Discussion of "Contextual Bandit Algorithms with Supervised Learning Guarantees"

H. Brendan McMahan
Google, Inc.

1 Contextual bandit problems

It is my pleasure to provide some commentary on the paper "Contextual Bandit Algorithms with Supervised Learning Guarantees" by Beygelzimer et. al. This discussion synthesizes my own impressions of the paper, as well as the comments of the anonymous reviewers.

For applied machine learning, finding an appropriate formulation of the problem is essential. Supervised learning is perhaps the most successful and extensively studied formulation in the field. It is broad enough to encompass many real-world problems, but is narrow enough that significant theoretical results are possible.

Nevertheless, it is becoming increasingly clear that supervised learning is too restrictive for some important applications. Supervised learning is about making good predictions, but often one cares more about the outcome of taking actions, in particular in environments where feedback is only received for the action taken. Applications in web search and advertising are important motivating examples: at the end of the day, what matters is what search results (or ads, or news results) we choose to show to users, and whether those users like (click) on those results. These web applications are particularly compelling because the scale of the problem and latencies required mandate an automated solution. The contextual bandit formulation captures both the measurement of success in terms of the real-world quantity of interest (clicks) as well as addressing the inherent explore/exploit trade-offs.

It is useful to think of two general methods for tackling the contextual bandits problem. The first imposes some structure on the set of possible policies, and then uses that structure to extract guarantees. For example, Langford and Zhang [2007] assume an oracle for solving the offline problem, and Auer [2003] assumes the rewards for each action are some linear function of the context vector. The second approach requires an enumeration of policies, and then applies an algorithm for multi-armed bandits with expert advice. The advantage of this approach is that if one can afford (or make efficient) the policy enumeration, no further assumptions on the policies or rewards are required. The present paper is of this latter style; it improves on earlier results, like those of Auer et al. [2002] and McMahan and Streeter [2009].

2 Contributions of the Present Work

The present paper's central contribution is a version of the Exp4 algorithm that offers high-probability guarantees on regret, as opposed to the bounds in expectation proved in the two papers just mentioned. This is accomplished via a generalization of known martingale tail bounds; this theorem seems likely to be useful in the analysis of other online algorithms.

They also show that in the case of a stochastic environment, one can obtain low regret against an infinite policy class as long as it has finite VC dimension. The trick is to construct a suitable finite sample of policies on which their algorithm, Exp4.p, can be applied.

While such an approach shows regret bounds are possible, for practical classes the naive implementation will not be efficient as the finite set of experts will still typically be exponential in size. In the experiments section, the authors demonstrate how Exp4.p can be simulated on an exponential set of experts using only polynomial space and time. Developing more general or powerful classes of policies where such efficient implementations are possible is a promising direction for future work, in some ways bridging the gap between the structural and enumeration approaches to the contextual bandits problem discussed earlier.

3 On the Experiments

I am pleased to see experiments on a large, real-world problem included in the paper. While these experi-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

ments are not the primary contribution, in some ways they raise more questions than answers (many beyond the scope of the present paper), and so provide a fertile topic for discussion. First, a few specific points:

- The interpretation of the experiments would be more clear if they compared to the algorithm of McMahan and Streeter [2009] (or the variant that uses the approach of Section 6), as the current results do not disambiguate between changes in performance due to this approach versus the changes that allow for high-probability bounds.

- Since the principal improvement of Exp4.p is a high-probability bound on regret, it would have been interesting to compare the empirical distribution on regret for the current algorithm versus Exp4. Presumably, while Exp4.p would pay a small penalty in terms of mean regret, the variance should be lower.

- The distinction between "learning CTR" and "deployment CTR" seems a bit artificial: bandit algorithms are designed to be run on all the data. Splitting the data into two buckets is effectively making a different explore/exploit trade-off than indicated by the theory.[1] How does splitting the data this way compare to tuning the individual algorithms to do more exploitation (e.g, by tuning the $p_{\min}$ parameter for Exp4.p)? Similarly, a natural straw-man is the algorithm uses the random policy in the "learning" bucket. I would like to see results for this approach.

This last point suggests a natural open question: if the splitting approach is preferable in practice, can one prove regret bounds for the *total* regret over both buckets by for example running Exp4.p with a larger value of $p_{\min}$ in the learning bucket? (This seems likely).

These points should not overshadow my primary comment on the experiments: it's great to see them, and many papers (some of mine included) have done less.

4 Looking Forward

It is clear that the study of supervised learning has benefited from the availability of standard evaluation datasets. Despite the clear practical applications, papers on bandit algorithms have included relatively less experimental work. The development and standardization of empirical evaluation methodologies for contextual bandit problems is an important problem for the community, as it should pave the way for greater real-world impact of the techniques developed. But as the present paper shows, several challenges arise:

- In real world bandit problems, by definition one only observes the outcome of the action taken. Thus, one must be clever in order to use real-world data to evaluate an arbitrary policy.

- The most compelling applications of contextual bandit algorithms are large-scale problems of significant commercial importance that involve user interactions. Thus, making such datasets publicly available is challenging for both business and privacy reasons.

These issues make establishing benchmark real-world datasets more difficult. Until this can be accomplished, perhaps there is a role for standardized synthetic problems? But then the choice of assumptions (stochastic? fully adversarial? structure of the experts?) can quickly bias the problem towards particular approaches.

Handling parameter tuning is also tricky: most contextual bandit algorithms have an explore/exploit trade-off parameter, with regret bounds that provide a suggested setting. However, for a particular problem this setting is often not optimal, and so plugging in a fixed setting for each algorithm may not give a fair comparison; on the other hand, optimizing over this parameter may lead to overfitting and might not be possible in real settings.

I expect contextual bandit problems to play an increasingly prominent role in machine learning research. The work in this paper is a clear step forward, but significant theoretical questions remain open, and there is still much work to be done in showing how real-world problems can be successfully formulated as contextual bandit problems. As more algorithms become evaluable, developing empirical as well as theoretical tools to help select the best algorithm for a particular application will become increasingly important.

References

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3, March 2003.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *NIPS 2007*, 2007.

H. Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In *COLT*, 2009.

[1] This may well be warranted for a specific problem — the theory often suggestions too much exploration in order to handle with the worst case.