Proceedings of Machine Learning Research vol 75:1–5, 2018 31st Annual Conference on Learning Theory

Online learning over a finite action set with limited switching

Jason Altschuler JASONALT@MIT.EDU
Massachusetts Institute of Technology

Kunal Talwar KUNAL@GOOGLE.COM
Google Brain

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the value of switching actions in the Prediction From Experts (PFE) problem and Adversarial Multi-Armed Bandits (MAB) problem. First, we revisit the well-studied and practically motivated setting of PFE with switching costs. Many algorithms are known to achieve the minimax optimal order of $O(\sqrt{T \log n})$ in *expectation* for both regret and number of switches, where T is the number of iterations and n the number of actions. However, no *high probability* guarantees are known. Our main technical contribution is the first algorithms which with high probability achieve this optimal order for both regret and number of switches. This settles an open problem of (Devroye et al., 2015), directly implies the first high probability guarantees for several problems of interest, and is efficiently adaptable to the related problem of online combinatorial optimization with limited switching.

Next, to investigate the value of switching actions at a more granular level, we introduce the setting of *switching budgets*, in which the algorithm is limited to $S \leq T$ switches between actions. This entails a limited number of free switches, in contrast to the unlimited number of expensive switches allowed in the switching cost setting. Using the above result and several reductions, we unify previous work and completely characterize the complexity of this switching budget setting up to small polylogarithmic factors: for both the PFE and MAB problems, for all switching budgets $S \leq T$, and for both expectation and high probability guarantees. For PFE, we show that the optimal rate is of order $\tilde{\Theta}(\sqrt{T \log n})$ for $S = \Omega(\sqrt{T \log n})$, and $\min(\tilde{\Theta}(\frac{T \log n}{S}), T)$ for $S = O(\sqrt{T \log n})$. Interestingly, the bandit setting does not exhibit such a phase transition; instead we show the minimax rate decays steadily as $\min(\tilde{\Theta}(\frac{T\sqrt{n}}{\sqrt{S}}), T)$ for all ranges of $S \leq T$. These results recover and generalize the known minimax rates for the (arbitrary) switching cost setting.

1. Introduction

Two classical problems in online learning are the *Prediction From Experts (PFE)* problem (Cesa-Bianchi et al., 1997; Cesa-Bianchi and Lugosi, 2006) and the Adversarial *Multi-Armed Bandit (MAB)* problem (Auer et al., 2002; Bubeck et al., 2012). These problems have received substantial attention due to their ability to model a variety of problems in machine learning, sequential decision making, online combinatorial optimization, online linear optimization, mathematical finance, and many more.

PFE and MAB are T-iteration repeated games between an algorithm (often called player or forecaster) and an adversary (often called nature). In each iteration $t \in \{1, \ldots, T\}$, the

. *Extended abstract. Full version appears as arXiv:1803.01548v2.*

algorithm selects an action i_t out of n possible actions, while the adversary simultaneously chooses a loss function over the actions $\ell_t : \{1, \ldots, n\} \to [0, 1]$. The algorithm then suffers the loss $\ell_t(i_t)$ for its action. The goal of the algorithm is to minimize its cumulative loss $\sum_{t=1}^{T} \ell_t(i_t)$ over the course of the game. Since the losses are at the adversary's disposal, one measures the cumulative loss of the algorithm against a more meaningful baseline: the cumulative loss of the *best action in hindsight*. The algorithm's *regret* is defined as the difference between these two quantities:

$$\text{Regret} := \sum_{t=1}^{T} \ell_t(i_t) - \min_{i^* \in [n]} \sum_{t=1}^{T} \ell_t(i^*)$$

The PFE and MAB problems differ in the feedback that the algorithm receives. In PFE, the algorithm is given *full-information feedback*: after the tth iteration it can observe the entire loss function ℓ_t. However in MAB, the algorithm is only granted *bandit feedback*: after the tth iteration, it can only observe the loss $\ell_t(i_t)$ of the action i_t it played.

Switching as a resource. Note that in the setup of PFE and MAB above, the algorithm can play a different action in each time step. In many applications, switching between different actions too often is undesirable. This motivates the idea of switching as a resource. This notion has attracted significant research interest in the past few years. The popular way to formalize this idea is the *c-**switching-cost*** setting, in which the algorithm incurs an additional loss of $c \geq 1$ each time it switches actions in consecutive iterations. We introduce the *S-**switching-budget*** setting, in which the algorithm can switch at most $S \in \{1, \ldots, T\}$ times in the T iterations. In words, the *switching-cost setting corresponds to expensive but unlimited switches*; whereas the *switching-budget setting corresponds to free but limited switches*. In this setting it can be shown that we cannot be competitive with respect to an *adaptive* adversary, and thus we focus on the *oblivious* adversary model, where the loss functions cannot depend on the algorithm's choices.

1.1. Previous work

Previous work on Prediction from Experts (details in Figure 1). In the classical (unconstrained) setting, the minimax regret $\Theta(\sqrt{T \log n})$ is well understood (Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Cesa-Bianchi et al., 1997). Moreover, this optimal regret rate is also achievable with high probability (Cesa-Bianchi and Lugosi, 2006).

The minimax rate is also well-understood in the c-switching cost setting. Recall that here the objective is "switching-cost-regret", which is defined as Regret $+ c \cdot$ (# switches). The minimax rate for expected switching-cost-regret is $\Theta(\sqrt{cT \log n})$ for PFE (Kalai and Vempala, 2005; Geulen et al., 2010; Devroye et al., 2015). In particular, these results give algorithms which achieve the optimal minimax order in *expectation* for both regret and number of switches. However, *no high-probability guarantees are known for switching-cost PFE*; this is raised as an open question by (Devroye et al., 2015).

For the S-switching-budget setting, even less is known. The best lower bound seems to be the unconstrained regret lower bound of $\Omega(\sqrt{T \log n})$. An upper bound of $O(\frac{T}{\sqrt{S}})$ follows from the Lazy Label Efficient Forecaster (Cesa-Bianchi et al., 2005). Existing minimax-optimal switching-cost algorithms such as Follow the Perturbed Leader (Kalai and

Vempala, 2005), Shrinking Dartboard (Geulen et al., 2010), and Prediction by Random Walk Perturbation (Devroye et al., 2015) do not apply to the switching-budget setting (even in expectation), since the number of times they switch is only bounded *in expectation*.

Previous work on Multi-Armed Bandits (details in Figure 2). In the unconstrained setting, the minimax rate $\Theta(\sqrt{Tn})$ is well understood (Auer et al., 2002; Audibert and Bubeck, 2010) and is achieveable with high probability (Audibert and Bubeck, 2010; Bubeck et al., 2012). For the c-switching cost setting, the minimax rate is known (up to a logarithmic factor in T) to be $\tilde{\Theta}(c^{1/3}T^{2/3}n^{1/3})$ for MAB (Arora et al., 2012; Dekel et al., 2014).

For the S-switching budget setting, a simple mini-batching reduction gives algorithms achieving the minimax rate in expectation and with high probability. Dekel et al. (2014) prove a lower bound of $\tilde{\Omega}(\frac{T}{\sqrt{S}})$ via a reduction to the switching-cost setting. However, this reduction does not get the correct dependence on the number of actions n and also loses track of polylogarithmic factors.

Table 1: Upper and lower bounds on the complexity of PFE in the different switching settings. Our new bounds are bolded.

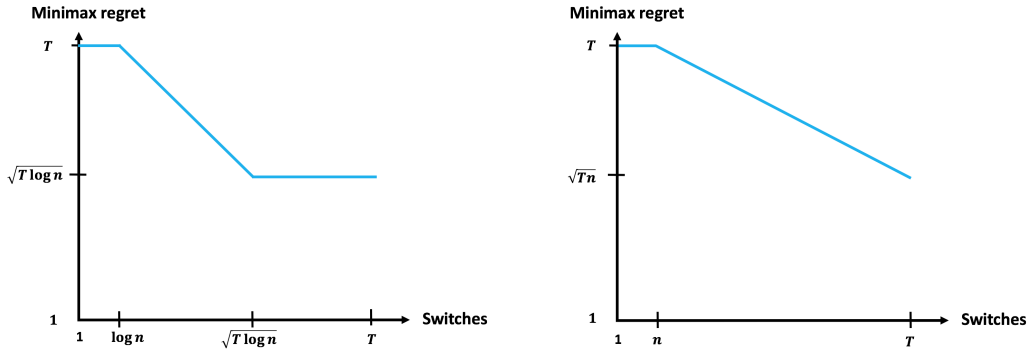
	LB on $\mathbb{E}$[Regret]	**UB on $\mathbb{E}$[Regret]**	**High prob. UB**
Unconstrained switching	$\sqrt{T\log n}$	$\sqrt{T\log n}$	$\sqrt{T\log\frac{n}{\delta}}$
c **switching cost**	$\sqrt{cT\log n}$	$\sqrt{cT\log n}$	$\boldsymbol{\sqrt{cT\log n\log\frac{1}{\delta}}}$
$S=\Omega(\sqrt{T\log n})$ **switching budget**	$\sqrt{T\log n}$	$\boldsymbol{\sqrt{T\log n}\log T}$	$\boldsymbol{\sqrt{T\log n}\log\frac{1}{\delta}}$
$S=O(\sqrt{T\log n})$ **switching budget**	$\boldsymbol{\frac{T\log n}{S}}$	$\boldsymbol{\frac{T\log n}{S}\log T}$	$\boldsymbol{\frac{T\log n}{S}\log\frac{1}{\delta}}$

Table 2: Upper and lower bounds on the complexity of MAB in the different switching settings. Our new bounds are bolded.

	LB on $\mathbb{E}$[Regret]	**UB on $\mathbb{E}$[Regret]**	**High prob. UB**
Unconstrained switching	$\sqrt{Tn}$	$\sqrt{Tn}$	$\sqrt{Tn}\frac{\log\frac{n}{\delta}}{\sqrt{\log n}}$
c **switching cost**	$\frac{c^{1/3}T^{2/3}n^{1/3}}{\log T}$	$c^{1/3}T^{2/3}n^{1/3}$	$c^{1/3}T^{2/3}n^{1/3}\frac{\log^{2/3}\frac{n}{\delta}}{\log^{1/3}n}$
S **switching budget**	$\boldsymbol{\frac{T\sqrt{n}}{\sqrt{S}\log^{3/2}T}}$	$\frac{T\sqrt{n}}{\sqrt{S}}$	$\frac{T\sqrt{n}}{\sqrt{S}}\frac{\log\frac{n}{\delta}}{\sqrt{\log n}}$

1.2. Our contributions

We present the first algorithms for switching-cost PFE that achieve the minimax optimal rate $O(\sqrt{cT\log n})$ with high probability. In fact, our results are more general: we give a framework to formulaically convert algorithms that work in expectation and fall under the Follow-the-Perturbed-Leader algorithmic umbrella, into algorithms that work with high probability. We also show how this framework extends to online combinatorial optimization,



(a) Switching-budget PFE. (b) Switching-budget MAB.

Figure 1: Complexity landscape of online learning over a finite action set with limited switching. Axes are plotted in log-log scale. Polylogarithmic factors in T are hidden for simplicity.

i.e. online linear optimization over a combinatorial polytope, where offline optimization can be done efficiently.

We also investigate the *switching budget* setting for the PFE and MAB problems. The above result and standard reductions allow us to completely characterize the complexity of this switching budget setting up to small polylogarithmic factors: for both the PFE and MAB problems, for all switching budgets $S \leq T$, and for both expectation and high probability guarantees. For PFE, we show the optimal rate is of order $\tilde{\Theta}(\sqrt{T \log n})$ for $S = \Omega(\sqrt{T \log n})$, and $\min(\tilde{\Theta}(\frac{T \log n}{S}), T)$ for $S = O(\sqrt{T \log n})$. Interestingly, the bandit setting does not exhibit such a phase transition; instead we show the minimax rate decays steadily as $\min(\tilde{\Theta}(\frac{T\sqrt{n}}{\sqrt{S}}), T)$ for all ranges of $S \leq T$.

.

2. Acknowledgements.

We are indebted to Elad Hazan for numerous fruitful discussions and for suggesting the switching-budget setting to us. We also thank Yoram Singer, Tomer Koren, David Martins, Vianney Perchet, and Jonathan Weed for helpful discussions.

Part of this work was done while JA was visiting the Simons Institute for the Theory of Computing, which was partially supported by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant #CCF-1740425. JA is also supported by NSF Graduate Research Fellowship 1122374.

References

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *ICML*, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1): 1–122, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44 (3):427–485, 1997.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467. ACM, 2014.

Luc Devroye, Gábor Lugosi, and Gergely Neu. Random-walk perturbations for online combinatorial optimization. *IEEE Transactions on Information Theory*, 61(7):4099–4106, 2015.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Sascha Geulen, Berthold Vöcking, and Melanie Winkler. Regret minimization for online buffering problems using the weighted majority algorithm. In *COLT*, pages 132–143, 2010.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.