A minimax and asymptotically optimal algorithm for stochastic bandits

Pierre Ménard Aurélien Garivier

Institut de Mathématiques de Toulouse

October 17, 2017

K-armed bandit problem

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

K-armed bandit problem

$$\nu = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_a), \dots, \mathcal{B}(\mu_K))$$

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

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

K-armed bandit problem

$$\nu = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_a), \dots, \mathcal{B}(\mu_K))$$

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

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

where $\mu^\star = \max\limits_{a=1,\dots,K} \mu_a$ and $N_a(T) = \sum_{t=1}^{T} \mathbb{I}_{\{A_t=a\}}$.

UCB algorithm

UCB algorithm:

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

$$A_{t+1} \in \arg\max_a \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{2\log(t)}{N_a(t)}}\,.$$

UCB algorithm

UCB algorithm:
Play each arm once, then for $K \leqslant t \leqslant T-1$ play:

$$A_{t+1} \in \arg\max_a \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{2\log(t)}{N_a(t)}}\,.$$

Regret bound: for all a such that $\mu^* - \mu_a > 0$

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{8}{(\mu^* - \mu_a)^2}\ln(T) + o\big(\ln(T)\big)\,,$$

UCB algorithm

UCB algorithm:
Play each arm once, then for $K \leqslant t \leqslant T-1$ play:

$$A_{t+1} \in \arg\max_a \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{2\log(t)}{N_a(t)}}\,.$$

Regret bound: for all a such that $\mu^* - \mu_a > 0$

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{8}{(\mu^* - \mu_a)^2}\ln(T) + o\big(\ln(T)\big)\,,$$

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

Asymptotic lower bound

Kullback-Leibler divergence :

$$\mathrm{kl}(p,q) := \mathrm{KL}(\mathcal{B}(p),\mathcal{B}(q)) = p\ln(p/q) + (1-p)\ln\left((1-p)/(1-q)\right)$$

Theorem (Asymptotic lower bound from Lai & Robbins)

For all consistent algorithms, for all suboptimal arms a (a such that $\mu^* - \mu_a > 0$),

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

Asymptotic lower bound

Kullback-Leibler divergence :

$$\mathrm{kl}(p,q) := \mathrm{KL}(\mathcal{B}(p),\mathcal{B}(q)) = p\ln(p/q) + (1-p)\ln\left((1-p)/(1-q)\right)$$

Theorem (Asymptotic lower bound from Lai & Robbins)

For all consistent algorithms, for all suboptimal arms a (a such that $\mu^* - \mu_a > 0$),

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

An algorithm is asymptotically optimal if (a such that $\mu^* - \mu_a > 0$)

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \leqslant \frac{1}{\mathrm{kl}(\mu_a,\mu^\star)} \underset{\text{Pinsker}}{\leqslant} \frac{2}{(\mu^* - \mu_a)^2}\,,$$

kl-UCB algorithm

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

$$A_{t+1} \in \arg\max_a \sup \left\{ \mu \in [0,1] \ : \ \mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \frac{\log(t) + 3\,\mathrm{loglog}(t)}{N_a(t)} \right\}.$$

Regret bound For a such that $\mu^* - \mu_a > 0$

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

$\Longrightarrow$ kl-UCB is asymptotically optimal.

Minimax optimality

Theorem (Minimax lower bound)

$$\sup_{\nu} R_T \geqslant C\sqrt{KT}\,.$$

An algorithm is minimax optimal if

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

UCB algorithm not minimax optimal

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

MOSS-algorithm

Algorithm:

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

$$A_{t+1} \in \arg\max_a \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{g(t,N_a(t))}{N_a(t)}}\,.$$

Algorithm	$g(t,n)$	Minimax opt.	Asymptotic opt.
UCB	$2\log(t)$	✗	✗
MOSS	$\log_+\left(\frac{T}{Kn}\right)$	✓	✗

kl-UCB^{++} algorithm

Generic algorithm:

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

$$A_{t+1} \in \arg\max_a \sup \left\{ \mu \in [0,1] \ : \ \mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \frac{g(t, N_a(t))}{N_a(t)} \right\}.$$

Algorithm	$g(T,n)$	Minimax opt.	Asymptotic opt.
kl-UCB	$\log(t) + 3\log\log(t)$	✗	✓
kl-UCB++	$\log_+\left(\frac{T}{Kn}\left(\log_+\left(\frac{T}{Kn}\right)^2 + 1\right)\right)$	✓	✓

Theorem (Optimality of kl-UCB^{++})

Minimax optimality:

$$R_T \leqslant 38\sqrt{KT} + K\,.$$

Asymptotic optimality: *For any sub-optimal arm a and any δ such that $\sqrt{11K/(2T)} \leqslant \delta \leqslant (\mu^* - \mu_a)/3$,*

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a + \delta, \mu^\star - \delta)} + O\left(\frac{\mathrm{loglog}(T)}{\delta^2}\right)$$

.

Sketch of proof: minimax bound

$$U_a(t) := \sup\left\{\mu \in [0,1] \ : \ \mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \frac{g(N_a(t))}{N_a(t)}\right\}.$$

Decomposition of the regret a^* optimal $\mu_{a^*} = \mu^*$

$$\begin{aligned} R_T &\leqslant K + \sum_{t=K}^{T-1} \mathbb{E}\big[\mu^\star - U_{A_{t+1}}(t) + U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big] \\ &\leqslant K + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[\mu^* - U_{a^*}(t)\big]}_{A} + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big]}_{B} \end{aligned}$$

Sketch of proof: minimax bound

$$U_a(t) := \sup\left\{\mu \in [0,1] \ : \ \mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \frac{g(N_a(t))}{N_a(t)}\right\}.$$

Decomposition of the regret a^* optimal $\mu_{a^*} = \mu^*$

$$R_T \leqslant K + \sum_{t=K}^{T-1} \mathbb{E}\big[\mu^\star - U_{A_{t+1}}(t) + U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big]$$

$$\leqslant K + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[\mu^* - U_{a^*}(t)\big]}_{A} + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big]}_{B}$$

B term: go back to MOSS-Index!

$$U_a(t) \leqslant B_a(t) = \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{g(N_a(t))}{2N_a(t)}}.$$

Sketch of proof: minimax bound

$$U_a(t) := \sup\left\{\mu \in [0,1] \ : \ \mathrm{kl}(\widehat{\mu}_a(t), \mu) \leqslant \frac{g(N_a(t))}{N_a(t)}\right\}.$$

Decomposition of the regret a^* optimal $\mu_{a^*} = \mu^*$

$$\begin{aligned} R_T &\leqslant K + \sum_{t=K}^{T-1} \mathbb{E}\big[\mu^\star - U_{A_{t+1}}(t) + U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big] \\ &\leqslant K + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[\mu^* - U_{a^*}(t)\big]}_{A} + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big]}_{B} \end{aligned}$$

B term: go back to MOSS-Index!

$$B \leqslant C_B\sqrt{KT}$$

Sketck of proof 2

A term: peeling trick.
Integrate the deviations $\delta_0 \sim \sqrt{K/T}$

$$E\big[\mu^* - U_{a^*}(t)\big] \leqslant \delta_0 + \int_{\delta_0}^{+\infty} \mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - u\big)\mathrm{d}u\,,$$

Split the probability

$$\mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - u\big) \leqslant$$
$$\mathbb{P}\big(\exists 1 \leqslant n \leqslant T,\ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^* - u) \geqslant g(n)/n\big) \leqslant$$
$$\underbrace{\mathbb{P}\big(\exists 1 \leqslant n \leqslant n_u,\ \ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^*) \geqslant g(n)/n\big)}_{A_1} + \underbrace{\mathbb{P}\big(\exists n_u \leqslant n \leqslant T,\ \ \widehat{\mu}_{a^*,n} \leqslant \mu^* - u\big)}_{A_2}$$

Sketck of proof 2

A term: peeling trick.

Integrate the deviations $\delta_0 \sim \sqrt{K/T}$

$$E\big[\mu^* - U_{a^*}(t)\big] \leqslant \delta_0 + \int_{\delta_0}^{+\infty} \mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - u\big)\mathrm{d}u\,,$$

Split the probability

$$\mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - u\big) \leqslant$$

$$\mathbb{P}\big(\exists 1 \leqslant n \leqslant T,\ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^* - u) \geqslant g(n)/n\big) \leqslant$$

$$\underbrace{\mathbb{P}\big(\exists 1 \leqslant n \leqslant n_u,\ \ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^*) \geqslant g(n)/n\big)}_{A_1} + \underbrace{\mathbb{P}\big(\exists n_u \leqslant n \leqslant T,\ \ \widehat{\mu}_{a^*,n} \leqslant \mu^* - u\big)}_{A_2}$$

With

$$g(n_u)/n_u \sim \frac{u^2}{2}$$

since

$$\mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^* - u) \geqslant g(n)/n \Rightarrow \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^*) \geqslant g(n)/n + u^2/2$$

Sketck of proof 2

A term: peeling trick.
Integrate the deviations $\delta_0 \sim \sqrt{K/T}$

$$E\big[\mu^* - U_{a^*}(t)\big] \leqslant \delta_0 + \int_{\delta_0}^{+\infty} \mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - u\big)\mathrm{d}u\,,$$

Split the probability

$$\mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - u\big) \leqslant$$
$$\mathbb{P}\big(\exists 1 \leqslant n \leqslant T,\ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^* - u) \geqslant g(n)/n\big) \leqslant$$
$$\underbrace{\mathbb{P}\big(\exists 1 \leqslant n \leqslant n_u,\ \ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^*) \geqslant g(n)/n\big)}_{A_1} + \underbrace{\mathbb{P}\big(\exists n_u \leqslant n \leqslant T,\ \ \widehat{\mu}_{a^*,n} \leqslant \mu^* - u\big)}_{A_2}$$

$$\left.\begin{array}{rl} A_1: & \text{Peeling trick} \\ A_2: & \text{Maximal inequality} \end{array}\right\} \Rightarrow A \leqslant C_A\sqrt{KT}$$

One-dimensional exponential family

Bandit problem:

$$\nu = \left(\nu_{\theta_1}, \dots, \nu_{\theta_a}, \dots, \nu_{\theta_K}\right),$$

one exponential family

$$\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) = \exp(x\theta - b(\theta)), \quad \text{where } b(\theta) = \log \int_{\mathbb{R}} e^{x\theta} \mathrm{d}\rho(x).$$

mean parametrization $\mu = b'(\theta)$, Pinsker-like inequality,

$$\mathrm{kl}(\mu, \mu') := \mathrm{KL}(\nu_\theta, \nu_{\theta'}) \geqslant \frac{1}{2V}(\mu - \mu')^2.$$

One-dimensional exponential family

Bandit problem:

$$\nu = \left(\nu_{\theta_1}, \dots, \nu_{\theta_a}, \dots, \nu_{\theta_K}\right),$$

one exponential family

$$\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) = \exp(x\theta - b(\theta)), \quad \text{where } b(\theta) = \log \int_{\mathbb{R}} e^{x\theta} \mathrm{d}\rho(x).$$

mean parametrization $\mu = b'(\theta)$, Pinsker-like inequality,

$$\mathrm{kl}(\mu, \mu') := \mathrm{KL}(\nu_\theta, \nu_{\theta'}) \geqslant \frac{1}{2V}(\mu - \mu')^2.$$

Regret bound for kl-UCB^{++}:

$$\mathbb{E}[N_a(T)] \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a + \delta, \mu^\star - \delta)} + O\left(\frac{\log\log(T)}{\delta^2}\right)$$

$$R_T \leqslant 76\sqrt{VKT} + (\mu^+ - \mu^-)K.$$

Open question: $\mathbb{P}[0,1]$

Bandit problem with $\nu_a \in \mathbb{P}[0,1]$:

$$\nu = \left(\nu_1, \ldots, \nu_a, \ldots, \nu_K\right),$$

Open question: $\mathbb{P}[0,1]$

Bandit problem with $\nu_a \in \mathbb{P}[0,1]$:

$$\nu = (\nu_1, \ldots, \nu_a, \ldots, \nu_K)\,,$$

Empirical measure: $\widehat{\nu}_a(t) = 1/N_a(t) \sum_{s=1}^{t} \delta_{Y_s} \mathbb{I}_{A_s=a}$

Algorithm:

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

$$A_{t+1} \in \arg\max_a \sup \left\{ E\nu'_a : \nu'_a \in \mathbb{P}[0,1],\ \mathrm{KL}(\widehat{\nu}_a(t), \nu'_a) \leqslant \frac{g(t, N_a(t))}{N_a(t)} \right\}.$$

Algorithm	$g(t,n)$	Minimax opt.	Asymptotic opt.
KL-UCB	$\log(t) + \mathrm{loglog}(t)$	✗	✓
KL-UCB+	$\log_+\left(\frac{T}{Kn}\right)$	?	?