The True Shape of Regret in Bandit problems

Aurélien Garivier, **Pierre Ménard**, Gilles Stoltz

July 5, 2016

Environment and strategy

- K arms bandit problem, $\nu = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_K))$ with $\mu_i \in (0,1)$.

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

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

- Information available at time t: $Y_{1:t} = (Y_1, \ldots, Y_t)$.

Environment and strategy

- K arms bandit problem, $\nu = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_K))$ with $\mu_i \in (0,1)$.

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

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

- Information available at time t: $Y_{1:t} = (Y_1, \ldots, Y_t)$.

- Goal of the player, minimize the expected regret :

$$R_{\nu,T} = T\mu^\star - \mathbb{E}_\nu\left[\sum_{t=1}^{T} Y_t\right] = \sum_{a=1}^{K} (\mu^\star - \mu_a)\, \mathbb{E}_\nu[N_a(T)] \,.$$

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

Theorem (Asymptotic lower bound from Lai & Robbins)

For all reasonable strategies (consistent), for all bandits problems ν, for all suboptimal arms a,

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

where kl the Kullback-Leibler divergence for Bernoulli distributions :

$$\forall p, q \in [0,1]^2, \qquad \mathrm{kl}(p,q) := p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \geqslant 2(p-q)^2 .$$

Theorem (Asymptotic lower bound from Lai & Robbins)

For all reasonable strategies (consistent), for all bandits problems ν, for all suboptimal arms a,

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

where kl the Kullback-Leibler divergence for Bernoulli distributions :

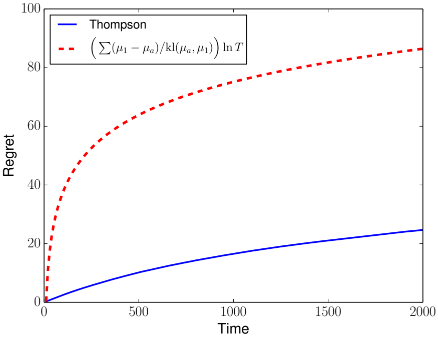
$$\forall p, q \in [0,1]^2, \qquad \mathrm{kl}(p,q) := p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \geqslant 2(p-q)^2 .$$

Theorem (UCB algorithm from Auer & Cesa-Bianchi & Fischer)

Algorithm UCB, for all bandits problems ν, for all suboptimal arm a:

$$\mathbb{E}_\nu\left[N_a(T)\right] \leqslant \frac{8\ln(T)}{(\mu^\star - \mu_a)^2} + 2$$

(right constant with KL-UCB algorithm from Cappe & al).



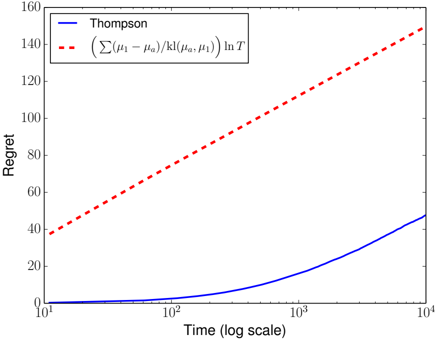


Figure : Bernoulli bandit problem with parameters : $(\mu_a)_{1\leqslant a\leqslant 6} = (0.05, 0.04, 0.02, 0.015, 0.01, 0.005)$

- Logarithmic regret for large T (**asymptotic** lower bound).
- Transition phase between.
- Linear regret for T small.

Consistent strategy

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 $\Delta_a > 0$, it satisfies $\mathbb{E}_\nu\left[N_a(T)\right] = o(T^\alpha)$ for all $0 < \alpha \leqslant 1$.

$$\nu = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_K)) \qquad \nu' = (\mathcal{B}(\mu'_1), .., \mathcal{B}(\mu'_K))$$

$$\boxed{\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)] \,\mathrm{kl}(\mu_a, \mu'_a) \geqslant \mathrm{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]) \,,} \tag{M}$$

where Z is a $\sigma(Y_{1:T})$–measurable random variable with values in $[0, 1]$.
Typically $Z = N_a(T)/T$.

$$\nu = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_K)) \qquad \nu' = (\mathcal{B}(\mu'_1), .., \mathcal{B}(\mu'_K))$$

$$\boxed{\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)]\, \mathrm{kl}(\mu_a, \mu'_a) \geqslant \mathrm{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z])\,,} \tag{M}$$

where Z is a $\sigma(Y_{1:T})$–measurable random variable with values in $[0,1]$. Typically $Z = N_a(T)/T$.

Sketch of proof :

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

$$\mathrm{KL}(\mathbb{P}_\nu^{Y_{1:T}}, \mathbb{P}_{\nu'}^{Y_{1:T}}) \geqslant \mathrm{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z])\,,$$

where :

- $\mathbb{P}_\nu^{Y_{1:T}}$ and $\mathbb{P}_{\nu'}^{Y_{1:T}}$ respective distributions of $Y_{1:T}$ under $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'}$
- chain rule for Kullback-Leibler divergences
- contraction of entropy

Proof.

- Contraction of entropy :

Let $V \sim \mathcal{U}[0,1]$ independent of $Y_{1:T}$, and the event $E = \{Z \geqslant V\}$ then

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\nu}^{Y_{1:T}}, \mathbb{P}_{\nu'}^{Y_{1:T}}) &= \mathrm{KL}(\mathbb{P}_{\nu}^{Y_{1:T}} \otimes \mathcal{U}[0,1],\ \mathbb{P}_{\nu'}^{Y_{1:T}} \otimes \mathcal{U}[0,1]) \\
&\geqslant \mathrm{KL}\Big((\mathbb{P}_{\nu}^{Y_{1:T}} \otimes \mathcal{U}[0,1])^{\mathbb{I}_E},\ (\mathbb{P}_{\nu'}^{Y_{1:T}} \otimes \mathcal{U}[0,1])^{\mathbb{I}_E} \Big) \\
&= \mathrm{kl}\big((\mathbb{P}_{\nu}^{Y_{1:T}} \otimes \mathcal{U}[0,1])(E),\ (\mathbb{P}_{\nu'}^{Y_{1:T}} \otimes \mathcal{U}[0,1])(E) \big) .
\end{aligned}$$

To conclude, for $\alpha = \nu$ or ν' (Fubini theorem):

$$(\mathbb{P}_{\alpha}^{Y_{1:T}} \otimes \mathcal{U}[0,1])(E) = \mathbb{E}_{\alpha}[Z] .$$

$\square$

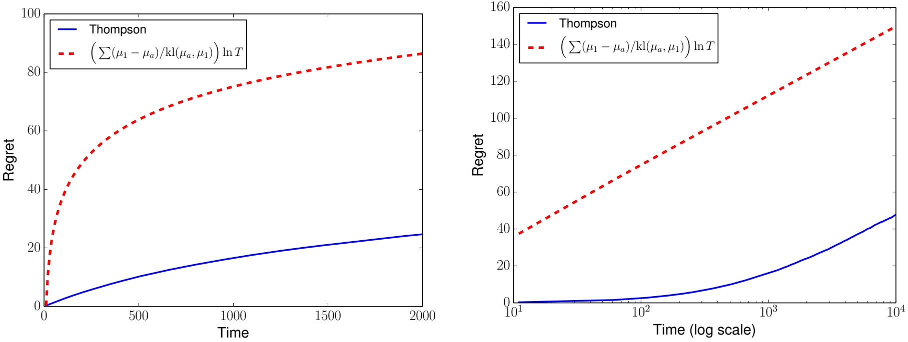


Figure : Bernoulli bandit problem with parameters : $(\mu_a)_{1\leqslant a\leqslant 6} = (0.05, 0.04, 0.02, 0.015, 0.01, 0.005)$

- Linear regret for T small.
- Logarithmic regret for large T (**asymptotic** lower bound).
- Transition phase between.

Absolute lower bound for small T

In the remainder of this section $\nu = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_K))$ with an unique optimal arm $i^\star$.

Absolute lower bound for small T

In the remainder of this section $\nu = (\mathcal{B}(\mu_1), .., \mathcal{B}(\mu_K))$ with an unique optimal arm $i^\star$.

Uniform strategy : pull an arm uniformly at random at each round.

Definition

A strategy is smarter than the uniform strategy if for all bandit problems ν, for all $T \geqslant 1$,

$$\mathbb{E}_\nu\left[N_{i^\star}(T)\right] \geqslant \frac{T}{K}$$

$$\mathbb{E}_\nu\left[N_a(T)\right] \leqslant \frac{T}{K} \quad \text{if } a \text{ suboptimal.}$$

Theorem

For all strategies that are smarter than the uniform strategy, for all bandit problems ν, for all suboptimal arms a, for all $T \geqslant 1$,

$$\mathbb{E}_\nu\left[N_a(T)\right] \geqslant \frac{T}{K}\left(1 - \sqrt{2T\mathrm{kl}(\mu_a, \mu^\star)}\right).$$

In particular,

$$\forall\, T \leqslant \frac{1}{8\mathrm{kl}(\mu_a, \mu^\star)}, \qquad \mathbb{E}_\nu\left[N_a(T)\right] \geqslant \frac{T}{2K}.$$

Linear regret

a suboptimal arm.

Modified bandit problem with $\mu'_a > \mu^\star$:

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

a suboptimal arm.
Modified bandit problem with $\mu'_a > \mu^\star$:

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

Main inequality (M),

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

$$\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T \leqslant 1/K \leqslant \mathbb{E}_{\nu'}\big[N_a(T)\big]/T\Big) \qquad \geqslant \mathrm{kl}\Big(\mathbb{E}_\nu[N_a(T)]/T,\ 1/K\Big)$$

$$\text{(Pinsker inequality)} \qquad \geqslant \frac{K}{2}\Big(\mathbb{E}_\nu[N_a(T)]/T - 1/K\Big)^2$$

a suboptimal arm.
Modified bandit problem with $\mu'_a > \mu^\star$:

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

Main inequality (M),

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

$$\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T \leqslant 1/K \leqslant \mathbb{E}_{\nu'}\big[N_a(T)\big]/T\Big) \qquad \geqslant \mathrm{kl}\Big(\mathbb{E}_\nu[N_a(T)]/T,\ 1/K\Big)$$

$$(\text{Pinsker inequality}) \qquad \geqslant \frac{K}{2}\Big(\mathbb{E}_\nu[N_a(T)]/T - 1/K\Big)^2$$

Still with $\mathbb{E}_\nu[N_a(T)]/T \leqslant 1/K$:

$$\mathrm{kl}(\mu_a, \mu'_a)\,T/K \geqslant \frac{K}{2}\Big(\mathbb{E}_\nu[N_a(T)]/T - 1/K\Big)^2$$

Collective lower bound for small T

Theorem

For all strategies that are smarter than the uniform strategy, for all bandit problems ν, for all suboptimal arms a,

$$\forall\, T \leqslant \frac{K\,?}{8\mathrm{kl}(\mu_a, \mu^\star)}, \qquad \mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{T}{2K}\,.$$

Collective lower bound for small T

Theorem

For all strategies that are smarter than the uniform strategy, for all bandit problems ν, for all suboptimal arms a,

$$\forall\, T \leqslant \frac{K\,?}{8\mathrm{kl}(\mu_a, \mu^\star)}, \qquad \mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{T}{2K}\,.$$

Theorem

Under weak (symmetry, ...) assumptions on the strategy, for all bandit problems ν,

$$\sum_{a \neq i^\star} \mathbb{E}_\nu\big[N_a(T)\big] \geqslant T\left(1 - \frac{1}{K} - \frac{\sqrt{2T\,\mathrm{kl}(\mu_a, \mu^\star)}}{K} - \frac{2T\,\mathrm{kl}(\mu_a, \mu^\star)}{K}\right).$$

Non-asymptotic bounds for large T

Theorem

For all reasonable strategies (refinement of consistence), for all bandit problems ν, for all suboptimal arms a,

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

with a closed-form expression for the last term.

Where, for T large enough

$$\mathrm{O}\Big(\ln(\ln T)\Big) = \frac{1}{(1-\mu^\star)\mathrm{kl}(\mu_a, \mu^\star)}(\ln T)^{-3} + C_{\psi,\mathcal{D}} H(\nu) \frac{\ln(T)^2}{T} + \ln\Big(K\, C_{\psi,\mathcal{D}} (\ln T)^9\Big) + \frac{\ln 2}{\mathrm{kl}(\mu_a, \mu^\star)}\,.$$

General bandit problems

$$\nu = (\nu_1, .., \nu_K) ,$$

$\nu_a \in \mathcal{D}$ a probability distribution and a real number x, we introduce

$$\mathcal{K}_{\text{inf}}(\nu_a, x) = \inf\Big\{\mathrm{KL}(\nu_a, \nu'_a) : \quad \nu'_a \in \mathcal{D} \;\; \text{and} \;\; E(\nu'_a) > x\Big\} ;$$

by convention, the infimum of the empty set equals $+\infty$.

$$\rightarrow \text{ replace } \mathrm{kl}(\mu_a, \mu^\star) \text{ by } \mathcal{K}_{\text{inf}}(\nu_a, \mu^\star)$$

References I

P. Auer, N. Cesa-Bianchi, and P. Fischer.
Finite-time analysis of the multiarmed bandit problem.
Machine Learning, 47(2-3):235–256, 2002.

A.N. Burnetas and M.N. Katehakis.
Optimal adaptive policies for sequential allocation problems.
Advances in Applied Mathematics, 17(2):122–142, 1996.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz.
Kullback-Leibler upper confidence bounds for optimal sequential allocation.
Annals of Statistics, 41(3):1516–1541, 2013.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz.
Explore first, exploit next: The true shape of regret in bandit problems.
arXiv preprint arXiv:1602.07182, 2016.

T. L. Lai and H. Robbins.
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6:4–22, 1985.