Dose Optimization via Multi-armed Bandits

I. Model Description

We consider the phase I trial involving K doses with a Maximum Tolerated Dose (MTD) denoted as θ. The performance of each dose is characterized by efficacy as well as toxicity. We model the efficacy and toxicity for a particular dose $d_k \in \mathcal{D}$ as Bernoulli random variables, whose probabilities are denoted as q_k and p_k respectively. At each round t (i.e., for the t-th patient), dose $I(t)$ is selected based on some policy and administered to the patient. The corresponding efficacy and toxicity are then observed. Binary outcomes X_t and Y_t are revealed, where $X_t = 1$ indicates that the dose level is effective and $Y_t = 1$ means a harmful side-effect has occurred, while $X_t = 0$ and $Y_t = 0$ indicates the dose level is not effective and no harmful effect occurs. The total number of patients, i.e., the whole budget, is fixed in advance.

The toxicity probability for different dose levels can be modeled as a simple one-parameter dose-response logistic model as [1]:

$$p_k(a) = \left(\frac{\tanh d_k + 1}{2}\right)^a, \tag{1}$$

where a can be seen as a global parameter for all the dose levels. Further, we make an assumption on the monotonicity and continuity of function (1).

Assumption 1. 1) *For each $d_k \in \mathcal{D}$ and $a, a' \in \mathcal{A}$ there exists $C_{1,k} > 0$ and $1 < \gamma_{1,k}$, such that:*

$$|p_k(a) - p_k(a')| \geq C_{1,k}|a - a'|^{\gamma_{1,k}}.$$

2) *For each $d_k \in \mathcal{D}$ and $a, a' \in \mathcal{A}$ there exists $C_{2,k} > 0$ and $0 < \gamma_{2,k} \leq 1$, such that:*

$$|p_k(a) - p_k(a')| \leq C_{2,k}|a - a'|^{\gamma_{2,k}}.$$

Under Assumption 1, the following proposition can also be deduced.

Proposition 2. 1) *For each $d_k \in \mathcal{D}$, the function $p_k(a)$ is invertible and the corresponding inverse function is denoted as $p_k^{-1}(\cdot)$;*

2) *For each $d_k \in \mathcal{D}$ and $a, a' \in \mathcal{A}$, there exists*

$$|p_k^{-1}(d) - p_k^{-1}(d')| \leq \bar{C}_{1,k}|d - d'|^{\bar{\gamma}_{1,k}}$$

under Assumption 1, where $\bar{\gamma}_{1,k} = \frac{1}{\gamma_{1,k}}$, $\bar{C}_{1,k} = (\frac{1}{C_{1,k}})^{\frac{1}{\gamma_{1,k}}}$.

We do not make any assumption on the efficacy model. This is in contrast to existing literature such as [2]. The reason is XXX (Sofia, please help us here).

A. Conservative Model

1) Problem Formulation: In this model the goal is to maximize the cumulative efficacy over the entire set of patients and simultaneously guarantee that the average toxicity observed from all the patients is kept under the MTD with a high probability. This can be explicitly written as:

$$\text{maximize } \sum_{t=1}^{n} X_t, \tag{2}$$

$$\text{subject to } \mathbb{P}\left[\frac{1}{n}\sum_{t=1}^{n} Y_t \leq \theta\right] \geq 1 - \delta. \tag{3}$$

The performance of the dose allocation policy can be evaluated using a regret formation as:

$$R(n) = q^* n - \mathbb{E}\left[\sum_{t=1}^{n} q_{I(t)}\right], \tag{4}$$

$$e(n) = \mathbb{P}\left[\frac{1}{n}\sum_{t=1}^{n} p_{I(t)} > \theta\right], \tag{5}$$

where $q^* = q_{k^*}$, with k^* representing the most suitable dose level satisfying $k^* = \arg\max_{k:p_k \le \theta} q_k$. The goal is to minimize $R(n)$ and keep $e(n) \le \delta$ at the same time.

Algorithm 1 Conservative Dose Allocation

Input: $p_k(a)$ for each $d_k \in \mathcal{D}$, θ as MTD,

Initialize: $t = 1, N_k(1) = 0, \hat{p}_k(1) = 0, \hat{q}_k(1) = 0$ for each $d_k \in \mathcal{D}$

1: **while** $t \le n$ **do**

2: Set the available set as $\mathcal{D}_1(t) = \{d_k \in \mathcal{D} : p_k(\hat{a}(t-1) - \alpha(t)) \le \theta\}$

3: Select dose level $I(t) = \arg\max_{d_k \in \mathcal{D}_1(t)} F(\hat{q}_k(t), N_k(t), t)$,

4: Observe the revealed outcomes X_t and Y_t, update the corresponding estimations:
$\hat{q}_{I(t)}(t) = \frac{\hat{q}_{I(t)}(t-1)N_{I(t)}(t-1)+X_t}{N_{I(t)}(t-1)+1}$, $\hat{p}_{I(t)}(t) = \frac{\hat{p}_{I(t)}(t-1)N_{I(t)}(t-1)+Y_t}{N_{I(t)}(t-1)+1}$, $N_{I(t)}(t) = N_{I(t)}(t-1) + 1$,

5: Update the estimation for the unknown parameter:
$\hat{a}_{I(t)}(t) = \arg\min |p_k(a) - \hat{p}_{I(t)}(t)|$, $w_k(t) = N_k(t)/t$, $\forall d_k \in \mathcal{D}$, $\hat{a}(t) = \sum_{k=1}^{K} w_k(t)\hat{a}_k(t)$,

6: $t = t + 1$.

7: **end while**

2) Algorithm Description: In the algorithm, the confidence interval of the estimated efficacy and toxicity are constructed as:

$$\text{KL-UCB index: } F(p, s, n) = \sup\{q \ge p : sI(p, q) \le \log(n) + \log\log(n)\}, \tag{6}$$

$$I(p, q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, \tag{7}$$

$$\text{UCB-1 index: } F(p, s, n) = p + \sqrt{\frac{c\log(n)}{s}}, \tag{8}$$

$$\alpha(t) = \bar{C}_1 K \left(\frac{\ln\frac{2K}{\delta}}{2t}\right)^{\frac{\gamma_1}{2}}. \tag{9}$$

3) Performance Analysis:

Lemma 3.

$$\mathbb{P}\left[p_k(\hat{a}(t-1) - \alpha(t)) > \theta\right] \le \delta, \textit{ for any } p_k(a) < \theta, \tag{10}$$

Proof.

$$\begin{aligned}
&\mathbb{P}[\hat{a}(t-1)-\alpha(t)>p_i^{-1}(\theta)] \le \mathbb{P}[\hat{a}(t-1)-\alpha(t)>a] = \mathbb{P}[\hat{a}(t-1)-a>\alpha(t)] \\
&\le \sum_{k=1}^{K} \mathbb{P}\left[|\hat{p}_k(t)-p_k(a)| > \left(\frac{\alpha(t)}{w_k(t)\bar{C}_1 K}\right)^{\gamma_1}\right] \\
&\le \sum_{k=1}^{K} 2\exp\left(-2N_k(t)\left(\frac{\alpha(t)}{w_k(t)\bar{C}_1 K}\right)^{2\gamma_1}\right) \\
&\le 2K\exp\left(-2\left(\frac{\alpha(t)}{\bar{C}_1 K}\right)^{2\gamma_1} t\right) = \delta.
\end{aligned}$$

□

Lemma 4. *If* $t > t_1 = \frac{1}{2}\left(\frac{\bar{C}_1 K}{\Delta_i - \epsilon}\right)^{2\gamma_1} \log\frac{2K}{\delta}$*,then:*

$$\mathbb{P}\left[p_k(\hat{a}(t-1)-\alpha(t)) \le \theta\right] \le \exp(-2t\epsilon^2), \text{ for any } p_k(a) > \theta, \tag{11}$$

Proof. Based on Hoeffding's Inequality, from inequality (11), we have:

$$\alpha(t) \le a - a_i(\theta) - \epsilon = \Delta_i - \epsilon,$$

where $a_i(\theta) = p_i^{-1}(\theta)$. When $t > t_1$, the above could be derived. □

Theorem 5.

Proof. We first decompose the regret as:

$$\begin{aligned}
R(n) &= \sum_{t=1}^{n} \mathbb{P}[k^* \notin \mathcal{D}_1(t)]\Delta_q + \mathbb{P}[k^* \in \mathcal{D}_1(t)]R_2(n) && (12)\\
&\le T\delta\Delta_q + R_2(n), && (13)\\
R_2(n) &= t_1 + (K-M)\sum_{t=1}^{n}\exp(-2t\epsilon^2) + \sum_{t=t1+1}^{n}\sum_{d_k:p_k\le\theta} I(t)=k && (14)\\
&\le t_1 + \frac{K-M}{2\epsilon^2} + \sum_{d_k:p_k\le\theta}\frac{q^*-q_k}{I(q_k,q^*)}\log(n) && (15)
\end{aligned}$$

□

Theorem 6.

$$\mathbb{P}\left[\frac{1}{n}\sum_{t=1}^{n} p_{I(t)} - \theta \le \epsilon\right] \ge 1-\delta$$

Proof.

$$\begin{aligned}
p_{I(t)}(a) - \theta &\le p_{I(t)}(a) - \theta + \theta - p_{I(t)}(a-\alpha(t)) \\
&\le C_2|a - \hat{a} + \alpha(t)|^{\gamma_2}, \\
\mathbb{P}[\hat{a}(t) - a > \alpha(t) + \epsilon] &\le \exp(-2t(\alpha(t)+\epsilon)^2) \\
\mathbb{P}\left[\frac{1}{n}\sum_{t=1}^{n} p_{I(t)}(a) - \theta < C_2\epsilon^{\gamma_2}\right] &\ge 1 - \exp(-2t(\alpha(t)+\epsilon)^2) \ge 1-\delta.
\end{aligned}$$

□

B. Aggressive Model

1) Problem formulation: In this model, we emphasize more on the efficacy. The goal is to maximize the cumulative efficacy over the entire set of patients, and minimize the number of times dose levels exceeding the MTD threshold are administered.

$$\begin{aligned} &\text{maximize } \sum_{t=1}^{n} X_t, \\ &\text{minimize } \sum_{k:p_k>\theta} N_k(n). \end{aligned} \tag{16}$$

Alternatively the goal can be set as to maximize the cumulative efficacy while identifying MTD with a given accuracy:

$$\begin{aligned} &\text{maximize } \sum_{t=1}^{n} X_t, \\ &\mathbb{P}\left(\hat{k}_n = \arg\min_{k\in\mathcal{K}} |p_k - \theta|\right) \geq 1-\delta. \end{aligned} \tag{17}$$

The expected cumulative regret is correspondingly set as:

$$\begin{aligned} R^{eff}(n) &= q^* n - \mathbb{E}\left[\sum_{t=1}^{n} q_{I(t)}\right], \\ N^{tox}(n) &= \sum_{k:p_k>\theta} \sum_{t=1}^{n} \mathbb{1}\{I(t)=k\}. \end{aligned}$$

where $q^* = q_{k^*}$ with k^* representing the most suitable dose level satisfying $k^* = \arg\max_{d_k\in\mathcal{D}} q_k$. According to (16), the goal of our policy is therefore to minimize $R^{eff}(n)$ while minimizing $N^{tox}(n)$ at the same time. This can be seen as a two objective bandit model with two regret minimization goals. As for (17), the goal is to minimize $R^{eff}(n)$ and guarantee the output dose is within the given accuracy within the given horizon, which can be seen as a combination of regret minimization and best arm identification with fixed confidence.

As efficacy is non-structured and toxicity could be seen as globally informative, the toxicity tends to be easier to be estimated accurately enough.

2) Algorithm Description: The following is the proposed algorithm for the aggressive model.

Algorithm 2 Aggressive Dose Allocation

Input: $p_k(a)$ for each $d_k \in \mathcal{D}$, θ as MTD,

Initialize: $t=1, N_k(1)=0, \hat{p}_k(1)=0, \hat{q}_k(1)=0$ for each $d_k \in \mathcal{D}$

while $t \leq n$ **do**

  Set optimal dose level in terms of efficacy as $d^1(t) = \arg\max F(\hat{q}_k(t), N_k(t), t)$,

  if Confidence interval of dose level $F(\hat{q}_{d^1(t)}(t), N_{d^1(t)}(t), t) - \hat{q}_{d^1(t)}(t) > \beta$ **then**

    Select dose level $I(t) = d^1(t)$

  else

    $I(t) = \arg\max_{d_k: p_k(\hat{a}(t)+\alpha_k(t))\leq\theta} F(\hat{q}_k(t), N_k(t), t)$

  end if

  Observe efficacy X_t and toxicity Y_t, update the estimations:
  $\hat{q}_{I(t)}(t) = \frac{\hat{q}_{I(t)}(t-1)N_{I(t)}(t-1)+X_t}{N_{I(t)}(t-1)+1}$, $\hat{p}_{I(t)}(t) = \frac{\hat{p}_{I(t)}(t-1)N_{I(t)}(t-1)+Y_t}{N_{I(t)}(t-1)+1}$, $N_{I(t)}(t) = N_{I(t)}(t-1)+1$,

  Update the estimation for the unknown parameter:
  $\hat{a}_{I(t)}(t) = \arg\min |p_{I(t)}(a) - \hat{p}_{I(t)}(t)|$, $w_k(t) = N_k(t)/t$, $\forall d_k \in \mathcal{D}$, $\hat{a}(t) = \sum_{k=1}^{K} w_k(t)\hat{a}_k(t)$,

  $t = t+1$.

end while

REFERENCES

[1] J. O'Quigley, M. Pepe, and L. Fisher, "Continual reassessment method: a practical design for phase 1 clinical trials in cancer," *Biometrics*, vol. 43, no. 1, pp. 33–48, 1990.

[2] M. Aziz, E. Kaufmann, and M.-K. Riviere, "On multi-armed bandit designs for phase i clinical trials," 2019, http://arxiv.org/abs/1903.07082v1.