Approximate information for efficient exploration-exploitation strategies

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

Approximate information for efficient exploration-exploitation strategies

Alex Barbier–Chebbah,* Christian L. Vestergaard, and Jean-Baptiste Masson
Institut Pasteur, Université Paris Cité, CNRS UMR 3571, Decision and Bayesian Computation, 75015 Paris, France.
(Dated: July 3, 2023)

This paper addresses the exploration-exploitation dilemma inherent in decision-making, focusing on multi-armed bandit problems. The problems involve an agent deciding whether to exploit current knowledge for immediate gains or explore new avenues for potential long-term rewards. We here introduce a novel algorithm, approximate information maximization (AIM), which employs an analytical approximation of the entropy gradient to choose which arm to pull at each point in time. AIM matches the performance of Infomax and Thompson sampling while also offering enhanced computational speed, determinism, and tractability. Empirical evaluation of AIM indicates its compliance with the Lai & Robbins asymptotic bound and demonstrates its robustness for a range of priors. Its expression is tunable, which allows for specific optimization in various settings.

Introduction. The exploration-exploitation dilemma is a fundamental challenge in decision-making. It arises when an agent must choose between exploiting its current knowledge to maximize immediate rewards or acquiring new information that may lead to greater long-term gains. This dilemma is ubiquitous in various fields, from anomaly detection [1] to the modelling of biological search strategies [2–4] and human decision-making [5–9].

The multi-armed bandit problem is a paradigmatic example of an explore-exploit problem and has been extensively studied and applied in a range of fields, including applied mathematics [10–16] to animal behavior [17], neuroscience [18–21], clinical trials [22–24], finance [25], epidemic control [26], and reinforcement-learning [27, 28], among others. In the multi-armed bandit problem, an agent is presented with a set of possible actions, or "arms", each associated with a probabilistic reward (akin to a multi-armed slot machines game). The agent must choose which arm to pull at each time step to maximize its cumulative reward over a fixed or infinite time horizon. Hence, at each time step, the agent can either play the arm with the better rewards to improve the knowledge on that arm or explore new arms to test if they would not lead to increased rewards.

In the following we begin with a brief introduction to the bandit problem, followed by a presentation of our novel approximate information procedure, completed with its corresponding analytical expression. We then provide empirical evidence of the procedure's efficacy before delving into a discussion of its various properties and implications.

We consider the classic multi-armed bandit setting [29]. At each point in time, t, an agent chooses an arm, A_t, between K different arms, $\mathbb{A} = \{1, 2, \dots, K\}$. The chosen arm, i_t returns a stochastic reward, X_t, drawn from a distribution whose mean, μ_{i_t}, is unknown to the agent [Fig. 1(a)]. The agent's goal is to maximize the cumulative reward (equivalently, minimize the cumulative regret) with no time horizon. Formally, we aim to minimize the expected regret [29], $\mathbb{E}[R(t)]$, with

$$R(t) = \mu^* t - \sum_{\tau=1}^{t} X_\tau. \tag{1}$$

The regret, $R(t)$, measures the cumulative difference between the rewards obtained by the algorithm and the expected reward that it would have obtained by choosing the best action. Optimal strategies, regardless of their details, are characterized by the following asymptotic bound (the Lai and Robbins bound) [30]:

$$\langle R(t) \rangle_{t\to\infty} \geq \beta \, \log(t), \tag{2}$$

where β is a constant factor that depends on the reward distributions.

Multiple strategies attain the Lai and Robbins bound [Eq. (2)]. Notably, the ϵ_n-greedy strategy [10], which plays the best current arm with probability $1 - \epsilon_n$ and randomly samples other arms with probability ϵ_n, with a time-varying ϵ_n; the Upper Confidence Bound-2 (UCB-2) algorithm [16], which relies on a tuned confidence index associated to each arm to decide which arm to play; Thompson sampling (proportional betting), which relies on sampling the action from the posterior distribution that it maximizes the expected reward. Importantly, methods such as the ϵ_n-greedy and UCB-based algorithms require parameter tuning to reach the Lai and Robbins bound, making them sensitive to uncertainties and variations of the prior information used for tuning.

Approximate information maximization for bandit problems. We aim here to develop a tractable, functional-based algorithm for the multi-armed bandit problem. Inspired by the Infomax principle [2, 31], we rely on the entropy as a functional to optimise to decide which arm to play. Contrary to classical bandit algorithms, the entropy encompasses the information carried by all arms in a single functional, thus characterising the global state of the game. More precisely, we aim to optimise S, the entropy of the posterior distribution of the value of the maximal reward, $p_{\max}$,

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

* alex.barbier-chebbah@pasteur.fr

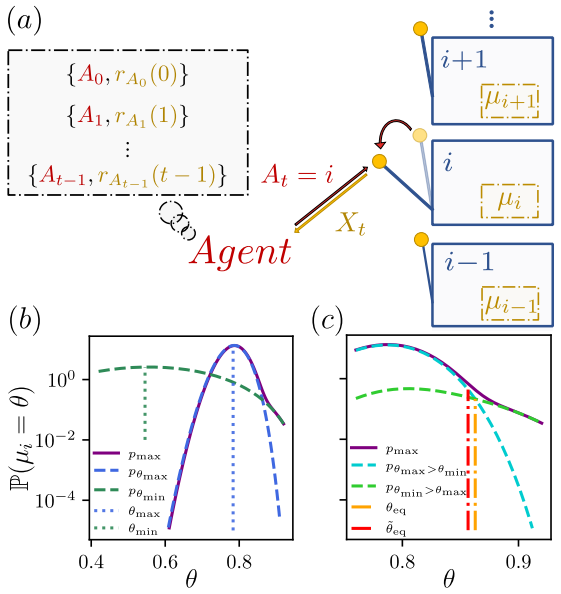


FIG. 1. **(a)** Illustration of the multi-armed bandit problem. At each time step t the agent chooses an action $i = A_t$ that returns a reward $r_i(t)$ drawn from a distribution of unknown mean μ_i. The agent's goal is to minimize the cumulative regret $R(t)$ [see Eq. (1)]. **(b)** Posterior distributions of bandit values after playing the 2-armed Bernoulli game with $r_1(t) = 5$, $n_1(t) = 9$, $r_2(t) = 41$, $n_2(t) = 192$, where $r_i(t)$, $n_i(t)$ are respectively the cumulative reward and number of draws of arm i. In blue, the posterior distribution, $p_{\theta_{\max}}$, of the reward of the current best arm. Vertical green and blue lines are the current average rewards of the suboptimal (denoted $\theta_{\min}$) and optimal arm ($\theta_{\max}$). In green, the posterior distribution, $p_{\theta_{\min}}$, of the current sub-optimal arm. In purple, the posterior distribution, $p_{\max}$, of the maximum reward of all arms. **(c)** Zoomed plot of **(b)** in the region where the posterior distribution of the maximal reward value transition from being dominated by $p_{\theta_{\max}>\theta_{\min}}$ to being dominated by $p_{\theta_{\min}>\theta_{\max}}$. In purple $p_{\max}$. In light blue, the probability, $p_{\theta_{\max}>\theta_{\min}}$, is that the optimal arm's gain is superior to the suboptimal arm. In light green, the probability, $p_{\theta_{\min}>\theta_{\max}}$, is that the gain of the suboptimal arm is superior to that of the optimal arm. The orange vertical line is the transition value θ_{eq} and the red vertical line its approximation $\tilde{\theta}_{\text{eq}}$ (see Supplemental Material S1 B & S2 B for its derivation).

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

$$p_{\max}(\theta) = \sum_{i=0}^{K} \mathbb{P}(\mu_i = \theta) \prod_{j \neq i} \mathbb{P}(\mu_j \leq \theta). \tag{4}$$

The entropy S summarizes the information about the state of the game and we require our algorithm to greedily optimise its gradient, i.e., to select the next arm according to:

$$\underset{i=1..K}{\text{argmin}} \langle S(t+1) - S(t) | A_{t+1} = i \rangle. \tag{5}$$

By doing so, the algorithm seeks to maximize the expected decrease in entropy, conditioned on the current knowledge of the game. This strategy has shown to be competitive with state-of-the-art algorithms and attain the Lai and Robbins bound [31].

However, while Eq. (5) can be numerically evaluated, it cannot be computed in closed form for most bandit problems. To obtain an algorithm that is both tractable and computationally efficient, a second functional approximating the entropy has to be derived.

Hence, we devise a set of approximations of both $p_{\max}$ and S to get a tractable algorithm. We develop our approach on the 2-armed bandit. We denote the arms according to their current mean rewards, respectively the maximum one by $i_{\max}$ (with expected reward $\theta_{\max}$) and the minimum by $i_{\min}$ (with $\theta_{\min}$). Note that the true expected reward of $i_{\max}$ may be smaller than that of $i_{\min}$ due to the stochasticity of the game.

Our approximate form of the entropy reads:

$$\tilde{S} = (1 - c_{\text{tail}})\tilde{S}_{\text{body}} + \tilde{S}_{\text{tail}} - (1 - c_{\text{tail}}) \ln(1 - c_{\text{tail}}). \tag{6}$$

It decomposes the entropy into three tractable terms corresponding to approximations made on $p_{\max}$. The first term, $\tilde{S}_{\text{body}}$, approximates the entropy of the mode of $p_{\max}$. The second, $\tilde{S}_{\text{tail}}$, captures the entropy of the tail (on the high reward side, see Fig. 1(b,c)) of $p_{\max}$. These approximate entropies are weighted by factors depending on c_{tail}, a corrective term that compensates for an extension of the integral boundaries in order to make the entropy evaluations analytically tractable (see Supplemental Material S1 B for details).

More precisely, the tail term reads:

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

where $\tilde{\theta}_{\text{eq}}$ is the approximation of θ_{eq}, the value of θ where the probability of being the maximum is identical for both arms (see red and orange curves on Fig. 1(c)), and $p_{\theta_{\min}}(\theta) = \mathbb{P}(\mu_{i_{\min}} = \theta)$ is the posterior probability of the current suboptimal arm having expected reward θ.

The approximate entropy of the main mode is split into two terms:

$$\begin{aligned} \tilde{S}_{\text{body}} = &-\int_{\Theta} p_{\theta_{\max}>\theta_{\min}}(\theta) \ln p_{\theta_{\max}}(\theta) d\theta \\ &- A_{\text{c}} \int_{\Theta} p_{\theta_{\min}>\theta_{\max}}(\theta) d\theta, \end{aligned} \tag{8}$$

where $p_{\theta_{\max}}(\theta)$ is the posterior probability at θ of the current optimal arm, $p_{\theta_i>\theta_j}(\theta) = \mathbb{P}(\mu_i = \theta, \mu_i \geq \mu_j)$ is the posterior probability for the expected reward θ of arm i to be larger than θ_j, and $A_{\text{c}} = 1.25889$ is a predetermined constant [see Eq. (S5) in Supplemental Material S1 A]. The first term in Eq. (8) is the leading-order term of the mode of $p_{\max}$, dominated by the current optimal arm, whereas the second term handles the corrections induced by the suboptimal arm in the vicinity of $\theta_{\max}$ (see Supplemental Material S1 A for details).

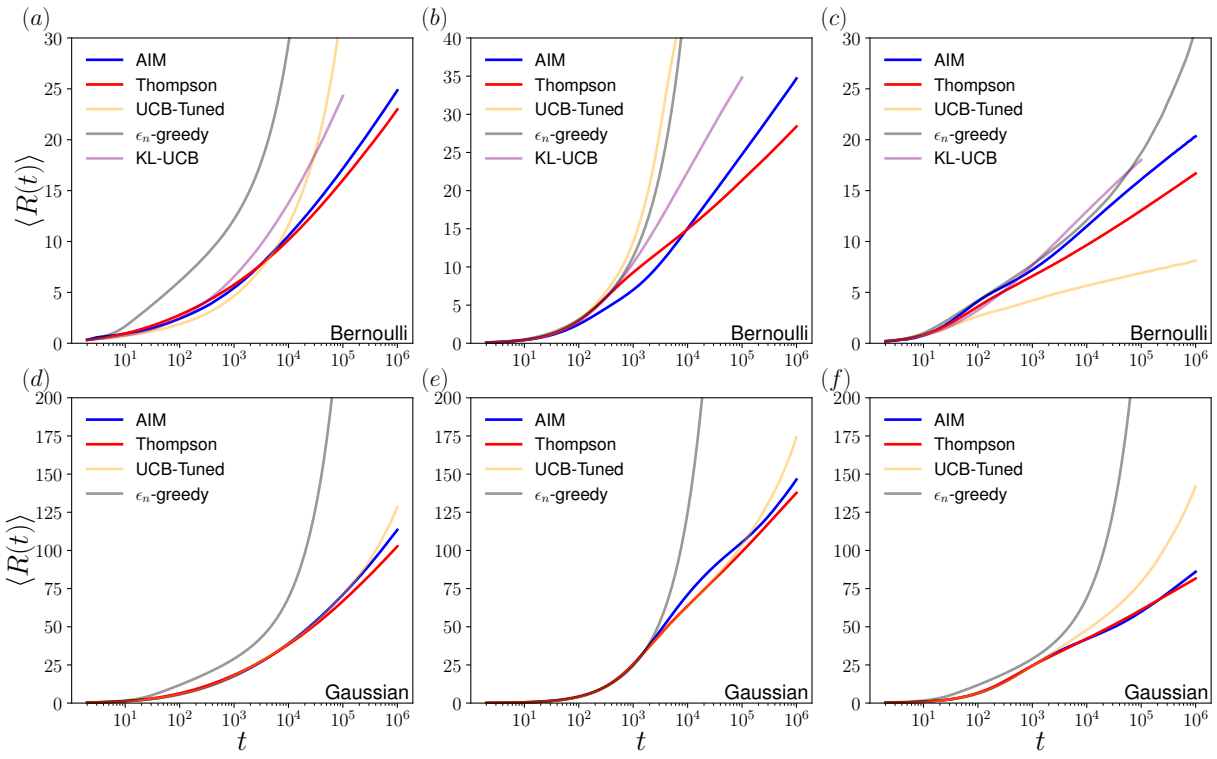


FIG. 2. Temporal evolution of the regret for Bernoulli **(a-c)** and Gaussian **(d-f)** 2-armed bandits. In blue AIM, in red Thompson sampling, in yellow UCB-tuned, in grey ϵ_n-greedy, and in purple KL-UCB. Details of simulations and the tuning required for some algorithms are provided in Supplemental Material S3 and S4. True parameters were drawn uniformly in $]0, 1[$ for both bandits in **(a,d)**, parameters were set to $\mu_1 = 0.7$, $\mu_1 = 0.8$ for **(b,e)** and to $\mu_1 = 0.1$, $\mu_1 = 0.3$ for **(c,f)**.

Finally, the third, corrective term in Eq. (7) is $c_{\text{tail}} = \int_{\tilde{\theta}_{\text{eq}}}^{\theta_{\text{sup}}} p_{\theta_{\min}}(\theta) d\theta$.

We propose approximate information maximization (AIM), an algorithm that consists in evaluating Eq. (6) for each arm in each time step t and choosing the one that minimizes the expected value of $\tilde{S}(t+1)$ according to Eq. (5). Depending on the reward distributions, and their associated Θ, the log dependencies inside $\tilde{S}_{\text{body}}$ and $\tilde{S}_{\text{tail}}$ can be integrated analytically or approximated by its long-time asymptote (see Supplemental Material S2 for a detailed deviation of all terms). Then, AIM provides a direct implementation following an analytically tractable expression.

Results. We demonstrate the performance of AIM on the paradigmatic Bernoulli bandits [10, 32, 33] and on Gaussian bandits [34] with unknown mean $\mu_i \in [0, 1]$ and unit variance. Supplementary Table S1 lists analytic expressions for the terms of $\tilde{S}$ [Eq. (6)] for each problem.

Figure 2 compares the performance of the AIM algorithm with other state-of-the-art algorithms on numerically generated data (see Supplemental Material S3 & S4 for implementation of AIM and other classic bandit strategies). For both Bernoulli and Gaussian bandits, AIM empirically follows the Lai & Robins bound, with a regret scaling as $\log(t)$. Its long time performance matches that of Thompson sampling while relying on a simple analytical formula. Additionally, AIM outperforms Thompson methods at intermediate times for challenging parameter configurations [Fig. 2(b)].

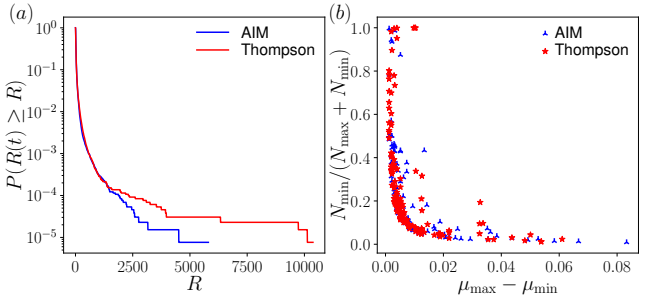


FIG. 3. Regret distribution and rare events. **(a)** The probability of obtaining a cumulative regret superior to R for both Thompson and AIM playing Bernoulli games with uniform priors and $N = 2^{17}$ realizations). AIM has an exponential decay similar to Thompson algorithm. **(b)** Fraction of the sub-dominant arm, drawn for high regret events (0.1%) and plotted along the mean difference $\mu_{\max} - \mu_{\min}$. In both **(a-b)** Thompson and AIM exhibit the same behaviour.

The following heuristic argument qualifies the optimal asymptotic scaling of AIM. Assuming $t \gg 1$ and $N_{\max} \gg N_{\min} \gg 1$, i.e., the best arm has been predominantly

pulled. Then, the variation along $N_{\min}$ and $N_{\max} = t - N_{\min}$ of the approximate entropy reads:

$$\frac{\partial \tilde{S}}{\partial N_{\min}} = (1 - c_{\text{tail}}) \frac{\partial \tilde{S}_{\text{body}}}{\partial N_{\min}} + \frac{\partial \tilde{S}_{\text{tail}}}{\partial N_{\min}} + .. \\ (-\tilde{S}_{\text{body}} + \ln(1 - c_{\text{tail}}) + 1) \frac{\partial c_{\text{tail}}}{\partial N_{\min}}, \tag{9}$$

To leading order, the minimum of Eq. (9) is found at $N_{\min} \sim \ln(t)/K_{\mathbb{B}}(\mu_{\min}, \mu_{\max})$. For Bernoulli bandits, where $K_{\mathbb{B}}(\mu_{\min}, \mu_{\max})$ is the Kullback-Leibler divergence between the reward distributions, thus recovering the Lai and Robbins bound (see derivation in Supplemental Material S5). Note that this derivation is not entirely rigorous as it assumes that, after a certain time, we can be sure that the optimal arm has been predominantly pulled. We checked this assumption by investigating the asymptotic behaviour of high cumulative regret events (Fig. 3), for which the sub-dominant arm has been drawn a non-negligible fraction of time. These events are exponentially rare and happen only for small $\mu_{\max} - \mu_{\min}$, which require exponentially long times to be distinguished (a behaviour that is shared by Thompson sampling).

Conclusion. In this study, we present a new approach, AIM, designed to effectively balance exploration and exploitation in multi-armed bandit problems. AIM employs an analytic approximation of the entropy gradient to select the optimal arm. This novel approach mirrors the performance of Infomax (see Supplemental Material S4 and Fig. S1), from which it is derived, while offering improved computational speed. It also parallels Thompson sampling in functionality, yet outperforms it in terms of being deterministic and more easily managed.

Empirical testing demonstrated that AIM complies with the Lai and Robbins bound and exhibits robustness to a broad spectrum of priors. Furthermore, since it relies on an analytic expression, AIM can easily be fine-tuned to optimise performance in various scenarios, while still satisfying the Lai and Robbins bounds. Specifically, tuned AIM is highly efficient for K-armed bandits with $K > 2$ (see Supplemental Material S6 and Fig. S2 for derivation and examples).

Due to its reliance on a single, analytically tractable functional expression, AIM proves adaptable for different bandit problems, particularly where other approaches may face efficiency constraints. Interesting future research directions include devising a rigorous proof of optimality, applying and optimising AIM to multi-armed problems with finite horizons, with insufficient time to sample all bandits, and its extension to Monte-Carlo path-planning schemes.

Acknowledgments. We thank Etienne Boursier for helpful discussions for optimality of AIM.

[1] K. Ding, J. Li, and H. Liu, in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2019), WSDM '19, pp. 357–365, ISBN 978-1-4503-5940-5.

[2] M. Vergassola, E. Villermaux, and B. I. Shraiman, Nature **445**, 406 (2007), number: 7126 Publisher: Nature Publishing Group.

[3] D. Martinez, L. Arhidi, E. Demondion, J.-B. Masson, and P. Lucas, Journal of Visualized Experiments : JoVE p. 51704 (2014).

[4] R. T. Cardé, Annual Review of Entomology **66**, 317 (2021).

[5] J. D. Cohen, S. M. McClure, and A. J. Yu, Philosophical Transactions of the Royal Society B: Biological Sciences **362**, 933 (2007), publisher: Royal Society.

[6] T. T. Hills, P. M. Todd, D. Lazer, A. D. Redish, and I. D. Couzin, Trends in Cognitive Sciences **19**, 46 (2015).

[7] K. Mehlhorn, B. R. Newell, P. M. Todd, M. D. Lee, K. Morgan, V. A. Braithwaite, D. Hausmann, K. Fiedler, and C. Gonzalez, Decision **2**, 191 (2015), place: US Publisher: Educational Publishing Foundation.

[8] K. Doya, *Bayesian Brain: Probabilistic Approaches to Neural Coding* (MIT Press, 2007), ISBN 978-0-262-04238-3, google-Books-ID: bsQMWXXHzrYC.

[9] M. Jepma and S. Nieuwenhuis, Journal of Cognitive Neuroscience **23**, 1587 (2011).

[10] A. Slivkins, Foundations and Trends® in Machine Learning **12**, 1 (2019), publisher: Now Publishers, Inc.

[11] J. C. Gittins, Journal of the Royal Statistical Society. Series B (Methodological) **41**, 148 (1979), publisher: [Royal Statistical Society, Wiley].

[12] L. Zhou, *A Survey on Contextual Multi-armed Bandits* (2016), arXiv:1508.03326 [cs].

[13] S. Bubeck, R. Munos, and G. Stoltz, Theoretical Computer Science **412**, 1832 (2011).

[14] M. Bayati, N. Hamidi, R. Johari, and K. Khosravi, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., 2020), vol. 33, pp. 1713–1723.

[15] D. Bouneffouf, I. Rish, and C. Aggarwal, in *2020 IEEE Congress on Evolutionary Computation (CEC)* (2020), pp. 1–8.

[16] P. Auer, N. Cesa-Bianchi, and P. Fischer, Machine Learning **47**, 235 (2002).

[17] J. Morimoto, Journal of Theoretical Biology **467**, 48 (2019).

[18] R. C. Wilson, E. Bonawitz, V. D. Costa, and R. B. Ebitz, Current Opinion in Behavioral Sciences **38**, 49 (2021).

[19] D. G. R. Tervo, M. Proskurin, M. Manakov, M. Kabra, A. Vollmer, K. Branson, and A. Y. Karpova, Cell **159**, 21 (2014).

[20] D. Bouneffouf, I. Rish, and G. A. Cecchi, in *Artificial General Intelligence*, edited by T. Everitt, B. Goertzel, and A. Potapov (Springer International Publishing, Cham, 2017), Lecture Notes in Computer Science, pp. 237–248, ISBN 978-3-319-63703-7.

[21] D. Marković, H. Stojić, S. Schwöbel, and S. J. Kiebel, Neural Networks **144**, 229 (2021).

[22] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau, in *Proceedings of the 3rd Machine Learning for Healthcare Conference* (PMLR, 2018), pp. 67–82, iSSN: 2640-3498.

[23] S. S. Villar, Probability in the Engineering and Informational Sciences **32**, 229 (2018), publisher: Cambridge University Press.

[24] S. S. Villar, J. Bowden, and J. Wason, Statistical Science **30**, 199 (2015), publisher: Institute of Mathematical Statistics.

[25] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha, in *Twenty-fourth international joint conference on artificial intelligence* (2015).

[26] B. Lin and D. Bouneffouf, in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (2022), pp. 1–8, iSSN: 1558-4739.

[27] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Nature **529**, 484 (2016).

[28] I. O. Ryzhov, W. B. Powell, and P. I. Frazier, Operations Research **60**, 180 (2012), publisher: INFORMS.

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning series (A Bradford Book, Cambridge, MA, USA, 1998), ISBN 978-0-262-19398-6.

[30] T. L. Lai and H. Robbins, Advances in Applied Mathematics **6**, 4 (1985).

[31] G. Reddy, A. Celani, and M. Vergassola, Journal of Statistical Physics **163**, 1454 (2016).

[32] S. Pilarski, S. Pilarski, and D. Varró, IEEE Transactions on Artificial Intelligence **2**, 2 (2021), conference Name: IEEE Transactions on Artificial Intelligence.

[33] W. R. Thompson, Biometrika **25**, 285 (1933), publisher: [Oxford University Press, Biometrika Trust].

[34] J. Honda and A. Takemura, in *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, edited by A. T. Kalai and M. Mohri (Omnipress, 2010), pp. 67–79.

[35] E. W. Ng and M. Geller, Journal of Research of the National Bureau of Standards, Section B: Mathematical Sciences **73B**, 1 (1969).

[36] A. Garivier and O. Cappé, Computing Research Repository - CORR (2011).

[37] E. Kaufmann, N. Korda, and R. Munos, in *Algorithmic Learning Theory*, edited by N. H. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann (Springer, Berlin, Heidelberg, 2012), Lecture Notes in Computer Science, pp. 199–213, ISBN 978-3-642-34106-9.

.

Supplemental material

S1. ENTROPY APPROXIMATION

Here, we derive the approximations $\tilde{S}_{\text{tail}}$ and $\tilde{S}_{\text{body}}$ constituting Eq. (6) in the main text. Equation (6) relies on the observation that that functional form of the posterior $p_{\max}$ can naturally be split in two distinct parts above and below the point $\tilde{\theta}_{\text{eq}} \approx \theta_{\text{eq}}$,

$$S = S_{\text{body}} + S_{\text{tail}}, \tag{S1}$$

with

$$S_{\text{body}} = -\int_{\theta_{\text{inf}}}^{\tilde{\theta}_{\text{eq}}} p_{\max}(\theta) \ln p_{\max}(\theta) d\theta\,, \quad S_{\text{tail}} = -\int_{\tilde{\theta}_{\text{eq}}}^{\theta_{\text{sup}}} p_{\max}(\theta) \ln p_{\max}(\theta) d\theta. \tag{S2}$$

The individual contribution, S_{body} and S_{tail} are easier to approximate using standard techniques than the full expression, S. We detail these approximations below.

A. Approximation of the main mode's contribution

The approximations leading to $\tilde{S}_{\text{body}}$ derives from decomposing S_{body} as:

$$S_{\text{body}} = -\int_{\theta_{\text{inf}}}^{\tilde{\theta}_{\text{eq}}} p_{\max} \ln\left(p_{\theta_{\max}>\theta_{\min}}\right) d\theta - \int_{\theta_{\text{inf}}}^{\tilde{\theta}_{\text{eq}}} p_{\max} \ln\left(1 + \frac{p_{\theta_{\min}>\theta_{\max}}}{p_{\theta_{\max}>\theta_{\min}}}\right) d\theta. \tag{S3}$$

To be able to perform the integration analytically, we extend the upper bound of the integrals from $\tilde{\theta}_{\text{eq}}$ to θ_{sup}. This requires neglecting the contribution from $p_{\theta_{\min}>\theta_{\max}}$ in the first term, resulting in the weight normalisation factor c_{tail} appearing in Eq. (6). We furthermore approximate $\ln(p_{\theta_{\max}>\theta_{\min}})$ by $\ln(p_{\theta_{\max}})$ in the first term. Next, we approximate the second term of Eq. (S3) by

$$\ln\left(1 + \frac{p_{\theta_{\min}>\theta_{\max}}}{p_{\theta_{\max}>\theta_{\min}}}\right) \approx A_{\text{c}} \frac{p_{\theta_{\min}>\theta_{\max}}}{p_{\theta_{\max}>\theta_{\min}} + p_{\theta_{\min}>\theta_{\max}}}, \tag{S4}$$

which is a variation of an approximation deduced from the Taylor series of the inverse hyperbolic tangent for $0 < \frac{p_{\theta_{\min}>\theta_{\max}}}{p_{\theta_{\max}>\theta_{\min}}} < 1$. First, note that this term should contribute significantly only when $\frac{p_{\theta_{\min}>\theta_{\max}}}{p_{\theta_{\max}>\theta_{\min}}} < 1$ since the entropy has already been partitioned. Thus, Eq. S4 choice is justified because it stays bounded even for $p_{\theta_{\min}>\theta_{\max}} \gg p_{\theta_{\max}>\theta_{\min}}$ which occurs since the integral bounds have been pushed above $\tilde{\theta}_{\text{eq}}$. Finally, A_{c} is obtained as the solution to:

$$\int_0^1 \ln(1+x) = A_{\text{c}} \int_0^1 \frac{x}{x+1}, \tag{S5}$$

leading to $A_{\text{c}} = 1.25889$. Taken altogether, this leads to Eq. (6).

B. Approximation of the tail contribution

The approximation expression for the tail contribution to the entropy, $\tilde{S}_{\text{tail}}$ [Eq. (7)] is obtained from $\tilde{S}_{\text{tail}}$ [Eq. (S2)] by neglecting the contribution from $i_{\max}$, i.e., approximating $p_{\max}$ by $p_{\theta_{\min}}$. This approximation requires our body-tail separator to precisely determine when the best arm contribution becomes sub-dominant. Rephrased differently, $\tilde{\theta}_{\text{eq}}$ approximates the transition value θ_{eq} where the current less expected arm will become more likely to be the maximum than the best expected arm. At long times, since $i_{\max}$ must be much more selected than the suboptimal one, we should observe a distribution $p_{\theta_{\max}}$ that is highly contracted compared to $p_{\theta_{\min}}$. This effect will result in a tail that is mostly dominated by $p_{\theta_{\min}}$ justifying our previous assumption.

S2. ANALYTICAL DERIVATION OF AIM

Here, we summarize all the steps leading to the analytic expressions used in AIM for 2-arms study case and exhibited in Supplementary Table S1.

	Bernoulli reward	Gaussian reward
θ_i, N_i	$\frac{r_i(t)+1}{n_i(t)+2}, n_i(t)+3$	$\frac{r_i(t)}{n_i(t)}, n_i(t)$
$V_i, \Delta, V_t,$	$\frac{\theta_i(1-\theta_i)}{N_i}, \theta_{\max}-\theta_{\min}, \frac{\theta_{\max}(1-\theta_{\max})}{N_{\max}}+\frac{\theta_{\min}(1-\theta_{\min})}{N_{\min}}$	$\frac{\sigma^2}{N_i}, \theta_{\max}-\theta_{\min}, \frac{\sigma^2}{N_{\max}}+\frac{\sigma^2}{N_{\min}}$
$\tilde{\theta}_{\text{eq}}$	$\theta_{\max}+\sqrt{2V_{\max}[N_{\min}K_{\mathbb{B}}(\theta_{\min},\theta_{\max})+\frac{1}{2}\ln\frac{N_{\max}}{N_{\min}}]}$	$\frac{N_{\max}\theta_{\max}-N_{\min}\theta_{\min}}{N_{\max}-N_{\min}}+\sqrt{\frac{4N_{\max}N_{\min}(\theta_{\max}-\theta_{\min})^2}{(N_{\max}-N_{\min})^2}+\frac{\sigma^2\ln\frac{N_{\max}}{N_{\min}}}{\lvert N_{\max}-N_{\min}\rvert}}$
c_{tail}	$1-I_{\tilde{\theta}_{\text{eq}}}(r_{\min}(t)+1, n_{\min}(t)-r_{\min}(t)+1)$	$\frac{1}{2}\text{erfc}\left(\frac{\sqrt{N_{\min}}(\tilde{\theta}_{\text{eq}}-\theta_{\min})}{\sqrt{2\sigma^2}}\right)$
$\tilde{S}_{\text{body}}$	$\frac{1}{4}\ln(2\pi V_{\max}e^{1-2A_{\text{c}}})+\frac{1}{4}\ln(2\pi V_{\max}e^{1+2A_{\text{c}}})\text{erf}\left[\frac{\Delta}{\sqrt{2V_t}}\right]$ $-\frac{\Delta V_{\max}}{2\sqrt{2\pi}V_{\text{t}}^{3/2}}e^{-\Delta^2/2V_{\text{t}}}$	$\frac{1}{4}\ln(2\pi V_{\max}e^{1-2A_{\text{c}}})+\frac{1}{4}\ln(2\pi V_{\max}e^{1+2A_{\text{c}}})\text{erf}\left[\frac{\Delta}{\sqrt{2V_t}}\right]$ $-\frac{\Delta V_{\max}}{2\sqrt{2\pi}V_{\text{t}}^{3/2}}e^{-\Delta^2/2V_{\text{t}}}$
$\tilde{S}_{\text{tail}}$	$(1-I_{\tilde{\theta}_{\text{eq}},m})[N_{\min}K_{\mathbb{B}}(\theta_{\min},\tilde{\theta}_{\text{eq}})+\frac{1}{2}\ln(2\pi V_{\min})]$	$\frac{1}{4}\ln(2\pi V_{\min}e)\text{erfc}\left(\frac{(\tilde{\theta}_{\text{eq}}-\theta_{\min})}{\sqrt{2V_{\min}}}\right)+\frac{(\tilde{\theta}_{\text{eq}}-\theta_{\min})}{2\sqrt{2\pi V_{\min}}}e^{-\frac{-(\tilde{\theta}_{\text{eq}}-\theta_{\min})^2}{2V_{\min}}}$

SUPPLEMENTARY TABLE S1. Analytic expressions for the terms of the approximate entropy Eq. (6) for Bernoulli (left) and Gaussian (right) reward distributions. To derive closed-form analytical expression for the Bernoulli bandits, we applied a Laplace approximation.

A. Gaussian approximation of the Beta posterior distribution

For the Bernoulli bandits, we approximate the Beta distributions by Gaussian distributions. To do so, we define θ_i, N_i such that

$$\mathbb{E}\left[X_{\mathcal{B}(r_i+1,n_i-r_i+1)}\right]=\frac{r_i+1}{n_i+2}=\theta_i, \tag{S6}$$

and

$$\begin{aligned}\text{Var}\left[X_{\mathcal{B}(r_i+1,n_i-r_i+1)}\right]&=\frac{r_i+1}{n_i+2}\left(1-\frac{n_i-r_i+1}{n_i+2}\right)\frac{1}{n_i+3}\\&=\frac{\theta_i(1-\theta_i)}{N_i}.\end{aligned} \tag{S7}$$

Thus, θ_i and N_i are respectively the mean and the number of draws that lead to a Gaussian approximation with the same two first moments as the true Beta distribution. (Note that for a Gaussian reward distribution, we have directly $\theta_i = r_i/n_i$ and $N_i = n_i$.)

B. The partitioning approximation

In this section we derive an approximation of the intersection point (defined above as $\tilde{\theta}_{\text{eq}}$) where the distributions $p_{\theta_{\max}>\theta_{\min}}$ and $p_{\theta_{\min}>\theta_{\max}}$ intersect at their highest value (if more than one solution exists).

We start with the case of Bernoulli bandits. The exact equation verified by the intersection point θ_{eq} is

$$e^{-N_{\max}K_{\mathbb{B}}(\theta_{\max},\theta_{\text{eq}})}\int_0^{\theta_{\text{eq}}}e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min},\theta')}d\theta'=e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min},\theta_{\text{eq}})}\int_0^{\theta_{\text{eq}}}e^{-N_{\max}K_{\mathbb{B}}(\theta_{\max},\theta')}d\theta'. \tag{S8}$$

Taking the logarithm of Eq. (S8) and normalizing the last term leads to

$$N_{\min}K_{\mathbb{B}}(\theta_{\min},\theta_{\text{eq}})-N_{\max}K_{\mathbb{B}}(\theta_{\max},\theta_{\text{eq}})+\frac{1}{2}\ln\frac{N_{\max}}{N_{\min}}+\ln\frac{\int_0^{\theta_{\text{eq}}}\sqrt{N_{\min}}e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min},\theta')}d\theta')}{\int_0^{\theta_{\text{eq}}}\sqrt{N_{\max}}e^{-N_{\max}K_{\mathbb{B}}(\theta_{\max},\theta')}d\theta'}=0. \tag{S9}$$

The distributions are uni-modal, and assuming that $(\theta_{\max}, N_{\max}) > (\theta_{\min}, N_{\min})$ and recalling that θ_{eq} is the highest intersection solution, we approximate θ_{eq} by neglecting the last term,

$$N_{\min}K_{\mathbb{B}}(\theta_{\min},\theta_{\text{eq}})-N_{\max}K_{\mathbb{B}}(\theta_{\max},\theta_{\text{eq}})+\frac{1}{2}\ln\frac{N_{\max}}{N_{\min}}\approx 0. \tag{S10}$$

In the long time limit $N_{\max} \gg N_{\min}$ and $\theta_{\rm eq}$ will be in the vicinity of $\theta_{\max}$ when the Gaussian expansion of the Kullback-Leibler divergence is relevant [in particular for $N_{\min} \sim O(\ln N_{\max})$]. Thus, we approximate $K_{\mathbb{B}}(\theta_{\min}, \theta_{\rm eq})$ by $K_{\mathbb{B}}(\theta_{\min}, \theta_{\max})$ and expand $K_{\mathbb{B}}(\theta_{\max}, \theta_{\rm eq})$ to lowest order in $\theta_{\rm eq}$, which leads to the expression given in Table S1,

$$\tilde{\theta}_{\rm eq} = \theta_{\max} + \sqrt{2V_{\max}\left[N_{\min} K_{\mathbb{B}}(\theta_{\min}, \theta_{\max}) + \frac{1}{2}\ln\frac{N_{\max}}{N_{\min}}\right]}, \tag{S11}$$

where $V_{\max} = \theta_{\max}(1-\theta_{\max})/N_{\max}$, which verifies $\tilde{\theta}_{\rm eq} - \theta_{\max} \sim o(N_{\max}^{-1/3})$, consistent with the Gaussian expansion of the Kullback-Leibler distance around $\theta_{\max}$.

We apply the same reasoning for Gaussian rewards, which leads to:

$$N_{\min}\frac{(\theta_{\rm eq}-\theta_{\min})^2}{2\sigma^2} - N_{\max}\frac{(\theta_{\rm eq}-\theta_{\max})^2}{2\sigma^2} + \frac{1}{2}\ln\frac{N_{\max}}{N_{\min}} \approx 0. \tag{S12}$$

Solving for $\theta_{\rm eq}$ leads to:

$$\tilde{\theta}_{\rm eq} = \frac{(N_{\max}\theta_{\max} - N_{\min}\theta_{\min})}{N_{\max}-N_{\min}} + \frac{2}{|N_{\max}-N_{\min}|} \times \sqrt{N_{\max}N_{\min}(\theta_{\max}-\theta_{\min})^2 + \sigma^2(N_{\max}-N_{\min})\ln\left(\frac{N_{\max}}{N_{\min}}\right)}. \tag{S13}$$

Note that the expressions, Eq. (S11) and Eq. (S13), rely on the assumption that $(\theta_{\max}, N_{\max}) > (\theta_{\min}, N_{\min})$. For $N_{\max} \leq N_{\min}$, the contributions from $p_{\theta_{\min}}$ and $p_{\theta_{\max}}$ do not intersect and $\tilde{\theta}_{\rm eq} = \theta_{\rm sup}$ (i.e., $\tilde{\theta}_{\rm eq} = 1$ and $\tilde{\theta}_{\rm eq} = +\infty$ for Bernoulli and Gaussian rewards, respectively), which means that the contribution from the tail is zero.

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

1. *Gaussian posterior distributions*

Here, we derive the $\tilde{S}_{\rm body}$ term given in Table S1 for Gaussian posterior distributions. Inserting the Gaussian form of the posterior into Eq. (8) gives:

$$\begin{aligned}\tilde{S}_{\rm body} = &-\int_{-\infty}^{+\infty}\frac{e^{-\Theta_{\max}^2}}{\sqrt{2\pi V_{\max}}}\frac{1}{2}\big[1+\mathrm{erf}(\Theta_{\min})\big]\left(-\frac{1}{2}\ln(2\pi V_{\max}) - \Theta_{\max}^2\right)\\ &-\int_{-\infty}^{+\infty} A_{\rm c}\frac{e^{-\Theta_{\min}^2}}{\sqrt{2\pi V_{\min}}}\frac{1}{2}\big[1+\mathrm{erf}(\Theta_{\max})\big]d\theta,\end{aligned} \tag{S14}$$

where V_i is the distribution's variance, $\Theta_{\max} = (\theta-\theta_{\max})/\sqrt{2V_{\max}}$, and $\Theta_{\min} = (\theta-\theta_{\min})/\sqrt{2V_{\min}}$. We integrate the constant part of the first term by use of the following identity [35]:

$$\int_{-\infty}^{\infty}\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\theta-\theta_1}{\sqrt{2V_1}}\right)\right]\frac{e^{-\frac{(\theta-\theta_2)^2}{2V_2}}}{\sqrt{2\pi V_2}} = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\theta_2-\theta_1}{\sqrt{2}\sqrt{V_2+V_1}}\right)\right], \tag{S15}$$

which leads to

$$\int_{-\infty}^{\infty}\frac{1}{2}\frac{e^{-\Theta_{\max}^2}}{\sqrt{2\pi V_{\max}}}\left[1+\mathrm{erf}(\Theta_{\min})\right]\frac{1}{2}\ln\left(2\pi V_{\max}\right)d\theta = \frac{1}{4}\ln\left(2\pi V_{\max}\right)\left[1+\mathrm{erf}\left(\frac{\theta_{\max}-\theta_{\min}}{\sqrt{2(V_{\max}+V_{\min})}}\right)\right]. \tag{S16}$$

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

$$\begin{aligned}\int_{-\infty}^{\infty}\Theta_{\max}^2\frac{1}{2}\left[1+\mathrm{erf}(\Theta_{\min})\right]\frac{e^{-\Theta_{\max}^2}}{\sqrt{2\pi V_{\max}}} &= \int_{-\infty}^{\infty}\frac{1}{4}\frac{e^{-\Theta_{\max}^2}}{\sqrt{2\pi V_{\max}}}\left[1+\mathrm{erf}(\Theta_{\min})\right] + \int_{-\infty}^{\infty}(\theta-\theta_{\max})\frac{1}{2}\frac{e^{-\Theta_{\max}^2}}{\sqrt{2\pi V_{\max}}}\frac{e^{-\Theta_{\min}^2}}{\sqrt{2\pi V_{\min}}}\\ &= \frac{1}{4}\left[1+\mathrm{erf}\left(\frac{\theta_{\max}-\theta_{\min}}{\sqrt{2(V_{\max}+V_{\min})}}\right)\right] + \frac{(\theta_{\min}-\theta_{\max})V_{\max}}{2\sqrt{2\pi}(V_{\max}+V_{\min})^{3/2}}e^{-\frac{(\theta_{\max}-\theta_{\min})^2}{2(V_{\max}+V_{\min})}},\end{aligned} \tag{S17}$$

where we also employed the identity of Eq. (S15).

Finally, the last term is also integrated using Eq. (S15), giving:

$$-A_{\rm c}\int_{-\infty}^{\infty}\frac{1}{2}\left[1+\mathrm{erf}(\Theta_{\rm max})\right]\frac{e^{-\Theta_{\rm min}^2}}{\sqrt{2\pi V_{\rm min}}}=-\frac{A_{\rm c}}{2}\left[1+\mathrm{erf}\left(\frac{\theta_{\rm min}-\theta_{\rm max}}{\sqrt{2(V_{\rm max}+V_{\rm min})}}\right)\right]. \tag{S18}$$

Combining Eq. (S16), Eq. (S17) and Eq. (S18) leads to the expression given in Table S1.

2. Bernoulli posterior distributions

Reminding S2 C, the analytic derivation of $\tilde{S}_{\rm body}$ made for Gaussian reward [Eq. (S16), Eq.(S17) and Eq. (S18)] can be extended to Bernoulli reward with θ_i and N_i thus obtained.

D. Closed-form expressions for the tail contribution

We conclude our approach by considering the tail contribution to the approximate entropy.

1. Gaussian posterior distribution

We first consider the Gaussian reward case for which the contribution from the tail can be derived exactly,

$$\begin{aligned}\tilde{S}_{\rm tail}&=\int_{\theta_{eq}}^{\infty}\frac{e^{-\Theta_{\rm min}^2}}{\sqrt{2\pi V_{\rm min}}}\left[\frac{1}{2}\ln(2\pi V_{\rm min})+\Theta_{\rm min}^2\right]\\&=\frac{1}{4}\ln(2\pi V_{\rm min}e)\mathrm{erfc}\left(\frac{\tilde{\theta}_{\rm eq}-\theta_{\rm min}}{\sqrt{2V_{\rm min}}}\right)+\frac{\tilde{\theta}_{\rm eq}-\theta_{\rm min}}{2\sqrt{2\pi V_{\rm min}}}e^{-\frac{(\tilde{\theta}_{\rm eq}-\theta_{\rm min})^2}{2V_{\rm min}}}.\end{aligned} \tag{S19}$$

2. Bernoulli posterior distribution

We now focus on the Bernoulli case for which $\tilde{S}_{\rm tail}$ requires a second approximation in order to get rid of the numerical integration. We have:

$$\begin{aligned}\tilde{S}_{\rm tail}&\approx -c_{\rm tail}\ln(p_{\theta_{\rm min}}(\tilde{\theta}_{\rm eq}))\\&\approx\left(1-I_{\tilde{\theta}_{\rm eq},m}\right)\left[N_{\rm min}K_{\mathbb{B}}(\theta_{\rm min},\tilde{\theta}_{\rm eq})+\frac{1}{2}\ln(2\pi V_{\rm min})\right],\end{aligned} \tag{S20}$$

where $I_{\tilde{\theta}_{\rm eq},m}$ is the normalized incomplete beta function evaluated at $\tilde{\theta}_{\rm eq}$ with parameters $r_{\rm min}+1$ and $n_{\rm min}-r_{\rm min}+1$. Of note, we have bounded $\ln(p_{\theta_{\rm min}}(\theta))$ by its value at $\tilde{\theta}_{\rm eq}$ and included only the leading order at large times. We remark that Eq. (S20) is one possible solution, but others would work as well as long as their leading order is given by $c_{\rm tail}$.

Applying Eq. (5) to the obtained analytic expression Eq. (6), leads to the AIM algorithm.

S3. AIM ALGORITHMS

Here, we summarize the algorithm procedures introduced in the main text.

A. Two-armed Gaussian bandit

1. Draw each arm once, update $r_i(t), n_i(t)$ and their associated θ_i, N_i according to Table S1.

2. For $t > 2$, sort the arms according to $\theta_i(t)$ to get $(\theta_{\min}(t), N_{\min}(t), n_{\min}(t), r_{\min}(t))$ and $(\theta_{\max}(t), N_{\max}(t), n_{\max}(t), r_{\max}(t))$ couples.

   - If $\theta_{\min}(t) = \theta_{\max}(t)$ then choose the arm which has been currently drawn the least. If $N_{\min}(t) = N_{\max}(t)$ choose randomly.
   - Otherwise, if $N_{\min}(t) \geq N_{\max}(t)$ then draw $i_{\max}$.
   - Otherwise, evaluate $\tilde{\theta}_{\text{eq}}$ according to Table S1. Then, evaluate the absolute value of the gradient of $\tilde{S}(r_i(t), n_i(t), r_j(t), n_j(t), \tilde{\theta}_{\text{eq}})$ along each arm according to:

$$\Delta_i \tilde{S} = \left| \frac{1}{2}\tilde{S}(r_i(t) + \theta_i(t) + \alpha\sigma, n_i(t)+1, ..) + \frac{1}{2}\tilde{S}(r_i(t) + \theta_i(t) - \alpha\sigma, n_i(t)+1, ..) - \tilde{S}(r_i(t), n_i(t), ..) \right|, \quad \text{(S21)}$$

     where the dots refer to constant variables ($r_j(t)$, $n_j(t)$, $\tilde{\theta}_{\text{eq}}$), $\alpha = 1$, and $\tilde{S}$ given by Eq. (6) with Table S1. Next, draw the arm with the highest gradient.

3. Update $r_i(t+1), n_i(t+1)$ according to the reward returned by the chosen arm.

Let us draw some additional observations. First, we stress that if $N_{\max} \leq N_{\min}$, the current best arm, $i_{\max}$, is automatically drawn. It is because the current best arm should always be played in this case (since the information about it is lesser than the current worst arm, $i_{\min}$). Next, $\tilde{S}$ also requires to sort its input values. Thus, $\theta_{\max}(t), N_{\max}(t), \theta_{\min}(t)$ and $N_{\min}(t)$ used in each $\tilde{S}$ evaluation are different from the ones used in the sorting step 2. However, $\tilde{\theta}_{\text{eq}}$ is shared among all $\tilde{S}$ evaluations. This avoids adding perturbations induced by the cutoff of the tail. Finally, one should note that Eq. S21 is a particular way to evaluate the gradient, but other approaches are possible (i.e., by modifying α or computing the higher derivative orders).

B. Two-armed Bernoulli bandit

For the Bernoulli bandit, most of the procedure is identical to the Gaussian case described above. One simply has to replace the expressions for the case of Gaussian rewards by those corresponding to Bernoulli rewards given in Table S1. The only difference in the procedure regards the gradient evaluation [Eq. (S21), which is replaced by:

$$\Delta_i \tilde{S} = \left| \frac{r_i(t)}{n_i(t)}\tilde{S}(r_i(t)+1, n_i(t)+1, ..) + \frac{n_i(t) - r_i(t)}{n_i(t)}\tilde{S}(r_i(t), n_i(t)+1, ..) - \tilde{S}(r_i(t), n_i(t), ..) \right|, \quad \text{(S22)}$$

where, as above, the two dots refer to constant variables $r_j(t)$, $n_j(t)$ and $\tilde{\theta}_{\text{eq}}$.

C. K > 2 armed Gaussian bandit

To further assess the efficiency of the AIM algorithm, we address the multi-armed case (with a number of arms $K > 2$). We notice that Eq. (6) is asymmetric in $(i_{\max}, i_i)$, which suggests that all but $i_{\max}$ could be decoupled in the general entropy expression by neglecting the correlations between subdominant arms. In practice, we thus propose to evaluate the $K-1$ gradients between each subdominant arm and $i_{\max}$ by the use of Eq. (6) with each subdominant arm in place of $i_{\min}$. The dominant arm $i_{\max}$ is pulled if all the gradient evaluations favor $i_{\max}$, and the subdominant arm with the highest absolute gradient is chosen if at least one gradient favors a subdominant arm. Thus, the AIM implementation for $K > 2$ reads:

1. Draw each arm once, and update $r_i(t), n_i(t)$ and their associated θ_i, N_i according to Table S1.

2. At each step $t > K$, determine the arm with the best empirical mean reward such that:

   - $\theta_{\max} > \theta_i \ \ \forall i \neq \max$,
   - or $i_{\max} = \underset{\{i:\theta_i = \theta_{\max}\}}{\text{argmax}}(N_i)$ if the dominant arm is not unique. If the maximal N_i is not unique, $i_{\max}$ is drawn randomly among $\{i : \theta_i = \theta_{\max} \wedge N_i = N_{\max}\}$.

3. Compare each arm i to $i_{\max}$ by computing the two following gradients:

$$\Delta_i\tilde{S} = \left|\frac{1}{2}\tilde{S}(r_i(t)+\theta_i(t)+\alpha\sigma, n_i(t)+1, ..) + \frac{1}{2}\tilde{S}(r_i(t)+\theta_i(t)-\alpha\sigma, n_i(t)+1, ..) - \tilde{S}(r_i(t), n_i(t), ..)\right|, \tag{S23}$$

where the two dots refer to constant variables ($r_{\max}(t)$, $n_{\max}(t)$, $\tilde{\theta}_{\rm eq}$), and

$$\Delta_{i,\max}\tilde{S} = \left|\frac{1}{2}\tilde{S}(r_{\max}(t)+\theta_{\max}(t)+\alpha\sigma, n_{\max}(t)+1, ..) + \frac{1}{2}\tilde{S}(r_{\max}(t)+\theta_{\max}(t)-\alpha\sigma, n_{\max}(t)+1, ..) - \tilde{S}(r_{\max}(t), n_{\max}(t), ..)\right|, \tag{S24}$$

where the two dots refer to constant variables $r_i(t)$, $n_i(t)$ and $\tilde{\theta}_{\rm eq}$, and $\tilde{\theta}_{\rm eq}$ is computed following Table S1 with $r_i(t), n_i(t), r_{\max}(t), n_{\max}(t)$.

4. Then, select the arm A_t such that:

- $A_t = i_{\max}$ if $\Delta_i\tilde{S} < \Delta_{i,\max}\tilde{S}, \forall i \neq \max$,
- or $A_t = \{\underset{\Delta_i\tilde{S}-\Delta_{i,\max}\tilde{S}\geq 0}{\rm argmax}(\Delta_i\tilde{S} - \Delta_{i,\max}\tilde{S})\}$ elsewise. If there is more than one solution A_t is drawn randomly among them.

5. Update $r_i(t+1), n_i(t+1)$ according to the reward returned by the chosen arm.

D. K > 2 armed Bernoulli bandit

For the multi-armed Bernoulli bandit, most of the procedure is identical to the multi-armed Gaussian algorithm described above. As for the two-armed case, the procedure is implemented by replacing the expressions by those for Bernoulli rewards given in Table S1. The expressions for the gradients are replaced by:

$$\Delta_i\tilde{S} = \left|\frac{r_i(t)}{n_i(t)}\tilde{S}(r_i(t)+1, n_i(t)+1, ..) + \frac{n_i(t)-r_i(t)}{n_i(t)}\tilde{S}(r_i(t), n_i(t)+1, ..) - \tilde{S}(r_i(t), n_i(t), ..)\right|, \tag{S25}$$

where the two dots refer to constant variables ($r_{\max}(t)$, $n_{\max}(t)$, $\tilde{\theta}_{\rm eq}$), and

$$\Delta_{i,\max}\tilde{S} = \left|\frac{r_{\max}(t)}{n_{\max}(t)}\tilde{S}(r_{\max}(t)+1, n_{\max}(t)+1, ..) + \frac{n_{\max}(t)-r_{\max}(t)}{n_{\max}(t)}\tilde{S}(r_{\max}(t), n_{\max}(t)+1, ..) - \tilde{S}(r_{\max}(t), n_{\max}(t), ..)\right|, \tag{S26}$$

where the two dots refer to constant variables $r_i(t)$, $n_i(t)$ and $\tilde{\theta}_{\rm eq}$.

S4. OTHER STATE-OF-THE-ART BANDIT ALGORITHMS

Here, we briefly review several baseline algorithms which provide a benchmark of our gradient method.

A. Epsilon-n-Greedy

This method is a variation of the ϵ-greedy strategy, and is one of the most widely used bandit algorithms due to its undeniable simplicity [29]. The ϵ-greedy strategy selects either a random arm with a probability ϵ or the current dominant arm otherwise. The $\epsilon_{\rm n}$-greedy strategy is a generalized form of this approach where the parameter ϵ is a time-dependent function $\epsilon(t) = \min\{1, c(\mu_1,\mu_2)K/(d^2t)\}$. The constant c is a hyperparameter of the method, which needs to be tuned for optimal performance. Here, we used $c = 10$ tuned for Bernoulli uniform priors and $c = 30$ for Gaussian uniform priors. Let us stress that $\epsilon_{\rm n}$-greedy relies on a priori knowledge of the distribution of $\{\mu_1, \mu_2\}$ in order to be effective.

B. UCB-Tuned

This method belong to the class of upper confidence bound (UCB) algorithms which pull the arm maximising a proxy function generally defined as $F_i = \theta_i + R_i$ where R_i bounds the regret to logarithmic growth. For UCB-tuned, $R_i$is given by:

$$R_i = c(\mu_1, \mu_2)\sqrt{\frac{\ln(t)}{n_i(t)}\min\left(\frac{1}{4}, s_i(t)\right)}, \;\; s_i(t) = \hat{\sigma_i}^2 + \sqrt{\frac{2\ln(t)}{n_i(t)}}, \tag{S27}$$

where $\hat{\sigma_i}^2$ is the reward variance and c a hyperparameter. Here, we rely on the optimised version of F_i proposed in [32] for Bernoulli reward:

$$F_i = \frac{r_i(t)+1}{n_i(t)+2} + c(\mu_1, \mu_2)\sqrt{\frac{\ln(t+2K)}{n_i(t)+2}\min\big(d, s_i(t)\big)}, \; s_i(t) = \frac{(r_i(t)+1)(n_i(t)-r_i(t)+1)}{(n_i(t)+2)^2(n_i(t)+3)} + \sqrt{\frac{2\ln(t+2K)}{n_i(t)+2}}, \tag{S28}$$

with $c = 0.73$ and $d = 0.19$. For Gaussian rewards we adapt (S27) with $c = 2.1$ and $\hat{\sigma_i}^2 = \frac{\sigma^2}{n_i(t)}$, although this is not necessarily optimal.

C. KL-UCB

This method is an another upper confidence bound (UCB) variant which has been especially designed for bounded reward, and in particular for Bernoulli distributed rewards where it reaches the Lai and Robbins bound [36]. For KL-UCB, F_i reads:

$$F_i = \max\Big\{\theta \in \Theta : N_i K_{\mathbb{B}}\left(\frac{r_i(t)}{n_i(t)}, \theta\right) \leq \ln(t) + c(\mu_1, \mu_2)\ln(\ln(t))\Big\}, \tag{S29}$$

where Θ denotes the definition interval of the posterior distribution. By testing various c values, we end up with $c(\mu_1, \mu_2) = 0.00001$.

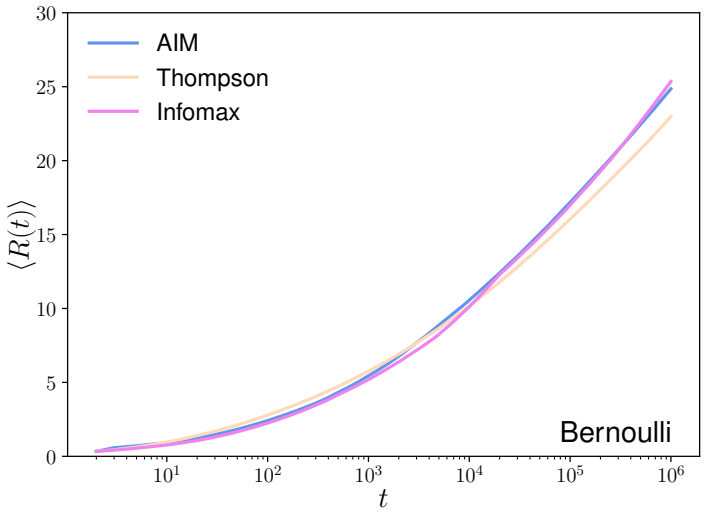
D. Thompson sampling

At each step, Thompson sampling [33, 37] stochastically selects an arm based on the posterior probability that it maximizes the expected reward. In practice, after drawing K random values according to each arms' posterior distribution, it picks the arm with the largest value:

$$A_t = \underset{i=1..K}{\operatorname{argmax}}\left(X_{\mathcal{B}(r_i+1, n_i-r_i+1)}\right). \tag{S30}$$

E. Infomax

At each step, Infomax relies on a greedy entropy minimization to decide the arm to be played. Here we adapt the Infomax [2, 31] algorithm by replicating the steps detailed in Supplementary Section S3 B but replacing $\tilde{S}$ by a numerical integration of S [Eq. (3)]. The regret performance of the Infomax algorithm thus obtained is compared with AIM and Thompson methods in Fig. S1.



SUPPLEMENTARY FIG. S1. Comparison of AIM and Thompson sampling with the Infomax algorithm based on the exact entropy. Mean regret as function of time for 2-armed Bernoulli bandits with parameters drawn uniformly in $]0,1[$. In light blue AIM , in orange Thompson, and in pink our Infomax gradient implementation (see S4 and S3) with the exact entropy obtained by numerical integration. The algorithms relying on the exact entropy or its analytic approximation show very similar performance.

S5. ASYMPTOTIC OPTIMALITY OF APPROXIMATE INFORMATION MAXIMISATION

Here, we provide details on the asymptotic behaviour of the entropy at fixed $\theta_{\max}$, $\theta_{\min}$, and $t \gg 1$. Recall that the derivative of Eq. (6), from which we seek to derive the asymptotic behavior of $N_{\min}$ by minimizing $\tilde{S}$, is given by:

$$\frac{\partial \tilde{S}}{\partial N_{\min}} = (1 - c_{\text{tail}}) \frac{\partial \tilde{S}_{\text{body}}}{\partial N_{\min}} + \frac{\partial \tilde{S}_{\text{tail}}}{\partial N_{\min}} + (-\tilde{S}_{\text{body}} + \ln(1 - c_{\text{tail}}) + 1) \frac{\partial c_{\text{tail}}}{\partial N_{\min}}. \tag{S31}$$

We will focus on Bernoulli reward distributions, and thus the terms of $\tilde{S}$ given by Table S1. Since the different terms share common derivatives, we first provide the $\tilde{\theta}_{\text{eq}}$ derivative. To leading order it is equal to:

$$\begin{aligned} \frac{\partial \tilde{\theta}_{\text{eq}}}{\partial N_{\min}} &= \frac{\theta_{\max}(1 - \theta_{\max})}{\tilde{\theta}_{\text{eq}} - \theta_{\max}} \frac{t(2N_{\min} K_{\mathbb{B}}(\theta_{\min}, \theta_{\max}) - 1) + N_{\min} \ln(t/N_{\min} - 1)}{2(t - N_{\min})^2 N_{\min}} \\ &= O\left(\frac{\ln(t)}{\sqrt{t N_{\min}}}, \sqrt{\frac{\ln(t)}{\sqrt{t}}} \right). \end{aligned} \tag{S32}$$

Next, we focus on the norm and main mode terms, c_{tail} and $\tilde{S}_{\text{body}}$, respectively. Since $V_{\max}, V_{\text{t}} \ll V_{\min}$, all the terms that depend exponentially on $V_{\max}$ or V_{t} are negligible to leading order. We thus obtain

$$\frac{\partial \tilde{S}_{\text{body}}}{\partial N_{\min}} \sim \frac{1}{2(t - N_{\min})} + O(\exp(-C_0 t)), \tag{S33}$$

with $C_0 > 0$.

Next, we consider the tail terms of which we propose to rewrite the regularized incomplete beta distribution as

$$\begin{aligned} c_{\text{tail}} &= 1 - I_{\tilde{\theta}_{\text{eq}}, m} \\ &= C(\hat{\theta}_{\min}, \hat{N}_{\min}) \sqrt{\hat{N}_{\min}} \int_{\tilde{\theta}_{\text{eq}}}^{1} e^{-\hat{N}_{\min} K_{\mathbb{B}}(\hat{\theta}_{\min}, \theta)} d\theta \\ &= C(\hat{\theta}_{\min}, \hat{N}_{\min}) \sqrt{\hat{N}_{\min}} e^{-\hat{N}_{\min} K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\text{eq}})} \int_{\tilde{\theta}_{\text{eq}}}^{1} e^{-\hat{N}_{\min} \left[\hat{\theta}_{\min} \ln(\frac{\tilde{\theta}_{\text{eq}}}{\theta}) + (1 - \hat{\theta}_{\min}) \ln(\frac{1 - \tilde{\theta}_{\text{eq}}}{1 - \theta}) \right]} d\theta, \end{aligned} \tag{S34}$$

where $\hat{\theta}_{\min} = \frac{r_{\min}}{n_{\min}}$, $\hat{N}_{\min} = n_{\min}$ and $C(\hat{\theta}_{\min}, \hat{N}_{\min}) \sim [2\pi\hat{\theta}_{\min}(1-\hat{\theta}_{\min})]^{-\frac{1}{2}} + O(\hat{N}_{\min}^{-\frac{1}{2}})$ by convergence of the Beta distribution to its Gaussian counterpart. Next, we partition the integral, which denotes I, at a cutoff $\mu_c = \hat{N}_{\min}(\theta_c - \tilde{\theta}_{\rm eq})$, leading to:

$$I = \int_{\theta_c}^{1} e^{-\hat{N}_{\min}\left[\hat{\theta}_{\min}\ln(\frac{\tilde{\theta}_{\rm eq}}{\theta}) + (1-\hat{\theta}_{\min})\ln(\frac{1-\tilde{\theta}_{\rm eq}}{1-\theta})\right]} d\theta + \int_{0}^{\mu_c} \frac{e^{\hat{N}_{\min}\left[\hat{\theta}_{\min}\ln(1+\frac{\mu}{\tilde{\theta}_{\rm eq}\hat{N}_{\min}}) + (1-\hat{\theta}_{\min})\ln(1-\frac{\mu}{\hat{N}_{\min}(1-\tilde{\theta}_{\rm eq})})\right]}}{\hat{N}_{\min}} d\mu. \tag{S35}$$

Taking $\mu_c \sim A\sqrt{\hat{N}_{\min}}$, we obtain a well-defined expansion of the second integral, which leads to:

$$\begin{aligned} I &= Ce^{\hat{N}_{\min}\left[\hat{\theta}_{\min}\ln(1+\frac{A}{\tilde{\theta}_{\rm eq}\sqrt{\hat{N}_{\min}}}) + (1-\hat{\theta}_{\min})\ln(1-\frac{A}{(1-\tilde{\theta}_{\rm eq})\sqrt{\hat{N}_{\min}}})\right]} + \int_0^{A\sqrt{\hat{N}_{\min}}} \frac{e^{\mu(\frac{\hat{\theta}_{\min}}{\tilde{\theta}_{\rm eq}} - \frac{1-\hat{\theta}_{\min}}{1-\tilde{\theta}_{\rm eq}})}}{\hat{N}_{\min}}\left[1 + A_{\min}\frac{\mu^2}{\hat{N}_{\min}} + O(\frac{\mu^3}{\hat{N}_{\min}^2})\right] d\mu \\ &= \frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{N_{\min}(\tilde{\theta}_{\rm eq} - \theta_{\min})} + O(N_{\min}^{-2}). \end{aligned} \tag{S36}$$

where the difference between $\theta_{\min}$ and $\hat{\theta}_{\min}$ is of the order $O(N_{\min}^{-1})$ and are then included in the second term.

By the use of Eq. (S36), we find

$$c_{\rm tail} \sim C_2 \frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{\sqrt{N_{\min}}(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}\left[1 + O(N_{\min}^{-1})\right], \tag{S37}$$

to leading order. Using this gives us the dominant order of the variation of $c_{\rm tail}$:

$$\begin{aligned} \frac{\partial c_{\rm tail}}{\partial N_{\min}} &= -\frac{C_2}{2N_{\min}^{\frac{3}{2}}}\frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})} - C_2\frac{\partial \tilde{\theta}_{\rm eq}}{\partial N_{\min}}\left[1 + \frac{\theta_{\min}(1-\theta_{\min})}{(\theta_{\min} - \tilde{\theta}_{\rm eq})^2}\right]\frac{e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}}{\sqrt{N_{\min}}} \\ &\quad - C_2\frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{\sqrt{N_{\min}}(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}\left[K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq}) + N_{\min}\frac{\partial \tilde{\theta}_{\rm eq}}{\partial N_{\min}}\left(\frac{1-\theta_{\min}}{1-\tilde{\theta}_{\rm eq}} - \frac{\theta_{\min}}{\tilde{\theta}_{\rm eq}}\right)\right] \\ &\quad + \frac{\partial C_2}{\partial N_{\min}}\frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{\sqrt{N_{\min}}(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})} + O\left(e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}N_{\min}^{-2}\right), \end{aligned} \tag{S38}$$

where $C_2 = Ce^{N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}e^{-\hat{N}_{\min}K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\rm eq})}$ accounts for the change of variable between $(\hat{\theta}_{\min}, \hat{N}_{\min})$ and $(\theta_{\min}, N_{\min})$. The derivative of C_2 is to leading order:

$$\frac{\partial C_2}{\partial N_{\min}} \sim C_2 O\left(\frac{1}{N_{\min}^{3/2}}\right). \tag{S39}$$

Combining Eq. (S39) with Eq. (S38) yields the leading order of the derivative of $c_{\rm tail}$:

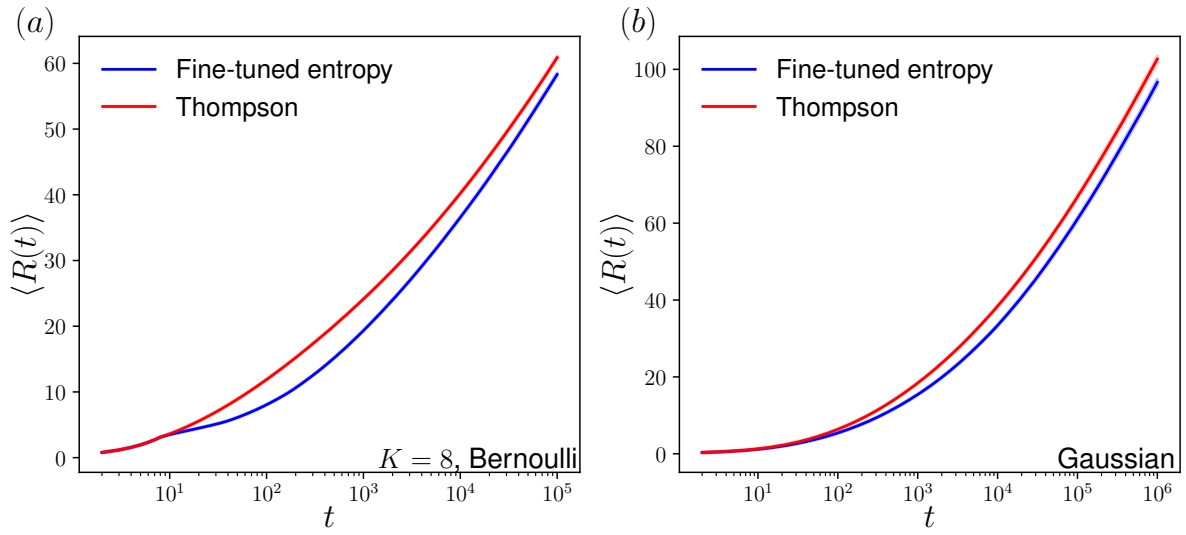
$$\frac{\partial c_{\rm tail}}{\partial N_{\min}} \sim -C_2 K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\rm eq})\frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{\sqrt{N_{\min}}(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}\left(1 + O\left(\frac{1}{N_{\min}}\right)\right). \tag{S40}$$

We finally expand $\tilde{S}_{\rm tail}$ to leading order, which yields:

$$\begin{aligned} \frac{\partial \tilde{S}_{\rm tail}}{\partial N_{\min}} &= \frac{\partial c_{\rm tail}}{\partial N_{\min}}[N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq}) + \frac{1}{2}\ln(2\pi V_{\min})] + c_{\rm tail}\left[N_{\min}\frac{\partial \tilde{\theta}_{\rm eq}}{\partial N_{\min}}\left(\frac{1-\theta_{\min}}{1-\tilde{\theta}_{\rm eq}} - \frac{\theta_{\min}}{\tilde{\theta}_{\rm eq}}\right) + K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq}) - \frac{1}{2N_{\min}}\right] \\ &= -C_2 K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\rm eq})^2\sqrt{N_{\min}}\frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}\left(1 + O\left(\frac{1}{N_{\min}}\right)\right). \end{aligned} \tag{S41}$$

Inserting the leading order terms of Eqs. (S33), (S38), and (S41) into Eq. (S31) leads to:

$$\begin{aligned} \frac{\partial \tilde{S}}{\partial N_{\min}} &= \frac{1}{2(t - N_{\min})} - C_2\sqrt{N_{\min}}K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\rm eq})^2\frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}\left(1 + O\left(\frac{1}{N_{\min}}\right)\right) \\ &\quad - \frac{1}{2}\ln(t)C_2 K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\rm eq})\frac{\tilde{\theta}_{\rm eq}(1-\tilde{\theta}_{\rm eq})}{\sqrt{N_{\min}}(\tilde{\theta}_{\rm eq} - \theta_{\min})} e^{-N_{\min}K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\rm eq})}\left(1 + O\left(\frac{1}{N_{\min}}\right)\right). \end{aligned} \tag{S42}$$



SUPPLEMENTARY FIG. S2. Mean regret for 8-armed Bernoulli **(a)** and Gaussian **(b)** bandits with parameters drawn uniformly in $]0, 1[$. In red Thompson sampling and in blue tuned AIM (see Section S6).

Finally, setting $\frac{\partial \tilde{S}}{\partial N_{\min}} = 0$ leads to:

$$\frac{1}{t} \sim 2C_2 \sqrt{N_{\min}} K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\text{eq}}) \frac{\tilde{\theta}_{\text{eq}}(1-\tilde{\theta}_{\text{eq}})}{(\tilde{\theta}_{\text{eq}} - \theta_{\min})} e^{-N_{\min} K_{\mathbb{B}}(\theta_{\min}, \tilde{\theta}_{\text{eq}})} \left(1 + \frac{\ln(t)}{2K_{\mathbb{B}}(\hat{\theta}_{\min}, \tilde{\theta}_{\text{eq}}) N_{\min}}\right), \tag{S43}$$

which, by taking the logarithm and noting that $\tilde{\theta}_{\text{eq}} \to \theta_{\max}$, leads to:

$$\ln(t) \sim N_{\min} K_{\mathbb{B}}(\theta_{\min}, \theta_{\max}) - \ln \left[2C_2 \sqrt{N_{\min}} K_{\mathbb{B}}(\hat{\theta}_{\min}, \theta_{\max}) \frac{\theta_{\max}(1-\theta_{\max})}{(\theta_{\max} - \theta_{\min})} \left(1 + \frac{\ln(t)}{2K_{\mathbb{B}}(\hat{\theta}_{\min}, \theta_{\max}) N_{\min}}\right) \right]. \tag{S44}$$

Hence, we obtain the Lai and Robbins relation for the AIM algorithm:

$$N_{\min} \sim \frac{\ln(t)}{K_{\mathbb{B}}(\theta_{\min}, \theta_{\max})} + o(\ln(t)). \tag{S45}$$

S6. TUNED APPROXIMATE INFORMATION MAXIMIZATION

Here, we detail how the AIM algorithm can be tuned for specific multi-armed bandit problems, showing its capacity to outperform Thompson sampling (see Fig. S2). We propose some empirically optimised variations to the functional form of the approximate entropy $\tilde{S}$ [Eq. (6) and Table S1] derived in the main text. These lead to a simplified version of the approximate entropy, which is adjusted to provide a better expected regret in case of expected rewards drawn according to a uniform prior while still respecting the Lai and Robbins bound.

A. Bernoulli rewards with $K > 2$

To obtain a tuned version of AIM for multi-armed bandits, we propose to simplify the main mode term by keeping the dominant term (when $\Delta/\sqrt{2V_t} \to \infty$) which we multiplied by two:

$$\tilde{S}_{\text{body}} = -\ln \left(2\pi \frac{\theta_{\max}(1-\theta_{\max})}{N_{\max}}\right). \tag{S46}$$

We also simplified the expression by neglecting the contribution from c_{tail} (i.e., letting $c_{\text{tail}} \to 0$). This leads to the tuned expression:

$$\tilde{S} \approx \tilde{S}_{\text{body}} + \tilde{S}_{\text{tail}}. \tag{S47}$$

Thus, the tuned version of AIM for multi-armed Bernoulli bandits consists in replacing $\tilde{S}$ functional in Eq. (S25) and Eq. (S26) by Eq. (S47).

B. Gaussian rewards with $K > 2$

For the multi-armed Gaussian bandits, we propose the following form:

$$\tilde{S}(\theta_{\max}, N_{\max}, N_{\min}, \theta_{\min}) = \frac{1}{8}\left[1 + \operatorname{erf}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min}}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right)\right] \ln\left(\frac{2\pi e^{1-2A_{\text{c}}}\sigma^2}{N_{\max}}\right) + 2\ln\left(\frac{2\pi e\sigma^2}{N_{\min}}\right) \operatorname{erfc}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min}}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right). \tag{S48}$$

Since, Eq. (S48) exhibits a simple closed-form expression, it is possible to derive an exact and explicit expression of its expected gradient for continuous Gaussian reward distributions,

$$\begin{aligned}\Delta_{\max,\min} S_{\text{p}} &= \int_{-\infty}^{\infty} \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\left[|S_{\text{p}}(\frac{\theta_{\max}N_{\max}+\mu}{N_{\max}+1}, N_{\max}+1, ..) - S_{\text{p}}(..)| - |S_{\text{p}}(.., \frac{\theta_{\min}N_{\min}+\mu}{N_{\min}+1}, N_{\min}+1) - S_{\text{p}}(..)|\right] d\mu \\ &= \int_{-\infty}^{\infty} \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\left[|\Delta_{\max} S_{\text{p}}| - |\Delta_{\min} S_{\text{p}}|\right],\end{aligned} \tag{S49}$$

where the two dots refer to constant variables. Noticing that the first term (variation along max) is independent of the integration variable, we obtain:

$$\Delta_{\max}\tilde{S} = -\frac{1}{8}\left[1 + \operatorname{erf}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min})}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right)\right] \ln\left(1 + \frac{1}{N_{\max}}\right). \tag{S50}$$

By use of the identity Eq. (S15) this leads to:

$$\begin{aligned}\Delta_{\min}\tilde{S} = \frac{1}{8}\ln\left(\frac{2\pi e^{1-2A_{\text{c}}}\sigma^2}{N_{\max}}\right)&\left[\operatorname{erf}\left(\frac{(\tilde{\theta}_{\text{eq}} - \theta_{\min})(N_{\min}+1)}{\sqrt{2\sigma^2}\sqrt{N_{\min}+2}}\right) - \operatorname{erf}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min}}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right)\right] \\ &+ 2\ln\left(\frac{2\pi\sigma^2 e^1}{N_{\min}+1}\right)\operatorname{erfc}\left(\frac{(\tilde{\theta}_{\text{eq}} - \theta_{\min})(N_{\min}+1)}{\sqrt{2\sigma^2}\sqrt{N_{\min}+2}}\right) - 2\ln\left(\frac{2\pi\sigma^2 e^1}{N_{\min}}\right)\operatorname{erfc}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min}}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right).\end{aligned} \tag{S51}$$

Combining S50 and S51 leads to the complete expression of the gradient:

$$\begin{aligned}\Delta_{\max,\min}\tilde{S} = \frac{1}{8}&\left[1 + \operatorname{erf}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min}}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right)\right] \ln\left(1 + \frac{1}{N_{\max}}\right) \\ &- \frac{1}{8}\ln\left(\frac{N_{\max}}{2\pi e^{1-2A_{\text{c}}}\sigma^2}\right)\left[\operatorname{erf}\left(\frac{(\tilde{\theta}_{\text{eq}} - \theta_{\min})(N_{\min}+1)}{\sqrt{2\sigma^2}\sqrt{N_{\min}+2}}\right) - \operatorname{erf}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min}}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right)\right] \\ &- 2\ln\left(\frac{N_{\min}+1}{2\pi\sigma^2 e^1}\right)\operatorname{erfc}\left(\frac{(\tilde{\theta}_{\text{eq}} - \theta_{\min})(N_{\min}+1)}{\sqrt{2\sigma^2}\sqrt{N_{\min}+2}}\right) + 2\ln\left(\frac{N_{\min}}{2\pi\sigma^2 e^1}\right)\operatorname{erfc}\left(\frac{\tilde{\theta}_{\text{eq}} - \theta_{\min}}{\sqrt{2\sigma^2 N_{\min}^{-1}}}\right).\end{aligned} \tag{S52}$$

Thus, the tuned and continuous version of AIM for Gaussian rewards consists in replacing gradient evaluation in Eq. (S21) by Eq. (S52)