Minimax policies for adversarial and stochastic bandits

Jean-Yves Audibert, Sébastien Bubeck

▶ **To cite this version:**

Jean-Yves Audibert, Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. COLT, Jun 2009, Montreal, Canada. pp.217-226. hal-00834882

HAL Id: hal-00834882
https://enpc.hal.science/hal-00834882

Submitted on 17 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Minimax policies for adversarial and stochastic bandits

Jean-Yves Audibert
Imagine, Université Paris Est
&
Willow, CNRS/ENS/INRIA, Paris, France
audibert@certis.enpc.fr

Sébastien Bubeck
SequeL Project, INRIA Lille
40 avenue Halley,
59650 Villeneuve d'Ascq, France
sebastien.bubeck@inria.fr

Abstract

We fill in a long open gap in the characterization of the minimax rate for the multi-armed bandit problem. Concretely, we remove an extraneous logarithmic factor in the previously known upper bound and propose a new family of randomized algorithms based on an implicit normalization, as well as a new analysis. We also consider the stochastic case, and prove that an appropriate modification of the upper confidence bound policy UCB1 (Auer et al., 2002) achieves the distribution-free optimal rate while still having a distribution-dependent rate logarithmic in the number of plays.

1 Introduction

In the multi-armed bandit problem, at each stage, an agent (or decision maker) chooses one action (or arm), and receives a reward from it. The agent aims at maximizing his rewards. Since he does not know the process generating the rewards, he needs to explore (try) the different actions and yet, exploit (concentrate its draws on) the seemingly most rewarding arms. The multi-armed bandit task has been first considered by Robbins (1952) and was originally motivated by a simplified view of clinical trials (in which an action consists in choosing a treatment and the reward depends on its efficiency on a patient).

To set the notation, let $K \geq 2$ be the number of actions (or arms) and $n \geq K$ be the time horizon. In both stochastic and adversarial K-armed bandit problems, the game between the agent and the environment goes as follows: At each time step $t \in \{1, \dots, n\}$, (i) the agent chooses a probability distribution p_t on a finite set $\{1, \dots, K\}$, (ii) the environment chooses a reward vector $g_t = (g_{1,t}, \dots, g_{K,t}) \in [0,1]^K$, and simultaneously (independently), the agent draws the action (or arm) I_t according to the distribution p_t, (iii) the agent only gets to see his own reward $g_{I_t,t}$. The goal of the decision maker is to maximize his cumulative reward $\sum_{t=1}^n g_{I_t,t}$.

The stochastic K-armed bandit problem is parameterized by a K-tuple of probability distributions $(\nu_1, \dots, \nu_K)$ on $[0,1]$. In this model, the components of the reward vector are i.i.d. realizations of respectively $\nu_1, \dots, \nu_K$. Besides the reward vectors at different times are independent.

The adversarial K-armed bandit problem is more general: The environment is much less constrained as it may choose a reward vector g_t as a function of the past decisions $I_1, \dots, I_{t-1}$ (and possibly of an independent randomization). A simple, but interesting, adversarial environment is obtained by considering deterministic reward vectors: in this case, known as the oblivious deterministic opponent, the environment is just parameterized by the nK real numbers $(g_{i,t})$. Note that in the adversarial environment, past gains have no reason to be representative of future ones in contrast to the stochastic setting in which confidence bounds on the mean reward of the arms can be deduced from the rewards obtained so far.

A policy is a strategy for choosing the drawing probability distribution p_t based on the history formed by the past plays and the associated rewards. So it is a function that maps any history to a probability distribution on $\{1, \dots, K\}$. We define the regret of a policy with respect to the best constant decision as

$$R_n = \max_{i=1,\dots,K} \mathbb{E} \sum_{t=1}^n \left(g_{i,t} - g_{I_t,t}\right).^1$$

To compare to the best constant decision is a reasonable target since it is well-known that (i) there exist randomized policies ensuring that R_n/n tends to zero as n goes to infinity, (ii) this convergence property would not hold if the maximum and the sum would be inverted in the definition of R_n.

The minimax rate of the expected regret is $\inf \sup R_n$, where the infimum is taken over all policies and the supremum over all K-tuples of probability distributions on $[0,1]$ for the stochastic case and over all adversarial environments with rewards in $[0,1]$ for the adversarial case. Auer et al. (1995) proved an upper bound on this quantity in the adversarial case (and thus also in the stochastic case) and a lower bound in the stochastic case (and thus also in the adversarial case). We recall here the results of Auer et al. (1995, 2003).

Theorem 1 *The EXP3 policy described in Fig.1 satisfies*

$$\sup R_n \leq 2.7\sqrt{nK \log K},$$

[1] In the case of an oblivious deterministic opponent this regret is equal to the strong regret: $\mathbb{E} \max_i \sum_{t=1}^n \left(g_{i,t} - g_{I_t,t}\right)$. In the case of an oblivious stochastic opponent (parametrized by nK probability distributions over $[0,1]$), one can prove that the difference between the regret and the strong regret is at most $\sqrt{n \log K}$. Thus at the price of this second order additional term, the bounds in Theorems 4 and 5 hold for the strong regret in the case of an oblivious opponent.

for $\eta = \min\left(0.8\sqrt{\frac{\log K}{nK}}, \frac{1}{K}\right)$*, where the supremum is taken over all adversarial environments. Besides we have*

$$\inf \sup R_n \geq \frac{1}{20}\sqrt{nK},$$

where the supremum is taken over all K*-tuple of probability distributions on* $[0,1]$.

EXP3 (Exploration-Exploitation with Exponential weights):

Parameter: $\eta \in (0, 1/K]$.
Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

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

(1) Draw an arm I_t from the probability distribution p_t.

(2) Compute the estimated gain for each arm: $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t=i}$ and update the estimated cumulative gain: $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.

(3) Compute the new probability distribution over the arms:

$$p_{i,t+1} = (1 - K\eta)\frac{\exp\left(\eta \tilde{G}_{i,t}\right)}{\sum_{k=1}^{K} \exp\left(\eta \tilde{G}_{k,t}\right)} + \eta.$$

Figure 1: A known policy for the adversarial case.

For stochastic environments, the goal is also to adapt to the simplicity of the bandit problem. To be more precise and set up the notation, let μ_i denote the expectation of ν_i. The suboptimality of an arm i is measured by $\Delta_i = \max_{j=1,\ldots,K} \mu_j - \mu_i$. The quantities Δ_i for suboptimal arms (i.e., arms having $\Delta_i > 0$) characterize the simplicity of the task. The larger they are, the easier it is to spot the best decision from a few observations. The upper bounds on the expected regret are usually stated in terms of the parameters Δ_i, and we say that they are distribution-dependent bounds. The UCB1 strategy (Auer et al., 2002), is known to be distribution-dependent optimal since the following upper and lower bounds hold.

Theorem 2 *The UCB1 policy satisfies*

$$R_n \leq 10\left(\sum_{i:\Delta_i>0} \frac{1}{\Delta_i}\right) \log n. \tag{1}$$

For any $\varepsilon > 0$*, for large enough* n*, there is no policy with*

$$R_n \leq \left(\sum_{i:\Delta_i>0} \frac{1}{(2+\varepsilon)\Delta_i}\right) \log n, \tag{2}$$

uniformly for all reward distributions $\nu_1, \ldots, \nu_K$.

The upper bound is a simple variant of (Auer et al., 2002, Theorem 1) while the lower bound comes from (Lai and Robbins, 1985, Theorem 1) applied to the parametric family of Bernoulli distributions. An easy modification of the proof of (1) gives

$$R_n \leq \max_{t_i \geq 0, \sum_i t_i = n} \sum_{i:\Delta_i>0} \min\left(\frac{10}{\Delta_i}\log n\,,\; t_i\Delta_i\right),$$

INF (Implicitly Normalized Forecaster):

Parameter: function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ satisfying (3).
Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

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

(1) Draw an arm I_t from the probability distribution p_t.

(2) Compute the estimated gain for each arm: $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t=i}$ and update the estimated cumulative gain: $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.

(3) Compute the normalization constant $C_t = C(\tilde{G}_t)$ where $\tilde{G}_t = (\tilde{G}_{1,t}, \ldots, \tilde{G}_{K,t})$.

(4) Compute the new probability distribution $p_{t+1} = (p_{1,t+1}, \ldots, p_{K,t+1})$ where

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t).$$

Figure 2: The proposed policy for the adversarial case.

which in the worst case (i.e., $\Delta_1 = 0$ and $\Delta_2 = \cdots = \Delta_K = \sqrt{10K(\log n)/n}$) is equal to $\sqrt{10n(K-1)\log n}$. This means that even in the stochastic bandit problem, there is a logarithmic gap between the lower and upper bounds.

Outline and contributions. Section 2 presents a new family of randomized policies to address the adversarial multi-armed bandit problem, and proves that many of them satisfy a regret of order $\sqrt{nK}$. This bridges the logarithmic gap between the known upper and lower bounds presented in Theorem 1.

Section 3 defines a policy achieving the best distribution-free regret $\sqrt{nK}$ for stochastic bandits as well as a distribution-dependent regret of order $K \log \frac{n}{K}$.

2 The adversarial case: $\sqrt{nK}$ regret

We start by defining a new class of randomized policies. Let us consider a function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ such that

$$\begin{array}{l} \psi \text{ increasing and continuously differentiable,} \\ \psi'/\psi \text{ nondecreasing,} \\ \lim_{u\to-\infty} \psi(u) < 1/K, \text{ and } \lim_{u\to 0} \psi(u) \geq 1. \end{array} \tag{3}$$

Lemma 3 *There exists a continuously differentiable function* $C : \mathbb{R}_+^K \to \mathbb{R}$ *satisfying for any* $x = (x_1, \ldots, x_K) \in \mathbb{R}_+^K$,

$$\max_{i=1,\ldots,K} x_i < C(x) \leq \max_{i=1,\ldots,K} x_i - \psi^{-1}(1/K)\,, \tag{4}$$

and

$$\sum_{i=1}^{K} \psi(x_i - C(x)) = 1. \tag{5}$$

Proof: See Appendix A. ■

The implicitly normalized forecaster (INF) is defined in Fig.2. Equality (5) makes the fourth step in Fig.2 legitimate.

From (4), $C(\tilde{G}_t)$ is roughly equal to $\max_{i=1,\dots,K} \tilde{G}_{i,t}$. This means that INF chooses the probability assigned to arm i as a function of the (estimated) regret. Note that, in spirit, it is similar to the traditional weighted average forecaster, see e.g. Section 2.1 of Cesa-Bianchi and Lugosi (2006), where the probabilities are proportional to a function of the difference between the (estimated) cumulative reward of arm i and the cumulative reward of the policy, which should be, for a well-performing policy, of order $C(\tilde{G}_t)$. Our main result is the following.

Theorem 4 *For any real $q > 1$, the Implicitly Normalized Forecaster with $\psi(x) = \frac{1}{K}\left(\frac{9\sqrt{qnK}}{-x}\right)^q + \frac{q^{q/(2q-2)}}{\sqrt{nK}}$ satisfies*

$$\sup R_n \leq \frac{37}{1-1/q}\sqrt{qnK}, \tag{6}$$

where the supremum is taken over all adversarial environments with rewards in $[0,1]$.

Proof: We put here the main lines of the proof. We need to lower bound $\sum_{t=1}^n g_{I_t,t} = \sum_{t=1}^n \sum_{i=1}^K p_{i,t}(\tilde{G}_{i,t} - \tilde{G}_{i,t-1})$. By an Abel transform, we come down to upper bounding $\sum_{t=1}^{n-1}\sum_{i=1}^K \tilde{G}_{i,t}(p_{i,t+1} - p_{i,t})$, which is equal to

$$\sum_{i=1}^K \sum_{t=1}^{n-1} \psi^{-1}(p_{i,t+1})(p_{i,t+1} - p_{i,t}).$$

Note that this striking last equality is closely linked to our specific class of randomized algorithms. Then we use a Taylor-Lagrange expansion, which makes appear

$$\sum_{t=1}^{n-1} \int_{p_{i,t}}^{p_{i,t+1}} \psi^{-1}(u)du = \int_{1/K}^{p_{i,n}} \psi^{-1}(u)du.$$

The difficulty is then to control the residual terms. See Appendix B. ■

For $q = 3$, inequality (6) gives $\sup R_n \leq 100\sqrt{nK}$. In fact, much better constants can be obtained by proper tuning. In Theorem 4, we take $\psi(x) = \alpha(-x)^{-q} + \beta$ for appropriate $\alpha > 0$ and $\beta > 0$. The role of β is to keep away from zero the probabilities of seemingly low rewarding arms. EXP3 contains a similar term. For low values of q, that is for $1 < q \leq 2$, it can be established (proof omitted here) that this additional term is not necessary to achieve a $\sqrt{nK}$ regret. For instance, for $\psi(x) = \frac{1}{K}\left(\frac{2.5\sqrt{nK}}{-x}\right)^{1.5}$, we have shown that $\sup R_n \leq 15\sqrt{nK}$.

Remark 1 For $\psi(u) = \eta + \exp(\eta u)$ with $\eta \in (0, 1/K]$, we have $\exp(-\eta C(x)) = (1 - K\eta)/\sum_{j=1}^K \exp(\eta x_j)$, so that INF reduces to EXP3. Except for this particular choice of ψ, the (normalizing) function C has usually no closed form expression, hence the name of the policy. However this does not lead to a major computational issue since, in the interval given by (4), $C(x)$ is the unique solution of $\phi(c) = 1$, where $\phi : c \mapsto \sum_{i=1}^K \psi(x_i - c)$ is a decreasing function.

Remark 2 The policies in the theorem, as the other ones presented in this paper, are not "anytime" since their implementation requires the knowledge of the horizon n. By using the doubling trick, see e.g. Section 2.3 of Cesa-Bianchi and Lugosi (2006), one can make them anytime.

3 The stochastic case

By considering the deterministic case when the rewards are $g_{i,t} = 1$ if $i = 1$ and $g_{i,t} = 0$ otherwise, it can be proved that the INF policies considered in Theorem 4, as well as EXP3, have an expected regret lower bounded by $\sqrt{nK}$. In this simple setting, and more generally in most of the stochastic multi-armed bandit problems, one would like to suffer a much smaller regret.

We recall that the suboptimality of an arm i is measured by $\Delta_i = \max_{j=1,\dots,K} \mu_j - \mu_i$. Our second contribution is to provide a strategy achieving a $\sqrt{nK}$ regret in the worst case, and a much smaller regret as soon as the Δ_i of the suboptimal arms are much larger than $\sqrt{K/n}$.

Let $\widehat{X}_{i,s}$ be the empirical mean of arm i after s draws of this arm. Let $T_i(t)$ denote the number of times we have drawn arm i on the first t rounds. In this section, we propose a policy inspired by the UCB1 policy (Auer et al., 2002), where each arm has an index measuring its performance, and at each round, we choose the arm having the highest index (see Fig.3). The index of an arm that has been drawn more than n/K times is simply the empirical mean of the rewards obtained from the arm. For the other arms, their index is an upper confidence bound on their mean reward, which, from Hoeffding's inequality, holds with high probability.

MOSS (Minimax Optimal Strategy in the Stochastic case):

For an arm i, define its index $B_{i,s}$ by

$$B_{i,s} = \widehat{X}_{i,s} + \sqrt{\frac{\max\left(\log(\frac{n}{Ks}), 0\right)}{s}}\,.$$

for $s \geq 1$ and $B_{i,0} = +\infty$.

At time t, draw an arm maximizing $B_{i,T_i(t-1)}$.

Figure 3: The proposed policy for the stochastic case.

Theorem 5 *MOSS satisfies*

$$\sup R_n \leq 49\sqrt{nK}, \tag{7}$$

where the supremum is taken over all K-tuple of probability distributions on $[0,1]$.

Proof: Here are the main steps. Without loss of generality consider that $\mu_1 \geq \ldots \geq \mu_K$, hence $\Delta_i = \mu_1 - \mu_i$. By Wald's identity, we have $R_n = \mathbb{E}\sum_i \Delta_i T_i(n)$. Tightly upper bounding R_n is difficult because of the heavy dependence between the random variables $T_i(n)$. To decouple the arms, we introduce the key thresholds $z_i = \mu_1 - \frac{\Delta_i}{2}$, and the r.v. $Z = \min_{1\leq s\leq n} B_{1,s}$ and $\tau_i = \min\{t : B_{i,t} < z_i\}$, and essentially prove that

$$R_n \leq \sum_i \Big\{\Delta_i \mathbb{E}\tau_i + n\mathbb{P}(Z < z_i)(\Delta_i - \Delta_{i-1})\Big\}.$$

The expectations $\mathbb{E}\tau_i = \sum_{\ell=0}^{+\infty} \mathbb{P}(\tau_i > \ell)$ are then bounded by using Hoeffding's inequality, and the probabilities

$\mathbb{P}(Z < z_i)$ are carefully upper bounded by using maximal inequalities and peeling arguments. See Appendix C. ■

Remark 3 A careful tuning of the constants in front and inside the logarithmic term of $B_{i,s}$ and of the thresholds used in the proof leads to $\sup R_n \leq 5.7\sqrt{nK}$. However, it makes the proof more intricate. So we will only prove (7).

The distribution-dependent upper bound for MOSS is the following.

Theorem 6 *MOSS satisfies*

$$R_n \leq 23K \sum_{i:\Delta_i>0} \frac{\max\left(\log\left(\frac{n\Delta_i^2}{K}\right), 1\right)}{\Delta_i}. \qquad (8)$$

Proof: It follows the same route as the previous proof. See Appendix D. ■

Theorem 6 contains a K factor, which does not appear in the logarithmic bound of UCB1 given in Theorem 2. This means that for fixed $\Delta_1, \ldots, \Delta_K$ and n going to infinity, UCB1 will perform better than MOSS (if the bounds are representative of the behaviour of the algorithms, which, we believe, is the case). Note that, despite the K factor, the bound has the right order in the critical case when $\Delta_1 = 0$, $\Delta_2 = \cdots = \Delta_K = \gamma\sqrt{K/n}$ for $\gamma \geq 3$. Indeed, in that case, MOSS satisfies $R_n \leq \frac{46\sqrt{nK}\log\gamma}{\gamma} \leq 17\sqrt{nK}$, whereas, for UCB1 and $\gamma = \sqrt{\log n}$, the best known bound contains a $\sqrt{\log n}$ factor.

To conclude, the following table summarizes the regret upper bounds of the policies discussed in this paper. The bounds for MOSS and UCB1 hold in the stochastic setting only. The bounds for INF and EXP3 hold in the adversarial setting (and thus also in the stochastic case).

UCB1	$\min\left(\sqrt{nK\log n}\ ,\ \sum_{i:\Delta_i>0} \frac{\log n}{\Delta_i}\right)$
MOSS	$\min\left(\sqrt{nK}\ ,\ \sum_{i:\Delta_i>0} \frac{K\log\left(2+n\Delta_i^2/K\right)}{\Delta_i}\right)$
EXP3	$\sqrt{nK\log K}$
INF	$\sqrt{nK}$

Table 1: regret upper bounds (up to a numerical constant factor) for different policies in the multi-armed bandit problem.

References

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. IEEE Computer Society Press, 1995.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

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

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58: 527–535, 1952.

A Proof of Lemma 3

Consider a fixed $x = (x_1, \ldots, x_K)$. The decreasing function $\phi : c \mapsto \sum_{i=1}^K \psi(x_i - c)$ satisfies

$$\lim_{c \to \max_{i=1,\ldots,K} x_i} \phi(c) > 1 \qquad \text{and} \qquad \lim_{c\to+\infty} \phi(c) < 1.$$

From the mean value theorem, there is a unique $C(x)$ satisfying $\phi(C(x)) = 1$. From the implicit function theorem, the mapping $x \mapsto C(x)$ is continuously differentiable.

B Proof of Theorem 4

We start with three lemmas which study INF in general for functions ψ such that (3) holds. The proof of Theorem 4 follows by applying the last lemma to the particular class of ψ we use. Let us set $\tilde{G}_0 = 0 \in \mathbb{R}_+^K$ and $C_0 = C(\tilde{G}_0)$.

Lemma 7 *Let*

$$A_T = \sup_{t\in\{1,\ldots,T-1\}} \frac{\psi'(\tilde{G}_{I_t,t} - C_t)}{\psi'(\tilde{G}_{I_t,t-1} - C_{t-1})}$$

and

$$B_T = \sup_{t\in\{1,\ldots,T-1\}, i\in\{1,\ldots,K\}} \frac{\psi'(\tilde{G}_{i,t-1} - C_{t-1})}{\psi'(\tilde{G}_{i,t} - C_t)}.$$

For any $T \geq 2$, INF satisfies

$$\begin{aligned} C_{T-1} - \sum_{t=1}^T g_{I_t,t} &\leq \sum_{i=1}^K p_{i,T}(-\psi^{-1})(p_{i,T}) \\ &+ \sum_{i=1}^K \int_{p_{i,T}}^{1/K} (-\psi^{-1})(u)du \\ &+ \frac{A_T(A_T+B_T)}{2} \sum_{t=1}^{T-1} \frac{\psi'(\tilde{G}_{I_t,t-1} - C_{t-1})}{\psi(\tilde{G}_{I_t,t-1} - C_{t-1})^2}. \end{aligned}$$

Proof: The proof is divided into five steps.

First step: Rewriting $\sum_{t=1}^T g_{I_t,t}$.

We start with the following equalities:

$$\sum_{t=1}^{T} g_{I_t,t} = \sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}\tilde{g}_{i,t} = \sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}(\tilde{G}_{i,t} - \tilde{G}_{i,t-1})$$
$$= \sum_{t=1}^{T-1}\sum_{i=1}^{K}(p_{i,t} - p_{i,t+1})\tilde{G}_{i,t} + \sum_{i=1}^{K} p_{i,T}\tilde{G}_{i,T},$$

where the last step comes from an Abel transformation. Now,

$$\sum_{i=1}^{K} p_{i,T}\tilde{G}_{i,T} \geq \sum_{i=1}^{K} p_{i,T}\tilde{G}_{i,T-1} = C_{T-1} + \sum_{i=1}^{K} p_{i,T}\psi^{-1}(p_{i,T}).$$

Hence, so far, we have proved

$$C_{T-1} - \sum_{t=1}^{T} g_{I_t,t} \leq \sum_{i=1}^{K} p_{i,T}(-\psi^{-1})(p_{i,T}) + \sum_{t=1}^{T-1}\sum_{i=1}^{K}(p_{i,t+1} - p_{i,t})\tilde{G}_{i,t}.$$

Let us rewrite the last sum by using $\tilde{G}_{i,t} = \psi^{-1}(p_{i,t+1}) + C_t$ and $\sum_{i=1}^{K}(p_{i,t+1} - p_{i,t})C_t = 0$. We obtain

$$C_{T-1} - \sum_{t=1}^{T} g_{I_t,t} \leq \sum_{i=1}^{K} p_{i,T}(-\psi^{-1})(p_{i,T}) + \sum_{t=1}^{T-1}\sum_{i=1}^{K}(p_{i,t+1} - p_{i,t})\psi^{-1}(p_{i,t+1}).$$

Second step: A Taylor-Lagrange expansion.

For $x \in [0,1]$ we define $f(x) = \int_0^x \psi^{-1}(u)du$. Remark that $f'(x) = \psi^{-1}(x)$ and $f''(x) = 1/\psi'(\psi^{-1}(x))$. Then by the Taylor-Lagrange formula, we know that for any i, there exists $p'_{i,t+1} \in [p_{i,t}, p_{i,t+1}]$ (with the convention $[a,b] = [b,a]$ when $a > b$) such that

$$f(p_{i,t}) = f(p_{i,t+1}) + (p_{i,t} - p_{i,t+1})f'(p_{i,t+1}) + \frac{(p_{i,t} - p_{i,t+1})^2}{2} f''(p'_{i,t+1}),$$

or, in other words:

$$(p_{i,t+1} - p_{i,t})\psi^{-1}(p_{i,t+1}) = \int_{p_{i,t+1}}^{p_{i,t}} (-\psi^{-1})(u)du + \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'(\psi^{-1}(p'_{i,t+1}))}.$$

Now remark that by summing over t we get the term

$$\int_{p_{i,T}}^{1/K} (-\psi^{-1})(u)du$$

which appears in the statement of the lemma. Moreover, using that $\psi' \circ \psi^{-1}$ is increasing (since ψ is increasing and convex from the assumption on ψ'/ψ) we have

$$1/\psi'\left\{\psi^{-1}(p'_{i,t+1})\right\} \leq 1/\psi'\left\{\psi^{-1}\left(\min(p_{i,t}, p_{i,t+1})\right)\right\}.$$

To finish the proof, we have to upper bound

$$\sum_{t=1}^{T-1}\sum_{i=1}^{K} \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'\left\{\psi^{-1}(\min(p_{i,t}, p_{i,t+1}))\right\}}. \tag{9}$$

Third step: Preliminary remarks to bound $(p_{i,t+1} - p_{i,t})^2$.

It is now convenient to consider the functions f_i and h_i defined for any $x \in \mathbb{R}_+^K$ by

$$f_i(x) = \psi(x_i - C(x)) \quad \text{and} \quad h_i(x) = \psi'(x_i - C(x)).$$

We are going to bound $p_{i,t+1} - p_{i,t} = f_i(\tilde{G}_t) - f_i(\tilde{G}_{t-1})$, and consequently (9), by using the mean value theorem. This step gathers preliminary results before applying the mean value theorem. First, we have

$$\frac{\partial f_i}{\partial x_j}(x) = \left(\mathbb{1}_{i=j} - \frac{\partial C}{\partial x_j}(x)\right) h_i(x).$$

Now, by definition of C, we have $\sum_{k=1}^{K} f_k(x) = 1$ and thus $\sum_{k=1}^{K} \frac{\partial f_k}{\partial x_j}(x) = 0$, which implies

$$\frac{\partial C}{\partial x_j}(x) = \frac{h_j(x)}{\sum_{k=1}^{K} h_k(x)} \geq 0, \tag{10}$$

and

$$\frac{\partial f_i}{\partial x_j}(x) = \left(\mathbb{1}_{i=j} - \frac{h_j(x)}{\sum_{k=1}^{K} h_k(x)}\right) h_i(x). \tag{11}$$

Let $x = (x_1, \ldots, x_K)$ and $x^j = (x_1, \ldots, x_j + \zeta, \ldots, x_K)$ where $\zeta \in \mathbb{R}_+$. From (10), we have for $i \neq j$,

$$x_i^j - C(x^j) \leq x_i - C(x).$$

Now since ψ and ψ' are increasing, we have for $i \neq j$

$$f_i(x^j) \leq f_i(x) \quad \text{and} \quad h_i(x^j) \leq h_i(x). \tag{12}$$

Moreover remark that

$$1 = \sum_{i=1}^{K} f_i(x^j) \leq f_j(x^j) + \sum_{i=1, i\neq j}^{K} f_i(x) = f_j(x^j) - f_j(x) + 1$$

and thus $f_j(x) \leq f_j(x^j)$ which in turn implies $x_j - C(x) \leq x_j^j - C(x^j)$ (since ψ^{-1} is increasing). Hence we have

$$f_j(x^j) \geq f_j(x) \quad \text{and} \quad h_j(x^j) \geq h_j(x). \tag{13}$$

Fourth step: Upper bounding $(p_{i,t+1} - p_{i,t})^2$.

Recall that $p_{i,t+1} - p_{i,t} = f_i(\tilde{G}_t) - f_i(\tilde{G}_{t-1})$ and that $\tilde{G}_t$ and $\tilde{G}_{t-1}$ only differs by $g_{I_t,t}/p_{I_t,t}$ at their I_t-th coordinate. Thus there exists $\zeta \in (0, g_{I_t,t}/p_{I_t,t})$ such that the following is true with $\tilde{G}'_t = (\tilde{G}_{1,t-1}, \ldots, \tilde{G}_{I_t,t-1} + \zeta, \ldots, \tilde{G}_{K,t-1})$,

$$p_{i,t+1} - p_{i,t} = f_i(\tilde{G}_t) - f_i(\tilde{G}_{t-1}) = \frac{g_{I_t,t}}{p_{I_t,t}} \frac{\partial f_i}{\partial x_{I_t}}(\tilde{G}'_t).$$

From (11) and since the rewards are in $[0,1]$, we obtain

$$\begin{aligned}
&(p_{i,t+1} - p_{i,t})^2 \\
&\leq \frac{1}{p_{I_t,t}^2}\left(\mathbb{1}_{i=I_t} - \frac{h_{I_t}(\tilde{G}'_t)}{\sum_{k=1}^{K} h_k(\tilde{G}'_t)}\right)^2 h_i(\tilde{G}'_t)^2 \\
&\leq \frac{1}{p_{I_t,t}^2} h_{I_t}(\tilde{G}'_t)^2 \mathbb{1}_{i=I_t} + \frac{1}{p_{I_t,t}^2} \frac{h_i(\tilde{G}'_t)^2 h_{I_t}(\tilde{G}'_t)^2}{\left(\sum_{k=1}^{K} h_k(\tilde{G}'_t)\right)^2} \mathbb{1}_{i\neq I_t} \\
&\leq \frac{h_{I_t}(\tilde{G}'_t)}{p_{I_t,t}^2}\left(h_i(\tilde{G}'_t)\mathbb{1}_{i=I_t} + \frac{h_i(\tilde{G}'_t)^2}{\sum_{k=1}^{K} h_k(\tilde{G}'_t)} \mathbb{1}_{i\neq I_t}\right) \\
&\leq \frac{h_{I_t}(\tilde{G}_t)}{p_{I_t,t}^2}\left(h_i(\tilde{G}_t)\mathbb{1}_{i=I_t} + \frac{h_i(\tilde{G}_{t-1}) h_i(\tilde{G}'_t)}{\sum_{k=1}^{K} h_k(\tilde{G}'_t)} \mathbb{1}_{i\neq I_t}\right),
\end{aligned}$$

where the last step comes from (12) and (13).

Fifth step: Bounding $\sum_{t=1}^{T-1}\sum_{i=1}^{K}\frac{(p_{i,t}-p_{i,t+1})^2}{2\psi'\{\psi^{-1}(\min(p_{i,t},p_{i,t+1}))\}}$.

Notice that (12) and (13) imply that

$$p_{I_t,t+1}\geq p_{I_t,t}\quad\text{and}\quad\text{for } i\neq I_t,\ p_{i,t+1}\leq p_{i,t}. \tag{14}$$

Thus we have

$$\begin{aligned}&\psi'\left[\psi^{-1}(\min(p_{i,t},p_{i,t+1}))\right]\\&\quad=\psi'\left\{\psi^{-1}(p_{i,t})\right\}\mathbb{1}_{i=I_t}+\psi'\left\{\psi^{-1}(p_{i,t+1})\right\}\mathbb{1}_{i\neq I_t}\\&\quad=h_i(\tilde{G}_{t-1})\mathbb{1}_{i=I_t}+h_i(\tilde{G}_t)\mathbb{1}_{i\neq I_t},\end{aligned}$$

hence

$$\begin{aligned}\frac{(p_{i,t}-p_{i,t+1})^2}{\psi'\left\{\psi^{-1}(\min(p_{i,t},p_{i,t+1}))\right\}}&\leq\frac{h_{I_t}(\tilde{G}_t)}{p_{I_t,t}^2}\left(\frac{h_i(\tilde{G}_t)}{h_i(\tilde{G}_{t-1})}\mathbb{1}_{i=I_t}\right.\\&\qquad\left.+\frac{h_i(\tilde{G}_{t-1})h_i(\tilde{G}'_t)}{h_i(\tilde{G}_t)\sum_{k=1}^{K}h_k(\tilde{G}'_t)}\mathbb{1}_{i\neq I_t}\right)\\&\leq\frac{h_{I_t}(\tilde{G}_t)}{p_{I_t,t}^2}\left(A_T\mathbb{1}_{i=I_t}+B_T\frac{h_i(\tilde{G}'_t)}{\sum_{k=1}^{K}h_k(\tilde{G}'_t)}\mathbb{1}_{i\neq I_t}\right),\end{aligned}$$

where we have used the definitions of A_T and B_T for the last inequality. It is now clear that by summing over i we get

$$\sum_{i=1}^{K}\frac{(p_{i,t}-p_{i,t+1})^2}{\psi'\left\{\psi^{-1}(\min(p_{i,t},p_{i,t+1}))\right\}}\leq(A_T+B_T)\frac{h_{I_t}(\tilde{G}_t)}{p_{I_t,t}^2}.$$

This last term is equal to

$$(A_T+B_T)\frac{h_{I_t}(\tilde{G}_t)}{f_{I_t}(\tilde{G}_{t-1})^2}\leq A_T(A_T+B_T)\frac{h_{I_t}(\tilde{G}_{t-1})}{f_{I_t}(\tilde{G}_{t-1})^2}$$

which concludes the proof. ■

In essence, Lemma 7 gives the result we want to apply for $\psi(x)=\frac{1}{K}\left(\frac{9\sqrt{qnK}}{-x}\right)^q+\frac{q^{q/(2q-2)}}{\sqrt{nK}}$. Unfortunately, we still need to answer positively to questions of the type: is the probability of drawing arm i at time t of the same order as the one at time $t+1$ (or more technically speaking, do A_T and B_T are bounded by constants)? While this seems an obvious point, it requires to our knowledge the unfortunately complicated and recursive arguments that follows.

Lemma 8 *Let $T\geq 2$. Assume that there exist $a>1, b>1$, $c>0$ such that $C_{T-1}\leq c$, and for any $x\in[-c,0)$,*

$$\psi'\left(x+1/\psi(x)\right)\leq a\psi'(x) \tag{15}$$

and

$$\psi'\left(x-ab/\psi(x)\right)\geq\frac{1}{b}\psi'(x). \tag{16}$$

Then we have $A_T\leq a$ and $B_T\leq b$.

Proof: Let $t\in\{1,\dots,T-1\}$. The quantities $\tilde{G}_{i,s}$ are nonnegative and nondecreasing as a function of the time s, hence from (10), $s\mapsto C_s$ is nondecreasing, and $C_{T-1}\leq c$ implies $\tilde{G}_{i,t-1}-C_{t-1}\geq -c$. Since the function ψ' is increasing (because ψ'/ψ is nondecreasing) and from (15), we have

$$\begin{aligned}&\psi'(\tilde{G}_{I_t,t}-C_t)\\&\quad=\psi'\left[\tilde{G}_{I_t,t-1}-C_t+g_{I_t,t}/\psi(\tilde{G}_{I_t,t-1}-C_{t-1})\right]\\&\quad\leq\psi'\left[\tilde{G}_{I_t,t-1}-C_{t-1}+1/\psi(\tilde{G}_{I_t,t-1}-C_{t-1})\right]\\&\quad\leq a\psi'(\tilde{G}_{I_t,t-1}-C_{t-1}).\end{aligned}$$

Thus we have proved $A_T\leq a$.

For the second inequality, from (14), we have $p_{I_t,t+1}\geq p_{I_t,t}$, and consequently $\psi(\tilde{G}_{I_t,t}-C_t)\geq\psi(\tilde{G}_{I_t,t-1}-C_{t-1})$. Let $\delta_t=C_t-C_{t-1}$. Since ψ is increasing, this implies $\tilde{g}_{I_t,t}-C_t\geq -C_{t-1}$, hence

$$\delta_t\leq\tilde{g}_{I_t,t}\leq 1/p_{I_t,t}=1/\psi(\tilde{G}_{I_t,t-1}-C_{t-1}). \tag{17}$$

Unfortunately, we need a more precise upper bound on δ_t because $1/p_{I_t,t}$ is large when the selected arm has low probability. We now prove that in fact, for all $i\in\{1,\dots,K\}$:

$$\delta_t\leq\frac{ab}{\psi(\tilde{G}_{i,t-1}-C_{t-1})}.$$

From (17), this holds for any i such that $\tilde{G}_{i,t-1}\leq\tilde{G}_{I_t,t-1}$. Now let i be such that $\tilde{G}_{i,t-1}>\tilde{G}_{I_t,t-1}$. From (14), we have

$$\begin{aligned}&p_{i,t+1}+p_{I_t,t+1}\\&\quad=1-\sum_{j\notin\{i,I_t\}}p_{j,t+1}\geq 1-\sum_{j\notin\{i,I_t\}}p_{j,t}=p_{i,t}+p_{I_t,t},\end{aligned}$$

which, by using $\tilde{G}_{i,t-1}=\tilde{G}_{i,t}$, implies

$$\begin{aligned}&\psi(\tilde{G}_{i,t-1}-C_{t-1})-\psi(\tilde{G}_{i,t-1}-C_t)\\&\quad\leq\ \psi(\tilde{G}_{I_t,t}-C_t)-\psi(\tilde{G}_{I_t,t-1}-C_{t-1}),\end{aligned} \tag{18}$$

Now, by using $\tilde{G}_{I_t,t}-C_t\geq\tilde{G}_{I_t,t-1}-C_{t-1}$ (which is a consequence of $p_{I_t,t+1}\geq p_{I_t,t}$), $A_T\leq a$, and since the functions ψ and ψ'/ψ (and therefore ψ') are nondecreasing, the mean value theorem gives:

$$\begin{aligned}&\psi(\tilde{G}_{I_t,t}-C_t)-\psi(\tilde{G}_{I_t,t-1}-C_{t-1})\\&\quad=\psi(\tilde{G}_{I_t,t-1}-C_{t-1}+\tilde{g}_{I_t,t}-\delta_t)-\psi(\tilde{G}_{I_t,t-1}-C_{t-1})\\&\quad\leq(\tilde{g}_{I_t,t}-\delta_t)\psi'(\tilde{G}_{I_t,t}-C_t)\leq\frac{\psi'(\tilde{G}_{I_t,t}-C_t)}{\psi(\tilde{G}_{I_t,t-1}-C_{t-1})}\\&\quad\leq a\frac{\psi'(\tilde{G}_{I_t,t-1}-C_{t-1})}{\psi(\tilde{G}_{I_t,t-1}-C_{t-1})}\leq a\frac{\psi'(\tilde{G}_{i,t-1}-C_{t-1})}{\psi(\tilde{G}_{i,t-1}-C_{t-1})}.\end{aligned} \tag{19}$$

On the other hand, from the mean value theorem and (16), we also have

$$\begin{aligned}&\psi(\tilde{G}_{i,t-1}-C_{t-1})\\&\quad-\psi\left(\tilde{G}_{i,t-1}-C_{t-1}-ab/\psi(\tilde{G}_{i,t-1}-C_{t-1})\right)\\&\quad\geq ab\frac{\psi'\left(\tilde{G}_{i,t-1}-C_{t-1}-ab/\psi(\tilde{G}_{i,t-1}-C_{t-1})\right)}{\psi(\tilde{G}_{i,t-1}-C_{t-1})}\\&\quad\geq a\frac{\psi'(\tilde{G}_{i,t-1}-C_{t-1})}{\psi(\tilde{G}_{i,t-1}-C_{t-1})},\end{aligned} \tag{20}$$

where we have used $\tilde{G}_{i,t-1} - C_{t-1} \geq -c$.

By combining inequalities (18), (19) and (20), we get

$$\begin{aligned}
&\psi\left(\tilde{G}_{i,t-1} - C_{t-1} - ab/\psi(\tilde{G}_{i,t-1} - C_{t-1})\right) \\
&\leq \psi(\tilde{G}_{i,t-1} - C_{t-1}) - a\frac{\psi'(\tilde{G}_{i,t-1} - C_{t-1})}{\psi(\tilde{G}_{i,t-1} - C_{t-1})} \\
&\leq \psi(\tilde{G}_{i,t-1} - C_t) = \psi(\tilde{G}_{i,t-1} - C_{t-1} - \delta_t)
\end{aligned}$$

and thus $\delta_t \leq ab/\psi(\tilde{G}_{i,t-1} - C_{t-1})$. We can now conclude the proof with the same method as for A_T. For $i \neq I_t$:

$$\begin{aligned}
\psi'(\tilde{G}_{i,t} - C_t) &= \psi'(\tilde{G}_{i,t-1} - C_{t-1} - \delta_t) \\
&\geq \psi'\left(\tilde{G}_{i,t-1} - C_{t-1} - ab/\psi(\tilde{G}_{i,t-1} - C_{t-1})\right) \\
&\geq \frac{1}{b}\psi'(\tilde{G}_{i,t-1} - C_{t-1})
\end{aligned}$$

where we have used (16) for the last inequality. ■

Lemma 9 *Assume that there exist $a > 1, b > 1, c > 0$, $c_1 > 0, c_2 > 0$ such that for any $x \in [-cn, 0)$ and any $q \in \mathbb{R}^K$ satisfying $\sum_{i=1}^K q_i = 1, q_i \geq \psi(-cn)$, we have:*

$$c \geq 1 + \frac{a(a+b)}{2}c_2 + c_1\sqrt{\frac{K}{n}} + ab\frac{K}{n}, \tag{21}$$

$$\psi(-cn) \leq 1/K, \tag{22}$$

$$-\sum_{i=1}^K q_i\psi^{-1}(q_i) - \int_{q_i}^{1/K} \psi^{-1}(u)du \leq c_1\sqrt{nK}, \tag{23}$$

$$\psi'(x)/\psi(x)^2 \leq c_2, \tag{24}$$

$$\psi'\left(x + 1/\psi(x)\right) \leq a\psi'(x), \tag{25}$$

$$\psi'\left(x - ab/\psi(x)\right) \geq \frac{1}{b}\psi'(x). \tag{26}$$

Then INF satisfies:

$$R_n \leq 1 + c_1\sqrt{nK} + \frac{a(a+b)}{2}\mathbb{E}\sum_{t=1}^{n-1}\sum_{i=1}^K \frac{\psi'(\tilde{G}_{i,t-1} - C_{t-1})}{\psi(\tilde{G}_{i,t-1} - C_{t-1})}.$$

Proof: Let $T = \min\{t \in \mathbb{N} : C_t > cn\}$. Since $p_{1,1} = \psi(-C_0) = 1/K$ and $\psi(-cn) \leq 1/K$ from (22), we know that $T \geq 1$. Since we have $C_{T-1} \leq cn$ and from the assumptions (25) and (26), we may apply Lemma 8 and obtain $A_T \leq a$ and $B_T \leq b$. So, from inequalities (23) and (24), and since the gains are bounded by 1, Lemma 7 implies:

$$C_{T-1} \leq T + c_1\sqrt{nK} + \frac{a(a+b)}{2}c_2T.$$

Now recall from the proof of Lemma 8 that we have $\delta_T = C_T - C_{T-1} \leq ab/p_{i,T}$ for any i and thus $C_T \leq C_{T-1} + abK$. Thus we have

$$C_T \leq abK + T + c_1\sqrt{nK} + \frac{a(a+b)}{2}c_2T.$$

If $T \leq n$, we obtain by using (21)

$$\begin{aligned}
C_T &\leq abK + n + c_1\sqrt{nK} + \frac{a(a+b)}{2}c_2n \\
&\leq \left(1 + \frac{a(a+b)}{2}c_2 + c_1\sqrt{\frac{K}{n}} + \frac{abK}{n}\right)n \leq cn,
\end{aligned}$$

which is impossible by definition of T. Thus, whatever arms are chosen, we always have $T > n$, and we can apply Lemmas 7 and 8 at time n and obtain

$$C_{n-1} - \sum_{t=1}^n g_{I_t,t} \leq c_1\sqrt{nK} + \frac{a(a+b)}{2}\sum_{t=1}^{n-1}\mathcal{S}_t,$$

with

$$\mathcal{S}_t = \frac{\psi'(\tilde{G}_{I_t,t-1} - C_{t-1})}{\psi(\tilde{G}_{I_t,t-1} - C_{t-1})^2}.$$

Now from (4), we have $C_{n-1} \geq \max_{i=1,\dots,K}\tilde{G}_{i,n-1}$, hence

$$\mathbb{E}C_{n-1} \geq \max_i \mathbb{E}\tilde{G}_{i,n-1} = \max_i \mathbb{E}\sum_{t=1}^{n-1} g_{i,t},$$

where we have used that the estimated gains are unbiased estimates of the true gains. Thus we have

$$R_n \leq 1 + c_1\sqrt{nK} + \frac{a(a+b)}{2}\sum_{t=1}^{n-1}\mathbb{E}\mathcal{S}_t.$$

We can now write

$$\begin{aligned}
\mathbb{E}\mathcal{S}_t &= \mathbb{E}\,\mathbb{E}_{I_t\sim p_t}\frac{\psi'(\tilde{G}_{I_t,t-1} - C_{t-1})}{p_{I_t,t}\psi(\tilde{G}_{I_t,t-1} - C_{t-1})} \\
&= \mathbb{E}\sum_{i=1}^K \frac{\psi'(\tilde{G}_{i,t-1} - C_{t-1})}{\psi(\tilde{G}_{i,t-1} - C_{t-1})},
\end{aligned}$$

which concludes the proof of the lemma. ■

We now use the particular choice of the function ψ given in Theorem 4. Let $\alpha = (9\sqrt{qnK})^q/K$ and $\beta = q^{\frac{q}{2(q-1)}}/\sqrt{nK}$. Thus we have $\psi(x) = \frac{\alpha}{(-x)^q} + \beta$. We can assume that

$$\frac{37}{1-1/q}\sqrt{qnK} \leq n, \tag{27}$$

since, otherwise, Theorem 4 is trivially true. We want to prove that the conditions of lemma 9 are satisfied with $a = 3/2, b = 2, c_2 = 1/9, c_1 = \frac{36\sqrt{q}}{1-1/q}$ and $c = \frac{18}{\sqrt{e}}\frac{1}{1-1/q}$.

First, we need to verify that condition (3) holds. The function ψ is indeed increasing and continuously differentiable. It is also easy to prove that ψ'/ψ is increasing. The only nontrivial part is to prove $\lim_{u\to-\infty}\psi(u) = \beta < 1/K$. We have $q^{\frac{q}{2(q-1)}} = \sqrt{q}\exp\left(\frac{\log q}{2(q-1)}\right) \leq \sqrt{eq}$. Thus we have $\beta \leq \sqrt{\frac{eq}{nK}}$. But by (27) we know that $\sqrt{qK/n} \leq \frac{1}{37} \leq \frac{1}{2\sqrt{e}}$ and thus $\beta \leq \frac{1}{2K}$. We will now check that (21) - (26) hold.

Inequality (21). The right side of (21) is upper bounded by $1.3 + \sqrt{q}\frac{36}{1-1/q}\sqrt{\frac{K}{n}} + 3\frac{K}{n}$ and c is lower bounded by 10.8. We trivially have $3\frac{K}{n} \leq 3\sqrt{K/n} \leq \frac{3\sqrt{q}}{1-1/q}\sqrt{K/n}$. Thus (21) holds if $9.5 \geq \frac{39\sqrt{q}}{1-1/q}\sqrt{K/n}$ and, from (27), this is true.

Inequality (22). Inequality (22) is implied by $\beta \leq 1/(2K)$ and $\alpha/(cn)^q \leq 1/(2K)$. We already proved that the first inequality is true. The second inequality boils down to

$$\sqrt{qnK} \leq \frac{1}{\sqrt{e}}\frac{n}{1-1/q},$$

which is true from (27).

Inequality (23). We have $-\psi^{-1}(x) = \left(\frac{\alpha}{x-\beta}\right)^{1/q}$. Thus

$$-\sum_{i=1}^{K} q_i \psi^{-1}(q_i) = -\sum_{i=1}^{K} \beta\psi^{-1}(q_i) - \sum_{i=1}^{K}(q_i-\beta)\psi^{-1}(q_i)$$
$$\leq \beta cnK + \alpha^{1/q}\sum_{i=1}^{K}(q_i-\beta)^{1-1/q}$$

since $q_i \geq \psi(-cn)$. From Holder's inequality, we get

$$\sum_{i=1}^{K}(q_i-\beta)^{\frac{q-1}{q}} \leq K^{1/q}\left(\sum_{i=1}^{K}(q_i-\beta)\right)^{\frac{q-1}{q}} \leq K^{1/q}.$$

We also need the following computations:

$$-\sum_{i=1}^{K}\int_{q_i}^{1/K}\psi^{-1}(u)du = \sum_{i=1}^{K}\int_{q_i}^{1/K}\left(\frac{\alpha}{u-\beta}\right)^{1/q}du$$
$$= \sum_{i=1}^{K}\alpha^{1/q}\left[\frac{(u-\beta)^{1-1/q}}{1-1/q}\right]_{q_i}^{1/K}$$
$$\leq K\frac{\alpha^{1/q}}{1-1/q}\left(\frac{1}{K}\right)^{1-1/q} = \frac{(\alpha K)^{1/q}}{1-1/q}$$

since $q_i \geq \psi(-cn) \geq \beta$ and $\beta \leq 1/K$.

Hence we need to verify that

$$c_1\sqrt{nK} \geq \beta cnK + (\alpha K)^{1/q} + (\alpha K)^{1/q}/(1-1/q),$$

which is true with our particular values.

Inequality (24). For any positive real numbers x, y, we have $x + y \geq \max(x,y) \geq x^{\frac{q-1}{2q}} y^{\frac{q+1}{2q}}$, hence

$$\psi^2(x) = \left(\frac{\alpha}{(-x)^q}+\beta\right)^2 \geq \beta^{\frac{q-1}{q}}\left(\frac{\alpha}{(-x)^q}\right)^{\frac{q+1}{q}}$$
$$= (\alpha\beta^{q-1})^{1/q}\frac{\alpha}{(-x)^{q+1}}$$
$$= \frac{(\alpha\beta^{q-1})^{1/q}}{q}\psi'(x).$$

With our particular values, we have

$$(\alpha\beta^{q-1})^{1/q} = 9q\,(n/K)^{1/q} \geq 9q$$

and thus (24) is satisfied with $c_2 = 1/9$.

Inequality (25). We have

$$\frac{\psi'(x)}{\psi'\left(x+1/\psi(x)\right)} = \left(1-\frac{1}{-x\psi(x)}\right)^{q+1},$$

and

$$-x\psi(x) = \frac{\alpha}{(-x)^{q-1}} + (-x\beta)$$
$$\geq (\alpha\beta^{q-1})^{1/q}\left((q-1)^{1/q} + (q-1)^{1/q-1}\right)$$
$$\geq (\alpha\beta^{q-1})^{1/q} \geq 9q.$$

Consequently, we have

$$\frac{\psi'(x)}{\psi'\left(x+1/\psi(x)\right)} \geq \left(1-\frac{1}{9q}\right)^{q+1}.$$

Now since $(1-1/u)^{u-1} \geq e^{-1}$ for any $u \geq 1$, we have

$$\left(1-\frac{1}{9q}\right)^{q-1} \geq \left(1-\frac{1}{9q}\right)^{(9q-1)/9} \geq e^{-1/9}.$$

In particular, we have

$$\psi'\left(x+\frac{1}{\psi(x)}\right) \leq e^{1/9}\left(1-\frac{1}{9q}\right)^{-2}\psi'(x) \leq \frac{3}{2}\psi'(x),$$

and thus inequality (25) is satisfied with $a = 3/2$.

Inequality (26). We have

$$\frac{\psi'(x)}{\psi'(x-\frac{ab}{\psi(x)})} \leq \left(1+\frac{ab}{9q}\right)^{q+1} \leq \exp\left(\frac{q+1}{9q}ab\right) \leq e^{\frac{2}{9}ab}$$

where we have used $\log(1+x) \leq x$. Now for $a = 3/2$ and $b = 2$, the last term is upper bounded by b, hence (26) holds.

Application of Lemma 9. One can easily check that

$$\frac{\psi'(x)}{\psi(x)} \leq \frac{q}{\alpha^{1/q}}\psi(x)^{1/q}.$$

Thus with Holder's inequality we get

$$\sum_{t=1}^{n-1}\sum_{i=1}^{K}\frac{\psi'(\tilde{G}_{i,t-1}-C_{t-1})}{\psi(\tilde{G}_{i,t-1}-C_{t-1})} \leq \frac{qn}{\alpha^{1/q}}K^{1-1/q} = \frac{\sqrt{qnK}}{9}.$$

Then we have

$$R_n \leq 1 + \frac{36\sqrt{qnK}}{1-1/q} + \frac{21\sqrt{qnK}}{72} \leq \frac{37}{1-1/q}\sqrt{qnK}.$$

C Proof of Theorem 5

We follow the steps described in the sketch of proof given right after Theorem 5. We may assume $\mu_1 \geq \ldots \geq \mu_K$.

First step: Decoupling the arms. For an arm k_0, we trivially have $\sum_{k=1}^{K}\Delta_k T_k(n) \leq n\Delta_{k_0} + \sum_{k=k_0+1}^{K}\Delta_k T_k(n)$. Let $\Delta_{K+1} = +\infty$, $z_k = \mu_1 - \frac{\Delta_k}{2}$ for $k_0 < k \leq K+1$ and $z_{k_0} = +\infty$. Define

$$Z = \min_{1\leq s\leq n} B_{1,s},$$

and

$$W_{j,k} = \mathbb{1}_{Z\in[z_{j+1},z_j)}\Delta_k T_k(n).$$

We have

$$\sum_{k=k_0+1}^{K}\Delta_k T_k(n) = \sum_{k=k_0+1}^{K}\sum_{j=k_0}^{K} W_{j,k}$$
$$= \sum_{j=k_0}^{K}\sum_{k=k_0+1}^{j} W_{j,k} + \sum_{j=k_0}^{K}\sum_{k=j+1}^{K} W_{j,k}. \quad (28)$$

An Abel transformation takes care of the first sum of (28):

$$\sum_{j=k_0}^{K}\sum_{k=k_0+1}^{j} W_{j,k} \leq \sum_{j=k_0}^{K}\mathbb{1}_{Z\in[z_{j+1},z_j)}n\Delta_j$$
$$= n\Delta_{k_0} + n\sum_{j=k_0+1}^{K}\mathbb{1}_{Z<z_j}(\Delta_j - \Delta_{j-1}). \quad (29)$$

To bound the second sum of (28), we introduce the stopping times $\tau_k = \min\{t : B_{k,t} < z_k\}$ and remark that, by definition of MOSS, we have $\{Z \geq z_k\} \subset \{T_k(n) \leq \tau_k\}$, since once we have pulled τ_k times arm k its index will always be lower than the index of arm 1. This implies

$$\sum_{j=k_0}^{K}\sum_{k=j+1}^{K} W_{j,k} = \sum_{k=k_0+1}^{K}\sum_{j=k_0}^{k-1} W_{j,k} = \sum_{k=k_0+1}^{K} \mathbb{1}_{Z\geq z_k}\Delta_k T_k(n) \leq \sum_{k=k_0+1}^{K} \tau_k\Delta_k. \quad (30)$$

Combining (28), (29) and (30) and taking the expectation, we get

$$R_n \leq 2n\Delta_{k_0} + \sum_{k=k_0+1}^{K}\Delta_k\mathbb{E}\tau_k + n\sum_{k=k_0+1}^{K}\mathbb{P}(Z<z_k)(\Delta_k-\Delta_{k-1}). \quad (31)$$

Let $\delta_0 = e^{1/16}\sqrt{\frac{K}{n}}$ and set k_0 such that $\Delta_{k_0} \leq \delta_0 < \Delta_{k_0+1}$.

Second step: Bounding $\mathbb{E}\tau_k$ for $k_0+1 \leq k \leq K$.

Let $\log_+(x) = \max(\log(x), 0)$. For $\ell_0 \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{E}\tau_k - \ell_0 &= \sum_{\ell=0}^{+\infty}\mathbb{P}(\tau_k > \ell) - \ell_0 \quad (32)\\ &\leq \sum_{\ell=\ell_0}^{+\infty}\mathbb{P}(\tau_k>\ell) = \sum_{\ell=\ell_0}^{+\infty}\mathbb{P}(\forall t\leq\ell, B_{k,t}>z_k)\\ &\leq \sum_{\ell=\ell_0}^{+\infty}\mathbb{P}\left(\widehat{X}_{k,\ell}-\mu_k \geq \frac{\Delta_k}{2} - \sqrt{\frac{\log_+(n/K\ell)}{\ell}}\right).\end{aligned}$$

Now let us take $\ell_0 = \lceil 8\log\left(\frac{n}{K}\Delta_k^2\right)/\Delta_k^2\rceil$ with $\lceil x\rceil$ the smallest integer larger than x. For $\ell \geq \ell_0$, since $k > k_0$, we have $\ell \geq \Delta_k^{-2}$, and thus $8\log_+(n/(K\ell)) \leq \ell\Delta_k^2$, hence

$$\frac{\Delta_k}{2} - \sqrt{\frac{\log_+(n/(K\ell))}{\ell}} \geq \frac{\Delta_k}{2} - \frac{\Delta_k}{\sqrt{8}} = c\Delta_k,$$

with $c = \frac{1}{2} - \frac{1}{\sqrt{8}}$. Therefore, by using Hoeffding's inequality and (32), we get

$$\begin{aligned}\mathbb{E}\tau_k - \ell_0 &\leq \sum_{\ell=\ell_0}^{+\infty}\mathbb{P}\left(\widehat{X}_{k,\ell}-\mu_k \geq c\Delta_k\right)\\ &\leq \sum_{\ell=\ell_0}^{+\infty}\exp\left(-2\ell(c\Delta_k)^2\right) = \frac{\exp\left(-2\ell_0(c\Delta_k)^2\right)}{1-\exp\left(-2(c\Delta_k)^2\right)}\\ &\leq \frac{1}{1-\exp\left(-2c^2\Delta_k^2\right)}. \quad (33)\end{aligned}$$

Plugging the value of ℓ_0, we obtain

$$\begin{aligned}\Delta_k\mathbb{E}\tau_k &\leq \Delta_k\left(1 + \frac{8\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k^2}\right) + \frac{\Delta_k}{1-\exp\left(-2c^2\Delta_k^2\right)}\\ &\leq 1 + 8\frac{\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k} + \frac{1}{2c^2(1-c^2)\Delta_k}, \quad (34)\end{aligned}$$

where the last step uses that, since $1-\exp(-x) \geq x - x^2/2$ for any $x \geq 0$, we have

$$\frac{1}{1-\exp\left(-2c^2\Delta_k^2\right)} \leq \frac{1}{2c^2\Delta_k^2 - 2c^4\Delta_k^4} \leq \frac{1}{2c^2\Delta_k^2(1-c^2)}$$

It is routine to check that $\frac{2}{e}\sqrt{\frac{n}{K}}$ is the maximum of $x \mapsto x^{-1}\log\left(\frac{n}{K}x^2\right)$. By using $\Delta_k \geq e^{1/16}\sqrt{K/n}$, we finally get

$$\begin{aligned}K\max_{k>k_0}\Delta_k\mathbb{E}\tau_k &\leq K + \left(\frac{16}{e} + \frac{e^{-1/16}}{2c^2(1-c^2)}\right)\sqrt{nK}\\ &\leq K + 28.3\sqrt{nK}. \quad (35)\end{aligned}$$

Third step: Bounding $n\sum_{k=k_0+1}^{K}\mathbb{P}(Z<z_k)(\Delta_k-\Delta_{k-1})$.

Let X_t denote the reward obtained by arm 1 when it is drawn for the t-th time. The random variables $X_1, X_2, \ldots$ are i.i.d. so that we have the maximal inequality (Hoeffding, 1963, inequality (2.17)): for any $x > 0$ and $m \geq 1$,

$$\mathbb{P}\left(\exists s \in \{1,\ldots,m\}, \sum_{t=1}^{s}(\mu_1 - X_t) > x\right) \leq \exp\left(-\frac{2x^2}{m}\right).$$

Since $z_k = \mu_1 - \Delta_k/2$ and since $u \mapsto \mathbb{P}(Z < \mu_1 - u/2)$ is a nonincreasing function, we have

$$\begin{aligned}&\sum_{k=k_0+1}^{K}\mathbb{P}(Z<z_k)(\Delta_k-\Delta_{k-1})\\ &\leq \quad \delta_0 - \Delta_{k_0} + \int_{\delta_0}^{1}\mathbb{P}\left(Z<\mu_1-\frac{u}{2}\right)du.\end{aligned}$$

For a fixed $u \in [\delta_0, 1]$ and $f(u) = 8\log\left(\sqrt{\frac{n}{K}}u\right)/u^2$, we have

$$\begin{aligned}&\mathbb{P}\left(Z<\mu_1-\frac{1}{2}u\right)\\ &= \mathbb{P}\left(\exists 1\leq s\leq n : \sum_{t=1}^{s}(\mu_1-X_t) > \sqrt{s\log_+\left(\frac{n}{Ks}\right)} + \frac{su}{2}\right)\\ &\leq \mathbb{P}\left(\exists 1\leq s\leq f(u) : \sum_{t=1}^{s}(\mu_1-X_t) > \sqrt{s\log\left(\frac{n}{Ks}\right)}\right)\\ &\qquad + \mathbb{P}\left(\exists f(u)<s\leq n : \sum_{t=1}^{s}(\mu_1-X_t) > \frac{su}{2}\right).\end{aligned}$$

For the first term we use a peeling argument with a geometric grid of the from $\frac{1}{2^{\ell+1}}f(u) \leq s \leq \frac{1}{2^\ell}f(u)$:

$$\begin{aligned}&\mathbb{P}\left(\exists 1\leq s\leq f(u) : \sum_{t=1}^{s}(\mu_1-X_t) > \sqrt{s\log\left(\frac{n}{Ks}\right)}\right)\\ &\leq \quad \sum_{\ell=0}^{+\infty}\mathbb{P}\left(\exists\frac{1}{2^{\ell+1}}f(u)\leq s\leq\frac{1}{2^\ell}f(u) :\right.\\ &\qquad\qquad \left.\sum_{t=1}^{s}(\mu_1-X_t) > \sqrt{\frac{f(u)}{2^{\ell+1}}\log\left(\frac{n2^\ell}{Kf(u)}\right)}\right)\\ &\leq \quad \sum_{\ell=0}^{+\infty}\exp\left(-2\frac{f(u)\frac{1}{2^{\ell+1}}\log\left(\frac{n2^\ell}{Kf(u)}\right)}{f(u)\frac{1}{2^\ell}}\right)\\ &= \quad \sum_{\ell=0}^{+\infty}\frac{Kf(u)}{n}\frac{1}{2^\ell} = 2\frac{Kf(u)}{n}.\end{aligned}$$

We integrate $f(u)$ now:

$$\int_{\delta_0}^{1} f(u)du = \left[\frac{8\log(e\sqrt{n/K}u)}{u}\right]_1^{\delta_0} \leq \frac{17\,e^{-1/16}}{2}\sqrt{n/K}.$$

For the second term we also use a peeling argument but with a geometric grid of the form $2^\ell f(u) \leq s \leq 2^{\ell+1} f(u)$:

$$\begin{aligned}
&\mathbb{P}\left(\exists s \in \{\lceil f(u)\rceil, \dots, n\} : \sum_{t=1}^{s}(\mu_1 - X_t) > \frac{su}{2}\right) \\
&\quad\leq \sum_{\ell=0}^{+\infty} \mathbb{P}\left(\exists 2^\ell f(u) \leq s \leq 2^{\ell+1} f(u) : \right. \\
&\qquad\qquad \left. \sum_{t=1}^{s}(\mu_1 - X_t) > 2^{\ell-1} f(u) u\right) \\
&\quad\leq \sum_{\ell=0}^{+\infty} \exp\left(-2\frac{\left(2^{\ell-1}f(u)u\right)^2}{f(u)2^{\ell+1}}\right) \\
&\quad= \sum_{\ell=0}^{+\infty} \exp\left(-2^\ell f(u)u^2/4\right) \\
&\quad\leq \sum_{\ell=0}^{+\infty} \exp\left(-(\ell+1)f(u)u^2/4\right) \\
&\quad= \frac{1}{\exp\left(f(u)u^2/4\right) - 1}.
\end{aligned}$$

From the choice of $f(u)$, this last term is upper bounded by $\frac{1}{nu^2/K - 1}$. Again we need to integrate this quantity. It is easy to show that

$$\int_{\delta_0}^{1} \frac{1}{nu^2/K - 1} du \leq \frac{1}{2}\log\left(\frac{e^{1/16}+1}{e^{1/16}-1}\right)\sqrt{\frac{K}{n}}.$$

All in all, we have

$$\begin{aligned}
&n\sum_{k=k_0+1}^{K} \mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1}) \\
&\leq n(\delta_0 - \Delta_{k_0}) + \left(17e^{-1/16} + \frac{1}{2}\log\left(\frac{e^{1/16}+1}{e^{1/16}-1}\right)\right)\sqrt{nK} \\
&\leq n(\delta_0 - \Delta_{k_0}) + 17.8\sqrt{nK}. \qquad (36)
\end{aligned}$$

Combining (31), (35) and (36), we get that

$$R_n \leq 48.3\sqrt{nK} + K.$$

Now, if $K \geq n/49^2$, we have $49\sqrt{nK} \geq n \geq R_n$. On the other hand, if $K < n/49^2$ then we also have $R_n \leq 49\sqrt{nK}$ since $K < \frac{1}{49}\sqrt{nK}$.

D Proof of Theorem 6

Without loss of generality, we again assume: $\mu_1 \geq \ldots \geq \mu_K$, and use the notation of the proof of Theorem 5. In particular, we recall that $z_k = \mu_1 - \frac{\Delta_k}{2}$, $Z = \min_{1\leq s\leq n} B_{1,s}$, and $\tau_k = \min\{t : B_{k,t} < z_k\}$. This time, we use k_0 such that $\Delta_{k_0} \leq \delta_0 < \Delta_{k_0+1}$ with $\delta_0 = 2\sqrt{K/n}$. We start with the following weakened version of (31):

$$R_n \leq 2n\Delta_{k_0} + \sum_{k=k_0+1}^{K} \Delta_k\big(\mathbb{E}\tau_k + n\mathbb{P}(Z < z_k)\big). \qquad (37)$$

Consider $k > k_0$. From (34), we have

$$\Delta_k \mathbb{E}\tau_k \leq 1 + 8\frac{\log\left(n\Delta_k^2/K\right)}{\Delta_k} + \frac{24}{\Delta_k}.$$

Now we have $\frac{n}{K}\Delta_k^2 \geq 4$, hence

$$\Delta_k \mathbb{E}\tau_k \leq 27\frac{\log(n\Delta_k^2/K)}{\Delta_k}.$$

For the last term in (37), we use the same cutting scheme and peeling argument as in the third step of the previous proof except that we take $u = \Delta_k$ and replace $f(u)$ with $v = \frac{4\log(n\Delta_k^2/K)}{\Delta_k^2}$. Introducing

$$A_1 = \mathbb{P}\left(\exists 1 \leq s \leq v : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{s\log\left(\frac{n}{Ks}\right)}\right),$$

and

$$A_2 = \mathbb{P}\left(\exists v < s \leq n : \sum_{t=1}^{s}(\mu_1 - X_t) > \frac{su}{2}\right),$$

we have

$$\mathbb{P}\left(Z < z_k\right) \leq A_1 + A_2.$$

The first peeling argument with a geometric grid of the from $\frac{1}{2^{\ell+1}}v \leq s \leq \frac{1}{2^\ell}v$ gives

$$A_1 \leq 2\frac{Kv}{n} = \frac{8K\log(n\Delta_k^2/K)}{n\Delta_k^2}.$$

The second peeling argument with a geometric grid of the form $2^\ell v \leq s \leq 2^{\ell+1}v$ gives

$$A_2 \leq \frac{1}{\exp\left(vu^2/4\right) - 1} = \frac{1}{n\Delta_k^2/K - 1} \leq \frac{4K}{3n\Delta_k^2},$$

where we again use $\frac{n}{K}\Delta_k^2 \geq 4$. By combining the previous inequalities, we get $n\Delta_k\mathbb{P}\left(Z < z_k\right) \leq \frac{9K\log(n\Delta_k^2/K)}{\Delta_k}$. Since $\Delta_{k_0} \leq 2\sqrt{K/n}$, by plugging the previous results in (37), we obtain

$$\begin{aligned}
R_n &\leq \frac{8K}{\Delta_{k_0}}\mathbb{1}_{\Delta_{k_0}>0} + \sum_{k=k_0+1}^{K} \frac{(27+9K)\log(n\Delta_k^2/K)}{\Delta_k} \\
&\leq 23K \sum_{k:\Delta_k>0} \frac{\max(\log(n\Delta_k^2/K), 1)}{\Delta_k}.
\end{aligned}$$

■

Acknowledgement

This work has been supported by the French National Research Agency (ANR) through the COSINUS program (ANR-08-COSI-004: EXPLO-RA project).