Letter

Approximate information for efficient exploration-exploitation strategies

Alex Barbier-Chebbah [1,2,*] Christian L. Vestergaard [1,2] and Jean-Baptiste Masson [1,2,†]
[1]*Institut Pasteur, Université Paris Cité, CNRS UMR 3571, Decision and Bayesian Computation, 75015 Paris, France*
[2]*Épimethée, Inria, 75012 Paris, France*

(Received 4 July 2023; accepted 29 January 2024; published 10 May 2024)

This paper addresses the exploration-exploitation dilemma inherent in decision-making, focusing on multi-armed bandit problems. These 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 class of algorithms, approximate information maximization (AIM), which employs a carefully chosen analytical approximation to the gradient of the entropy to choose which arm to pull at each point in time. AIM matches the performance of Thompson sampling, which is known to be asymptotically optimal, as well as that of Infomax from which it derives. AIM thus retains the advantages of Infomax while also offering enhanced computational speed, tractability, and ease of implementation. In particular, we demonstrate how to apply it to a 50-armed bandit game. Its expression is tunable, which allows for specific optimization in various settings, making it possible to surpass the performance of Thompson sampling at short and intermediary times.

DOI: 10.1103/PhysRevE.109.L052105

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 modeling of biological search strategies [2–4] and human decision-making [5–9].

The multiarmed 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], animal behavior [17], neuroscience [18–21], clinical trials [22–24], finance [25], epidemic control [26], and reinforcement learning [27,28], among others. In the multiarmed bandit problem, an agent is presented with a set of possible actions, or "arms," each associated with a probabilistic reward (akin to a multiarmed slot machine 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 best-observed average reward or explore other arms to test if they would not lead to increased rewards.

The information maximization principle has emerged as an effective decision-making strategy and has demonstrated its applicability to different partially known and fluctuating environments [29–31]. Specifically, its original application to target search is termed infotaxis [2], and its application to classical bandit settings was coined Infomax [32]. It has shown empirical state-of-the-art performance.

Here, our goal is to build a class of algorithms based on this information maximization principle with a focus on analytical tractability, computational efficiency, and extendability, all the while demonstrating its robustness for a range of priors.

In the following, we begin with a brief introduction to the bandit problem, followed by a presentation of the information-maximization principle. We then introduce our procedure for approximating the entropy analytically, leading to the approximate information maximization (AIM) strategy. We finally provide empirical evidence of AIM's efficiency before delving into a discussion of its various properties and implications.

Multiarmed bandit problems. We consider the classic multiarmed bandit setting [33]. At each point in time, t, an agent chooses an arm, A_t, between K different arms, $\mathbb{A} = \{1, 2, \ldots, K\}$. 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 [33] $\mathbb{E}[R(t)]$ with

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

where μ^* is the expected reward of the best arm. 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) [34]:

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

where β is a constant factor that depends on the reward distributions. For the two-armed Bernoulli and Gaussian bandit games we consider below, it is given by $\beta = (\mu^* - \mu_2)/D_{\mathrm{KL}}(\mu_2, \mu^*)$, where μ_2 is the expected

*alex.barbier-chebbah@pasteur.fr
†jbmasson@pasteur.fr

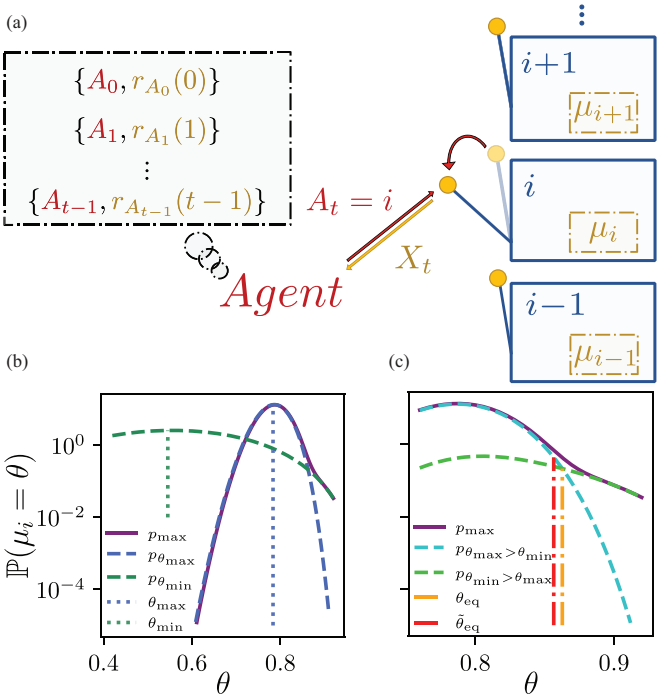

FIG. 1. (a) Illustration of the multiarmed bandit problem. At each time step t the agent chooses an action, $i = A_t$, that returns a reward, $r_i(t)$, drawn from a distribution with unknown mean μ_i. The agent's goal is to minimize the cumulative regret $R(t)$ [see Eq. (1)]. (b) Posterior distributions of bandit values after playing the two-armed Bernoulli game for 201 rounds with $r_1(t) = 5$, $n_1(t) = 9$, $r_2(t) = 41$, and $n_2(t) = 192$, where $r_i(t)$ and $n_i(t)$ are respectively the cumulative reward and the 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 posterior mean rewards of the suboptimal arm (denoted $\theta_{\min}$) and the optimal arm ($\theta_{\max}$). In green: The posterior distribution $p_{\theta_{\min}}$ of the current suboptimal arm. In purple: The posterior distribution $p_{\max}$ of the maximum reward of all arms. Vertical green and blue lines are the posterior mean rewards of the suboptimal arm (denoted $\theta_{\min}$) and the optimal arm ($\theta_{\max}$) (c) Zoomed plot of panel (b) in the region where the posterior distribution of the maximal reward value transitions 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 $p_{\theta_{\max}>\theta_{\min}}$ that the optimal arm's gain is superior to the suboptimal arm. In light green: The probability $p_{\theta_{\min}>\theta_{\max}}$ 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, Secs. S1 B and S2 B [35]).

reward of the worse arm, and $D_{\text{KL}}(\mu_2, \mu^*) = \int \log[p(x;\mu_2)/p(x;\mu^*)]p(x;\mu_2)dx$ is the Kullback-Leibler divergence of the reward distribution of the worse arm, $p(x;\mu_2)$, from that of the better one, $p(x;\mu^*)$.

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 with each arm to decide which arm to play; and Thompson sampling (proportional betting), which relies on sampling the action from the posterior distribution so 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.

Information-maximization principle. Information maximization aims to maximize, in each step, the information gain on a given quantity encapsulating the relevant information about the system [2,32]. We briefly review the fundamentals of the information-maximization strategy specifically adapted to the bandit game, where it is called Infomax [32]. Contrary to classic bandit algorithms, Infomax relies on the entropy to encompass the information carried by all arms in a single functional, thus characterizing the global state of the game. More precisely, an information-maximization strategy that was found to be effective in Ref. [32], termed Info-p, aims to optimize S, the entropy of the posterior distribution of the value of the maximal reward $p_{\max}$,

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

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

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

The entropy S summarizes the information about the state of the game and an information-maximization algorithm greedily optimizes its gradient, i.e., selects the next arm according to

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

By doing so, the algorithm seeks to maximize the expected decrease in entropy, conditioned on the current knowledge of the game. This strategy has been shown empirically to be competitive with state-of-the-art algorithms and to attain the Lai and Robbins bound [32].

Approximate information maximization. While Eq. (5) can be numerically evaluated, it cannot be computed in closed form for most bandit problems. This makes it computationally demanding and makes it difficult to extend the strategy to more complex bandit problems. To obtain an algorithm that is both tractable and computationally efficient, a second functional approximating the entropy thus 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 two-armed bandit. We denote the arms according to their current empirical mean rewards, respectively, the better empirical one by $i_{\max}$ (with expected reward $\theta_{\max}$) and the worse empirical one by $i_{\min}$ (with expected reward $\theta_{\min}$). Note that $\theta_{\max}$ may be smaller than $\theta_{\min}$ 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 main mass of $p_{\max}$

centered around its mode. The second term, $\tilde{S}_{\text{tail}}$, captures the entropy of the tail of $p_{\max}$ [corresponding to high rewards, see Figs. 1(b) and 1(c)]. 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 analytically tractable (see Supplemental Material, S1 [35], for details).

More precisely, the tail term reads as

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

where $\tilde{\theta}_{\text{eq}}$ is an approximation of θ_{eq}, the value of θ where the probability of having the maximum reward is identical for both arms [see red and orange curves in 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:

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

where $p_{\theta_{\max}}(\theta)$ is the posterior probability at θ of the current optimal arm, $p_{\theta_i > \theta_j}(\theta) = \mathbb{P}(\mu_i = \theta, \mu_i \geqslant \mu_j)$ is the posterior probability for the expected reward θ of arm i to be larger than θ_j, and $A_{\text{c}} = 1.258\,89$ is a fixed constant that comes from approximating the logarithm by a rational function (see Eq. (S6) in Supplemental Material, Sec. S1 A [35]). 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, Sec. S1 A [35], for details). Finally, the third, corrective term in Eq. (7) is $c_{\text{tail}} = \int_{\tilde{\theta}_{\text{eq}}}^{\theta_{\text{sup}}} p_{\theta_{\min}}(\theta) d\theta$.

AIM consists in evaluating Eq. (6) for each arm in each time step t and choosing the one that will maximize the decrease of the approximate entropy, similar to Infomax but with the exact entropy replaced by $\tilde{S}(t+1)$. 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, Sec. S2 [35], for a detailed deviation of all terms). To prevent any entrapment scenario due to finite-time-step evaluation, we also replace the gradient minimization of Eq. (6) by the maximization of its absolute value (see Supplemental Material, Sec. S3 [35], for details). Doing so and keeping the dominant terms of $\tilde{S}_{\text{body}}$ and $\tilde{S}_{\text{tail}}$, it leads to a robust principle to provide optimal algorithms independently of the reward distributions.

Numerical performance. We demonstrate the performance of AIM on the paradigmatic Bernoulli bandits [10,36,37] and on Gaussian bandits [38] with the unknown mean $\mu_i \in [0, 1]$ and unit variance. Supplemental Material, Table S1 [35] 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, Secs. S3 and S4 [35], for implementation of AIM and other classic bandit strategies). For both Bernoulli and Gaussian bandits, AIM empirically follows the Lai and Robins bound, with a regret scaling as $\log(t)$. Its long-time performance matches that of exact information maximization (Infomax) and Thompson sampling while relying on a deterministic analytical formula. Additionally, AIM outperforms Thompson sampling at intermediate times for challenging parameter configurations, similar to Infomax [Fig. 2(b)]. These observations highlight the ability of our approximation to accurately capture the significant contribution of both S and Infomax, thus maintaining efficiency at all timescales. Specifically, we emphasize that these terms must provide substantial information to facilitate effective decision-making at short times when both arms are evolving fast and exchanging ranks often.

Asymptotic performance. Empirical evidence indicates that AIM and Infomax both attain the Lai and Robbins bound. For Infomax, this is supported theoretically by long-time scaling arguments derived in Ref. [32]. We apply a similar heuristic development which shows that the dominant contributions of the isolated terms of Eq. (6) correspond to those of Infomax at large times, ensuring that AIM retains the same asymptotic behavior as Infomax, i.e., that the logarithmic slope observed in Fig. 2 will asymptotically follow the Lai and Robbins optimal prefactor.

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 $N_{\min} \sim \ln(t)/D_{\text{KL}}(\mu_{\min}, \mu_{\max})$ for Bernoulli bandits, where $D_{\text{KL}}(\mu_{\min}, \mu_{\max})$ is the Kullback-Leibler divergence between the reward distributions, thus recovering the Lai and Robbins bound (see derivation in Supplemental Material, Sec. S5 [35]). Surprisingly, retaining only the asymptotically dominant terms of $\tilde{S}_{\text{body}}$ and $\tilde{S}_{\text{tail}}$ is enough to obtain a regret that is only slightly higher than that of Eq. (6) and Infomax even at short times (see Supplemental Material, Fig. S1 [35]), underlining their importance in these algorithms.

Note that our 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 behavior of high cumulative regret events (Fig. 3), for which the subdominant arm has been drawn a non-negligible fraction of time. These events happen only for small differences between $\mu_{\max}$ and $\mu_{\min}$, which require exponentially long times to be distinguished (a behavior that is shared with Thompson sampling). Finally, it also confirms that significant gaps between $\mu_{\min}$ and $\mu_{\max}$ consistently lead to small $N_{\min}$, indicating the absence of entrapment scenarios.

Fine-tuning and extension to K-armed bandits. Due to its reliance on a closed-form expression, AIM is easy to extend to more complex bandit games and it is furthermore possible

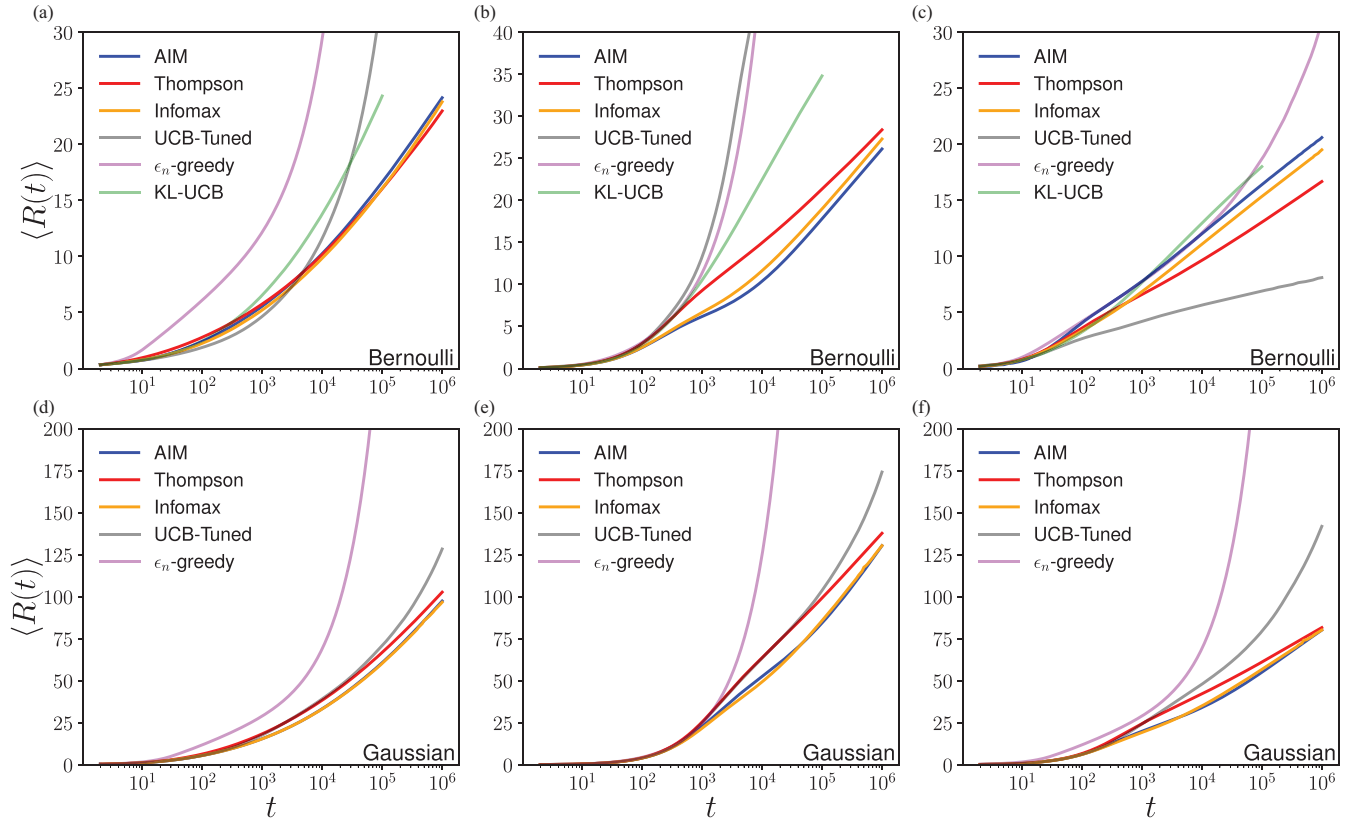

FIG. 2. Temporal evolution of the regret for Bernoulli [panels (a)–(c)] and Gaussian [panels (d)–(f)] two-armed bandits. In blue, AIM; in red, Thompson sampling; in yellow, Infomax; in gray, UCB-tuned; in purple, ϵ_n-greedy; and in light green, KL-UCB. Details of simulations and the tuning required for some algorithms are provided in the Supplemental Material, Secs. S3 and S4 [35]. True parameters were drawn uniformly in]0,1[for both bandits in panels (a) and (d), and parameters were set to $\mu_1 = 0.7$ and $\mu_2 = 0.8$ for panels (b) and (e) and to $\mu_1 = 0.1$ and $\mu_2 = 0.3$ for panels (c) and (f).

to fine-tune its performance to specific settings (see Supplemental Material, S6 [35], for details). As an example, we apply AIM to 50-armed bandits with Bernoulli rewards and 2-armed bandits with Gaussian reward distributions, which are computationally challenging to tackle using Infomax, and where a tuned version of AIM is able to outperform Thompson sampling at short and intermediary times (Fig. 4).

Conclusion. This study presents a new approach, termed approximate information maximization (AIM), designed to efficiently balance exploration and exploitation in multiarmed bandit problems. AIM employs an analytic approximation of the entropy gradient to select the optimal arm. This approach mirrors the performance of Infomax (see Supplemental Material, Sec. S4 [35], and Fig. 2), from which it is derived, while offering improved computational speed (see

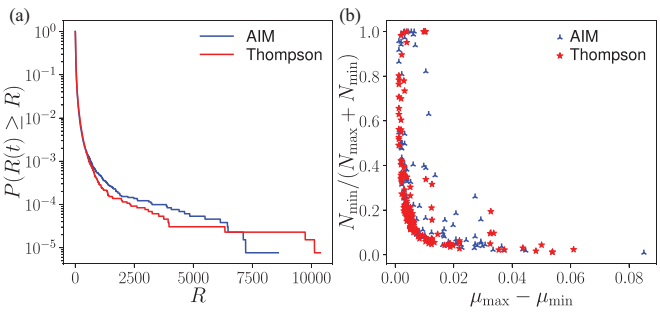

FIG. 3. Regret distribution and rare events. (a) The probability of obtaining a cumulative regret superior to R for both Thompson and AIM playing Bernoulli games with uniform priors and $N = 2^{17}$ realizations). AIM shows a decay similar to that obtained by the Thompson algorithm. (b) Fraction of the subdominant arm, is drawn for high-regret events (0.1%) as function of the mean difference $\mu_{\max} - \mu_{\min}$. In both panels (a) and (b), Thompson and AIM exhibit the same behavior.

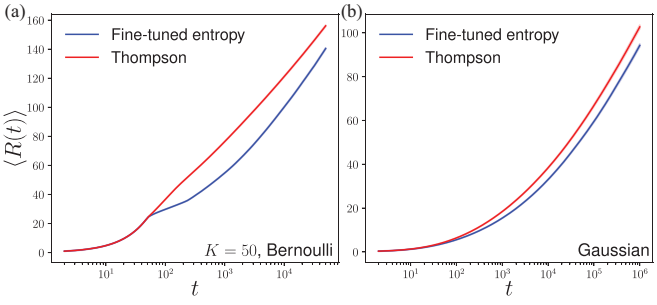

FIG. 4. Mean regret for 50-armed Bernoulli (a) and two-armed Gaussian (b) bandits with parameters drawn uniformly in]0,1[. In red Thompson sampling and in blue tuned AIM (see Supplemental Material, Sec. S6 [35]).

Supplemental Material, Sec. S3 E [35]). 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 optimize performance in various scenarios, while still satisfying the Lai and Robbins bounds.

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

Acknowledgments. We thank E. Boursier for helpful discussions on the optimality of AIM and asymptotic arguments' interpretation from physics and mathematics points of view. We acknowledge the help of the HPC Core Facility of the Institut Pasteur with this work. This study was funded by the INCEPTION project (PIA/ANR-16-CONV-0005), the "*Investissements d'avenir*" program managed by *Agence Nationale de la Recherche*, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), and Agence Nationale de la Recherche ANR-20-CE45-0021.

[1] K. Ding, J. Li, and H. Liu, in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19 (Association for Computing Machinery, New York, 2019), pp. 357–365.
[2] M. Vergassola, E. Villermaux, and B. I. Shraiman, Nature (London) **445**, 406 (2007).
[3] D. Martinez, L. Arhidi, E. Demondion, J.-B. Masson, and P. Lucas, J. Visualized Exp. **4**, e51704 (2014).
[4] R. T. Cardé, Annu. Rev. Entomol. **66**, 317 (2021).
[5] J. D. Cohen, S. M. McClure, and A. J. Yu, Philos. Trans. R. Soc., B **362**, 933 (2007).
[6] T. T. Hills, P. M. Todd, D. Lazer, A. D. Redish, and I. D. Couzin, Trends Cognit. Sci. **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).
[8] K. Doya, *Bayesian Brain: Probabilistic Approaches to Neural Coding* (MIT, Cambridge, MA, 2007).
[9] M. Jepma and S. Nieuwenhuis, J. Cognit. Neurosci. **23**, 1587 (2011).
[10] A. Slivkins, Foundations and Trends® in Machine Learning **12**, 1 (2019).
[11] J. C. Gittins, J. R. Stat. Soc. B **41**, 148 (1979).
[12] L. Zhou, arXiv:1508.03326.
[13] S. Bubeck, R. Munos, and G. Stoltz, Theor. Comput. Sci. **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, United States, 2020), Vol. 33, pp. 1713–1723.
[15] D. Bouneffouf, I. Rish, and C. Aggarwal, in *2020 IEEE Congress on Evolutionary Computation (CEC)* (IEEE, New York, 2020), pp. 1–8.
[16] P. Auer, N. Cesa-Bianchi, and P. Fischer, Mach. Learn. **47**, 235 (2002).
[17] J. Morimoto, J. Theor. Biol. **467**, 48 (2019).
[18] R. C. Wilson, E. Bonawitz, V. D. Costa, and R. B. Ebitz, Curr. Opin. Behav. Sci. **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, Lecture Notes in Computer Science (Springer, Cham, 2017), pp. 237–248.
[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), Vol. 85, pp. 67–82.
[23] S. S. Villar, J. Bowden, and J. Wason, Stat. Sci. **30**, 199 (2015).
[24] S. S. Villar, Probab. Eng. Inf. Sci. **32**, 229 (2018).
[25] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha, in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence* (The International Joint Conference on Artificial Intelligence, United States, 2015).
[26] B. Lin and D. Bouneffouf, in *2022 Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (IEEE, New York, 2022), pp. 1–8.
[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 (London) **529**, 484 (2016).
[28] I. O. Ryzhov, W. B. Powell, and P. I. Frazier, Oper. Res. **60**, 180 (2012).
[29] G. Reddy, V. N. Murthy, and M. Vergassola, Annu. Rev. Condens. Matter Phys. **13**, 191 (2022).
[30] S. Zhang, D. Martinez, and J.-B. Masson, Front. Rob. AI **2** (2015).
[31] J.-B. Masson, Proc. Natl. Acad. Sci. USA **110**, 11261 (2013).
[32] G. Reddy, A. Celani, and M. Vergassola, J. Stat. Phys. **163**, 1454 (2016).
[33] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning Series (MIT Press, Cambridge, MA, 1998).
[34] T. L. Lai and H. Robbins, Adv. Appl. Math. **6**, 4 (1985).
[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.109.L052105 for calculations details and algorithm implementations supporting the main text, which includes Refs. [2,32,33,36,37,39,40,41].
[36] S. Pilarski, S. Pilarski, and D. Varró, IEEE Trans. Artif. Intell. **2**, 2 (2021).

[37] W. R. Thompson, Biometrika **25**, 285 (1933).
[38] 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, Catonsville, Maryland, 2010), pp. 67–79.
[39] A. Garivier and O. Cappé, Computing Research Repository - CORR, 2011.
[40] E. Kaufmann, N. Korda, and R. Munos, in *Algorithmic Learning Theory*, edited by N. H. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann, Lecture Notes in Computer Science (Springer, Berlin, 2012), pp. 199–213.
[41] E. W. Ng and M. Geller, J. Res. Natl. Bur. Stand., Sect. B. **73B**, 1 (1969).