Dose Finding via Multi Arm Bandits: A Review

Riddhiman Bhattacharya

Department of Biostatistics
Duke University

October 4, 2024

Adaptive Clinical Trials

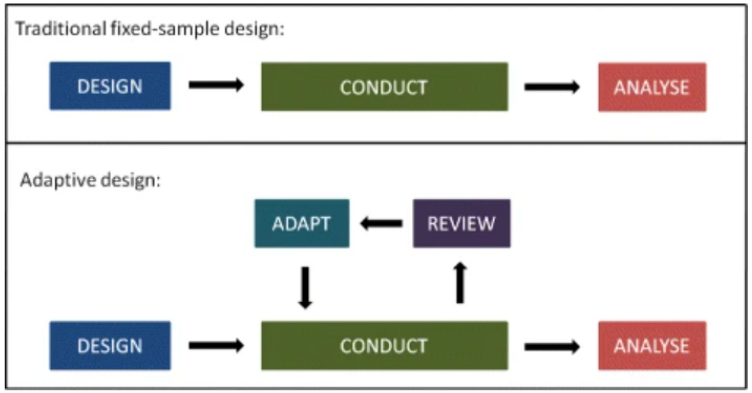
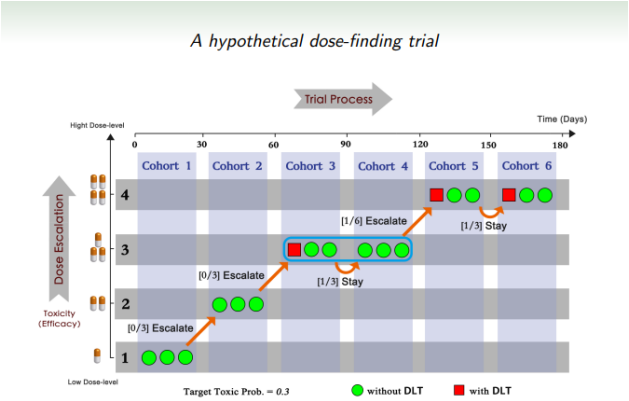


Figure: Schematic of a traditional clinical trial design with fixed sample size, and an adaptive design with pre-specified review(s) and adaptation(s)

- Can make clinical trials more flexible.
- Are often more efficient, informative and ethical than fixed design trials.

Dose Escalation Models

- Dose escalation in phaseI/II of clinical trials entails finding maxmimum tolerated dose (MTD) among increasing dose levels.



- Depending on toxic effects observed currently, decision is taken whether to escalate the dose or not.
- Has high prevalence in toxic treatments- e.g. chaemotheraphy in cancer trials, etc.

Dose Escalation Models

- K dose levels chosen by physicians via preliminiary experiments.
- p_k-toxicity probability, unknown.
- θ-pre-specified target toxicity probability. Usually between .2 and .35 for clinical trials.
- MTD$=k^* = \text{argmin}_{k \in \{1,2,3,\cdots,K\}} |\theta - p_k|$.
- Implicit Assumption: efficacy increasing with toxicity.

MTD: Background

- MTD identification proceeds sequentially.
- At round t a dose $D_t \in \{1, 2, \cdots, K\}$ is selected and administered to a patient for whom a toxicity response is observed.
- A binary outcome X_t is revealed indicating toxicity or not- $X_t = 1$ implies toxicity and $X_t = 0$ implies not.
- For fixed design trials nCRM, BOIN, mTP, etc are used for dose finding.

MTD in Adaptive Clinical Trials

- Key difference between fixed designs and adaptive-sampling scheme.
- Fixed designs- random sample gives inferential findings.
- Adaptive designs-sampling and inference/learning happens in a balanced manner based on data history.
- Reinforcement learning tailor made in adaptive setting.

Reinforcement Learning

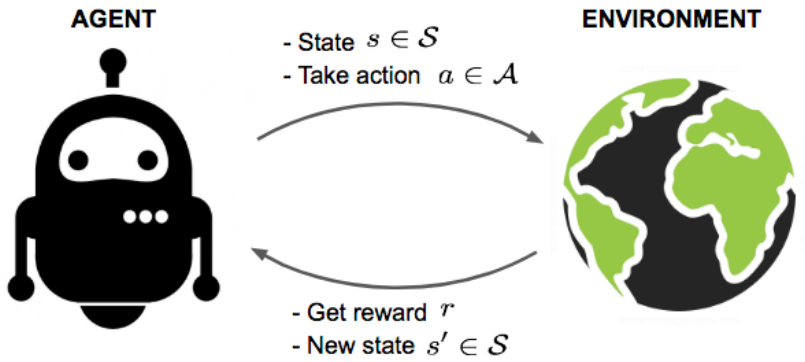


Figure: RL Illustration

- Person/agent interacts with environment to know more about it.
- Two types of Reinforcement Learning- Online Learning and Offline Learning-focus on online.
- Two different approaches: MDPs and bandits- focus on bandits.

Thompson Sampling

- First algorithm for bandits is Thompson sampling, 1933.
- Thompson showed empirical findings.
- Bayesian approach to bandits.
- Positives- known theory and more stable than UCB (recent work by Jeevi eta al.).
- Negatives-mostly intractable posteriors leading to approximate sampling (recent work by Michael Jordan's group).

Thompson Sampling

Input: Bayesian bandit environment $(\mathcal{E}, \mathcal{B}(\mathcal{E}), Q, P)$.
for: $t = 1, 2, \cdots, n$ **do**
Sample $\nu_t \sim Q(\cdot \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1})$
Choose $A_t = \text{argmax}_{i \in [k]} \mu_i(\nu_t)$.
end for

Algorithm 1: Thompson Sampling Algorithm

- Key idea: Given data, sample from posterior and take action which maximizes the average posterior reward given the sample.

Thompson Sampling: Example

Input: Bayesian bandit environment $(\mathcal{E}, \mathcal{B}(\mathcal{E}), Q, P)$.
for: $t = 1, 2, \cdots, n$ **do**
for: $k = 1, 2, \cdots, K$ **do**
Sample $\hat{\theta}_k \sim \text{beta}(\alpha_k, \beta_k)$ (some hyper-parameters for the k-th arm.
end for
Choose $A_t = \text{argmax}_k \hat{\theta}_k$.
Pull A_t to get reward r_{A_t}.
Update: $(\alpha_{A_t}, \beta_{A_t}) \leftarrow (\alpha_{A_t} + r_{A_t}, \beta_{A_t} + 1 - r_{A_t})$.

Algorithm 2: Thompson Sampling: Bernoulli Bandit

- Bernoulli bandit where at each stage we sample the success probability of each arm and sample the reward from the arm with the highest success probability.

Regret of Thompson Sampling

- Thompson Sampling begets two regrets-frequentist and Bayesian.
- The frequentist regret

$$R_n(\pi, \eta) = n\mu^* - \sum_{t=1}^{T} \mathbb{E}[X_t]$$

-bandit instance dependnet.

- The Bayesian bandit regret is given as

$$\mathrm{BR}_n(\pi, Q) = \int_{\mathcal{E}} R_n(\pi, \eta) dQ(\eta)$$

-average over the bandit instance.

Bandit Model for MTD

- At round t select dose $D_t \in \{1, 2, 3, \cdots, K\}$.
- A binary outcome X_t is revealed where $X_t = 1$ implies toxicity and is 0 o.w.
- $X_t \sim \text{Ber}(p_{D_t})$, independent of previous observations.
- $N_k(t) = \sum_{s=1}^{t} \mathbf{1}_{\{D_s = k\}}$ number of times dose k is selected.

Bandit Model for MTD

- Prior distribution on $\mathbf{p} = (p_1, p_2, \cdots, p_K)$ is $\Pi^0 = \prod_{i=1}^{K} \pi_k^0$ with $\pi_k^0 = \text{Unif}([0,1])$.
- Generate a dose at each time instance $\forall k \theta_k(t) \sim \pi_k^t$
- $D_{t+1} = \arg\min_k |\theta_k(t) - \theta|$.

Bandit Model for MTD

- Under an identifiability condition for the optimal dose one has

$$\mathbb{E}[N_k(n)] \leq O(\log n).$$

- Further, one has

$$\lim \inf_{n \to \infty} \frac{\mathbb{E}[N_k(n)]}{\log n} \geq \frac{1}{kl(p_k, d_k^*)}$$

where d_k^* is the gap and p_k is the toxicity probability.

- Finally,

$$\mathbb{P}\left(\hat{k}_n \neq k^*\right) = O(\log n).$$

Key Takeaways

- Thompson Sampling works well in dose escalation model adaptive designs.
- Produces sub-linear regret. Tight bound.
- Correct dose estimated at end of trial with high probability.