The Thompson sampling algorithm and applications to contextual bandits

Prithvijit Chakrabarty
Department of Computer Science
University of Massachusetts
Amherst, MA 01002
pchakrabarty@umass.edu

Abstract

Thompson sampling is a Bayesian algorithm to handle the exploration/exploitation dilemma in the multi-armed bandit (MAB) problem. Despite a number of empirical studies demonstrating its effectiveness in practice, there is limited theoretical understanding of its worst case performance. Contextual bandits is a generalization of the MAB problem, which requires learning a policy instead of the best arm. This report discusses the regret bound on applying Thompson sampling applied to contextual bandits. This is an open problem. Previous theoretical results on the algorithm either focused on the stochastic MAB or made strong assumptions on the policy class used in the contextual setting. This report briefly describes the key ideas used for the proofs in these results. It also discusses the difficulties in applying them directly in the general contextual bandits setting, with possible workarounds.

1 Introduction

The exploration-exploitation dilemma is a problem that arises in many settings which require learning a decision strategy. This is formally modeled as the multi-armed bandit (MAB) problem. The "arms" refer to the various actions a learner might take (belong to a set of arms $\mathcal{A}$). Every action is associated with a reward. The goal of the learner is to maximize its rewards over time. At each timestep t, the learner selects an arm $a_t \in \mathcal{A}$ and receives the corresponding reward $r_t(a_t)$. The goal of the learner is to maximize its reward over time. A related quantity is the regret $\mathcal{R}(T)$, which defined as the difference between the reward obtained by the learner ($r_t(a_t)$) and the maximum possible reward ($r_t(a_t^*)$, where $a*_t$ is the appropriate action at time t).

$$\mathcal{R}(T) = \sum_{t=1}^{T} r(a_t^*) - r_t(a_t)$$

The learner can maximizing the reward by minimizing the regret. There are many variants of the MAB problem and algorithms to solve them. This report discusses the Thompson Sampling algorithm and its possible to contextual bandits.

1.1 Thompson Sampling

Thompson Sampling is a heuristic for solving the explore/exploit problem. Initially proposed in Thompson [1933] for drug trials, interest in the algorithm rose again with the popularity of reinforcement and fast learning algorithms which require effective exploration schemes. Ortega and Braun [2010],Strens [2000] applied the algorithm in reinforcement learning. Wyatt [2001] studied a

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

closely related algorithm in the reinforcement learning setting. Thompson sampling is a Bayesian algorithm. At every timestep the algorithm plays the best arm according to its current estimates (the prior). On receiving a reward r, it updates its estimate (the posterior) using Bayes rule:

$$P[\theta|r] = \frac{P[r|\theta]P[\theta]}{\int P[r|\theta]P[\theta]d\theta}$$

If the learner receives Bernoulli feedback ($r \in \{0, 1\}$), a convenient choice for the posterior distribution is the Beta distribution as it is the *conjugate prior*. In this case, the posterior for an arm a can be maintained simply as a beta distribution whose parameters are the number of times that a succeeded or failed.

Algorithm 1: Thompson Sampling for Beta-Bernoulli bandits

For each arm a, set $S_a = F_a = 1$ (success and failure count)
foreach $t = 0, ..., T$ **do**
    For each arm $a \in \mathcal{A}$ sample reward estimate $r_t(a) \sim \beta(S_a, F_a)$
    Play $a_t = \text{argmax}(r_t(a))$
    Observe reward $r_t(a_t)$
    Update posterior parameters: $S_{a_t} = S_{a_t} + r_t(a_t), F_{a_t} = F_{a_t} + (1 - r_t(a_t))$

Though there are provably optimal alternatives to Thompson sampling (the UCB family of algorithms), studies such as Graepel et al. [2010], Scott [2010] Granmo [2010] and show that it often performs better in practice. It is also easier to implement than most exploration schemes and allows incorporating prior beliefs without modification to the algorithm. Chapelle and Li [2011] demonstrates empirically that the Thompson Sampling outperformed LinUCB in training a logistic regression model for news article recommendation. Despite this rise in interest, there was no theoretical understanding of the algorithm until 2012. Agrawal and Goyal [2012] derived the first guarantees for the algorithm, showing that for a 2-armed stochastic bandit problem, the regret is bounded by

$$E[\mathcal{R}(T)] \leq \mathcal{O}\Big(\frac{\ln T}{\Delta} + \frac{1}{\Delta^3}\Big)$$

and for N-armed bandits, the regret is

$$E[\mathcal{R}(T)] \leq \mathcal{O}\Big(\Big(\sum_{i=2}^{N} \frac{1}{\Delta_i^2}\Big)^2 \ln T\Big)$$

Here, Δ_i refers to the suboptimality of an arm, i.e., $\Delta_i = \mu_i - \mu^*$, where μ_i is the mean reward for arm i and μ^* is the mean reward for the best arm. After this, Kaufmann et al. [2012] proved that this bound is close to the lower bound given by Lai and Robbin for bandit problems Lai and Robbins [1985], proving that for the stochastic multi-armed bandits Thompson Sampling is indeed optimal.

1.2 Contextual bandits

Contextual bandits is a variation of the MAB problem in which the learner is given extra context or "side information" at every timestep. The rewards each arm will produce depend on this information and the appropriate arm to be played can be computed using the context and a policy. The learner is required to find a good policy while receiving partial feedback (feedback only on the arm it selected).

Formally, at every timestep t, the learner observes a context x_t drawn from a context space $\mathcal{X}$ and plays one of N arms $a_t \in \mathcal{A}$ (if the arms have numerical labels, $\mathcal{A} = [N]$). It then receives a reward $r_t(a_t) \in [0, 1]$, depending its choice of arm. The goal is to learn a policy $\pi : \mathcal{X} \to \mathcal{A}$ which maps a context to the appropriate arm. We assume that the learner competes within a policy class Π, and must learn the best possible policy π^* in this class. In this setting, the maximizing the reward is equivalent to minimizing the regret $\mathcal{R}(T)$:

$$\mathcal{R}(T) = \sum_{t=1}^{T} r_t(\pi^*(x_t)) - r_t(a_t)$$

The contextual bandits problem was introduced by Langford and Zhang [2008]. In 2014, Agrawal and Goyal [2013] proved regret bounds for Thompson sampling applied to contextual bandits with linear payoffs (the reward vector can be computed from the context: $r_t = w^T x_t$).

$$\begin{aligned}\mathcal{R}(T) &\leq \mathcal{O}(d\sqrt{T\log(N)}(\ln(T) + \sqrt{\ln(T)ln(\frac{1}{\delta})})) \\ &= \tilde{\mathcal{O}}(d\sqrt{T\log(N)})\end{aligned}$$

Contextual bandits with linear payoffs have a lower bound of $\Omega(d\sqrt{T})$, which was given in Dani et al. [2008], when the number of arms are allowed to be infinite. In the finite N armed setting, Chu et al. [2011] proved a lower bound to be $\Omega(\sqrt{dT})$ when $d^2 \leq T$.

The regret bound for Thompson sampling differ from the lower bound by a factor of $\sqrt{\log(N)}$. If the number of arms is exponential in d, the regret bound above would be $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$, which differs from the optimal a factor of $\sqrt{d}$. There are algorithms which do achieve the theoretical lower bound. For example, Bubeck et al. [2012], give an algorithm which achieves a regret bound of $O(\sqrt{dT\log(N)})$ for finite arms and $O(d\sqrt{T})$ for infinite arms. Agrawal and Goyal [2013] suggests that the factor of $\sqrt{d}$ is the what the algorithm pays for its efficiency. They note that Bubeck et al. [2012] needs to maintain a distribution linear in the number of arms, while Chu et al. [2011] and Dani et al. [2008] effectively require solving an NP complete problem at every round. They also suggest that this factor might not be eliminated for any efficient algorithm to this problem. As an example, they cite is Algorithm 3.2 in Dani et al. [2008] which also suffers the extra factor in its regret bound.

It is worth noting that this bound holds under the realizability condition, i.e., there exists an ideal w^*, such that $w^{*T}x_t$ will always result in the correct estimate of the reward. The realizability assumption has been quite common in related work on contextual bandits Filippi et al. [2010], Auer [2002], Chu et al. [2011] all assume the realizability. Finding the regret bound of Thompson sampling for contextual bandits without this assumption is an open problem.

2 Problem setting

There are N arms. At each timestep t, the learner is presented with a context $x_t \in \mathcal{X}$. The learner has access to a function class F, consisting of functions $f_\theta : \mathcal{X} \to \mathbb{R}$. Here, $\theta \in \Theta$ is a parameter vector which can be used to completely specify any function in F. Thus, sampling a value from a distribution over Θ corresponds to sampling a function f_θ from F. There exists a reward function $r : \mathcal{X} \to \mathbb{R}$ per arm which correctly predicts the reward given the context. Also, for each arm a, θ_a^* is the parameter vector for the function in F which performs best in predicting its reward (closest to the true function r).

Algorithm 2: Thompson Sampling for contextual bandits

For each arm a, set posterior $P_a[\theta|r] = Uniform(\Theta)$
foreach $t = 0, ..., T$ **do**
  Observer context x_t
  For each a = 1,...,N sample a model $\theta_a(t) \sim P_a[\theta|r]$
  Play $a_t = \text{argmax} f_{\theta_a(t)}(x_t)$
  Observe reward $r = r_t(a_t, x_t)$
  Update posterior for the selected arm: $P_{a_t}[\theta|r, x_t] \propto P_{a_t}[r|\theta, x_t].P_{a_t}[\theta]$

Note that this is an equivalent, but slightly different formulation from the standard contextual bandits, where the policies directly map contexts to arms. In this setting, f_θ behaves as a value function. The output of f_θ followed by the argmax operation collectively behaves as the policy.

3 Challenges in proof

This section discusses the difficulties in finding the regret bound for Thompson sampling without any assumption on the policy class. The proof techniques in Agrawal and Goyal [2012] used for the

multi-armed bandits are used reference. These also form the basis of the proof for linear contextual bandits Agrawal and Goyal [2013]. The presented difficulties prevent us applying those techniques directly in the contextual bandits setting.

3.1 Posterior computation

Consider using a normal distribution as the likelihood function. Then, $P[r_t|\theta, x_t] = \mathcal{N}(f_\theta(x_t), \sigma^2)$. For simplicity, let $\sigma = 1$. This yields:

$$P[r_t|\theta, x_t] = \mathcal{N}(f_\theta(x_t), 1) \propto \exp\Big(-\frac{(r_t - f_\theta(x_t))^2}{2}\Big)$$

Now, at each timestep t, the algorithm updates the posterior as:

$$P[\theta|r_t, x_t] \propto P[r_t|\theta, x_t].P[\theta]$$

Unrolling this in time, at timestep T, we have:

$$P[\theta|r_T, x_T] \propto \prod_{t=1}^{T} P[r_t|\theta, x_t].P[\theta_0]$$

$$= \Big(\prod_{t=1}^{T} \exp\Big(-\frac{(r_t - f_\theta(x_t))^2}{2}\Big)\Big).P[\theta_0]$$

$$= \exp\Big(-\frac{1}{2}\sum_{t=1}^{T}(r_t - f_\theta(x_t))^2\Big).P[\theta_0] \tag{1}$$

Further simplifying this expression requires making assumptions on the form of the function class F. This poses a problem, as we will not have a closed-form solution for the posterior distribution. In Agrawal and Goyal [2013], an important element in the proof was using the form $f_\theta(x_t) = \theta^T x_t$ (due to the linear payoff assumption) to show that the posterior will also be a normal distribution. the

Possible workaround Without simplifying the expression for the posterior, it is clear that this procedure behaves like a weighting algorithm. It places a high probability mass on functions in F which reduce the squared error on the observed data (the term $-\frac{1}{2}\sum_{t=1}^{T}(r_t - f_\theta(x_t))^2$ in equation (1) is essentially the squared loss that f_θ suffers on the observed samples).

In Agrawal and Goyal [2012], the proof used the Hoeffding inequality to derive the minimum number of times an arm must be played to get a good estimate of the mean reward:

$$L = \frac{\ln T}{\Delta^2}$$

After L plays of an arm, its posterior will be concentrated around the true mean reward. This is not applicable here, as we are not estimating the mean. However, as the algorithm is effectively minimizing the squared loss and placing a high probability mass over the best models, we can lower bound the number of times an arm should be played with the sample complexity of the function class F.

Using a fat shattering dimension F is essentially a class of regression functions (mapping contexts to rewards in $\mathbb{R}$). Thus, we cannot use the VC dimension to estimate the sample complexity. I am not sure if this is correct, but for such a class, $P_{dim}(F)$, the Pollard pseduo dimension (or a fat shattering dimension at scale ϵ) of the F can be used to estimate the sample complexity Sample complexity [2010]. As F maps $[0, 1]$, we have:

$$N_F = \mathcal{O}\Big(\frac{P_{dim}(F)\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta}}{\epsilon^2}\Big) \tag{2}$$

Further, consider the 2-armed scenario, let $\Delta = \min_{x_t} f_{\theta_0^*}(x_t) - f_{\theta_1^*}(x_t)$. Δ is the minimum difference in true rewards of the arms across all possible contexts. Thus, to pick the correct arm, we must have an estimator f_θ with an error margin $\epsilon \leq \frac{\Delta}{4}$. Using this and setting $\delta = \frac{1}{T}$, we have:

$$N_F = \mathcal{O}\Big(\frac{P_{dim}(F)\ln\frac{4}{\Delta} + \ln T}{\Delta^2}\Big) \tag{3}$$

This lower bound is similar to that for the multi-armed bandits (L), with an extra term dependent on F.

Using the Rademacher complexity An alternative approach might be to use the Rademacher complexity of the function class F. Let $\mathcal{R}$ denote the Rademacher complexity of F with the squared loss. With N samples, we have the generalization bound:

$$\epsilon \leq \mathcal{O}\left(\mathcal{R} + \sqrt{\frac{\log \frac{4}{\delta}}{N}}\right) \tag{4}$$

Solving this for N, we get an expression for N_F

$$N = \frac{\log \frac{4}{\delta}}{(\epsilon - \mathcal{R})^2}$$

Again, setting $\delta = \frac{1}{T}$ and $\epsilon = \frac{\Delta}{4}$, we get a bound similar to the expression for L, but with an extra term dependent on the function class F:

$$N = \mathcal{O}\Big(\frac{\log T}{(\Delta - \mathcal{R})^2}\Big) \tag{5}$$

3.1.1 Lack of ordering in the parameter space

Given the minimum number of plays of each arm, the proof in Agrawal and Goyal [2012] proceeds by case analysis. Consider one possible case: a 2-armed stochastic bandit problem where arm 1 has been played L times. The posterior for this arm will be concentrated around its true mean. Thus, the reward samples drawn for this arm will be close to μ_1. Now, we can approximate the probability of selecting arm 2 by computing $\mathbb{P}[\theta_2 \geq \mu_1]$, where θ_2 is a sample drawn from the posterior on arm 2 (let this be $Q_2(r)$). As the Q_2 is a distribution over the reward space, we have $\mathbb{P}[\theta_2 \geq \mu_1] = \int_{\mu_1}^{\infty} Q_2(r)dr$. With the closed form expression for Q_2, this can be computed or approximated (Agrawal and Goyal [2012] show that for Beta-Bernoulli bandits, this can be approximated with a geometric random variable).

Now, consider the same case in the contextual setting. At timestep T, arm 1 has been played N_F times. Thus, a parameter vector θ_1 sampled from this arm will accurately model the reward, i.e., $f_{\theta_1}(x_t) \approx r_1(x_t)$. Also, let the true reward for the second arm be $r_2(x_t)$ such that $r_1(x_t) - r_2(x_t) = \Delta(x_t)$. The probability of playing arm 2 will be:

$$\begin{aligned} &\mathbb{P}[f_{\theta_2}(x_t) \geq f_{\theta_1}(x_t)] \\ &\approx \mathbb{P}[f_{\theta_2}(x_t) \geq r_1(x_t)] \\ &= \mathbb{P}[f_{\theta_2}(x_t) \geq r_2(x_t) + \Delta(x_t)] \\ &= \mathbb{P}[f_{\theta_2}(x_t) - r_2(x_t) \geq \Delta(x_t)] \end{aligned} \tag{6}$$

Thus, computing the probability of playing arm 2 is equivalent to computing the following:
If we learn a function class F to minimize the squared error on n samples (n is less than the sample complexity of F), what is the probability that the error on a new sample will be greater than $\Delta(x_t)$.

I am not sure if there is a way to compute this. Intuitively, this seems to be related to the mistake bound for learning the function class F.

Possible workaround We can try to find an expression for ϵ, given the number of samples. This is where the Rademacher complexity is easier to use than the shattering number. If an arm has been played n times, rearranging equation (4) gives:

$$n\epsilon^2 + P_{dim}(F) \ln \epsilon \leq \ln \frac{1}{\delta}$$

which is a transcendental equation and is difficult to solve (though we can get an approximate upper bound for ϵ). On the other hand, with the Rademacher complexity, equation (4) directly gives an upper

bound for ϵ. Now, from equation (6), we are interested in the quantity $f_{\theta_2}(x_t) - r_2(x_t)$. However, ϵ will be equivalent to $|f_{\theta_2}(x_t) - r_2(x_t)|$.

Intuitively, a generalization bound will give the probability that the distance between the prediction and the true value is greater than a threshold. However, for the case analysis we require, we specifically need the probability of the prediction being greater than or less than the true value.

I am not sure how this can be handled. If there is a way to compute this, then we can proceed along the lines of the case analysis and to bound the regret till the timestep when the models on both arms have been learned.

4 Conclusion

This report described the Thompson sampling algorithm and its application to the contextual bandits problem. The algorithm is modeled as a training algorithm to learn a policy. A Gaussian likelihood function is used, and no assumptions were made on the policy class.

The proof techniques used to analyze the algorithm in other settings cannot be applied to this problem. In particular, the posterior computation is difficult without knowledge of the policy class. A workaround is suggested which may be useful in finding a proof without explicitly computing the posterior. A second problem arises in estimating the probability of selecting one arm over the other. With generalization bounds, we can estimate the probability of an error, which is will bound the distance to the true value. However, we specifically require bounds on the probability the estimates are greater or less than a certain threshold. This quantity is hard to compute. If there is a workaround, it may be used to continue the proof along the same lines. If not, we require a different approach for the prove the regret bound.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Annual Conference on Learning Theory*, volume 23, pages 41–1. Microtome, 2012.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

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.

Thore Graepel, Joaquin Q Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 13–20, 2010.

Ole-Christoffer Granmo. Solving two-armed bernoulli bandit problems using a bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*, 3(2):207–234, 2010.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*, volume 12, pages 199–213. Springer, 2012.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

Pedro A Ortega and Daniel A Braun. Linearly parametrized bandits. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

Sample complexity. Sample complexity — Wikipedia, the free encyclopedia, 2010. URL `https://en.wikipedia.org/wiki/Sample_complexity`. [Online; accessed 19-December-2017].

Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, pages 943–950, 2000.

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.

Jeremy L Wyatt. Exploration control in reinforcement learning using optimistic model selection. In *ICML*, pages 593–600, 2001.