Adaptation to the range in K–armed stochastic bandits

Gilles Stoltz

Laboratoire de mathématiques d'Orsay

Joint work with Hédi Hadiji, now at University of Amsterdam

K–armed stochastic bandits

Framework and statement of regret bounds

K probability distributions $\nu_1, \ldots, \nu_K$
with expectations $\mu_1, \ldots, \mu_K$ $\longrightarrow$ $\mu^\star = \max_{a \in [K]} \mu_a$

At each round $t = 1, 2, \ldots,$

1. Statistician picks arm $A_t \in [K]$
2. She gets a reward Y_t drawn according to ν_{A_t}
3. This is the only feedback she receives

$\longrightarrow$ Exploration–exploitation dilemma
estimate the ν_a vs. get high rewards Y_t

Pseudo-regret:

$$R_T = \sum_{t=1}^{T} \bigl(\mu^\star - \mathbb{E}[Y_t]\bigr) = \sum_{t=1}^{T} \bigl(\mu^\star - \mathbb{E}[\mu_{A_t}]\bigr)$$

$$= \sum_{a \in [K]} \left((\mu^\star - \mu_a)\, \mathbb{E}\!\left[\sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}} \right] \right) = \sum_{a \in [K]} (\mu^\star - \mu_a)\, \mathbb{E}\bigl[N_a(T)\bigr]$$

Model: $\nu_1, \ldots, \nu_K$ are distributions over $[0, 1]$

A classical strategy: UCB [upper confidence bound]

Auer, Cesa-Bianchi and Fisher [2002]

For $t \geqslant K$, pick $$A_{t+1} \in \mathop{\arg\max}_{a \in [K]} \left\{ \widehat{\mu}_a(t) + \sqrt{\frac{2 \ln t}{N_a(t)}} \right\}$$

Exploitation: cf. empirical mean $\widehat{\mu}_a(t)$

Exploration: cf. $\sqrt{2 \ln t / N_a(t)}$ favors arms a not pulled often

Two types of regret bounds

– Distribution-dependent bound: $$R_T \lesssim \sum_{a : \mu_a < \mu^\star} \frac{8 \ln T}{\mu^\star - \mu_a}$$

– Distribution-free bound: $$\sup_{\nu_1, \ldots, \nu_K} R_T \lesssim \sqrt{8KT \ln T}$$

Model: $\nu_1, \ldots, \nu_K$ are distributions over $[0, 1]$

Optimal bounds read as follows:

Distribution-free bound: $\sup_{\nu_1,\ldots,\nu_K} R_T$ at best $\Theta\big(\sqrt{KT}\big)$

Upper bound $K + 45\sqrt{KT}$ for the MOSS strategy by Audibert and Bubeck [2009]

Lower bound $(1/20)\sqrt{KT}$ by Auer, Cesa-Bianchi, Freund and Schapire [2002]

Distribution-dependent bound: $\displaystyle\sum_{a:\mu_a<\mu^\star} \frac{\mu^\star - \mu_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} \ln T - \Theta(\ln \ln T)$

where $\mathcal{K}_{\inf}(\nu_a, \mu^\star) = \inf\big\{\mathrm{KL}(\nu_a, \nu_a') : E(\nu_a') > \mu^\star\big\}$

References: Lai and Robbins [1985], Burnetas and Katehakis [1996], Honda and Takemura [2015], Garivier, Ménard and Stoltz [2019], among others

Both bounds can be achieved simultaneously!

By combining the MOSS strategy and the KL-UCB strategy by Cappé et al. [2013]; see the KL-UCB-switch strategy by Garivier, Hadiji, Ménard, Stoltz [submitted]

Proofs of the regret lower bounds on $[0, 1]$

(At least, high-level ideas...)

Proof ideas for the lower bounds

Strategy ψ: maps $H_t = (Y_1, \ldots, Y_t) \mapsto A_{t+1} = \psi_t(H_t)$

Change of measure: compare distributions of H_T
under $\underline{\nu} = (\nu_1, \ldots, \nu_K)$ vs. $\underline{\nu}' = (\nu'_1, \ldots, \nu'_K)$

Fundamental inequality: performs an implicit change of measure

Reference: Lai and Robbins [1985], Auer et al. [2002], Garivier et al. [2019]

For all Z taking values in $[0,1]$ and $\sigma(H_T)$–measurable,

$$\sum_{a \in [K]} \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \, \mathrm{KL}(\nu_a, \nu'_a) = \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{H_T}, \mathbb{P}_{\underline{\nu}'}^{H_T}\big)$$

$$\geqslant \mathrm{kl}\big(\mathbb{E}_{\underline{\nu}}[Z], \mathbb{E}_{\underline{\nu}'}[Z]\big)$$

where $\mathrm{kl}(p,q) = \mathrm{KL}\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big)$

Later use: $\underline{\nu}'$ only differs from $\underline{\nu}$ at some a, with $Z = N_a(T)/T$

Distribution-free lower bound, for distributions over $[0,1]$

Problem $\underline{\nu}_0 = \big(\mathrm{Ber}(1/2)\big)_{a\in[K]}$ vs. $\underline{\nu}_k = \Big(\mathrm{Ber}\big(1/2+\varepsilon\,\mathbb{I}_{\{a=k\}}\big)\Big)_{a\in[K]}$

$$R_T \stackrel{\text{def}}{=} \sum_{a\neq k} \varepsilon\,\mathbb{E}_{\underline{\nu}_k}\big[N_a(T)\big] = T\varepsilon\Big(1-\mathbb{E}_{\underline{\nu}_k}\big[N_k(T)/T\big]\Big)$$

Thus, $\displaystyle \sup_{\underline{\nu}} R_T \geqslant \sup_{\varepsilon\in(0,1)} \max_{k\in[K]} T\varepsilon\Big(1-\mathbb{E}_{\underline{\nu}_k}\big[N_k(T)/T\big]\Big)$

Fundamental inequality, with $k\in[K]$ such that $\mathbb{E}_{\underline{\nu}_0}\big[N_k(T)/T\big]\leqslant 1/K$

+ Pinsker's inequality with $Z = N_k(T)/T$

$$2\Big(\mathbb{E}_{\underline{\nu}_k}\big[N_k(T)/T\big]-\mathbb{E}_{\underline{\nu}_0}\big[N_k(T)/T\big]\Big)^2 \leqslant \mathrm{kl}\big(\mathbb{E}_{\underline{\nu}_0}[Z],\,\mathbb{E}_{\underline{\nu}_k}[Z]\big)$$
$$\leqslant \underbrace{\mathbb{E}_{\underline{\nu}_0}\big[N_k(T)\big]}_{\leqslant T/K}\;\underbrace{\mathrm{KL}\big(\mathrm{Ber}(1/2),\,\mathrm{Ber}(1/2+\varepsilon)\big)}_{=-\ln(1-4\varepsilon^2)/2\,\leqslant\,2.5\varepsilon^2}$$

Thus, $\displaystyle \sup_{\underline{\nu}} R_T \geqslant \sup_{\varepsilon\in(0,1/4)} T\varepsilon\Big(1-1/K-\varepsilon\sqrt{1.25\,T/K}\Big) \geqslant \Theta\big(\sqrt{KT}\big)$

Distribution-dependent bound: $R_T = \sum_{a \in [K]} (\mu^\star - \mu_a)\, \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]$

We lower bound each $\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]$ for a fixed a with $\mu_a < \mu^\star$; let ν'_a with $\mu_a > \mu^\star$

Problems $\underline{\nu} = (\nu_a)_{a\in[K]}$ vs. $\underline{\nu}' = (\nu_1, \ldots, \nu_{a-1}, \nu'_a, \nu_{a+1}, \ldots, \nu_K)$

Fundamental inequality
on "good" strategies $\qquad \forall \alpha \in (0,1],\ \ \mathbb{E}[N_k(T)] = o(T^\alpha)$ for subopt. k
& lower bound on kl $\qquad \mathrm{kl}(p,q) \geqslant (1-p)\ln\big(1/(1-q)\big) - \ln 2$

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) \geqslant \mathrm{kl}\Big(\overbrace{\mathbb{E}_{\underline{\nu}}\big[N_a(T)/T\big]}^{=o(1)},\ \mathbb{E}_{\underline{\nu}'}\big[N_a(T)/T\big]\Big)$$
$$\gtrsim \ln\Big(1/\big(1-\mathbb{E}_{\underline{\nu}'}\big[N_a(T)/T\big]\big)\Big)$$

Since $\mathbb{E}_{\underline{\nu}'}\big[N_a(T)/T\big] = 1 - \sum_{k\neq a} \mathbb{E}_{\underline{\nu}'}\big[N_k(T)/T\big] \gtrsim 1 - T^{\alpha-1}$, we get:

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) \gtrsim \ln T^{1-\alpha}$$

Distribution-dependent bound: $$R_T = \sum_{a \in [K]} \bigl(\mu^\star - \mu_a\bigr)\, \mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]$$

We lower bound each $\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]$ for a fixed a with $\mu_a < \mu^\star$; let ν'_a with $\mu_a > \mu^\star$

$$\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]\,\mathrm{KL}(\nu_a, \nu'_a) \gtrsim \ln T^{1-\alpha}, \qquad \text{that is,} \qquad \frac{\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]\,\mathrm{KL}(\nu_a, \nu'_a)}{\ln T} \gtrsim 1 - \alpha \to 1$$

Therefore, "good" strategies can ensure, at best:

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]}{\ln T} \geqslant \sup_{\nu'_a : \mu'_a > \mu^\star} \frac{1}{\mathrm{KL}(\nu_a, \nu'_a)} \stackrel{\text{def}}{=} \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$$

By summing over suboptimal arms:

$$\liminf_{T \to \infty} \frac{R_T}{\ln T} \geqslant \sum_{a \in [K]} \frac{\mu^\star - \mu_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$$

How do we prove the fundamental inequality?

For all Z taking values in $[0,1]$ and $\sigma(H_T)$–measurable,

$$\sum_{a \in [K]} \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \, \mathrm{KL}(\nu_a, \nu'_a) = \mathrm{KL}\big(\mathbb{P}^{H_T}_{\underline{\nu}}, \, \mathbb{P}^{H_T}_{\underline{\nu}'}\big) \geqslant \mathrm{kl}\big(\mathbb{E}_{\underline{\nu}}[Z], \, \mathbb{E}_{\underline{\nu}'}[Z]\big)$$

Equality: chain rule for KL

$$H_t = (Y_1, \ldots, Y_t) \mapsto A_{t+1} = \psi_t(H_t)$$

and $Y_{t+1} \mid H_t \sim \nu_{A_{t+1}}$

$$\begin{aligned} \mathrm{KL}\big(\mathbb{P}^{H_{t+1}}_{\underline{\nu}}, \, \mathbb{P}^{H_{t+1}}_{\underline{\nu}'}\big) &= \mathrm{KL}\big(\mathbb{P}^{H_t}_{\underline{\nu}}, \, \mathbb{P}^{H_t}_{\underline{\nu}'}\big) + \mathbb{E}_{\underline{\nu}}\big[\mathrm{KL}(\nu_{A_{t+1}}, \, \nu'_{A_{t+1}})\big] \\ &= \mathrm{KL}\big(\mathbb{P}^{H_t}_{\underline{\nu}}, \, \mathbb{P}^{H_t}_{\underline{\nu}'}\big) + \mathbb{E}_{\underline{\nu}}\left[\sum_{a \in [K]} \mathrm{KL}(\nu_a, \, \nu'_a) \, \mathbb{I}_{\{A_{t+1} = a\}}\right] \end{aligned}$$

Conclude by induction

How do we prove the fundamental inequality?

For all Z taking values in $[0,1]$ and $\sigma(H_T)$–measurable,

$$\sum_{a\in[K]} \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) = \mathrm{KL}\Big(\mathbb{P}^{H_T}_{\underline{\nu}},\,\mathbb{P}^{H_T}_{\underline{\nu}'}\Big) \geqslant \mathrm{kl}\Big(\mathbb{E}_{\underline{\nu}}[Z],\,\mathbb{E}_{\underline{\nu}'}[Z]\Big)$$

Inequality: data-processing inequality $\qquad \mathrm{KL}(\mathbb{P}^X,\mathbb{Q}^X) \leqslant \mathrm{KL}(\mathbb{P},\mathbb{Q})$

First: $$\mathrm{KL}\Big(\mathbb{P}^{H_T}_{\underline{\nu}},\,\mathbb{P}^{H_T}_{\underline{\nu}'}\Big) \geqslant \mathrm{KL}\Big(\mathbb{P}^{Z}_{\underline{\nu}},\,\mathbb{P}^{Z}_{\underline{\nu}'}\Big) = \mathrm{KL}\Big(\mathbb{P}^{Z}_{\underline{\nu}}\otimes\mathfrak{m},\,\mathbb{P}^{Z}_{\underline{\nu}'}\otimes\mathfrak{m}\Big)$$

with $\mathfrak{m}$ the Lebesgue measure on $[0,1]$

Second: $$\mathrm{KL}\Big(\mathbb{P}^{Z}_{\underline{\nu}}\otimes\mathfrak{m},\,\mathbb{P}^{Z}_{\underline{\nu}'}\otimes\mathfrak{m}\Big) \geqslant \mathrm{KL}\bigg(\underbrace{\big(\mathbb{P}^{Z}_{\underline{\nu}}\otimes\mathfrak{m}\big)^{\mathbb{I}_E}}_{=\mathrm{Ber}(\mathbb{E}_{\underline{\nu}}[Z])},\,\big(\mathbb{P}^{Z}_{\underline{\nu}'}\otimes\mathfrak{m}\big)^{\mathbb{I}_E}\bigg)$$

with $E = \big\{(z,x) : z \leqslant x\big\}$, yielding

$$\mathbb{P}^{Z}_{\underline{\nu}}\otimes\mathfrak{m}(E) = \int \mathbb{I}_{\{z\leqslant x\}}\,\mathrm{d}\mathfrak{m}(x)\,\mathrm{d}\mathbb{P}^{Z}_{\underline{\nu}}(z) = \int z\,\mathrm{d}\mathbb{P}^{Z}_{\underline{\nu}}(z) = \mathbb{E}_{\underline{\nu}}[Z]$$

I call the second application "Ménard's lemma" (see Garivier, Ménard and Stoltz, 2019)

Adaptation to the range

Bounded but unknown range

Reference for the final part of this talk: Hadiji and Stoltz [2020]

That is, model: $$\mathcal{D} = \bigcup_{m,M : m<M} \mathcal{D}_{m,M}$$

where $\mathcal{D}_{m,M}$ set of distributions ν over a given interval $[m, M]$

What changes?

Same distribution-free lower bound:

$\Theta\big((M-m)\sqrt{KT}\big)$ by rescaling

Any worsening due to ignorance of the range? No! (or almost)

Different distribution-dependent lower bound:

$R_T / \ln T \to +\infty$ as $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}) = 0$

But any rate $\gg \ln T$ may be achieved

Focus on the UCB strategy

With a known range $[m, M]$, reads (knowledge of the range is key!)

$$A_{t+1} \in \mathop{\arg\max}_{a \in [K]} \left\{ \widehat{\mu}_a(t) + (M - m)\sqrt{\frac{2 \ln t}{N_a(t)}} \right\}$$

Extension to an unknown range:

$$A_{t+1} \in \mathop{\arg\max}_{a \in [K]} \left\{ \widehat{\mu}_a(t) + \sqrt{\frac{\varphi(t)}{N_a(t)}} \right\}$$

where $\ln t \ll \varphi(t) \ll t$

Guarantee: for all bandit problems $\nu_1, \ldots, \nu_K$ in $\mathcal{D}$,

$$\limsup \frac{R_T}{\varphi(T)} < +\infty$$

$\Phi_{\mathrm{dep}} = \varphi$ is the corresponding distribution-dependent rate for adaptation to the range

Distribution-free rate for adaptation to the range

$\Phi_{\text{free}} : \mathbb{N} \to (0, +\infty)$ such that

$\forall m < M$,
$\forall \nu_1, \ldots, \nu_K$ in $\mathcal{D}_{m,M}$,
$\forall T \geqslant 1$,

$$R_T \leqslant (M - m)\Phi_{\text{free}}(T)$$

By the lower bound proved for $[m, M] = [0, 1]$:

$$\Phi_{\text{free}}(T) \geqslant \Theta\bigl(\sqrt{KT}\bigr)$$

AdaHedge on estimated payoffs + mixing achieves

$$\Phi_{\text{free}}(T) \approx 7(M - m)\sqrt{TK \ln K}$$

Reference for AdaHedge: Cesa-Bianchi, Mansour, Stoltz [2005, 2007] and De Rooij, van Erven, Grünwald, Koolen [2014]

Note: $\sqrt{\ln K}$ shaved off (with different strategy) when M is known

What about simultaneous bounds?

Reminder for known range: $\ln T$ and $\sqrt{T}$ rates for regret upper bounds

Theorem: If $\Phi_{\text{free}}(T) \ll T$ then $\Phi_{\text{dep}}(T) \times \Phi_{\text{free}}(T) \geqslant \Theta(T)$

Example: $\Phi_{\text{free}}(T) = \Theta(\sqrt{T})$ now forces $\Phi_{\text{dep}}(T) \geqslant \Theta(\sqrt{T})$

$\rightarrow$ We finally exhibit some price for adaptation!

AdaHedge on estimated payoffs + mixing simultaneously achieves

$$\Phi_{\text{free}}(T) = \Theta(\sqrt{T}) \quad \text{and} \quad \Phi_{\text{dep}}(T) = \Theta(\sqrt{T})$$

Analysis heavily based on Seldin and Lugosi [2017]

Actually, all pairs $\Phi_{\text{free}}(T) = \Theta(T^{\alpha})$ and $\Phi_{\text{dep}}(T) = \Theta(T^{1-\alpha})$
with $\alpha \in [1/2,\ 1)$ may be achieved, by setting the mixing factor properly

Next page: proof of the theorem above, consisting in showing

$$\big(\Phi_{\text{free}}(T)/T\big)\, \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \gtrsim \text{cst}$$

We lower bound each $\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]$ for a fixed a with $\mu_a < \mu^\star$

Problems $\underline{\nu}$, $\underline{\nu}'$ only differing at $\nu'_a = (1-\varepsilon)\nu_a + \varepsilon\,\delta_{\mu_a + c/\varepsilon}$

such that $\nu_a \perp \delta_{\mu_a+c/\varepsilon}$ and $\mu'_a > \mu^\star$

$f = \dfrac{\mathrm{d}\nu_a}{\mathrm{d}\nu'_a} = \dfrac{1}{1-\varepsilon}$ so that $\mathrm{KL}(\nu_a, \nu'_a) = \mathbb{E}_{\nu_a}[\ln f] \approx \varepsilon$

Fundamental inequality and $\mathrm{kl}(p,q) \gtrsim (1-p)\ln\big(1/(1-q)\big)$

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \overbrace{\mathrm{KL}(\nu_a, \nu'_a)}^{\approx\varepsilon} \geqslant \mathrm{kl}\Big(\overbrace{\mathbb{E}_{\underline{\nu}}\big[N_a(T)/T\big]}^{=o(1)}, \mathbb{E}_{\underline{\nu}'}\big[N_a(T)/T\big]\Big)$$
$$\gtrsim \ln\Big(1/\big(1-\mathbb{E}_{\underline{\nu}'}\big[N_a(T)/T\big]\big)\Big)$$

Indeed: $(\mu^\star - \mu_a)\,\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \leqslant R_T(\underline{\nu}) \leqslant (M-m)\,\Phi_{\mathrm{free}}(T) \ll T$

Similarly: $\ln\Big(1/\big(1-\mathbb{E}_{\underline{\nu}'}\big[N_a(T)/T\big]\big)\Big) \gtrsim \ln\big(c'\,\Phi_{\mathrm{free}}(T)\,/(T\varepsilon)\big)$

As: $(\mu'_a - \mu^\star)\,\Big(T - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]\Big) \leqslant R_T(\underline{\nu}') \leqslant (M + c/\varepsilon - m)\,\Phi_{\mathrm{free}}(T)$

Picking $\varepsilon \sim \Phi_{\mathrm{free}}(T)/T$: $\big(\Phi_{\mathrm{free}}(T)/T\big)\,\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \gtrsim \mathrm{cst}$

This (technical) proof shows that the main issue is the lack of an upper end M on the range; the lower end m did not change

When M is known, adaptation to m is not so difficult

The DMED strategy by Honda and Takemura [2015] gets the optimal $\ln T/\mathcal{K}_{\inf} < +\infty$ distribution-dependent bound

A variation on the INF strategy by Audibert and Bubeck [2009] gets $\Phi_{\text{free}}(T) = \Theta\big(\sqrt{KT}\big)$

On the contrary, the knowledge of m comes with no advantage:

All impossibility results of this section still hold!