JMLR: Workshop and Conference Proceedings vol 23 (2012) 43.1–43.3 25th Annual Conference on Learning Theory

Open Problem: Regret Bounds for Thompson Sampling

Lihong Li LIHONG@YAHOO-INC.COM
Yahoo! Research
Santa Clara, CA 95054

Olivier Chapelle* O.CHAPELLE@CRITEO.COM
Criteo
Palo Alto, CA 94301

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

1. Background

Contextual multi-armed bandits (Langford and Zhang, 2008) have received substantial interests in recent years due to their wide applications on the Internet, such as new recommendation and advertising. The fundamental challenge here is to balance exploration and exploitation so that the total payoff collected by an algorithm approaches that of an optimal strategy. Exploration techniques like ϵ-greedy, UCB (upper confidence bound), and their many variants have been extensively studied. Interestingly, one of the oldest exploration heuristics, dated back to Thompson (1933), has not been popular in the literature until recently when researchers started to realize its effectiveness in critical real-world applications (Scott, 2010; Graepel et al., 2010; May and Leslie, 2011; Chapelle and Li, 2012).

This heuristic, known as *Thompson sampling*, fulfills the principle of "probability matching," which states that an arm is chosen with the probability that it is the optimal one. A generic description is given in Algorithm 1, where the algorithm maintains a posterior distribution $P(\theta|D)$ over a parameter space Θ that defines a set of greedy policies. At every step, a random model θ^t is drawn from the posterior, and the greedy action according to the payoff predictions of θ^t is chosen.

Algorithm 1 Thompson sampling (adapted from Chapelle and Li (2012))

Initialize observed data set: $D \leftarrow \emptyset$
for $t = 1, \ldots, T$ **do**
  Observe context x_t
  Draw $\theta^t \in \Theta$ according to $P(\theta|D)$, and select $a_t = \arg\max_a \mathbb{E}_r \left[r|x_t, a, \theta^t\right]$
  Observe payoff $r_t(a_t)$, and augment observed data set $D \leftarrow D \cup (x_t, a_t, r_t(a_t))$
end for

Thompson sampling has a number of significant advantages in practice. First, it can be easily combined with Bayesian approaches and complicated parametric models (Graepel et al., 2010; Chapelle and Li, 2012); in contrast, popular exploration strategies like UCB are often hard to derive except in special cases like (generalized) linear models. Second, a number of recent empirical studies, including those conducted on large-scale, real-world problems, have shown the algorithm is highly effective for balancing exploration and exploitation (Scott, 2010; Graepel et al., 2010;

* Work done while author was at Yahoo Research.

May and Leslie, 2011; Chapelle and Li, 2012). Furthermore, Thompson sampling appears to be more robust to observation delays of payoffs, compared to deterministic exploration strategies like UCB (Chapelle and Li, 2012).

2. Known Results

In contrast to the promising empirical findings, there has not been many theoretical results. Our notion of performance here is *regret*—the difference between the total payoffs achieved by the algorithm and the highest achievable payoffs. Formally, let $r_t(a)$ denote the payoff at the t-th step if arm a is chosen, and let $a_t(\theta) = \arg\max_a \mathbb{E}\left[r|x_t, a, \theta\right]$ be the greedy arm if payoffs are accurately predicted by model θ. The T-step regret is then given by $R(T) = \max_\theta \sum_{t=1}^{T} \left(r_t(a_t(\theta)) - r_t\right)$.

The earliest theoretical results are asymptotic in nature (Granmo, 2010; May et al., 2011). These results ensure that, under certain natural assumptions, Thompson sampling converges to an optimal arm-choosing policy, which may only imply $R(T) = o(T)$. While these results are important for showing the fundamental correctness of the algorithm, they do not tell how fast the algorithm "learns" after a finite number of steps.

More recently, Agrawal and Goyal (2011) give the first non-trivial regret upper bound for the special case of the traditional (non-contextual) K-armed bandits where Beta distributions are used as the prior. For $K = 2$, their upper bound is $O(\ln T/\Delta)$ and matches the well-known lower bound of Lai and Robbins (1985), where Δ is the difference in the expected payoff between the optimal arm and the suboptimal one. For general $K > 2$, their upper bound still scales logarithmically in T, but the multiplicative constant, $\left(\sum_{i=2}^{K} \Delta_i^{-2}\right)^2$, is significantly worse.

3. Open Questions

Given the encouraging empirical results, the lack of theoretical understanding of Thompson sampling implies a number of open questions. A sample of them are given below.

First, for the traditional (non-contextual) K-armed bandits, there is a substantial gap between the state-of-the-art result of Agrawal and Goyal (2011) and the lower bound of Lai and Robbins (1985). Inspired by numerical simulations (Chapelle and Li, 2012), we conjecture that Thompson sampling's regret actually matches the lower bound and is indeed optimal: $O(\sum_{i=2}^{K} \Delta_i^{-1} \cdot \ln T)$. Furthermore, it remains open whether one can find a problem-independent regret bound that does not depend on $1/\Delta_i$ – a quantity that can be arbitrarily small. An ideal answer would be $O(\sqrt{T})$, in light of existing results (Auer et al., 2002).

Second, there is no published regret bound in the contextual setting, which is arguably a more useful setting in practice for generalization reasons (Langford and Zhang, 2008). Given the similarity between Bayes' formula and the exponential update rule, some of the techniques from the online-learning literature may be useful. A straightforward application of the techniques, however, does not seem to produce the strong upper bound of $O(\sqrt{T})$. One obstacle is the possible under-exploration of the optimal θ^* if it is given a low prior probability; in this case, the algorithm may have to wait a long time before seeing one data point to increase the posterior of θ^*. Another family of recent algorithms get around this problem by explicitly "massaging" the selection probabilities of promising θ values, which unfortunately can be computationally expensive (Dudík et al., 2011; Agarwal et al., 2012). One obvious fix is to mix Thompson sampling with uniform exploration, as

done in EXP4 (Auer et al., 2002), but the regret bound becomes $O(T^{2/3})$, no better than the simple epoch-greedy algorithm (Langford and Zhang, 2008).

Acknowledgments

We appreciate helpful discussions with John Langford and Miroslav Dudík.

References

A. Agarwal, M. Dudík, S. Kale, J. Langford, and R. E. Schapire. Contextual bandit learning under the realizability assumption. In *AISTATS*, 2012.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2012.

M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *UAI*, pages 169–178, 2011.

T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *ICML*, pages 13–20, 2010.

O.-C. Granmo. Solving two-armed bernoulli bandit problems using a bayesian learning automaton. *Int'l Journal of Intellient Computing and Cybernetics*, 3(2):207–234, 2010.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems 20*, pages 1096–1103, 2008.

B. C. May and D.S. Leslie. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:02, Dept. of Mathematics, Univ. of Bristol, 2011.

B. C. May, N. Korda, A. Lee, and D.S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:01, Dept. of Mathematics, Univ. of Bristol, 2011.

S. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658, 2010.

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.