Lifting the Information Ratio: An Information-Theoretic Analysis of Thompson Sampling for Contextual Bandits

Presenter: Xuanfei Ren*

January 31, 2023

1 Abstract

The paper adapt the information-theoretic perspective of [2] to the contextual setting by introducing a new concept of information ratio based on the mutual information between the unknown model parameter and the observed loss. And the main goal is to bound the regret in terms of the entropy of the prior distribution through a remarkably simple proof, and with no structural assumptions on the likelihood or the prior. After proving the general results, it is mentioned in this paper that several specific binary loss bandits and linear gaussian loss bandit have good regret bounds.

2 Questions to ask

Question 1. *What is the difference between the information ratio based TS and the original TS algorithm?*

Actually, information ratio just consider a new way to analysis the regret of TS algorithm, the algorithm is the same. Using information ratio can bound the regret in terms of the entropy of the prior distribution and the upper bound of so-called information ratio.

Question 2. *What is the benefit of using this new notion?*

I think it provides a new framework for algorithmic regret analysis. In the future if we want to prove a regret bound for TS, we can first think if we can find upper bound for this information ratio and maybe the entropy of the prior distribution.

In the main section, I use red color to denote some things we may improve this method. Just my personal thinking, due to limited knowledge may not be the right idea

3 Preliminaries

3.1 Notation

We consider a parametric class of contextual bandits with parameters space Θ, context space $\mathcal{X}$, and K actions. To each parameter $\theta \in \Theta$ there corresponds a contextual bandit with loss distribution $P_{\theta,x,a}$ for each context $x \in \mathcal{X}$ and action $a \in \mathcal{A}$, with the mean loss of the distribution denoted by $l(\theta, x, a)$.

We study the problem of regret minimization in the Bayesian setting. In this setting, the environment secretly samples a parameter θ^* from a known prior distribution Q_1 over Θ. We assume that the agent has full knowledge of the prior and the likelihood model $P_{\theta,x,a}$. The goal of the agent is to minimize the expected sum of losses. In the Bayesian setting, this is equivalent to minimizing the Bayesian regret, defined as follows:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(l(\theta^*, X_t, A_t) - l(\theta^*, X_t, A_t^*))\right],$$

*University of Science and Technology of China; email: xuanfeiren@mail.ustc.edu.cn or xuanfeir@gmail.com

where A_t^* is the optiminal action for round t.

Furthermore, let $\mathcal{F}_t = \sigma(X_1, A_1, L_1, \ldots, X_t, A_t, L_t)$. We use Q_t to denote the distribution of the unknown parameter θ^* conditional on the past history $\mathcal{F}_{t-1}$. We denoted by $\pi(\cdot|X_t)$ the distribution over the agent's actions conditional on X_t and $\mathcal{F}_{t-1}$, and call it agent's policy. Finally, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}, X_t]$ and $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_{t-1}, X_t]$.

In Thompson Sampling, an important fact is $\mathbb{P}_t[A_t = a] = \mathbb{P}_t[A_t^* = a]$ and $(\theta, A_t) \overset{\mathrm{d}}{=} (\theta^*, A_t^*)$.

3.2 Basic information theory concepts

The Shannon entropy of X is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x).$$

Fact 1. $0 \leq H(X) \leq \log(|\mathcal{X}|)$.

For two probability measures P and Q, if P is absolutely continuous with respect to Q, the *Kullback-Leibler divergence* between them is

$$D(P||Q) = \int \log \frac{dP}{dQ} dP = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Fact 2. *$D(P||Q) \geq 0$ with equality if and only if $P = Q$ P-a.s.*

Mutual information is defined by

$$I(X;Y) = D(P(X,Y)||P(X)P(Y)) = \mathbb{E}_X[D(P(Y|X)||P(Y))].$$

Fact 3 (Data processing inequality)**.** *$I(X;Y) \geq I(X; g(Y))$.*

3.3 Information ratio

Definition 3.1 (Information ratio)**.** *Informally, the information ratio measures the tradeoff between achieving low regret and gaining information about the identity of the optimal action A^* (which is a deterministic function of θ^* in the standard multi-armed bandit setting). The formal definition is given by*

$$\rho_t^* = \frac{(\mathbb{E}_t[l(\theta^*, A_t) - l(\theta^*, A^*)])^2}{I_t(A^*; (A_t, L_t))}.$$

But in contextual bandit, the optimal action A_t^* changes from round to round, influenced by the context X_t, so information about A^* is useless.

Definition 3.2 (Lifted information ratio)**.**

$$\rho_t = \frac{(\mathbb{E}_t[l(\theta^*, A_t) - l(\theta^*, A^*)])^2}{I_t(\theta^*; L_t)}.$$

We can use

$$\theta \to A_t \to L_t$$

to describe the relationship between *theta*, A_t and L_t, the arrow means given A_t, *theta* and L_t are conditional independent. Then the data processing inequality implies that the information gain about θ^* is always smaller than that about A_t^* , which in turn implies that ρ_t is greater than ρ_t^*.

As our analysis will establish, a bounded lifted information ratio guarantees low regret, and we will show that the ratio itself can be bounded reasonably under conditions similar to the ones required by the analysis of [2].

4 Main results

Theorem 4.1. *Assume Q_1 is supported on the countable set $\Theta_1 \subseteq \Theta$ and that the lifted information ratio for all rounds t satisfies $\rho_t \leq \rho$ for some $\rho > 0$. Then, the Bayesian regret of TS after T rounds can be bounded as*

$$R_T \leq \sqrt{\rho T H(\theta^*)}.$$

Using this theorem, if we can find upper bounds for ρ and $H(\theta^*)$, then we find upper bound for the regret.

Lemma 4.2. *Suppose that the losses are binary and $|\mathcal{A}| = K$. Then, the lifted information ratio of Thompson sampling satisfies $\rho_t \leq 2K$ for all $t \geq 1$.*

We now instantiate our bounds in two well-studied settings for Bernoulli bandits. We start from the fully unstructured case, assuming finite actions and finitely supported prior. The following regret bound follows direst from Theorem 4 and Lemma 4.2.

Theorem 4.3. *Consider a contextual bandit with K actions and binary losses, and suppose Θ_1, the support of Q_1, is finite with $|\Theta_1| = N$. Then, the Bayesian regret of TS satisfies:*

$$R_T \leq \sqrt{2KT \log N}.$$

Unfortunately, the Shannon entropy can be unbounded for distributions with infinite support, which is in fact the typical situation that one encouters in practice. To address this concern, we develop a more general result, that holds for a broader family of distributions.

In the following, (Θ, ϱ) is a metric space with metric $\varrho : \Theta^2 \to \mathbb{R}$. We make the following regularity assumption on the likelihood function $P_{\theta,x,a}$:

Assumption 1 (log-Lipschitz)**.** *There exists a constant $C > 0$ such that for any $\theta, \theta' \in \Theta_1$, $|\log P_{\theta,x,a} - \log P_{\theta',x,a}| \leq C\varrho(\theta, \theta')$ holds for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $L \in \{0, 1\}$.*

Under this assumption, we can state a variant of Theorem 1 that applies to metric parameter spaces:

Theorem 4.4. *Assume (Θ, ϱ) is a metric space, and Q_1 is supported on $\Theta_1 \subseteq \Theta$ with ϵ-coverning number $\mathcal{N}_\epsilon(\Theta_1, \varrho)$. Let assumption hold, and assume the lifted information ratio for all round t satisfies $\rho_t \leq \rho$ for some $\rho > 0$. Then, the Bayesian regret of TS after T rounds can be bounded as*

$$R_T \leq \sqrt{\rho T \min_\epsilon (\log \mathcal{N}_\epsilon(\Theta_1, \varrho) + 2\epsilon CT)}$$

When the Shannon entropy is bounded, we can use it to bound the regret. And the artical find one way to deal with unbounded entropy problem by adding Lipschitz assumption.

One potential direction of improvement is to find new ways to bound $\sum_{t=1}^{T} I_t(\theta^*, L_t)$.

4.1 Logistic bandits

In this model, the losses are generated by a Bernoulli distribution as $L_t(\theta, x, a) \sim Ber(\sigma(f_\theta(x, a)))$, where $\sigma(z) = 1/(1 + e^{-z})$ is sigmoid function.

See Theorem 4 and Corollary 1 in [1] for the results.

4.2 Linear bandits

We can consider two types of linear bandits. And the regret analysis is similar. The first one supposes the losses are binary, and the expected losses are linear functions of the form $l(\theta, x, a) =< \theta, \phi(x, a) >$, see Lemma 2 in [1].

Another type is linear bandits with Gaussian noise. $L_t \sim \mathcal{N}(l(\theta^*, X_t, A_t), \sigma^2)$. See Lemma 3 in [1].

Another potential direction is apply the method to more types of bandits. For one thing, we can use this lifted information ratio to deal with other types of contextual bandit. For another, we can develop new theory about information ratio to deal with bandits beyond basic and contextual bandits. For example, we may find another type of information ratio.

I don't have so much knowledge about types of bandits beyond these, and we can have a discussion here.

References

[1] Gergely Neu, Julia Olkhovskaya, Matteo Papini, and Ludovic Schwartz. Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits. *arXiv preprint arXiv:2205.13924*, 2022.

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