (empirical best arm)
2. $B_T = \arg\max_a N_a(T)$ (arm selected the most)
3. $B_T = A_\tau$ where τ is a random index in $\{1, \dots, T\}$

Under ③,

$$\mathbb{E}[p_\star - p_{B_T}] = \sum_{a \neq a_\star} (p_\star - p_a) \frac{\mathbb{E}[N_a(T)]}{T} = O\left(\frac{\log(T)}{T}\right)$$

$$\Rightarrow \quad \mathbb{P}(B_T \neq a_\star) = O\left(\frac{\log(T)}{T}\right)$$

➜ the error probability can actually decay much faster...

The fixed budget setting

Goal: propose an arm B_T which minimizes $\mathbb{P}(B_T \neq a_\star)$

State-of-the-art bandit algorithms are based on eliminations.

Successive Rejects [Audibert et al., 2010]

- $K-1$ stages
- at the end of each stage, discard* one arm
- B_T is the last surviving arm

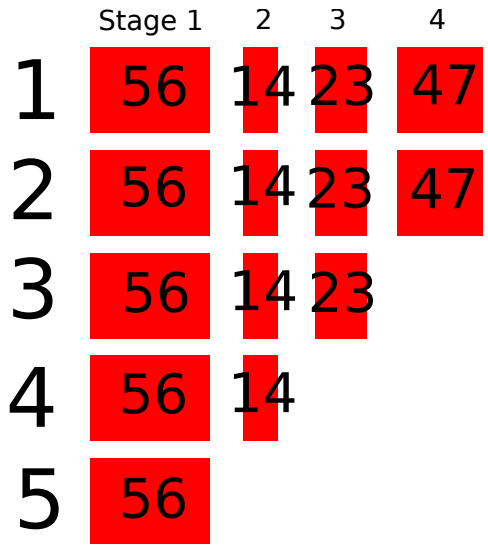
Sequential Halving [Karnin et al., 2013]

- $\lceil \log_2(K) \rceil$ stages
- at the end of each stage, discard* half of the arms
- B_T is the last surviving arm

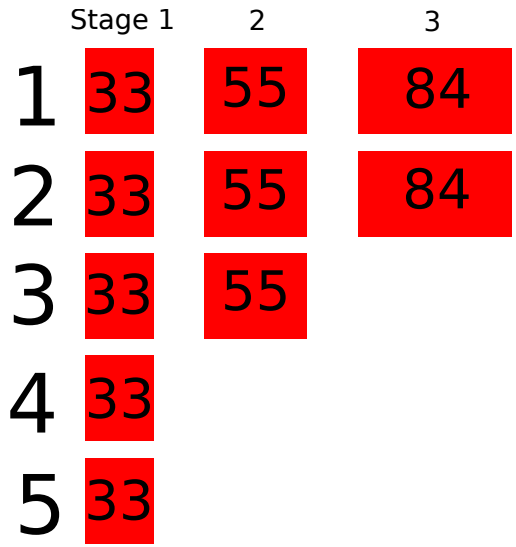
* remove arm(s) whose empirical mean(s) are the smallest

Fixed budget algorithms

Illustration of the allocation for $T = 500, K = 5$



Successive Rejects



Sequential Halving

Theorem [Audibert et al., 2010, Karnin et al., 2013]

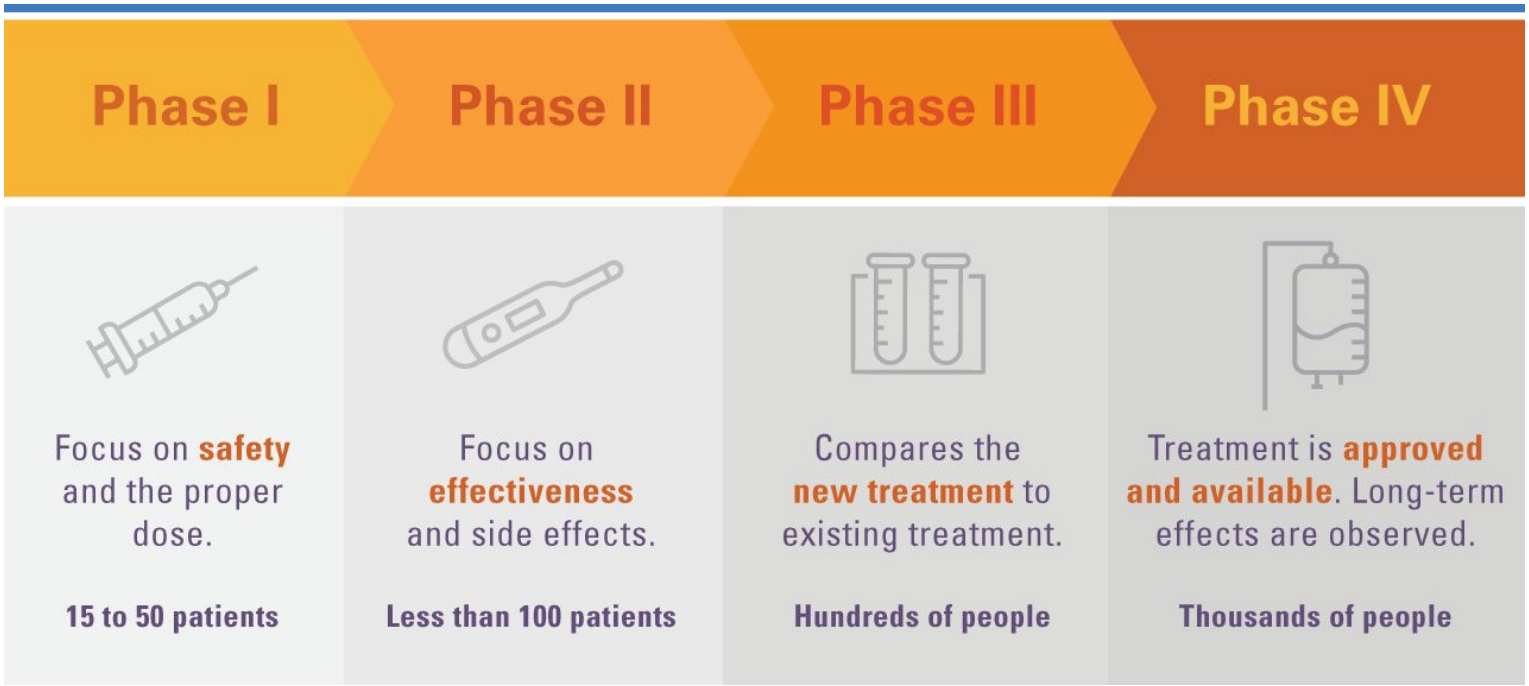
Both algorithms satisfy, for all $\boldsymbol{p} = (p_1, \ldots, p_K)$,

$$\mathbb{P}\left(B_T \neq a_\star\right) \leq C_K \exp\left(-\frac{C \times T}{H(\boldsymbol{p}) \log(K)}\right), \text{ with } H(\boldsymbol{p}) = \sum_{a \neq a_\star} \frac{1}{(p_\star - p_a)^2}$$

➜ error decays exponentially with a near-optimal exponent

Outline

The four phases of clinical trials



Early stage trials are about finding the right dose (or combinations of doses) of a given treatment.

Sequential dose-finding protocol

	Dose 1	Dose 2	$\cdots$	Dose K
toxicity probability	p_1	p_2	$\cdots$	p_K
efficacy probability	eff_1	eff_2	$\cdots$	eff_K

After selecting a dose $D_t \in \{1, \ldots, K\}$ ("arm") for patient t,

- observe whether un-desired side effects occur: $X_t \sim \mathcal{B}(p_{D_t})$

$$\mathbb{P}(X_t = 1 | D_t = a) = p_a \quad \mathbb{P}(X_t = 0 | D_t = a) = 1 - p_a$$

- observe whether the dose is efficient: $Y_t \sim \mathcal{B}(\mathrm{eff}_{D_t})$
  (in phase I/II designs only)

Depending on the context the optimal dose can have a different definition, but the *treatment* versus *identification* dilemma remains.

Adapting Thompson Sampling

Given toxicity and efficacy probabilities

$$\begin{aligned} \boldsymbol{p} &= (p_1, \ldots, p_K) \\ \boldsymbol{eff} &= (\text{eff}_1, \ldots, \text{eff}_K) \end{aligned}$$

define an optimal dose $a_\star = \text{OPT}\left(\boldsymbol{p}, \boldsymbol{eff}\right)$.

Thompson Sampling

Π_t: posterior distribution on $(\boldsymbol{p}, \boldsymbol{eff})$ after t rounds

At round $t+1$:

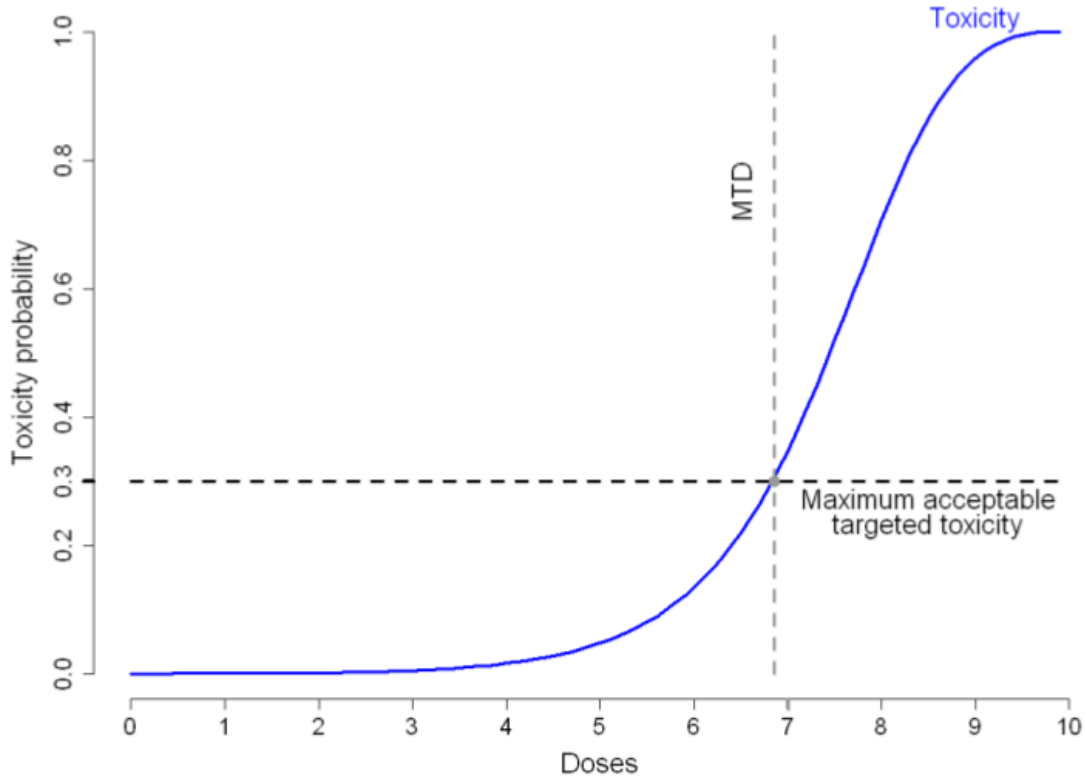
- sample $\left(\widetilde{\boldsymbol{p}}(t), \widetilde{\boldsymbol{eff}}(t)\right) \sim \Pi_t$
- $D_{t+1} = \text{OPT}\left(\widetilde{\boldsymbol{p}}(t), \widetilde{\boldsymbol{eff}}(t)\right)$

Outline

Maximum Tolerated Dose

Given a threshold θ, the Maximum Tolerated Dose is

$$\mathrm{MTD}(\boldsymbol{p}) = \arg\min_{a \in [K]} |p_a - \theta|$$



Context: phase I trials in oncology, assuming increasing efficacy

Choosing a prior

Parametric assumption: given two parameters $\beta_0, \beta_1 \in \mathbb{R}$,

$$p_a(\beta_0, \beta_1) = \frac{1}{1 + e^{-\beta_0 - \beta_1 u_a}}$$

u_a: effective dose (some carefully chosen parameter)

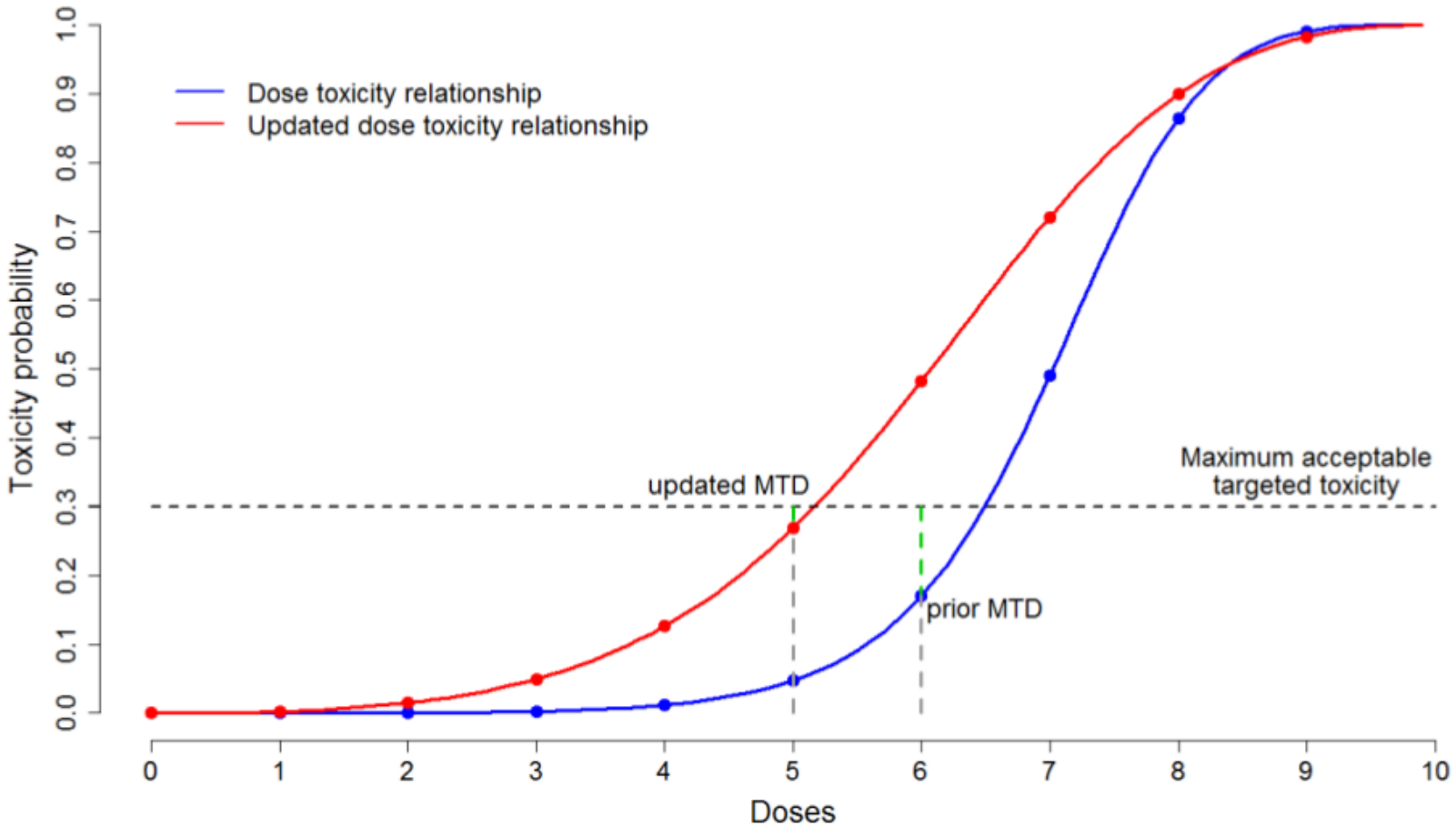
➜ enforces increasing toxicities

Bayesian model: $(\beta_0, \beta_1) \sim \pi$, e.g.

$$\beta_0 \sim \mathcal{N}(0, 100) \quad \text{and} \quad \beta_1 \sim \mathrm{Exp}(1).$$

➜ the posterior distribution π_t on (β_0, β_1) can be sampled from using, e.g. Hamiltonian Monte-Carlo methods

Illustration of the posterior update



source: Marie-Karelle Riviere (PhD thesis)

Thompson Sampling versus the CRM

Thompson Sampling

$$\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t)\right) \sim \pi_t,$$

$$D_{t+1}^{\mathsf{TS}} \in \arg\min_{a \in [K]} \left| \theta - p_a\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t)\right) \right|$$

Continual Reassesment Method (CRM) [O'Quingley et al., 1990]

$$\hat{\beta}_i(t) = \int_{\mathbb{R}} \beta_i d\pi_t(\beta_0, \beta_1) \quad \text{(posterior mean)}$$

$$D_{t+1}^{\mathsf{CRM}} \in \arg\min_{a \in [K]} \left| \theta - p_a\left(\hat{\beta}_0(t), \hat{\beta}_1(t)\right) \right|$$

➜ compared to the existing CRM, TS is adding exploration

Empirical performance

Too much exploration may be un-ethical ➜ two variants of TS restricting the set of doses that can be chosen

$$\boldsymbol{T = 36}\ \textbf{patients}\ ,\ K = 6\ \text{doses}\ ,\ \theta = 0.3$$

Sc. 5: Tox prob	0.10	0.25	0.40	0.50	0.65	0.75	0.10	0.25	0.40	0.50	0.65	0.75
3 + 3 [3.1]	20.6	30.8	24.2	15.3	5.1	0.8	-	-	-	-	-	-
CRM	4.8	49.7	39.0	6.5	0.1	0.0	17.8 (18.2)	38.3 (27.4)	30.9 (23.9)	9.0 (14.8)	2.4 (5.5)	1.7 (4.0)
TS	4.3	**50.7**	39.4	5.4	0.1	0.1	26.3 (17.6)	31.2 (17.5)	22.3 (16.0)	8.8 (11.4)	3.2 (5.4)	8.2 (7.2)
TS(ϵ)	4.8	**52.2**	36.5	6.2	0.2	0.0	18.8 (19.3)	**41.2** (27.1)	29.7 (24.4)	7.3 (13.7)	1.4 (4.2)	1.6 (3.9)
TS_A	3.0	**50.8**	36.4	7.0	1.6	1.1	29.6 (20.0)	**40.1** (18.8)	23.4 (18.5)	6.1 (11.0)	0.8 (3.2)	0.1 (1.1)
Independent TS	24.3	32.6	21.4	14.6	5.4	1.6	19.4 (10.5)	22.6 (10.8)	19.1 (10.0)	16.0 (9.1)	12.5 (7.0)	10.4 (5.5)

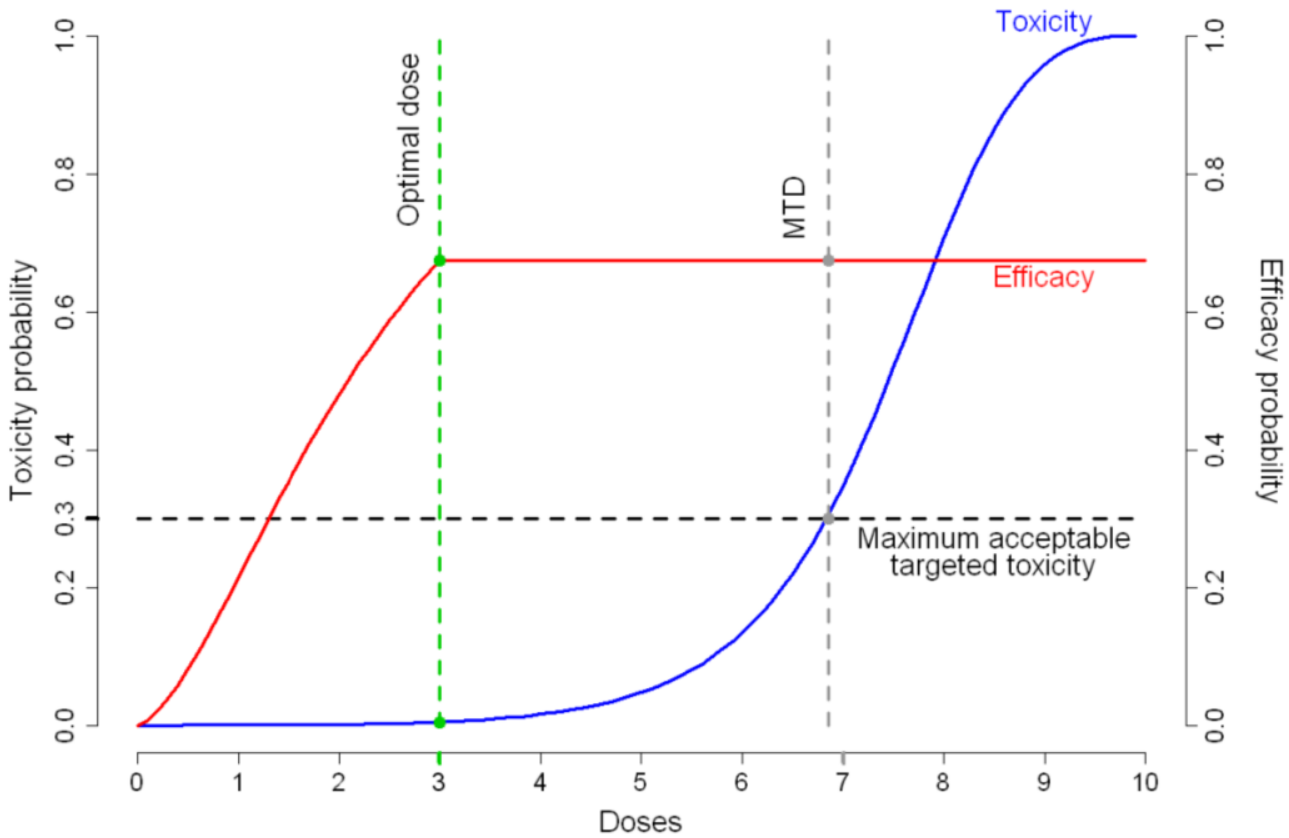
% of recommendation (left) and allocation (right)
(average over 2000 repetitions)

[Aziz et al., 2021]

Outline

A two-dimensional structured bandit

For certain agents, a plateau of efficacy is observed, which motivates the search of the Minimal Effective Dose (MED)



$$\mathrm{MED}\,(\boldsymbol{p}, \boldsymbol{eff}) = \min \left\{ a \in [K] : \mathrm{eff}_a = \max_{\ell : p_\ell \leq \theta} \mathrm{eff}_\ell \right\}$$

Choosing a prior

Toxicity: $p_a(\beta_0, \beta_1) = \frac{1}{1+e^{-[\beta_0+\beta_1 u_a]}}$

$$\beta_0 \sim \mathcal{N}(0, 100), \quad \beta_1 \sim \text{Exp}(1)$$

Efficacy: τ indicates the beginning of the plateau

$$\text{eff}_a(\gamma_0, \gamma_1, \tau) = \frac{1}{1 + e^{-[\gamma_0+\gamma_1(v_a \mathbb{1}(a<\tau)+v_\tau \mathbb{1}(a\geq\tau))]}}$$

$$\gamma_0 \sim \mathcal{N}(0, 100), \quad \gamma_1 \sim \text{Exp}(1), \quad \tau \sim (1/K, \ldots, 1/K).$$

Thompson Sampling

$$\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t), \tilde{\gamma}_0(t), \tilde{\gamma}_1(t), \tilde{\tau}(t)\right) \sim \pi_t,$$

$$D_{t+1}^{\text{TS}} \in \text{MED}\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t), \tilde{\gamma}_0(t), \tilde{\gamma}_1(t), \tilde{\tau}(t)\right)$$

Results

Competitive results wrt. the state-of the art MTA-RA algorithm [Riviere et al., 2017]

$$\boldsymbol{T} = 60 \textbf{ patients},\ K = 6 \text{ doses},\ \theta = 0.35$$

Table 4: Results for MED identification (part 1/3).

Algorithm	E-Stop	Recommended						Allocated					
		1	2	3	4	5	6	1	2	3	4	5	6
Sc. 1: Tox prob		0.01	0.05	0.15	0.2	0.45	0.6	0.01	0.05	0.15	0.2	0.45	0.6
Sc. 1: Eff prob		0.1	0.35	0.6	0.6	0.6	0.6	0.1	0.35	0.6	0.6	0.6	0.6
MTA-RA	0.4	0.4	7.0	54.9	29.1	7.4	0.8	7.1	14.2	37.9	24.9	12.9	2.5
								(3.8)	(13.9)	(24.4)	(18.8)	(13.6)	(4.9)
TS	0.9	0.1	9.7	**57.6**	27.0	4.2	0.4	10.6	18.4	31.9	23.8	10.0	4.4
								(5.7)	(11.0)	(14.4)	(13.2)	(8.0)	(4.5)
TS_A	0.9	0.3	9.6	**59.4**	26.1	3.5	0.2	10.7	20.7	35.7	23.9	7.3	0.9
								(5.4)	(12.9)	(14.9)	(14.1)	(8.1)	(2.7)

% of recommendation (left) and allocation (right)
(average over 2000 repetitions)

[Aziz et al., 2021]

Conclusion

Thompson Sampling is a flexible design that can be used

- whenever there is a proper notion of optimal dose/treatment
- whenever there is a posterior that can be sampled from

But its strong theoretical guarantees only hold for simple (e.g. product) priors and large sample sizes...

Insights from the bandit literature:

- we need exploration (*do we always?*)
- it is not possible to be optimal for treatment and identification at the same time (*but are existing designs good trade-offs?*)


Agrawal, S. and Goyal, N. (2013).
Further Optimal Regret Bounds for Thompson Sampling.
In Proceedings of the 16th Conference on Artificial Intelligence and Statistics.


Audibert, J.-Y., Bubeck, S., and Munos, R. (2010).
Best Arm Identification in Multi-armed Bandits.
In Proceedings of the 23rd Conference on Learning Theory.


Aziz, M., Kaufmann, E., and Riviere, M. (2021).
On multi-armed bandit designs for dose-finding clinical trials.
Journal of Machine Learning Research, 22(14):1–38.


Chapelle, O. and Li, L. (2011).
An empirical evaluation of Thompson Sampling.
In Advances in Neural Information Processing Systems.


Karnin, Z., Koren, T., and Somekh, O. (2013).
Almost optimal Exploration in multi-armed bandits.
In International Conference on Machine Learning (ICML).


Kaufmann, E., Korda, N., and Munos, R. (2012).
Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis.
In Proceedings of the 23rd conference on Algorithmic Learning Theory.


Lai, T. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6(1):4–22.

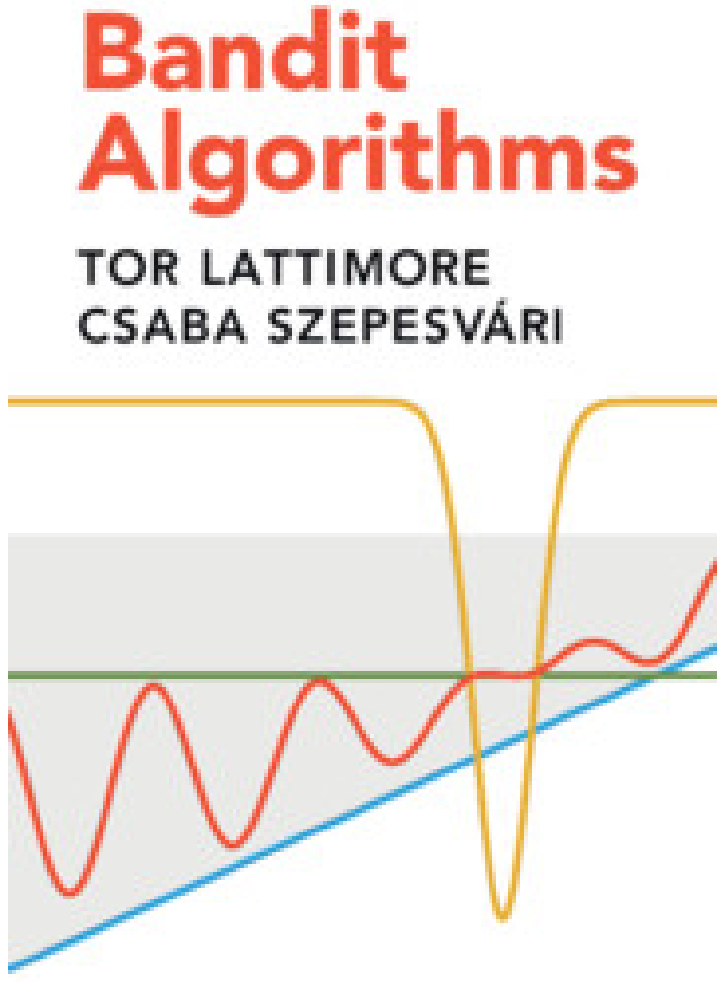

Lattimore, T. and Szepesvari, C. (2019).
Bandit Algorithms.
Cambridge University Press.


O'Quingley, J., Pepe, M., and Fisher, L. (1990).
Continual reassessment method: A practical design for phase I clinical trials in cancer.
Biometrics, 46(1):33–48.


Riviere, M.-K., Yuan, Y., Jourdan, J.-H., Dubois, F., and Zohar, S. (2017).
Phase i/ii dose-finding design for molecularly targeted agent: Plateau determination using adaptive randomization.
Statistical Methods in Medical Research.


Scott, S. (2010).
A modern Bayesian look at the multi-armed bandit.
Applied Stochastic Models in Business and Industry, 26:639–658.


Thompson, W. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25:285–294.



The Bandit Book

by [Lattimore and Szepesvari, 2019]

Details on the variants of Thompson Sampling

TS(ε) outputs a dose that belongs to the set

$$\left\{a \in [K] : \left| p_a(\hat{\beta}_0(t), \hat{\beta}_1(t)) - p_{\mathrm{MTD}(\hat{\beta}_0(t), \hat{\beta}_1(t))}(\hat{\beta}_0(t), \hat{\beta}_1(t)) \right| \leq \varepsilon \right\}$$

$$(\varepsilon = 0.05)$$

TS_A outputs a dose that belongs to the set

$$\left\{a \in [K] : \mathbb{P}_{(\beta_0, \beta_1) \sim \pi_t} \left(p_a(\beta_0, \beta_1) > p_{\mathrm{MTD}(\beta_0, \beta_1)}(\beta_0, \beta_1) \right) \leq c_1 \right\}$$

$$(c_1 = 0.8)$$