Introduction to Stochastic Multi-armed Bandit

Pierre Ménard

February 14, 2018

K-armed bandit problem: parametric setting

Bernoulli rewards:

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

 ...  ... 

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

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

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 $\quad \Rightarrow R_T \sim T$

- Second idea: pull the current best empirical arm,

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

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

Ideas of strategy

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

- Second idea: pull the current best empirical arm,

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

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

$\Rightarrow$ 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 $A_t \in \operatorname{argmax}_{a\in\{1,\cdots,K\}} U_a^{\mathrm{UCB}}(t)$.

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 $\delta = 1/t$

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

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 $x < \mu$

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

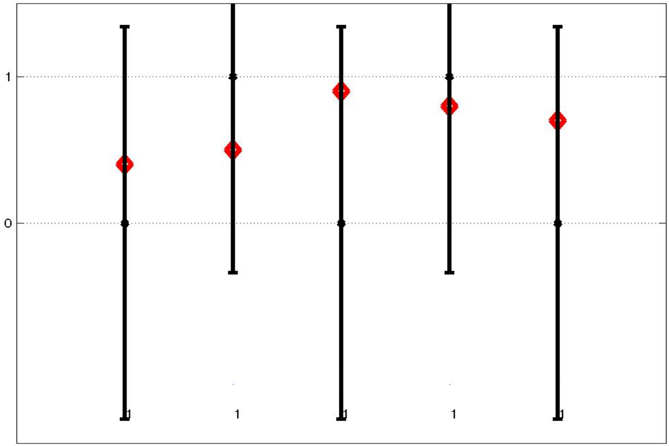
With probability at least $1 - \delta$

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

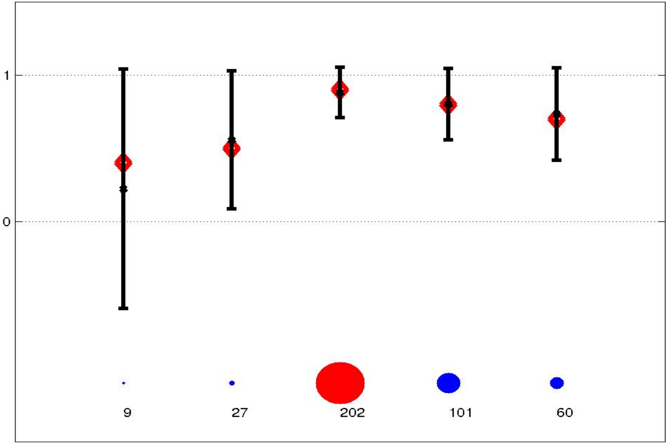
UCB index $\delta = 1/t$

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

UCB in action



UCB in action



Regret bound

Theorem

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

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

Regret bound

Theorem

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

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

therefore (Chain rule)

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

Regret bound

Theorem

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

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

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 $\mathbb{E}[N_a(T)] = o(T^\alpha)$ 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 $\mathbb{E}[N_a(T)] = o(T^\alpha)$ 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: $Y^{1:t} = (Y_1, \cdots, Y_t)$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Get the right constant: kl-UCB algorithm

Algorithm 2: The kl-UCB algorithm.

Initialization: Pull each arm of $\{1, .., K\}$ once.

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

1. Compute for each arm a the upper confidence bound

$$U_a^{kl}(t) = \sup \left\{ \mu' \geqslant \widehat{\mu}_a(t) : \ \mathrm{kl}\big(\widehat{\mu}_a(t), \mu'\big) \leqslant \frac{\log(t)}{N_a(t)} \right\}.$$

2. Play $A_t \in \mathrm{argmax}_{a \in \{1,..,K\}} U_a(t)$.

Get the right constant: kl-UCB algorithm

Theorem

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

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

Lower bound

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

K-armed bandit problem: non-parametric setting

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

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

. . .

. . .

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

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

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

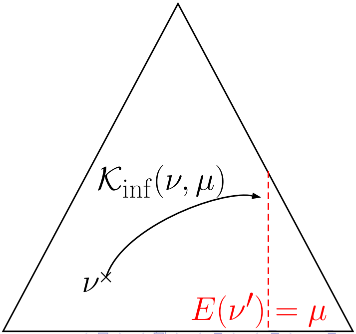
Lower bound

Theorem (Asymptotic lower)

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

KL-UCB algorithm

Algorithm 3: The KL-UCB algorithm.

Initialization: Pull each arm of $\{1, .., K\}$ once.

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

1. Compute for each arm a the upper confidence bound

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

2. Play $A_t \in \mathrm{argmax}_{a\in\{1,..,K\}} U_a(t)$.

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

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

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) ?

Usually we want to control

$$T\,\mathbb{P}\Big(\mathcal{K}_{\inf}(\widehat{\nu}_n, E(\nu)) \geqslant \log(T)\Big)$$

Variational formula

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

Variational formula

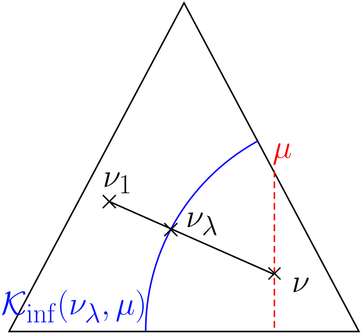
$$\mathcal{K}_{\inf}(\nu,\mu) = \max_{0 \leqslant \lambda \leqslant 1} \mathbb{E}_\nu\left[\ln\left(1 - \lambda \frac{X-\mu}{1-\mu}\right)\right].$$

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

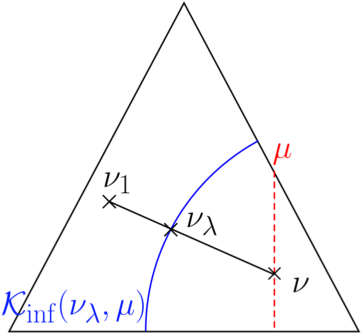
$$\mathcal{K}_{\mathsf{inf}}(\nu, \mu) = -\min_{0 \leqslant \lambda \leqslant 1} \mathrm{KL}(\nu, \nu_\lambda) = 0.$$

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:

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



Asymptotic optimality of KL-UCB algorithm

Theorem

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

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

Lower bound

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