On the notion of Complexity in the Stochastic Multi-armed Bandit Problems

Pierre Ménard

Supervised by Aurélien Garivier and Gilles Stoltz

July 3, 2018

Chapters addressed in this presentation:

- Explore First, Exploit Next: The True Shape of Regret in Bandit Problems
- kl-UCB Algorithms for Exponential Families
- KL-UCB Algorithms for Bounded Rewards
- Thresholding Bandit for Dose-ranging: The Impact of Monotonicity
- Fano's inequality for random variables

K-armed bandit problem: parametric setting

Bernoulli rewards:

$$\underline{\nu} = \quad (\mathcal{B}(\mu_1), \quad \cdots \quad \mathcal{B}(\mu_a) \quad \cdots \quad , \mathcal{B}(\mu_K))$$

$\cdots$

$\cdots$

Game: for each round $1 \leqslant t \leqslant T$:

1. Player pulls arm $A_t \in \{1, .., K\}$.
2. He gets a reward $Y_t \sim \mathcal{B}(\mu_{A_t})$.

Regret

Player wants to maximize

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] ,$$

equivalently, minimize his regret

$$R_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] ,$$

where $\mu^\star = \max_{a=1,\dots,K} \mu_a$.

Regret

Player wants to maximize

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] ,$$

equivalently, minimize his regret

$$R_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] ,$$

where $\mu^\star = \max_{a=1,\dots,K} \mu_a$.
Chain rule

$$R_T = \sum_{a=1}^{K} (\mu^\star - \mu_a)\, \mathbb{E}\left[N_a(T)\right]$$

where $N_a(T) = \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}}$.

Regret

Player wants to maximize

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right],$$

equivalently, minimize his regret

$$R_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right],$$

where $\mu^\star = \max_{a=1,\dots,K} \mu_a$.
Chain rule

$$R_T = \sum_{a=1}^{K} (\mu^\star - \mu_a)\, \mathbb{E}[N_a(T)] (\sim T \text{ worst case}),$$

where $N_a(T) = \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}}$.

Ideas of strategy

- First idea: pull an arm uniformly at random at each round.
  $\Rightarrow$ Exploration $\Rightarrow R_T \sim T$

Ideas of strategy

- First idea: pull an arm uniformly at random at each round.
  $\Rightarrow$ Exploration $\Rightarrow R_T \sim T$

- Second idea: pull the current best empirical arm,

$$A_{t+1} = \operatorname{argmax}_{a \in \{1,\cdots,K\}} \widehat{\mu}_{a,N_a(t)} \qquad \widehat{\mu}_{a,N_a(t)} = \sum_{s=1}^{t} Y_s \, \mathbb{I}_{A_t=a} / N_a(t)$$

  $\Rightarrow$ Exploitation $\Rightarrow R_T \sim T$

Ideas of strategy

- First idea: pull an arm uniformly at random at each round.
  $\Rightarrow$ Exploration $\Rightarrow R_T \sim T$

- Second idea: pull the current best empirical arm,

$$A_{t+1} = \operatorname{argmax}_{a \in \{1,\cdots,K\}} \widehat{\mu}_{a,N_a(t)} \qquad \widehat{\mu}_{a,N_a(t)} = \sum_{s=1}^{t} Y_s \, \mathbb{I}_{A_t=a} / N_a(t)$$

$\Rightarrow$ Exploitation $\Rightarrow R_T \sim T$

$\Rightarrow$ Exploration-Exploitation tradeoff

$\Rightarrow R_T \sim \log(T)$

UCB algorithm

Algorithm 1: UCB

Initialization: Play each arm once.

For $t = K$ to $T - 1$, **do**

1. Compute for each arm a the upper confidence bound

$$U_a^{\mathrm{UCB}}(t) = \underbrace{\widehat{\mu}_{a,N_a(t)}}_{\text{Exploitation}} + \underbrace{\sqrt{\frac{\log(T)}{2N_a(t)}}}_{\text{Exploration}}$$

2. Play $A_{t+1} \in \operatorname{argmax}_{a \in \{1,\cdots,K\}} U_a^{\mathrm{UCB}}(t)$.

Upper Confident Bound

$$X_1, \cdots, X_n \text{ i.i.d.} \sim \mathcal{B}(\mu) \text{ with } \widehat{\mu}_n = \sum_{k=1}^{n} X_k / n$$

Hoeffding inequality for $x < \mu$

$$\mathbb{P}(\widehat{\mu}_n < x) \leqslant e^{-2n(x-\mu)^2}.$$

With probability at least $1 - \delta$

$$\mu \leqslant \widehat{\mu}_n + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Upper Confident Bound

$$X_1, \cdots, X_n \text{ i.i.d.} \sim \mathcal{B}(\mu) \text{ with } \widehat{\mu}_n = \sum_{k=1}^{n} X_k / n$$

Hoeffding inequality for $x < \mu$

$$\mathbb{P}(\widehat{\mu}_n < x) \leqslant e^{-2n(x-\mu)^2} .$$

With probability at least $1 - \delta$

$$\mu \leqslant \widehat{\mu}_n + \sqrt{\frac{\log(1/\delta)}{2n}} .$$

UCB index $\delta = 1/T$

$$U_a^{\mathrm{UCB}}(t) = \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{\log(T)}{2N_a(t)}}$$

Upper Confident Bound

$$X_1, \cdots, X_n \text{ i.i.d.} \sim \mathcal{B}(\mu) \text{ with } \widehat{\mu}_n = \sum_{k=1}^{n} X_k / n$$

Hoeffding inequality for $x < \mu$

$$\mathbb{P}(\widehat{\mu}_n < x) \leqslant e^{-2n(x-\mu)^2} .$$

With probability at least $1 - \delta$

$$\mu \leqslant \widehat{\mu}_n + \sqrt{\frac{\log(1/\delta)}{2n}} .$$

UCB index $\delta = 1/T$

$$U_a^{\mathrm{UCB}}(t) = \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{\log(T)}{2N_a(t)}}$$

Regret bound

Theorem

For the UCB algorithm, for all a such that $\mu^\star - \mu_a > 0$

$$\mathbb{E}[N_a(T)] \leqslant \frac{1}{2(\mu^\star - \mu_a)^2} \log(T) + o(\log(T)),$$

Regret bound

Theorem

For the UCB algorithm, for all a such that $\mu^\star - \mu_a > 0$

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{1}{2(\mu^\star - \mu_a)^2}\log(T) + o\big(\log(T)\big)\,,$$

therefore (Chain rule)

$$R_T \leqslant \sum_{a:\ \mu^\star > \mu_a} \frac{1}{2(\mu^\star - \mu_a)}\log(T) + o\big(\log(T)\big)\,.$$

Regret bound

Theorem

For the UCB algorithm, for all a such that $\mu^\star - \mu_a > 0$

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{1}{2(\mu^\star - \mu_a)^2}\log(T) + o\big(\log(T)\big)\,,$$

therefore (Chain rule)

$$R_T \leqslant \sum_{a:\ \mu^\star > \mu_a} \frac{1}{2(\mu^\star - \mu_a)}\log(T) + o\big(\log(T)\big)\,.$$

Is that the best we can do? $\Rightarrow$ Lower bound

Kullback-Leibler divergence

For two probability distributions P and Q

$$\mathrm{KL}(P, Q) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q & \text{if } P \ll Q \\ +\infty & \text{else.} \end{cases}$$

Example with Bernoulli

$$\mathrm{kl}(p, q) := \mathrm{KL}(\mathcal{B}(p), \mathcal{B}(q)) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

An asymptotic lower bound

Strategy which always pulls the same arm $\Rightarrow$ assumptions on the strategy.

An asymptotic lower bound

Strategy which always pulls the same arm $\Rightarrow$ assumptions on the strategy.

Definition

A strategy is consistent if for all bandit problems ν, for all suboptimal arms a, i.e., for all arms a such that $\mu^\star - \mu_a > 0$, it satisfies $\mathbb{E}\big[N_a(T)\big] = o(T^\alpha)$ for all $0 < \alpha \leqslant 1$.

An asymptotic lower bound

Strategy which always pulls the same arm $\Rightarrow$ assumptions on the strategy.

Definition

A strategy is consistent if for all bandit problems ν, for all suboptimal arms a, i.e., for all arms a such that $\mu^\star - \mu_a > 0$, it satisfies $\mathbb{E}\big[N_a(T)\big] = o(T^\alpha)$ for all $0 < \alpha \leqslant 1$.

Theorem (Asymptotic lower bound from Lai & Robbins)

For all consistent strategies, for all suboptimal arms a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)}.$$

Sketch of proof 1/2

a suboptimal arm ($\mu^\star - \mu_a > 0$).
Modified bandit problem with $\mu'_a > \mu^\star$:

$$\underline{\nu} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_a), .., \mathcal{B}(\mu_K))$$
$$\underline{\nu'} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu'_a), .., \mathcal{B}(\mu_K))$$

Information at time t: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

Sketch of proof 1/2

a suboptimal arm ($\mu^\star - \mu_a > 0$).
Modified bandit problem with $\mu'_a > \mu^\star$:

$$\underline{\nu} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_a), .., \mathcal{B}(\mu_K))$$
$$\underline{\nu'} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu'_a), .., \mathcal{B}(\mu_K))$$

Information at time t: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

$$\mathbb{E}_{\underline{\nu}}[N_a(T)]\, \mathrm{kl}(\mu_a, \mu'_a) = \mathrm{KL}(\mathbb{P}_{\underline{\nu}}^{Y_{1:T}}, \mathbb{P}_{\underline{\nu'}}^{Y_{1:T}})$$

Chain rule

Sketch of proof 1/2

a suboptimal arm ($\mu^\star - \mu_a > 0$).
Modified bandit problem with $\mu'_a > \mu^\star$:

$$\underline{\nu} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_a), .., \mathcal{B}(\mu_K))$$
$$\underline{\nu'} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu'_a), .., \mathcal{B}(\mu_K))$$

Information at time t: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

$$\mathbb{E}_{\underline{\nu}}[N_a(T)]\, \mathrm{kl}(\mu_a, \mu'_a) = \mathrm{KL}(\mathbb{P}_{\underline{\nu}}^{Y_{1:T}}, \mathbb{P}_{\underline{\nu'}}^{Y_{1:T}})$$

contraction of entropy

$$\geqslant \mathrm{KL}(\mathbb{P}_{\underline{\nu}}^{N_a(T)/T}, \mathbb{P}_{\underline{\nu'}}^{N_a(T)/T})$$

Sketch of proof 1/2

a suboptimal arm ($\mu^\star - \mu_a > 0$).
Modified bandit problem with $\mu'_a > \mu^\star$:

$$\underline{\nu} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_a), .., \mathcal{B}(\mu_K))$$
$$\underline{\nu'} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu'_a), .., \mathcal{B}(\mu_K))$$

Information at time t: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

$$\mathbb{E}_{\underline{\nu}}[N_a(T)]\,\mathrm{kl}(\mu_a, \mu'_a) = \mathrm{KL}(\mathbb{P}_{\underline{\nu}}^{Y_{1:T}}, \mathbb{P}_{\underline{\nu}'}^{Y_{1:T}})$$
$$\text{contraction of entropy} \qquad \geqslant \mathrm{KL}(\mathbb{P}_{\underline{\nu}}^{N_a(T)/T}, \mathbb{P}_{\underline{\nu}'}^{N_a(T)/T})$$
$$\text{projection} \qquad \geqslant \mathrm{kl}\Big(\mathbb{E}_{\underline{\nu}}[N_a(T)]/T, \mathbb{E}_{\underline{\nu}'}[N_a(T)]/T\Big)$$

Sketch of proof 1/2

a suboptimal arm ($\mu^\star - \mu_a > 0$).
Modified bandit problem with $\mu'_a > \mu^\star$:

$$\underline{\nu} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_a), .., \mathcal{B}(\mu_K))$$
$$\underline{\nu}' = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu'_a), .., \mathcal{B}(\mu_K))$$

Information at time t: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

$$\mathbb{E}_{\underline{\nu}}[N_a(T)]\,\mathrm{kl}(\mu_a, \mu'_a) = \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{Y_{1:T}}, \mathbb{P}_{\underline{\nu}'}^{Y_{1:T}}\big)$$

contraction of entropy
$$\geqslant \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{N_a(T)/T}, \mathbb{P}_{\underline{\nu}'}^{N_a(T)/T}\big)$$

projection
$$\geqslant \mathrm{kl}\Big(\mathbb{E}_{\underline{\nu}}[N_a(T)]/T, \mathbb{E}_{\underline{\nu}'}[N_a(T)]/T\Big)$$

$\mathrm{kl}(p,q) \geqslant p\log(1/q) - \log(2)$
$$\geqslant \Big(1 - \mathbb{E}_{\underline{\nu}}[N_a(T)]/T\Big)\log\frac{T}{T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]} - \log(2)$$

Sketch of proof 1/2

a suboptimal arm ($\mu^\star - \mu_a > 0$).
Modified bandit problem with $\mu'_a > \mu^\star$:

$$\underline{\nu} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_a), .., \mathcal{B}(\mu_K))$$
$$\underline{\nu}' = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu'_a), .., \mathcal{B}(\mu_K))$$

Information at time t: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

$$\mathbb{E}_{\underline{\nu}}[N_a(T)]\,\mathrm{kl}(\mu_a, \mu'_a) = \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{Y_{1:T}}, \mathbb{P}_{\underline{\nu}'}^{Y_{1:T}}\big)$$

contraction of entropy
$$\geqslant \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{N_a(T)/T}, \mathbb{P}_{\underline{\nu}'}^{N_a(T)/T}\big)$$

projection
$$\geqslant \mathrm{kl}\Big(\mathbb{E}_{\underline{\nu}}[N_a(T)]/T, \mathbb{E}_{\underline{\nu}'}[N_a(T)]/T\Big)$$

Consistent
$$\geqslant \Big(1 - \underbrace{\mathbb{E}_{\underline{\nu}}[N_a(T)]/T}_{o(1)}\Big) \log \frac{T}{\underbrace{T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]}_{O(T^\alpha)}} - \log(2)$$

Sketch of proof 1/2

a suboptimal arm ($\mu^\star - \mu_a > 0$).
Modified bandit problem with $\mu'_a > \mu^\star$:

$$\underline{\nu} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_a), .., \mathcal{B}(\mu_K))$$
$$\underline{\nu'} = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu'_a), .., \mathcal{B}(\mu_K))$$

Information at time t: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

$$\begin{aligned}
\mathbb{E}_{\underline{\nu}}[N_a(T)]\,\mathrm{kl}(\mu_a, \mu'_a) &= \mathrm{KL}(\mathbb{P}_{\underline{\nu}}^{Y_{1:T}}, \mathbb{P}_{\underline{\nu'}}^{Y_{1:T}}) \\
\text{contraction of entropy} \qquad &\geqslant \mathrm{KL}(\mathbb{P}_{\underline{\nu}}^{N_a(T)/T}, \mathbb{P}_{\underline{\nu'}}^{N_a(T)/T}) \\
\text{projection} \qquad &\geqslant \mathrm{kl}\Big(\mathbb{E}_{\underline{\nu}}[N_a(T)]/T, \mathbb{E}_{\underline{\nu'}}[N_a(T)]/T\Big) \\
&\gtrsim (1-\alpha)\log(T) - \log(2)
\end{aligned}$$

For all $\alpha \in (0, 1]$:

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log T} \geqslant \frac{1-\alpha}{\mathrm{kl}(\mu_a, \mu'_a)}\,.$$

For all $\alpha \in (0, 1]$:

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log T} \geqslant \frac{1-\alpha}{\mathrm{kl}(\mu_a, \mu'_a)} .$$

Very generic tools, can be used to prove :
-minimax/problem dependent lower bound on the Bayesian posterior concentration rates
-minimax lower bound for sparse adversarial bandits
-lower bounds in best arm identification

Sub-optimality of UCB

UCB

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leqslant \frac{1}{2(\mu_a - \mu^\star)^2},$$

Lower bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)}.$$

Pinsker inequality

$$\mathrm{kl}(\mu_a, \mu^\star) \geqslant 2(\mu_a - \mu^\star)^2$$

Chernoff Bound

$$X_1, \cdots, X_n \text{ i.i.d.} \sim \mathcal{B}(\mu) \text{ with } \widehat{\mu}_n = \sum_{k=1}^{n} X_k / n$$

Chernoff inequality for $x < \mu$

$$\mathbb{P}(\widehat{\mu}_n < x) \leqslant e^{-n\mathrm{kl}(x,\mu)} \underset{\text{Pinsker}}{\leqslant} e^{-2n(x-\mu)^2}$$

Chernoff Bound

$$X_1, \cdots, X_n \text{ i.i.d.} \sim \mathcal{B}(\mu) \text{ with } \widehat{\mu}_n = \sum_{k=1}^n X_k / n$$

Chernoff inequality for $x < \mu$

$$\mathbb{P}(\widehat{\mu}_n < x) \leqslant e^{-n\mathrm{kl}(x,\mu)} \underset{\text{Pinsker}}{\leqslant} e^{-2n(x-\mu)^2}$$

Inverting for $u = \mathrm{kl}(x, \mu)$

$$\mathbb{P}(\widehat{\mu}_n < \mu \text{ and } \mathrm{kl}(\widehat{\mu}_n, \mu) > u) \leqslant e^{-nu}$$

New upper confidence bound, with probability at least $1 - \delta$

$$\widehat{\mu}_n \geqslant \mu \text{ or } \mathrm{kl}(\widehat{\mu}_n, \mu) \leqslant \frac{\log(1/\delta)}{n}$$

$$\mu \leqslant \sup \left\{ \mu' \geqslant \widehat{\mu}_n \ : \ \mathrm{kl}(\widehat{\mu}_n, \mu') \leqslant \frac{\log(1/\delta)}{n} \right\}$$

Get the right constant: kl-UCB algorithm

kl-UCB algorithm:

Play each arm once, then for $K \leqslant t \leqslant T-1$ play according to the index:

$$U_a^{\mathrm{kl}}(t) = \sup\left\{\mu \geqslant \widehat{\mu}_a(t) \ : \ N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \log(T)\right\}.$$

Get the right constant: kl-UCB algorithm

kl-UCB algorithm:
Play each arm once, then for $K \leqslant t \leqslant T-1$ play according to the index:

$$U_a^{\mathrm{kl}}(t) = \sup\left\{ \mu \geqslant \widehat{\mu}_a(t) \ : \ N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \log(T) \right\}.$$

Regret upper bound:
For the kl-UCB algorithm, for all a such that $\mu^\star - \mu_a > 0$

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{1}{\mathrm{kl}(\mu_a,\, \mu^\star)} \log(T) + o\big(\log(T)\big)\,,$$

Lower bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}\big[N_a(T)\big]}{\log T} \geqslant \frac{1}{\mathrm{kl}(\mu_a,\, \mu^\star)}\,.$$

Minimax optimality

Theorem (Minimax lower bound)

$$\sup_{\underline{\nu}} R_{T,\underline{\nu}} \geqslant C\sqrt{KT}\,.$$

An algorithm is minimax optimal if

$$R_T \leqslant C'\sqrt{KT}\,.$$

UCB algorithm is not minimax optimal

$$R_T \leqslant C''\sqrt{KT\log(T)}\,.$$

MOSS-algorithm

MOSS index:

$$U_a^{\mathrm{M}}(t) = \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{\log_+\Big(T/(KN_a(t))\Big)}{2N_a(t)}}.$$

MOSS algorithm is minimax optimal, i.e.

$$R_T \leqslant 17\sqrt{KT} + K\,,$$

but is not asymptotically optimal (as UCB)...

A minimax and asymptotically optimal algorithm

kl-UCB^{++} index:

$$U_a^{\mathrm{kl}++}(t) = \sup\left\{\mu \geqslant \widehat{\mu}_a(t) \ : \ N_a(t)\mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \log_+\Big(T/(KN_a(t))\Big)\right\}.$$

A minimax and asymptotically optimal algorithm

kl-UCB^{++} index:

$$U_a^{\mathrm{kl}++}(t) = \sup\left\{ \mu \geqslant \widehat{\mu}_a(t) \ : \ N_a(t)\mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \log_+\Big(T/(KN_a(t)) \Big) \right\}.$$

kl-UCB^{++} algorithm is minimax and asymptotically optimal:

Minimax optimality:

$$R_T \leqslant 17\sqrt{KT} + K \, .$$

Asymptotic optimality: For any sub-optimal arm a,

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^\star)} + o\big(\log(T)\big) \, .$$

Second order optimality

With a stronger assumption:

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{C\log(T)}{\Delta_a^2}\,,$$

refinement of the Lai and Robbins lower bound

$$\mathbb{E}\big[N_a(T)\big] \geqslant \frac{\log T}{\mathrm{kl}(\mu_a, \mu^\star)} - \mathrm{O}\big(\mathrm{loglog}(T)\big)\,.$$

Second order optimality

With a stronger assumption:

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{C\log(T)}{\Delta_a^2}\,,$$

refinement of the Lai and Robbins lower bound

$$\mathbb{E}\big[N_a(T)\big] \geqslant \frac{\log T}{\mathrm{kl}(\mu_a,\mu^\star)} - \mathrm{O}\big(\mathrm{loglog}(T)\big)\,.$$

For the kl-UCB^{++} algorithm, for any sub-optimal arm a,

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T) - \mathrm{loglog}(T)}{\mathrm{kl}(\mu_a,\mu^\star)} + \mathrm{O}(1)\,.$$

An intuition for the exploration function

Bandit problems $\underline{\nu}$ and $\underline{\nu}'$ with $\underline{\nu} = \underline{\nu}'$ except $\mu'_a > \mu^\star$. Lower bound:

$$\mathbb{E}_{\underline{\nu}}[N_a(T)]\,\mathrm{kl}(\mu_a, \mu'_a) \geqslant \left(1 - \frac{\mathbb{E}_{\underline{\nu}}[N_a(T)]}{T}\right) \log \frac{T}{\sum_{b \neq a} \mathbb{E}_{\underline{\nu}'}[N_b(T)]} - \ln 2\,.$$

An intuition for the exploration function

Bandit problems $\underline{\nu}$ and $\underline{\nu}'$ with $\underline{\nu} = \underline{\nu}'$ except $\mu'_a > \mu^\star$. Lower bound:

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{kl}(\mu_a, \mu'_a) \geqslant \left(1 - \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T}\right) \log \frac{T}{\sum_{b\neq a} \mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big]} - \ln 2\,.$$

Approximations:

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \approx N_a(T)\ \left(\approx \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^\star)}\right), \quad \mu_a \approx \widehat{\mu}_a(T), \quad \sum_{b\neq a} \mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big] \approx K N_a(T)\,.$$

An intuition for the exploration function

Bandit problems $\underline{\nu}$ and $\underline{\nu}'$ with $\underline{\nu} = \underline{\nu}'$ except $\mu'_a > \mu^\star$. Lower bound:

$$N_a(T)\,\mathrm{kl}(\widehat{\mu}_a(T), \mu'_a) \gtrsim \log \frac{T}{\sum_{b \neq a} \mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big]}\,.$$

Approximations:

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \approx N_a(T) \left(\approx \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^\star)}\right), \quad \mu_a \approx \widehat{\mu}_a(T), \quad \sum_{b \neq a} \mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big] \approx K N_a(T)\,.$$

An intuition for the exploration function

Bandit problems $\underline{\nu}$ and $\underline{\nu}'$ with $\underline{\nu} = \underline{\nu}'$ except $\mu'_a > \mu^\star$. Lower bound:

$$N_a(T)\,\mathrm{kl}(\widehat{\mu}_a(T), \mu'_a) \gtrsim \log \frac{T}{\sum_{b\neq a} \mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big]}\,.$$

Approximations:

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \approx N_a(T)\ \left(\approx \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^\star)}\right), \quad \mu_a \approx \widehat{\mu}_a(T), \quad \sum_{b\neq a} \mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big] \approx K N_a(T)\,.$$

Index:

$$U_a^{\mathrm{kl}++}(T) = \sup\left\{\mu \geqslant \widehat{\mu}_a(T)\ :\ N_a(T)\,\mathrm{kl}\big(\widehat{\mu}_a(T), \mu\big) \leqslant \log \frac{T}{K N_a(T)}\right\}$$

$$= \inf\left\{\mu \geqslant \widehat{\mu}_a(T)\ :\ N_a(T)\,\mathrm{kl}\big(\widehat{\mu}_a(T), \mu\big) \geqslant \log \frac{T}{K N_a(T)}\right\},$$

K-armed bandit problem: non-parametric setting

Bounded rewards: $\nu_a \in \mathcal{P}[0,1]$

$$\underline{\nu} = (\nu_1, \quad \cdots \quad \nu_a \quad \cdots \quad , \nu_K)$$

Game: for each round $1 \leqslant t \leqslant T$:

1. Player pulls arm $A_t \in \{1, .., K\}$.
2. He gets a reward $Y_t \sim \nu_{A_t}$.

$$\mu_a = E(\nu_a)$$

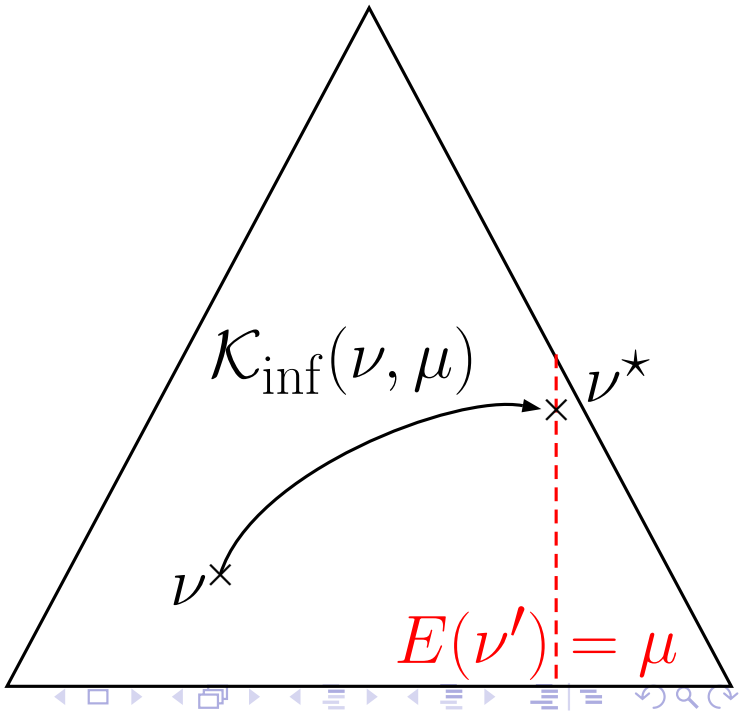
Lower bound

Theorem (Asymptotic lower)

For all consistent strategies, for all arms a such that $\mu^\star - E(\nu_a) > 0$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}.$$

$$\mathcal{K}_{\inf}(\nu, \mu) := \inf\left\{\mathrm{KL}(\nu, \nu') : \ E(\nu') > \mu\right\}$$



Lower bound

Theorem (Asymptotic lower)

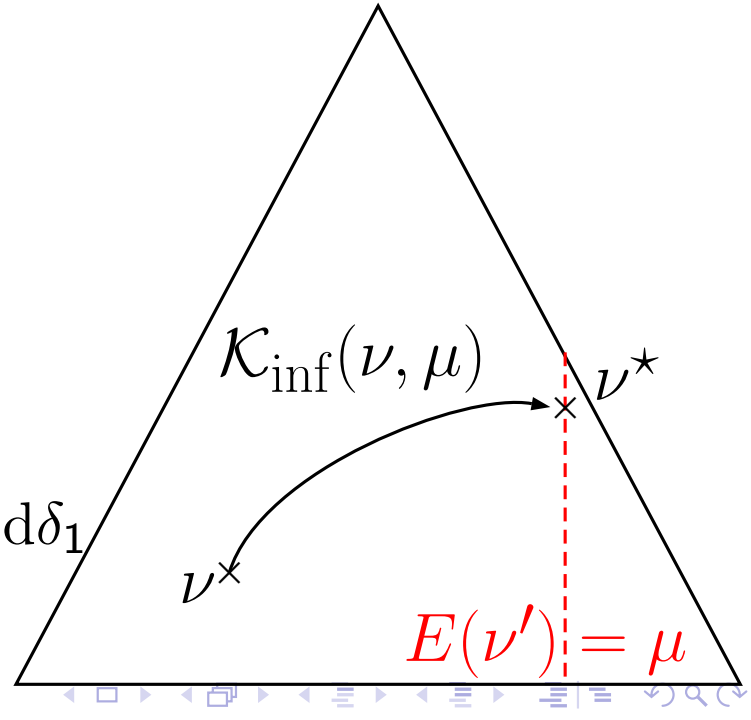
For all consistent strategies, for all arms a such that $\mu^\star - E(\nu_a) > 0$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}.$$

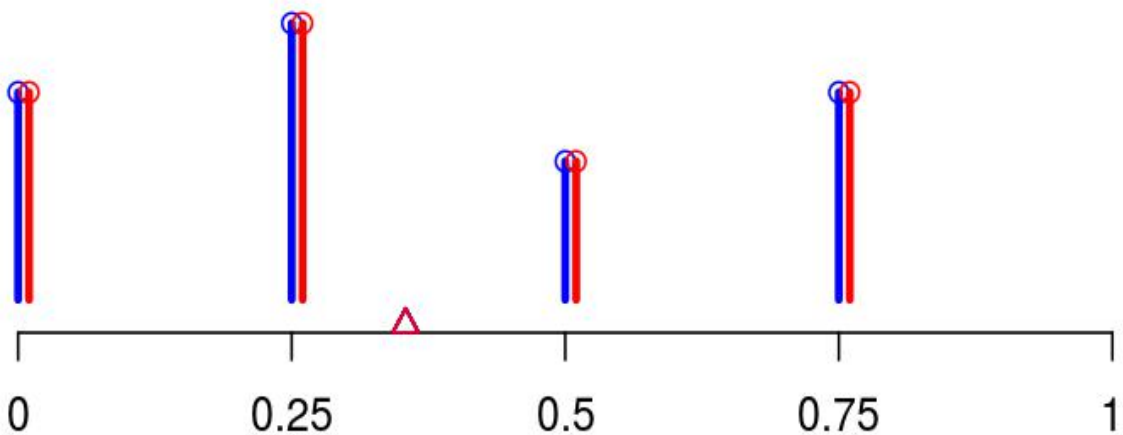
$$\begin{aligned}\mathcal{K}_{\inf}(\nu,\mu) &:= \inf\left\{\mathrm{KL}(\nu,\nu') : \ E(\nu') > \mu\right\} \\ &= \mathrm{KL}(\nu,\nu^\star)\end{aligned}$$

$\nu^\star$ of the form, for a certain $\lambda^\star \in [0,1]$

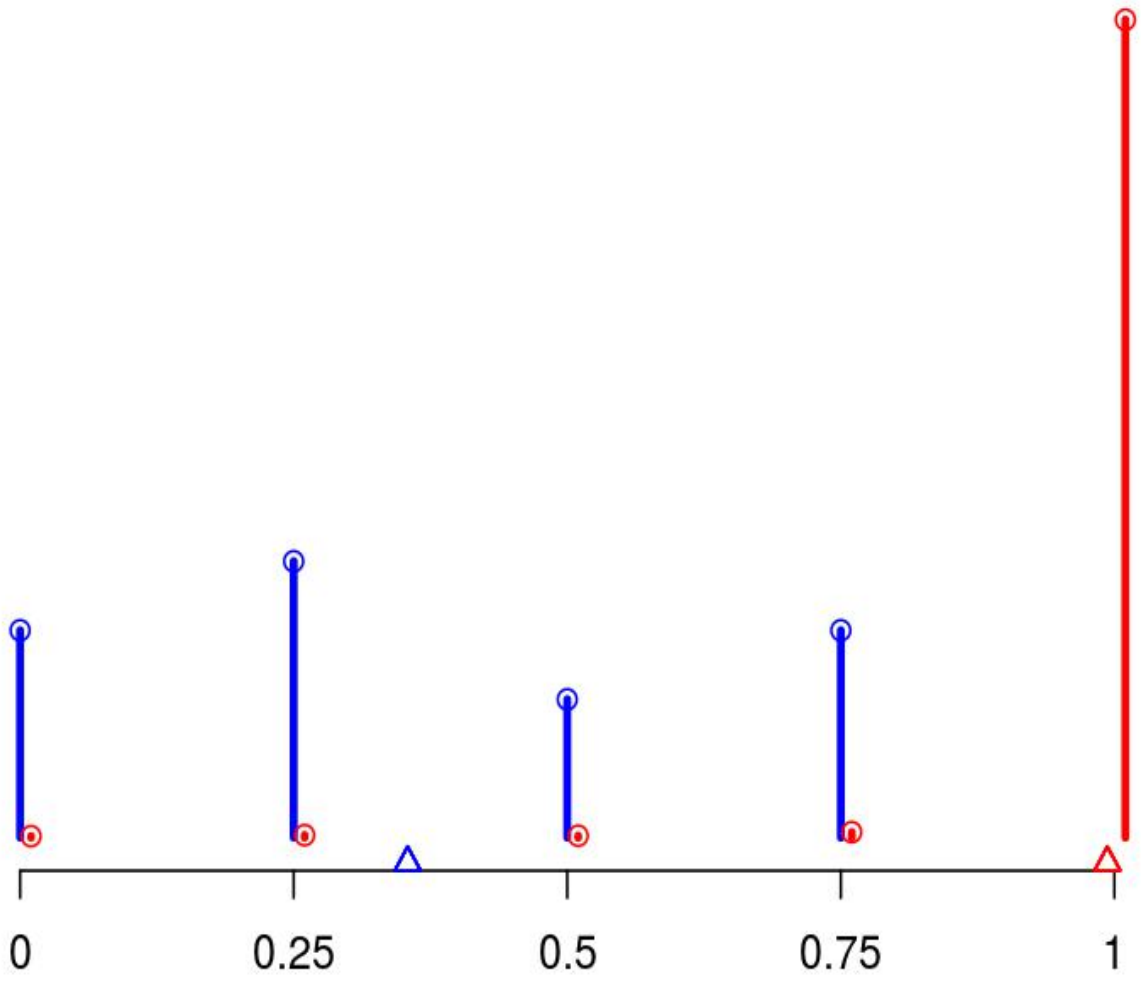
$$\mathrm{d}\nu^\star = \frac{1}{1-\lambda^\star\frac{x-\mu}{1-\mu}}\,\mathrm{d}\nu + \left(1-\mathbb{E}_\nu\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right]\right)\mathrm{d}\delta_1$$



$$E(\nu) \leqslant E(\nu^\star) = \mu$$



$$E(\nu) \leqslant E(\nu^\star) = \mu$$



Sub-optimality of kl-UCB

kl-UCB

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leqslant \frac{1}{\mathrm{kl}(\mu_a,\, \mu^\star)}\,,$$

Lower bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

Pseudo-Pinsker inequality

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star) \geqslant \mathrm{kl}(E(\nu_a),\, \mu^\star)$$

Sub-optimality of kl-UCB

kl-UCB

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leqslant \frac{1}{\mathrm{kl}(\mu_a,\, \mu^\star)}\,,$$

Lower bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

Pseudo-Pinsker inequality

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star) \geqslant \mathrm{kl}(E(\nu_a),\, \mu^\star)$$

Reduction to kl for Bernoulli:

$$\mathcal{K}_{\inf}(\mathcal{B}(\mu_a), \mu^\star) = \mathrm{kl}(\mu_a, \mu^\star)$$

Sub-optimality of kl-UCB

kl-UCB

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leqslant \frac{1}{\mathrm{kl}(\mu_a,\, \mu^\star)}\,,$$

Lower bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

Pseudo-Pinsker inequality

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star) \geqslant \mathrm{kl}(E(\nu_a),\, \mu^\star)$$

$$\underbrace{\inf\Big\{\mathrm{KL}(\nu,\nu') :\ E(\nu') > \mu\Big\}}_{=\mathcal{K}_{\inf}(\nu,\mu)} \geqslant \underbrace{\inf\Big\{\mathrm{KL}(\nu'',\nu') :\ E(\nu') > \mu,\ E(\nu'') = E(\nu)\Big\}}_{=\mathrm{kl}\left(E(\nu),\mu\right)}$$

Index ?

Move from empirical mean $\widehat{\mu}_n$ to empirical distribution $\widehat{\nu}_n = 1/n \sum_{k=1}^{n} \delta_{X_k}$

New index

$$U_a^{\mathrm{kl}}(t) = \sup \left\{ \mu' \geqslant \widehat{\mu}_a(t) : \ \mu' \in [0,1], \, \mathrm{kl}(\widehat{\mu}_a(t), \, \mu') \leqslant \frac{\log(T)}{N_a(t)} \right\}$$

$$U_a^{\mathrm{KL}}(t) = \sup \left\{ E\nu' \geqslant E\left(\widehat{\nu}_a(t)\right) : \ \nu' \in \mathcal{P}[0,1], \ \ \mathrm{KL}(\widehat{\nu}_a(t), \, \nu') \leqslant \frac{\log(T)}{N_a(t)} \right\}$$

Index ?

Move from empirical mean $\widehat{\mu}_n$ to empirical distribution $\widehat{\nu}_n = 1/n \sum_{k=1}^{n} \delta_{X_k}$

New index

$$U_a^{\mathrm{kl}}(t) = \sup \left\{ \mu' \geqslant \widehat{\mu}_a(t) : \ \mu' \in [0,1], \, \mathrm{kl}(\widehat{\mu}_a(t), \, \mu') \leqslant \frac{\log(T)}{N_a(t)} \right\}$$

$$U_a^{\mathrm{KL}}(t) = \sup \left\{ E\nu' \geqslant E\big(\widehat{\nu}_a(t)\big) : \ \nu' \in \mathcal{P}[0,1], \ \ \mathrm{KL}(\widehat{\nu}_a(t), \, \nu') \leqslant \frac{\log(T)}{N_a(t)} \right\}$$

$$= \sup \left\{ \mu' : \ \mu' \in [0,1], \, \mu' \geqslant \widehat{\mu}_a(t), \ \ \mathcal{K}_{\mathsf{inf}}(\widehat{\nu}_a(t), \, \mu') \leqslant \frac{\log(T)}{N_a(t)} \right\}.$$

An asymptotically optimal algorithm

KL-UCB index:

$$U_a^{\mathrm{KL}}(t) = \sup\left\{\mu' \geqslant \widehat{\mu}_a(t) : \ N_a(t)\, \mathcal{K}_{\inf}\left(\widehat{\nu}_a(t),\, \mu'\right) \leqslant \log(T)\right\}.$$

An asymptotically optimal algorithm

KL-UCB index:

$$U_a^{\mathrm{KL}}(t) = \sup\left\{\mu' \geqslant \widehat{\mu}_a(t) : \ N_a(t)\, \mathcal{K}_{\inf}(\widehat{\nu}_a(t),\, \mu') \leqslant \log(T)\right\}.$$

Upper bound on the regret: for all a such that $\mu^\star - E(\mu_a) > 0$

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a,\, \mu^*)} \log(T) + o(\log(T))\,,$$

Lower bound

$$\liminf_{T\to\infty} \ \frac{\mathbb{E}\big[N_a(T)\big]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a,\, \mu^\star)}\,.$$

Non-parametric upper confidence bound

$$X_1, \cdots, X_n \text{ i.i.d.} \sim \nu \text{ with } \widehat{\nu}_n = \sum_{k=1}^{n} \delta_{X_k}/n\,.$$

Deviations of kl

$$\mathbb{P}\Big(\widehat{\mu}_n < E(\nu) \text{ and } \mathrm{kl}(\widehat{\mu}_n, E(\nu)) > u\Big) \leqslant e^{-nu}$$

Deviations of $\mathcal{K}_{\mathrm{inf}}$

$$\mathbb{P}\Big(\mathcal{K}_{\mathrm{inf}}(\widehat{\nu}_n, E(\nu)) > u\Big) \leqslant e(n+3)e^{-nu}\,.$$

Non-parametric upper confidence bound

$$X_1, \cdots, X_n \text{ i.i.d.} \sim \nu \text{ with } \widehat{\nu}_n = \sum_{k=1}^{n} \delta_{X_k}/n\,.$$

Deviations of kl

$$\mathbb{P}\Big(\widehat{\mu}_n < E(\nu) \text{ and } \mathrm{kl}(\widehat{\mu}_n, E(\nu)) > u\Big) \leqslant e^{-nu}$$

Deviations of $\mathcal{K}_{\inf}$

$$\mathbb{P}\Big(\mathcal{K}_{\inf}(\widehat{\nu}_n, E(\nu)) > u\Big) \leqslant e\cancel{(n+3)}e^{-nu}\,.$$

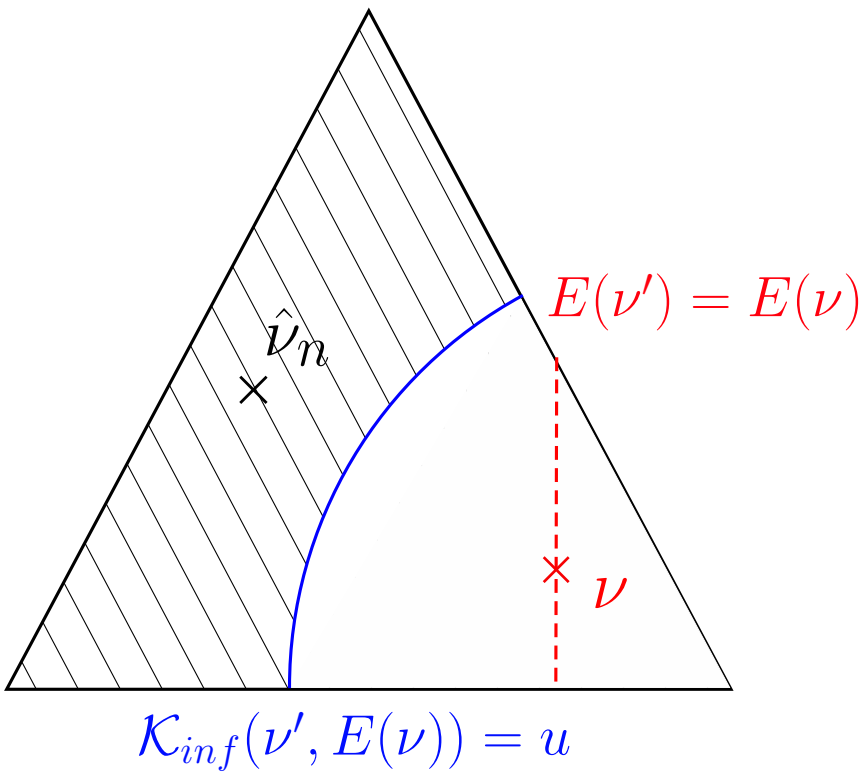
Open question: remove the factor (n+3) ?

Variational formula

$$\mathcal{K}_{\inf}(\nu,\mu) = \max_{0\leqslant\lambda\leqslant 1} \mathbb{E}_\nu\Bigg[\underbrace{\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)}_{:=f_\lambda(X)}\Bigg]$$

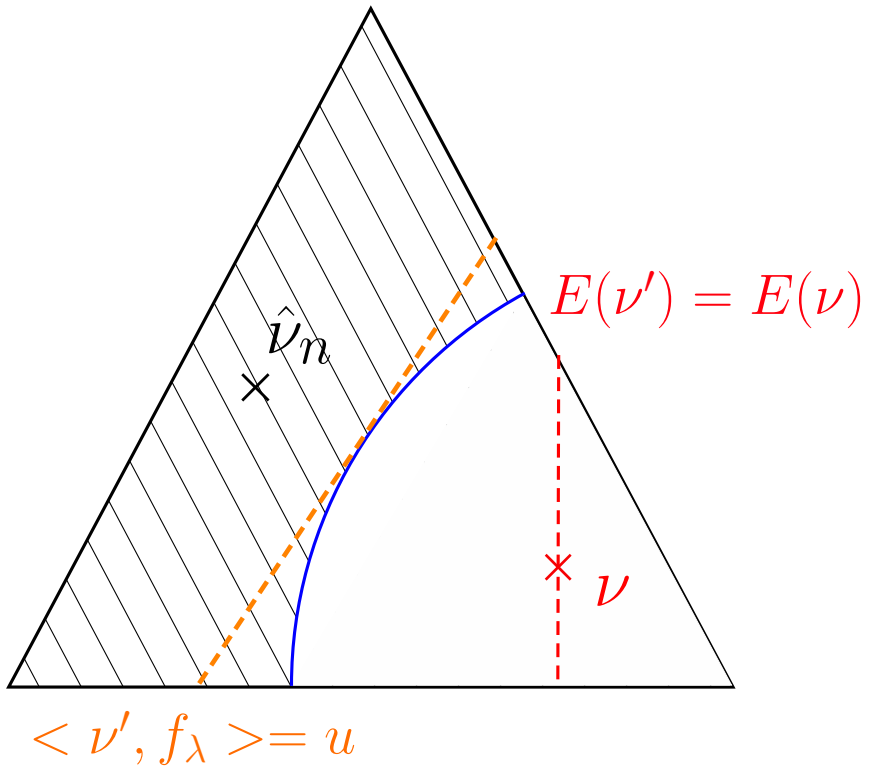
Variational formula

$$\mathcal{K}_{\inf}(\widehat{\nu}_n, E(\nu)) = \max_{0 \leqslant \lambda \leqslant 1} \mathbb{E}_{\widehat{\nu}_n}\Big[\underbrace{\log\Big(1 - \lambda \frac{X - E(\nu)}{1 - E(\nu)}\Big)}_{:= f_\lambda(X)}\Big]$$



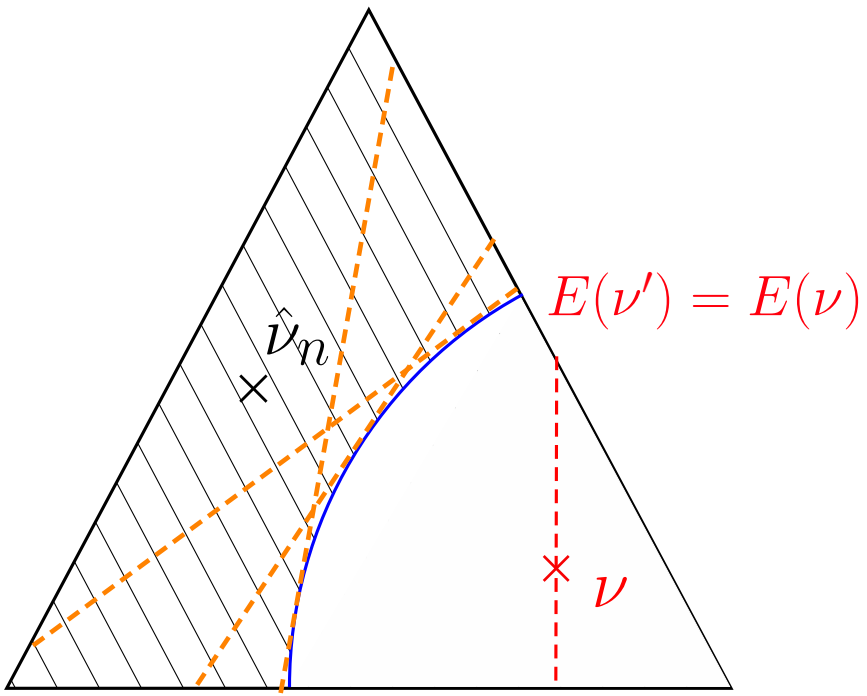
Variational formula

$$\mathcal{K}_{\inf}(\widehat{\nu}_n, E(\nu)) = \max_{0 \leqslant \lambda \leqslant 1} < \widehat{\nu}_n, f_\lambda >$$



Variational formula

$$\mathcal{K}_{\inf}(\widehat{\nu}_n, E(\nu)) = \max_{0 \leqslant \lambda \leqslant 1} < \widehat{\nu}_n, f_\lambda >$$



Variational formula: Worst family

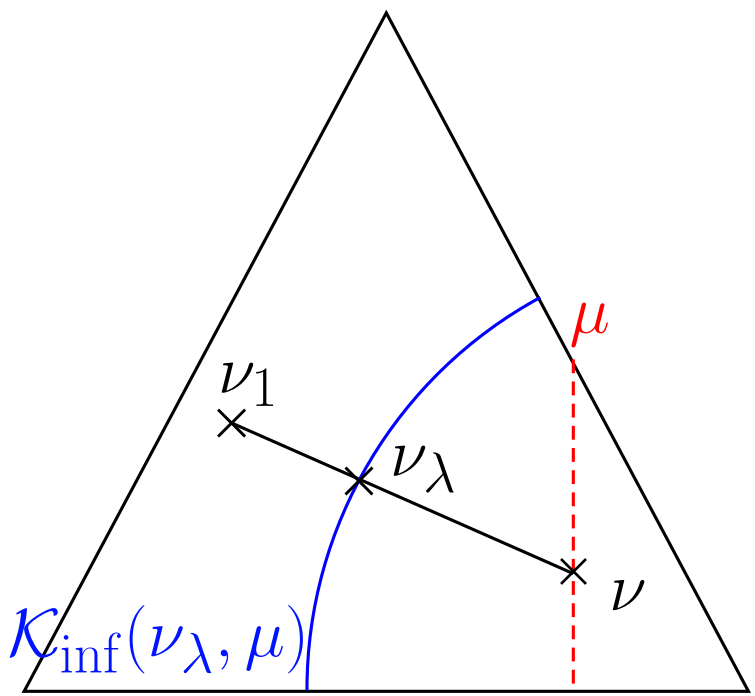
$$\mathcal{K}_{\inf}(\nu,\mu) = \max_{0\leqslant\lambda\leqslant 1} \mathbb{E}_\nu\left[\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\right].$$

If $E(\nu)=\mu$. Convex family of probability distributions: $\frac{\mathrm{d}\nu_\lambda}{\mathrm{d}\nu} = \left(1-\lambda\frac{x-\mu}{1-\mu}\right)$

$$\nu_\lambda = \lambda\nu_1 + (1-\lambda)\nu$$

Worst family for ν:

$$\mathcal{K}_{\inf}(\nu_\lambda,\mu) = \mathrm{KL}(\nu_\lambda,\nu)$$



Variational formula: Worst family

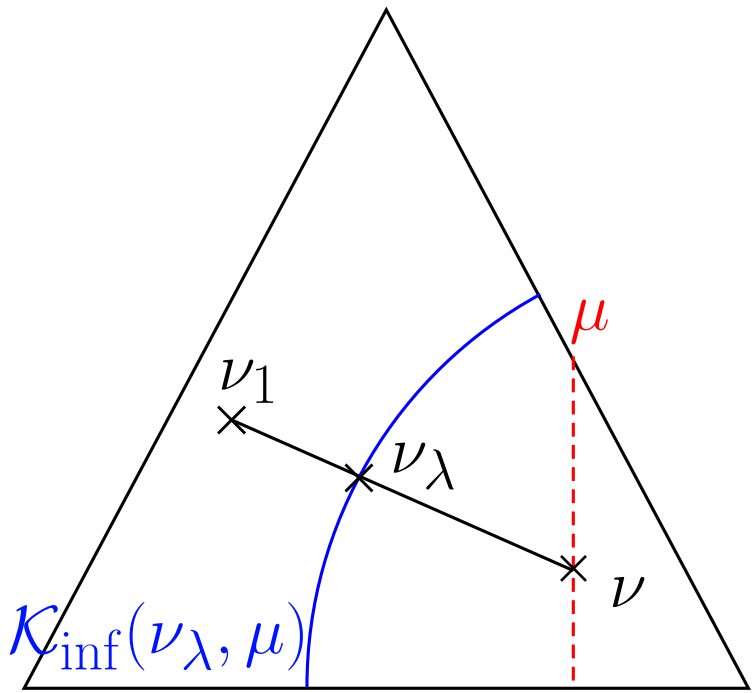
$$\mathcal{K}_{\inf}(\nu, \mu) = - \min_{0 \leqslant \lambda \leqslant 1} \mathrm{KL}(\nu, \nu_\lambda) = 0.$$

If $E(\nu) = \mu$. Convex family of probability distributions: $\frac{\mathrm{d}\nu_\lambda}{\mathrm{d}\nu} = \left(1 - \lambda \frac{x-\mu}{1-\mu}\right)$

$$\nu_\lambda = \lambda \nu_1 + (1-\lambda)\nu$$

Worst family for ν:

$$\mathcal{K}_{\inf}(\nu_\lambda, \mu) = \mathrm{KL}(\nu_\lambda, \nu)$$



Thompson Sampling for arms in $\mathcal{P}[0,1]$?

$\pi_a^0 = \mathcal{D}(\alpha, H)$ Dirichlet process with $H \in \mathcal{P}[0,1]$.

Algorithm 2: Thompson sampling

Parameter: prior $\Pi^0 = (\pi_1^0, \ldots, \pi_K^0)$

For $t = 0$ to $T-1$, **do**

1. **For** $a = 1$ to K, **do**
   Sample $\nu_a(t) \sim \pi_a^t$.
2. Play $A_t \in \operatorname{argmax}_{a \in \{1,\cdots,K\}} E(\nu_a(t))$, update the posterior Π^{t+1}.

Is Thompson Sampling asymptotically optimal ?

A minimax and asymptotically optimal algorithm

KL-UCB^{++} index:

$$U_a^{\mathrm{KL}++}(t) := \sup\left\{ \mu \in [0,1] \ : \ N_a(t)\, \mathcal{K}_{\inf}\left(\widehat{\nu}_a(t), \mu\right) \leqslant \log_+\left(\frac{T}{K N_a(t)}\right) \right\}.$$

A minimax and asymptotically optimal algorithm

KL-UCB^{++} index:

$$U_a^{\mathrm{KL}++}(t) := \sup\left\{ \mu \in [0,1] \ : \ N_a(t)\, \mathcal{K}_{\inf}\left(\widehat{\nu}_a(t), \mu\right) \leqslant \log_+\left(\frac{T}{K N_a(t)}\right) \right\}.$$

KL-UCB-switch index:

$$U_a^{\text{KL-S}}(t) := \begin{cases} U_a^{\mathrm{KL}++}(t) & \text{si } N_a(t) \leqslant \lfloor (T/K)^{1/5} \rfloor \\ U_a^{\mathrm{M}}(t) & \text{si } N_a(t) > \lfloor (T/K)^{1/5} \rfloor \end{cases},$$

A minimax and asymptotically optimal algorithm

KL-UCB^{++} index:

$$U_a^{\mathrm{KL}++}(t) := \sup\left\{ \mu \in [0,1] \ : \ N_a(t)\, \mathcal{K}_{\inf}\left(\widehat{\nu}_a(t), \mu\right) \leqslant \log_+\left(\frac{T}{K N_a(t)}\right) \right\}.$$

KL-UCB-switch index:

$$U_a^{\text{KL-S}}(t) := \begin{cases} U_a^{\mathrm{KL}++}(t) & \text{si } N_a(t) \leqslant \lfloor (T/K)^{1/5} \rfloor \\ U_a^{\mathrm{M}}(t) & \text{si } N_a(t) > \lfloor (T/K)^{1/5} \rfloor \end{cases},$$

KL-UCB-switch algorithm is minimax and asymptotically optimal:

Minimax optimality:

$$R_T \leqslant 25\sqrt{KT} + K\,.$$

Asymptotic optimality: For any sub-optimal arm a

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + o\big(\log(T)\big)$$

.

A minimax and asymptotically optimal algorithm

KL-UCB^{++} index:

$$U_a^{\mathrm{KL}++}(t) := \sup\left\{ \mu \in [0,1] \ : \ N_a(t)\, \mathcal{K}_{\inf}\left(\widehat{\nu}_a(t), \mu\right) \leqslant \log_+\left(\frac{T}{K N_a(t)}\right) \right\}.$$

KL-UCB-switch index:

$$U_a^{\text{KL-S}}(t) := \begin{cases} U_a^{\mathrm{KL}++}(t) & \text{si } N_a(t) \leqslant \lfloor (T/K)^{1/5} \rfloor \\ U_a^{\mathrm{M}}(t) & \text{si } N_a(t) > \lfloor (T/K)^{1/5} \rfloor \end{cases},$$

Open question, does the same theorem hold for KL-UCB^{++} ?

Minimax optimality:

$$R_T \leqslant 25\sqrt{KT} + K\,.$$

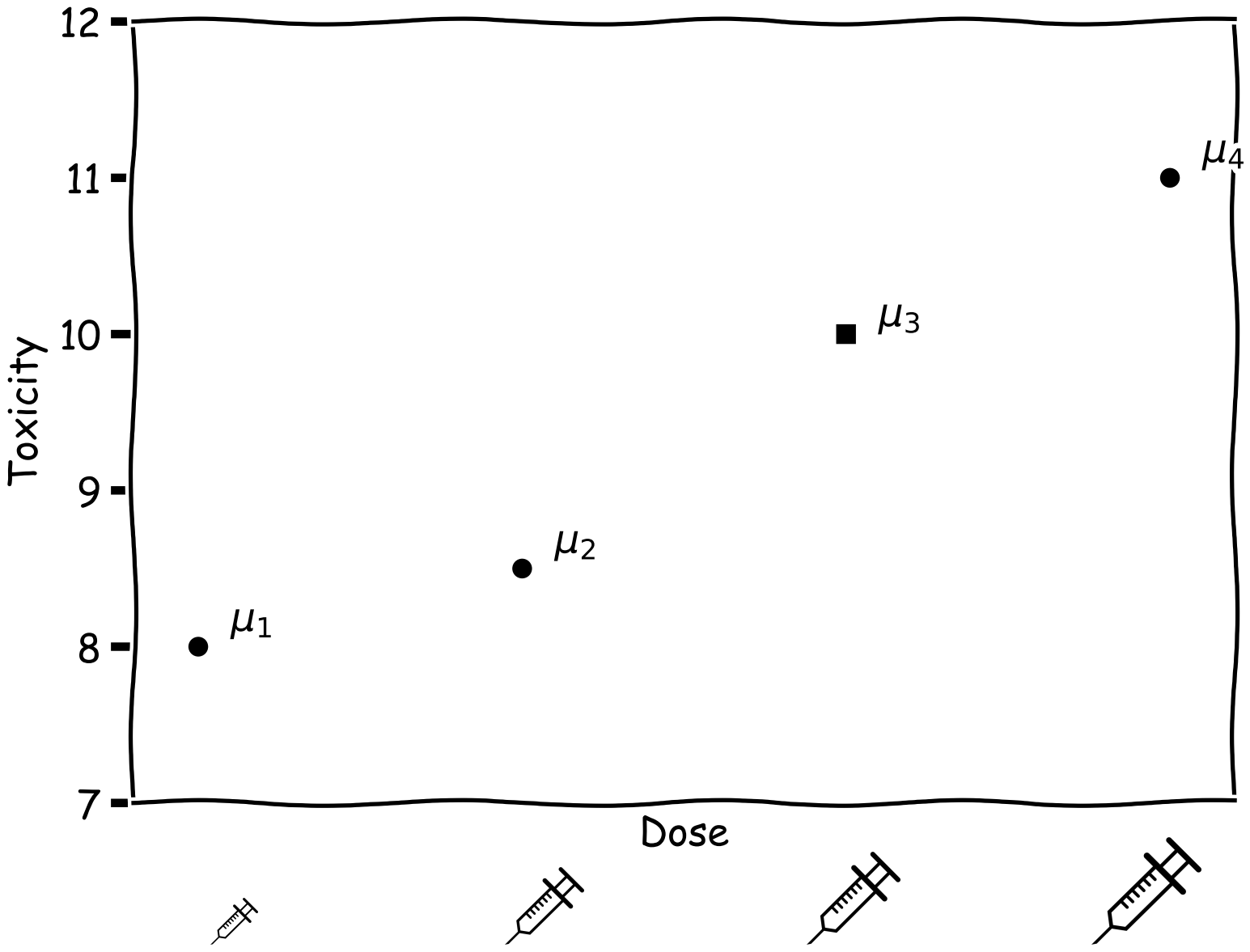
Asymptotic optimality: For any sub-optimal arm a

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + o\big(\log(T)\big)$$

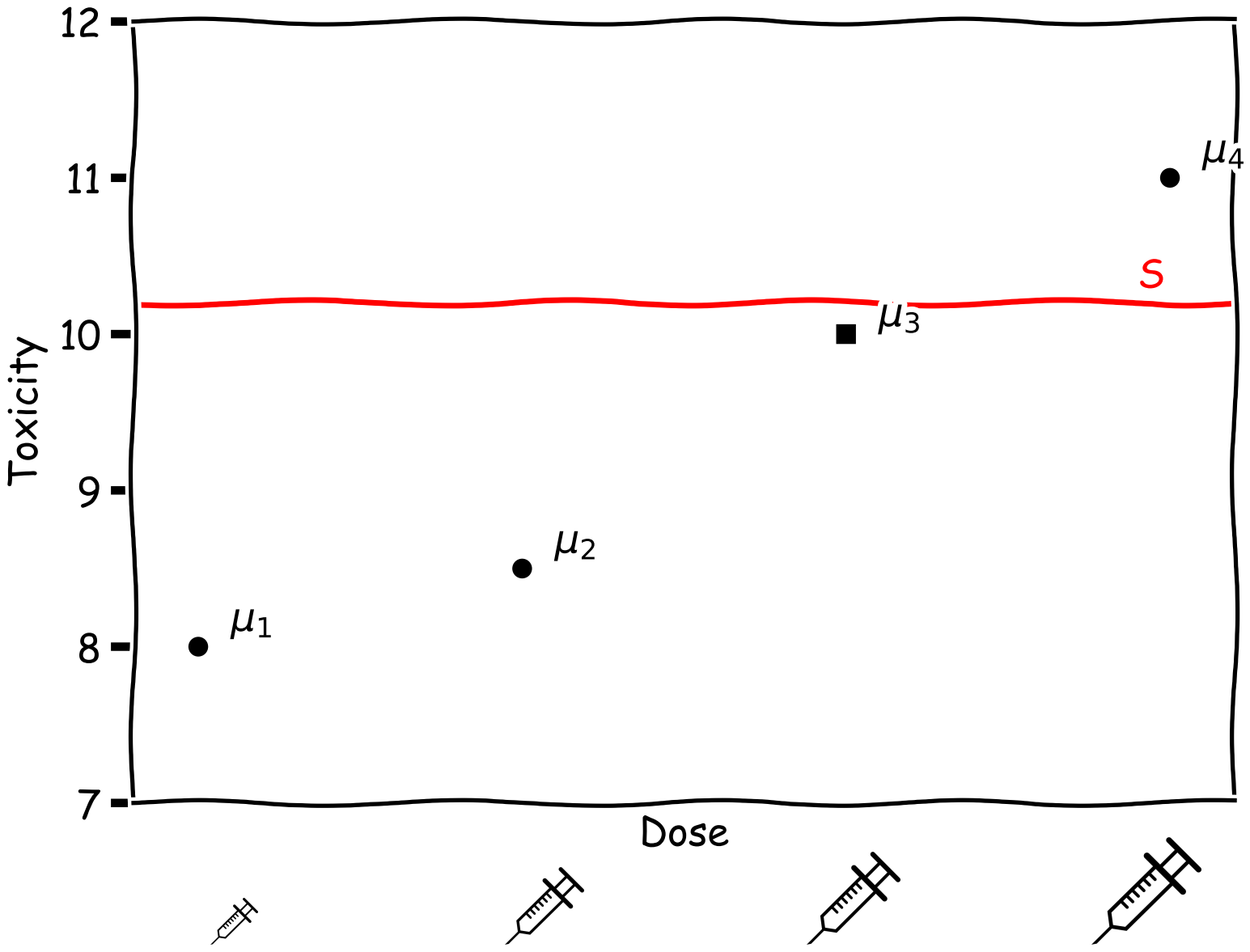
.

Merci pour votre attention !

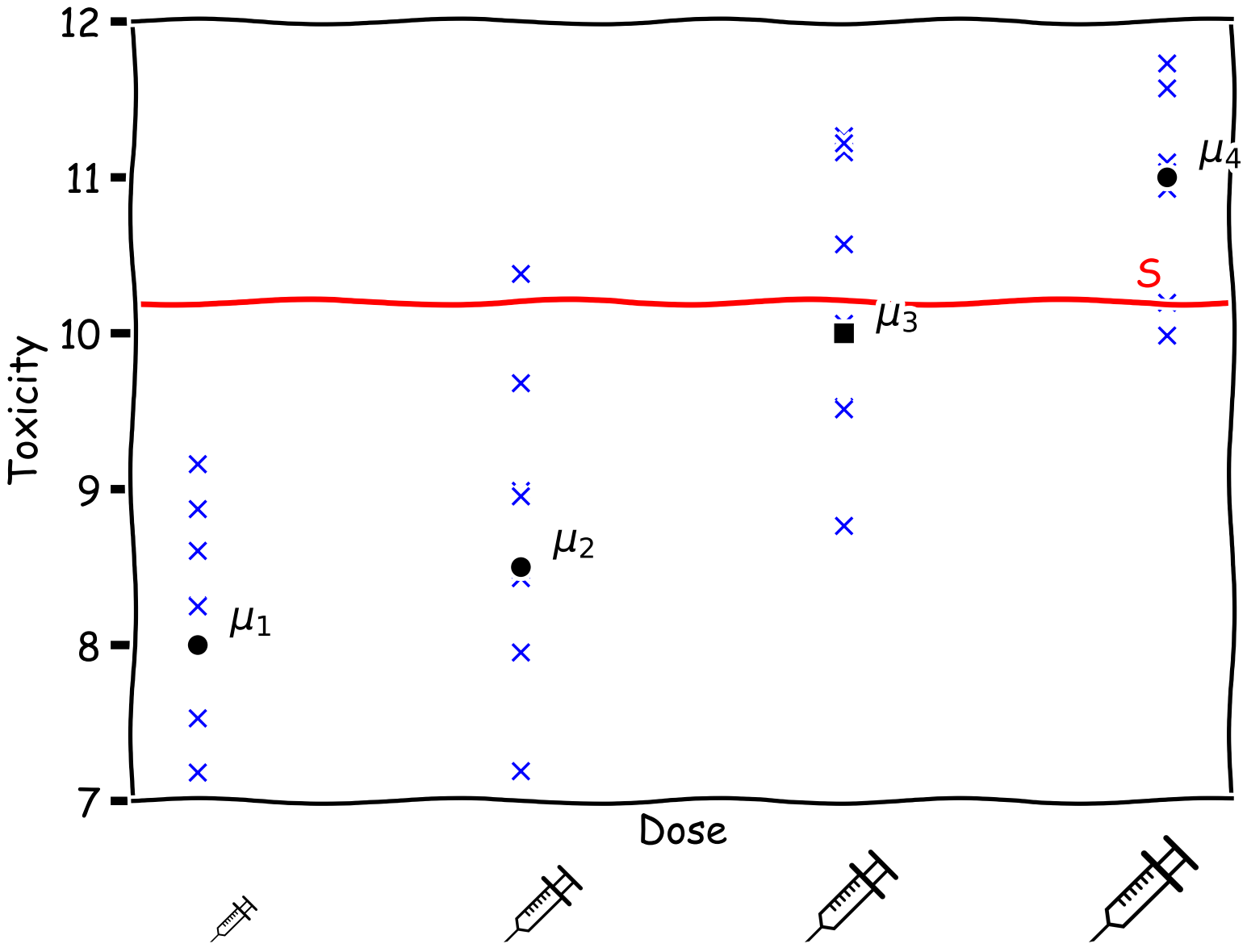
Dose-ranging



Dose-ranging



Dose-ranging



Best arm identification

$$\boldsymbol{\mu} = \quad (\mathcal{N}(\mu_1, 1) \quad \cdots \quad \mathcal{N}(\mu_a, 1) \quad \cdots \quad \mathcal{N}(\mu_K, 1))$$


$\cdots$

$\cdots$


Best arm identification

$$\begin{array}{rccccc} \boldsymbol{\mu} = & (\mathcal{N}(\mu_1, 1) & \cdots & \mathcal{N}(\mu_a, 1) & \cdots & \mathcal{N}(\mu_K, 1)) \\ \sim & [\mu_1 & \cdots & \mu_a & \cdots & \mu_K] \end{array}$$

Optimal arm (dose): $a^*_{\boldsymbol{\mu}} \in \arg\min_{1 \leqslant a \leqslant K} |\mu_a - S|$

Best arm identification

$$\begin{array}{rccccc} \boldsymbol{\mu} = & (\mathcal{N}(\mu_1, 1) & \cdots & \mathcal{N}(\mu_a, 1) & \cdots & \mathcal{N}(\mu_K, 1)) \\ \sim & [\mu_1 & \cdots & \mu_a & \cdots & \mu_K] \end{array}$$

Optimal arm (dose): $a^*_{\boldsymbol{\mu}} \in \arg\min_{1 \leqslant a \leqslant K} |\mu_a - S|$

Game: while $t < \tau$:

1. Player pulls arm (dose) $A_t \in \{1, .., K\}$.
2. He gets an observation (toxicity) $Y_t \sim \mathcal{N}(\mu_{A_t}, 1)$.

Predict best arm $\widehat{a}_\tau$.

Best arm identification

$$\begin{array}{rccccc} \boldsymbol{\mu} = & (\mathcal{N}(\mu_1, 1) & \cdots & \mathcal{N}(\mu_a, 1) & \cdots & \mathcal{N}(\mu_K, 1)) \\ \sim & [\mu_1 & \cdots & \mu_a & \cdots & \mu_K] \end{array}$$

Optimal arm (dose): $a^*_{\boldsymbol{\mu}} \in \arg\min_{1 \leqslant a \leqslant K} |\mu_a - S|$

Game: while $t < \tau$:

1. Player pulls arm (dose) $A_t \in \{1, .., K\}$.
2. He gets an observation (toxicity) $Y_t \sim \mathcal{N}(\mu_{A_t}, 1)$.

Predict best arm $\widehat{a}_\tau$.

δ-**correct algorithm.** ($\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$ information available at step t)

- a **sampling rule** $(A_t)_{t \geqslant 1}$, where A_t is $\mathcal{F}_{t-1}$-measurable;
- a **stopping rule** τ, stopping time with respect to the filtration $(\mathcal{F}_t)_{t \geqslant 1}$;
- a $\mathcal{F}_\tau$-measurable **decision rule** $\widehat{a}_\tau$;

An algorithm is δ-correct if $\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_{\tau_\delta} \neq a^*_{\boldsymbol{\mu}}) \leqslant \delta$ and $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < +\infty) = 1$.

Goal: find a δ-correct algorithm that minimize $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$.

$\rightarrow$ lower bound on $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$?

Lower bound

$\mathcal{M} = \{\boldsymbol{\mu} \in \mathbb{R}^K : a^*_{\boldsymbol{\mu}} \text{ is unique}\}$ $\quad \mathcal{I} = \{\boldsymbol{\mu} \in \mathcal{M} : \mu_1 < \ldots < \mu_K\}$

Alternative set for $\mathcal{S} \in \{\mathcal{M}, \mathcal{I}\}$: $Alt(\boldsymbol{\mu}, \mathcal{S}) := \{\boldsymbol{\lambda} \in \mathcal{S} : a^*_{\boldsymbol{\lambda}} \neq a^*_{\boldsymbol{\mu}}\}$.

Lower bound

$$\mathcal{M} = \{\boldsymbol{\mu} \in \mathbb{R}^K : a^*_{\boldsymbol{\mu}} \text{ is unique}\} \qquad \mathcal{I} = \{\boldsymbol{\mu} \in \mathcal{M} : \mu_1 < \ldots < \mu_K\}$$

Alternative set for $\mathcal{S} \in \{\mathcal{M}, \mathcal{I}\}$: $\mathcal{A}lt(\boldsymbol{\mu}, \mathcal{S}) := \{\boldsymbol{\lambda} \in \mathcal{S} : a^*_{\boldsymbol{\lambda}} \neq a^*_{\boldsymbol{\mu}}\}$.

Theorem

Let $\mathcal{S} \in \{\mathcal{M}, \mathcal{I}\}$. For all δ-correct algorithm,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geqslant T^*_{\mathcal{S}}(\boldsymbol{\mu}),$$

where the characteristic time is

$$T^*_{\mathcal{S}}(\boldsymbol{\mu})^{-1} = \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt(\boldsymbol{\mu}, \mathcal{S})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2},$$

where Σ_K is the simplex of dimension $K - 1$.

Key quantities

Characteristic time:

$$T^*_{\mathcal{S}}(\boldsymbol{\mu})^{-1} = \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt(\boldsymbol{\mu}, \mathcal{S})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \,.$$

Optimal weights:

$$\omega^*(\boldsymbol{\mu}) := \mathrm{argmax}_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in (\mu, \mathcal{S})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \,.$$

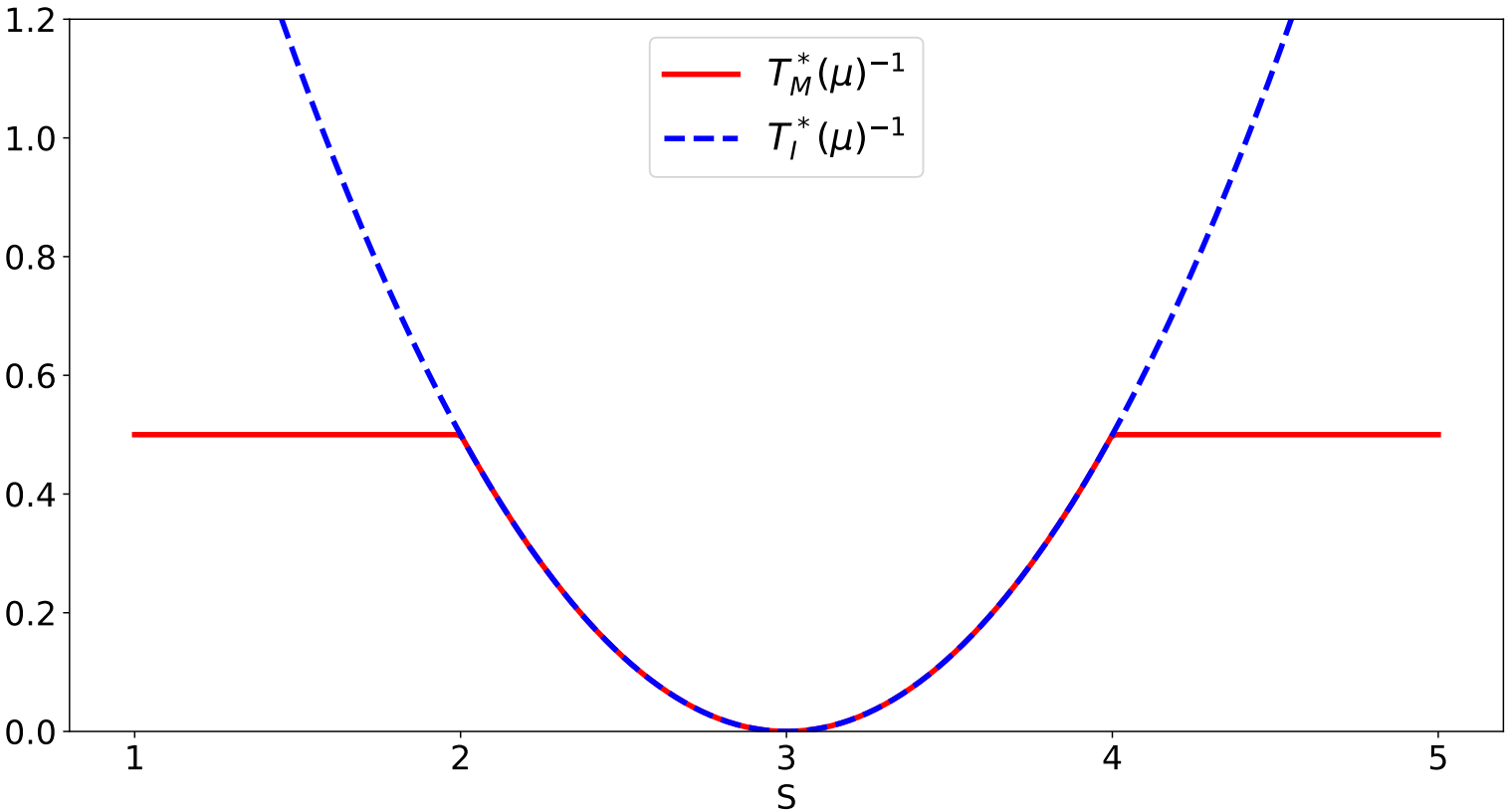
where Σ_K simplex of dimension $K - 1$.

Two arms

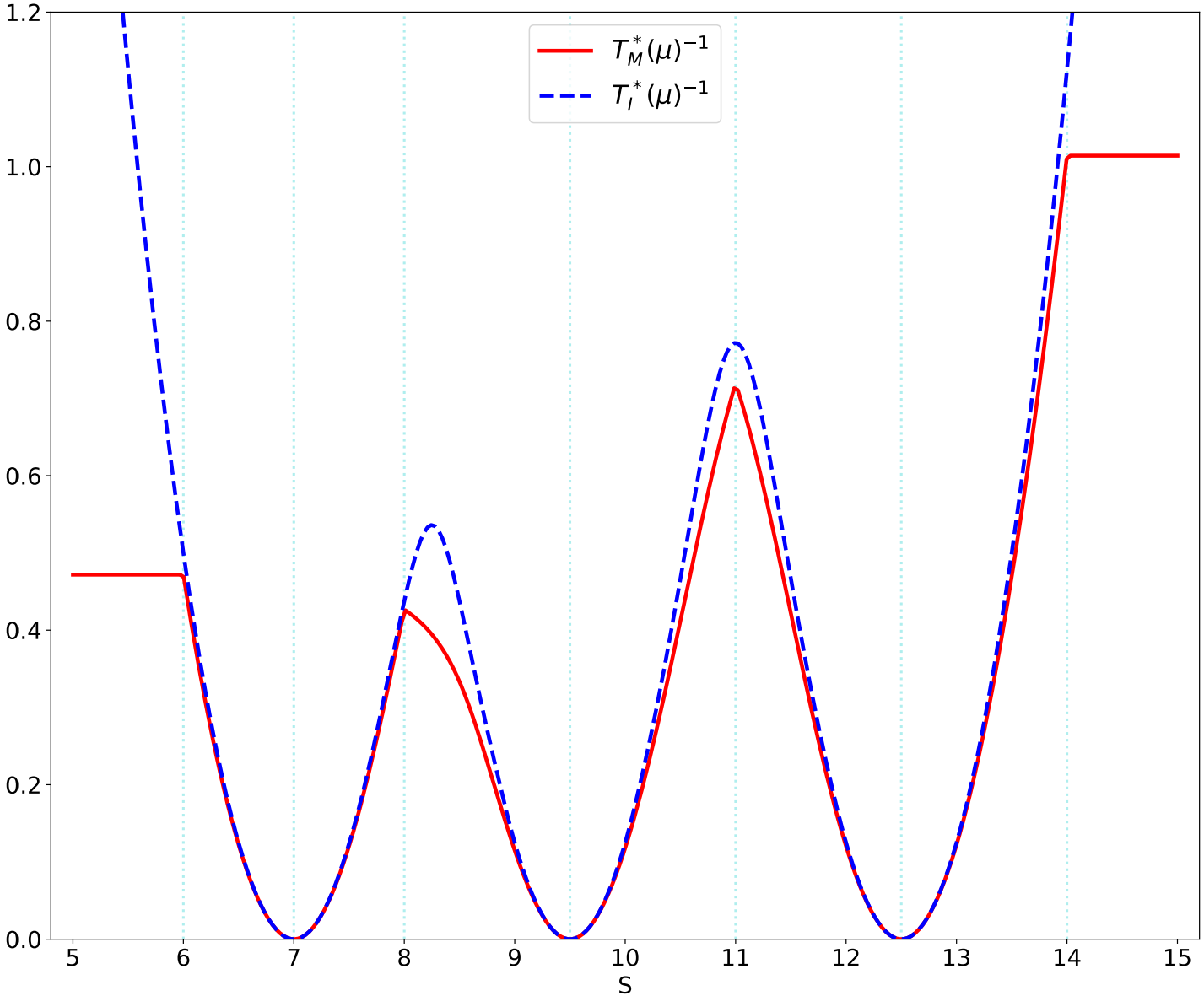
If $K = 2$,

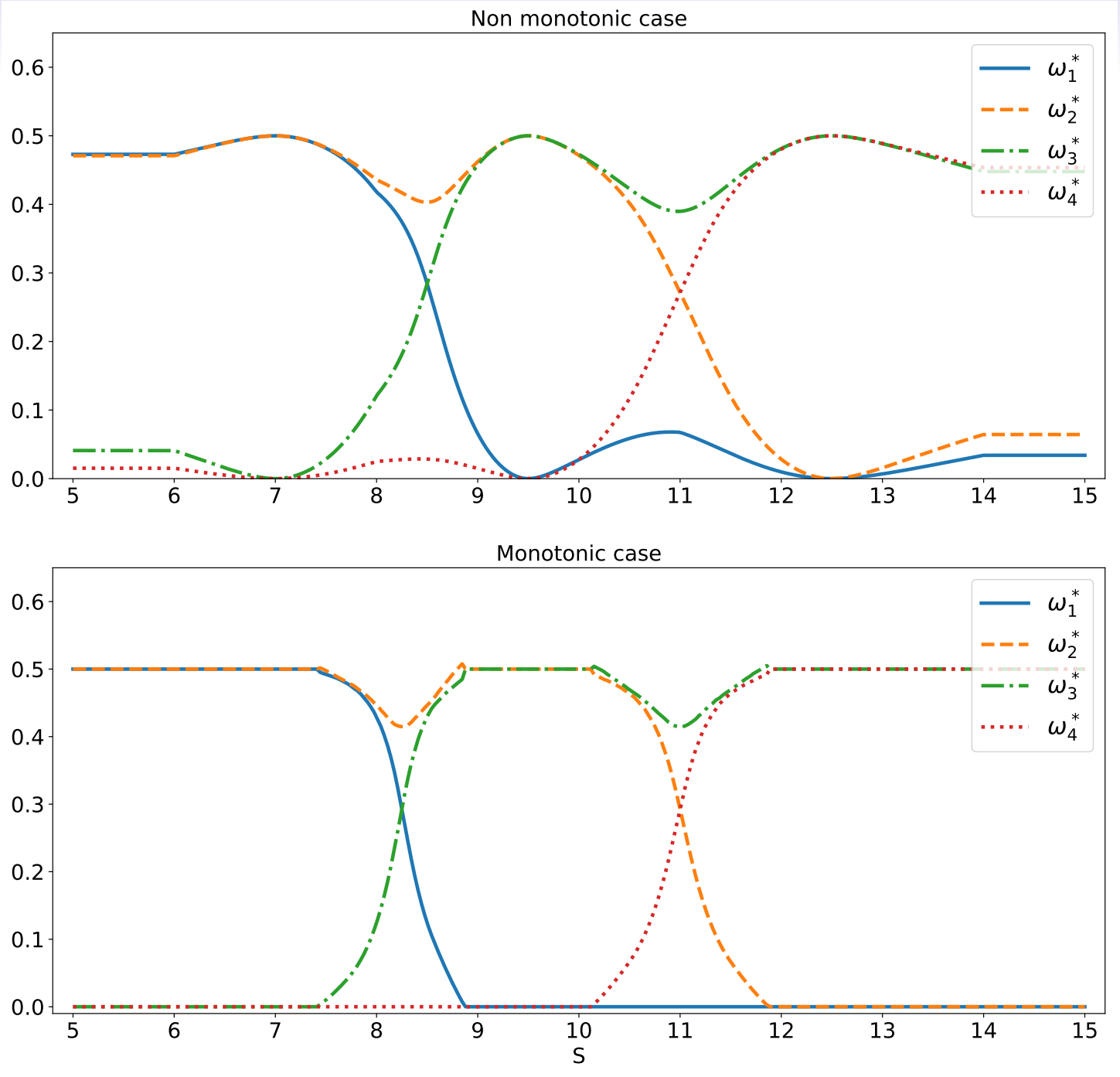
$$T^*_{\mathcal{I}}(\boldsymbol{\mu})^{-1} = (2S - \mu_1 - \mu_2)^2/8$$
$$T^*_{\mathcal{M}}(\boldsymbol{\mu})^{-1} = \min\Big((2S - \mu_1 - \mu_2)^2, (\mu_1 - \mu_2)^2\Big)/8\,.$$

$\boldsymbol{\mu} = [2,\ 4]$



K arms: $\boldsymbol{\mu} = [6,\ 8,\ 11,\ 14]$




Non monotonic case
Monotonic case
ω_1^*
ω_2^*
ω_3^*
ω_4^*
0.0
0.1
0.2
0.3
0.4
0.5
0.6
5
6
7
8
9
10
11
12
13
14
15
S

Asymptotically optimal algorithm

Algorithm 3: Direct-tracking

Sampling rule

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\operatorname{argmin}}\ N_a(t) \text{ if one } N_a(t) \text{ "too small"} & (\textit{forced exploration}) \\ \underset{1 \leqslant a \leqslant K}{\operatorname{argmax}}\ \omega_a^*(\widehat{\boldsymbol{\mu}}(t)) - N_a(t)/t & (\textit{direct tracking}) \end{cases}$$

Stopping rule

$$\tau_\delta = \inf \left\{ t \in \mathbb{N}^* : \widehat{\boldsymbol{\mu}}(t) \in \mathcal{M} \text{ and } \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt\left(\widehat{\boldsymbol{\mu}}(t), \mathcal{S}\right)} \sum_{a=1}^{K} N_a(t) \frac{(\widehat{\mu}_a(t) - \lambda_a)^2}{2} > \beta(t, \delta) \right\}$$

Decision rule

$$\widehat{a}_\tau \in \underset{1 \leqslant a \leqslant K}{\arg\min} \left| \widehat{\mu}_a(\tau) - S \right|.$$

Theorem (Asymptotic optimality)

For $\mathcal{S} \in \{\mathcal{I}, \mathcal{M}\}$, for $\beta(t,\delta) = \log(tC/\delta) + (3K+2)\log\log(tC/\delta)$
*Direct-tracking is δ-**correct on** $\mathcal{S}$ and **asymptotically optimal**, i.e.*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leqslant T^*_{\mathcal{S}}(\boldsymbol{\mu})\,.$$

where

$$C := e^{K+1}\left(\frac{2}{K}\right)^K (2(3K+2))^{3K}\,\frac{4}{\log(3)}\,.$$

Theorem (Asymptotic optimality)

For $\mathcal{S} \in \{\mathcal{I}, \mathcal{M}\}$, for $\beta(t,\delta) = \log(tC/\delta) + (3K+2)\log\log(tC/\delta)$
*Direct-tracking is δ-**correct on** $\mathcal{S}$ and **asymptotically optimal**, i.e.*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\mu}[\tau_\delta]}{\log(1/\delta)} \leqslant T^*_{\mathcal{S}}(\mu)\,.$$

One key tool :

Theorem (Concentration inequality for KL divergences)

For all $\delta \geqslant (K+1)$ and $t \in \mathbb{N}^*$ we have

$$\mathbb{P}\left(\sum_{a=1}^{K} N_a(t)\frac{(\widehat{\mu}_a(t)-\mu_a)^2}{2} \geqslant \delta\right) \leqslant e^{K+1}\left(\frac{2\delta(\delta\log(t)+1)}{K}\right)^K e^{-\delta}\,.$$

Practical implementation

How to compute the optimal weights ?

$$\omega^*(\boldsymbol{\mu}) = \mathrm{argmax}_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt(\boldsymbol{\mu}, \mathcal{S})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} .$$

Non-monotonous case: $\mathcal{S} = \mathcal{M}$

$\rightarrow$ boils down to solve one scalar equation $\rightarrow$ Fast !

Monotonous case: $\mathcal{S} = \mathcal{I}$

$\mathcal{I}_b := \{\boldsymbol{\lambda} \in \mathcal{I}, a^*_\lambda = b\}$

$$F : \omega \mapsto \inf_{\lambda \in \mathcal{A}lt(\mu, \mathcal{I})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} = \min_{b \neq a^*_\mu} \inf_{\lambda \in \mathcal{I}_b} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}, \quad (1)$$

F is concave $\rightarrow$ sub-gradient ascent on the simplex.

Monotonous case: $\mathcal{S} = \mathcal{I}$

$\mathcal{I}_b := \{\boldsymbol{\lambda} \in \mathcal{I}, a^*_\lambda = b\}$

$$F : \omega \mapsto \inf_{\lambda \in \mathcal{A}lt(\mu, \mathcal{I})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} = \min_{b \neq a^*_\mu} \inf_{\lambda \in \mathcal{I}_b} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}, \quad (1)$$

F is concave $\rightarrow$ sub-gradient ascent on the simplex.
Optimal alternative in $\overline{\mathcal{I}}_b$:

$$\boldsymbol{\lambda}^b := \arg\min_{\lambda \in \overline{\mathcal{I}}_b} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2},$$

Sub-gradient of F at ω

$$\partial F(\omega) = \operatorname*{Conv}_{b \in B_{Opt}} \left[\frac{(\mu_a - \lambda^b_a)^2}{2} \right]_{a \in \{1, \cdots, K\}},$$

where Conv denotes the convex hull and B_{Opt} the set of points that attain the minimum in (1).

How to compute λ^b ?

$$\boldsymbol{\lambda}^b := \underset{\lambda \in \overline{\mathcal{I}}_b}{\arg\min} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2},$$

$\sim$ to compute the Unimodal regression $\widehat{\boldsymbol{\lambda}}^b$ of $\boldsymbol{\mu}'$ with mode at b:

$$\widehat{\boldsymbol{\lambda}}^b = \underset{\substack{\lambda'_1 \leqslant \ldots \leqslant \lambda'_b \\ \lambda'_K \leqslant \ldots \leqslant \lambda'_b}}{\arg\min} \sum_{a=1}^{K} \omega_a \frac{(\mu'_a - \lambda'_a)^2}{2}.$$

Unimodal regression can be efficiently computed via isotonic regressions (Pool Adjacent Violators Algorithm) in $O(K)$.

$\rightarrow$ sub-gradient in $O(K^2)$.

Open questions

- Extend to any one-exponential family ? For example

$$\boldsymbol{\mu} = (\mathcal{B}(\mu_1), \ldots, \mathcal{B}(\mu_a), \ldots, \mathcal{B}(\mu_K)),$$

- Compute the sub-gradient in $O(K)$.
- Find algorithm with better theoretical/practical properties ("finite δ" bound).