Bypassing the Monster: A Faster and Simpler Optimal Algorithm for Contextual Bandits under Realizability

David Simchi-Levi DSLEVI@MIT.EDU
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Yunzong Xu YXU@MIT.EDU
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract

We consider the general (stochastic) contextual bandit problem under the realizability assumption, i.e., the expected reward, as a function of contexts and actions, belongs to a general function class $\mathcal{F}$. We design a fast and simple algorithm that achieves the statistically optimal regret with only $O(\log T)$ calls to an offline least-squares regression oracle across all T rounds. The number of oracle calls can be further reduced to $O(\log \log T)$ if T is known in advance. Our results provide the first universal and optimal reduction from contextual bandits to offline regression, solving an important open problem in contextual bandits. A direct consequence of our results is that any advances in offline regression immediately translate to contextual bandits, statistically and computationally. This leads to faster algorithms and improved regret guarantees for broader classes of contextual bandit problems.

Keywords: Contextual Bandits, Least Squares Regression, Computational Efficiency, Statistical Optimality, Online-to-Offline Reduction, Fast Algorithm

1. Introduction

The contextual bandit problem is a fundamental framework for online decision making and interactive machine learning, with diverse applications ranging from healthcare (Tewari and Murphy 2017; Bastani and Bayati 2020) to electronic commerce (Li et al. 2010; Agarwal et al. 2016), see a NIPS 2013 tutorial (link) for the theoretical background, and a recent ICML 2017 tutorial (link) for further illustrations on its practical importance.

Broadly speaking, approaches to (stochastic) contextual bandits can be classified into two categories (see Foster et al. 2018): *realizability-based* approaches which rely on weak or strong assumptions on the model representation, and *agnostic* approaches which are completely model-free. While many different contextual bandit algorithms (realizability-based or agnostic) have been proposed over the past twenty years, most of them suffer from either theoretical or practical issues (see Bietti et al. 2018). Existing realizability-based algorithms built on upper confidence bounds (e.g., Filippi et al. 2010; Abbasi-Yadkori et al. 2011; Chu et al. 2011; Li et al. 2017) and Thompson sampling (e.g., Agrawal and Goyal

2013; Russo et al. 2018) rely on strong assumptions on the model representation and are only tractable for specific parametrized families of models like generalized linear models. Meanwhile, agnostic algorithms that make no assumption on the model representation (e.g., Dudik et al. 2011; Agarwal et al. 2014) may lead to overly conservative exploration in practice (Bietti et al. 2018), and their reliance on an *offline cost-sensitive classification oracle* as a subroutine typically causes implementation difficulties as the oracle itself is computationally intractable in general (Klivans and Sherstov 2009). At this moment, designing a provably optimal contextual bandit algorithm that is applicable for large-scale real-world deployments is still widely deemed a very challenging task (see Agarwal et al. 2016; Foster and Rakhlin 2020).

Recently, Foster et al. (2018) propose a (realizability-based) approach that solves contextual bandits with general model representations (i.e., general function classes) using an *offline regression oracle* — an oracle that can typically be implemented efficiently and has wide availability for numerous function classes due to its core role in modern machine learning. In particular, the (weighted) least-squares regression oracle assumed in the algorithm of Foster et al. (2018) is highly practical as it has a strongly convex loss function and is amenable to gradient-based methods. As Foster et al. (2018) point out, designing offline-regression-oracle-based algorithms is a promising direction for making contextual bandits practical, as they seem to combine the advantages of both realizability-based and agnostic algorithms: they are general and flexible enough to work with any given function class, while using a more realistic and reasonable oracle than the computationally-expensive classification oracle. Indeed, according to multiple experiments and extensive empirical evaluations conducted by Bietti et al. (2018) and Foster et al. (2018), the algorithm of Foster et al. (2018) "works the best overall" among existing contextual bandit approaches.

Despite its empirical success, the algorithm of Foster et al. (2018) is, however, theoretically sub-optimal — it could incur $\widetilde{\Omega}(T)$ regret in the worst case. Whether the optimal regret of contextual bandits can be attained via an offline-regression-oracle-based algorithm is listed as an open problem in Foster et al. (2018). In fact, this problem has been open to the bandit community since 2012 — it dates back to Agarwal et al. (2012), where the authors propose a computationally *inefficient* contextual bandit algorithm that achieves the optimal $O(\sqrt{KT \log |\mathcal{F}|})$ regret for a general *finite* function class $\mathcal{F}$, but leave designing computationally tractable algorithms as an open problem.

More recently, Foster and Rakhlin (2020) propose an algorithm that achieves the optimal regret for contextual bandits assuming access to an *online regression oracle* (which is *not* an offline optimization oracle and has to work with an adaptive adversary). Their finding that contextual bandits can be completely reduced to online regression is novel and important, and their result is also very general: it requires only the minimal realizability assumption, and holds true even when the contexts are chosen adversarially. However, compared with access to an offline regression oracle, access to an online regression oracle is a much stronger (and relatively restrictive) assumption. In particular, optimal and efficient algorithms for online regression are only known for specific function classes, much less than those known for offline regression. Whether the optimal regret of contextual bandits can be attained via a reduction to an offline regression oracle is listed as an open problem again in Foster and Rakhlin (2020).

1.1 Our Contributions

In this paper, we study the following open question which is repeatedly mentioned in the contextual bandit literature (Agarwal et al. 2012; Foster et al. 2018; Foster and Rakhlin 2020):

Is there an offline-regression-oracle-based algorithm that achieves the optimal regret for general contextual bandits?

Our paper provides the first resolution to this problem. Specifically, we provide the first optimal black-box reduction from contextual bandits to offline regression, with only the minimal realizability assumption. The significance of this result is that it reduces contextual bandits, a prominent online decision-making problem, to offline regression, a very basic and common offline optimization task that serves as the building block of modern machine learning. A direct consequence of this result is that any advances in solving offline regression problems immediately translate to contextual bandits, statistically and computationally. Note that such online-to-offline reduction is highly nontrivial (and impossible without specialized structures) for online learning problems in general (Hazan and Koren 2016).

Our reduction is accomplished by a surprisingly fast and simple algorithm that achieves the optimal $O(\sqrt{KT\log(|\mathcal{F}|T)})$ regret for general finite function class $\mathcal{F}$ with only $O(\log T)$ calls to an offline least-squares regression oracle over T rounds. The number of oracle calls can be further reduced to $O(\log\log T)$ if T is known. Notably, this can be understood as a "triply exponential" speedup over previous work: (1) compared with the previously known regret-optimal algorithm Agarwal et al. (2012) for this setting, which requires enumerating over $\mathcal{F}$ at each round, our algorithm accesses the function class only through a least-squares regression oracle, thus typically avoids an exponential cost at each round; (2) compared with the classification-oracle-based algorithm of Agarwal et al. (2014) which requires $\widetilde{O}(\sqrt{KT/\log|\mathcal{F}|})$ calls to a computationally expensive classification oracle, our algorithm requires only $O(\log T)$ calls to a simple regression oracle, which implies an exponential speedup over all existing provably optimal oracle-efficient algorithms, even when we ignore the difference between regression and classification oracles; (3) when the number of rounds T is known in advance, our algorithm can further reduce the number of oracle calls to $O(\log\log T)$, which is an exponential speedup by itself. Our algorithm is thus highly practical.

The statistical analysis of our algorithm is also quite interesting. Unlike existing analysis of other realizability-based algorithms in the literature, we do not directly analyze the decision outcomes of our algorithm — instead, we find a dual interpretation of our algorithm as sequentially maintaining a *dense* distribution over *all* (possibly *improper*) policies, where a policy is defined as a deterministic decision function mapping contexts to actions. We analyze how the realizability assumption enables us to establish uniform-convergence-type results for some *implicit* quantities in the universal policy space, regardless of the huge capacity of the universal policy space. Note that while the dual interpretation itself is not easy to compute in the universal policy space, it is only applied for the purpose of analysis and has nothing to do with our original algorithm's implementation. Through this lens, we find that our algorithm's dual interpretation satisfies a series of sufficient conditions for optimal contextual bandit learning. Our identified sufficient conditions for optimal

contextual bandit learning in the universal policy space are built on the previous work of Dudik et al. (2011), Agarwal et al. (2012) and Agarwal et al. (2014) — the first one is colloquially referred to as the "monster paper" by its authors due to its huge complexity (link), and the third one is titled as "taming the monster" by its authors due to its improved computational efficiency. Since our algorithm achieves all the conditions required for regret optimality in the universal policy space in a completely *implicit* way (which means that all the requirements are automatically satisfied without explicit computation), our algorithm comes with significantly reduced computational cost compared with previous work (thanks to the realizability assumption), and we thus title our paper as "bypassing the monster".

Finally, we extend all the above results to the case of general infinite function class $\mathcal{F}$. For this case, we state our results in a more general way: for any function class $\mathcal{F}$, given an arbitrary offline regression oracle with an arbitrary offline *estimation error* guarantee, we propose a fast and simple contextual bandit algorithm whose regret can be bounded by a function of the offline estimation error, through only $O(\log T)$ calls (or $O(\log \log T)$ calls if T is known) to the offline regression oracle. We show that our algorithm is statistically optimal as long as the offline regression oracle is statistically optimal. Notably, the above results provide a universal and optimal "converter" from results of offline/batch learning with general function classes to results of contextual bandits with general function classes. This leads to improved algorithms with tighter regret bounds for many existing contextual bandit problems, as well as practical algorithms for many new contextual bandit problems (e.g., contextual bandits with neural networks).

Overall, our algorithm is fast and simple, and our analysis is quite general. We believe that our algorithm has the potential to be implemented on a large scale, and our approach may contribute to deeper understanding of contextual bandits.

The reminder of the paper is omitted in this extended abstract. Please see https://arxiv.org/abs/2003.12699 for the full version of this paper.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320, 2011.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *AISTATS*, pages 19–26, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *ICML*, pages 1638–1646, 2014.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pages 127–135, 2013.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, pages 208–214, 2011.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, page 169–178, 2011.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *NIPS*, pages 586–594, 2010.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *ICML*, pages 1539–1548, 2018.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *STOC*, pages 128–141, 2016.

Adam Klivans and Alexander Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

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

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1): 1–96, 2018.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.