On the Performance of Thompson Sampling on Logistic Bandits

Shi Dong SDONG15@STANFORD.EDU
Tengyu Ma TENGYUMA@STANFORD.EDU
Benjamin Van Roy BVR@STANFORD.EDU
Stanford University

Editors: Alina Beygelzimer and Daniel Hsu

[1] Abstract

We study the logistic bandit, in which rewards are binary with success probability $\exp(\beta a^\top \theta)/(1+\exp(\beta a^\top \theta))$ and actions a and coefficients θ are within the d-dimensional unit ball. While prior regret bounds for algorithms that address the logistic bandit exhibit exponential dependence on the slope parameter β, we establish a regret bound for Thompson sampling that is independent of β. Specifically, we establish that, when the set of feasible actions is identical to the set of possible coefficient vectors, the Bayesian regret of Thompson sampling is $\tilde{O}(d\sqrt{T})$. We also establish a $\tilde{O}(\sqrt{d\eta T}/\lambda)$ bound that applies more broadly, where λ is the worst-case optimal log-odds[2] and η is the "fragility dimension," a new statistic we define to capture the degree to which an optimal action for one model fails to satisfice for others. We demonstrate that the fragility dimension plays an essential role by showing that, for any $\epsilon > 0$, no algorithm can achieve $\mathrm{poly}(d, 1/\lambda) \cdot T^{1-\epsilon}$ regret.

Keywords: bandits, Thompson sampling, logistic regression, regret bounds.

1. Introduction

In the *logistic bandit* an agent observes a binary reward after each action, with outcome probabilities governed by a logistic function:

$$\mathbb{P}\left(\text{reward} = 1 \middle| \text{action} = a\right) = \frac{e^{\beta a^\top \theta}}{1 + e^{\beta a^\top \theta}}.$$

Each action a and parameter vector θ is a vector within the d-dimensional unit ball. The agent initially knows the scale parameter β but is uncertain about the coefficient vector θ. The problem of learning to improve action selection over repeated interactions is sometimes referred to as the *logistic bandit problem* or *online logistic regression*.

The logistic bandit serves as a model for a wide range of applications. One example is the problem of personalized recommendation, in which a service provider successively recommends content, receiving only binary responses from users, indicating "like" or "dislike." A growing literature treats the design and analysis of action selection algorithms for the logistic bandit. Upper-confidence-bound (UCB) algorithms have been analyzed in Filippi et al. (2010); Li et al. (2017); Russo and Van Roy (2013), while Thompson sampling (Thompson (1933)) was treated in Russo

[1] Extended abstract. Full version is available on arXiv.

[2] In defining "log-odds," we use base e^β rather than e. As a result, the term "log-odds" throughout this article refers to $a^\top \theta$ instead of $\beta a^\top \theta$.

Algorithm	Regret Upper Bound	Notes
GLM-UCB (Filippi et al. (2010))	$O\left(e^{\beta} \cdot d \cdot T^{1/2} \log^{3/2} T\right)$	Frequentist bound.
A variation of GLM-UCB (Russo and Van Roy (2013))	$O\left(e^{\beta} \log \beta \cdot d \cdot T^{1/2}\right)$	Bayesian bound.
SupCB-GLM (Li et al. (2017))	$O\left(e^{\beta} \cdot (d \log K)^{1/2} \cdot T^{1/2} \log T\right)$	Frequentist bound, K is the number of actions.
Thompson Sampling (Russo and Van Roy (2014b))	$O\left(e^{\beta} \cdot d \cdot T^{1/2} \log^{3/2} T\right)$	Bayesian bound.
Thompson Sampling (Abeille and Lazaric (2017))	$O\left(e^{\beta} \cdot d^{3/2} \log^{1/2} d \cdot T^{1/2} \log^{3/2} T\right)$	Frequentist bound.
Thompson Sampling (this work)	$O\left(\lambda^{-1} \cdot \left(d(\eta \vee d)\right)^{1/2} \cdot T^{1/2} \log^{1/2} T\right)$	Bayesian bound, λ and η are independent of β.

Table 1: Comparison of various results on logistic bandits. The upper bound in this work depends on β-independent parameters λ and η. Readers are referred to the full version of this paper for detailed definitions of the parameters. We use the notation $a \vee b = \max\{a, b\}$.

and Van Roy (2014b) and Abeille and Lazaric (2017). Each of these algorithms has been shown to converge on the optimal action with time dependence $\tilde{O}(1/\sqrt{T})$, where $\tilde{O}$ ignores poly-logarithmic factors. However, previous analyses leave open the possibility that the convergence time increases exponentially with the parameter β, which seems counterintuitive. In particular, as β increases, distinctions between good and bad actions become more definitive, which should make them easier to learn.

To shed light on this issue, we build on an information-theoretic line of analysis, which was first proposed in Russo and Van Roy (2016) and further developed in Bubeck and Eldan (2016) and Dong and Van Roy (2018). A critical device here is the *information ratio*, which quantifies the one-stage trade-off between exploration and exploitation. The information ratio has also motivated the design of efficient bandit algorithms, as in Russo and Van Roy (2014a), Russo and Van Roy (2018) and Liu et al. (2018). While prior bounds on the information ratio pertain only to independent or linear bandits, in this work we develop a new technique for bounding the information ratio of a logistic bandit. This leads to a stronger regret bound and insight into the role of β.

Our Contributions. Let $\mathcal{A}$ and Θ be the set of feasible actions and the support of θ, respectively. Under an assumption that $\mathcal{A} = \Theta$, we establish a $\tilde{O}(d\sqrt{T})$ bound on Bayesian regret. This bound scales with the dimension d, but notably exhibits no dependence on β or the number of feasible actions. We then generalize this bound, relaxing the assumption that $\mathcal{A} = \Theta$ while introducing dependence on two statistics of the these sets: the *worst-case optimal log-odds* $\lambda = \min_{\theta \in \Theta} \max_{a \in \mathcal{A}} \alpha^\top \theta$ and the *fragility dimension* η, which is the number of possible models such that the optimal action for each yields success probability no greater than 50% for any other. Assuming $\lambda > 0$, we establish a $\tilde{O}(\sqrt{d\eta T}/\lambda)$ bound on Bayesian regret. We also demonstrate that the fragility dimension plays an essential role, as for any function f, polynomial p, and $\epsilon > 0$, any algorithm for the logistic bandit cannot achieve Bayesian regret uniformly bounded by $f(\lambda)p(d)T^{1-\epsilon}$. We believe that, although η can grow exponentially with d, in most relevant contexts η should scale at most linearly with d.

The assumption that the worst-case optimal log-odds are positive may be restrictive. This is equivalent to assuming that the for each possible model, the optimal action yields more than 50% probability of success. However, this assumption is essential, since it ensures that the fragility dimension is well-defined. When the worst-case optimal log-odds are negative, the geometry of action and parameter sets plays a less significant role than parameter β, therefore we conjecture that the exponential dependence on β is inevitable. This could be an interesting direction for future research.

References

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Sébastien Bubeck and Ronen Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589, 2016.

Shi Dong and Benjamin Van Roy. An information-theoretic analysis for Thompson sampling with many actions. In *Advances in Neural Information Processing Systems*, 2018.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080, 2017.

Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014a.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014b.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *arXiv preprint arXiv:1803.02855*, 2018.

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