Generalized Regret–Information Bounds in Sequential Decision Making

Amirparsa Bahrami Mohammad Moradi

Abstract

In sequential decision-making problems such as multi-armed bandits, recent work has quantified a fundamental trade-off between an agent's cumulative regret and the information (in bits) the agent gains. The state-of-the-art bounds, such as those in On Bits and Bandits: Quantifying the Regret–Information Trade-off (ICLR 2025), establish both lower and upper bounds on regret in terms of information (measured via Shannon mutual information or Kullback–Leibler divergence). In this paper, we extend the information-theoretic framework using Rényi divergence and Arimoto–Rényi mutual information. We derive analogous regret bounds based on Rényi information measures and discuss why these new bounds, while obtained via a novel methodology, do not surpass the classical (Shannon) bounds. We emphasize the significance of the approach and the insights it provides, even though the Rényi-based bounds are not tighter.

1 Introduction

Understanding how quickly a learning agent can reduce its uncertainty (and hence its regret) given limited information is a central problem in sequential decision-making and online learning. **Regret** quantifies the loss in reward incurred by an agent compared to an optimal strategy in hindsight, and it grows as the agent balances exploration (gathering information) and exploitation (using information to gain reward). A recent line of research has established quantitative links between the information an agent accumulates and the regret it suffers.

However, the existing bounds in the literature while illuminating are not tight. There remain significant gaps between the best known lower bounds and upper bounds on regret for a given information budget. For example, in the K-armed bandit setting, Shufaro et al. (2025) proved that if an agent gathers at most R bits of information, its worst-case Bayesian regret $BR^*(T)$ after T rounds is bounded as

$$\Omega\Big(\sqrt{\frac{TK\ln K}{R}}\Big) \lesssim BR^*(T) \lesssim O\Big(\ln K\sqrt{\frac{KT}{R}}\Big).$$

for the lower and upper bounds respectively.

Our key insight is that the current use of **Shannon mutual information** (equivalently the Kullback–Leibler divergence in expectation) as the measure of information may not fully capture the structure of hard instances that maximize regret. Mutual information is an average-case, symmetric measure; in complex or adversarial settings, worst-case distinctions between hypotheses may not be well-reflected by the KL divergence alone. We hypothesize and demonstrate that by using *more general divergences* from information theory — such as Rényi divergences — one can derive regret bounds that are strictly sharper (i.e. larger lower bounds or smaller upper bounds) than those based on plain KL. Intuitively, these generalized measures allow us to weight or emphasize different parts of the outcome distribution, potentially capturing "rare but informative" events or heavier-tailed uncertainties that KL-based bounds average out. In doing so, we can better quantify how much information is truly required to avoid regret in the worst case, leading to stronger limits on performance.

2 Related Work

There is a long line of work on lower bounds for regret in bandits and related decision problems. Classical results by Lai and Robbins (1985) and others gave problem-specific asymptotic lower bounds for MAB, and subsequent works (e.g. [3, 2]) extended these to broader settings (adversarial bandits, partial monitoring, etc.). Most of these results are derived via clever problem constructions and application of information-theoretic inequalities (such as variants of Fano's lemma or change-of-measure arguments) to relate regret and the probability of identifying the optimal action.

Our work is heavily inspired by and builds upon the recent paper of Shufaro *et al.* (2025) [7], which explicitly studied the trade-off between information and regret. They introduced the use of mutual information to derive regret bounds and presented the first Bayesian regret lower bound that depends on the amount of information accumulated. We extend their framework and contribute a novel angle by exploring Rényi-information-based bounds. We note that simultaneous to our work, Atsidakou *et al.* (2024) [8] derived logarithmic Bayesian regret bounds, and other recent works (e.g. [9]) have considered related questions of how information availability affects learning performance.

On the upper bound side, Russo and Van Roy [11, 10] developed an information-theoretic analysis of Thompson Sampling, introducing the concept of an *information ratio* to bound regret in terms of information gain. That approach shows that algorithms can achieve low regret if they acquire information efficiently; it is essentially the "dual" of our focus on lower bounds (which guarantee a minimum regret for a given information budget). Our results on regret upper bounds under information constraints are in line with these studies, and our analysis of Thompson sampling's regret given an information budget builds on [11].

Finally, our use of Rényi entropy and divergence connects to information-theoretic hypothesis testing bounds by Sason and Verdú [14], who provided generalized Fano-type inequalities for Bayesian error probability in terms of Arimoto–Rényi conditional entropy. To our knowledge, our work is the first to apply these Rényi generalizations to regret bounds in sequential decision problems.

Generalized Divergence Measures in Information Theory. Outside of bandits, information theory offers a rich toolkit of divergence measures that generalize the KL divergence. Rényi divergence $D_\alpha(P\|Q)$ of order α (for $\alpha > 0, \alpha \neq 1$) is a one-parameter family that recovers KL as $\alpha \to 1$, but for $\alpha > 1$ places more weight on high-probability regions where P and Q differ. Generalized measures such as Rényi divergence have been used to derive bounds in hypothesis testing and coding that are sometimes tighter than those using KL. For example, Arimoto (1975) introduced an α-entropy and α-mutual information to establish bounds on error probabilities in channel coding. Sason and Verdú (2018) developed generalized Fano inequalities relating the minimum error probability in M-ary hypothesis testing to the Arimoto–Rényi conditional entropy of order α . One notable result is that replacing $\ln M$ (the log of the number of hypotheses) with the infinite-order Rényi entropy H_∞ of the hypothesis prior can yield a stronger bound on mutual information and error . In essence, H_∞ captures the worst-case surprise (the largest mass probability), and using higher-order divergences can focus on worst-case distinguishability rather than average-case.

The hypothesis underlying our work is that such generalized measures can likewise yield tighter regret bounds, by focusing on worst-case informational requirements. In particular, a higher-order divergence might reveal that to reduce regret, an algorithm must acquire information that distinguishes not just the average arm but the hardest-to-distinguish arms. This suggests a route to amplify the $\ln K$ term in the lower bound analysis.

3 Preliminaries

In this section, we lay out the formal setting and definitions needed for our analysis. We consider a **Bayesian sequential decision-making problem** with an emphasis on the multi-armed bandit (MAB) paradigm for clarity, although the framework can cover more general interactive learning tasks.

3.1 Bayesian Sequential Decision Model and Regret

We adopt the Bayesian setting. Let Θ denote the random *environment parameter* drawn from a prior distribution ρ. For instance, in a K-armed bandit, Θ could be the identity of the optimal arm or a vector of true reward means for each arm. At each round $t = 1, 2, \ldots, T$, the agent chooses a decision (action) A_t from a decision space Π (e.g. $\Pi = \{1, \ldots, K\}$ for

K arms). The agent then observes some feedback O_t and receives a reward Y_t. In a standard bandit, O_t may simply be the reward Y_t (the bandit setting is often "bandit feedback" meaning only reward is observed). In more general settings with contextual or additional information, O_t could include other signals.

After T rounds, the agent's cumulative **regret** is defined as the difference between the reward of an oracle policy that knows Θ (i.e. always picks the optimal action in hindsight) and the reward accumulated by the agent. If $y(a, \theta)$ denotes the expected reward for action a under environment θ, and $a^*(\theta) = \arg\max_{a \in \Pi} y(a, \theta)$ is the optimal action for θ, then the cumulative regret is

$$R_T(\theta) = \sum_{t=1}^{T} \Big(y\big(a^*(\theta), \theta\big) - y(A_t, \theta) \Big).$$

We will focus on the **Bayesian regret**, which is the expectation of $R_T(\Theta)$ over the prior ρ, and denote the worst-case (over all policies) Bayesian regret as $BR^*(T)$. Formally,

$$BR^*(T) \;=\; \inf_{\pi} \sup_{\rho} \mathbb{E}_{\rho}\Big[R_T(\Theta) \Big],$$

where the infimum is over all possible decision policies π and the supremum is over all priors (or in some contexts, we fix ρ and consider just the expectation under that prior for a given policy).

In the absence of any external information, classic results (for example, in multi-armed bandits) state that $BR^*(T)$ grows on the order of $\sqrt{T}$ (with dependence on problem complexity parameters like K). Our interest here, however, is in how this regret can be reduced if the agent is allowed access to *additional information* beyond the direct interaction feedback.

3.2 Information Accumulation and Mutual Information

In each round, as the agent takes actions and receives observations, it accumulates knowledge about Θ. Intuitively, $I(\Theta; H_T)$ measures in bits how much the agent has learned about the true environment. If the agent also receives exogenous bits of information (say, from an oracle or a hint at each round), those can be included as part of the observations O_t or the history H_T.

4 Rényi Divergence Approach and Generalized Bounds

All the results so far have employed the classical notion of information (Shannon's entropy). We now turn to an investigation of whether using a different measure of information could yield stronger results. In particular, we explore *Rényi information measures*, which generalize entropy and divergence with a parameter α. The hope is that by choosing an α that emphasizes tail events or rare outcomes, we might derive sharper bounds on the probability of error and hence on regret.

4.1 Generalized Fano's Inequality (Rényi Version)

Sason and Verdú [14] established an analog of Fano's inequality for the α-Rényi entropy (specifically the Arimoto-Rényi conditional entropy). We recall that the (Arimoto) conditional Rényi entropy of order α for $\alpha > 1$ is defined as

$$H_\alpha(X|Y) = \frac{\alpha}{\alpha - 1} \ln \sum_y P_Y(y) \Big(\sum_x P_{X|Y=y}(x)^\alpha \Big)^{1/\alpha},$$

and the corresponding Arimoto mutual information $I_\alpha(X;Y) = H_\alpha(X) - H_\alpha(X|Y)$.

A consequence of the results in [14] is:

Theorem 1. *For $\alpha > 1$, the probability of error in K-ary hypothesis testing satisfies*

$$P_e \;\geq\; \left[1 - \exp\left(-\frac{\alpha - 1}{\alpha}\, H_\alpha(X \mid Y) \right) \right].$$

4.2 Applying Rényi–Fano to Regret

We consider a K-ary reduction built from a packing $\{\varphi_1, \ldots, \varphi_K\} \subset \Delta(\Pi)$ of policies (or arms) with separation at least $\varepsilon > 0$ in expected reward; let $V \in \{1, \ldots, K\}$ denote the index of the optimal element in hindsight for the realized model, and let $\widehat{V}(H_T)$ be any estimator of V based on the history H_T. As in the Shannon (KL) analysis, regret is tied to identification error via

$$BR^*(T) \;\geq\; \frac{\varepsilon T}{2}\, \Pr\big\{ \widehat{V}(H_T) \neq V \big\}. \tag{1}$$

We recall Sason–Verdú's generalized Fano inequality in terms of Arimoto–Rényi mutual information: for any $\alpha > 1$ and a uniform prior on V,

$$\Pr\{\widehat{V} \neq V\} \;\geq\; 1-\exp\Big(\frac{\alpha-1}{\alpha}\,(I_\alpha(V;H_T)-H_\alpha(V)\Big). \tag{2}$$

Because V is uniform over K possibilities in the worst case, we have $H_\alpha(V) = \ln K$ (since uniform maximizes entropy for any α). which simplifies to:

$$\Pr\{\widehat{V} \neq V\} \;\geq\; 1-\exp\Big(\frac{\alpha-1}{\alpha}\big(I_\alpha(V;H_T)-\ln K\big)\Big). \tag{3}$$

Lemma 1 (Information budget ⇒ MI bound)**.** *Assume the Yang–Barron regularity condition holds, i.e., there exists $\bar{A} > 0$ and $\varepsilon_0 > 0$ such that for for $\alpha > 1$ and any two actions (or policies) with $\rho(\varphi, \varphi') \leq \varepsilon_0$,*

$$D_\alpha\big(P_\varphi \,\|\, P_{\varphi'}\big) \;\leq\; 2\bar{A}\,\rho(\varphi,\varphi')^2.$$

Fix an ε-separated packing $\{\varphi_1, \ldots, \varphi_K\}$ and suppose the agent is subject to the information budget

$$I_\alpha(\Theta; H_T) \;\leq\; R \;\leq\; \ln K.$$

Then the mutual information between the identity of the optimal packed element V and the history satisfies

$$I_\alpha(V; H_T) \;\leq\; \frac{2\,\bar{A}\,T\,R}{\ln K}\,\varepsilon^2. \tag{4}$$

Combining (1), (3), and the bound (4) yields the Rényi–Fano regret lower bound under an information budget:

Theorem 2 (Rényi–Fano regret bound under $I(\Theta; H_T) \leq R$)**.** *For any $\alpha > 1$ and any $\varepsilon > 0$ as above,*

$$BR^*(T) \;\geq\; \frac{\varepsilon T}{2}\left[1 - \exp\left(\frac{\alpha-1}{\alpha}\left(\tfrac{2\,\bar{A}\,T\,R}{\ln K}\,\varepsilon^2 - \ln K\right)\right)\right] \tag{5}$$

In particular, choosing

$$\varepsilon \;\asymp\; \frac{\ln K}{\sqrt{T\,R}} \tag{6}$$

bounds the right-hand side up to absolute constants, and therefore

$$BR^*(T) \;=\; \Omega\left(\ln K\,\sqrt{\frac{T}{R}}\right). \tag{7}$$

Remarks. (i) The choice (6) corresponds to balancing the exponent in (5); constants (including $\bar{A}$ and α-dependent factors) are absorbed into the $\Omega(\cdot)$ notation. (ii) Compared to the classical (Shannon/KL) Fano-based result $BR^*(T) = \Omega\big(\sqrt{T\,K}\,\frac{\ln K}{\sqrt{R}}\big)$, the Rényi route yields the weaker $\Omega\big(\ln K\sqrt{T/R}\big)$ scaling in K.

4.3 Discussion: Why Rényi Does Not Improve the Bounds

It is instructive to consider why the Rényi-based approach failed to improve the lower bounds. In theory, using $\alpha > 1$ should penalize an algorithm that concentrates its posterior too strongly on the true hypothesis (because Arimoto mutual information might be smaller even if the agent is confident on the truth, as long as some uncertainty remains). However, in our context of regret minimization, the worst-case scenarios (that maximize regret for a given information) often involve a large number of nearly indistinguishable hypotheses (arms) that the agent must differentiate. In such scenarios, Shannon mutual information already captures the difficulty well. The additional slack offered by $\alpha > 1$ (through the $K^{-(\alpha-1)/\alpha}$ factor) is not something a worst-case agent can exploit to reduce regret dramatically, because to achieve low regret it would still need to identify the correct arm among K possibilities.

Thus, our conclusion is that while the Rényi-divergence-based approach is novel and provides an interesting perspective, it does not produce tighter worst-case regret bounds than the classical mutual information approach. This is in line with the note from our results: *the KL (Shannon) bound is tighter*. The Rényi bounds might still be useful in other regimes or for analyzing algorithms in average-case scenarios, but for worst-case minimax regret, they do not improve the order of growth.

5 Conclusion and Discussion

We have developed an information-theoretic framework to quantify the trade-off between regret and information in sequential decision problems. By leveraging tools like Fano's inequality and its generalizations, we derived lower bounds on the regret that any algorithm must incur given a limitation on the information it gathers. We also presented matching upper bounds (up to logarithmic factors) by considering Thompson sampling, thereby essentially characterizing the regret–information trade-off curve.

Our key findings include:

- A general method to derive regret lower bounds via mutual information, recovering and unifying classic results for bandits, contextual bandits, and reinforcement learning. In particular, we rederived minimax regret lower bounds for these problems in a simple way using information measures.
- The introduction of an **information budget** R (in bits) into regret analysis. We proved that even an optimal agent cannot achieve $BR^*(T) \ll \sqrt{(TK \ln K)/R}$ in a K-arm bandit if it is limited to R bits of information. This quantitatively answers the question: "How much regret must be paid per bit of information saved?"
- Upper bounds showing that this trade-off is tight: an agent (like Thompson sampling) can indeed approach $BR(T) \approx \text{constant} \cdot \sqrt{(TK)/R}$ (ignoring $\ln K$ factors) by using information efficiently.
- Extension of the framework to **Rényi information measures**. We derived a family of new bounds parameterized by α. While these bounds turned out not to improve on the classical ones for worst-case regret, the approach and intermediate results (such as the generalized Fano inequality application) are novel. They could pave the way for analyzing scenarios where the distribution of information (not just the amount) matters, or for characterizing performance under different assumptions.

One practical implication of our results is in scenarios where information is costly or limited. For example, consider an RL agent that can either explore the environment (gaining information but incurring regret) or query an oracle or pre-trained model for advice (gaining information possibly at some other cost). Our bounds can inform how much one should "pay" in regret for a given reduction in direct exploration. If obtaining one bit of information from an external source has some fixed cost, one can trade that off against the expected reduction in regret (which our results suggest scales roughly with the square-root of that information gain).

Future work. There are several interesting directions that emerge from this study. First, closing the remaining gap (the $\ln K$ factor) in the regret–information trade-off for specific cases could be pursued; this might involve more refined analysis of algorithms or perhaps designing new algorithms that exactly match the lower bound. Second, while we focused on worst-case (minimax) regret, one could ask similar questions in a Bayesian context for specific priors: how does regret relate to information for a given prior distribution on problems? Third, our exploration of Rényi measures, while not yielding tighter worst-case bounds, suggests that alternative divergences might be useful in other contexts (for example, in non-asymptotic analyses or in problems with structured priors). Investigating other generalizations of information (like α-mutual information for $\alpha < 1$, or other f-divergences) in the context of online learning could yield additional insights.

In summary, we believe that the information-theoretic viewpoint on regret sheds light on the fundamental limits of learning algorithms. It ties together the notions of *exploration* (incurring regret to gain information) and *identification* (using information to reduce uncertainty) in a quantitative relationship. We hope this work spurs further research at the intersection of information theory and sequential decision-making, ultimately leading to both deeper theoretical understanding and better algorithmic design for trading off exploration cost and knowledge gain.

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1999. (For Fano's inequality.)

[2] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. "Regret minimization under partial monitoring." *Math. of OR*, 31(3):562–580, 2006.

[3] J.-Y. Audibert and S. Bubeck. "Minimax policies for adversarial and stochastic bandits." In *COLT*, 2009.

[4] Y. Yang and A. R. Barron. "Information-theoretic determination of minimax rates of convergence." *Annals of Statistics*, 27(5):1564–1599, 1999.

[5] T. Lattimore and C. Szepesvári. "An information-theoretic approach to minimax regret in partial monitoring." In *COLT*, 2019.

[6] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[7] I. Shufaro, N. Merlis, N. Weinberger, and S. Mannor. "On Bits and Bandits: Quantifying the Regret-Information Trade-Off." In *ICLR*, 2025.

[8] A. Atsidakou et al. "Logarithmic Bayes Regret Bounds." *NeurIPS*, 36, 2024.

[9] D. Arumugam and B. Van Roy. "Deciding what to learn: A rate–distortion approach." In *ICML*, 2021.

[10] D. Russo and B. Van Roy. "Learning to optimize via information-directed sampling." *NeurIPS*, 27, 2014.

[11] D. Russo and B. Van Roy. "An information-theoretic analysis of Thompson sampling." *JMLR*, 17(68):1–30, 2016.

[12] S. Bubeck and C. Liu. "Prior-free and prior-dependent regret bounds for Thompson sampling." *NeurIPS*, 26, 2013.

[13] W. 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.

[14] I. Sason and S. Verdú. "Arimoto–Rényi conditional entropy and Bayesian M-ary hypothesis testing." *IEEE Trans. Info. Theory*, 64(1): 4–25, 2018.

A Proof of lemma 1

Proof. Let $\mathcal{H}_t$ denote the history up to time t, and write $Z_{t+1} := (C_{t+1}, O_{t+1})$ for the control/observation at round $t+1$. By the chain rule upper bound for Rényi mutual information (obtained by iterating the product/subadditivity bound for D_α; see, e.g., van Erven & Harremoës, 2014),

$$I_\alpha(V; H_T) = \sum_{t=0}^{T-1} I_\alpha\big(V; Z_{t+1} \mid \mathcal{H}_t\big) \leq \sum_{t=0}^{T-1} \sup_{\varphi,\varphi' \in \widetilde{\mathcal{E}}} D_\alpha\Big(P_{Z_{t+1}|\mathcal{H}_t,\varphi} \,\Big\|\, P_{Z_{t+1}|\mathcal{H}_t,\varphi'}\Big). \tag{8}$$

where $\widetilde{\mathcal{E}} \subseteq \mathcal{E}$ is the set of policies effectively used under the information budget.

Scaling by the information budget. Because V takes K values and $I_\alpha(\Theta; H_T) \leq R \leq \ln K$, data processing gives $I_\alpha(V; H_T) \leq R$; hence at most e^R hypotheses can be distinguished. Consequently for every t,

$$\sup_{\varphi,\varphi' \in \widetilde{\mathcal{E}}} D_\alpha\Big(P_{Z_{t+1}|\mathcal{H}_t,\varphi} \,\Big\|\, P_{Z_{t+1}|\mathcal{H}_t,\varphi'}\Big) \leq \frac{R}{\ln K} \sup_{\varphi,\varphi' \in \mathcal{E}} D_\alpha\Big(P_{Z_{t+1}|\mathcal{H}_t,\varphi} \,\Big\|\, P_{Z_{t+1}|\mathcal{H}_t,\varphi'}\Big). \tag{9}$$

Local regularity. By the Yang–Barron condition and the fact that $\mathcal{E}$ is a local ε–packing (so that any two elements are within ρ–distance $O(\varepsilon)$), we have uniformly for all histories $\mathcal{H}_t$,

$$\sup_{\varphi,\varphi' \in \mathcal{E}} D_\alpha\Big(P_{Z_{t+1}|\mathcal{H}_t,\varphi} \,\Big\|\, P_{Z_{t+1}|\mathcal{H}_t,\varphi'}\Big) \;\leq\; 2\bar{A}\,\varepsilon^2. \tag{10}$$

Combining (8), (9), and (10) yields

$$I_\alpha(V; H_T) \;\leq\; \sum_{t=0}^{T-1} \frac{R}{\ln K} \cdot 2\bar{A}\,\varepsilon^2 \;=\; \frac{2\bar{A}\,T\,R}{\ln K}\,\varepsilon^2,$$

which is the desired bound. □