Approximate information maximization for bandit games

Alex Barbier-Chebbah, Christian L. Vestergaard, Jean-Baptiste Masson, Etienne Boursier

Approximate information maximization for bandit games

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

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

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

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

Abstract

Entropy maximization and free energy minimization are general physical principles for modeling the dynamics of various physical 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). Built 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 an entropy to forecast the information gain of each action and greedily choose the one with the largest information gain. This method yields strong performances in classical bandit settings. Motivated by its empirical success, we prove its asymptotic optimality for the two-armed bandit problem with Gaussian rewards. Owing to its ability to encompass the system's properties in a global physical functional, this approach can be efficiently adapted to more complex bandit settings, calling for further investigation of information maximization approaches for multi-armed bandit problems.

Preprint.

1 Introduction

Multi-armed bandit problems have raised a vast interest in the past decades. They embody the challenge of balancing exploration and exploitation and have been applied to various different 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 finer 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 setting, 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 as the exploration-exploitation trade-off. An agent can exploit its current knowledge to optimize gains by drawing the current empirically best arm or exploring other arms to potentially increase future gains.

In the infinite horizon setting, 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 associate a tuned confidence index to each arm, Thompson sampling (Kaufmann et al., 2012a; Agrawal and Goyal, 2013) relies on sampling the action from the posterior distribution that maximizes the expected reward, and 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 current available information, they do not aim directly to acquire more information while considering the impact of future draws on the expected reward gain. We highlight however the approach of Russo and Van Roy (2014), which relies on a measure of the information gain of the optimal actions. However, like DMED, this method explicitly balances information gain with expected losses induced by exploration, and the efficiency of purely information-maximizing approaches remains to be proven.

Information-maximization approaches consist of building a decision-making strategy where the agent tries to maximize their information about one or a set of relevant stochastic variables. This principle has 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. These domains include, e.g., robotics applications (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 case of bandits, information-based strategies have shown promising empirical results, and heuristic arguments support their asymptotic optimality (Reddy et al., 2016; Barbier-Chebbah et al., 2023). In this context, we aim to leverage new strategies derived from this information acquisition principle without the burden of numerical computational evaluation of complex functionals and rigorously prove their efficiency.

Contributions. Our main contribution is the introduction of a new class of asymptotically optimal algorithms that rely on approximations of a functional representing the currently available information 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 two-armed bandit problem with Gaussian rewards, for which we derive a simple approximate information maximization algorithm (AIM) and provide an upper bound on its pseudo-regret, ensuring that AIM is asymptotically optimal. Indeed, the information from each arm is mixed in a unique entropy functional, which shows promise for tackling more complex bandits settings such as linear bandits, or in the presence of a large amount of untested arms *aka* the many arms problem. Our main motivation is thus to design an analytic 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, which shows strong performances, as illustrated numerically.

Organization. In Section 2, we briefly review the two-armed bandit setting. Section 3 presents the general principle of information maximization approaches originally inspired from both information bottleneck principles and navigation in turbulent plumes. Section 4 upper bounds the regret of AIM, which attains Lai and Robbins asymptotic bound. The performance of AIM is numerically compared with known baselines on a few examples Section 5. Finally, Section 6 discusses extensions of AIM to various bandit settings.

2 Setting

We consider the classical two-armed stochastic bandit. In each round t, the agent selects an arm $a_t \in [K] = \{1, \dots, K\}$ among a set of $K = 2$ 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 $\mu_k \in [0, 1]$. The goal of the agent is then to maximize its cumulative reward, or equivalently to minimize its pseudo-regret up to round T defined as

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

where $\mu^* = \max(\mu_1, \mu_2)$. 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 pseudo-regret is lower-bounded for any uniformly good policy by Lai et al. (1985) as

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

where $k^* = \mathrm{argmax}_{i=1,2}\, \mu_i, k = \mathrm{argmin}_{i=1,2}\, \mu_i$ and $D_{\mathrm{KL}}(\nu_k \| \nu_{k^*})$ denotes the Kullback-Leibler divergence between the reward distributions of the arms k and k^*. 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 $D_{\mathrm{KL}}(\nu_1 \| \nu_2) = (\mu_1 - \mu_2)^2/(2\sigma^2)$.

3 Infomax strategies

We introduce here entropy-based strategies and their underlying physical principles. We then detail the ap-

proximations leading to the implementation of an analytic and simplified entropy functional, which is the basis of the AIM algorithm.

3.1 Algorithm design principle: physical intuition

We aim to design a functional that encompasses the current available knowledge of the full system. Inspired by the information maximization principle (Vergassola et al., 2007; Reddy et al., 2016) which has revealed effective in taxis strategies where the agent needs to find an emitting odour source (Martinez et al., 2014; Cardé, 2021; Murlis et al., 1992), we rely on an entropic functional for policy decision. More precisely, we choose $S_{\max}$, the entropy of the posterior distribution of the value of the maximal mean reward, denoted $p_{\max}$. Formally, $S_{\max}$ reads as

$$S_{\max} = -\int_{\Theta} p_{\max}(\theta) \ln p_{\max}(\theta) \mathrm{d}\theta, \tag{3}$$

where $\Theta = [\mu_{\inf}, \mu_{\sup}]$ is the support of $p_{\max}$ (which depends on the nature of the game and can be infinite) and

$$\begin{aligned} p_{\max}(\theta)\mathrm{d}\theta &= \mathrm{d}\mathbb{P}\left(\max_k \mu_k = \theta \mid \mathcal{F}_{t-1}\right) \\ &= \sum_{k=1}^{K} \mathrm{d}\mathbb{P}(\mu_k = \theta|\mathcal{F}_{t-1}) \prod_{j\neq k} \mathbb{P}(\mu_j \leq \theta|\mathcal{F}_{t-1}), \end{aligned} \tag{4}$$

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

Of note, $p_{\max}$ includes the arms' priors and depends on the reward distributions.[1] Similarly to Thompson sampling, it relies on a Bayesian representation of the environment. Yet, given past observations, it distinguishes itself by providing a deterministic decision procedure. The entropy encompasses a measure of the information carried by all arms in a single functional, characterising a global state description of the game.

Our policy aims at minimizing the entropy of the system. For that, it greedily chooses the arm providing the largest decrease in entropy, conditioned on the current knowledge of the game,

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

We stress that $S_{\max}$ does not quantify the available information about the best arm but rather about the value of the average reward of the best arm. One could have assumed that a policy based on the information gain about the best arm's identity could have been more efficient. However, such a policy tends to to overexplore, leading to a linear regret, as shown in (Reddy et al., 2016).

Therefore, approaches based on the information on the best arm's mean reward fix this concern by including the expected regret in the functional to favor exploitation (Russo and Van Roy, 2014). Furthermore, we argue that by definition of $p_{\max}$, the information carried by the arms' posteriors are sufficiently mixed to ensure an optimal behavior that we will prove in Section 4. Since the policy aims to maximise the information about the best arm's mean, it mainly pulls the current best arm to learn more about its value. On the contrary, best arm identification policies pull worse empirical arms more often because they are only concerned about the arms' order.

The information maximization policy based on Eq. (5) has been empirically shown to be competitive with state-of-the-art algorithms (Reddy et al., 2016) and robust to prior variations (Reddy et al., 2016; Barbier-Chebbah et al., 2023) on classical bandit settings. However, while Eq. (5) can be numerically evaluated, it cannot be computed in closed form, preventing the gradient from being analytically tractable. This makes it complicated to theoretically bound the two-armed setting but also prevents its extension to more complicated bandit settings. Additionally, it also induces a computational cost which isdisadvantageous when considering a large number of arms. Finally, the integral form of $S_{\max}$ prevents fine-tuning that could reveal itself crucial to obtain or surpass empirical state of the art performances.

A second simplified and analytic functional mirroring $S_{\max}$ has to be derived to address these concerns. This analytic 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 Towards an analytical approximation

Here, we devise a set of approximations of $p_{\max}$ and $S_{\max}$ to get a tractable analytical algorithm. Given that the better empirical arm and the worse empirical arm have notably distinct contributions to $p_{\max}$, we approximate $p_{\max}$ while taking into account the current arms' order. We sort them based on their present posterior means, labelling the highest one as M_t (with an empirical reward of $\hat{\mu}_{M_t}$) and the worse empirical

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

one as m_t (with $\hat{\mu}_{m_t}$). Of note, M_t might differ from the actual optimal arm k^* due to the randomness in the observed rewards. Our algorithm focuses on approximating Eqs. (3) and (4) when the best empirical arm has already been extensively drawn more often than the other arm.

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 $p_{\max}$, while the second term, $\tilde{S}_{\text{tail}}$, quantifies the information carried by the tail of $p_{\max}$ (corresponding to high rewards, see Fig. 1). Each of these terms then corresponds to a part of the entropy where the dominant term of Eq. (4) is distinct (see Appendix A.1 for details).

More precisely, the tail term reads as:

$$\tilde{S}_{\text{tail}} = -\int_{\tilde{\mu}_{\text{eq}}}^{\mu_{\text{sup}}} p_{m_t}(\theta) \ln p_{m_t}(\theta) \mathrm{d}\theta, \tag{7}$$

where $\tilde{\mu}_{\text{eq}}$, defined in Eq. (9) below, approximates $\bar{\mu}_{\text{eq}}$, the value of θ where both arms have the same probability of being the maximal one (see red and orange curves in Fig. 1(b)), and $p_{m_t}(\theta) = \frac{\mathrm{d}\mathbb{P}(\mu_{m_t}=\theta|\mathcal{F}_t)}{\mathrm{d}\theta}$ is the posterior density of the current worse arm evaluated at θ. Roughly, because the better empirical arm has been largely drawn, $p_{M_t}(\theta)$ decays faster than $p_{m_t}(\theta)$ resulting on a tail term (see Eq. (7)) whose main contribution is the worse empirical arm. The approximate entropy of the body component reads as:

$$\tilde{S}_{\text{body}} = -\int_{\Theta} C_{m_t}(\theta) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta, \tag{8}$$

where $p_{M_t}(\theta)$ is the posterior density at θ of the better empirical arm and $C_{m_t}(\theta) = \mathbb{P}(\theta > \mu_{m_t} \mid \mathcal{F}_t)$ is the cumulative probability of the mean of the worse empirical arm. Eq. (8) is the leading-order term of the mode of $p_{\max}$, which is mainly contributed to by the better empirical arm.

This approximation of Eq. (4) is good when the best empirical arm has been extensively drawn with respect to the worst empirical one, which corresponds to the situation encountered at infinity for uniformly good algorithms. Surprisingly, when both arms have instead been pulled roughly the same number of times, the approximation captured by Eq. (6) is still accurate enough to provide a high-performance decision scheme.

We denote by $N_k(t)$ and $\hat{\mu}_k(t)$ the number of times the arm k has been pulled and its empirical mean at time t, respectively. When clear from context, we omit the dependence on t for simplicity.

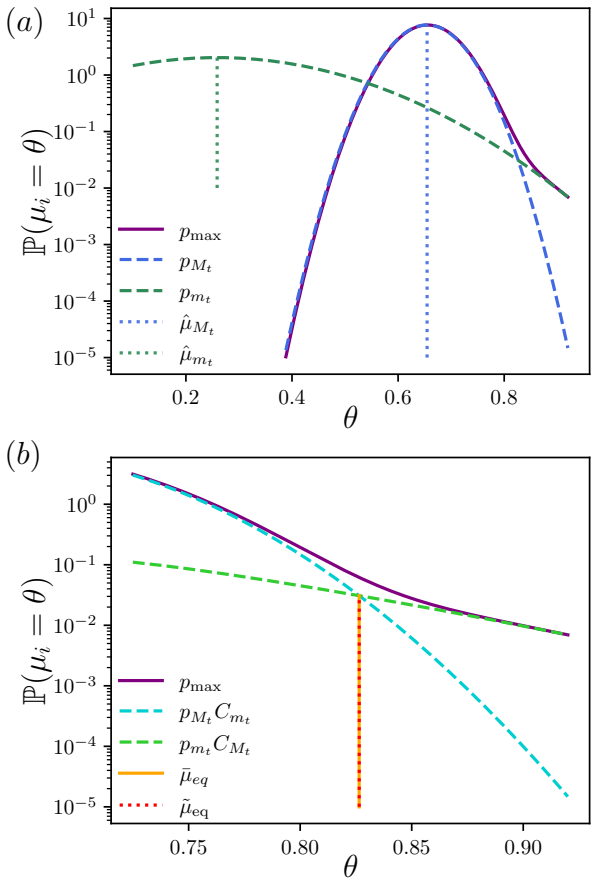

Figure 1: **(a)** Posterior distributions of a two-armed bandit with Gaussian rewards. The dotted lines represent the individual posterior distributions of each arm while the continuous line represents the posterior of the maximum mean reward of all arms (Eq. (4)). **(b)** Zoom of **(a)** in the region where both arms have the same posterior probability of being the best one. $p_{M_t}C_{m_t}$ ($p_{m_t}C_{M_t}$) is the probability that the maximal value is given by the better (worse) empirical arm.

For Gaussian reward distributions, one can first derive an analytic expression for $\tilde{\mu}_{\text{eq}}$ (see Appendix A.2 for details):

$$\tilde{\mu}_{\text{eq}} = \hat{\mu}_{M_t} + \frac{N_{m_t}(\hat{\mu}_{M_t} - \hat{\mu}_{m_t})}{N_{M_t} - N_{m_t}} + .. \\ \sqrt{\frac{N_{M_t} N_{m_t} (\hat{\mu}_{M_t} - \hat{\mu}_{m_t})^2}{(N_{M_t} - N_{m_t})^2} + \frac{\sigma^2}{N_{M_t} - N_{m_t}} \ln\left(\frac{N_{M_t}}{N_{m_t}}\right)}. \tag{9}$$

Hence, for Gaussian reward distributions Eq. (7) can

be rewritten as (see Appendix A.4):

$$\tilde{S}_{\text{tail}} = \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{m_t}})\text{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right) + \frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{2\sqrt{2\pi\sigma^2}} e^{-\frac{-N_{m_t}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}. \tag{10}$$

Similarly for $\tilde{S}_{\text{body}}$, we obtain for Gaussian rewards (see Appendix A.3):

$$\tilde{S}_{\text{body}} = \frac{1}{2}\ln\left(\frac{2\pi\sigma^2 e}{N_{M_t}}\right)\times .. \left[1-\frac{1}{2}\text{erfc}\left(\frac{\sqrt{N_{m_t}N_{M_t}}(\hat{\mu}_{M_t}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2(N_{m_t}+N_{M_t})}}\right)\right] - \frac{\sqrt{N_{M_t}}N_{m_t}^{3/2}(\hat{\mu}_{M_t}-\hat{\mu}_{m_t})}{2\sigma\sqrt{2\pi}(N_{M_t}+N_{m_t})^{3/2}} e^{-\frac{N_{m_t}N_{M_t}(\hat{\mu}_{M_t}-\hat{\mu}_{m_t})^2}{2\sigma^2(N_{m_t}+N_{M_t})}}. \tag{11}$$

At this stage, even if we have already obtained a closed-form expression for $S_{\max}$, it remains too involved to directly compute its exact (discrete) gradient for our decision policy. To finally derive a simplified gradient, we opt to retain asymptotic expressions of Eq. (10) and Eq. (11) and of the obtained gradient (see Appendix A.5 for derivation details). Finally, the expression of our approximate difference of gradients of the entropy functional reads:

$$\Delta = \frac{1}{2}\ln(\frac{N_{m_t}}{N_{m_t}+1}) + \frac{1}{4N_{M_t}}\text{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right) + \frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\pi\sigma^2}} e^{-\frac{N_{m_t}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}} \times \left[\frac{2N_{M_t}^2-3N_{m_t}^2}{4N_{m_t}N_{M_t}^2} + \frac{1}{4}\ln\left(\frac{N_{M_t}}{2\pi\sigma^2 e}\right)\frac{N_{m_t}^3+N_{M_t}^2}{N_{m_t}^2N_{M_t}^2} + \frac{(N_{m_t}^3+N_{M_t}^2)(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})^2}{4\sigma^2 N_{m_t}N_{M_t}^2}\right]. \tag{12}$$

In words, Δ approximates the difference

$$\Delta \approx \mathbb{E}\left[S_{\max}(t) \mid \mathcal{F}_{t-1}, a_t = M_t\right] - \mathbb{E}\left[S_{\max}(t) \mid \mathcal{F}_{t-1}, a_t = m_t\right],$$

which is directly related to the greedy choice maximizing the entropy decrease, described in Eq. (5).

The decision procedure can be summarized as follows: if Eq. (12) is negative, the better empirical arm is chosen as it reduces the most our entropy approximation in expectation. Inversely, if Eq. (12) is positive, the worse empirical arm is chosen.

In conclusion, we have derived an analytical expression obtained from the information measure of the maximum posterior. Furthermore, it allows us to isolate an analytically tractable gradient acting as a decision procedure that eluded previous approximated information derivations. We now provide the full AIM implementation and bound its regret in the next section.

3.3 Approximate information maximization algorithm

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

Algorithm 1: 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_t ← argmin_{k=1,2} μ̂_k;
    if N_{M_t} ≤ N_{m_t} then
        a_t ← M_t
    else
        Evaluate Δ following Eq. (12) ;
        if Δ < 0 then
            a_t ← M_t
        else
            a_t ← m_t
    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
```

The better empirical arm is drawn by default if $N_{M_t} \leq N_{m_t}$. In such a case, both entropy components in Eq. (6) are mainly contributed to by M_t since the better empirical arm has been less selected than the worse empirical one.

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 two-armed bandits problem with Gaussian rewards.

Theorem 1. *For Gaussian reward distributions with unit variance, the regret of AIM satisfies for any mean vector* (μ_1, μ_2)

$$\limsup_{T\to\infty} \frac{R(T)}{\ln(T)} \leq \frac{2\sigma^2}{|\mu_1-\mu_2|}.$$

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 B. We briefly sketch the proof idea below, and the complete proof is deferred to Appendix B.

Sketch of the proof. We assume without loss of generality here and in the proof that $\mu_1 > \mu_2$. Although very different in the demonstration, the structure of the proof is borrowed from Kaufmann et al. (2012a). In particular, the first main step is showing that the optimal arm is pulled at least a polynomial (in t) number of times with high probability. This result holds because otherwise, the contribution of the arm 1 in the tail of the information entropy would dominate the contribution of the arm 2 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.

From there, we only need to work in the asymptotic regime where both arms 1 and 2 are pulled a lot of times. We can ignore low-probability events and restrict the analysis to cases where empirical means of the arms do not significantly deviate from the true means. An important property of the entropy is that it approximates the behaviour of the bound of Lai et al. (1985). More precisely, in the asymptotic regime, the difference Δ of the entropy gradients almost behaves as

$$\Delta \approx -\frac{1}{2N_{M_t}} + \mathrm{poly}(N_{m_t}) e^{-\frac{N_{m_t}(\mu_1-\mu_2)^2}{2\sigma^2}}, \qquad (13)$$

where $\mathrm{poly}(N_{m_t})$ is some positive polynomial in N_{m_t}. In the asymptotic regime, we thus naturally have N_{m_t} of order $\frac{2\sigma^2 \ln T}{(\mu_1-\mu_2)^2}$. □

Moreover, the required form of the entropy for the proof is very general. As long as we are guaranteed that the optimal arm will be pulled a significant amount of times with high probability and that the asymptotic regime behaves as Eq. (13), the algorithm yields an optimal regret. Hence, Theorem 1 should hold for a large family of entropy approximations (and likely for free energy generalization 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 we devised here allows tuned formulas, in which individual terms ensure asymptotic optimality and the coefficients in front of them can be adjusted to improve short-time performance.

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 D.

We start by considering two arms, with Gaussian rewards (Honda and Takemura, 2010) of unit variance and means μ_k drawn uniformly in $[0, 1]$. 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 and Thompson sampling (Kaufmann et al., 2012b; Pilarski et al., 2021; Cappé et al., 2013) (see Appendix D.4 for detailed descriptions of the algorithms and an overview of other classic bandit strategies). 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. AIM particularly outperforms Thompson sampling in instances where the arms are difficult to distinguish due to their mean rewards being close (see examples in Appendix D.5.1).

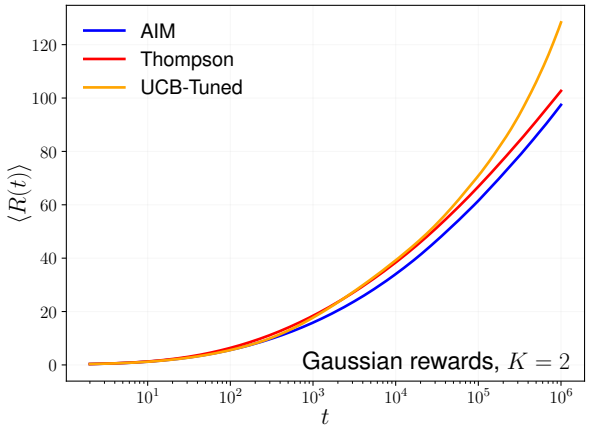

Figure 2: Evolution of the Bayesian regret for 2-armed bandit with Gaussian rewards under a uniform mean prior. The regret is averaged over more than 10^5 runs.

AIM yields strong performance in the two-armed Gaussian rewards case, as predicted by our theoretical analysis. We now aim at extending our method to other bandit settings. Figs. 3 and 4 present the performance of AIM when adapted to Bernoulli reward distributions and bandits with more than two arms. These adaptations are described in detail in Section 6 below.

Fig. 3 considers Bernoulli distributed rewards (Pilarski et al., 2021) with arm means drawn uniformly in $[0, 1]$. The performance of AIM is comparable to Thompson

sampling here. Additionally, AIM shows comparable performance to Thompson sampling for close mean rewards (see Appendix D.5.2).

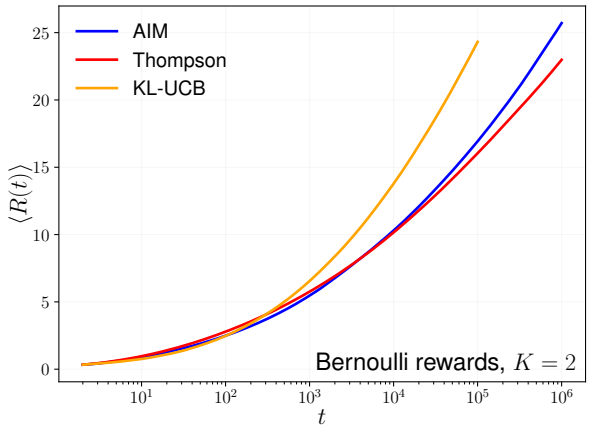

Figure 3: Evolution of the Bayesian regret for 2-armed bandit with Bernoulli rewards under a uniform mean prior. The regret is averaged over more than 10^5 runs.

Fig. 4 investigates AIM, for 50 arms with Bernoulli rewards and means drawn uniformly in $[0, 1]$. Its 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 time as Thompson sampling.

These observations support the robustness of AIM and its potential for extensions to more complex bandit settings.

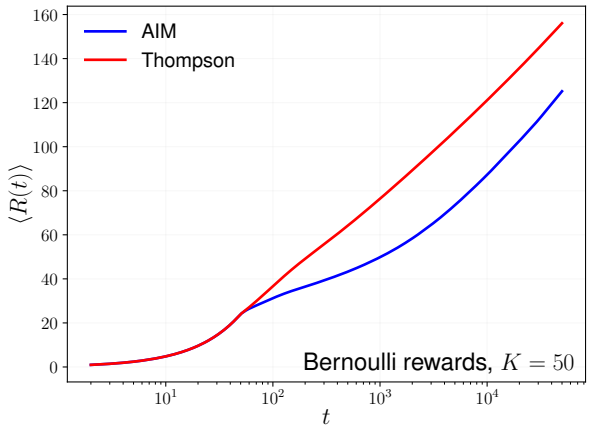

Figure 4: Evolution of the Bayesian regret for 50-armed bandit with Bernoulli rewards under a uniform mean prior. The regret is averaged over 4×10^4 runs.

6 Extensions

We have applied our approach (above) to Bernouilli bandits with many arms, where it shows strong empirical performances (see Figs. 3 and 4). This section describes the extensions of AIM to these cases and discusses potential extensions to more general bandit settings.

Exponential family bandits. Since Eq. (4) explicitly relies on arm 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 C.1 for details on such distributions), an asymptotic and analytic expression of the entropy can be derived for the case of uniform priors (see Appendix C for more details), yielding

$$\tilde{S}_{\max} = \frac{1}{2}\ln(2\pi\bar{\sigma}_i^2)\left[1 - \frac{e^{-N_{m_t}\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})}}{N_{m_t}\partial_2\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})\sqrt{2\pi\bar{\sigma}^2_{m_t}}}\right] + \frac{\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})e^{-N_{m_t}\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})}}{\partial_2\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})\sqrt{2\pi\bar{\sigma}^2_{m_t}}}. \tag{14}$$

Here $\mathrm{KL}(\hat{\mu}_i, \tilde{\mu}_{\mathrm{eq}})$ is the Kullback-Leibler divergence between the reward distributions, parameterized by $\hat{\mu}_i$ and $\tilde{\mu}_{\mathrm{eq}}$, respectively, 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 on their asymptotic simplifications obtained afterwards with Laplace's method. Our implementation of AIM to Bernoulli rewards (a specific case of the exponential family) with Eq. (14) shows comparable performance to state-of-the-art algorithms (see Fig. 3), supporting its adaptability to general settings. We believe that AIM should be optimal for all exponential family reward distributions as well as general prior distributions (see Appendix C for a detailed discussion).

Extension to $K > 2$. The functional form of Eq. (3) offers a straightforward extension to AIM for more than two arms (Reddy et al., 2016). As for the previous settings, a similar set of simplifications have to be derived to extract an analytic expression. We will again consider the best empirical arm as the main component of the $p_{\max}$ mode and approximate the worse empirical arms contributions as essentially concentrated in the tail. We thus keep a partitioning similar to Eq. (6) as

$$\tilde{S}_{\mathrm{tail}} = -\sum_{i \neq M_t}^{K} \int_{\tilde{\mu}_{\mathrm{eq,i}}}^{\mu_{\sup}} p_i(\theta)\ln p_i(\theta)\mathrm{d}\theta, \tag{15}$$

where $\tilde{\mu}_{\text{eq,i}}$ approximates the value where arm i has the same probability of being the maximum as the better empirical arm and

$$\tilde{S}_{\text{body}} = -\int_{\Theta} \big(1 - \sum_{i \neq M_t}^{K} [1 - C_i(\theta)]\big) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta. \tag{16}$$

Notably, Eqs. (15) and (16) summations come from the expansion at the lower order of the worse empirical arms contributions (see Appendix C.8 for derivation details). Hence, Eqs. (10), (11) and (14) can be directly used to get tractable expressions for any finite arms setting. An implementation[2] for Bernoulli rewards with $K = 50$ is displayed Fig. 4, where AIM shows comparable performances to Thompson Sampling on large times and strongly outperform Thompson sampling on short time scales. Finally, since Eqs. (15) and (16) resemble the expression in the case of two arms, we believe that AIM should also be optimal for $K > 2$ arms and that similar proof techniques can be used.

Other bandits settings. At last, we provide a quick overview of several more complex bandit settings for which the information maximization should build to the specific bandit structure of the problem to provide efficient algorithms. First, we emphasize that AIM partition between body/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, let us stress that information can also be quantified for unpulled arms, which may prove crucial when facing a large number of arms. The agent could quantify the information of the "reservoir" of unpulled arms, to anticipate the information gain of exploring these unpulled arms. Additionally, if the agent has access to the remaining time, it can not only evaluate the expected information gain when pulling an arm for a single round, but instead evaluate the information gain of multiple pulls of the same arm. We believe that such a consideration might be pivotal when facing many arms, since the limited amount of time does not allow to pull all the arms (Bayati et al., 2020) sufficiently. Thirdly, one can consider linear bandits, where arms are correlated with each other (Li et al., 2010). Because pulling a specific direction directly provides side information on correlated directions, the shared information gain could be leveraged by information-based methods to yield strong performances. Finally, one could consider 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, thus making the information maximization principle an inherent candidate when a suitable entropy is derived from the underlying bandit structure and problem objective. A last advantage of AIM lies in its possible extension to multiple constraints that would be introduced using Lagrange multipliers (or borrowed from physics reasoning by defining a free energy), further improving its adaptability to various settings and to specific requirements.

[2]We refer to Appendix D.3 for implementation details.

7 Conclusion

This paper introduces a new algorithm class, AIM, which leverages approximate information maximization 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 schemes. It enables easily tunable and tractable algorithms, which we prove to be optimal in for two-armed Gaussian bandits. Numerical experiments for Bernoulli rewards with two or several arms emphasise 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. 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 some wisely selected information measure can efficiently apprehend the arms' structure and correlations.

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.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambert w function. *Advances in Computational mathematics*, 5:329–359, 1996.

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.

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.

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.

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.

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. *J Stat Phys*, 163(6):1454–1476, June 2016. ISSN 1572-9613. 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.

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.

A 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.

A.1 The partitioning approximation

We start by commenting on the partition scheme and the approximations leading to the body/tail expressions summarized below.

$$\tilde{S}_{\rm tail} = -\int_{\tilde{\mu}_{\rm eq}}^{\mu_{\rm sup}} p_{m_t}(\theta)\ln p_{m_t}(\theta)d\theta, \text{ and } \tilde{S}_{\rm body} = -\int_{\Theta} p_{M_t}(\theta)C_{m_t}(\theta)\ln p_{M_t}(\theta)\mathrm{d}\theta. \tag{17}$$

We first remind $p_{\max}(\theta)$ expression with the arms order (M_t,m_t)

$$p_{\max}(\theta) = \left[C_{m_t}(\theta)p_{M_t}(\theta) + C_{M_t}(\theta)p_{m_t}(\theta)\right]. \tag{18}$$

We aim to keep the leading orders of $p_{\max}(\theta)$ when $N_{M_t} \gg N_{m_t} \gg 1$ and $\hat{\mu}_{M_t} > \hat{\mu}_{m_t}$. 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 $N_{M_t} > N_{m_t}$, $p_{M_t}(\theta)$ is more concentrated than $p_{m_t}(\theta)$, resulting in the dominance of the second term in the distribution tail (for the high reward side).

Thus, by defining the intersection point $\tilde{\mu}_{\rm eq}$ (where $C_{m_t}(\tilde{\mu}_{\rm eq})p_{M_t}(\tilde{\mu}_{\rm eq}) = C_{M_t}(\tilde{\mu}_{\rm eq})p_{m_t}(\tilde{\mu}_{\rm eq})$), we decompose the entropy in the body/tail components defined in the main text. In the asymptotic regime, $\tilde{\mu}_{\rm eq}$ will verify $p_{m_t}(\theta) \gg p_{M_t}(\theta)$ for $\theta > \tilde{\mu}_{\rm eq}$ and $p_{m_t}(\theta) \ll p_{M_t}(\theta)$ for $\theta < \tilde{\mu}_{\rm eq}$. We neglect the transition regime where $\tilde{\mu}_{\rm eq} \sim \theta$ where both distributions are of the same order because it is narrow (in the asymptotic regime) and have a very little influence on the total value of the entropy.

We now consider the tail where $\theta > \tilde{\mu}_{\rm eq}$ implying $p_{m_t}(\theta) \gg p_{M_t}(\theta)$. Because $N_{m_t} \gg 1$, we get $C_{m_t}(\theta) \approx 1$. Noticing that $C_{M_t}(\theta) > C_{m_t}(\theta)$, we also get $C_{M_t}(\theta) \approx 1$. Applying these scaling ratios, we get rid of all the subdominant terms in $\tilde{S}_{\rm tail}$:

$$\begin{aligned}
-\tilde{S}_{\rm tail} &\sim \int_{\tilde{\mu}_{\rm eq}}^{\mu_{\rm sup}} \left[C_{m_t}(\theta)p_{M_t}(\theta) + C_{M_t}(\theta)p_{m_t}(\theta)\right]\log\left[C_{m_t}(\theta)p_{M_t}(\theta) + C_{M_t}(\theta)p_{m_t}(\theta)\right]\mathrm{d}\theta \\
&\sim \int_{\tilde{\mu}_{\rm eq}}^{\mu_{\rm sup}} C_{M_t}(\theta)p_{m_t}(\theta)\log\left[C_{M_t}(\theta)p_{m_t}(\theta)\right] + C_{m_t}(\theta)p_{M_t}(\theta) + o(C_{m_t}(\theta)p_{M_t}(\theta))\mathrm{d}\theta \\
&\qquad + \int_{\tilde{\mu}_{\rm eq}}^{\mu_{\rm sup}} C_{m_t}(\theta)p_{M_t}(\theta)\log\left[C_{m_t}(\theta)p_{M_t}(\theta) + C_{M_t}(\theta)p_{m_t}(\theta)\right]\mathrm{d}\theta \\
&\sim \int_{\tilde{\mu}_{\rm eq}}^{\mu_{\rm sup}} p_{m_t}(\theta)\log\left[p_{m_t}(\theta)\right]\left[1 + \mathcal{O}\left(1 - C_{M_t}(\tilde{\mu}_{\rm eq})\right)\right] + \mathcal{O}\left(p_{M_t}(\theta)\log[p_{m_t}(\theta)]\right)\mathrm{d}\theta, \\
&\sim \int_{\tilde{\mu}_{\rm eq}}^{\mu_{\rm sup}} p_{m_t}(\theta)\log\left[p_{m_t}(\theta)\right]\mathrm{d}\theta
\end{aligned} \tag{19}$$

where we used that $C_{M_t}(\theta) = 1 - \int_{\theta}^{\mu_{\rm sup}} p_{M_t}(\theta)d\theta < C_{M_t}(\tilde{\mu}_{\rm eq})$.

We finally consider the body term. Noticing that $p_{M_t}(\theta) \gg p_{m_t}(\theta)$, $\tilde{S}_{\rm body}$ reads

$$
\begin{aligned}
-\tilde{S}_{\text{body}} &\sim \int_{\mu_{\text{inf}}}^{\tilde{\mu}_{\text{eq}}} [C_{m_t}(\theta)p_{M_t}(\theta) + C_{M_t}(\theta)p_{m_t}(\theta)] \log [C_{m_t}(\theta)p_{M_t}(\theta) + C_{M_t}(\theta)p_{m_t}(\theta)] \, \mathrm{d}\theta \\
&\sim \int_{\mu_{\text{inf}}}^{\tilde{\mu}_{\text{eq}}} C_{m_t}(\theta)p_{M_t}(\theta) \log [C_{m_t}(\theta)p_{M_t}(\theta)] + C_{M_t}(\theta)p_{m_t}(\theta) + o(C_{M_t}(\theta)p_{m_t}(\theta))d\theta \\
&\qquad + \int_{\mu_{\text{inf}}}^{\tilde{\mu}_{\text{eq}}} C_{M_t}(\theta)p_{m_t}(\theta) \log [C_{m_t}(\theta)p_{M_t}(\theta) + C_{M_t}(\theta)p_{m_t}(\theta)] \, \mathrm{d}\theta \\
&\sim \int_{\mu_{\text{inf}}}^{\tilde{\mu}_{\text{eq}}} C_{m_t}(\theta)p_{M_t}(\theta) \log [p_{M_t}(\theta)] \, [1 + o(1)] + o(p_{M_t}(\theta))\mathrm{d}\theta,
\end{aligned}
\tag{20}
$$

where we have used that $p_{M_t}(\theta) \gg 1$ around the vicinity of $\hat{\mu}_{M_t}$. Since, the inner term of the integral is negligible for $\theta > \tilde{\mu}_{\text{eq}}$ (by definition of $\tilde{\mu}_{\text{eq}}$), we extend the upper bound of Eq. (20) to the full integration domain without loss of consistency. It will simplify the derivation of an analytical expression for the body component in the next section. Taking altogether the leading orders of Eqs. (19) and (20) gives Eq. (17).

A.2 Asymptotic of the intersection point

In this section we derive the asymptotic expression of the intersection point (defined above as $\tilde{\mu}_{\text{eq}}$) where the distributions $C_{m_t}(\tilde{\mu}_{\text{eq}})p_{M_t}(\tilde{\mu}_{\text{eq}})$ and $C_{M_t}(\tilde{\mu}_{\text{eq}})p_{m_t}(\tilde{\mu}_{\text{eq}})$ intersect at their highest value (if they intersect more than once). Here we consider Gaussian rewards. The exact equation verified by the intersection point $\bar{\mu}_{\text{eq}}$ is

$$
\frac{\sqrt{N_{M_t}} e^{-\frac{N_{M_t}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{1}{2} \left[1 + \text{erf}\left(\frac{\sqrt{N_{m_t}}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2}} \right) \right] = \frac{\sqrt{N_{m_t}} e^{-\frac{N_{m_t}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{m_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{1}{2} \left[1 + \text{erf}\left(\frac{\sqrt{N_{M_t}}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{M_t})}{\sqrt{2\sigma^2}} \right) \right].
\tag{21}
$$

Taking the logarithm of Eq. (21) and normalizing the last term leads to

$$
\frac{N_{m_t}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{m_t})^2}{2\sigma^2} - \frac{N_{M_t}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{M_t})^2}{2\sigma^2} + \frac{1}{2} \ln \frac{N_{M_t}}{N_{m_t}} + \ln \left[\frac{1 + \text{erf}\left(\frac{\sqrt{N_{m_t}}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2}} \right)}{1 + \text{erf}\left(\frac{\sqrt{N_{M_t}}(\bar{\mu}_{\text{eq}} - \hat{\mu}_{M_t})}{\sqrt{2\sigma^2}} \right)} \right] = 0.
\tag{22}
$$

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

$$
\tilde{\mu}_{\text{eq}} = \hat{\mu}_{M_t} + \frac{N_{m_t}(\hat{\mu}_{M_t} - \hat{\mu}_{m_t})}{N_{M_t} - N_{m_t}} + \sqrt{\frac{N_{M_t} N_{m_t}}{(N_{M_t} - N_{m_t})^2}(\hat{\mu}_{M_t} - \hat{\mu}_{m_t})^2 + \frac{\sigma^2}{N_{M_t} - N_{m_t}} \ln \left(\frac{N_{M_t}}{N_{m_t}} \right)}.
\tag{23}
$$

Note that Eq. (23) expression relies on both $\hat{\mu}_{M_t} > \hat{\mu}_{m_t}$ and $N_{M_t} > N_{m_t}$. For $N_{M_t} \leq N_{m_t}$, even if the above $\tilde{\mu}_{\text{eq}}$ can be computed, it doesn't quantify the tail contribution. As a matter of fact, for $N_{M_t} \leq N_{m_t}$, the tail is always dominated by p_{M_t} which means that it has been already included in the main mode $\tilde{S}_{\text{body}}$. Then, in this specific configuration, we take $\tilde{S}_{\text{tail}}$ equals to 0 and $\tilde{\mu}_{\text{eq}} = \mu_{\text{sup}}$.

A.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:

$$\tilde{S}_{\text{body}} = -\int_{-\infty}^{+\infty} \frac{\sqrt{N_{M_t}} e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{1}{2}\left[1+\text{erf}\left(\frac{\sqrt{N_{m_t}}(\theta-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)\right]\left(-\frac{1}{2}\ln(\frac{2\pi\sigma^2}{N_{M_t}}) - \frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}\right)\text{d}\theta, \tag{24}$$

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

$$\int_{-\infty}^{\infty}\left[1+\text{erf}\left(\frac{\theta-\theta_1}{\sqrt{2V_1}}\right)\right]\frac{e^{-\frac{(\theta-\theta_2)^2}{2V_2}}}{\sqrt{2\pi V_2}}\text{d}\theta = \left[1+\text{erf}\left(\frac{\theta_2-\theta_1}{\sqrt{2}\sqrt{V_2+V_1}}\right)\right], \tag{25}$$

which leads to

$$\int_{-\infty}^{+\infty} \frac{\sqrt{N_{M_t}} e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\left[1+\text{erf}\left(\frac{\sqrt{N_{m_t}}(\theta-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)\right]\text{d}\theta = \left[1+\text{erf}\left(\frac{\hat{\mu}_{M_t}-\hat{\mu}_{m_t}}{\sqrt{2\sigma^2(\frac{1}{N_{M_t}}+\frac{1}{N_{m_t}})}}\right)\right]. \tag{26}$$

Next, we integrate by parts the second term to obtain:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}\frac{\sqrt{N_{M_t}} e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\frac{1}{2}\left[1+\text{erf}\left(\frac{\sqrt{N_{m_t}}(\theta-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)\right]\text{d}\theta \\
&= \int_{-\infty}^{\infty}\frac{1}{4}\frac{\sqrt{N_{M_t}} e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\left[1+\text{erf}\left(\frac{\sqrt{N_{m_t}}(\theta-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)\right] + \frac{(\theta-\hat{\mu}_{M_t})}{2}\frac{\sqrt{N_{M_t}N_{m_t}} e^{-\frac{N_{M_t}(\theta-\hat{\mu}_{M_t})^2}{2\sigma^2}-\frac{N_{m_t}(\theta-\hat{\mu}_{m_t})^2}{2\sigma^2}}}{2\pi\sigma^2}\text{d}\theta \\
&= \frac{1}{4}\left[1+\text{erf}\left(\frac{\hat{\mu}_{M_t}-\hat{\mu}_{m_t}}{\sqrt{2\sigma^2(\frac{1}{N_{M_t}}+\frac{1}{N_{m_t}})}}\right)\right] + \frac{(\hat{\mu}_{m_t}-\hat{\mu}_{M_t})\sigma^2}{2N_{M_t}\sqrt{2\pi}(\frac{\sigma^2}{N_{M_t}}+\frac{\sigma^2}{N_{m_t}})^{3/2}} e^{-\frac{(\hat{\mu}_{M_t}-\hat{\mu}_{m_t})^2}{2\sigma^2(\frac{1}{N_{M_t}}+\frac{1}{N_{m_t}})}},
\end{aligned} \tag{27}$$

where we also rely on the identity of Eq. (25).

Combining Eqs. (26) and (27) leads to the analytical expression of the body component.

$$\tilde{S}_{\text{body}} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1-\frac{1}{2}\text{erfc}\left(\frac{\sqrt{N_{m_t}N_{M_t}}(\hat{\mu}_{M_t}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2(N_{m_t}+N_{M_t})}}\right)\right] - \frac{\sqrt{N_{M_t}}N_{m_t}^{3/2}(\hat{\mu}_{M_t}-\hat{\mu}_{m_t})}{2\sigma\sqrt{2\pi}(N_{M_t}+N_{m_t})^{3/2}} e^{-\frac{N_{m_t}N_{M_t}(\hat{\mu}_{M_t}-\hat{\mu}_{m_t})^2}{2\sigma^2(N_{m_t}+N_{M_t})}}. \tag{28}$$

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

$$\tilde{S}_{\text{b}} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1-\frac{1}{2}\text{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)\right]. \tag{29}$$

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

A.4 Close form and asymptotic expression for the entropy tail

The contribution from the tail can be derived exactly and reads

$$\begin{aligned}\tilde{S}_{\text{tail}} &= \int_{\tilde{\mu}_{\text{eq}}}^{\infty} \frac{\sqrt{N_{m_t}} e^{-\frac{N_{m_t}(\theta-\hat{\mu}_{m_t})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left[\frac{1}{2}\ln(\frac{2\pi\sigma^2}{N_{m_t}}) + \frac{N_{m_t}(\theta-\hat{\mu}_{m_t})^2}{2\sigma^2}\right] \mathrm{d}\theta \\ &= \frac{1}{4}\ln\left(\frac{2\pi\sigma^2 e}{N_{m_t}}\right)\mathrm{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right) + \frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{2\sqrt{2\pi\sigma^2}} e^{-\frac{N_{m_t}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}.\end{aligned} \tag{30}$$

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

$$\tilde{S}_{\text{t}} = \frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{2\sqrt{2\pi\sigma^2}} e^{-\frac{N_{m_t}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}. \tag{31}$$

Taken altogether, Eqs. (29) and (31) leads to the desired simplified approximation of the entropy:

$$\tilde{S}_{\text{m}} = \tilde{S}_{\text{b}} + \tilde{S}_{\text{t}}. \tag{32}$$

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

Since, Eqs. (29) and (31) 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. The posterior of the reward obtained at time $N_{m_t} + N_{M_t} + 1$ is approximated as a Gaussian of variance σ^2 and centred around $\hat{\mu}_{M_t}$, leading for the increment evaluation to:

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

For the sake of simplicity, we neglect the variations of all the subdominant terms inside $\tilde{\mu}_{\text{eq}}$ meaning we approximate $\tilde{\mu}_{\text{eq}}$ as $\tilde{\mu}_{\text{eq}}(\hat{\mu}_{M_t} + \frac{\mu}{N_{M_t}+1}, N_{M_t}+1, \hat{\mu}_{m_t}, N_{m_t}) \approx \tilde{\mu}_{\text{eq}}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_{m_t}, N_{m_t}) + \frac{\mu}{N_{M_t}+1}$, 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. (25), one can show that the gradient of the body component $\Delta_{M_t}\tilde{S}_{\max}$ rewrites as

$$\Delta_{M_t}\tilde{S}_{\text{b}} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}+1})\left[1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}\sqrt{1+\frac{N_{m_t}}{(N_{M_t}+1)^2}}}\right)\right] - \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left[1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)\right]. \tag{34}$$

Next, we consider the increment of the tail component along the better empirical arm:

$$\begin{aligned}\Delta_{M_t}\tilde{S}_{\text{t}} &= \int_{-\infty}^{\infty} \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{\sqrt{N_{m_t}}(\frac{\mu}{N_{M_t}+1} + \tilde{\mu}_{\text{eq}} - \hat{\mu}_{m_t})}{2\sqrt{2\pi\sigma^2}} e^{-\frac{N_{m_t}(\tilde{\mu}_{\text{eq}}+\frac{\mu}{N_{M_t}+1}-\hat{\mu}_{m_t})^2}{2\sigma^2}} - \tilde{S}_{\text{t}}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_{m_t}, N_{m_t})\mathrm{d}\mu \\ &= e^{-N_{m_t}\frac{(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2(1+\frac{N_{m_t}}{(1+N_{M_t})^2})}} \sqrt{\frac{N_{m_t}}{8\pi\sigma^2}} \frac{(\tilde{\mu}_{\text{eq}}-\hat{\mu}_{m_t})}{(1+\frac{N_{m_t}}{(N_{M_t}+1)^2})^{3/2}} - \tilde{S}_{\text{t}}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_{m_t}, N_{m_t}).\end{aligned} \tag{35}$$

Next, we consider the increment evaluation along the worst empirical arm.

$$\Delta_{m_t}\tilde{S}_{\rm m} = \int_{-\infty}^{\infty} \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left[\tilde{S}_{\rm m}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_{m_t} + \frac{\mu}{N_{m_t}+1}, N_{m_t}+1) - \tilde{S}_{\rm m}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_{m_t}, N_{m_t})\right] {\rm d}\mu. \tag{36}$$

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

$$\Delta_{m_t}\tilde{S}_{\rm b} = \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}}) \left[1 - {\rm erfc}\left(\frac{(N_{m_t}+1)(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2(N_{m_t}+2)}}\right)\right] - \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}}) \left[1 - \frac{1}{2}{\rm erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)\right]. \tag{37}$$

Finally, we consider the last component of the increment along the worst empirical arm:

$$\begin{aligned}\Delta_{m_t}\tilde{S}_{\rm t} &= \int_{-\infty}^{\infty} \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{\sqrt{N_{m_t}+1}(\frac{\mu}{N_{m_t}+1} + \tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{2\sqrt{2\pi\sigma^2}} e^{-\frac{(N_{m_t}+1)(\tilde{\mu}_{\rm eq} + \frac{\mu}{N_{m_t}+1} - \hat{\mu}_{m_t})^2}{2\sigma^2}} - \tilde{S}_{\rm t}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_{m_t}, N_{m_t}){\rm d}\mu \\ &= e^{-\frac{(N_{m_t}+1)^2}{(N_{m_t}+2)}\frac{(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})^2}{2\sigma^2}} \frac{(1+N_{m_t})^2(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{8\pi\sigma^2}(2+N_{m_t})^{3/2}}. - \tilde{S}_{\rm t}(\hat{\mu}_{M_t}, N_{M_t}, \hat{\mu}_{m_t}, N_{m_t})\end{aligned} \tag{38}$$

Taken altogether, Eqs. (34), (35), (37) and (38) leads to the final analytical expression of the increment:

$$\begin{aligned}\Delta_{M_t}\tilde{S}_{\rm m} - \Delta_{m_t}\tilde{S}_{\rm m} = &\frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}+1}) \left[1 - \frac{1}{2}{\rm erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2}\sqrt{1+\frac{N_{m_t}}{(N_{M_t}+1)^2}}}\right)\right] \\ &- \frac{1}{2}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}}) \left[1 - \frac{1}{2}{\rm erfc}\left(\frac{(N_{m_t}+1)(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2(N_{m_t}+2)}}\right)\right] \\ &+ e^{-N_{m_t}\frac{(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})^2}{2\sigma^2(1+\frac{N_{m_t}}{(1+N_{M_t})^2})}} \sqrt{\frac{N_{m_t}}{8\pi\sigma^2}} \frac{(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{(1+\frac{N_{m_t}}{(N_{M_t}+1)^2})^{3/2}} - e^{-\frac{(N_{m_t}+1)^2}{(N_{m_t}+2)}\frac{(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})^2}{2\sigma^2}} \frac{(1+N_{m_t})^2(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{8\pi\sigma^2}(2+N_{m_t})^{3/2}},\end{aligned} \tag{39}$$

Which rewrites as:

$$\begin{aligned}\Delta_{M_t}\tilde{S}_{\rm m} - \Delta_{m_t}\tilde{S}_{\rm m} = &\frac{1}{2}\ln(\frac{N_{M_t}}{N_{M_t}+1}) \\ &- \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}+1}){\rm erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2}\sqrt{1+\frac{N_{m_t}}{(N_{M_t}+1)^2}}}\right) + \frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}}){\rm erfc}\left(\frac{(N_{m_t}+1)(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{2\sigma^2(N_{m_t}+2)}}\right) \\ &+ e^{-N_{m_t}\frac{(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})^2}{2\sigma^2(1+\frac{N_{m_t}}{(1+N_{M_t})^2})}} \sqrt{\frac{N_{m_t}}{8\pi\sigma^2}} \frac{(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{(1+\frac{N_{m_t}}{(N_{M_t}+1)^2})^{3/2}} - e^{-\frac{(N_{m_t}+1)^2}{(N_{m_t}+2)}\frac{(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})^2}{2\sigma^2}} \frac{(1+N_{m_t})^2(\tilde{\mu}_{\rm eq} - \hat{\mu}_{m_t})}{\sqrt{8\pi\sigma^2}(2+N_{m_t})^{3/2}}.\end{aligned} \tag{40}$$

To finally obtain a simplified expression, we make an expansion in the first order for each component of the last two terms. We consider the second term denoted as T_2:

$$
\begin{aligned}
T_2 &= -\frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}+1})\mathrm{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}\sqrt{1+\frac{N_{m_t}}{(N_{M_t}+1)^2}}}\right)+\frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\mathrm{erfc}\left(\frac{(N_{m_t}+1)(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2(N_{m_t}+2)}}\right)\\
&\approx \frac{1}{4N_{M_t}}\mathrm{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)-\frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_{M_t}})\left(\frac{1}{N_{m_t}^2}+\frac{N_{m_t}}{N_{M_t}^2}\right)\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\pi\sigma^2}}e^{-\frac{N_{m_t}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}.
\end{aligned}
\tag{41}
$$

Finally, the last term denoted as T_3 reads:

$$
\begin{aligned}
T_3 &= e^{-N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2(1+\frac{N_{m_t}}{(1+N_{M_t})^2})}}\sqrt{\frac{N_{m_t}}{8\pi\sigma^2}}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{(1+\frac{N_{m_t}}{(N_{M_t}+1)^2})^{3/2}}-e^{-\frac{(N_{m_t}+1)^2}{(N_{m_t}+2)}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}\frac{(1+N_{m_t})^2(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{8\pi\sigma^2}(2+N_{m_t})^{3/2}}\\
&\approx -\frac{3N_{m_t}}{2N_{M_t}^2}e^{-N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}\sqrt{\frac{N_{m_t}}{8\pi\sigma^2}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})+\frac{1}{N_{m_t}}e^{-N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{8\pi\sigma^2}}\\
&\quad+\frac{N_{m_t}}{N_{M_t}^2}N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}e^{-N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}\sqrt{\frac{N_{m_t}}{8\pi\sigma^2}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})\\
&\quad+\frac{1}{N_{m_t}^2}N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}e^{-N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}\sqrt{\frac{N_{m_t}}{8\pi\sigma^2}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})\\
&\approx e^{-N_{m_t}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}\sqrt{\frac{N_{m_t}}{8\pi\sigma^2}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})\left[\frac{1}{N_{m_t}}-\frac{3N_{m_t}}{2N_{M_t}^2}+\frac{N_{m_t}^2}{N_{M_t}^2}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}+\frac{1}{N_{m_t}}\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}\right].
\end{aligned}
\tag{42}
$$

Taken altogether, we finally obtain the following simplified increment used for AIM:

$$
\begin{aligned}
\Delta = \frac{1}{2}\ln(\frac{N_{M_t}}{N_{M_t}+1})+\frac{1}{4N_{M_t}}\mathrm{erfc}\left(\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\sigma^2}}\right)+\frac{\sqrt{N_{m_t}}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})}{\sqrt{2\pi\sigma^2}}e^{-\frac{N_{m_t}(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{2\sigma^2}}\times\\
\left[\frac{1}{4}\ln(\frac{N_{M_t}}{2\pi\sigma^2 e})\left(\frac{1}{N_{m_t}^2}+\frac{N_{m_t}}{N_{M_t}^2}\right)+\frac{1}{2N_{m_t}}-\frac{3N_{m_t}}{4N_{M_t}^2}+\left(\frac{N_{m_t}^2}{N_{M_t}^2}+\frac{1}{N_{m_t}}\right)\frac{(\tilde{\mu}_{\mathrm{eq}}-\hat{\mu}_{m_t})^2}{4\sigma^2}\right].
\end{aligned}
\tag{43}
$$

B 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 two-armed bandits with Gaussian rewards of unit variance, for any $\varepsilon \in (0, \frac{1}{2})$, there exists a constant $C(|\mu_1-\mu_2|,\varepsilon)\in\mathbb{R}$ depending solely on $|\mu_1-\mu_2|$ and ε such that for any $T\in\mathbb{N}$*

$$
R(T) \le \frac{2\sigma^2\ \ln T}{(1-\varepsilon)|\mu_1-\mu_2|}+\frac{3\sigma^2\ \ln\ln T}{(1-\varepsilon)|\mu_1-\mu_2|}+C(|\mu_1-\mu_2|,\varepsilon).
$$

Proof. We assume in the whole proof and without loss of generality that $\mu_1>\mu_2$. For $\Delta_2=\mu_1-\mu_2$, the regret can then be written as

$$
R(T)=\Delta_2\mathbb{E}\left[N_2(T)\right]=\Delta_2\sum_{t=1}^{T}\mathbb{P}(a_t=2).
$$

For some $b \in (0,1)$, we decompose this expectation in 4 terms as follows

$$\mathbb{E}\left[N_2(T)\right] \leq \sum_{t=1}^{T} \mathbb{P}\left(N_1(t) \leq t^b\right) + \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_2(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}}, a_t = 2, N_1(t) \geq t^b\right)$$
$$+ \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_1(t) \leq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}}, N_1(t) \geq t^b\right) + \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_1(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}} \geq \hat{\mu}_2(t), a_t = 2, N_1(t) \geq t^b\right).$$

This inequality comes simply by noticing the event $\{a_t = 2\}$ 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 $C(b, \Delta_2)$ depending solely on b and Δ_2, so that

$$\mathbb{E}\left[N_2(T)\right] \leq \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_1(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}} \geq \hat{\mu}_2(t), a_t = 2, N_1(t) \geq t^b\right) + C(b, \Delta_2).$$

Thanks to Lemma 5, there exist constants $t(\Delta_2), n(\Delta_2), c_4(\Delta_2)$ depending solely on Δ_2 such that

$$\sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_1(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}} \geq \hat{\mu}_2(t), a_t = 2, N_1(t) \geq t^b\right) \leq t(\Delta_2) + \sum_{t=1}^{T} \mathbb{P}\left(\mathcal{E}_1(t)\right) + \mathbb{P}\left(\mathcal{E}_2(t)\right) + \mathbb{P}\left(\mathcal{E}_3(t)\right),$$

where

$$\mathcal{E}_1(t) = \{N_2(t) \leq n(\Delta_2), a_t = 2\},$$
$$\mathcal{E}_2(t) = \{\mu_2 - \hat{\mu}_2(t) \leq -\varepsilon\Delta_2, a_t = 2\},$$
$$\mathcal{E}_3(t) = \{N_2(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_2^2}\left(\ln t + \frac{3}{2}\ln\ln t\right) + c_4(\Delta_2), a_t = 2\}.$$

Let us now bound indivdually the sum corresponding to each of these 3 events. The first bound holds directly:

$$\sum_{t=1}^{T} \mathbb{P}(\mathcal{E}_1(t)) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{N_2(t) \leq n(\Delta_2), a_t = 2\}\right]$$
$$\leq n(\Delta_2).$$

The second one can be bounded using Hoeffding's inequality. Indeed, for independent random variables $Z_2(n) \sim \mathcal{N}(\mu_2, \sigma^2)$, it reads as:

$$\sum_{t=1}^{T} \mathbb{P}\left(\mu_2 - \hat{\mu}_2(t) \leq -\varepsilon\Delta_2, a_t = 2\right) \leq \sum_{n=1}^{T} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_2(i) - \mu_2 \geq \varepsilon\Delta_2\right)$$
$$\leq \sum_{n=1}^{T} e^{-\frac{n\varepsilon^2\Delta_2^2}{2\sigma^2}}$$
$$\leq \frac{1}{e^{\frac{\varepsilon^2\Delta_2^2}{2\sigma^2}} - 1}.$$

The bound of the last term is similar to the bound for $\mathcal{E}_1(t)$, so that we get:

$$\sum_{t=1}^{T} \mathbb{P}(\mathcal{E}_3(t)) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_2^2}\left(\ln T + \frac{3}{2}\ln\ln T\right) + c_4(\Delta_2) + 1.$$

Wrapping up everything finally yields that for some constant $C(\Delta_2, \varepsilon)$ depending solely on Δ_2, ε,

$$R(T) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_2} \ln T + \frac{3\sigma^2}{(1-2\varepsilon)^2\Delta_2} \ln\ln T + C(\Delta_2, \varepsilon).$$

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

B.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 assume in this whole section that $\mu_1 > \mu_2$.

Lemma 1. *For any $b \in (0,1)$, there exists a constant $C_0(b, \Delta_2)$ depending solely on b and Δ_2 such that*

$$\sum_{t=1}^{\infty} \mathbb{P}(N_1(t) \leq t^b) \leq C_0(b, \Delta_2).$$

Proof. Let $b \in (0,1)$ and $t_0(b, \Delta_2)$ a large constant that depends solely on b and Δ_2. In the remaining of the proof, we assume at some points that $t_0(b, \Delta_2)$ is chosen large enough (but only larger than a threshold depending on b and Δ_2) such that some inequalities hold.

Assume that for $t \geq t_0(b, \Delta_2)$, $N_1(t) \leq t^b$. Necessarily, this implies the arm 2 is pulled at some time $t' \leq t$ where $N_1(t') \leq t^b$ and $N_2(t') \geq t - t^b - 1$. We choose $t_0(b, \Delta_2)$, so that it implies $N_2(t') > N_1(t')$. We thus necessarily have $\hat{\mu}_2(t') > \hat{\mu}_1(t')$. The arm 2 is thus pulled at time t' because $S_2 \geq S_1$ (i.e. $\Delta < 0$), where

$$S_2 = \frac{1}{2} \ln \left(1 + \frac{1}{N_2(t')} \right),$$

$$S_1 = \frac{1}{4N_2(t')} \operatorname{erfc} \left(\frac{\sqrt{N_1(t')}(\tilde{\mu}_{\text{eq}} - \hat{\mu}_1)}{\sqrt{2\sigma^2}} \right) + \frac{(\tilde{\mu}_{\text{eq}} - \hat{\mu}_1)}{\sqrt{2\pi\sigma^2 N_1(t')}} \times$$
$$e^{-\frac{N_1(t')(\tilde{\mu}_{\text{eq}} - \hat{\mu}_1)^2}{2\sigma^2}} \left(-\frac{1}{4} \ln(\frac{2\pi\sigma^2 e}{N_2(t')}) \left(\frac{1}{N_1(t')} + \frac{N_1(t')^2}{N_2(t')^2} \right) + \frac{1}{2} - \frac{3N_1(t')^2}{4N_2(t')^2} + (1 + \frac{N_1(t')^3}{N_2(t')^2}) \frac{(\tilde{\mu}_{\text{eq}} - \hat{\mu}_1)^2}{4\sigma^2} \right).$$

To simplify, note that $S_2 \leq \frac{1}{2N_2(t')}$. Moreover since $N_2(t') \geq t - t^b - 1 \geq 2\pi e^4 \sigma^2$ for a large enough choice of $t_0(b, \Delta_2)$, S_1 can be easily lower bounded as

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

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

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

Recall that $N_2(t') \geq t - t^b - 1$, so that Eq. (44) can be rewritten as

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

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

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

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

$$\tilde{x} \geq \sqrt{\frac{\ln(\frac{t - t^b - 1}{t^b})}{t - 2t^b - 1}} = \Omega \left(\sqrt{\frac{(1-b)\ln(t)}{t}} \right).$$

Moreover by subadditivity of the square root:

$$\tilde{x} \leq \sqrt{\frac{N_1(t')}{2\sigma^2}} (\hat{\mu}_2 - \hat{\mu}_1) \left(1 + \frac{N_1(t') + \sqrt{N_1(t') N_2(t')}}{N_2(t') - N_1(t')} \right) + \sqrt{\frac{N_1(t') \ln(\frac{N_2(t')}{N_1(t')})}{2(N_2(t') - N_1(t'))}} \tag{47}$$

$$\leq \sqrt{\frac{N_1(t')}{2\sigma^2}} (\hat{\mu}_2 - \hat{\mu}_1) \left(1 + \mathcal{O}\left(t^{\frac{b-1}{2}} \right) \right) + \mathcal{O}\left(\sqrt{\ln(t)} t^{\frac{b-1}{2}} \right). \tag{48}$$

Let us now consider the events

$$\mathcal{H}_1(t) \coloneqq \left\{ \exists s \leq t, \hat{\mu}_1(s) - \mu_1 \leq -\sqrt{\frac{2\sigma^2(\ln(t) - \ln\ln(t))}{N_1(s)}} - \frac{\Delta_2}{3} \right\}, \tag{49}$$

$$\mathcal{H}_2(t) \coloneqq \left\{ \exists s \leq t, t - t^b - 1 \leq N_2(s) \leq t \text{ and } \hat{\mu}_2(s) - \mu_2 \geq \frac{\Delta_2}{3} \right\}. \tag{50}$$

Assume in the following that $\neg\mathcal{H}_1(t) \cap \neg\mathcal{H}_2(t)$. This implies that

$$\hat{\mu}_2 - \hat{\mu}_1 \leq -\frac{\Delta_2}{3} + \sqrt{\frac{2\sigma^2(\ln(t) - \ln\ln(t))}{N_1(s)}}. \tag{51}$$

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 $\tilde{x} \leq \sqrt{\ln(t) - \ln\ln(t)} + \mathcal{O}\left(\sqrt{\frac{\ln(t)}{t^{1-b}}}\right)$. Using the lower and upper bounds on $\tilde{x}$, we have thanks to Eq. (46) that under $\neg\mathcal{H}_1(t) \cap \neg\mathcal{H}_2(t)$,

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

For a large enough choice of $t_0(b, \Delta_2)$, this last equality along with Eq. (51) actually yield $\hat{\mu}_2 - \hat{\mu}_1 < 0$, which contradicts the beginning of the proof. By contradiction, we thus showed the following event inclusion for $t \geq t_0(b, \Delta)$:

$$\{N_1(t) \leq t^b\} \subset \mathcal{H}_1(t) \cup \mathcal{H}_2(t). \tag{52}$$

Lemma 1 then follows, thanks to Lemma 2 below,

$$\sum_{t=1}^{\infty} \mathbb{P}(N_1(t) \leq t^b) \leq t_0(b, \Delta_2) + \sum_{t=t_0(b,\Delta_2)+1}^{\infty} \mathbb{P}(\mathcal{H}_1(t)) + \mathbb{P}(\mathcal{H}_2(t)).$$

□

Lemma 2. *For any $b \in (0,1)$, the events $\mathcal{H}_1(t), \mathcal{H}_2(t)$ defined in Eqs. (49) and (50), there exist constants c_1 and c_2 depending solely on Δ_2 and b such that*

$$\sum_{t=1}^{\infty} \mathbb{P}(\mathcal{H}_1(t)) \leq c_1 \quad \text{and} \quad \sum_{t=1}^{\infty} \mathbb{P}(\mathcal{H}_2(t)) \leq c_2.$$

Proof. The two bounds directly result from Hoeffding's inequality. Consider independent random variables $(Z_1(n))_{n\in N}, (Z_2(n))_{n\in N}$ where $Z_k(n) \sim \mathcal{N}(\mu_k, \sigma^2)$. Let us first bound the probability of $\mathcal{H}_2(t)$, which is simpler.

$$\begin{aligned}
\mathbb{P}(\mathcal{H}_2(t)) &\leq \sum_{n=\lceil t-t^b-1\rceil}^{t} \mathbb{P}\left(\sum_{i=1}^{n}(Z_2(i) - \mu_2) \geq \frac{n\Delta}{3}\right) \\
&\leq \sum_{n=\lceil t-t^b-1\rceil}^{t} e^{-\frac{n\Delta^2}{18\sigma^2}} \\
&\leq \frac{e^{-\frac{\lceil t-t^b-1\rceil\Delta_2^2}{18\sigma^2}}}{1 - e^{-\frac{\Delta_2^2}{18\sigma^2}}}.
\end{aligned}$$

The second inequality of Lemma 2 then follows by noting that the last term is summable over t for $b \in (0,1)$.

For the second bound, we also have by Hoeffding's inequality

$$\begin{aligned}\mathbb{P}(\mathcal{H}_1(t)) &\leq \sum_{n=1}^{\infty} \mathbb{P}\left(\sum_{i=1}^{n}(Z_1(i)-\mu_1) \leq -\sqrt{2n\sigma^2(\ln(t)-\ln\ln(t))} - \frac{n\Delta_2}{3}\right)\\ &\leq \sum_{n=1}^{\infty} \exp\left(-\ln(t)+\ln\ln(t) - \sqrt{2n\sigma^2(\ln(t)-\ln\ln(t))}\frac{\Delta_2}{3\sigma^2} - \frac{n\Delta_2^2}{18\sigma^2}\right)\\ &\leq \frac{\ln(t)}{t}\exp\left(-\sqrt{2(\ln(t)-\ln\ln(t))}\frac{\Delta_2}{3\sqrt{\sigma^2}}\right)\sum_{n=1}^{\infty} e^{-\frac{n\Delta_2^2}{18\sigma^2}}.\end{aligned}$$

The last sum is obviously finite. Moreover, $\sqrt{2(\ln(t)-\ln\ln(t))} = \Omega(\alpha\ln\ln(t))$ for any $\alpha > 0$, so that $\exp\left(-\sqrt{2(\ln(t)-\ln\ln(t))}\frac{\Delta_2}{3\sqrt{\sigma^2}}\right) = \mathcal{O}\left(\frac{1}{\ln^{\alpha}(t)}\right)$ 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 $b \in (0,1)$, there exists a constant $C_1(b,\Delta_2)$ depending solely on b and Δ_2 such that*

$$\sum_{t=1}^{\infty}\mathbb{P}\left(\hat{\mu}_2(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2\ln t}{N_1(t)}}, a_t = 2, N_1(t) \geq t^b\right) \leq C_1(b,\Delta_2).$$

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

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{P}\left(\hat{\mu}_2(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2\ln t}{N_1(t)}}, a_t = 2, N_1(t) \geq t^b\right) &\leq \sum_{t=1}^{T}\sum_{n=0}^{t}\mathbb{P}\left(\hat{\mu}_2(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2\ln t}{t^b}}, N_2(t)=n, N_2(t+1)=n+1\right)\\ &\leq \sum_{n=0}^{T}\sum_{t=n}^{T}\mathbb{P}\left(\underbrace{\hat{\mu}_2(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2\min_{s\geq n}\ln s}{s^b}}, N_2(t)=n, N_2(t+1)=n+1}_{:=\mathcal{G}_1(t,n)}\right).\end{aligned}$$

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

$$\sum_{t=n}^{T}\mathbb{P}(\mathcal{G}_1(t,n)) = \mathbb{P}\left(\exists t \in [n,T], \hat{\mu}_2(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2\min_{s\geq n}\ln s}{s^b}}, N_2(t)=n\right).$$

For independent random variables $Z_2(n) \sim \mathcal{N}(\mu_2,\sigma^2)$, we have by independence of the X_t and a_t, and then by Hoeffding inequality:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{P}\left(\hat{\mu}_2(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2\ln t}{N_1(t)}}, a_t = 2, N_1(t) \geq t^b\right) &\leq 1 + \sum_{n=1}^{T}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_2(i)-\mu_2) \geq \Delta_2 - \sqrt{\frac{6\sigma^2\min_{s\geq n}\ln s}{s^b}}\right),\\ &\leq 1 + \sum_{n=1}^{T}\exp\left(-\frac{n\left(\Delta_2 - \sqrt{\frac{6\sigma^2\min_{s\geq n}\ln s}{s^b}})\right)^2}{2\sigma^2}\right).\end{aligned}$$

Obviously, this sum can be bounded for any $T \in \mathbb{N}$ by a constant solely depending on Δ_2 and b. □

Lemma 4. *For any $b \in (0,1)$, there exists a universal constant C_2 such that*

$$\sum_{t=0}^{\infty}\mathbb{P}\left(\hat{\mu}_1(t) \leq \mu_1 - \sqrt{\frac{6\sigma^2\ln t}{N_1(t)}}, N_1(t) \geq t^b)\right) \leq C_2.$$

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

$$\mathbb{P}(N_1(t)\frac{(\hat{\mu}_1(t)-\mu_1)^2}{2\sigma^2} \geq (1+\alpha)\ln t) \leq 2\left\lceil\frac{\ln t}{\ln(1+\eta)}\right\rceil t^{-(1-\frac{\eta^2}{16})(1+\alpha)} \quad \text{for any } t \in \mathbb{N}^* \text{ and } \alpha,\eta > 0.$$

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

$$\mathbb{P}\left(\hat{\mu}_1(t) \leq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}}\right) \leq 2\frac{\ln(t)+1}{\ln(3)} t^{-\frac{9}{4}}.$$

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 inclusion of the considered events. □

For any $b \in (0,1)$, Lemma 5 below gives an event inclusion for the event

$$\mathcal{E}(t) \coloneqq \left\{\hat{\mu}_1(t) \geq \mu_1 - \sqrt{\frac{6\sigma^2 \ln t}{N_1(t)}} \geq \hat{\mu}_2(t), a_t = 2, N_1(t) \geq t^b\right\}. \tag{53}$$

Lemma 5. *There exist constants $t(\Delta_2)$, $n(\Delta_2)$ and $c_3(\Delta_2)$ depending solely on Δ_2 such that for any $t \geq t(\Delta_2)$ and $\varepsilon \in (0, \frac{1}{3})$,*

$$\begin{aligned}\mathcal{E}(t) \subset \{N_2(t) \leq n(\Delta_2), a_t = 2\} \cup \{\mu_2 - \hat{\mu}_2(t) \leq -\varepsilon\Delta_2, a_t = 2\} \\ \cup \{N_2(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_2^2}\big(\ln t + \frac{3}{2}\ln\ln t\big) + c_4(\Delta_2), a_t = 2\}.\end{aligned}$$

Proof. Assume in the following that $\mathcal{E}(t)$ holds for some $t \geq t(\Delta_2)$. First of all, $\hat{\mu}_1(t) \geq \mu_2(t)$ and $a_t = 2$ so that $N_1(t) \geq \frac{t}{2} \geq N_2(t)$. Moreover as we pulled the second arm, $S_2 \geq S_1$ where

$$\begin{aligned}S_1 &= \frac{1}{2}\ln(1 + \frac{1}{N_1(t)}) \\ S_2 &= \frac{1}{N_1(t)}\mathrm{erfc}\left(\frac{\sqrt{N_2(t)(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2)^2}}{\sqrt{2\sigma^2}}\right) + \frac{(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_1)\sqrt{N_2(t)}}{\sqrt{2\pi\sigma^2}} e^{-\frac{N_2(t)(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2)^2}{2\sigma^2}} \times \\ &\qquad \left(-\frac{1}{4}\ln(\frac{2\pi\sigma^2 e}{N_1(t)}))(\frac{1}{N_2(t)^2} + \frac{N_2(t)}{N_1(t)^2}) + \frac{1}{2N_2(t)} - \frac{3N_2(t)}{4N_1(t)^2} + (1 + \frac{N_2(t)^2}{N_1(t)^2})\frac{(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2)^2}{4\sigma^2}\right).\end{aligned}$$

In particular, for a large enough choice of the constant $t(\Delta_2)$,

$$S_2 \leq \frac{(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2)\sqrt{N_2(t)}}{\sqrt{2\pi\sigma^2}} e^{-\frac{N_2(t)(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2)^2}{2\sigma^2}} \left[\ln t + \frac{(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2)^2}{4\sigma^2}\right] + \frac{1}{4N_1(t)}\mathrm{erfc}\left(\frac{\sqrt{N_2(t)(\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2)^2}}{\sqrt{2\sigma^2}}\right).$$

Also, $S_1 \geq \frac{1}{t} - \frac{1}{t^2}$ since $N_1(t) \geq \frac{t}{2}$ and $\ln(1+x) \geq x - \frac{x^2}{2}$ for $x \in [0,1]$.

Now assume that both $\mu_2 - \hat{\mu}_2(t) \geq -\varepsilon\Delta_2$ and $N_2(t) \geq n(\Delta_2)$ for some constant $n(\Delta_2)$ depending only on Δ_2. It then holds

$$\begin{aligned}\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2(t) &\geq \hat{\mu}_1(t) - \hat{\mu}_2(t) \\ &\geq \Delta_2 + \mu_2 - \hat{\mu}_2(t) - \sqrt{\frac{12\sigma^2 \ln t}{t}} \\ &\geq (1-\varepsilon)\Delta_2 - \sqrt{\frac{12\sigma^2 \ln t}{t}}.\end{aligned}$$

Again, we can choose $t(\Delta_2)$ large enough so that $\tilde{\mu}_{\mathrm{eq}} - \hat{\mu}_2(t) \geq (1-2\varepsilon)\Delta_2$. Moreover, note that the functions $\mathrm{erfc}, x \mapsto xe^{-x^2}, x \mapsto x^3 e^{-x^2}$ are all decreasing on an interval of the form $[M, +\infty]$. As a consequence, we can choose $n(\Delta_2)$ large enough so that $\frac{\sqrt{n(\Delta_2)((1-2\varepsilon)\Delta_2)^2}}{\sqrt{2\sigma^2}} \geq M$. As $N_2(t) \geq n(\Delta_2)$, we then have

$$S_2 \leq \frac{(1-2\varepsilon)\Delta_2}{\sqrt{2\pi\sigma^2}}\sqrt{N_2(t)} e^{-\frac{N_2(t)(1-2\varepsilon)^2\Delta_2^2}{2\sigma^2}} \left[\ln t + \frac{(1-2\varepsilon)^2\Delta_2^2}{4\sigma^2} + 1\right].$$

The inequality $S_2 \geq S_1$ then implies, thanks to the above bounds:

$$\frac{(1-2\varepsilon)\Delta_2}{\sqrt{2\pi\sigma^2}}\sqrt{N_2(t)}e^{-\frac{N_2(t)(1-2\varepsilon)^2\Delta_2^2}{2\sigma^2}}\left[\ln t + \frac{\Delta_2^2+4\sigma^2}{4\sigma^2}\right] \geq \frac{1}{t} - \frac{1}{t^2}$$
$$\sqrt{x}e^{-x} \geq \left(\frac{1}{t} - \frac{1}{t^2}\right)\left(\frac{\sqrt{\pi}}{\ln t + \frac{\Delta_2^2+4\sigma^2}{4\sigma^2}}\right),$$

where $x = \frac{(1-2\varepsilon)^2\Delta_2^2}{2\sigma^2}N_2(t)$. Moreover, for a large enough choice of $n(\Delta_2)$, $x \geq 1$.

The inverse of the function $x \mapsto \sqrt{x}e^{-x}$ on $[1,\infty)$ is given by the function $y \mapsto -\frac{1}{2}W_{-1}(-2y^2)$, where W_{-1} denotes the negative branch of the Lambert W function, which is decreasing on the interval of interest. As a consequence, the previous inequality yields for $y = \left(\frac{1}{t} - \frac{1}{t^2}\right)\left(\frac{\sqrt{\pi}}{\ln t + \frac{\Delta_2^2+4\sigma^2}{4\sigma^2}}\right)$:

$$x \leq -\frac{1}{2}W_{-1}(-2y^2)$$
$$N_2(t) \leq -\frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_2^2}\frac{1}{2}W_{-1}(-2y^2).$$

Moreover, classical results on the Lambert function (Corless et al., 1996) yield that, $W_{-1}(z) \geq \ln(-z) - \ln(-\ln(-z)) + o(1)$. In particular here:

$$N_2(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_2^2}\left(\ln t + \frac{3}{2}\ln\ln t + \mathcal{O}(1)\right),$$

where the $\mathcal{O}$ hides constants depending in Δ_2. This concludes the proof of Lemma 5 as we just shown that if $\mathcal{E}(t)$ holds, at least one of the three following events hold:

- $N_2(t) \leq n(\Delta_2)$
- $\mu_2 - \hat{\mu}_2(t) \leq -\varepsilon\Delta_2$
- $N_2(t) \leq \frac{2\sigma^2}{(1-2\varepsilon)^2\Delta_2^2}\left(\ln t + \frac{3}{2}\ln\ln t + \mathcal{O}(1)\right)$.

□

C Generalization of the information maximization approximation

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

C.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 analytic 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., writing as

$$f(x|\theta) = A(x)\exp\big(T(x)\theta - F(\theta)\big), \tag{54}$$

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

$$\mathrm{KL}(\theta,\theta') = F(\theta') - F(\theta) - F'(\theta)(\theta'-\theta), \tag{55}$$

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

For an unknown prior $\pi(\theta)$, after $(x_1, \dots, x_n)$ reward realizations, the associated posterior distribution on θ, denoted p, reads

$$p(\theta|x_1,..,x_n) = \frac{1}{C}\pi(\theta)\exp\left(\theta\sum_{k=1}^{n} T(x_k) - nF(\theta)\right), \tag{56}$$

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, denoted $\hat{\mu}_l$, which verifies

$$\sum_{i=1}^{n} T(x_i) = nF'(\hat{\mu}_l) - \frac{\pi'(\hat{\mu}_l)}{\pi(\hat{\mu}_l)}. \tag{57}$$

Replacing the sum in Eq. (56) leads to

$$\begin{aligned} p(\theta|x_1,..,x_n) &= \frac{1}{C}\pi(\theta)\exp\left(\theta nF'(\hat{\mu}_l) - \theta\frac{\pi'(\hat{\mu}_l)}{\pi(\hat{\mu}_l)} - nF(\theta)\right) \\ &= \frac{e^{n\hat{\mu}_l F'(\hat{\mu}_l) - nF(\hat{\mu}_l)}}{C}\pi(\theta)e^{-\theta\frac{\pi'(\hat{\mu}_l)}{\pi(\hat{\mu}_l)}}e^{-n\mathrm{KL}(\hat{\mu}_l,\theta)} \\ &= \frac{1}{C_2}\pi(\theta)e^{-\theta\frac{\pi'(\hat{\mu}_l)}{\pi(\hat{\mu}_l)}}e^{-n\mathrm{KL}(\hat{\mu}_l,\theta)}. \end{aligned} \tag{58}$$

where C_2 also acts as a normalization constant of Eq. (58). For $n \gg 1$, the distribution concentrates in the vicinity of $\hat{\mu}_l$ from which we will derive the asymptotic scaling of C_2. We then integrate Eq. (58) after change of variable $\theta(u) = \hat{\mu}_l + \frac{u}{\sqrt{n}}$:

$$\begin{aligned} 1 = \int_\Theta p(\theta|x_1,..,x_n)\mathrm{d}\theta = &\int_{-(\theta_b-\hat{\mu}_l)\sqrt{n}}^{(\theta_b-\hat{\mu}_l)\sqrt{n}} \frac{1}{C_2\sqrt{n}}\pi(\hat{\mu}_l + \frac{u}{\sqrt{n}})e^{-(\hat{\mu}_l+\frac{u}{\sqrt{n}})\frac{\pi'(\hat{\mu}_l+\frac{u}{\sqrt{n}})}{\pi(\hat{\mu}_l+\frac{u}{\sqrt{n}})}}e^{-n\mathrm{KL}(\hat{\mu}_l,\hat{\mu}_l+\frac{u}{\sqrt{n}})}\mathrm{d}u \\ &+ \int_{\mu_{\inf}}^{-\theta_b} p(\theta|x_1,..,x_n)\mathrm{d}\theta + \int_{\theta_b}^{\mu_{\sup}} p(\theta|x_1,..,x_n)\mathrm{d}\theta \end{aligned} \tag{59}$$

Taking $(\theta_b - \hat{\mu}_l) \sim n^{-b}$ with $b < 1/2$, we get rid of the tail components in the asymptotic limit. Secondly, by noticing that $\frac{F''(\hat{\mu}_l)}{2} = \lim_{\theta\to\hat{\mu}_l} K(\hat{\mu}_l,\theta)/|\theta - \hat{\mu}_l|^2$ from Eq. (55), we make an expansion at the lower order of the Kullback-Leibler divergence which gives

$$1 = \lim_{\theta\to\hat{\mu}_l}\int_{-(\theta_b-\hat{\mu}_l)\sqrt{n}}^{(\theta_b-\hat{\mu}_l)\sqrt{n}} \frac{1}{C_2\sqrt{n}}\pi(\hat{\mu}_l)e^{-\hat{\mu}_l\frac{\pi'(\hat{\mu}_l)}{\pi(\hat{\mu}_l)}}e^{-\frac{F''(\hat{\mu}_l)u^2}{2}\hat{\mu}_l + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)}\mathrm{d}u. \tag{60}$$

Then, we obtain

$$C_2 \sim \frac{\sqrt{2\pi}}{\sqrt{nF''(\hat{\mu}_l)}}\pi(\hat{\mu}_l)e^{-\hat{\mu}_l\frac{\pi'(\hat{\mu}_l)}{\pi(\hat{\mu}_l)}}. \tag{61}$$

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

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

$$p_i(\theta) \underset{N_i\to\infty}{\sim} \sqrt{\frac{1}{2\pi\bar{\sigma}_i^2}}H(\theta,\hat{\mu}_l)e^{-N_i\mathrm{KL}(\hat{\mu}_l,\theta)}, \tag{62}$$

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

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

C.2 The partitioning approximation

Since all the steps leading to the partitioning approximations are independent of the reward distribution type. We retain the identical expressions of $\tilde{S}_{\text{tail}}$ and $\tilde{S}_{\text{body}}$ given in Appendix A.1.

C.3 Asymptotic of the intersection point

By use of Eq. (62) the equation verified by the intersection point $\bar{\mu}_{\text{eq}}$ now asymptotically reads

$$\frac{e^{-N_{M_t}\text{KL}(\hat{\mu}_{M_t},\bar{\mu}_{\text{eq}})}}{\sqrt{2\pi\bar{\sigma}^2_{M_t}}}\int_{\mu_{\text{inf}}}^{\bar{\mu}_{\text{eq}}}\frac{e^{-N_{m_t}\text{KL}(\hat{\mu}_{m_t},\theta')}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}}\text{d}\theta' = \frac{e^{-N_{m_t}\text{KL}(\hat{\mu}_{m_t},\bar{\mu}_{\text{eq}})}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}}\int_{\mu_{\text{inf}}}^{\bar{\mu}_{\text{eq}}}\frac{e^{-N_{M_t}\text{KL}(\hat{\mu}_{M_t},\theta')}}{\sqrt{2\pi\bar{\sigma}^2_{M_t}}}\text{d}\theta'. \tag{63}$$

Taking the logarithm of Eq. (63) leads to

$$N_{m_t}\text{KL}(\hat{\mu}_{m_t},\bar{\mu}_{\text{eq}}) - N_{M_t}\text{KL}(\hat{\mu}_{M_t},\bar{\mu}_{\text{eq}}) + \frac{1}{2}\ln\frac{\bar{\sigma}^2_{m_t}}{\bar{\sigma}^2_{M_t}} + \ln\frac{\int_{\mu_{\text{inf}}}^{\bar{\mu}_{\text{eq}}}\sqrt{\bar{\sigma}^2_{M_t}}e^{-N_{m_t}\text{KL}(\hat{\mu}_{m_t},\theta')}\text{d}\theta'}{\int_{\mu_{\text{inf}}}^{\bar{\mu}_{\text{eq}}}\sqrt{\bar{\sigma}^2_{m_t}}e^{-N_{M_t}\text{KL}(\hat{\mu}_{M_t},\theta')}\text{d}\theta'} = 0. \tag{64}$$

with identical arguments to the ones exposed in Appendix A.2, we approximate $\bar{\mu}_{\text{eq}}$ by neglecting the last term. Furthermore, in the considered asymptotic scaling $N_{M_t} \gg N_{m_t}$, $\bar{\mu}_{\text{eq}}$ will be in the vicinity of $\hat{\mu}_{M_t}$ where a the Gaussian expansion of the Kullback-Leibler divergence is relevant (see Eq. (59)). Thus, we approximate $\text{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\text{eq}})$ by $\text{KL}(\hat{\mu}_{m_t},\hat{\mu}_{M_t})$ and expand $\text{KL}(\hat{\mu}_{M_t},\bar{\mu}_{\text{eq}})$ to lowest order in $\tilde{\mu}_{\text{eq}}$ (with $\bar{\sigma}_i^2 \sim F''(\hat{\mu}_l)^{-1}N_i^{-1}$), leading to:

$$\tilde{\mu}_{\text{eq}} = \hat{\mu}_{M_t} + \sqrt{2\bar{\sigma}^2_{M_t}\left[N_{m_t}\text{KL}(\hat{\mu}_{m_t},\hat{\mu}_{M_t}) + \frac{1}{2}\ln\frac{\bar{\sigma}^2_{m_t}}{\bar{\sigma}^2_{M_t}}\right]}. \tag{65}$$

C.4 Generalisation of the main mode's contribution

We start by reminding the body component expression

$$\tilde{S}_{\text{body}} = -\int_{\Theta} p_{M_t}(\theta)C_{m_t}(\theta)\ln p_{M_t}(\theta)\text{d}\theta. \tag{66}$$

Without any additional information on KL expression, Eq. (66) cannot be computed in a closed form. Then, we will rely on the asymptotic scaling $N_{M_t} \gg N_{m_t} \gg 1$ to provide a tractable expression. First, we neglect $C_{m_t}(\theta)$ variations in Eq. (66) integral by evaluating it at $\tilde{\mu}_{\text{eq}}$. Then, by noticing that the resulting integral is the entropy of the better empirical arm mean, we approximate it by the leading order proportional to $\ln(2\pi\bar{\sigma}^2_{M_t})/2$:

$$\tilde{S}_{\text{body}} \approx \frac{1}{2}\ln(2\pi\bar{\sigma}_i^2)\left[1 - \int_{\bar{\mu}_{\text{eq}}}^{\mu_{\text{sup}}}\frac{e^{-N_{m_t}\text{KL}(\hat{\mu}_{m_t},\theta')}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}}\text{d}\theta'\right]. \tag{67}$$

We finally consider the last integral in Eq. (67). By noticing that it is concentrated in the vicinity of $\tilde{\mu}_{\text{eq}}$ for $N_{m_t} \gg 1$, we do a Taylor expansion of $\text{KL}(\hat{\mu}_{m_t},\theta')$ at $\tilde{\mu}_{\text{eq}}$ to obtain

$$\begin{aligned}\int_{\bar{\mu}_{\text{eq}}}^{\mu_{\text{sup}}}\frac{e^{-N_{m_t}\text{KL}(\hat{\mu}_{m_t},\theta')}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}}\text{d}\theta &\approx \frac{e^{-N_{m_t}\text{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\text{eq}})}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}}\int_{\bar{\mu}_{\text{eq}}}^{\mu_{\text{sup}}} e^{-N_{m_t}(\theta'-\tilde{\mu}_{\text{eq}})\partial_2\text{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\text{eq}})}\text{d}\theta' \\ &\approx \frac{e^{-N_{m_t}\text{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\text{eq}})}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}N_{m_t}\partial_2\text{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\text{eq}})}.\end{aligned} \tag{68}$$

Injecting Eq. (68) expression in Eq. (67) leads to the expected body component

$$\tilde{S}_{\rm b} = \frac{1}{2}\ln(2\pi\bar{\sigma}_i^2)\left[1 - \frac{e^{-N_{m_t}\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}N_{m_t}\partial_2\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})}\right]. \tag{69}$$

C.5 Generalised expression for the entropy tail

We start by reminding the tail component expression

$$\tilde{S}_{\rm tail} = -\int_{\tilde{\mu}_{\rm eq}}^{\mu_{\rm sup}} p_{m_t}(\theta)\ln p_{m_t}(\theta)\mathrm{d}\theta. \tag{70}$$

As for Eq. (68), we make a Taylor expansion of $\mathrm{KL}(\hat{\mu}_{m_t},\theta')$ at $\tilde{\mu}_{\rm eq}$ in the exponential term to obtain

$$\tilde{S}_{\rm t} = -\ln p_{m_t}(\tilde{\mu}_{\rm eq})\frac{e^{-N_{m_t}\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})}}{\sqrt{2\pi\bar{\sigma}^2_{m_t}}N_{m_t}\partial_2\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})}. \tag{71}$$

Keeping the leading order of $-\ln p_{m_t}(\tilde{\mu}_{\rm eq}) \sim N_{m_t}\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})$ leads to the expected tail expression used in the main text.

C.6 Generalised form of the entropy approximation

To summarize, by combining Eqs. (69) and (71) we obtain an asymptotic expression from exponential family bandits with uniform prior:

$$\tilde{S}_{\max} = \frac{1}{2}\ln(2\pi\bar{\sigma}_i^2)\left[1 - \frac{e^{-N_{m_t}\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})}}{N_{m_t}\partial_2\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})\sqrt{2\pi\bar{\sigma}^2_{m_t}}N_{m_t}}\right] + \frac{\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})e^{-N_{m_t}\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})}}{\partial_2\mathrm{KL}(\hat{\mu}_{m_t},\tilde{\mu}_{\rm eq})\sqrt{2\pi\bar{\sigma}^2_{m_t}}}. \tag{72}$$

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

Note, all these steps can be adapted to non-uniform priors (in particular by multiplying the tail by the prior effects evaluated in $\tilde{\mu}_{\rm eq}$). Finally, let us draw that our approximation scheme holds for any posterior distributions verifying Eq. (62), a property we believe to be shared for more general reward distributions.

C.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. (72) 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 can't be rendered analytical, 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 be also considered.

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

C.8 Extension of information maximization for more than two arms

Here, we briefly justify the entropy approximations for more than two arms leading to the body/tail components given in the main text.

Most of the approximations are similar to the one derived in Appendix A.1 discussion. We continue to denote the better empirical arm as M_t and look for a valid approximations in the regime $N_{M_t} \gg N_i \gg 1, \forall i \neq M_t$. We start by rewriting the exact entropy expression isolating M_t:

$$\begin{aligned} S_{\max} = &- \int_\Theta p_{M_t}(\theta) \left[\prod_{j\neq M_t} C_j(\theta)\right] \ln\left(p_{M_t}(\theta) \prod_{j\neq M_t} C_j(\theta) + \sum_{i\neq M_t} C_{M_t}(\theta) p_i(\theta) \prod_{j\neq i, M_t} C_j(\theta)\right) \mathrm{d}\theta \\ &- \sum_{i\neq M_t} \int_\Theta p_i(\theta) C_{M_t}(\theta) \left[\prod_{j\neq i, M_t} C_j(\theta)\right] \ln\left(p_{M_t}(\theta) \prod_{j\neq M_t} C_j(\theta) + \sum_{i\neq M_t} C_{M_t}(\theta) p_i(\theta) \prod_{j\neq i, M_t} C_j(\theta)\right) \mathrm{d}\theta. \end{aligned} \tag{73}$$

We define $\tilde{\mu}_{\mathrm{eq,i}}$, which approximates the value where arm i has the same probability of being the maximum as the best empirical arm. Next, we consider the first term for $\theta < \min(\{\tilde{\mu}_{\mathrm{eq,i}}, i \neq M_t\})$, we assume to neglect all the inner terms inside the logarithm which is then dominated by $p_{\max}$. Next, by noticing that $C_i(\theta) \approx 1$ in the vicinity of $\hat{\mu}_{M_t}$, we make a first order expansion of the product along all the worse empirical arms. Finally, extending the upper bound of the integral leads to the body expression given in the main text:

$$\tilde{S}_{\mathrm{body}} = - \int_\Theta \big(1 - \sum_{i\neq M_t}^{K} [1 - C_i(\theta)]\big) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta. \tag{74}$$

Then, we consider the additional terms (each denoted as i) in Eq. (73). First, due to $C_{M_t}(\theta)$, each term of the sum is negligible for $\theta < \tilde{\mu}_{\mathrm{eq,i}}$. Then, we approximate all the cumulatives by one since we are considering the tail contribution. Finally, to get a simplified expression for the increment, we assume to neglect all the posterior distributions except for $p_i(\theta)$ inside the logarithmic of the i-th term. One could legitimately remark that some of these posterior distributions ($j \neq i, M_t$) are not necessarily negligible compared to $p_i(\theta)$ at a given θ. However, this cross information between current suboptimal arms is not valuable regarding the decision procedure (which largely resumes as balancing between exploiting the best empirical solution compared to exploring worse empirical arms), while unnecessarily complicating the increment evaluation.

Taken altogether, we obtain the full expression of the entropy

$$\tilde{S}_{\max} = - \int_\Theta \big(1 - \sum_{i\neq M_t}^{K} [1 - C_i(\theta)]\big) p_{M_t}(\theta) \ln p_{M_t}(\theta) \mathrm{d}\theta - \sum_{i\neq M_t}^{K} \int_{\tilde{\mu}_{\mathrm{eq,i}}}^{\mu_{\sup}} p_i(\theta) \ln p_i(\theta) d\theta. \tag{75}$$

D 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.

D.1 Numerical settings:

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

For the two-arm cases in Figs. 2 and 3 the arm means are chosen from a uniform grid in $[0, 1]$ using a Sobol sequence (we have avoided $\{0, 1\}$ values but it has no impact on the obtained results). Regret is averaged over

more than 10^5 events and observed during 10^6 steps to attest the logarithmic scaling. 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 Thompson sampling to allow a direct comparison.

For the fifty-armed case in Fig. 4, the arm means are drawn on a uniform prior and regret is averaged over 4.10^4 events and observed during 5.10^4 steps.

Finally for close arm means in Figs. 5 and 6, 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. Regret is averaged over 10^5 events and observed during 10^6 steps.

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.

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

D.2 AIM implementation details

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

D.2.1 Information maximization approximation for Bernoulli rewards

We denote $\bar{\mu}_i$ as the posterior mean with

$$\mathbb{E}\left[X_{\mathcal{B}(r_i+1,N_i-r_i+1)}\right] = \frac{r_i+1}{N_i+2} = \bar{\mu}_i, \tag{76}$$

where r_i is the cumulative reward at time t , N_i the number of draws and $X_{\mathcal{B}(a,b)}$ is variable following a Beta distribution of parameters (a, b). Then, we also denote $\bar{N}_i$ as verifying

$$\begin{aligned}\mathrm{Var}\left[X_{\mathcal{B}(r_i+1,N_i-r_i+1)}\right] &= \frac{r_i+1}{N_i+2}\frac{N_i-r_i+1}{N_i+2}\frac{1}{N_i+3} \\ &= \frac{\bar{\mu}_i(1-\bar{\mu}_i)}{\bar{N}_i},\end{aligned} \tag{77}$$

i.e., $\bar{N}_i = N_i + 3$.

For Bernoulli reward, we asses to approximate the gradient as follow:

$$\begin{aligned}\Delta_i|\tilde{S}_{\max}| = \Bigg|&\frac{\bar{\mu}_i(\bar{N}_i-1)-1}{\bar{N}_i-3}\tilde{S}_{\max}(\frac{\bar{\mu}_i\bar{N}_i+1-\bar{\mu}_i}{\bar{N}_i},\bar{N}_i+1,\bar{\mu}_j,\bar{N}_j) \\ &+\frac{\bar{N}_i-2-\bar{\mu}_i(\bar{N}_i-1)}{\bar{N}_i-3}\tilde{S}_{\max}(\frac{\bar{\mu}_i(\bar{N}_i-1)}{\bar{N}_i},\bar{N}_i+1,\bar{\mu}_j,\bar{N}_j)-\tilde{S}_{\max}(\bar{\mu}_i,\bar{N}_i,\bar{\mu}_j,\bar{N}_j)\Bigg|,\end{aligned} \tag{78}$$

with $\tilde{S}_{\max}$ given by Eq. (72) reading:

$$\begin{aligned}\tilde{S}_{\max}(\bar{\mu}_{M_t},\bar{N}_{M_t},\bar{\mu}_{m_t},\bar{N}_{m_t}) = &\left(1-\frac{e^{-\bar{N}_{m_t}\mathrm{KL}(\bar{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})}}{\sqrt{N_{m_t}}\partial_2\mathrm{KL}(\bar{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})\sqrt{2\pi\bar{\mu}_{m_t}(1-\bar{\mu}_{m_t})}}\right)\frac{1}{2}\ln\left(\frac{2\pi\bar{\mu}_{M_t}(1-\bar{\mu}_{M_t})}{\bar{N}_{M_t}}\right) \\ &+\frac{\sqrt{\bar{N}_{m_t}}\mathrm{KL}(\bar{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})e^{-\bar{N}_{m_t}\mathrm{KL}(\bar{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})}}{\partial_2\mathrm{KL}(\bar{\mu}_{m_t},\tilde{\mu}_{\mathrm{eq}})\sqrt{2\pi\bar{\mu}_{m_t}(1-\bar{\mu}_{m_t})}},\end{aligned} \tag{79}$$

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 $1-\frac{\bar{\mu}_i(\bar{N}_i-1)-1}{\bar{N}_i-3}$.

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 C.7 for further discussion).

Algorithm 2: 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 */`
  $M_t \leftarrow \text{argmax}_{k=1,2}\, \bar{\mu}_k$, $m_t \leftarrow \text{argmin}_{k=1,2}\, \bar{\mu}_k$;
  if $N_{M_t} \leq N_{m_t}$ **then**
    $a_t \leftarrow M_t$
  else
    Evaluate $\Delta = \Delta_{M_t}|\tilde{S}_{\max}| - \Delta_{m_t}|\tilde{S}_{\max}|$ following Eq. (78) ;
    if $\Delta > 0$ **then**
      $a_t \leftarrow M_t$
    else
      $a_t \leftarrow m_t$
  Pull a_t and observe $X_t(a_t)$
  `/* Update statistics */`
  $\bar{\mu}_{a_t} \leftarrow \frac{\bar{\mu}_{a_t}(\bar{N}_{a_t}-1)+X_t}{\bar{N}_{a_t}}$, $\bar{N}_{a_t} \leftarrow \bar{N}_{a_t}+1$

Let us draw some additional observation on the practical implementation of the code. First in then gradient evaluation of Δ_i following Eq. (78). If ones finds a $\tilde{\mu}_{\text{eq}}$ value to be undefined (because $N_{m_t} > N_{M_t}$ 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 $S_{\max} = \frac{1}{2}\ln\left(\frac{2\pi\bar{\mu}_{M_t}(1-\bar{\mu}_{M_t})}{\bar{N}_{M_t}}\right)$.

Secondly, at large times, noticing the better empirical arm is drawn extensively, one can build on entropy increment structure to speed up AIM performances. 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. 2 still returns M_t, then it ensures that all increment evaluations between $[t, t+T]$ of Alg. 2 will always return M_t independently of its returned rewards. Then, by use of a dichotomy search on the variable T, ones can diminish the number of increment evaluation of AIM at large times, thus improving AIM time performances.

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

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

$$\tilde{S}_{\max} = -\int_\Theta \big(1-\sum_{i\neq M_t}^{K}[1-C_i(\theta)]\big)p_{M_t}(\theta)\ln p_{M_t}(\theta)\mathrm{d}\theta - \sum_{i\neq M_t}^{K}\int_{\tilde{\mu}_{\text{eq,i}}}^{\mu_{\sup}} p_i(\theta)\ln p_i(\theta)\mathrm{d}\theta. \tag{80}$$

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

$$\Delta_i|\tilde{S}_{\max}| = \Delta_i\left|-\int_\Theta C_i(\theta)p_{M_t}(\theta)\ln p_{M_t}(\theta)\mathrm{d}\theta - \int_{\tilde{\mu}_{\text{eq,i}}}^{\mu_{\sup}} p_i(\theta)\ln p_i(\theta)\mathrm{d}\theta\right|, \tag{81}$$

which is exactly the increment evaluated in the two-armed case given in Eq. (79).

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. (79) we obtain

$$\Delta_{M_t} S_{\max} = \left| 1 - \sum_{i \neq M_t}^{K} \frac{e^{-\bar{N}_{m_t} \mathrm{KL}(\bar{\mu}_{m_t}, \tilde{\mu}_{\mathrm{eq}})}}{\sqrt{\bar{N}_{m_t}} \partial_2 \mathrm{KL}(\bar{\mu}_{m_t}, \tilde{\mu}_{\mathrm{eq}}) \sqrt{2\pi \bar{\mu}_{m_t}(1-\bar{\mu}_{m_t})}} \right| \Delta_{M_t} \left| H(\bar{\mu}_{M_t}, \bar{N}_{M_t}) \right|, \tag{82}$$

where

$$\begin{aligned} \Delta_{M_t} \left| H(\bar{\mu}_i, \bar{N}_i) \right| = \Bigg| & \frac{\bar{\mu}_i(\bar{N}_i - 1) - 1}{\bar{N}_i - 3} H\left(\frac{\bar{\mu}_i \bar{N}_i + 1 - \bar{\mu}_i}{\bar{N}_i}, \bar{N}_i + 1, \bar{\mu}_j, \bar{N}_j\right) \\ & + \frac{\bar{N}_i - 2 - \bar{\mu}_i(\bar{N}_i - 1)}{\bar{N}_i - 3} H\left(\frac{\bar{\mu}_i(\bar{N}_i - 1)}{\bar{N}_i}, \bar{N}_i + 1, \bar{\mu}_j, \bar{N}_j\right) - H(\bar{\mu}_i, \bar{N}_i, \bar{\mu}_j, \bar{N}_j) \Bigg|, \end{aligned} \tag{83}$$

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 3: 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 */`
  $M_t \leftarrow \mathrm{argmax}_{k=\{1,..,K\}} \bar{\mu}_k$; Evaluate $\Delta_{M_t} \tilde{S}_{\max}$ following Eq. (82) ;
  Evaluate $m_t = argmax(\Delta_i |\tilde{S}_{\max}|, i \neq M_t)$ with $\Delta_i |\tilde{S}_{\max}|$ following Eq. (78) ;
  if $\Delta_{M_t} \tilde{S}_{\max} > \Delta_{m_t} |\tilde{S}_{\max}|$ **then**
    $a_t \leftarrow M_t$
  else
    $a_t \leftarrow m_t$
  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. (78), if ones finds a $\tilde{\mu}_{\mathrm{eq,i}}$ value undefined (because $N_{m_t} > N_{M_t}$ or $\tilde{\mu}_{\mathrm{eq,i}} > 1$ which is unusable for Bernoulli reward), then, $\tilde{\mu}_{\mathrm{eq,i}}$ 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 $M_t \leftarrow \mathrm{argmax}_{k=\{1,..,K\}} \bar{\mu}_k$ has multiple solution, we suggest choosing the one displaying the lowest number of draws.

D.4 Overview of investigated classical bandit algorithms

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

D.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:

$$R_i = c(\mu_1, \mu_2) \sqrt{\frac{\ln(t)}{N_i(t)} \min\left(\frac{1}{4}, s_i(t)\right)}, \ \ s_i(t) = \hat{\sigma_i}^2 + \sqrt{\frac{2\ln(t)}{N_i(t)}}, \tag{84}$$

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. (84), we end up with $c = 2.1$ and $\hat{\sigma_i}^2 = \frac{\sigma^2}{N_i(t)}$.

D.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:

$$F_i = \max\left\{\theta \in \Theta : Ni(t)\mathrm{KL}\left(\frac{r_i(t)}{N_i(t)}, \theta\right) \leq \ln(t) + c(\mu_1, \mu_2)\ln(\ln(t))\right\}, \tag{85}$$

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

D.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 maximizing the expected reward. In practice, it draws K random values according to each arm mean's posterior distribution and selects the arm with the highest sampled value:

$$a_t = \underset{i=1..K}{\mathrm{argmax}}\Big(Z_i\left(\hat{\mu}_i(t), N_i(t)\right)\Big). \tag{86}$$

where $Z_i(t)$ is drawn according to the posterior distribution of the arm mean. Here we used an uniform prior on $[0, 1]$ for Bernoulli rewards and a uniform prior on $\mathbb{Z}$ for Gaussian rewards to provide a direct comparison with AIM.

D.5 Additional experiments

D.5.1 Information maximization approximation for Gaussian 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 Gaussian reward distributions. Then, AIM shows state-of-the-art performance comparable to Thompson sampling even when arms mean rewards are difficult to distinguish.

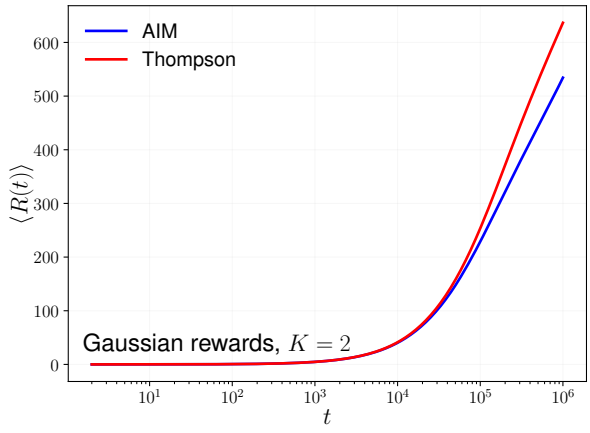

Figure 5: Temporal evolution of the regret for 2-armed bandit with Gaussian rewards ($\sigma = 1$) for close mean parameters. In blue AIM, in red Thompson sampling. Arm mean reward values are fixed with $\mu_1 = 0.8$ and $\mu_2 = 0.79$, the regret is obtained by averaging over 10^5 realizations.

D.5.2 Information maximization approximation for Bernoulli rewards and close arms

For completeness, we provide in Fig. 6 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.

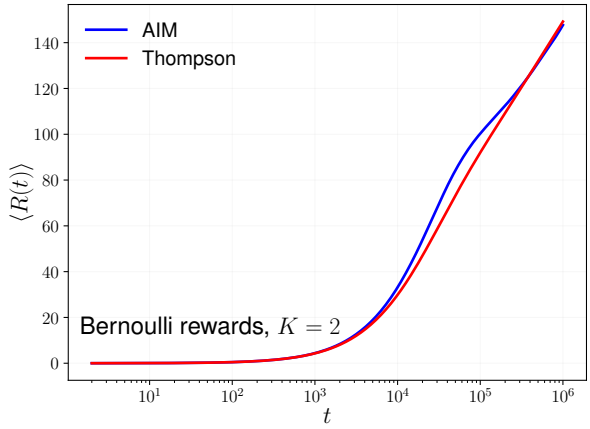

Figure 6: Temporal evolution of the regret for 2-armed bandit with Bernoulli rewards for close mean parameters. In blue AIM, in red Thompson sampling. Arm mean reward values are fixed with $\mu_1 = 0.8$ and $\mu_2 = 0.79$, the regret is obtained by averaging over 10^5 realizations.