ミニマックス最適かつベイズ最適な最適腕識別

Minimax and Bayes Optimal Best-Arm Identification
arXiv Preprint. URL: https://arxiv.org/abs/2506.24007.

Masahiro Kato
Mizuho-DL Financial Technology

MIZUHO

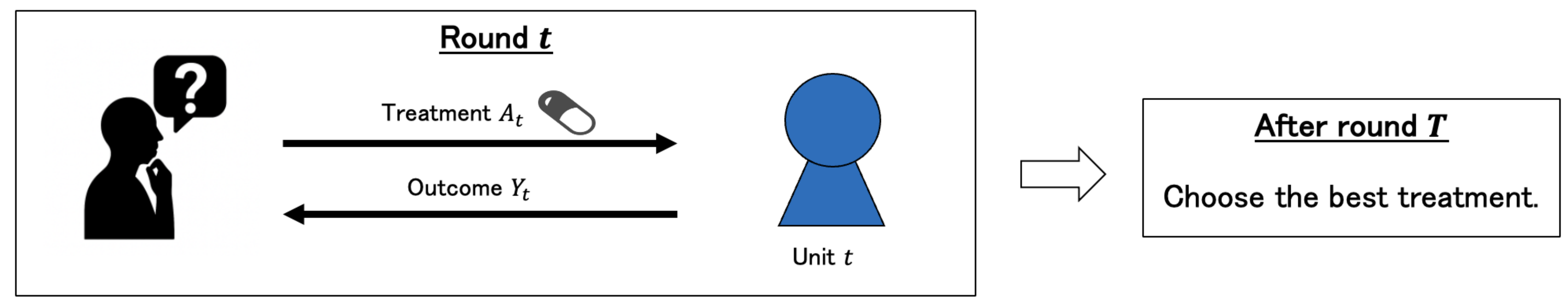
Best–Arm Identification (BAI) with a Fixed Budget

Setup

- K **Arms** indexed by $1,2,\ldots K$. Also referred to as **treatment arms** or **arms**.
- **Potential outcome.** $Y_a \in \mathbb{R}$.
- **Distribution.** Y_a follows a distribution $P_{\boldsymbol{\mu}}$, where $\boldsymbol{\mu} = (\mu_a)_{a\in[K]} \in \mathbb{R}^K$.
  - The mean of Y_a is μ_a (and only the mean can take different values).
  - For simplicity, let $P_{\boldsymbol{\mu}}$ be a Gaussian distribution with variance σ_a^2.
- **Goal.** Find the following **best arm** $a^*_{\boldsymbol{\mu}} \coloneqq \arg\max_{a\in\{1,2,\ldots,K\}} \mu_a$ efficiently:

Procedure

- **Adaptive experimental design with T rounds**
  1. **Allocation phase**: in each round $t = 1,2,\ldots,T$:
     - Draw arm $A_t \in [K] \coloneqq \{1,2,\ldots,K\}$ using $\{(A_s, Y_s)\}_{s=1}^{t-1}$.
     - Observe the outcome $Y_t = \sum_{a\in[K]} 1[A_t = a] Y_{a,t}$.
  2. **Decision-making phase**: at the end of the experiment:
     - Obtain an estimator $\hat{a}_T$ of the the best arm using $\{(X_t, A_t, Y_t)\}_{t=1}^{T}$.



Regret

- **Expected simple regret:** $\mathrm{Regret}_{\boldsymbol{\mu}} \coloneqq \mathbb{E}_{\boldsymbol{\mu}}\left[Y_{a^*_{\mu}} - Y_{\hat{a}_T}\right]$.
- Our goal is to minimize the expected simple regret.

Evaluation framework

- **Minimax analysis.**
  - Evaluate the regret for the worst-case $\boldsymbol{\mu}$.
  - We evaluate $\sup_{\boldsymbol{\mu}\in\mathbb{R}^K} \mathrm{Regret}_{\boldsymbol{\mu}}$
- **Bayes analysis.**
  - A prior Π is given.
  - Evaluate the regret by summing them weighted by the prior.
  - We evaluate $\int_{\mathbb{R}^K} \mathrm{Regret}_{\boldsymbol{\mu}}\, d\Pi(\mu)$.

Two-Stage (TS)–Empirical–Best–Arm (EBA) Strategy

- **Two-Stage-Empirical-Best-Arm (TS-EBA) Strategy $\delta^{\mathrm{TS}-EBA}$.**

1. Allocation phase

Split T into the first rT rounds and the second $(1-r)T$ rounds.

① **First stage**:

- Allocate each arm with equivalent ratio $\frac{rT}{K}$
- We select candidates of the best arms as $\hat{\mathcal{S}} \subseteq [K]$.
- Obtain an estimator $\hat{\sigma}^2_{a,rT}$ of the variance σ^2_a. (sample variance).

② **Second stage** (we do not allocate arm $a \notin \hat{\mathcal{S}}$ from this stage):

- If $|\hat{\mathcal{S}}| = 1$, return arm $a \in |\hat{\mathcal{S}}|$ as an estimate of the best arm.
- If $|\hat{\mathcal{S}}| = 2$, allocate arm $a \in \hat{\mathcal{S}}$ with probability $w_{a,t} \coloneqq \frac{\hat{\sigma}_{a,rT}}{\hat{\sigma}_{1,rT}+\hat{\sigma}_{2,rT}}$.
- If $|\hat{\mathcal{S}}| \geq 3$, allocate arm $a \in \hat{\mathcal{S}}$ with probability $w_{a,t} \coloneqq \frac{\hat{\sigma}^2_{a,rT}}{\sum_{b\in\hat{\mathcal{S}}}\hat{\sigma}^2_{b,rT}}$.

2. Decision-making phase:

- Return $\hat{a}_T = \arg\max_{a\in[K]} \hat{\mu}_{a,T}$.
  - $\hat{\mu}_{a,T} \coloneqq \frac{1}{\sum_{t=1}^T 1[A_t=a]} \sum_{t=1}^T 1[A_t = a]\, Y_t$ is the sample mean.

✓Remark: How to select candidates $\hat{\mathcal{S}}$?

Let $v_{a,qT} \coloneqq \sqrt{\frac{\log(T)}{T}}\,\hat{\sigma}_{a,rT}$, and $\hat{a}_{rT} = \arg\max_{a\in[K]} \hat{\mu}_{a,rT}$.
Construct lower and upper confidence bounds as $\hat{\mu}_{a,rT} - v_{a,qT}$ and $\hat{\mu}_{a,rT} + v_{a,qT}$.
Define $\hat{\mathcal{S}}$ as $\hat{\mathcal{S}} \coloneqq \{a \in [K] : \hat{\mu}_{a,rT} + v_{a,qT} \geq \hat{\mu}_{\hat{a}_{rT},rT} - v_{\hat{a}_{rT},rT}\}$.

Minimax and Bayes Optimality

- For the TS-EBA strategy, we prove the minimax and Bayes optimality.
- Let $\mathcal{E}$ be a set of consistent strategies. For any $\delta \in \mathcal{E}$, it holds that

$\mathbb{P}_\mu(\hat{a}^\delta_T \neq a^*_\mu) \to 0$.

Minimax optimality.

$$\begin{aligned}
&\lim_{T\to\infty} \sup_{\mu\in\mathbb{R}^K} \sqrt{T}\,\mathrm{Regret}_\mu^{\delta^{\mathrm{TS-EBA}}}\\
&\leq \max\left\{\frac{1}{\sqrt{e}}\max_{a\neq b}(\sigma_a + \sigma_b), \frac{K-1}{K}\sqrt{\sum_{a\in[K]}\sigma_a^2\log(K)}\right\}\\
&\leq \inf_{\delta\in\mathcal{E}} \lim_{T\to\infty} \sup_{\mu\in\mathbb{R}^K} \sqrt{T}\,\mathrm{Regret}_\mu^\delta.
\end{aligned}$$

Bayes optimality.

$$\begin{aligned}
&\lim_{T\to\infty} T \int_{\mu\in\mathbb{R}^K} \mathrm{Regret}_\mu^{\delta^{\mathrm{TS-EBA}}}\, d\Pi(\mu)\\
&\leq 4 \sum_{a\in[K]} \int_{\mathbb{R}^{K-1}} \sigma^{2*}_{\backslash\{a\}}\, h_a\left(\mu^*_{\backslash\{a\}} \mid \boldsymbol{\mu}_{\backslash\{a\}}\right) dH^{\backslash\{a\}}\left(\boldsymbol{\mu}_{\backslash\{a\}}\right)\\
&\leq \inf_{\delta\in\mathcal{E}} \lim_{T\to\infty} T \int_{\mu\in\mathbb{R}^K} \mathrm{Regret}_\mu^\delta\, d\Pi(\mu),
\end{aligned}$$

- $\sigma^{2*}_{\backslash\{a\}} = \sigma^2_{b^*}$ and $\mu^*_{\backslash\{a\}} = \mu_{b^*}$, where b^* be $\arg\max_{c\in[K]\backslash\{a\}} \mu_a$.
- $\boldsymbol{\mu}_{\backslash\{a\}}$ is a parameter vector which removes μ_a from $\boldsymbol{\mu}$. $H^{\backslash\{a\}}\left(\boldsymbol{\mu}_{\backslash\{a\}}\right)$ is a prior of $\boldsymbol{\mu}_{\backslash\{a\}}$.

Regret Decomposition

- The simple regret can be written as $\mathrm{Regret}_\mu = \mathbb{E}_\mu\left[Y_{a^*_\mu}\right] - \mathbb{E}_\mu\left[Y_{\hat{a}_T}\right] = \sum_{b\neq a^*_\mu}\left(\mu_{a^*_\mu} - \mu_b\right)\mathbb{P}_\mu(\hat{a}_T = b)$.
  - $\mathbb{P}_\mu(\hat{a}_T \neq a^*_\mu)$ roughly upper bounded by $\sum_{b\neq a^*_\mu} \exp\left(-C^*\, T\left(\mu_{a^*_\mu} - \mu_b\right)^2\right)$ for some constant $C^* > 0$.

Regret analysis: For simplicity, let $K = 2$. Define $\Delta_\mu = \mu_1 - \mu_2$.

- For some constant $C > 0$, $\mathrm{Regret}_\mu = \Delta_\mu(\hat{a}_T \neq a^*_\mu)$ can be written as $\mathrm{Regret}_\mu \approx \Delta_\mu \exp\left(-CT(\Delta_\mu)^2\right)$

  The dependency of Δ_μ on T affects the evaluation of the regret.

  - Δ_μ converges to zero with an order slower than $1/\sqrt{T}$ → For some increasing function g, $\mathrm{Regret}_\mu \approx \exp(-g(T))$.
  - Δ_μ converges to zero with an order of $1/\sqrt{T}$ → For some $C > 0$, $\mathrm{Regret}_\mu \approx \frac{C}{\sqrt{T}}$.
  - Δ_μ converges to zero with an order faster than $1/\sqrt{T}$ → $\mathrm{Regret}_\mu \approx o(1/\sqrt{T})$

→ In the worst case, $\mathrm{Regret}_\mu \approx \frac{C}{\sqrt{T}}$, where $\Delta_\mu = O(1\sqrt{T})$).