Approximate information maximization for bandit games

Alex Barbier–Chebbah
Institut Pasteur, Université Paris Cité,
CNRS UMR 3571, Paris France
Epiméthée INRIA, Paris France

Chrisitian L. Vestergaard
Institut Pasteur, Université Paris Cité,
CNRS UMR 3571, Paris France
Epiméthée INRIA, Paris France

Jean-Baptiste Masson
Institut Pasteur, Université Paris Cité,
CNRS UMR 3571, Paris France
Epiméthée INRIA, Paris France

Etienne Boursier
INRIA, Université Paris Saclay
LMO, Orsay, France

Abstract

Entropy maximization and free energy minimization are general principles for modeling dynamic systems. Notable examples include modeling decision-making within the brain using the free-energy principle, optimizing the accuracy-complexity trade-off when accessing hidden variables with the information bottleneck principle (Tishby et al., 2000), and navigation in random environments using information maximization (Vergassola et al., 2007). Building on this principle, we propose a new class of bandit algorithms that maximize an approximation to the information of a key variable within the system. To this end, we develop an approximated, analytical physics-based representation of the entropy to forecast the information gain of each action and greedily choose the one with the largest information gain. This method yields strong performance in classical bandit settings. Motivated by its empirical success, we prove its asymptotic optimality for the multi-armed bandit problem with Gaussian rewards. Since it encompasses the system's properties in a single, global functional, this approach can be efficiently adapted to more complex bandit settings. This calls for further investigation of information maximization approaches for bandit problems.

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 Introduction

Multi-armed bandit problems have attracted wide attention in the past decades. They embody the challenge of balancing exploration and exploitation and have been applied to various settings such as online recommendation (Bresler et al., 2014), medical trials (Thompson, 1933), dynamic pricing (Den Boer, 2015), and reinforcement learning-based decision making (Silver et al., 2016; Ryzhov et al., 2012). Besides the classic stochastic version of the multi-armed bandit problem, many subsequent extensions have been developed, providing richer models for specific applications. These extensions include linear bandits (Li et al., 2010), many-armed bandits (Bayati et al., 2020), and pure exploration problems such as thresholding bandits (Locatelli et al., 2016) or top-K bandits (Kalyanakrishnan et al., 2012; Kaufmann et al., 2016).

In the classic bandit game, an agent chooses an arm at each time step and observes a stochastic reward. Since they only observe the payoff of the chosen arm, the agent should regularly explore suboptimal arms. This is often referred to as the exploration-exploitation trade-off. An agent can exploit its current knowledge to optimize gains by drawing the current empirically best arm, or it can explore other arms to potentially increase future gains.

Optimal strategies are characterized asymptotically by the Lai and Robbins bound (Lai et al., 1985). Among them, upper confidence bound (UCB, Auer, 2000; Garivier and Cappé, 2011) methods greedily pull the arm maximizing some tuned confidence index; Thompson sampling (Kaufmann et al., 2012a; Agrawal and Goyal, 2013) relies on sampling mean rewards from a posterior distribution and chooses the arm with the largest random sample; deterministic minimum empirical divergence (DMED, Honda and Takemura, 2010) builds on a balance between the maximum likelihood of an arm being the best and the posterior expectation of the regret.

Even if these approaches efficiently utilize currently available information, they do not aim directly to acquire more information. We highlight, however, the information directed sampling approach (IDS) of Russo and Van Roy (2014), which relies on a measure of the information gain of the optimal actions. By leveraging an information measure that consistently captures the specific problem structure, IDS can address general classes of problems, particularly those with a complex information structure where classic bandit methods fall short. Surprisingly, IDS can even outperform UCB and Thompson sampling in classic bandit problems. However, like DMED, this method explicitly balances information gain with expected losses induced by exploration, and the efficiency of pure information-maximizing strategies thus remains to be proven.

Information-maximization approaches provide a decision-making strategy in which the agent tries to maximize information about one or more relevant stochastic variables. The information-maximizing principle has been shown to be efficient in a broad range of domains (Helias and Dahmen, 2020; Parr et al., 2022; Hernández-Lobato et al., 2015; Vergassola et al., 2007) where decisions have to be taken in fluctuating or unknown environments. Its applications include robotics (Zhang et al., 2015), where the ability to share approximate information improves collective decisions, and the search for olfactory sources in turbulent flows (Masson, 2013; Reddy et al., 2022).

In the specific setting of bandit problems, information maximization has shown promising empirical results, and heuristic arguments support its asymptotic optimality (Reddy et al., 2016; Barbier-Chebbah et al., 2023). As IDS, information maximization leverages the information structure to provide a versatile decision framework with the capability to address various bandit settings. However, the efficiency of such a "pure exploration" strategy in terms of regret minimization has yet to be proven, and it has been previously argued that it would result in a linear regret (Russo and Van Roy, 2014). Moreover, current information-based algorithms often rely on numerical integration of a complex functional, leading to high computational costs, a significant challenge that information-based methods must overcome. In this context, we aim to leverage new, computationally efficient strategies derived from information maximization principles focusing on global observables, i.e., variables depending on more than one arm, and to rigorously prove their efficiency.

Contributions. Our main contribution is introducing a new class of asymptotically optimal algorithms that rely on approximations of a functional representing the current information of interest about the whole bandit system. This approach is based on the entropy of the posterior mean value of the best arm, for which we provide an approximate expression to enable robust, easily tunable, and extendable algorithms with a direct analytical formulation. We focus here on the multi-armed bandit problem, for which we derive a simple approximate information maximization algorithm (AIM) and provide an upper bound on its expected regret, ensuring that AIM is asymptotically optimal. The information from each arm is incorporated in a unique entropy functional, which shows promise for tackling more complex bandit settings such as linear or many-armed bandits. Thus, our main motivation is to design an analytical functional-based algorithmic principle, which can potentially address problems with more correlated information structures in the future. Additionally, another strength of AIM lies in its short-time behavior, where it shows strong performances as we illustrate numerically.

Organization. In Section 2, we briefly review the K-armed bandit problem. Section 3 presents the general principle of information maximization, originally inspired by both the information bottleneck principle and navigation in turbulent plumes (Masson, 2013; Reddy et al., 2022). Section 4 upper bounds the regret of AIM, showing it attains Lai and Robbin's asymptotic bound. In Section 5, the performance of AIM is numerically compared with known baselines on multiple examples. Finally, Section 6 discusses extensions of AIM to various bandit settings.

2 Setting

We consider the classic K-armed stochastic bandit game. In each round t, the agent selects an arm $a_t \in [K] = \{1, \dots, K\}$ among a set of K choices solely based on the rewards of the previously pulled arms. The chosen arm k then returns a stochastic reward $X_t(k)$, drawn independently of the previous rounds, according to a distribution ν_k of mean μ_k. We denote by $N_k(t)$ the number of times the arm k has been pulled. When clear from context, we omit the dependence on t for simplicity.

The goal of the agent is to maximize its cumulative reward, or equivalently, to minimize its expected regret up to round T, defined as

$$R(T) = \mu^* T - \sum_{t=1}^{T} \mathbb{E}[\mu_{a_t}], \tag{1}$$

where $\mu^* = \max_{i \in [K]} \mu_i$. Hence, the agent will optimize its choice of a_t relying on the previous observations up to t. For a large family of reward distributions, the asymptotic regret is lower-bounded for any uniformly good policy by

$$\liminf_{T\to\infty} \frac{R(T)}{\ln(T)} \geq \sum_{k,\mu_k<\mu^*} \frac{\mu^* - \mu_k}{D_{\mathrm{KL}}(\nu_k \| \nu_{k^*})}, \tag{2}$$

where $k^* \in \mathrm{argmax}_{i\in[K]}\, \mu_i$,, and $D_{\mathrm{KL}}(\nu_k \| \nu_{k^*})$ denotes the Kullback-Leibler divergence between the reward distributions of the arms k and k^* (Lai et al., 1985). In the particular case of Gaussian rewards with equal variances, i.e., $\nu_i = \mathcal{N}(\mu_i, \sigma^2)$, the Kullback-Leibler divergence is $D_{\mathrm{KL}}(\nu_k \| \nu_{k^*}) = (\mu^* - \mu_k)^2/(2\sigma^2)$.

3 Information maximization strategies

Here, we introduce entropy-based, information maximization strategies and their underlying physical principles, which serve as the basis for the AIM algorithms. We then detail an approximation scheme leading to an analytical and simplified entropy functional and illustrate it with the case of Gaussian rewards.

3.1 Algorithm design principle: physical intuition

The information maximization principle (Vergassola et al., 2007; Reddy et al., 2016) has shown to be effective in taxis strategies where a mobile agent needs to find an emitting odour source (Martinez et al., 2014; Cardé, 2021; Murlis et al., 1992). This principle can be summarized as: identify a well-chosen quantity that captures the state of the game, calculate its associated entropy which quantifies the information accessible to the agent, and select the action (i.e., the "arm") that minimizes this entropy thereby providing the most information. To get an efficient strategy, we thus aim to design a functional encompassing the current available information of the full system.

For the classic bandit setting, we rely on an entropy functional for policy decision. More precisely, we choose $S_{\max}$, the entropy of the posterior distribution $p_{\max}$ of the value of the maximal mean reward. With independent priors[1], the posterior distribution of the value of the maximal mean reward can be expressed as

$$\begin{aligned} p_{\max}(\theta)\mathrm{d}\theta &= \mathrm{d}\mathbb{P}\left(\max_k \mu_k = \theta \mid \mathcal{F}_{t-1}\right) \\ &= \sum_{k=1}^{K} \mathrm{d}\mathbb{P}(\mu_k = \theta | \mathcal{F}_{t-1}) \prod_{j\neq k} \mathbb{P}(\mu_j \leq \theta | \mathcal{F}_{t-1}), \end{aligned} \tag{3}$$

where $\mathcal{F}_{t-1} = \sigma(X_1(a_1), \ldots, X_{t-1}(a_{t-1}))$ denotes the filtration associated to the observations up to time $t-1$. The associated entropy reads

$$S_{\max} = -\int_{\Theta} p_{\max}(\theta) \ln p_{\max}(\theta) \mathrm{d}\theta, \tag{4}$$

where $\Theta = [\mu_{\inf}, \mu_{\sup}]$ is the support of $p_{\max}$ (which depends on the nature of the game and can be infinite). Note that, as exemplified by Eq. (3), $p_{\max}$ includes the arms' priors and directly depends on the reward distributions.[2] The entropy, $S_{\max}$, is a measure of the information carried by all arms in a single functional, providing a global state description of the game.

Our information maximization policy aims to minimize the entropy of $p_{\max}$. For that, it greedily chooses the arm providing the largest expected decrease in entropy, conditioned on the current knowledge of the game. This is achieved by choosing the arm minimizing the entropy increment:

$$\underset{k\in[K]}{\mathrm{argmin}}\, \mathbb{E}\left[S_{\max}(t) - S_{\max}(t-1) \mid \mathcal{F}_{t-1}, a_t = k\right]. \tag{5}$$

Similar to Thompson sampling, it relies on a Bayesian representation. Yet, it distinguishes itself by providing a deterministic decision procedure given past observations. We stress that $S_{\max}$ quantifies the available information about the average reward of the best arm. This choice contrasts with using the entropy of the probability of the best arm, which is known to overexplore and is suboptimal for regret minimization (Reddy et al., 2016). Because of this suboptimality, existing approaches based on the information on the best arm fix this concern by including the expected regret in the functional to favor exploitation (Russo and Van Roy, 2014; Kirschner and Krause, 2018a,b; Foster et al., 2021; Xu and Zeevi, 2023) (see Appendix A for a detailed comparison with IDS).

Hence, the entropy choice Eq. (4) is crucial, as it determines how contributions from different arms to the posterior distribution combine to achieve an optimal balance. As shown in Fig. 2, the main body of the posterior distribution captures most cases where the

[1] Even though our approach may be derived for any prior distribution for the arms' mean rewards we use a uniform prior which admits a simpler expression for the entropy.

[2] In the remainder of the paper, we consider an improper uniform prior over $\mathbb{R}$, as often considered with Gaussian rewards.

better empirical arm is indeed optimal. To learn the maximal mean value, a sound information-based algorithm must spend most of its time gathering information from this main body component. In contrast, the posterior tail addresses rare events where a seemingly suboptimal arm might prove to be the best. However, the information gain from the posterior tail is often lower, as it deals with rare events, naturally leading to the desired exploration-exploitation ratio. Furthermore, as we prove in Section 4, by the definition of $p_{\max}$, the information carried by the arms' posteriors ensures an optimal behavior of the policy. In particular, since it aims to maximize the information about the best arm's mean, it mainly pulls the current best arm to learn more about its value, balancing exploration and exploitation. On the contrary, policies solely aiming to identify the best arm pull worse empirical arms more often because they are only concerned about the arms' order, leading to over-exploration.

The information maximization policy based on Eq. (5) has been empirically shown to be competitive with state-of-the-art algorithms in classic bandit games (Reddy et al., 2016; Barbier-Chebbah et al., 2023). However, while Eq. (5) can be numerically evaluated, it cannot be computed in closed form in general, preventing the policy from being analytically tractable.

This entails two important practical drawbacks. First, it makes it difficult to theoretically bound the regret, hindering the extension of its principles to more complicated bandit settings. Second, it induces a high computational cost (a trait shared with IDS Russo and Van Roy, 2014). This in particular becomes disadvantageous when considering a large number of arms and at large times (when $p_{\max}$ is peaked), where one has to manage vanishing numerical precision, making the numerical integration even longer.

To address these concerns, we derive in Section 3.2 a simplified and analytical functional mirroring $S_{\max}$ and its increments. This analytical result strengthens the information maximization principle, both by providing novel algorithms that are analytical, tractable, and computationally efficient while conserving the main advantages of the exact entropy (Reddy et al., 2016) and by making theoretical analysis tractable.

3.2 Main elements of the analytical entropy approximation

In this section, we devise a set of approximations of $p_{\max}$ and $S_{\max}$ to obtain a tractable analytical algorithm. Given that the best empirical arm and the worst empirical arms have notably distinct contributions to $p_{\max}$ (Fig. 1(a)), we approximate $p_{\max}$ while considering the current arms' order. We sort them based on their current posterior means, labeling the highest one as M_t (with an empirical reward of $\hat{\mu}_{M_t}$) and $\mathcal{A}_t = [K] \setminus \{M_t\}$ the set of worst empirical arms. Of course, M_t might differ from the actual optimal arm k^* due to the randomness in the observed rewards. We focus on approximating Eqs. (3) and (4) when the best empirical arm has already been drawn much more often than the other arms.

The entropy is then decomposed into two tractable terms corresponding to distinct behaviors of $p_{\max}(\theta)$ when θ varies:

$$\tilde{S}_{\max} = \tilde{S}_{\text{body}} + \tilde{S}_{\text{tail}}. \tag{6}$$

The first term, $\tilde{S}_{\text{body}}$, approximates the contribution around the mode of $p_{\max}$ where the maximum mean value is most likely to occur. The second term, $\tilde{S}_{\text{tail}}$, quantifies rare events referred to as the tail of $p_{\max}$ (corresponding to high mean rewards, see Fig. 1).

Each of these terms then corresponds to a part of the entropy where the dominant term of Eq. (3) is distinct, yielding a distinct expression for each (see Appendix B.1 for details).

More precisely, by denoting by $p_i(\theta)\mathrm{d}\theta = \mathrm{d}\mathbb{P}(\mu_i = \theta \mid \mathcal{F}_t)$ the mean posterior density of the associated arm i, the tail term is approximated as

$$\tilde{S}_{\text{tail}} = -\sum_{m\in\mathcal{A}_t} \int_{\tilde{\mu}_{\text{eq},m}}^{\mu_{\sup}} p_m(\theta) \ln p_m(\theta) \mathrm{d}\theta, \tag{7}$$

where $\tilde{\mu}_{\text{eq},m}$, given in Appendix B.6, approximates $\bar{\mu}_{\text{eq},m}$, the value of θ where the empirical best arm M_t and the selected worse arm m have the same probability of being the true best arm (see red and orange curves in Fig. 1(b)). Here, $p_m(\theta)$ is the posterior density of the current worse arm evaluated at θ. Note that we have neglected the posterior density of the best empirical arm. Roughly, because M_t has predominantly been drawn, $p_{M_t}(\theta)$ decays faster than $p_m(\theta)$, thereby exhibiting a negligible contribution to the tail (by definition of $\tilde{\mu}_{\text{eq},m}$). The approximate entropy of the body component is

$$\tilde{S}_{\text{body}} = -\int_{\Theta} \big(1-\sum_{m\in\mathcal{A}_t} [1-C_m(\theta)]\big) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta, \tag{8}$$

where $C_i(\theta) = \mathbb{P}(\theta > \mu_i \mid \mathcal{F}_t)$ is the cumulative posterior probability of the mean of the arm i. Eq. (8) is the leading-order term of the mode of $p_{\max}$, which is mainly contributed to by the best empirical arm.

These approximations of Eq. (3) are valid when the best empirical arm has been drawn extensively compared to the worse empirical ones, corresponding to the situation encountered asymptotically for uniformly

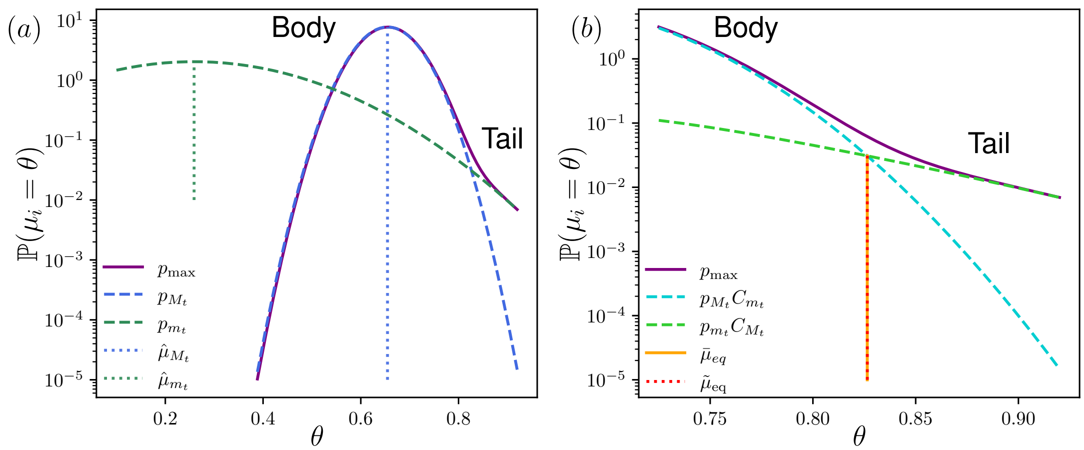

Figure 1: **(a)** Posterior distributions of a two-armed bandit with Gaussian rewards. The dotted lines are the individual posterior distributions of each arm, p_{M_t} and p_m, while the continuous line represents the posterior of the maximum mean reward of all arms, $p_{\max}$ (Eq. (3)). **(b)** Zoom of **(a)** around the point $\bar{\mu}_{\mathrm{eq},m}$ where both arms have the same posterior probability of being the best one. $p_{M_t}C_m$ ($p_m C_{M_t}$) is the probability that the maximal value is given by the better (worse) empirical arm, and $\tilde{\mu}_{\mathrm{eq},m}$ approximates $\bar{\mu}_{\mathrm{eq},m}$ given in Appendix B.6.

good algorithms. Surprisingly, the approximation captured by Eqs. (7) and (8) is still accurate enough outside this asymptotic regime to provide a high-performance decision scheme.

The previous expressions hold independently of the reward distribution, and general approximations can be derived for exponential family bandits. Yet, for Gaussian reward distributions, Eqs. (7) and (8) can be computed exactly (see Appendix B.4) providing a more convenient closed-form expression of $\tilde{S}_{\max}$. Finally, we aim to obtain a closed-form expression of Eq. (5), which drives the decision process. To achieve a simplified expression, we focus solely on its asymptotic form (see Appendix B.5 for derivation details). This leads to a comparison of the expected entropy increments when choosing a worse empirical arm or the best empirical arm:

$$
\begin{aligned}
\Delta_{M_t,k} = & \frac{1}{2}\ln\Big(\frac{N_{M_t}}{N_{M_t}+1}\Big) \\
& + \frac{1}{2N_{M_t}} \min\left(\frac{1}{2}\sum_{m}^{\mathcal{A}_t} \mathrm{erfc}\left(\delta\tilde{\mu}_{\mathrm{eq},m}\right), 1 - \frac{1}{K}\right) \\
& + Q(N_k^{-1}, \ln(N_{M_t}), \delta\tilde{\mu}_{\mathrm{eq},k}) e^{-\delta\tilde{\mu}^2_{\mathrm{eq},k}} \\
& + \sum_{m}^{\mathcal{A}_t} P(N_m^{1/2}, N_{M_t}^{-1}, \ln(N_{M_t}), \delta\tilde{\mu}_{\mathrm{eq},m}) e^{-\delta\tilde{\mu}^2_{\mathrm{eq},m}}
\end{aligned}
\tag{9}
$$

where Q and P are polynomials given in Appendix B.6 and $\delta\tilde{\mu}_{\mathrm{eq},i} = \frac{\sqrt{N_i}(\tilde{\mu}_{\mathrm{eq},i} - \hat{\mu}_i)}{\sqrt{2\sigma^2}}$ are standardized variables with $\tilde{\mu}_{\mathrm{eq},i}$ given in Appendix B.6. More generally, $\Delta_{M_t,k}$ approximates the difference

$$
\begin{aligned}
\Delta_{M_t,k} \approx \; & \mathbb{E}\left[S_{\max}(t+1) \mid \mathcal{F}_t, a_{t+1} = M_t\right] \\
& - \mathbb{E}\left[S_{\max}(t+1) \mid \mathcal{F}_t, a_{t+1} = k\right],
\end{aligned}
$$

which is directly related to greedily maximizing the entropy decrease, described in Eq. (5). The decision process can be summarized as follows: if $\Delta_{M_t,k}$ is negative for all $k \in \mathcal{A}_t$, the better empirical arm is chosen as it reduces the most the expected value of the approximate entropy. Inversely, if at least one value $\Delta_{M_t,k}$ is positive, the arm k maximizing $\Delta_{M_t,k}$ is chosen.

In conclusion, we have derived an approximate, analytical expression for the information available about the maximum expected reward of all arms. This allows us to evaluate the expected information gain with an efficient numerical implementation by eliminating numerical integrals, substantially improving the computational speed and enabling an efficient implementation of the information maximization strategy. In Reddy et al. (2016), the decision process was not tractable, as it requires a numerical evaluation of Eq. (5) at each step, making a rigorous proof attempt infeasible. Similarly, the algorithm proposed in Barbier-Chebbah et al. (2023) lacks a final increment expression such as Eq. (9), requiring additional computation steps and hindering both theoretical guarantees and tractability. Consequently, those algorithms could not ensure that the best arm is sufficiently pulled without incurring excessive regret. As shown in Section 4, we provide such a guarantee, confirming that the asymptotic regime is reached with high probability. Even though the final derivation Eq. (9) is based on Gaussian rewards, the resulting algorithm performs well for any sub-Gaussian distribution. More importantly, it demonstrates that retaining only the dominant terms of $S_{\max}$ and making rough approximations still yields a valid algorithm. This last observation calls for extending this design principle, which subsequently derives an information-based strategy with the entropy increment approximated by its leading order, to more complicated bandit settings. We now provide the full implementation of AIM and bound its regret in the next section.

3.3 Approximate information maximization algorithm

The pseudo-code for the AIM algorithm is presented in Alg. 1 below.

Algorithm 1: AIM for K Gaussian arms

Draw each arm once, observe reward $X_t(t)$ and initialize the statistics $\hat{\mu}_t = X_t(t)$
for $t = K+1$ **to** T **do** // Arm selection
  $M_t \leftarrow \text{argmax}_{k\in[K]}\, \hat{\mu}_k$
  $\mathcal{A}_t = [K] \setminus \{M_t\}$
  if $N_{M_t} \leq \textit{argmax}_{k\in\mathcal{A}_t} N_k$ **then** $a_t \leftarrow M_t$
  else
    Evaluate $m = \text{argmax}_{k\in\mathcal{A}_t \neq M_t} \Delta_{M_t,k}$ following Eq. (9)
    if $\Delta_{M_t,m} \leq 0$ **then** $a_t \leftarrow M_t$
    else $a_t \leftarrow m$
    Pull a_t and observe $X_t(a_t)$
  $\hat{\mu}_{a_t} \leftarrow \frac{\hat{\mu}_{a_t} N_{a_t} + X_t(a_t)}{N_{a_t}+1}$, $N_{a_t} \leftarrow N_{a_t} + 1$

The best empirical arm is drawn by default if there exists at least one worse empirical arm m that has been drawn more frequently (i.e., if $N_{M_t} \leq N_m$). In such a case, both entropy components in Eq. (6) are mainly contributed to by M_t.

4 Regret bound

This section provides theoretical guarantees on the performance of AIM. More precisely, Theorem 1 below states that AIM is asymptotically optimal on the multi-armed bandit problem with sub-Gaussian rewards.

Theorem 1. *For σ^2 sub-Gaussian reward distributions, the regret of AIM satisfies for any mean vector $\boldsymbol{\mu} \in \mathbb{R}^K$*

$$\limsup_{T\to\infty} \frac{R(T)}{\ln(T)} \leq \sum_{k,\mu_k<\mu^*} \frac{2\sigma^2}{\mu^* - \mu_k},$$

where $\mu^* = \max_{k\in[K]} \mu_k$.

With Gaussian rewards, the asymptotic regret of AIM thus exactly reaches the lower bound of Lai et al. (1985) given by Eq. (2). A non-asymptotic version of Theorem 1 is given by Theorem 2 in Appendix C. We briefly sketch the proof idea below and refer to Appendix C for the complete proof.

Sketch of the proof. We assume for sake of clarity in this sketch that $\mu_1 > \mu_k$ for any $k \geq 2$. The structure of the proof is similar to the one found in Kaufmann et al. (2012a). In particular, the first main step shows that the optimal arm is pulled at least $\sqrt{t}$ times with high probability. This result holds because otherwise, the contribution of arm 1 to the tail of the distribution would dominate the contribution of other arms in the approximate information. In that case, pulling the first arm would naturally lead to a larger decrease in entropy, which ensures that the optimal arm is always pulled a significant amount of times.

Then, we only need to work in the asymptotic regime where arm 1 is pulled at least $\sqrt{t}$ times and we aim at bounding the number of pulls on the arm $k \geq 2$. Additionally, we restrict ourselves to a large number (in $\log(T)$) of pulls on arm k and automatically count the pulls before that point in the regret. As a consequence, we can show that with high probability:

$$\hat{\mu}_{M_t} \geq \mu^* - \sqrt{\frac{6\sigma^2 \ln t}{\sqrt{t}}} \quad \text{and} \quad \hat{\mu}_k \leq \mu_k + \varepsilon$$

for some arbitrary $\varepsilon > 0$. An important property of entropy is that it approximates the behavior of the bound of Lai et al. (1985). More precisely, in the asymptotic regime, the difference of the entropy increments behaves as

$$\Delta_{M_t,k} \approx -\frac{1}{2N_{M_t}} + Q(N_k) e^{-\frac{N_k(\mu_1-\mu_k)^2}{2\sigma^2}} + \sum_{i\neq M_t} P(N_i) e^{-\frac{N_i(\mu_1-\mu_i)^2}{2\sigma^2}}, \tag{10}$$

where Q_k and P_i are polynomials that also depend on extra variables (see Eq. (9)). Manipulating these polynomial terms altogether is intricate, but we can still show that if the arm k is pulled, this means the term $e^{-\frac{N_k(\mu_1-\mu_k)^2}{2\sigma^2}}$ dominates the other exponents in the sum of Eq. (10). This then implies that N_k is of order at most $\frac{2\sigma^2 \ln T}{(\mu_1-\mu_k)^2}$, as arm k is only pulled if $\Delta_{M_t,k} \geq 0$. □

Our policy is deterministic at each time step while displaying a logarithmic regret, showing that intuitions from Russo and Van Roy (2014) of linear regrets for stationary (in the sense they only depend on the posterior distribution) deterministic algorithms was inexact.

Moreover, our regret bound is frequentist, in opposition to the Bayesian regret bound obtained for IDS (Russo and Van Roy, 2014).[3] As a consequence, AIM does not need a well-specified prior: using a uniform prior, as done in our work, is a well-suited choice. Also, the required form of the entropy for the proof is general. The algorithm yields an optimal regret as long as we are guaranteed that the optimal arm is pulled a significant amount of times with high probability and that the asymptotic regime behaves as Eq. (10). Hence, Theorem 1 will hold for a large family of entropy approximations (and likely for generalizations to free energies too, as in Masson, 2013) as long as the approximation is accurate enough to not yield trivial behaviors in the short time regime. Additionally, the approximate framework devised here allows fine-tuning the formulas to improve short-time performance all the while ensuring asymptotic optimality by keeping the correct asymptotic terms.

5 Experiments

This section investigates the empirical performance of AIM (Alg. 1) on numerical examples. All details of the numerical experiments are given in Appendix E.

We start by considering two arms with Gaussian rewards (Honda and Takemura, 2010) of unit variance and means μ_k drawn uniformly from $[0, 1]$. Fig. 2 compares the Bayesian regret (i.e., the regret averaged over all values of (μ_1, μ_2) in $[0, 1] \times [0, 1]$) of Alg. 1 with the state-of-the-art algorithms UCB-tuned, Thompson sampling, Thompson sampling+, KLUCB++, and MED (Kaufmann et al., 2012b; Pilarski et al., 2021; Cappé et al., 2013; Jin et al., 2022; Honda and Takemura, 2011; Ménard and Garivier, 2017). We refer to Appendix E.4 for an overview and detailed descriptions of these bandit algorithms. The Bayesian regret of AIM empirically scales as $\log(T)$. Its long-time performance matches Thompson sampling, as implied by Theorem 1, while relying on a (conditionally) deterministic decision process. Additionally, AIM outperforms Thompson sampling at both short and intermediate times (see Appendix E.5.3 for finer measurements). AIM particularly outperforms Thompson sampling when the arms are difficult to distinguish due to their mean rewards being close (see examples in Appendix E.5.1 with single instance regret experiments).

AIM yields strong performance in both two-armed Gaussian and 50-armed Gaussian rewards case, as predicted by our theoretical analysis. We now aim to extend our method to other bandit settings. Fig. 3 presents the performance of AIM when adapted to Bernoulli rewards (Pilarski et al., 2021) with arm means drawn uniformly in $[0, 1]$. This adaptation is described in detail in Section 6 below. The performance of AIM is comparable to Thompson sampling here. Additionally, AIM performs comparably to Thompson sampling for close mean rewards (see Appendix E.5.2). Additionally, for 50 arms with Bernoulli rewards AIM's short-time efficiency is comparable to Thompson sampling, and it is significantly more efficient at intermediary times while showing the same logarithmic scaling at long times as Thompson sampling.

Hence, our algorithm shows strong empirical performances compared to state-of-the-art baselines for both Bernoulli and Gaussian rewards while providing outstanding effectiveness when facing multiple arms with Bernoulli rewards. Experiments suggest that AIM displays the same typical worst-case regret as Thompson sampling (which is minimax optimal up to $\sqrt{\ln K}$ for sub-Gaussian rewards), but proving a theoretical distribution-free bound remains challenging and left for future work. Of note, similar observations are drawn in Appendices E.5.1 and E.5.2 for non-Bayesian versions of the regret, with fixed bandit instances.

Taken altogether, our results show that taking physics-based approximations to get a tractable and simplified entropy-based functional still leads to a high-performance algorithm. These observations support the robustness of AIM and call for the extension of its design principles to more complex bandit settings.

6 Extensions

We applied our information maximisation approach to Bernoulli bandits both with two and with many arms, where it shows strong empirical performances (see Fig. 3 above). This section describes the extensions of AIM to this case and discusses potential extensions to more general bandit settings.

Exponential family bandits. Since Eq. (3) explicitly relies on the arms' posterior distributions, information maximization methods can be directly extended to various reward distributions. In particular, when the reward distributions belong to the exponential family (see Korda et al., 2013, and Appendix D.1 for details on such distributions), an asymptotic and analytical expression of the entropy can be derived for the case of uniform priors (see Appendix D for more

[3] Kirschner et al. (2021a) later proposed an adaptation of IDS that has a frequentist, instance dependent optimal regret bound. The development of such an adaptation however required a long investigation and significant changes in the original IDS algorithm.

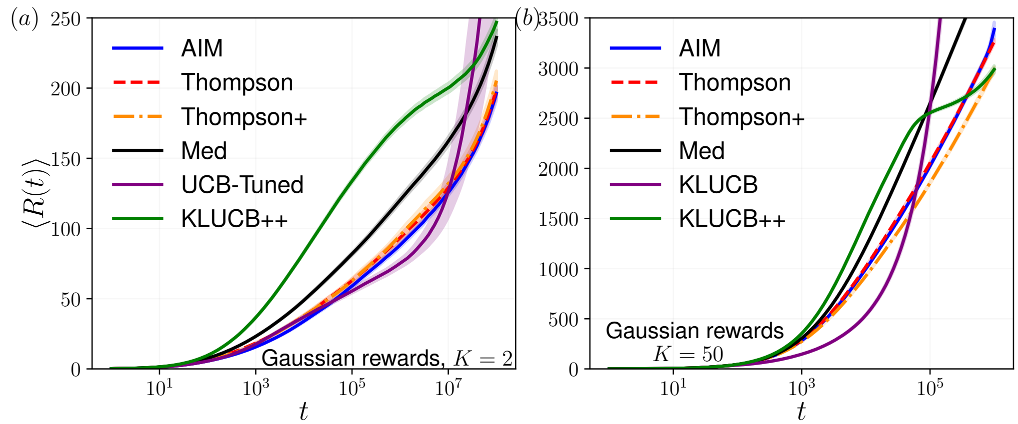

Figure 2: Evolution of the Bayesian regret for (a) 2-armed and (b) 50-armed bandit with Gaussian rewards under a uniform mean prior. Regret is averaged over 8000 for (a) and 2000 runs for (b) Confidence intervals show the standard deviation.

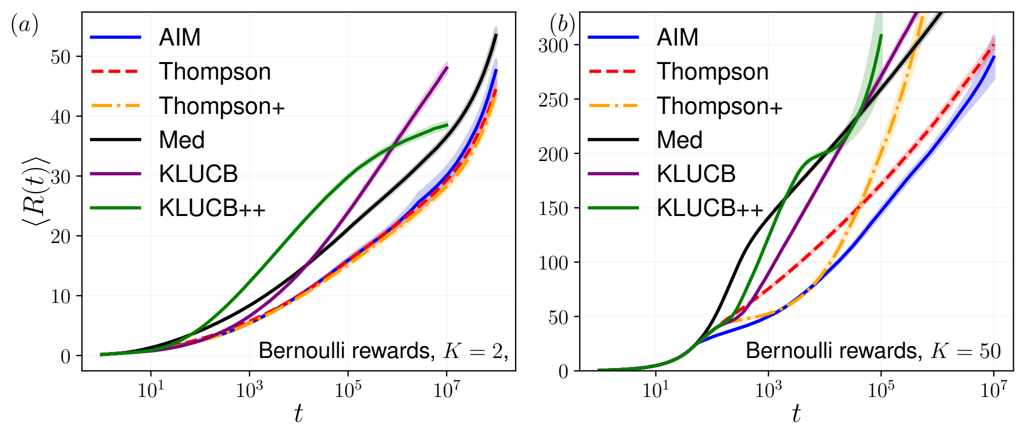

Figure 3: Evolution of the Bayesian regret for (a) 2-armed and (b) 50-armed bandit with Bernoulli rewards under a uniform mean prior. The regret is averaged over 16000 runs for (a) and 2000 runs for (b). Confidence intervals show the standard deviation.

details), yielding

$$\tilde{S}_{\max} = \frac{1}{2}\ln(2\pi\bar{\sigma}_i^2)\left[1 - \frac{e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})}}{N_m \partial_2 \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})\sqrt{2\pi\bar{\sigma}_m^2}}\right] + \frac{\mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}}) e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})}}{\partial_2 \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})\sqrt{2\pi\bar{\sigma}_m^2}}. \tag{11}$$

Here $\mathrm{KL}(\hat{\theta}_i, \tilde{\theta}_{\mathrm{eq}})$ is the Kullback-Leibler divergence between the reward distribution parameterized by $\hat{\theta}_i$ and $\tilde{\theta}_{\mathrm{eq}}$ where θ is the family parameter, and $\partial_2\mathrm{KL}$ denotes its derivative w.r.t. the second variable. All the steps leading to Eqs. (7) and (8) in Section 3 are not specific to Gaussian rewards. The main difference lies in their asymptotic simplifications obtained afterwards with Laplace's method which differ from Barbier-Chebbah et al. (2023), providing a coherent generalization for exponential family bandits. Our implementation of AIM to Bernoulli rewards (a specific case of the exponential family) with Eq. (11) shows comparable performance to state-of-the-art algorithms (see Fig. 2), supporting its adaptability to general settings. We believe that AIM should be optimal for all exponential family reward distributions and general prior distributions and that similar proof techniques can be used (see Appendix D for a detailed discussion). However, significant work still remains to ensure that the asymptotic regime, where all arms have been sufficiently drawn, is reached for any reward distribution and will be addressed in future work.

Other bandit settings. We provide a quick overview of several other bandit settings for which approximate information maximization, adapted to the specific bandit problem, should provide efficient algorithms. Here, we briefly discuss the entropy functional that needs to be selected based on the problem at hand. The design of specific approximations will be addressed in future work. First, we emphasize that AIM's partitioning between body and tail components remains relevant even when dealing with heavy-tailed (Lee et al., 2023) or non-parametric reward distributions (Baudry et al., 2020). It should thus be able to provide strong guarantees in these settings, similarly to Thompson sampling. Secondly, information can also be quantified for unpulled arms, which is crucial when handling many arms. Especially if the agent is aware of the remaining time, it can assess the maximum information gain possible from pulling an untested arm. Such consideration may be pivotal when facing many arms with a limited amount of time (Bayati et al., 2020). Thirdly, in linear bandits, where arms are correlated with each other (Li et al., 2010), the shared information gain could be leveraged by information-based methods to yield strong performances. To prevent over-exploration of sub-optimal arms, particularly those in negative directions that still provide useful information, we propose using distinct entropy functionals depending on the arm being evaluated. Finally, information maximization is an inherent candidate for pure exploration problems (Bubeck et al., 2011; Locatelli et al., 2016; Kalyanakrishnan et al., 2012) where the agent's goal is directly linked to an information gain. Preliminary results on all these problems indicate that our design principle is promising, but theoretical guarantees are required to confirm its effectiveness.

7 Conclusion

This paper introduces a new algorithm class, approximate information maximization (AIM), which leverages the approximate entropy of the whole bandit system to achieve optimal regret performances. This approach builds on the entropy of the posterior of the arms' maximal mean, from which we extract a simplified and analytical functional at the core of the decision scheme. It enables easily tunable and tractable algorithms, which we prove to be optimal for multi-armed Gaussian bandits. Numerical experiments for Bernoulli rewards with two or several arms emphasize the robustness and efficiency of AIM. An additional strength of AIM lies in its efficiency at short times and when the arms have close mean rewards where it outperforms existing state of the art. Further research should focus on adjusting the information maximization framework to more complex bandit settings, including many-armed bandits, linear bandits, and thresholding bandits, where appropriately selected information measures can efficiently apprehend the games' structure and correlations.

Acknowledgments. We acknowledge the help of the HPC Core Facility of the Institut Pasteur for this work. This study was funded by the INCEPTION project (PIA/ANR-16-CONV-0005), the *"Investissements d'avenir"* program managed by *Agence Nationale de la Recherche*, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), and *Agence Nationale de la Recherche* ANR-20-CE45-0021.

References

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135. PMLR, May 2013.

P. Auer. Using upper confidence bounds for online learning. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 270–279, Redondo Beach, CA, USA, 2000. IEEE Comput. Soc. ISBN 978-0-7695-0850-4. doi: 10.1109/SFCS.2000.892116.

Alex Barbier-Chebbah, Christian L. Vestergaard, and Jean-Baptiste Masson. Approximate information for efficient exploration-exploitation strategies, July 2023.

Dorian Baudry, Emilie Kaufmann, and Odalric-Ambrym Maillard. Sub-sampling for efficient non-parametric bandit exploration. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 5468–5478, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

Mohsen Bayati, Nima Hamidi, Ramesh Johari, and Khashayar Khosravi. Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. *Advances in Neural Information Processing Systems*, 33:1713–1723, 2020.

Guy Bresler, George H Chen, and Devavrat Shah. A latent source model for online collaborative filtering. *Advances in neural information processing systems*, 27, 2014.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412 (19):1832–1852, April 2011. ISSN 0304-3975. doi: 10.1016/j.tcs.2010.12.059.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, June 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1119.

Ring T. Cardé. Navigation Along Windborne Plumes of Pheromone and Resource-Linked Odors. *Annual Review of Entomology*, 66(1):317–336, 2021. doi: 10.1146/annurev-ento-011019-024932.

Arnoud V. Den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The Statistical Complexity of Interactive Decision Making, July 2023.

Aurélien Garivier. Informational confidence bounds for self-normalized averages and applications. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2013.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*, volume 970 of *Lecture Notes in Physics*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-46443-1 978-3-030-46444-8. doi: 10.1007/978-3-030-46444-8.

José Miguel Hernández-Lobato, Michael A. Gelbart, Matthew W. Hoffman, Ryan P. Adams, and Zoubin Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1699–1707, Lille, France, July 2015. JMLR.org.

Junya Honda and Akimichi Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 67–79. Omnipress, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Mach Learn*, 85 (3):361–391, December 2011. ISSN 1573-0565. doi: 10.1007/s10994-011-5257-4.

Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Anima Anandkumar. Finite-Time Regret of Thompson Sampling Algorithms for Exponential Family Multi-Armed Bandits. *Advances in Neural Information Processing Systems*, 35:38475–38487, December 2022.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC Subset Selection in Stochastic Multi-armed Bandits. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1, January 2012.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 199–213, Berlin, Heidelberg, 2012b. Springer. ISBN 978-3-642-34106-9. doi: 10.1007/978-3-642-34106-9_18.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42, January 2016. ISSN 1532-4435.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018a.

Johannes Kirschner and Andreas Krause. Information Directed Sampling and Bandits with Heteroscedastic Noise. In *Proceedings of the 31st Conference On Learning Theory*, pages 358–384. PMLR, July 2018b.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021a.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvari. Asymptotically Optimal Information-Directed Sampling. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2777–2821. PMLR, July 2021b.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pages 1448–1456, Red Hook, NY, USA, December 2013. Curran Associates Inc.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Jongyeong Lee, Junya Honda, Chao-Kai Chiang, and Masashi Sugiyama. Optimality of thompson sampling with noninformative priors for pareto bandits. *arXiv preprint arXiv:2302.01544*, 2023.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the Thresholding Bandit Problem. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1690–1698. PMLR, June 2016.

Dominique Martinez, Lotfi Arhidi, Elodie Demondion, Jean-Baptiste Masson, and Philippe Lucas. Using Insect Electroantennogram Sensors on Autonomous Robots for Olfactory Searches. *JoVE (Journal of Visualized Experiments)*, (90):e51704, August 2014. ISSN 1940-087X. doi: 10.3791/51704.

Jean-Baptiste Masson. Olfactory searches with limited space perception. *Proceedings of the National Academy of Sciences*, 110(28):11261–11266, July 2013. doi: 10.1073/pnas.1221091110.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, October 2017.

John Murlis, Joseph S. Elkinton, and Ring T. Cardé. Odor Plumes and How Insects Use Them. *Annual Review of Entomology*, 37(1):505–532, January 1992. doi: 10.1146/annurev.en.37.010192.002445.

Edward W. Ng and Murray Geller. A table of integrals of the Error functions. *J. RES. NATL. BUR. STAN. SECT. B. MATH. SCI.*, 73B(1):1, January 1969. ISSN 0098-8979. doi: 10.6028/jres.073B.001.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior.* The MIT Press, March 2022. ISBN 978-0-262-36997-8. doi: 10.7551/mitpress/12441.001.0001.

Sebastian Pilarski, Slawomir Pilarski, and Dániel Varró. Optimal Policy for Bernoulli Bandits: Computation and Algorithm Gauge. *IEEE Transactions on Artificial Intelligence*, 2(1):2–17, February 2021. ISSN 2691-4581. doi: 10.1109/TAI.2021.3074122.

Gautam Reddy, Antonio Celani, and Massimo Vergassola. Infomax Strategies for an Optimal Balance Between Exploration and Exploitation. *Journal of Statistical Physics*, 163(6):1454–1476, April 2016. doi: 10.1007/s10955-016-1521-0.

Gautam Reddy, Venkatesh N. Murthy, and Massimo Vergassola. Olfactory Sensing and Navigation in Turbulent Environments. *Annual Review of Condensed Matter Physics*, 13:191–213, March 2022. doi: 10.1146/annurev-conmatphys-031720-032754.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling. In Z. Ghahramani, M. Welling, C. Cortes,

N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Ilya O. Ryzhov, Warren B. Powell, and Peter I. Frazier. The Knowledge Gradient Algorithm for a General Class of Online Learning Problems. *Operations Research*, 60(1):180–195, February 2012. ISSN 0030-364X. doi: 10.1287/opre.1110.0999.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, January 2016. ISSN 1476-4687. doi: 10.1038/nature16961.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, April 2000.

Massimo Vergassola, Emmanuel Villermaux, and Boris I. Shraiman. 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126): 406–409, January 2007. ISSN 1476-4687. doi: 10.1038/nature05464.

Yunbei Xu and Assaf Zeevi. Bayesian Design Principles for Frequentist Sequential Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38768–38800. PMLR, July 2023.

Siqi Zhang, Dominique Martinez, and Jean-Baptiste Masson. Multi-Robot Searching with Sparse Binary Cues and Limited Space Perception. *Frontiers in Robotics and AI*, 2, 2015. ISSN 2296-9144.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, the setting is detailed in Section 2, all algorithms and the specific numerical settings of each experiment are detailed in Appendix E. Extension to more complex bandit settings can be found in Section 6 and Appendix D
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, we provide a short description on the compute resources and time execution in Appendix E.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, in the supplementary material, we provide commented code implementing our algorithm for bandit games with more than two arms for both Bernoulli and Gaussian rewards. For Gaussian rewards, other baseline algorithms are also implemented.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, the setting is detailed in Section 2, we give a detailed proof in Appendix C and a brief overview of the proof Section 4
   (b) Complete proofs of all theoretical results. Yes, the full proof is given in the supplementary material Appendix C
   (c) Clear explanations of any assumptions. Yes, the setting is detailed in Section 2, we give a detailed proof in Appendix C and a brief overview of the proof Section 4 as it requirements.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, in the supplementary material, we provide commented code implementing our algorithm for bandit games with more than two arms for both Bernoulli and Gaussian rewards. All algorithms and the specific numerical settings of each experiment are detailed in Appendix E
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, all algorithms and the specific numerical settings of each experiment are detailed in Appendix E.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, we provide a short description on the error bars and statistics in Section 2 as in each figure caption. We provide a clear description of the process to draw the arm means and compare the main results with additional experiments involving more realizations.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, we provide a short description on the compute resources in Appendix E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Appendix

Table of Contents

A Contrast with information-directed sampling **14**

B Towards an analytical approximation of the entropy **15**
B.1 The partitioning approximation 15
B.2 Asymptotics of the intersection point 16
B.3 Closed-form expressions for the main mode's contribution 17
B.4 Closed form and asymptotic expression for the tail's entropy 18
B.5 Derivation of the increment for the closed-form expression of entropy 18
B.6 Final expression for the increment comparison 21

C Proof of Theorem 1 **21**
C.1 Auxiliary Lemmas 23

D Generalization of the information maximization approximation **28**
D.1 Asymptotic expression for exponential family rewards 28
D.2 The partitioning approximation 29
D.3 Asymptotic intersection point 29
D.4 Generalization of the main mode's contribution 30
D.5 Generalized expression for the entropy tail 30
D.6 Generalized form of the entropy approximation 31
D.7 Derivation of the increment for the closed-form expression of entropy 31

E Numerical experiments **31**
E.1 Numerical settings 31
E.2 AIM implementation details 32
E.3 Information maximization approximation for Bernoulli rewards with more than two arms 34
E.4 Overview of baseline bandit algorithms 35
E.5 Additional experiments 36

A Contrast with information-directed sampling

Here, we briefly recapitulate the main similarities and dissimilarities that AIM shares with IDS.

We aim to explore information-based design principles, since we are convinced of their potential, especially on more structured problems, as already illustrated by Russo's seminal work Russo and Van Roy (2014). Hence, this work shares a lineage with IDS and, more generally, with information-based decision algorithms (infotaxis, the information bottleneck, etc.). Our work complements previous information-based methods Kirschner and Krause (2018b); Xu and Zeevi (2023); Foster et al. (2023) in MAB by:

1) Not requiring a "free energy" formulation that balances exploration with instantaneous regrets. Designing a regret-minimizing "pure exploration" based strategy is one of the main challenges of our work. This avoids irrelevant information acquisition by focusing on a global observable and is the origin of the following critical differences with Russo and Van Roy (2014) and other related methods.

2) AIM policy is deterministic at each time step and shows that Russo's intuition of linear regrets in stationary (in the sense they only depend on the posterior distribution) deterministic algorithms was inexact (see Example 1 of their arXiv version).

3) The simplification scheme developed above leads to an efficient numerical implementation by eliminating numerical integrals, substantially improving computational speed, a crucial challenge for information methods, as also stressed in Russo's paper.

4) AIM regret bounds are frequentist, independent of the priors, and instance-dependent, compared to Bayesian regret and distribution-free in Russo and Van Roy (2014). To our knowledge, instance-dependent regret bounds

have only been obtained in Kirschner et al. (2021b) for a version of an information-based method.

B Towards an analytical approximation of the entropy

In this section, we recapitulate all the steps leading to the analytical expression constitutive of our AIM algorithm. We stress that it involves exact derivations but also simplifications to considerably simplify the final form of AIM. Therefore, alternative approaches could lead to a slightly different version of AIM. However, our chosen method retains the essential features that emerge in the asymptotic regime while providing a simple version of AIM.

B.1 The partitioning approximation

We start by commenting on the partition scheme and the approximations leading to the following body/tail expressions. We first recall the expression for $p_{\max}(\theta)$, with the arms ordered along M_t,

$$p_{\max}(\theta) = \left(p_{M_t}(\theta) \prod_{m}^{\mathcal{A}_t} C_m(\theta) + \sum_{m}^{\mathcal{A}_t} C_{M_t}(\theta) p_m(\theta) \prod_{j \neq m}^{\mathcal{A}_t} C_j(\theta) \right). \tag{12}$$

where we remind that $C_m(\theta)$ is the cumulative posterior probability of the mean of arm m

Because of its dependency along all arms, there is no unique dominant term in Eq. (12), and distinct regimes emerge depending on θ and the state of the game. Then, we assume to isolate distinct regimes contributing asymptotically to the entropy while significantly simplifying them. It will considerably simplify the derivation of an analytical expression for the body/tail components in the next section. The next paragraphs will then present heuristic arguments justifying our simplification scheme.

We start by rewriting the exact entropy expression isolating M_t:

$$\begin{aligned} S_{\max} = &- \int_{\Theta} p_{M_t}(\theta) \prod_{m}^{\mathcal{A}_t} C_j(\theta) \ln \left(p_{M_t}(\theta) \prod_{j}^{\mathcal{A}_t} C_j(\theta) + \sum_{m}^{\mathcal{A}_t} p_m(\theta) C_{M_t}(\theta) \prod_{j \neq m}^{\mathcal{A}_t} C_j(\theta) \right) \mathrm{d}\theta \\ &- \sum_{m}^{\mathcal{A}_t} \int_{\Theta} p_i(\theta) C_{M_t}(\theta) \prod_{j \neq m}^{\mathcal{A}_t} C_j(\theta) \ln \left(p_{M_t}(\theta) \prod_{j}^{\mathcal{A}_t} C_j(\theta) + \sum_{m}^{\mathcal{A}_t} p_m(\theta) C_{M_t}(\theta) \prod_{j \neq m}^{\mathcal{A}_t} C_j(\theta) \right) \mathrm{d}\theta. \end{aligned} \tag{13}$$

Let us briefly comment on the different contributions to Eq. (13). We aim to keep the leading orders of $p_{\max}(\theta)$ when $N_{M_t} \gg N_m \gg 1$ and $\hat{\mu}_{M_t} > \hat{\mu}_m$ for all m in the set of current worse empirical arms $\mathcal{A}_t$. Here, the posterior distributions are assumed uni-modal. The first term is the leading order in the vicinity of the mode of $\hat{\mu}_{M_t}$. Also, since $N_{M_t} > N_m$, $p_{M_t}(\theta)$ is more concentrated than all $p_m(\theta)$, resulting in the dominance of the second term in the distribution's tail (i.e., for high rewards).

We now decompose the entropy in the body/tail components defined in the main text, the first term of Eq. (13) will form the body component, and the sum over all worse empirical arms will compose the tail. We now define $\bar{\mu}_{\mathrm{eq},m}$ the intersection point associated to the arm m verifying $C_m(\bar{\mu}_{\mathrm{eq},m}) p_{M_t}(\bar{\mu}_{\mathrm{eq},m}) = C_{M_t}(\bar{\mu}_{\mathrm{eq},m}) p_m(\bar{\mu}_{\mathrm{eq},m})$. Then, in the asymptotic regime, $\bar{\mu}_{\mathrm{eq},m}$ will verify $p_m(\theta) \gg p_{M_t}(\theta)$ for $\theta > \bar{\mu}_{\mathrm{eq},m}$ and $p_m(\theta) \ll p_{M_t}(\theta)$ for $\theta < \bar{\mu}_{\mathrm{eq},m}$. Again, we will assume to neglect the transition regime where $\bar{\mu}_{\mathrm{eq},m} \sim \theta$ where both distributions are of the same order because it is narrow (in the asymptotic regime) and has very little influence on the total value of the entropy.

To get the body component, we consider the first term of Eq. (13). We neglect all the inner terms inside the logarithm which is then dominated by M_t. Next, by noticing that $C_i(\theta) \approx 1$ is in the vicinity of $\hat{\mu}_{M_t}$, we make a first-order expansion of the remaining product along all the worse empirical arms. Since the inner term of the body component is negligible for $\theta > \min(\{\bar{\mu}_{\mathrm{eq},m}, m \in \mathcal{A}_t\})$ (because of its dependency along p_{M_t}), we ignore that our simplification is no more valid in this specific regime without loss of consistency. Taken together we obtain the body expression of the main text:

$$\tilde{S}_{\mathrm{body}} = - \int_{\Theta} \big(1 - \sum_{m}^{\mathcal{A}_t} [1 - C_m(\theta)]\big) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta. \tag{14}$$

Then, we consider the additional terms (each denoted as m) in Eq. (13). First, each term of the sum is negligible to the first one for $\theta < \bar{\mu}_{\mathrm{eq},m}$, we then only keep the upper part of the integral where $\theta > \bar{\mu}_{\mathrm{eq},m}$. Because $N_m \gg 1$ and $\theta > \min(\{\bar{\mu}_{\mathrm{eq},m}, i \neq M_t\}) > \hat{\mu}_{M_t} > \hat{\mu}_j$, we approximate all the cumulative by one. Finally, to get a simplified expression for the increment, we assume to neglect all the posterior distributions except for $p_i(\theta)$ inside the logarithm of the i-th term and approximate $\bar{\mu}_{\mathrm{eq},m}$ (see next section) which leads to the tail expression:

$$\tilde{S}_{\mathrm{tail}} = -\sum_{m}^{\mathcal{A}_t} \int_{\tilde{\mu}_{\mathrm{eq},m}}^{\mu_{\mathrm{sup}}} p_m(\theta) \ln p_m(\theta) \mathrm{d}\theta. \tag{15}$$

note that some of these posterior distributions ($j \neq m, M_t$) are not negligible compared to $p_m(\theta)$ at a given θ. However, this cross-information between current suboptimal arms is asymptotically negligible regarding the decision procedure (which largely resumes as balancing exploiting the best empirical solution compared to exploring worse empirical arms) while unnecessarily complicating the increment evaluation.

Finally, we obtain the full expression of the entropy approximation:

$$\tilde{S}_{\max} = -\int_{\Theta} \left(1 - \sum_{m}^{\mathcal{A}_t} [1 - C_m(\theta)]\right) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta - \sum_{m}^{\mathcal{A}_t} \int_{\tilde{\mu}_{\mathrm{eq},m}}^{\mu_{\mathrm{sup}}} p_i(\theta) \ln p_i(\theta) \mathrm{d}\theta. \tag{16}$$

B.2 Asymptotics of the intersection point

In this section, we derive the asymptotic expression of the intersection point (defined above as $\tilde{\mu}_{\mathrm{eq},m}$) where the distributions $C_m(\tilde{\mu}_{\mathrm{eq},m}) p_{M_t}(\tilde{\mu}_{\mathrm{eq},m})$ and $C_{M_t}(\tilde{\mu}_{\mathrm{eq},m}) p_m(\tilde{\mu}_{\mathrm{eq},m})$ intersect (at their highest value if they intersect more than once). Here, we consider Gaussian rewards and the intersection between M_t and a given worse empirical arm denoted m. The exact equation verified by the intersection point $\bar{\mu}_{\mathrm{eq},m}$ is:

$$\begin{aligned} \frac{\sqrt{N_{M_t}} e^{-\frac{N_{M_t}(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{1}{2} \left[1 + \mathrm{erf}\left(\frac{\sqrt{N_m}(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)\right] = \\ \frac{\sqrt{N_m} e^{-\frac{N_m(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_m)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{1}{2} \left[1 + \mathrm{erf}\left(\frac{\sqrt{N_{M_t}}(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_{M_t})}{\sqrt{2\sigma^2}}\right)\right]. \end{aligned} \tag{17}$$

Taking the logarithm of Eq. (17) and normalizing the last term leads to:

$$\frac{N_m(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_m)^2}{2\sigma^2} - \frac{N_{M_t}(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_{M_t})^2}{2\sigma^2} + \frac{1}{2} \ln \frac{N_{M_t}}{N_m} + \ln \left[\frac{1 + \mathrm{erf}\left(\frac{\sqrt{N_m}(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)}{1 + \mathrm{erf}\left(\frac{\sqrt{N_{M_t}}(\bar{\mu}_{\mathrm{eq},m} - \hat{\mu}_{M_t})}{\sqrt{2\sigma^2}}\right)}\right] = 0. \tag{18}$$

The distributions are uni-modal, and assuming that $\hat{\mu}_{M_t} > \hat{\mu}_m$, $N_{M_t} > N_m$ and recalling that $\bar{\mu}_{\mathrm{eq},m}$ is the highest intersection, we get that $\bar{\mu}_{\mathrm{eq},m} > \hat{\mu}_{M_t} > \hat{\mu}_m$. Both error functions are then bounded in $[0, 1]$, making the last term bounded as well. We then approximate $\bar{\mu}_{\mathrm{eq},m}$ with $\tilde{\mu}_{\mathrm{eq},m}$ by neglecting the last term, which leads to the following solution:

$$\tilde{\mu}_{\mathrm{eq},m} = \hat{\mu}_{M_t} + \frac{N_m(\hat{\mu}_{M_t} - \hat{\mu}_m)}{N_{M_t} - N_m} + \sqrt{\frac{N_{M_t} N_m}{(N_{M_t} - N_m)^2} (\hat{\mu}_{M_t} - \hat{\mu}_m)^2 + \frac{\sigma^2}{N_{M_t} - N_m} \ln\left(\frac{N_{M_t}}{N_m}\right)}. \tag{19}$$

Note that Eq. (19) relies on both $\hat{\mu}_{M_t} > \hat{\mu}_m$ and $N_{M_t} > N_m$. For $N_{M_t} \leq N_m$, even if the above $\tilde{\mu}_{\mathrm{eq},m}$ can be computed, it does not quantify the tail contribution. As a matter of fact, for $N_{M_t} \leq N_m$, the tail is always dominated by p_{M_t}, which means that it has already been included in the main mode $\tilde{S}_{\mathrm{body}}$. Then, in this specific configuration, we take the contribution of the arm m to $\tilde{S}_{\mathrm{tail}}$ equals to 0 (in other words $\tilde{\mu}_{\mathrm{eq},m} = \mu_{\mathrm{sup}}$).

B.3 Closed-form expressions for the main mode's contribution

Here, we derive the $\tilde{S}_{\text{body}}$ expression given in the main text for Gaussian rewards distribution. Inserting the Gaussian form of the posterior into Eq. (8) gives:

$$\begin{aligned}\tilde{S}_{\text{body}} = -\int_{-\infty}^{+\infty} \frac{\sqrt{N_{M_t}}e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left(-\frac{1}{2}\ln(\frac{2\pi\sigma^2}{N_{M_t}}) - \frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}\right) \times \\ \left(1 - \sum_m^{\mathcal{A}_t} \frac{1}{2}\left[1 - \text{erf}\left(\frac{\sqrt{N_m}(\theta-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)\right]\right) \mathrm{d}\theta,\end{aligned} \tag{20}$$

We integrate the constant part of the first term, denoted T_1 by the use of the following identity (Ng and Geller, 1969):

$$\int_{-\infty}^{\infty} \left[1 + \text{erf}\left(\frac{\theta-\theta_1}{\sqrt{2V_1}}\right)\right] \frac{e^{-\frac{(\theta-\theta_2)^2}{2V_2}}}{\sqrt{2\pi V_2}} \mathrm{d}\theta = \left[1 + \text{erf}\left(\frac{\theta_2-\theta_1}{\sqrt{2}\sqrt{V_2+V_1}}\right)\right], \tag{21}$$

which leads to

$$\begin{aligned}T_1 &= \frac{1}{2}\ln\left(\frac{2\pi\sigma^2}{N_{M_t}}\right) \int_{-\infty}^{+\infty} \frac{\sqrt{N_{M_t}}e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left[1 - \frac{1}{2}\sum_m^{\mathcal{A}_t} 1 - \text{erf}\left(\frac{\sqrt{N_m}(\theta-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)\right] \mathrm{d}\theta \\ &= \frac{1}{2}\ln\left(\frac{2\pi\sigma^2}{N_{M_t}}\right)\left(1 - \sum_m^{\mathcal{A}_t}\frac{1}{2}\left[1 - \text{erf}\left(\frac{\sqrt{N_m}(\hat{\mu}_{M_t}-\hat{\mu}_m)}{\sqrt{2\sigma^2(\frac{1}{N_{M_t}}+\frac{1}{N_m})}}\right)\right]\right) \\ &= \frac{1}{2}\ln\left(\frac{2\pi\sigma^2}{N_{M_t}}\right)\left(1 - \sum_m^{\mathcal{A}_t}\frac{1}{2}\text{erfc}\left[\frac{\sqrt{N_m}(\hat{\mu}_{M_t}-\hat{\mu}_m)}{\sqrt{2\sigma^2(\frac{1}{N_{M_t}}+\frac{1}{N_m})}}\right]\right)\end{aligned} \tag{22}$$

Next, we separate the second term in two parts $T_{2,1}$ and $T_{2,2}$, first :

$$T_{2,1} = \int_{-\infty}^{+\infty} \frac{\sqrt{N_{M_t}}e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}\mathrm{d}\theta = \frac{1}{2}. \tag{23}$$

Then, we integrate by parts the remaining term $T_{2,2}$ to obtain:

$$\begin{aligned}T_{2,2} &= -\sum_m^{\mathcal{A}_t}\int_{-\infty}^{\infty} \frac{N_{M_t}^{3/2}(\theta-\hat{\mu}_{M_t})^2}{4\sigma^2} \frac{e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left[1 - \text{erf}\left(\frac{\sqrt{N_m}(\theta-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)\right] \mathrm{d}\theta \\ &= -\sum_m^{\mathcal{A}_t}\int_{-\infty}^{\infty} \frac{1}{4}\frac{\sqrt{N_{M_t}}e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left[1 - \text{erf}\left(\frac{\sqrt{N_m}(\theta-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)\right] \\ &\qquad - \frac{(\theta-\hat{\mu}_{M_t})}{2}\frac{\sqrt{N_{M_t}N_m}e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}-\frac{N_m(\theta-\hat{\mu}_m)^2}{2\sigma^2}}}{2\pi\sigma^2}\mathrm{d}\theta \\ &= -\sum_{M_t}^{\mathcal{A}_t}\frac{1}{4}\text{erfc}\left(\frac{\hat{\mu}_{M_t}-\hat{\mu}_m}{\sqrt{2\sigma^2(\frac{1}{N_{M_t}}+\frac{1}{N_m})}}\right) - \frac{(\hat{\mu}_{M_t}-\hat{\mu}_m)\sigma^2}{2N_{M_t}\sqrt{2\pi}(\frac{\sigma^2}{N_{M_t}}+\frac{\sigma^2}{N_m})^{3/2}}e^{-\frac{(\hat{\mu}_{M_t}-\hat{\mu}_m)^2}{2\sigma^2(\frac{1}{N_{M_t}}+\frac{1}{N_m})}},\end{aligned} \tag{24}$$

where we also rely on the identity of Eq. (21).

Combining Eqs. (22) to (24) leads to the analytical expression of the body component.

$$\tilde{S}_{\text{body}} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1 - \frac{1}{2}\sum_m^{\mathcal{A}_t} \text{erfc}\left(\frac{\sqrt{N_m N_{M_t}}(\hat{\mu}_{M_t} - \hat{\mu}_m)}{\sqrt{2\sigma^2(N_m + N_{M_t})}}\right)\right] - \sum_m^{\mathcal{A}_t} \frac{\sqrt{N_{M_t}} N_m^{3/2}(\hat{\mu}_{M_t} - \hat{\mu}_m)}{2\sigma\sqrt{2\pi}(N_{M_t} + N_m)^{3/2}} e^{-\frac{N_m N_{M_t}(\hat{\mu}_{M_t} - \hat{\mu}_m)^2}{2\sigma^2(N_m + N_{M_t})}}. \tag{25}$$

To finally get an asymptotic and simplified expression of the body component, we neglect the second term. Then, since $\tilde{\mu}_{\text{eq},m} \underset{N_{M_t}\to\infty}{\longrightarrow} \hat{\mu}_{M_t}$ and $N_m \ll N_{M_t}$ asymptotically, we approximate the first term as:

$$\tilde{S}_{\text{b}} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1 - \frac{1}{2}\sum_m^{\mathcal{A}_t} \text{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\text{eq},m} - \hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)\right]. \tag{26}$$

This last approximation will enable us to provide an analytically tractable gradient without altering the asymptotic behavior expected at large times for the entropy measure.

B.4 Closed form and asymptotic expression for the tail's entropy

The contribution from the tail can be derived exactly and reads:

$$\begin{aligned}\tilde{S}_{\text{tail}} &= \sum_m^{\mathcal{A}_t} \int_{\tilde{\mu}_{\text{eq},m}}^{\infty} \frac{\sqrt{N_m} e^{-\frac{N_m(\theta - \hat{\mu}_m)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left[\frac{1}{2}\ln(\frac{2\pi\sigma^2}{N_m}) + \frac{N_m(\theta - \hat{\mu}_m)^2}{2\sigma^2}\right] \mathrm{d}\theta \\ &= \sum_m^{\mathcal{A}_t} \frac{1}{4}\ln\left(\frac{2\pi\sigma^2 e}{N_m}\right) \text{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\text{eq},m} - \hat{\mu}_m)}{\sqrt{2\sigma^2}}\right) \\ &\qquad + \frac{\sqrt{N_m}(\tilde{\mu}_{\text{eq},m} - \hat{\mu}_m)}{2\sqrt{2\pi\sigma^2}} e^{-\frac{N_m(\tilde{\mu}_{\text{eq},m} - \hat{\mu}_m)^2}{2\sigma^2}}.\end{aligned} \tag{27}$$

To get a simplified analytical expression of the tail component, we only keep the second term since it dominates the others asymptotically,

$$\tilde{S}_{\text{t}} = \sum_m^{\mathcal{A}_t} \frac{\sqrt{N_m}(\tilde{\mu}_{\text{eq},m} - \hat{\mu}_m)}{2\sqrt{2\pi\sigma^2}} e^{-\frac{N_m(\tilde{\mu}_{\text{eq},m} - \hat{\mu}_m)^2}{2\sigma^2}}. \tag{28}$$

Taken altogether, Eqs. (26) and (28) lead to the desired simplified approximation of the entropy:

$$\tilde{S}_{\text{m}} = \tilde{S}_{\text{b}} + \tilde{S}_{\text{t}}. \tag{29}$$

B.5 Derivation of the increment for the closed-form expression of entropy

Since Eqs. (26) and (28) exhibit simple closed-form expressions, it becomes possible to derive an explicit expression of its expected increment. Here, we again consider continuous Gaussian reward distributions.

We start by deriving the increment along the better empirical arm, $\Delta_{M_t}\tilde{S}_{\text{m}}$. The posterior of the reward obtained at time $t+1$ is approximated as a Gaussian of variance σ^2 and centred around $\hat{\mu}_{M_t}$, leading to:

$$\Delta_{M_t}\tilde{S}_{\text{m}} = \int_{-\infty}^{\infty} \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left[\tilde{S}_{\text{m}}(\hat{\mu}_{M_t} + \frac{\mu}{N_{M_t}+1}, N_{M_t}+1, ...) - \tilde{S}_{\text{m}}(\hat{\mu}_{M_t}, N_{M_t}, ...)\right] \mathrm{d}\mu. \tag{30}$$

where the dots runs over all the worse empirical arms variables remaining constant when the best empirical arm is drawn at time $t+1$.

For the sake of simplicity, we neglect the variations of all the subdominant terms inside all $\tilde{\mu}_{\mathrm{eq},m}$ meaning we approximate them as $\tilde{\mu}_{\mathrm{eq},m}(\hat{\mu}_{M_t}+\frac{\mu}{N_{M_t}+1}, N_{M_t}+1, \hat{\mu}_m, N_m) \approx \tilde{\mu}_{\mathrm{eq},m}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_m, N_m)+\frac{\mu}{N_{M_t}+1}$, after observing a reward μ when pulling the arm M_t for the $(N_{M_t}+1)$th time.

By use of the identity Eq. (21), the gradient of the body component $\Delta_{M_t}\tilde{S}_{\mathrm{b}}$ can be rewritten as:

$$\begin{aligned}\Delta_{M_t}\tilde{S}_{\mathrm{b}} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}+1})\left[1-\frac{1}{2}\sum_m^{\mathcal{A}_t}\mathrm{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{\sqrt{2\sigma^2}\sqrt{1+\frac{N_m}{(N_{M_t}+1)^2}}}\right)\right]\\ -\frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1-\frac{1}{2}\sum_m^{\mathcal{A}_t}\mathrm{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right)\right].\end{aligned} \tag{31}$$

The increment of the tail component along the better empirical arm can be calculated as:

$$\begin{aligned}\Delta_{M_t}\tilde{S}_{\mathrm{t}} &= \sum_m^{\mathcal{A}_t}\int_{-\infty}^{\infty}\frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\frac{\sqrt{N_m}(\frac{\mu}{N_{M_t}+1}+\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_m(\tilde{\mu}_{\mathrm{eq},m}+\frac{\mu}{N_{M_t}+1}-\hat{\mu}_m)^2}{2\sigma^2}}\mathrm{d}\mu\\ &\qquad -\sum_m^{\mathcal{A}_t}\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_m(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}}\\ &= \sum_m^{\mathcal{A}_t} e^{-N_m\frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2\left(1+\frac{N_m}{(1+N_{M_t})^2}\right)}}\sqrt{\frac{N_m}{8\pi\sigma^2}}\frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{(1+\frac{N_m}{(N_{M_t}+1)^2})^{3/2}}\\ &\qquad -\sum_m^{\mathcal{A}_t}\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_m(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}}.\end{aligned} \tag{32}$$

Next, we consider the increment evaluation along a given worse empirical arm denoted by k,

$$\Delta_k\tilde{S}_{\mathrm{m}} = \int_{-\infty}^{\infty}\frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\left[\tilde{S}_{\mathrm{m}}(...,\hat{\mu}_k+\frac{\mu}{N_k+1}, N_k+1,...)-\tilde{S}_{\mathrm{m}}(...,\hat{\mu}_k,N_k,...)\right]\mathrm{d}\mu. \tag{33}$$

We here also neglect the variations of the subdominant term inside $\tilde{\mu}_{\mathrm{eq},m}$. We start by considering the increment of the body component:

$$\begin{aligned}\Delta_k\tilde{S}_{\mathrm{b}} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1-\frac{1}{2}\mathrm{erfc}\left(\frac{(N_k+1)(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{\sqrt{2\sigma^2(N_k+2)}}\right)\right]\\ -\frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1-\frac{1}{2}\mathrm{erfc}\left(\frac{\sqrt{N_k}(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{\sqrt{2\sigma^2}}\right)\right].\end{aligned} \tag{34}$$

Of note, all other terms in the sum independent of index k remain constant, showing no increment. Finally, we consider the associated tail component of the increment along k:

$$\begin{aligned}\Delta_k\tilde{S}_{\mathrm{t}} &= \int_{-\infty}^{\infty}\frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\frac{\sqrt{N_k+1}(\frac{\mu}{N_k+1}+\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{(N_k+1)(\tilde{\mu}_{\mathrm{eq},k}+\frac{\mu}{N_k+1}-\hat{\mu}_k)^2}{2\sigma^2}}\mathrm{d}\mu\\ &\qquad -\frac{\sqrt{N_k}(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_k(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\\ &= e^{-\frac{(N_k+1)^2}{(N_k+2)}\frac{(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\frac{(1+N_k)^2(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{\sqrt{8\pi\sigma^2}(2+N_k)^{3/2}}\\ &\qquad -\frac{\sqrt{N_k}(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_k(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\end{aligned} \tag{35}$$

As for the body increment, all other terms in the sum independent of index k remain constant, showing no increment.

Taken altogether, Eqs. (31), (32), (34) and (35) lead to the final analytical expression of the increment:

$$
\begin{aligned}
\Delta_{M_t,k} =& \frac{1}{2}\ln(\frac{N_{M_t}}{N_{M_t}+1}) - \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}+1})\sum_m^{\mathcal{A}_t}\operatorname{erfc}\left[\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{\sqrt{2\sigma^2(1+\frac{N_m}{(N_{M_t}+1)^2})}}\right] \\
&+ \frac{1}{4}\ln(\frac{2\pi\sigma^2}{N_{M_t}})\sum_m^{\mathcal{A}_t}\operatorname{erfc}\left[\sqrt{N_m}\frac{\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m}{\sqrt{2\sigma^2}}\right] \\
&+ \sum_m^{\mathcal{A}_t} e^{-N_m\frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2(1+\frac{N_m}{(1+N_{M_t})^2})}}\sqrt{\frac{N_m}{8\pi\sigma^2}}\frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{(1+\frac{N_m}{(N_{M_t}+1)^2})^{3/2}} \\
&- \sum_m^{\mathcal{A}_t}\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{\sqrt{8\pi\sigma^2}}e^{-\frac{N_m(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}} \\
&+ \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[\operatorname{erfc}\left(\frac{(N_k+1)(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{\sqrt{2\sigma^2(2+N_k)}}\right) - \operatorname{erfc}\left(\sqrt{N_k}\frac{\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k}{\sqrt{2\sigma^2}}\right)\right] \\
&- e^{-\frac{(N_k+1)^2}{(N_k+2)}\frac{(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\frac{(1+N_k)^2\,(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{\sqrt{8\pi\sigma^2}(2+N_k)^{3/2}} + \frac{\sqrt{N_k}(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_k(\tilde{\mu}_{\mathrm{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}
\end{aligned}
\tag{36}
$$

To obtain a simplified expression, we expand to the first order each component of the different components of Eq. (36) denoted T_1, T_2, T_3, T_4. The former is given by:

$$
\begin{aligned}
T_1 =& -\frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}+1})\sum_m^{\mathcal{A}_t}\operatorname{erfc}\left[\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{\sqrt{2\sigma^2(1+\frac{N_m}{(N_{M_t}+1)^2})}}\right] \\
&\qquad + \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\sum_{m\neq M_t}^{K}\operatorname{erfc}\left[\sqrt{N_m}\frac{\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m}{\sqrt{2\sigma^2}}\right] \\
\approx& \sum_m^{\mathcal{A}_t}\frac{1}{4N_{M_t}}\operatorname{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right) \\
&\qquad - \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\frac{N_m}{N_{M_t}^2}\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{\sqrt{2\pi\sigma^2}}e^{-\frac{N_m(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}}.
\end{aligned}
\tag{37}
$$

Next we consider the second component, which reads:

$$
\begin{aligned}
T_2 =& \sum_m^{\mathcal{A}_t} e^{-N_m\frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2(1+\frac{N_m}{(1+N_{M_t})^2})}}\sqrt{\frac{N_m}{8\pi\sigma^2}}\frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{(1+\frac{N_m}{(N_{M_t}+1)^2})^{3/2}} \\
&\qquad - \sum_m^{\mathcal{A}_t}\frac{\sqrt{N_m}(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_m(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}} \\
\approx& \sum_m^{\mathcal{A}_t} e^{-N_m\frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}}\sqrt{\frac{N_m}{8\pi\sigma^2}}\,(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)\left[-\frac{3}{2}\frac{N_m}{N_{M_t}^2} + \frac{(\tilde{\mu}_{\mathrm{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}\frac{N_m^2}{N_{M_t}^2}\right].
\end{aligned}
\tag{38}
$$

Next we consider the third term, denoted T_3, which reads:

$$
\begin{aligned}
T_3 &= \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[\text{erfc}\left(\frac{(N_k+1)(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)}{\sqrt{2\sigma^2(2+N_k)}}\right)-\text{erfc}\left(\sqrt{N_k}\frac{\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k}{\sqrt{2\sigma^2}}\right)\right] \\
&\approx -\frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\frac{1}{N_k^2}\left(\frac{\sqrt{N_k}(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)}{\sqrt{2\pi\sigma^2}}\right)e^{-\frac{N_k(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}
\end{aligned} \tag{39}
$$

Finally, the last term T_4 reads

$$
\begin{aligned}
T_4 &= -e^{-\frac{(N_k+1)^2}{(N_k+2)}\frac{(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\frac{(1+N_k)^2(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)}{\sqrt{8\pi\sigma^2}(2+N_k)^{3/2}}+\frac{\sqrt{N_k}(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{N_k(\tilde{\mu}_{\text{eq}}-\hat{\mu}_k)^2}{2\sigma^2}} \\
&\approx e^{-N_k\frac{(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\sqrt{\frac{N_k}{8\pi\sigma^2}}(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)\left[\frac{1}{N_k}+\frac{1}{N_k}\frac{(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}\right].
\end{aligned} \tag{40}
$$

Taken altogether, we finally obtain the following simplified increment :

$$
\begin{aligned}
\widetilde{\Delta}_{M_t,k} &= \frac{1}{2}\ln(\frac{N_{M_t}}{N_{M_t}+1})+\frac{1}{2N_{M_t}}\sum_m^{\mathcal{A}_t}\frac{1}{2}\text{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right) \\
&+\sum_m^{\mathcal{A}_t}\frac{N_m^{3/2}(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)}{\sqrt{2\pi\sigma^2}N_{M_t}^2}e^{-N_m\frac{(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}}\left[\frac{1}{4}\ln(\frac{N_{M_t}}{2\pi\sigma^2 e})-\frac{3}{4}+\frac{N_m(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)^2}{4\sigma^2}\right] \\
&+\frac{\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k}{\sqrt{2\pi\sigma^2 N_k}}e^{-N_k\frac{(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\left[\frac{1}{4N_k}\ln(\frac{N_{M_t}}{2\pi\sigma^2 e})+\frac{1}{2}+\frac{(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{4\sigma^2}\right]
\end{aligned} \tag{41}
$$

By noticing that the sum of the second term should account for the tail contribution along the body increment, it shouldn't be allowed to be superior to one. Then we assume to bound it by taking the minimum compared to $1-1/K$.

B.6 Final expression for the increment comparison

Taken altogether, it leads to the expression used for AIM for multiple Gaussian arms:

$$
\begin{aligned}
\Delta_{M_t,k} &= \frac{1}{2}\ln(\frac{N_{M_t}}{N_{M_t}+1})+\frac{1}{2N_{M_t}}\min\left(\sum_m^{\mathcal{A}_t}\frac{1}{2}\text{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)}{\sqrt{2\sigma^2}}\right),1-\frac{1}{K}\right) \\
&+\sum_m^{\mathcal{A}_t}\frac{N_m^{3/2}(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)}{\sqrt{2\pi\sigma^2}N_{M_t}^2}e^{-N_m\frac{(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)^2}{2\sigma^2}}\left[\frac{1}{4}\ln(\frac{N_{M_t}}{2\pi\sigma^2 e})-\frac{3}{4}+\frac{N_m(\tilde{\mu}_{\text{eq},m}-\hat{\mu}_m)^2}{4\sigma^2}\right] \\
&+\frac{\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k}{\sqrt{2\pi\sigma^2 N_k}}e^{-N_k\frac{(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{2\sigma^2}}\left[\frac{1}{4N_k}\ln(\frac{N_{M_t}}{2\pi\sigma^2 e})+\frac{1}{2}+\frac{(\tilde{\mu}_{\text{eq},k}-\hat{\mu}_k)^2}{4\sigma^2}\right]
\end{aligned} \tag{42}
$$

with

$$
\tilde{\mu}_{\text{eq},i} = \hat{\mu}_{M_t}+\frac{N_i(\hat{\mu}_{M_t}-\hat{\mu}_i)}{N_{M_t}-N_i}+\sqrt{\frac{N_{M_t}N_i(\hat{\mu}_{M_t}-\hat{\mu}_i)^2}{(N_{M_t}-N_i)^2}+\frac{\sigma^2\ln\left(\frac{N_{M_t}}{N_i}\right)}{N_{M_t}-N_i}}. \tag{43}
$$

C Proof of Theorem 1

This section provides the complete proof of Theorem 1. More precisely, it proves the more refined Theorem 2 below.

Theorem 2. *For any multi-armed bandits with σ^2 sub-Gaussian rewards and mean vector $\boldsymbol{\mu} \in \mathbb{R}^K$, for any $\varepsilon \in (0, \frac{1}{2})$, there exists a constant $C(\boldsymbol{\mu}, \varepsilon) \in \mathbb{R}$ depending solely on $\boldsymbol{\mu}$ and ε such that for any $T \in \mathbb{N}$*

$$R(T) \leq \sum_{k, \mu_k < \mu^*} \left[\frac{2\sigma^2 \ \ln T}{(1-\varepsilon)(\mu^* - \mu_k)} + \frac{2\sigma^2 \ \ln\ln T}{(1-\varepsilon)(\mu^* - \mu_k)} \right] + C(\boldsymbol{\mu}, \varepsilon).$$

Proof. We denote in the whole proof $\mathcal{M}^* = \{k \in [K] \mid \mu_k = \mu^*\}$. For $\Delta_k = \mu^* - \mu_k$, the regret can then be written as

$$R(T) = \sum_{k, \Delta_k > 0} \Delta_k \mathbb{E}\left[N_k(T)\right].$$

We decompose this expectation in 4 terms as follows

$$\begin{aligned}\mathbb{E}\left[N_k(T)\right] \leq \sum_{t=1}^{T} \mathbb{P}\left(\forall i \in \mathcal{M}^*, N_i(t) \leq \sqrt{t}\right) + \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_k(t) \geq \mu^* - \sqrt{\frac{6\sigma^2 \ln t}{\sqrt{t}}}, a_t = k\right) \\ + \sum_{t=1}^{T} \mathbb{P}\left(\exists i \in \mathcal{M}^*, \hat{\mu}_i(t) \leq \mu_i - \sqrt{\frac{6\sigma^2 \ln t}{N_i(t)}}\right) + \sum_{t=1}^{T} \mathbb{P}\left(\mathcal{E}_k(t)\right),\end{aligned}$$

where

$$\mathcal{E}_k(t) \coloneqq \left\{ \exists i \in \mathcal{M}^*, N_i(t) \geq \sqrt{t} \text{ and } \hat{\mu}_i(t) \geq \mu^* - \sqrt{\frac{6\sigma^2 \ln t}{N_i(t)}} \geq \hat{\mu}_k(t), a_t = k \right\}.$$

This inequality comes simply by noticing the event $\{a_t = k\}$ is included in the union of the 4 other events. Lemmas 1, 3 and 4 allow to respectively bound the first, second and third sums by a constant $C(\boldsymbol{\mu})$ depending solely on $\boldsymbol{\mu}$, so that

$$\mathbb{E}\left[N_k(T)\right] \leq \sum_{t=1}^{T} \mathbb{P}\left(\mathcal{E}_k(t)\right) + C(\boldsymbol{\mu}).$$

Thanks to Lemma 5, there exist constants $t(\boldsymbol{\mu}), n(\boldsymbol{\mu})$ depending solely on K and Δ_k such that

$$\sum_{t=1}^{T} \mathbb{P}\left(\mathcal{E}_k(t)\right) \leq t(\boldsymbol{\mu}) + \sum_{t=1}^{T} \mathbb{P}\left(\mathcal{G}_1(t)\right) + \mathbb{P}\left(\mathcal{G}_2(t)\right),$$

where

$$\begin{aligned}\mathcal{G}_1(t) &= \{\mu_k - \hat{\mu}_k(t) \leq -\varepsilon\Delta_k, a_t = k\}, \\ \mathcal{G}_2(t) &= \{N_k(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_k^2}\big(\ln t + \ln\ln t\big) + n(\boldsymbol{\mu}), a_t = k\}.\end{aligned}$$

Now, we bound individually the sum corresponding to each of these 2 events. The first one can be bounded using Hoeffding's inequality. Indeed, for independent random variables $Z_k(n) \sim \mathcal{N}(\mu_k, \sigma^2)$, it reads as:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}\left(\mu_k - \hat{\mu}_k(t) \leq -\varepsilon\Delta_k, a_t = k\right) &\leq \sum_{n=1}^{T} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_k(i) - \mu_k \geq \varepsilon\Delta_k\right) \\ &\leq \sum_{n=1}^{T} e^{-\frac{n\varepsilon^2\Delta_k^2}{2\sigma^2}} \\ &\leq \frac{1}{e^{\frac{\varepsilon^2\Delta_k^2}{2\sigma^2}} - 1}.\end{aligned}$$

The bound of the second term is bounded as

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}(\mathcal{E}_k(t)) &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{N_k(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_k^2}\big(\ln t + \ln\ln t\big) + n(\boldsymbol{\mu}), a_t = k\}\right] \\ &\leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_k^2}\big(\ln T + \ln\ln T\big) + n(\boldsymbol{\mu}) + 1.\end{aligned}$$

Wrapping up everything finally yields that for some constant $C(\boldsymbol{\mu}, \varepsilon)$ depending solely on $\boldsymbol{\mu}, \varepsilon$,

$$R(T) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2 \Delta_k} (\ln T + \ln \ln T) + C(\boldsymbol{\mu}, \varepsilon).$$

This concludes the proof of Theorem 2 with the reparameterization $\varepsilon \leftarrow 1 - (1-2\varepsilon)^2$. □

C.1 Auxiliary Lemmas

Similarly to the proof of Thompson sampling, the first part of the proof shows that the optimal arm is at least pulled a polynomial number of times with high probability. We recall that we denote in this whole section $\mathcal{M}^* = \arg\max_k \mu_k$.

Lemma 1. *There exists a constant $C_0(\boldsymbol{\mu})$ depending solely on the mean vector $\boldsymbol{\mu}$ such that*

$$\sum_{t=1}^{\infty} \mathbb{P}(\forall i \in \mathcal{M}^*, N_i(t) \leq \sqrt{t}) \leq C_0(\boldsymbol{\mu}).$$

Proof. Let $t_0(\boldsymbol{\mu})$ be a large constant that depends solely on $\boldsymbol{\mu}$. In the remaining of the proof, we assume at some points that $t_0(\boldsymbol{\mu})$ is chosen large enough (but only larger than a threshold depending on $\boldsymbol{\mu}$) such that some inequalities hold. We also assume in the following, without loss of generality, that $\mu_1 = \mu^*$, i.e., $1 \in \mathcal{M}^*$.

Assume that for $t \geq t_0(\boldsymbol{\mu})$, $N_i(t) \leq \sqrt{t}$ for all $i \in \mathcal{M}^*$. Let then k be the most pulled arm at time t, i.e., $k \in \operatorname{argmax}_j N_j(t)$ (if multiple arms maximise the number of pulls, we select the one such that its last pull happened the earliest). Necessarily $N_k(t) \geq \frac{t}{K}$. We can choose $t_0(\boldsymbol{\mu})$ large enough so that $\frac{t}{K} > \sqrt{t}$ and thus $\Delta_k > 0$. Let $t' \leq t$ be the last time k was pulled. By design, k also maximised the number of pulls then, so that $k \in \operatorname{argmax}_j \hat{\mu}_j(t')$. Moreover, $N_i(t') \leq \sqrt{t}$ for all $i \in \mathcal{M}^*$ and $N_k(t') \geq \frac{t}{K} - 1$. For $t_0(\boldsymbol{\mu})$ large enough, this yields $N_k(t') \geq N_i(t')$ for all $i \in \mathcal{M}^*$ and $a_{t'} = k$. The arm k is thus pulled at time t', in particular because $S_k \geq S_1$ (i.e., $\Delta_{k,1} S \leq 0$), where

$$S_k = \frac{1}{2} \ln\left(1 + \frac{1}{N_k(t')}\right) - \frac{1}{2N_k(t')} \min\left(\frac{1}{2}\sum_{i \neq k} \operatorname{erfc}\left(\frac{\sqrt{N_i(t')}(\tilde{\mu}_{\mathrm{eq},i} - \hat{\mu}_i)}{\sqrt{2\sigma^2}}\right), 1 - \frac{1}{K}\right),$$

$$S_1 = \sum_{i \neq k} \sqrt{\frac{N_i(t')}{2\pi\sigma^2}} (\tilde{\mu}_{\mathrm{eq},i} - \hat{\mu}_i) e^{-N_i(t') \frac{(\tilde{\mu}_{\mathrm{eq},i} - \hat{\mu}_i)^2}{2\sigma^2}} \left[\frac{1}{4} \ln\left(\frac{N_k(t')}{2\pi\sigma^2 e}\right) \frac{N_i(t')}{N_k^2(t')} - \frac{3}{4} \frac{N_i(t')}{N_k^2(t')} + \frac{(\tilde{\mu}_{\mathrm{eq},i} - \hat{\mu}_i)^2}{4\sigma^2} \frac{N_i^2(t')}{N_k^2(t')}\right]$$
$$+ e^{-N_1(t') \frac{(\tilde{\mu}_{\mathrm{eq},1} - \hat{\mu}_1)^2}{2\sigma^2}} \sqrt{\frac{N_1(t')}{2\pi\sigma^2}} (\tilde{\mu}_{\mathrm{eq},1} - \hat{\mu}_1) \left[\frac{1}{4} \ln\left(\frac{N_k(t')}{2\pi\sigma^2 e}\right) \frac{1}{N_1^2(t')} + \frac{1}{2N_1(t')} + \frac{1}{N_1(t')} \frac{(\tilde{\mu}_{\mathrm{eq},1} - \hat{\mu}_1)^2}{4\sigma^2}\right].$$

To simplify, note that $S_k \leq \frac{1}{2N_k(t')}$. Moreover since $N_k(t') \geq \frac{t}{K} - 1 \geq 2\pi e^4 \sigma^2$ for a large enough choice of $t_0(\boldsymbol{\mu})$, S_1 can be easily lower bounded as

$$S_1 \geq \frac{1}{2} \frac{(\tilde{\mu}_{\mathrm{eq},1} - \hat{\mu}_1)}{\sqrt{2\pi\sigma^2 N_1(t')}} e^{-\frac{N_1(t')(\tilde{\mu}_{\mathrm{eq},1} - \hat{\mu}_1)^2}{2\sigma^2}}.$$

So we finally have the following inequality at time t':

$$\frac{1}{N_2(t')} \geq \frac{(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_1)}{\sqrt{2\pi\sigma^2 N_1(t')}} e^{-\frac{N_1(t')(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_1)^2}{2\sigma^2}}. \tag{44}$$

Recall that $N_2(t') \geq \frac{t}{K} - 1$, so that Eq. (44) can be rewritten as

$$N_1(t') \geq \left(\frac{t}{K} - 1\right) \frac{\tilde{x}}{\sqrt{\pi}} e^{-\tilde{x}^2}, \tag{45}$$

where $\tilde{x} = \frac{\sqrt{N_1(t')}(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_1)}{\sqrt{2\sigma^2}}$. In the following, we will show that $\tilde{x} \in [\tilde{x}_{\min}, \tilde{x}_{\max}] \subset \mathbb{R}_+$. By analysing the variations of $x \mapsto x e^{-x^2}$, this will imply that

$$N_1(t') \geq \frac{\frac{t}{K} - 1}{\sqrt{\pi}} \min\{\tilde{x}_{\min} e^{-\tilde{x}_{\min}^2}, \tilde{x}_{\max} e^{-\tilde{x}_{\max}^2}\}. \tag{46}$$

For the lower bound, the definition of $\tilde{\mu}_{\mathrm{eq},1}$ and the fact that $N_1(t') \geq 1$ directly implies that

$$\tilde{x} \geq \sqrt{\frac{\ln(\frac{N_k(t')}{N_1(t')})}{2(\frac{N_k(t')}{N_1(t')} - 1)}} = \Omega\left(\sqrt{\frac{\ln(t)}{t}}\right).$$

Moreover, by subadditivity of the square root:

$$\tilde{x} \leq \sqrt{\frac{N_1(t')}{2\sigma^2}}(\hat{\mu}_2 - \hat{\mu}_1)\left(1 + \frac{N_1(t') + \sqrt{N_1(t')N_k(t')}}{N_k(t') - N_1(t')}\right) + \sqrt{\frac{N_1(t')\ln(\frac{N_k(t')}{N_1(t')})}{2(N_k(t') - N_1(t'))}} \tag{47}$$

$$\leq \sqrt{\frac{N_1(t')}{2\sigma^2}}(\hat{\mu}_k - \hat{\mu}_1)\left(1 + Kt^{-\frac{1}{4}}\right) + \mathcal{O}\left(\frac{\sqrt{K\ln(t)}}{t^{\frac{1}{4}}}\right). \tag{48}$$

Let us now consider the events, for $\Delta_{\min} = \min_{j,\Delta_j>0}\Delta_j$,

$$\mathcal{H}_*(t) \coloneqq \left\{\exists i \in \mathcal{M}^*, \exists s \leq t, \hat{\mu}_i(s) - \mu_i \leq -\sqrt{\frac{2\sigma^2(\ln(t) - \ln\ln(t))}{N_i(s)}} - \frac{\Delta_{\min}}{3}\right\}, \tag{49}$$

$$\mathcal{H}_k(t) \coloneqq \left\{\exists s \leq t, \frac{t}{K} - 1 \leq N_k(s) \leq t \text{ and } \hat{\mu}_k(s) - \mu_k \geq \frac{\Delta_k}{3}\right\}. \tag{50}$$

Assume in the following that $\neg\mathcal{H}_*(t) \cap \neg\mathcal{H}_k(t)$. This implies that

$$\hat{\mu}_k - \hat{\mu}_1 \leq -\frac{\Delta_k}{3} + \sqrt{\frac{2\sigma^2(\ln(t) - \ln\ln(t))}{N_1(s)}}. \tag{51}$$

In particular,

$$\sqrt{\frac{N_1(t')}{2\sigma^2}}(\hat{\mu}_2 - \hat{\mu}_1) \leq \sqrt{\ln(t) - \ln\ln(t)},$$

which implies that $\tilde{x} \leq \sqrt{\ln(t) - \ln\ln(t)} + \mathcal{O}\left(K\sqrt{\ln(t)}t^{-\frac{1}{4}}\right)$. Using the lower and upper bounds on $\tilde{x}$, we have thanks to Eq. (46) that under $\neg\mathcal{H}_*(t) \cap \neg\mathcal{H}_k(t)$,

$$N_1(t') = \Omega(\frac{\ln^{\frac{3}{2}}(t)}{K}).$$

For a large enough choice of $t_0(\boldsymbol{\mu})$, this last equality along with Eq. (51) actually yield $\hat{\mu}_k - \hat{\mu}_1 < 0$, which contradicts the beginning of the proof (k being best empirical arm at time t'). By contradiction, we thus showed the following event inclusion for $t \geq t_0(\boldsymbol{\mu})$:

$$\left\{\forall i \in \mathcal{M}^*, N_i(t) \leq \sqrt{t}\right\} \subset \mathcal{H}_*(t) \cup \mathcal{H}_k(t). \tag{52}$$

Lemma 1 then follows, thanks to Lemma 2 below,

$$\sum_{t=1}^{\infty}\mathbb{P}(\forall i \in \mathcal{M}^*, N_i(t) \leq \sqrt{t}) \leq t_0(\boldsymbol{\mu}) + \sum_{t=t_0(\boldsymbol{\mu})+1}^{\infty}\mathbb{P}(\mathcal{H}_*(t)) + \mathbb{P}(\mathcal{H}_k(t)).$$

□

Lemma 2. *For any $b \in (0,1)$ and the events $\mathcal{H}_*(t), \mathcal{H}_k(t)$ defined in Eqs. (49) and (50), there exist constants c_1 and c_2 depending solely on $\boldsymbol{\mu}$ such that*

$$\sum_{t=1}^{\infty}\mathbb{P}(\mathcal{H}_*(t)) \leq c_1 \quad \text{and} \quad \sum_{t=1}^{\infty}\mathbb{P}(\mathcal{H}_k(t)) \leq c_2 \quad \textit{for any } k \notin \mathcal{M}^*.$$

Proof. The two bounds directly result from Hoeffding's inequality. Consider independent random variables $(Z_j(n))_{n\in N, j\in[K]}$ where $Z_j(n) \sim \mathcal{N}(\mu_j, \sigma^2)$. Let us first bound the probability of $\mathcal{H}_k(t)$, which is simpler.

$$\begin{aligned}\mathbb{P}(\mathcal{H}_k(t)) &\leq \sum_{n=\lceil t-t^b-1\rceil}^{t} \mathbb{P}\left(\sum_{i=1}^{n}(Z_k(i)-\mu_k) \geq \frac{n\Delta_k}{3}\right)\\ &\leq \sum_{n=\lceil \frac{t}{K}-1\rceil}^{t} e^{-\frac{n\Delta_k^2}{18\sigma^2}}\\ &\leq \frac{e^{-\frac{\lceil \frac{t}{K}-1\rceil\Delta_k^2}{18\sigma^2}}}{1-e^{-\frac{\Delta_k^2}{18\sigma^2}}}.\end{aligned}$$

The second inequality of Lemma 2 then follows by noting that the last term is summable over t. For the second bound, we also have by Hoeffding's inequality

$$\begin{aligned}\mathbb{P}(\mathcal{H}_*(t)) &\leq \sum_{j\in\mathcal{M}^*}\sum_{n=1}^{\infty}\mathbb{P}\left(\sum_{i=1}^{n}(Z_j(i)-\mu_j) \leq -\sqrt{2n\sigma^2(\ln(t)-\ln\ln(t))} - \frac{n\Delta_{\min}}{3}\right)\\ &\leq \sum_{j\in\mathcal{M}^*}\sum_{n=1}^{\infty}\exp\left(-\ln(t)+\ln\ln(t)-\sqrt{2n\sigma^2(\ln(t)-\ln\ln(t))}\frac{\Delta_{\min}}{3\sigma^2}-\frac{n\Delta_{\min}^2}{18\sigma^2}\right)\\ &\leq |\mathcal{M}^*|\frac{\ln(t)}{t}\exp\left(-\sqrt{2(\ln(t)-\ln\ln(t))}\frac{\Delta_{\min}}{3\sqrt{\sigma^2}}\right)\sum_{n=1}^{\infty}e^{-\frac{n\Delta_{\min}^2}{18\sigma^2}}.\end{aligned}$$

The last sum is obviously finite. Moreover, $\sqrt{2(\ln(t)-\ln\ln(t))} = \omega(\ln\ln(t))$, so that $\exp\left(-\sqrt{2(\ln(t)-\ln\ln(t))}\frac{\Delta_{\min}}{3\sqrt{\sigma^2}}\right) = \mathcal{O}\left(\frac{1}{\ln^{\alpha}(t)}\right)$ for any $\alpha > 0$. By comparison with series of the form $\frac{1}{n\ln^{\alpha}(n)}$, the term $\frac{\ln(t)}{t}\exp\left(-\sqrt{2(\ln(t)-\ln\ln(t))}\frac{\Delta_2}{3\sqrt{\sigma^2}}\right)$ is summable over t, which leads to the first bound of Lemma 2. □

Lemma 3. *For any $k \notin \mathcal{M}^*$, there exists a constant $C_1(\boldsymbol{\mu})$ depending solely on $\boldsymbol{\mu}$ such that*

$$\sum_{t=1}^{\infty}\mathbb{P}\left(\hat{\mu}_k(t) \geq \mu^* - \sqrt{\frac{6\sigma^2\ln t}{\sqrt{t}}}, a_t = k\right) \leq C_1(\boldsymbol{\mu}).$$

Proof. A union bound on the sum yields for any $T \in \mathbb{N}$

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{P}\left(\hat{\mu}_2(t) \geq \mu^* - \sqrt{\frac{6\sigma^2\ln t}{\sqrt{t}}}, a_t = k\right)\\ &\qquad\leq \sum_{t=1}^{T}\sum_{n=0}^{t}\mathbb{P}\left(\hat{\mu}_k(t) \geq \mu^* - \sqrt{\frac{6\sigma^2\ln t}{\sqrt{t}}}, N_k(t)=n, N_k(t+1)=n+1\right)\\ &\qquad\leq \sum_{n=0}^{T}\sum_{t=n}^{T}\mathbb{P}\left(\underbrace{\hat{\mu}_k(t) \geq \mu^* - \sqrt{\frac{6\sigma^2\min_{s\geq n}\ln s}{\sqrt{s}}}, N_k(t)=n, N_k(t+1)=n+1}_{:=\mathcal{G}_1(t,n)}\right).\end{aligned}$$

Now note that the $\mathcal{G}_1(t,n)$ are disjoint for different t. In particular,

$$\sum_{t=n}^{T}\mathbb{P}(\mathcal{G}_1(t,n)) = \mathbb{P}\left(\exists t\in[n,T], \hat{\mu}_k(t) \geq \mu^* - \sqrt{\frac{6\sigma^2\min_{s\geq n}\ln s}{s^b}}, N_2(t)=n\right).$$

For independent random variables $Z_k(n) \sim \mathcal{N}(\mu_k, \sigma^2)$, we have by the independence of the X_t and a_t, and then by the Hoeffding inequality:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_k(t) \geq \mu^* - \sqrt{\frac{6\sigma^2 \ln t}{\sqrt{t}}}, a_t = k\right) &\leq 1 + \sum_{n=1}^{T} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_k(i) - \mu_k) \geq \Delta_k - \sqrt{\frac{6\sigma^2 \min_{s\geq n} \ln s}{\sqrt{s}}}\right), \\ &\leq 1 + \sum_{n=1}^{T} \exp\left(-\frac{n\left(\Delta_k - \sqrt{\frac{6\sigma^2 \min_{s\geq n} \ln s}{\sqrt{s}}})\right)^2}{2\sigma^2}\right).\end{aligned}$$

Obviously, this sum can be bounded for any $T \in \mathbb{N}$ by a constant solely depending on Δ_k. □

Lemma 4. *For any $i \in [K]$, there exists a universal constant C_2 such that*

$$\sum_{t=0}^{\infty} \mathbb{P}\left(\hat{\mu}_i(t) \leq \mu_i - \sqrt{\frac{6\sigma^2 \ln t}{N_i(t)}})\right) \leq C_2.$$

Proof. This is a direct consequence of Garivier (2013), which states that for σ^2 sub-Gaussian rewards:

$$\mathbb{P}(N_i(t)\frac{(\hat{\mu}_i(t) - \mu_i)^2}{2\sigma^2} \geq (1+\alpha)\ln t) \leq 2\left\lceil \frac{\ln t}{\ln(1+\eta)} \right\rceil t^{-(1-\frac{\eta^2}{16})(1+\alpha)} \quad \text{for any } t \in \mathbb{N}^* \text{ and } \alpha, \eta > 0.$$

In particular with $\alpha = 2 = \eta$, this implies

$$\mathbb{P}\left(\hat{\mu}_i(t) \leq \mu_i - \sqrt{\frac{6\sigma^2 \ln t}{N_i(t)}}\right) \leq 2\frac{\ln(t) + 1}{\ln(3)} t^{-\frac{9}{4}}.$$

This term is obviously summable so that there exists a constant C_2 such that

$$\sum_{t=1}^{\infty} \mathbb{P}\left(\hat{\mu}_1(t) \leq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}}\right) \leq C_2.$$

Lemma 4 directly follows by the inclusion of the considered events. □

For any $k \notin \mathcal{M}^*$, Lemma 5 below gives an event inclusion for the event $\mathcal{E}_k(t)$ that we recall here,

$$\mathcal{E}_k(t) := \left\{\exists i \in \mathcal{M}^*, N_i(t) \geq \sqrt{t} \text{ and } \hat{\mu}_i(t) \geq \mu^* - \sqrt{\frac{6\sigma^2 \ln t}{N_i(t)}} \geq \hat{\mu}_k(t), a_t = k,\right\}.$$

Lemma 5. *There exist constants $t(\boldsymbol{\mu})$ and $n(\boldsymbol{\mu})$ depending solely on $\boldsymbol{\mu}$ such that for any $k \notin \mathcal{M}^*, t \geq t(\boldsymbol{\mu})$ and $\varepsilon \in (0, \frac{1}{3})$,*

$$\mathcal{E}_k(t) \subset \{\mu_k - \hat{\mu}_k(t) \leq -\varepsilon\Delta_k, a_t = k\} \cup \{N_k(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_k^2}\big(\ln t + \ln\ln t\big) + n(\boldsymbol{\mu}), a_t = k\}.$$

Proof. Assume in the following that $\mathcal{E}_k(t)$ holds for some $t \geq t(\boldsymbol{\mu})$. Let $i \in [K]$ be an arm maximising the empirical mean at time t. Necessarily $\hat{\mu}_i(t) \geq \mu^* - \sqrt{\frac{6\sigma^2 \ln t}{N_i(t)}}$. Moreover, $a_t = k$ so that i also maximises the number of pulls, in particular $N_i(t) \geq \frac{t}{K}$. Moreover, as we pull the arm k, $S_k \geq S_i$ where

$$\begin{aligned} S_i &= \frac{1}{2}\ln\left(1 + \frac{1}{N_i(t)}\right) - \frac{1}{2N_i(t)}\min\left(\frac{1}{2}\sum_{j\neq i}\text{erfc}\left(\frac{\sqrt{N_j(t)}(\tilde{\mu}_{\text{eq},j} - \hat{\mu}_j)}{\sqrt{2\sigma^2}}\right), 1 - \frac{1}{K}\right), \\ S_k &= g_k(t)Q_k(t) + \sum_{j\neq i} g_j(t)P_j(t), \end{aligned} \tag{53}$$

where for all $j \neq i$

$$g_j(t) = \sqrt{\frac{N_j(t)}{2\pi\sigma^2}}\,(\tilde{\mu}_{\text{eq},j} - \hat{\mu}_j)\,e^{-N_j(t)\frac{(\tilde{\mu}_{\text{eq},j}-\hat{\mu}_j)^2}{2\sigma^2}},$$

$$P_j(t) = \left[\frac{1}{4}\ln(\frac{N_i(t)}{2\pi\sigma^2 e})\frac{N_j(t)}{N_i^2(t)} - \frac{3}{4}\frac{N_j(t)}{N_i^2(t)} + \frac{(\tilde{\mu}_{\text{eq},j} - \hat{\mu}_j)^2}{4\sigma^2}\frac{N_j^2(t)}{N_i^2(t)}\right]$$

$$\text{and } Q_j(t) = \left[\frac{1}{4}\ln(\frac{N_i(t)}{2\pi\sigma^2 e})\frac{1}{N_j^2(t)} + \frac{1}{2N_j(t)} + \frac{1}{N_j(t)}\frac{(\tilde{\mu}_{\text{eq},j} - \hat{\mu}_j)^2}{4\sigma^2}\right].$$

Also note that as we pull the arm k, we have for any $j \leq i$,

$$g_j(t)Q_j(t) \leq g_k(t)Q_k(t). \tag{54}$$

As a consequence, we can write for any $\delta > 0$ and $\tilde{x}_j = \sqrt{\frac{N_j(t)}{2\sigma^2}}(\tilde{\mu}_{\text{eq},j} - \hat{\mu}_j)$:

$$\begin{aligned}
g_j(t)P_j(t) &\leq g_j(t)\left(\ln(t)\frac{N_j(t)}{4N_i^2(t)}\right) + \tilde{x}_j^3 e^{-\tilde{x}_j^2}\frac{N_j(t)}{2N_i^2(t)} \\
&\leq g_j(t)\left(2\ln(t) + \frac{1}{\delta}\right)\frac{N_j(t)}{2N_i^2(t)} + \frac{N_j(t)}{2N_i^2(t)}\delta,
\end{aligned}$$

where we used the fact that $\tilde{x}_j^3 e^{-\tilde{x}_j^2} \leq \frac{\tilde{x}_j e^{-\tilde{x}_j^2}}{\delta} + \delta$. Moreover, note that for $t(\boldsymbol{\mu})$ large enough, $Q_j(t) \geq \frac{1}{2N_j(t)}$. As a consequence,

$$\begin{aligned}
g_j(t)P_j(t) &= g_j(t)\left(2\ln(t) + \frac{1}{\delta}\right)\frac{N_j(t)}{2N_i^2(t)Q_j(t)}Q_j(t) + \frac{N_j(t)}{2N_i^2(t)}\delta \\
&\leq Q_j(t)g_j(t)\left(2\ln(t) + \frac{1}{\delta}\right) + \frac{K\delta}{2t} \\
&\leq Q_k(t)g_k(t)\left(2\ln(t) + \frac{1}{\delta}\right) + \frac{K\delta}{2t},
\end{aligned}$$

where the last inequality comes from Eq. (54). In particular,

$$S_k \leq K\left(2\ln(t) + \frac{1}{\delta}\right)Q_k(t)g_k(t) + \frac{K^2\delta}{2t}. \tag{55}$$

Also, $S_i \geq \frac{1}{2t} - \frac{K^2}{4t^2}$ since $N_i(t) \geq \frac{t}{K}$ and $\ln(1+x) \geq x - \frac{x^2}{2}$ for $x \in [0,1]$.

Now assume that $\mu_k - \hat{\mu}_k(t) \geq -\varepsilon\Delta_k$. It then holds

$$\begin{aligned}
\tilde{\mu}_{\text{eq},k} - \hat{\mu}_k(t) &\geq \hat{\mu}_i(t) - \hat{\mu}_k(t) \\
&\geq \Delta_k + \mu_k - \hat{\mu}_k(t) - \sqrt{\frac{6\sigma^2\ln t}{\sqrt{t}}} \\
&\geq (1-\varepsilon)\Delta_k - \sqrt{\frac{6\sigma^2\ln t}{\sqrt{t}}}.
\end{aligned}$$

Again, we can choose $t(\boldsymbol{\mu})$ large enough so that $\tilde{\mu}_{\text{eq},k} - \hat{\mu}_k(t) \geq (1-2\varepsilon)\Delta_k$. Moreover, note that the functions $x \mapsto \frac{e^{-x^2}}{x}, x \mapsto xe^{-x^2}, x \mapsto x^3e^{-x^2}$ are all decreasing on an interval of the form $[M, +\infty]$. As a consequence, we can choose $n(\boldsymbol{\mu})$ large enough so that $\frac{\sqrt{n(\boldsymbol{\mu})((1-2\varepsilon)\Delta_k)^2}}{\sqrt{2\sigma^2}} \geq M$. If $N_k(t) \geq n(\boldsymbol{\mu})$, we then have from Eq. (55), for a constant $c(K, \Delta_k)$ solely depending on K and Δ_k:

$$S_k \leq c(K,\Delta_k)e^{-\frac{N_k(t)(1-2\varepsilon)^2\Delta_k^2}{2\sigma^2}}\left[\ln t + \frac{1}{\delta}\right] + \frac{K^2\delta}{2t}.$$

The inequality $S_k \geq S_i$ then implies, thanks to the above bounds:

$$c(K, \Delta_k) e^{-\frac{N_k(t)(1-2\varepsilon)^2 \Delta_k^2}{2\sigma^2}} \left[\ln t + \frac{1}{\delta} \right] \geq \frac{1 - K^2 \delta}{2t} - \frac{K^2}{4t^2}.$$

In particular, for $\delta = \frac{1}{2K^2}$,

$$N_k(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2 \Delta_k^2} \left(\ln t + \ln \ln t + \mathcal{O}(1) \right),$$

where the $\mathcal{O}$ hides constants depending in K and Δ_k. This concludes the proof of Lemma 5 as we just shown that if $\mathcal{E}_k(t)$ holds, at least one of the two following events holds when $N_k(t) \geq n(\boldsymbol{\mu})$:

- $\mu_k - \hat{\mu}_k(t) \leq -\varepsilon \Delta_k$
- $N_k(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2 \Delta_k^2} \left(\ln t + \ln \ln t + \mathcal{O}(1) \right)$.

□

D Generalization of the information maximization approximation

In this section, we will generalize the approach derived in Appendix B to bandit settings with a reward distribution belonging to the exponential family. We will retrace all the previous steps made in Appendix B, insisting on the differences with the Gaussian reward case. We will also discuss bandit settings with non-uniform priors and with more than two arms.

D.1 Asymptotic expression for exponential family rewards

We derive an asymptotic expression for the one-dimensional canonical exponential family from which we will derive an analytical approximation of the entropy. We thus focus on a reward distribution density f with respect to some reference measure ν belonging to some one-dimensional canonical exponential family, i.e.,

$$f(x|\theta) = A(x) \exp\big(T(x)\theta - F(\theta)\big), \tag{56}$$

where F is twice differentiable and strictly convex. Additionally, let us recall that the Kullback-Leibler divergence verifies: (Korda et al., 2013):

$$\mathrm{KL}(\theta, \theta') = F(\theta') - F(\theta) - F'(\theta)(\theta' - \theta), \tag{57}$$

where $\mathrm{KL}(\theta, \theta')$ is the Kullback-Leibler divergence between the reward distribution parameterized by θ and the one parameterized by θ'.

Given a prior $\pi(\theta)$ and the reward realizations $(x_1, \ldots, x_n)$, the associated posterior distribution on θ, denoted p, reads:

$$p(\theta | x_1, .., x_n) = \frac{1}{C} \pi(\theta) \exp\left(\theta \sum_{k=1}^{n} T(x_k) - nF(\theta) \right), \tag{58}$$

where $C = \int \pi(\theta) \exp\left(\theta \sum T(x_k) - nF(\theta)\right) \mathrm{d}\theta$ is a normalization constant. Next, we derive the maximum a posteriori for the parameter θ, denoted $\hat{\theta}_l$, which verifies:

$$\sum_{i=1}^{n} T(x_i) = nF'(\hat{\theta}_l) - \frac{\pi'(\hat{\theta}_l)}{\pi(\hat{\theta}_l)}. \tag{59}$$

At this stage, we assume that there exits such $\hat{\theta}_l$ verifying Eq. (59). In practice, for a reward distribution that does not meet this criteria, one can replace $\hat{\theta}_l$ by a series $\hat{\theta}_{n,l}$ which, for sufficiently large values of n, asymptotically conforms to the aforementioned definition. For example, a Bernoulli arm that consistently fails under a uniform prior, will result in an undefined $\hat{\theta}_l$. To address this, one may redefine $\hat{\theta}_l$ such that $(1 + \sum i = 1^n T(x_i)) = (n+2)F'(\hat{\theta}_l)$, effectively replacing the empirical mean in Eq. (59) with the posterior mean.

Replacing the sum in Eq. (58) leads to

$$
\begin{aligned}
p(\theta|x_1,..,x_n) &= \frac{1}{C}\pi(\theta)\exp\left(\theta n F'(\hat{\theta}_l) - \theta\frac{\pi'(\hat{\theta}_l)}{\pi(\hat{\theta}_l)} - nF(\theta)\right)\\
&= \frac{e^{n\hat{\theta}_l F'(\hat{\theta}_l) - nF(\hat{\theta}_l)}}{C}\pi(\theta)e^{-\theta\frac{\pi'(\hat{\theta}_l)}{\pi(\hat{\theta}_l)}}e^{-n\mathrm{KL}(\hat{\theta}_l,\theta)}\\
&= \frac{1}{C_2}\pi(\theta)e^{-\theta\frac{\pi'(\hat{\theta}_l)}{\pi(\hat{\theta}_l)}}e^{-n\mathrm{KL}(\hat{\theta}_l,\theta)},
\end{aligned}
\tag{60}
$$

where C_2 also acts as a normalization constant of Eq. (60). For $n \gg 1$, the distribution concentrates in the vicinity of $\hat{\theta}_l$ from which we will derive the asymptotic scaling of C_2. We then integrate Eq. (60) after a change of variable $\theta(u) = \hat{\theta}_l + \frac{u}{\sqrt{n}}$,

$$
\begin{aligned}
1 = \int_\Theta p(\theta|x_1,..,x_n)\mathrm{d}\theta = &\int_{-(\theta_b-\hat{\theta}_l)\sqrt{n}}^{(\theta_b-\hat{\theta}_l)\sqrt{n}} \frac{1}{C_2\sqrt{n}}\pi(\hat{\theta}_l + \frac{u}{\sqrt{n}})e^{-(\hat{\theta}_l+\frac{u}{\sqrt{n}})\frac{\pi'(\hat{\theta}_l)}{\pi(\hat{\theta}_l)}}e^{-n\mathrm{KL}(\hat{\theta}_l,\hat{\theta}_l+\frac{u}{\sqrt{n}})}\mathrm{d}u\\
&+ \int_{\mu_{\inf}}^{-\theta_b} p(\theta|x_1,..,x_n)\mathrm{d}\theta + \int_{\theta_b}^{\mu_{\sup}} p(\theta|x_1,..,x_n).\mathrm{d}\theta
\end{aligned}
\tag{61}
$$

Taking $(\theta_b - \hat{\theta}_l) \sim n^{-b}$ with $b < 1/2$, we get rid of the tail components in the asymptotic limit. Secondly, by noticing that $\frac{F''(\hat{\theta}_l)}{2} = \lim_{\theta\to\hat{\theta}_l} K(\hat{\theta}_l,\theta)/|\theta-\hat{\theta}_l|^2$ from Eq. (57), we make an expansion to the lowest order of the Kullback-Leibler divergence, which gives:

$$
1 = \lim_{\theta\to\hat{\theta}_l}\int_{-(\theta_b-\hat{\theta}_l)\sqrt{n}}^{(\theta_b-\hat{\theta}_l)\sqrt{n}} \frac{1}{C_2\sqrt{n}}\pi(\hat{\theta}_l)e^{-\hat{\theta}_l\frac{\pi'(\hat{\theta}_l)}{\pi(\hat{\theta}_l)}}e^{-\frac{F''(\hat{\theta}_l)u^2}{2}\hat{\theta}_l + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)}\mathrm{d}u. \tag{62}
$$

Thus, we obtain:

$$
C_2 \sim \frac{\sqrt{2\pi}}{\sqrt{nF''(\hat{\theta}_l)}}\pi(\hat{\theta}_l)e^{-\hat{\theta}_l\frac{\pi'(\hat{\theta}_l)}{\pi(\hat{\theta}_l)}}. \tag{63}
$$

Of note, the gaussian limit also gives that $\bar{\sigma}_i^2 \sim F''(\hat{\theta}_l)^{-1}N_i^{-1}$.

Thus, we assume to develop an approximation scheme for a posterior distribution p_i asymptotically verifying:

$$
p_i(\theta) \underset{N_i\to\infty}{\sim} \sqrt{\frac{1}{2\pi\bar{\sigma}_i^2}}H(\theta,\hat{\theta}_l)e^{-N_i\mathrm{KL}(\hat{\theta}_l,\theta)}, \tag{64}
$$

where H is a function accounting for the prior distribution. For the following, we take a uniform prior on Θ, which leads to $H(\theta,\hat{\theta}_l) = 1$.

In the following, we will denote $\hat{\mu}_{M_t}$ and $\hat{\mu}_m$ as the maximum a posteriori estimates associated to their respective arms (instead of the empirical means).

D.2 The partitioning approximation

Since all the steps leading to the partitioning approximation are independent of the type of reward distribution, $\tilde{S}_{\mathrm{tail}}$ and $\tilde{S}_{\mathrm{body}}$ have the same general form as given in Appendix B.1. Here, we consider all the distributions under θ parameter for which we replace $\hat{\mu}_{M_t}, \hat{\mu}_m$ and $\tilde{\mu}_{\mathrm{eq}}$ by there equivalents $\hat{\theta}_{m_t}, \hat{\theta}_{M_t}$ and $\tilde{\theta}_{\mathrm{eq}}$.

D.3 Asymptotic intersection point

By use of Eq. (64), the equation verified by the intersection point $\bar{\theta}_{\mathrm{eq}}$ asymptotically reads:

$$
\frac{e^{-N_{M_t}\mathrm{KL}(\hat{\theta}_{M_t},\bar{\theta}_{\mathrm{eq}})}}{\sqrt{2\pi\bar{\sigma}_{M_t}^2}}\int_{\mu_{\inf}}^{\bar{\theta}_{\mathrm{eq}}}\frac{e^{-N_m\mathrm{KL}(\hat{\theta}_{m_t},\theta')}}{\sqrt{2\pi\bar{\sigma}_m^2}}\mathrm{d}\theta' = \frac{e^{-N_m\mathrm{KL}(\hat{\theta}_{m_t},\bar{\theta}_{\mathrm{eq}})}}{\sqrt{2\pi\bar{\sigma}_m^2}}\int_{\mu_{\inf}}^{\bar{\theta}_{\mathrm{eq}}}\frac{e^{-N_{M_t}\mathrm{KL}(\hat{\theta}_{M_t},\theta')}}{\sqrt{2\pi\bar{\sigma}_{M_t}^2}}\mathrm{d}\theta'. \tag{65}
$$

Taking the logarithm of Eq. (65) leads to

$$N_m \mathrm{KL}(\hat{\theta}_{m_t}, \bar{\theta}_{\mathrm{eq}}) - N_{M_t} \mathrm{KL}(\hat{\theta}_{M_t}, \bar{\theta}_{\mathrm{eq}}) + \frac{1}{2} \ln \frac{\bar{\sigma}_m^2}{\bar{\sigma}_{M_t}^2} + \ln \frac{\int_{\mu_{\inf}}^{\bar{\theta}_{\mathrm{eq}}} \sqrt{\bar{\sigma}_{M_t}^2} e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \theta')} \mathrm{d}\theta'}{\int_{\mu_{\inf}}^{\bar{\theta}_{\mathrm{eq}}} \sqrt{\bar{\sigma}_m^2} e^{-N_{M_t} \mathrm{KL}(\hat{\theta}_{M_t}, \theta')} \mathrm{d}\theta'} = 0. \tag{66}$$

Employing the same arguments as the ones exposed in Appendix B.2, we approximate $\bar{\theta}_{\mathrm{eq}}$ by neglecting the last term. Furthermore, in the considered asymptotic scaling regime ($N_{M_t} \gg N_m$), $\bar{\theta}_{\mathrm{eq}}$ will be in the vicinity of $\hat{\theta}_{M_t}$ where a Gaussian expansion of the Kullback-Leibler divergence is relevant (see Eq. (61)). Thus, we approximate $\mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})$ by $\mathrm{KL}(\hat{\theta}_{m_t}, \hat{\theta}_{M_t})$ and expand $\mathrm{KL}(\hat{\theta}_{M_t}, \bar{\theta}_{\mathrm{eq}})$ to lowest order in $\tilde{\theta}_{\mathrm{eq}}$ (with $\bar{\sigma}_i^2 \sim F''(\hat{\theta}_l)^{-1} N_i^{-1}$), leading to:

$$\tilde{\theta}_{\mathrm{eq}} = \hat{\theta}_{M_t} + \sqrt{2\bar{\sigma}_{M_t}^2 \left[N_m \mathrm{KL}(\hat{\theta}_{m_t}, \hat{\theta}_{M_t}) + \frac{1}{2} \ln \frac{\bar{\sigma}_m^2}{\bar{\sigma}_{M_t}^2} \right]}. \tag{67}$$

D.4 Generalization of the main mode's contribution

We start by recalling the expression for the body component of the entropy:

$$\tilde{S}_{\mathrm{body}} = - \int_{\Theta} p_{M_t}(\theta) C_m(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta. \tag{68}$$

Without any additional information on the expression for KL, Eq. (68) cannot be computed in a closed form. Thus, we will rely on the asymptotic scaling $N_{M_t} \gg N_m \gg 1$ to provide a tractable expression. First, we neglect variations of $C_m(\theta)$ in Eq. (68) integral by evaluating it at $\tilde{\mu}_{\mathrm{eq}}$. Then, by noticing that the resulting integral is the entropy of the better empirical arm's mean, we approximate it by its leading order, proportional to $\ln(2\pi\bar{\sigma}_{M_t}^2)/2$:

$$\tilde{S}_{\mathrm{body}} \approx \frac{1}{2} \ln(2\pi\bar{\sigma}_{M_t}^2) \left[1 - \int_{\tilde{\theta}_{\mathrm{eq}}}^{\mu_{\sup}} \frac{e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \theta')}}{\sqrt{2\pi\bar{\sigma}_m^2}} \mathrm{d}\theta' \right]. \tag{69}$$

We finally consider the last integral in Eq. (69). By noticing that it is concentrated in the vicinity of $\tilde{\mu}_{\mathrm{eq}}$ for $N_m \gg 1$, we Taylor expand $\mathrm{KL}(\hat{\theta}_{m_t}, \theta')$ at $\tilde{\mu}_{\mathrm{eq}}$ to obtain:

$$\begin{aligned} \int_{\tilde{\theta}_{\mathrm{eq}}}^{\mu_{\sup}} \frac{e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \theta')}}{\sqrt{2\pi\bar{\sigma}_m^2}} \mathrm{d}\theta &\approx \frac{e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\mu}_{\mathrm{eq}})}}{\sqrt{2\pi\bar{\sigma}_m^2}} \int_{\tilde{\theta}_{\mathrm{eq}}}^{\mu_{\sup}} e^{-N_m (\theta' - \tilde{\mu}_{\mathrm{eq}}) \partial_2 \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\mu}_{\mathrm{eq}})} \mathrm{d}\theta' \\ &\approx \frac{e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\mu}_{\mathrm{eq}})}}{\sqrt{2\pi\bar{\sigma}_m^2} N_m \partial_2 \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\mu}_{\mathrm{eq}})}. \end{aligned} \tag{70}$$

Inserting Eq. (70) into Eq. (69) leads to the body component:

$$\tilde{S}_{\mathrm{b}} = \frac{1}{2} \ln(2\pi\bar{\sigma}_{M_t}^2) \left[1 - \frac{e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})}}{\sqrt{2\pi\bar{\sigma}_m^2} N_m \partial_2 \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})} \right]. \tag{71}$$

D.5 Generalized expression for the entropy tail

We start by recalling the expression for the tail component:

$$\tilde{S}_{\mathrm{tail}} = - \int_{\tilde{\theta}_{\mathrm{eq}}}^{\mu_{\sup}} p_m(\theta) \ln p_m(\theta) \mathrm{d}\theta. \tag{72}$$

As for Eq. (70), we Taylor expand $\mathrm{KL}(\hat{\theta}_{m_t}, \theta')$ at $\tilde{\theta}_{\mathrm{eq}}$ in the exponential term to obtain:

$$\tilde{S}_{\mathrm{t}} = - \ln p_m(\tilde{\theta}_{\mathrm{eq}}) \frac{e^{-N_m \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})}}{\sqrt{2\pi\bar{\sigma}_m^2} N_m \partial_2 \mathrm{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\mathrm{eq}})}. \tag{73}$$

Keeping the leading order of $-\ln p_m(\tilde{\theta}_{\text{eq}}) \sim N_m \text{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\text{eq}})$ leads to the expected tail expression used in the main text.

D.6 Generalized form of the entropy approximation

To summarize, by combining Eqs. (71) and (73) we obtain an asymptotic expression for exponential family bandits with a uniform prior:

$$\tilde{S}_{\max} = \frac{1}{2}\ln(2\pi\bar{\sigma}^2_{M_t})\left[1 - \frac{e^{-N_m \text{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\text{eq}})}}{N_m \partial_2 \text{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\text{eq}})\sqrt{2\pi\bar{\sigma}^2_m}}\right] + \frac{\text{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\text{eq}})e^{-N_m \text{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\text{eq}})}}{\partial_2 \text{KL}(\hat{\theta}_{m_t}, \tilde{\theta}_{\text{eq}})\sqrt{2\pi\bar{\sigma}^2_m}}. \tag{74}$$

Finally, depending implementation, we propose for convenience to replace in $\tilde{S}_{\max}$ the maximum a posteriori estimates of each arm by either their empirical mean, their mean posterior values or the maximum of the log-likelihood. This does not alter the algorithm's efficiency in practice, while it may simplify the implementation procedure for specific reward distributions.

Note that all these steps can be adapted to non-uniform priors (in particular by multiplying the tail by the prior effects evaluated in $\tilde{\theta}_{\text{eq}}$). Finally, let us underline that our approximation scheme holds for any posterior distributions verifying Eq. (64), a property we believe to be shared for more general reward distributions.

D.7 Derivation of the increment for the closed-form expression of entropy

First, we stress there is no unique guideline to compute the expected increment of Eq. (74), and multiple solutions emerge depending on the type of the reward distribution. In particular, if the reward distribution is continuous, one could integrate the increment as it has been done for Gaussian rewards above. But, if the integration cannot be solved analytically, one could approximate the increments by taking discrete reward values of the order of $\pm\sigma$. Similarly, if the reward takes discrete values, the increments are already discrete, but asymptotic simplifications or taking the continuous limit can also be considered.

Finally, if the increment evaluation is discrete or approximated, one could encounter rare cases where the algorithm gets trapped. It could occur when the algorithm observes a worse suboptimal arm close to the best empirical arm when it has already extensively been drawn. Because the entropy could increase drastically if an arm inversion occurs, the gradient signs may occasionally switch, leading to the failure of the minimization procedure. To prevent such cases, we change the decision procedure by maximizing the entropy variation rather than its direct minimization. An example is given for the implementation of Bernoulli rewards in the next section.

Lastly, depending on the reward distribution it may be straightforward to express the increments along the usual empirical or posterior mean as opposed to the family parameter $\hat{\theta}$. Often, this can be achieved through a basic variable transformation. The Bernoulli distribution example provided below serves as an illustration of this approach.

E Numerical experiments

Here, we provide all the information regarding numerical experiments. This includes details on the numerical settings, implementation details for AIM in the investigated settings, an overview of investigated classical bandit algorithms, and additional experiments focusing on close-arm means.

E.1 Numerical settings

In Fig. 1, the posterior distributions are drawn with $\hat{\mu}_{M_t} \approx 0.65$, $N_1(t) = 374$, $\hat{\mu}_m \approx 0.29$, $N_m = 26$, where μ_i and $N_i(t)$ are, respectively, the empirical mean and number of draws of arm i and have been obtained with the AIM algorithm.

For the Gaussian two-armed cases in Fig. 2 the arm means are chosen from a uniform grid in $(0,1)\times(0,1)$ using a Sobol sequence (we have avoided the values 0 and 1 but it has no impact on the obtained results). The regret is averaged over 8192 games and observed during 10^8 steps to attest the logarithmic scaling. For Bernoulli rewards with two-armed Fig. 3, the regret is averaged over 16384 games and observed during 10^8 steps. It is worth noting that for Gaussian rewards, the prior information of arm means being only between 0 and 1 is not given to AIM nor to the Thompson sampling algorithm to allow a direct comparison.

For the fifty-armed case in Figs. 2 and 3, the arm means are drawn from a uniform prior, and the regret is averaged over 2000 games and observed during 10^6 steps in Fig. 2 and 10^7 in Fig. 3.

For close arm means in Figs. 4 and 5, the mean values are fixed with $\mu_1 = 0.79$ and $\mu_2 = 0.8$, but this prior information is not given to the investigated algorithms. The regret is averaged over 10^5 games and observed during 10^6 steps.

For the two-armed cases in Figs. 6 and 7, the arm means are chosen from a uniform grid in $(0,1)\times(0,1)$ using a Sobol sequence. The regret is averaged over more than 10^5 games and observed during 10^6 steps to enhance measurement accuracy.

Finally, in the fifty-armed case in Fig. 8, the arm means are drawn from a uniform prior, and the regret is averaged over 4.10^4 games and observed during 5.10^4 steps to enhance measurement accuracy.

Of note, for all the experiments, seed values are not shared throughout the algorithms. To obtain a sufficient number of runs, the code was parallelized on a cluster (asynchronously), with each run operating independently while ensuring that seed values are not common between runs. Because it relies on an analytical expression, AIM shows an execution time of the same order of Thompson sampling (measured three times slower for two-armed Bernoulli rewards).

For completeness, an implementation of AIM for both Bernoulli and Gaussian rewards and more than two arms are given in the supplementary material (`AIM Bernoulli bandits and AIM Gaussian` folders).

E.2 AIM implementation details

Here, we recap below AIM setups for the different settings evoked in the main text.

E.2.1 Approximate information maximization for the two arm Gaussian rewards

Specifically for the two-armed case, one can simplifies the expressions given in the main text. $\tilde{\mu}_{\rm eq}$ defined the value of θ where both arms have the same probability of being the maximal one and reads

$$\tilde{\mu}_{\rm eq} = \hat{\mu}_{M_t} + \frac{N_m(\hat{\mu}_{M_t} - \hat{\mu}_m)}{N_{M_t} - N_m} + \sqrt{\frac{N_{M_t}N_m(\hat{\mu}_{M_t} - \hat{\mu}_m)^2}{(N_{M_t} - N_m)^2} + \frac{\sigma^2}{N_{M_t} - N_m}\ln\left(\frac{N_{M_t}}{N_m}\right)}. \tag{75}$$

Hence, following identical approximations of the ones derived in Appendix B.1for Gaussian reward distributions, the tail component given by Eq. (7) simplifies into:

$$\tilde{S}_{\rm tail} = \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_m})\mathrm{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\rm eq} - \hat{\mu}_m)}{\sqrt{2\sigma^2}}\right) + \frac{\sqrt{N_m}(\tilde{\mu}_{\rm eq} - \hat{\mu}_m)}{2\sqrt{2\pi\sigma^2}}e^{-\frac{-N_m(\tilde{\mu}_{\rm eq} - \hat{\mu}_m)^2}{2\sigma^2}}. \tag{76}$$

Similarly for $\tilde{S}_{\rm body}$, we obtain:

$$\begin{aligned}\tilde{S}_{\rm body} = \frac{1}{2}\ln\left(\frac{2\pi\sigma^2 e}{N_{M_t}}\right) \times \left[1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\sqrt{N_m N_{M_t}}(\hat{\mu}_{M_t} - \hat{\mu}_m)}{\sqrt{2\sigma^2(N_m + N_{M_t})}}\right)\right] \\ - \frac{\sqrt{N_{M_t}}N_m^{3/2}(\hat{\mu}_{M_t} - \hat{\mu}_m)}{2\sigma\sqrt{2\pi}(N_{M_t} + N_m)^{3/2}}e^{-\frac{N_m N_{M_t}(\hat{\mu}_{M_t} - \hat{\mu}_m)^2}{2\sigma^2(N_m + N_{M_t})}}.\end{aligned} \tag{77}$$

Finally, it allows us to derive the approximation of the gradient difference of the entropy for the two-armed case:

$$\begin{aligned}\Delta = \frac{1}{2}\ln(\frac{N_m}{N_m + 1}) + \frac{1}{4N_{M_t}}\mathrm{erfc}\left(\frac{\sqrt{N_m}(\tilde{\mu}_{\rm eq} - \hat{\mu}_m)}{\sqrt{2\sigma^2}}\right) + \frac{\sqrt{N_m}(\tilde{\mu}_{\rm eq} - \hat{\mu}_m)}{\sqrt{2\pi\sigma^2}}\times \\ e^{-\frac{N_m(\tilde{\mu}_{\rm eq} - \hat{\mu}_m)^2}{2\sigma^2}}\left[\frac{2N_{M_t}^2 - 3N_m^2}{4N_m N_{M_t}^2} + \frac{1}{4}\ln\left(\frac{N_{M_t}}{2\pi\sigma^2 e}\right)\frac{N_m^3 + N_{M_t}^2}{N_m^2 N_{M_t}^2} + \frac{(N_m^3 + N_{M_t}^2)(\tilde{\mu}_{\rm eq} - \hat{\mu}_m)^2}{4\sigma^2 N_m N_{M_t}^2}\right],\end{aligned} \tag{78}$$

and the associated pseudo-code used for Figs. 2, 4 and 6 is presented in Alg. 2 below.

Algorithm 2: AIM Algorithm for 2 Gaussian arm

```
Draw each arm once; observe reward X_t(t) and update statistics μ̂_t ← X_t(t), N_t ← 1 ∀t ∈ {1,2}
for t = 3 to T do
    /* Arm selection                                                    */
    M_t ← argmax_{k=1,2} μ̂_k, m ← argmin_{k=1,2} μ̂_k;
    if N_{M_t} ≤ N_m then
        a_t ← M_t
    else
        Evaluate Δ following Eq. (9) ;
        if Δ < 0 then
            a_t ← M_t
        else
            a_t ← m
    Pull a_t and observe X_t(a_t)
    /* Update statistics                                                */
    μ̂_{a_t} ← (μ̂_{a_t} N_{a_t} + X_t(a_t)) / (N_{a_t} + 1), N_{a_t} ← N_{a_t} + 1
```

E.2.2 Approximate information maximization for Bernoulli rewards

We denote by $\bar{\mu}_i$ the posterior mean, given by:

$$\mathbb{E}\left[X_{\mathcal{B}(r_i+1,N_i-r_i+1)}\right] = \frac{r_i+1}{N_i+2} = \bar{\mu}_i, \tag{79}$$

where r_i is the cumulative reward at time t , N_i the number of draws, and $X_{\mathcal{B}(a,b)}$ follows a Beta distribution with parameters (a,b). The variance verifies:

$$\begin{aligned}\mathrm{Var}\left[X_{\mathcal{B}(r_i+1,N_i-r_i+1)}\right] &= \frac{r_i+1}{N_i+2}\frac{N_i-r_i+1}{N_i+2}\frac{1}{N_i+3}\\ &= \frac{\bar{\mu}_i(1-\bar{\mu}_i)}{\bar{N}_i},\end{aligned} \tag{80}$$

where $\bar{N}_i = N_i + 3$.

For Bernoulli rewards, we approximate the gradient as follows:

$$\begin{aligned}|\Delta_i \tilde{S}_{\max}| = \Bigg| &\frac{\bar{\mu}_i(\bar{N}_i-1)-1}{\bar{N}_i-3}\tilde{S}_{\max}\Big(\frac{\bar{\mu}_i\bar{N}_i+1-\bar{\mu}_i}{\bar{N}_i}, \bar{N}_i+1, \bar{\mu}_j, \bar{N}_j\Big)\\ &+\frac{\bar{N}_i-2-\bar{\mu}_i(\bar{N}_i-1)}{\bar{N}_i-3}\tilde{S}_{\max}\Big(\frac{\bar{\mu}_i(\bar{N}_i-1)}{\bar{N}_i}, \bar{N}_i+1, \bar{\mu}_j, \bar{N}_j\Big) - \tilde{S}_{\max}(\bar{\mu}_i, \bar{N}_i, \bar{\mu}_j, \bar{N}_j)\Bigg|,\end{aligned} \tag{81}$$

with $\tilde{S}_{\max}$ given by Eq. (74) expressed along μ with $\mu = \frac{e^\theta}{1+e^\theta}$. For Bernoulli rewards the equation reads:

$$\begin{aligned}\tilde{S}_{\max}(\bar{\mu}_{M_t}, \bar{N}_{M_t}, \bar{\mu}_m, \bar{N}_m) = &\left(1 - \frac{e^{-\bar{N}_m \mathrm{KL}(\bar{\mu}_m, \tilde{\mu}_{\mathrm{eq}})}}{\sqrt{\bar{N}_m}\partial_2 \mathrm{KL}(\bar{\mu}_m, \tilde{\mu}_{\mathrm{eq}})\sqrt{2\pi\bar{\mu}_m(1-\bar{\mu}_m)}}\right)\frac{1}{2}\ln\left(\frac{2\pi\bar{\mu}_{M_t}(1-\bar{\mu}_{M_t})}{\bar{N}_{M_t}}\right)\\ &+\frac{\sqrt{\bar{N}_m}\mathrm{KL}(\bar{\mu}_m, \tilde{\mu}_{\mathrm{eq}})e^{-\bar{N}_m \mathrm{KL}(\bar{\mu}_m, \tilde{\mu}_{\mathrm{eq}})}}{\partial_2 \mathrm{KL}(\bar{\mu}_m, \tilde{\mu}_{\mathrm{eq}})\sqrt{2\pi\bar{\mu}_m(1-\bar{\mu}_m)}},\end{aligned} \tag{82}$$

with $\mathrm{KL}(\theta, \theta') = \theta \ln(\theta/\theta') + (1-\theta)\ln([1-\theta]/[1-\theta'])$ and $\sigma_i^2 = \frac{\bar{\mu}_i(1-\bar{\mu}_i)}{\bar{N}_i}$.

Briefly, the expected gradient is evaluated along arm i with a returned reward equal to 1 with probability $\frac{\bar{\mu}_i(\bar{N}_i-1)-1}{\bar{N}_i-3}$ (which is the empirical mean) or equal to 0 with probability $1 - \frac{\bar{\mu}_i(\bar{N}_i-1)-1}{\bar{N}_i-3}$.

Algorithm 3: AIM Algorithm for 2 Bernoulli arm

Draw each arm once; observe reward $X_t(t)$ and update statistics $\bar{\mu}_t \leftarrow \frac{X_t(t)+1}{3}, \bar{N}_t \leftarrow 4 \; \forall t \in \{1, 2\}$
for $t = 3$ **to** T **do**
  /* Arm selection */
  $M_t \leftarrow \text{argmax}_{k=1,2} \bar{\mu}_k$, $m \leftarrow \text{argmin}_{k=1,2} \bar{\mu}_k$;
  if $N_{M_t} \leq N_m$ **then**
    $a_t \leftarrow M_t$
  else
    Evaluate $\Delta = |\Delta_{M_t} \tilde{S}_{\max}| - |\Delta_m \tilde{S}_{\max}|$ following Eq. (81) ;
    if $\Delta > 0$ **then**
      $a_t \leftarrow M_t$
    else
      $a_t \leftarrow m$
  Pull a_t and observe $X_t(a_t)$
  /* Update statistics */
  $\bar{\mu}_{a_t} \leftarrow \frac{\bar{\mu}_{a_t}(\bar{N}_{a_t}-1)+X_t}{\bar{N}_{a_t}}$, $\bar{N}_{a_t} \leftarrow \bar{N}_{a_t} + 1$

Of note, by adding absolute values, we seek to maximize the entropy variation rather than its direct minimization to avoid falling into an entrapment scenario (see Appendix D.7 for further discussion).

We draw some additional observations on the practical implementation of the code. First, in the gradient evaluation of Δ_i following Eq. (81) we may find a $\tilde{\mu}_{\text{eq}}$ value to be undefined (because $N_m > N_{M_t}$ or $\tilde{\mu}_{\text{eq},i} > 1$), which is unusable for Bernoulli rewards. In this case, $\tilde{\mu}_{\text{eq},i}$ is taken to be equal to 1, resulting in $S_{\max} = \frac{1}{2} \ln \left(\frac{2\pi \bar{\mu}_{M_t}(1-\bar{\mu}_{M_t})}{\bar{N}_{M_t}} \right)$.

Second, at large times, noticing that the better empirical arm is drawn extensively, one can increment the algorithm by multiple steps at a time to speed up AIM . Indeed, let us assume that the better empirical arm is drawn T times successively while always returning a null reward, which is the worst scenario for the returned reward of the better empirical arm. Then, if the increment evaluation at $t + T$ of Alg. 3 still returns M_t, then it ensures that all increment evaluations between $[t, t + T]$ of Alg. 3 will always return M_t independently of its returned rewards. Then, using a dichotomy search on the variable T, we can diminish the number of increment evaluations of AIM at large times, thus improving AIM's performance.

E.3 Information maximization approximation for Bernoulli rewards with more than two arms

We start by reminding the obtained entropy approximation for more than two arms:

$$\tilde{S}_{\max} = -\int_{\Theta} \big(1 - \sum_{i \neq M_t}^{K} [1 - C_i(\theta)]\big) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta - \sum_{i \neq M_t}^{K} \int_{\tilde{\mu}_{\text{eq},i}}^{\mu_{\text{sup}}} p_i(\theta) \ln p_i(\theta) \mathrm{d}\theta. \tag{83}$$

We first consider the increment along a worse empirical arm, which simplifies :

$$|\Delta_i \tilde{S}_{\max}| = \Delta_i \bigg[-\int_{\Theta} C_i(\theta) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta - \int_{\tilde{\mu}_{\text{eq},i}}^{\mu_{\text{sup}}} p_i(\theta) \ln p_i(\theta) \mathrm{d}\theta \bigg], \tag{84}$$

which is exactly the increment evaluated in the two-armed case given in Eq. (82).

Finally, we consider the increment along the better empirical arm. For simplicity, we neglect $\tilde{\mu}_{\text{eq},i}$ variations for the increments evaluation. By use of Eq. (82) we obtain

$$|\Delta_{M_t} S_{\max}| = \left| 1 - \sum_{i \neq M_t}^{K} \frac{e^{-\bar{N}_m \mathrm{KL}(\bar{\mu}_i, \tilde{\mu}_{\text{eq}})}}{\sqrt{N_m} \partial_2 \mathrm{KL}(\bar{\mu}_i, \tilde{\mu}_{\text{eq}}) \sqrt{2\pi \bar{\mu}_i (1 - \bar{\mu}_i)}} \right| \left| \Delta_{M_t} H(\bar{\mu}_{M_t}, \bar{N}_{M_t}) \right|, \tag{85}$$

where

$$\left|\Delta_{M_t} H(\bar{\mu}_i, \bar{N}_i)\right| = \left|\frac{\bar{\mu}_i(\bar{N}_i - 1) - 1}{\bar{N}_i - 3} H(\frac{\bar{\mu}_i \bar{N}_i + 1 - \bar{\mu}_i}{\bar{N}_i}, \bar{N}_i + 1, \bar{\mu}_j, \bar{N}_j) + \frac{\bar{N}_i - 2 - \bar{\mu}_i(\bar{N}_i - 1)}{\bar{N}_i - 3} H(\frac{\bar{\mu}_i(\bar{N}_i - 1)}{\bar{N}_i}, \bar{N}_i + 1, \bar{\mu}_j, \bar{N}_j) - H(\bar{\mu}_i, \bar{N}_i, \bar{\mu}_j, \bar{N}_j)\right|, \tag{86}$$

with $H(\bar{\mu}_{M_t}, \bar{N}_{M_t}) = \frac{1}{2} \ln\left(\frac{2\pi\bar{\mu}_{M_t}(1-\bar{\mu}_{M_t})}{\bar{N}_{M_t}}\right)$.

Algorithm 4: AIM Algorithm for $K > 2$ Bernoulli arm

Draw each arm once; observe reward $X_t(t)$ and update statistics $\bar{\mu}_t \leftarrow \frac{X_t(t)+1}{3}, \bar{N}_t \leftarrow 4\ \forall t \in \{1, ..K\}$
for $t = K + 1$ **to** T **do**
  /* Arm selection */
  $M_t \leftarrow \operatorname{argmax}_{k=\{1,..,K\}} \bar{\mu}_k$; Evaluate $\Delta_{M_t} \tilde{S}_{\max}$ following Eq. (85) ;
  Evaluate $m = argmax(\Delta_i |\tilde{S}_{\max}|, i \neq M_t)$ with $\Delta_i |\tilde{S}_{\max}|$ following Eq. (81) ;
  if $\Delta_{M_t} |\tilde{S}_{\max}| > \Delta_m |\tilde{S}_{\max}|$ **then**
    $a_t \leftarrow M_t$
  else
    $a_t \leftarrow m$
  Pull a_t and observe $X_t(a_t)$
  /* Update statistics */
  $\bar{\mu}_{a_t} \leftarrow \frac{\bar{\mu}_{a_t}(\bar{N}_{a_t}-1)+X_t}{\bar{N}_{a_t}}$, $\bar{N}_{a_t} \leftarrow \bar{N}_{a_t} + 1$

Of note in the gradient evaluation of Δ_i following Eq. (81), if ones finds a $\tilde{\mu}_{\text{eq},i}$ value undefined (because $N_m > N_{M_t}$ or $\tilde{\mu}_{\text{eq},i} > 1$ which is unusable for Bernoulli reward), then, $\tilde{\mu}_{\text{eq},i}$ is taken to be equal to 1 resulting in $S_{\max} = \frac{1}{2} \ln\left(\frac{2\pi\bar{\mu}_{M_t}(1-\bar{\mu}_{M_t})}{\bar{N}_{M_t}}\right)$. Finally, if $M_t \leftarrow \operatorname{argmax}_{k=\{1,..,K\}} \bar{\mu}_k$ has multiple solutions, we suggest choosing the one displaying the lowest number of draws.

E.4 Overview of baseline bandit algorithms

Here, we briefly review several baseline algorithms and their chosen parameters to provide a benchmark of our information maximization method.

E.4.1 UCB-Tuned

This algorithm falls under the category of upper confidence bound (UCB) algorithms, which select the arm maximizing a proxy function typically defined as $F_i = \hat{\mu}_i + B_i$. For UCB-tuned, B_i is given by:

$$R_i = c(\mu_1, \mu_2) \sqrt{\frac{\ln(t)}{N_i(t)} \min\left(\frac{1}{4}, s_i(t)\right)},\ \ s_i(t) = \hat{\sigma_i}^2 + \sqrt{\frac{2\ln(t)}{N_i(t)}}, \tag{87}$$

where $\hat{\sigma_i}^2$ is the reward variance and c a hyperparameter. For Gaussian rewards, by testing various c values for uniform priors in Eq. (87), we end up with $c = 2.1$ and $\hat{\sigma_i}^2 = \frac{\sigma^2}{N_i(t)}$.

E.4.2 KL-UCB

This algorithm is another variant of the upper confidence bound (UCB) class specifically designed for bounded rewards. In particular, it is known to be optimal for Bernoulli distributed rewards (Garivier and Cappé, 2011; Cappé et al., 2013). For KL-UCB, F_i is expressed as follows:

$$F_i = \max\Big\{\theta \in \Theta : N_i(t)\mathrm{KL}\left(\frac{r_i(t)}{N_i(t)}, \theta\right) \leq \ln(t) + c(\mu_1, \mu_2)\ln(\ln(t))\Big\}, \tag{88}$$

where Θ denotes the definition interval of the posterior distribution. By testing various c values for uniform priors, we end up with $c(\mu_1, \mu_2) = 0.00001$ ($c = 0$ for the 50-armed Gaussian setting and $c = 10^{-6}$ for the 2-armed Bernoulli setting). Of note, the maximum is found using a dichotomy method using a precision of 10^{-5} and a maximum number of iterations of 50.

For KL-UCB++ (Ménard and Garivier, 2017), the function F_i is expressed as follows

$$F_i = \max\Big\{\theta \in \Theta : N_i(t)\mathrm{KL}\left(\frac{r_i(t)}{N_i(t)}, \theta\right) \leq \ln_+\left(\frac{T}{KN_i(t)(t)}\ln_+^2\left(\frac{T}{KN_i(t)(t)}\right) + 1\right)\Big\}, \tag{89}$$

where $\ln_+(x) = \max(\ln(x), 0)$, and T is the stopping time of the bandit game. Therefore, KLUCB++ is not an anytime algorithm, but it still underperforms when compared to AIM and Thompson sampling.

E.4.3 Thompson sampling

At each step, Thompson sampling (Thompson, 1933; Kaufmann et al., 2012a,b) selects an arm at random, based on the posterior probability that is maximizes the expected reward. In practice, it draws K random values according to each arm's mean posterior distribution and selects the arm with the highest sampled value as:

$$a_t = \underset{i=1..K}{\operatorname{argmax}}\Big(Z_i\left(\hat{\mu}_i(t), N_i(t)\right)\Big), \tag{90}$$

where $Z_i(t)$ is drawn according to the posterior distribution of the ith arm's mean. Here, we used a uniform prior on $[0, 1]$ for Bernoulli rewards and a uniform prior on $\mathbb{R}$ for Gaussian rewards to provide a direct comparison with AIM.

Finally, for Thompson sampling plus (Jin et al., 2022), denoted TS+, each sampled value for the comparison is drawn according to $Z_i\left(\hat{\mu}_i(t), N_i(t)\right)$ with a probability $1/K$ or taken equal to $\hat{\mu}_i(t)$, otherwise.

E.4.4 MED

At each step, the minimal empirical divergence (MED) algorithm (Honda and Takemura, 2011), selects an arm at random, based on a tailored distribution building on the Kullback-Leibler distance to the better empirical arm. In practice, the arm $a_t = i$ is be drawn with a probability:

$$p_i = \frac{\exp\left[-N_i(t)\mathrm{KL}\left(\frac{r_i(t)}{N_i(t)}, \bar{\mu}_{M_t}\right)\right]}{\sum_{j=0}^{K}\exp\left[-N_j(t)\mathrm{KL}\left(\frac{r_j(t)}{N_j(t)}, \bar{\mu}_{M_t}\right)\right]}. \tag{91}$$

E.5 Additional experiments

E.5.1 Approximate information maximization for Gaussian rewards and close arms

For completeness, we provide in Fig. 4 below regret performances in which the arms' mean values are close ($\Delta\mu = 0.01$), and are thus difficult to distinguish, for Gaussian reward distributions. Here, AIM shows state-of-the-art performance comparable to Thompson sampling, even outperforming it at longer times.

E.5.2 Approximate information maximization for Bernoulli rewards and close arms

For completeness, we provide in Fig. 5 below regret performances in which the arms mean value are close ($\Delta\mu = 0.01$) for Bernoulli reward distributions. As for Gaussian rewards, AIM shows state-of-the-art performance comparable to Thompson sampling even when arms mean rewards are difficult to distinguish.

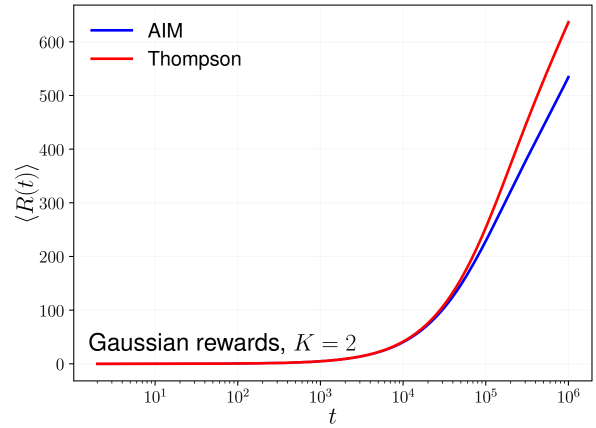

Figure 4: Temporal evolution of the regret for 2-armed bandit with Gaussian rewards ($\sigma = 1$) for close mean parameters. In blue AIM, in red Thompson sampling. Arm mean reward values are fixed with $\mu_1 = 0.8$ and $\mu_2 = 0.79$, the regret is obtained by averaging over 10^5 realizations.

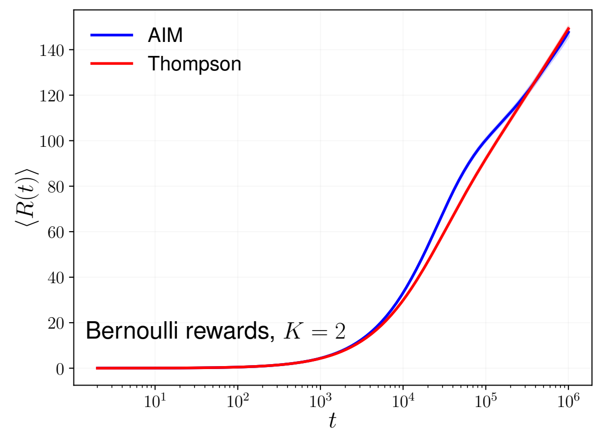

Figure 5: Temporal evolution of the regret for 2-armed bandit with Bernoulli rewards for close mean parameters. In blue AIM, in red Thompson sampling. Arm mean reward values are fixed with $\mu_1 = 0.8$ and $\mu_2 = 0.79$, the regret is obtained by averaging over 10^5 realizations. Confidence intervals shows the standard deviation.

E.5.3 Investigating approximate information maximization for large simulation volumes

To refine the numerical investigation of AIM's regret performance, we replicate the experiments of the main text using a larger volume of simulations (but on shorter timescales). For the 2-armed bandit with Gaussian and Bernoulli rewards, the regret performance under a uniform prior is averaged over more than 10^5 runs. Similarly, for the 50-armed bandit with Bernoulli rewards, the regret is averaged over 4×10^4. This leads to the results shown in Figs. 6 to 8, confirming the results of Figs. 2 to 3.

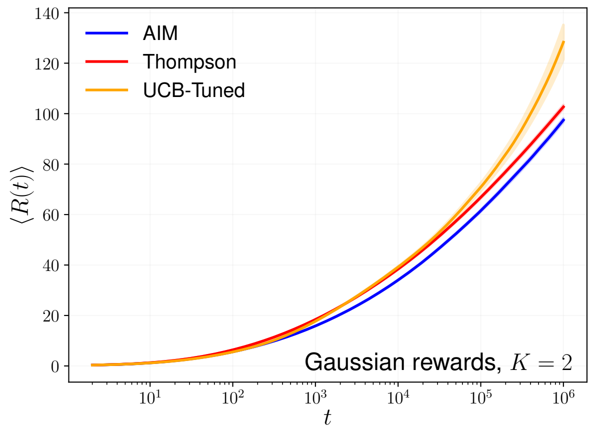

Figure 6: Evolution of the Bayesian regret for 2-armed bandit with Gaussian rewards under a uniform mean prior. The regret is averaged over more than 10^5 runs. Confidence intervals shows the standard deviation. Confidence intervals show the standard deviation.

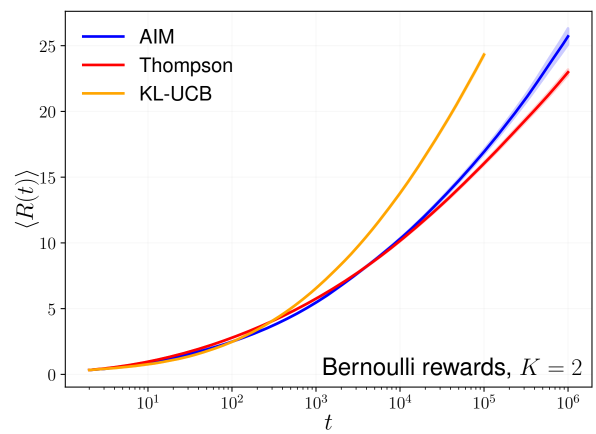

Figure 7: Evolution of the Bayesian regret for 2-armed bandit with Bernoulli rewards under a uniform mean prior. The regret is averaged over more than 10^5 runs. Confidence intervals show the standard deviation.

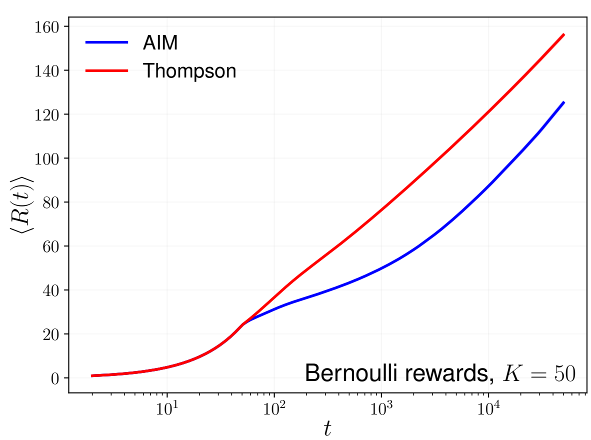

Figure 8: Evolution of the Bayesian regret for 50-armed bandit with Bernoulli rewards under a uniform mean prior. The regret is averaged over 4×10^4 runs. Confidence intervals show the standard deviation.