A Dimension-free Algorithm for Contextual Continuum-armed Bandits

Preprint · July 2019

3 authors:

Wenhao Li
University of Technology Sydney

Ningyuan Chen
University of Toronto

L. Jeff Hong
Fudan University

A Dimension-free Algorithm for Contextual Continuum-armed Bandits

Wenhao Li
Department of Management Sciences
City University of Hong Kong
wenhaoli6-c@my.cityu.edu.hk

Ningyuan Chen
Department of Industrial Engineering & Decision Analytics
The Hong Kong University of Science and Technology
nychen@ust.hk

L.Jeff Hong
School of Management and School of Data Science
Fudan University
hong_liu@fudan.edu.cn

Abstract

In contextual continuum-armed bandits, the contexts x and the arms y are both continuous and drawn from high-dimensional spaces. The payoff function to learn $f(x,y)$ does not have a particular parametric form. The literature has shown that for Lipschitz-continuous functions, the optimal regret is $\tilde{O}(T^{\frac{d_x+d_y+1}{d_x+d_y+2}})$, where d_x and d_y are the dimensions of contexts and arms, and thus suffers from the curse of dimensionality. We develop an algorithm that achieves regret $\tilde{O}(T^{\frac{d_x+1}{d_x+2}})$ when f is globally concave in y. The global concavity is a common assumption in many applications. The algorithm is based on stochastic approximation and estimates the gradient information in an online fashion. Our results generate a valuable insight that the curse of dimensionality of the arms can be overcome with some mild structures of the payoff function.

1 Introduction

The multi-armed bandits (MAB) deal with a class of sequential decision making problems [Lai and Robbins, 1985, Auer et al., 2002a, Bubeck et al., 2012]. Without knowing the payoff of each decision, the decision maker chooses a decision from a set of alternatives (arms) in each epoch based on the past history. The observed random payoff associated with the chosen decision can be used to learn in future epochs. The goal is to maximize the total payoff over a finite horizon. The MAB setting has been introduced in Robbins [1952] and studied intensively since then in statistics, computer sciences, operations research, and economics.

Recently, the contextual bandits problems have received attentions of many scholars. Before making decision in each epoch, the decision maker receives a context that can be used to infer the payoff and suggest which arm to pull. Contextual bandits are motivated by advertisement placement on webpages. Upon observing the user profile (context), the firm needs to decide which advertisement to place (arm) that may interest the user. The number of clicks is the payoff to maximize.

Preprint. Under review.

In this paper, we consider contextual continuum-armed bandits. Compared to the classic MAB setting, both decision and context are drawn from continuous spaces in our problem. This is motivated by personalized pricing in operations research and marketing. A firm sells multiple products over a finite selling season via dynamically adjusted prices. For each coming customer, the firm observes her profile such as education background, zip code, age and purchasing history. Then the firm decides a personalized price vector for the customer. The total payoff is the revenue extracted from a finite number of customers. The continuous nature and high dimensionality of the customer profile and pricing motivate the contextual continuum-armed bandit problem.

Prior Work. There is a body of literature on continuum-armed bandits [Agrawal, 1995, Pandey et al., 2007, Kleinberg and Slivkins, 2010, Maillard and Munos, 2010]. Kleinberg [2005] studies the case that the mean payoff function satisfies a Hölder continuous property with constant α. This work proves a worst-case lower bound $O(T^{\frac{\alpha+1}{2\alpha+1}})$ for the regret of any algorithm when the arms set is one dimensional and proposes a uniform discretization algorithm achieving a regret of order $\tilde{O}(T^{\frac{\alpha+1}{2\alpha+1}})$, nearly tight with the lower bound. Auer et al. [2007] also study the one-dimensional arm set. Under the condition that payoff functions have finitely many maxima and a non-vanishing, continuous second-order derivative at all maxima, their algorithm achieves the regret order $\tilde{O}(\sqrt{T})$. Kleinberg et al. [2008] consider the multi-dimensional case and generalize the Lipschitz bandit problem to metric space. They present an algorithm obtaining the regret $\tilde{O}(T^{\frac{d+1}{d+2}})$ where d is the covering dimension of arm space, kindly capturing the sparsity of arm space. The same regret is achieved by Bubeck et al. [2011], but d is the packing dimension instead. They further demonstrate that the smoothness of the mean payoff function around its maximum can be used to reduce the packing dimension. Regret bounds independent of the dimension of the arm space are obtained by Cope [2009], Agarwal et al. [2013]. Cope [2009] shows an asymptotic regret bound of size $O(\sqrt{T})$ if the payoff functions are unimodal, three times continuously differentiable and its derivative is well behaved at its maximal. Agarwal et al. [2013] assume globally convex and Lipschitz payoff functions and achieving a regret $\tilde{O}(poly(d)\sqrt{T})$ with high probability.

Another stream of related literature is contextual bandits. Woodroofe [1979], Wang et al. [2005], Rigollet and Zeevi [2010], Perchet and Rigollet [2013] study contextual MAB with stochastic payoffs, under the name *bandits with covariates*: the context is a random variable correlated with the payoffs. They consider the case of finitely discrete arms. On the other hand, Slivkins [2014], Lu et al. [2009] consider continuous arms and assume Lipschitz continuity for both the arm and context space. They prove a lower bound $O(T^{\frac{\hat{d}_x+\hat{d}_y+1}{\hat{d}_x+\hat{d}_y+2}})$ for the regret of any algorithm where $\hat{d}_x, \hat{d}_y$ are the packing dimensions of the context and arms space respectively. Lu et al. [2009] presents a uniformly partition algorithm obtaining nearly tight regret upper bound $\tilde{O}(T^{\frac{d_x+d_y+1}{d_x+d_y+2}})$ where d_x, d_y are covering dimensions of the context and arms space. The same regret bound can be achieved by the uniform partition and a bandit-with-expert-advice algorithm such as EXP4 [Auer et al., 2002b] or NEXP [McMahan and Streeter, 2009]. The uniform partition is used to define an expert whose advise is simply an arbitrary arm for each set of the partition. Slivkins [2014] proposes an adaptive zooming algorithm so that frequently occurring contexts and high-paying arms structure can be used to improve practical performance. There other versions of contextual bandit problems, such as linear bandits [Auer, 2002, Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011], contextual bandits with policy sets [Auer et al., 2002b, Langford and Zhang, 2008, Agarwal et al., 2012, Dudik et al., 2011].

In operations research, many papers have focused on dynamic pricing and demand learning [Besbes and Zeevi, 2009, 2015, Broder and Rusmevichientong, 2012, den Boer and Zwart, 2013, den Boer, 2015]. Recently Chen and Gallego [2018] consider personalized pricing of a single product to customers. Besides Lipschitz continuity in arms and context space, they further assume smoothness and local concavity at the unique maximizer of the payoff function. Their algorithm achieves the near-optimal regret in their setting, $\tilde{O}(T^{\frac{d_x+2}{d_x+4}})$, slightly better than the $\tilde{O}(T^{\frac{d_x+d_y+1}{d_x+d_y+2}})$ when $d_y = 1$. In a recent paper, Chen and Shi [2019] consider multi-product pricing problem with inventory constraints. Similar to our setting, they assume global concavity and propose an algorithm which also depends on the online learning of gradients and achieves the regret bound of $\tilde{O}(T^{\frac{4}{5}})$. The regret is independent of the dimension of the arm space, confirming the insights provided in this paper. However, the rate of regret bound does not seem to be optimal and the contextual information is not considered.

Main results and contributions. According to Slivkins [2014] and references therein, the optimal regret for the contextual continuum-armed bandits is $\tilde{O}(T^{\frac{d_x+d_y+1}{d_x+d_y+2}})$, when the payoff function $f(x,y)$ is Lipschitz continuous and d_x, d_y are the dimensions of the context and arm space. After imposing the structure property that $f(x,y)$ is globally concave in the decision variable y when fixing x, we provide an algorithm that achieves the regret $O(d_y^{\frac{d_x+6}{2(d_x+2)}} T^{\frac{d_x+1}{d_x+2}})$. Compared to the previous bound, the dimensionality d_y does not increase the regret exponentially when T increases. The mitigation of the curse of the dimensionality can improve the performance of the algorithm significantly in practice. For example, in the context-free setting ($d_x = 0$), the regret of a ten-dimensional decision variable ($d_y = 10$) is $\tilde{O}(T^{11/12})$ for previous algorithms, while a mere $\tilde{O}(\sqrt{T})$ for our algorithm. On the other hand, global concavity is a mild assumption, which is commonly assumed in various applications. Therefore, the improvement in regret does not come with a substantial sacrifice in the generality of the formulation.

The algorithm is based on binning the contextual space and applying stochastic gradient descent or stochastic approximation in each bin to learn the optimal decision. Such algorithms are popular in machine learning [Bottou, 2010, Shalev-Shwartz and Srebro, 2008, Shalev-Shwartz et al., 2009, Duchi and Singer, 2009]. In the case of concave functions, the classic algorithms in the learning literature such as UCB or Thompson sampling fail to take into account the special structure and do not seem to be able to achieve the optimal regret. Instead, gradient-based algorithms, which do not perform well for general functions due to multiple local maxima, can guarantee a surprising dimension-free regret in our setting. Our results thus convey the message that problem and domain-specific algorithmic design for learning problems may be helpful and beneficial.

2 Problem Formulation

The domain of the unknown payoff function $f(x,y)$ is $x \in \mathcal{X} \triangleq [0,1]^{d_x}$ and $y \in \mathcal{Y} \triangleq [0,1]^{d_y}$. One can interpret $\mathcal{X}$ and $\mathcal{Y}$ to be the normalization of some compact sets. Let $\mathcal{T} = \{1, 2, \ldots, T\}$ denote the sequence of decision epochs faced by the decision maker. At the beginning of each epoch $t \in \mathcal{T}$, the contextual information $x_t \in \mathcal{X}$ is revealed to the decision maker. The contextual information is drawn independently from some unknown distribution[1] and therefore denoted by X_t. Then the decision maker chooses an arm y_t in $\mathcal{Y}$. The payoff in epoch t is a random variable Z_t, whose mean is $\mathbb{E}[Z_t | X_t, y_t] = f(X_t, y_t)$. We require Z_t to be independent across epochs given X_t and y_t.

Regret. If the payoff function were known, then the optimal decision and average payoff given context x are

$$y^*(x) \triangleq \arg\max_{y \in \mathcal{Y}} f(x,y), \quad f^*(x) = \max_{y \in \mathcal{Y}} f(x,y),$$

which is referred to as the oracle. Since the decision maker does not have access to the unknown function, the total payoff is always lower than that of the oracle in expectation. A standard performance metric of an algorithm is defined as the regret incurred compared to the oracle.

$$R(T) \triangleq \sum_{t=1}^{T} \mathbb{E}\left[f^*(X_t) - f(X_t, y_t)\right]. \tag{1}$$

Note that the decision made in epoch t, y_t, is also random, even though the decision maker does not use active randomization. This is because at each epoch t, the decision maker may rely on the information revealed so far to make decisions. Therefore, y_t may depend on the observed contexts $\{X_s\}_{s=1}^{t}$, the adopted decisions $\{y_s\}_{s=1}^{t-1}$ and the realized payoffs $\{Z_s\}_{s=1}^{t-1}$. Since f is unknown to the decision maker, the objective is thus to design an algorithm that achieves small regret for a wide class of functions as $T \to \infty$. One can expect that if f is an arbitrary function such as discontinuous ones, then no algorithm can achieve small regret. Next, we specify the assumptions that f has to satisfy.

[1] Slivkins [2014] assumes that the context arrivals x_t are fixed before the first round. We follow Perchet and Rigollet [2013] and assume that X_t are i.i.d.

2.1 Assumptions

We now present a set of assumptions in our setting, which are required to guarantee the existence and good behavior of the gradient estimates. They are not required by Lipschitz bandits [Slivkins, 2014]. Most assumptions are mild.

Assumption 1 (Twice differentiable)**.** *For all $x \in \mathcal{X}$, the function $f(x, \cdot)$ is twice continuously differentiable on the arm space $\mathcal{Y}$.*

Besides the existence of gradient, We assume strong concavity to ensure the global convergence of gradient descent and the uniqueness of the optimal solution.

Assumption 2 (Strong concavity)**.** *There exists a constant $M_1 > 0$ such that $f(x, y_1) \le f(x, y_2) + (y_1 - y_2)^T \frac{\partial}{\partial y} f(x, y_2) - \frac{1}{2} M_1 \|y_1 - y_2\|_2^2$ for all $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$.*

An immediate implication of Assumption 2 is the unique optimal solution $y^*(x) = \arg\max_{y \in \mathcal{Y}} f(x, y)$ for any context x. The following assumption makes sure that $y^*(x)$ is in the interior of $\mathcal{Y}$, which implies that $\frac{\partial}{\partial y} f(x, y^*(x)) = 0$.

Assumption 3 (Interior optimal solution)**.** *For any $x \in \mathcal{X}$, the optimal solution $y^*(x) = \arg\max_{y \in \mathcal{Y}} f(x, y)$ satisfies $y^*(x) \in int(\mathcal{Y})$.*

Assumption 3 is imposed mainly for technical purposes. In practice, one may also extend the domain of y to ensure an interior optimal solution. The next assumption imposes some regularity (Hölder condition) on the context space.

Assumption 4 (Hölder continuity of the context)**.** *For every $y \in \mathcal{Y}$, the function $f(\cdot, y)$ is Hölder continuous in $\mathcal{X}$, i.e. $|f(x_1, y) - f(x_2, y)| \le M_2 \|x_1 - x_2\|_2^\alpha$ with constant $M_2 > 0$ and $0 < \alpha \le 1$.*

Hölder continuity is a generalization of Lipschitz continuity. It is easy to see that for $\alpha = 1$, it is equivalent to Lipschitz continuity. A Similar condition is also imposed in Perchet and Rigollet [2013]. The next assumption makes sure that the random payoff Z behaves normally, which is standard in the literature.

Assumption 5 (Finite second moment)**.** *For any given x and y, the random payoff Z has finite second moment, i.e., there exists a uniform constant $M_3 > 0$ such that $\mathbb{E}\left[Z^2 | x, y\right] \le M_3$.*

At the beginning of the horizon, the following information is revealed to the decision maker: the domain of the context $\mathcal{X}$ and the decision variable $\mathcal{Y}$, the length of the horizon T, and the constant M_1 defined in the Assumption 2.

3 Our Algorithm

There are two components of our algorithm. To deal with the contexts, we partition the context space into rectangular bins. When the partition is designed carefully, we are able to conduct context-free learning in each of the bin without significantly increasing the regret. That is, treat the contexts generated in the same bin equally. This idea is also adopted by Lu et al. [2009], Rigollet and Zeevi [2010], Perchet and Rigollet [2013], Chen and Gallego [2018]. To find the optimal solution $y^*(x)$ when the context falls into a particular bin, we use stochastic approximation and the estimated gradient to find the maximum. Next we elaborate on the two components separately.

Binning the context space. Discretization and local approximations are probably the most popular method to deal with nonparametric estimation. Utilizing Assumption 4, one can expect that $f(x_1, \cdot)$ and $f(x_2, \cdot)$ tend to behave similarly, including close maximal values and optimal solutions, when $\|x_1 - x_2\|_2$ is small. Following this intuition, we partition the context space as follows. We divide each of the d_x dimensions of the context space into K equal intervals. As a result, the context space $\mathcal{X}$ divided into K^{d_x} identical hyper-rectangles, referred to as bins. The partition $\mathcal{B}_K \triangleq \{B_1, \dots, B_{K^{d_x}}\}$ is thus a collection of bins of the following form: for $k = (k_1, \dots, k_{d_x}) \in \{1, \dots, K\}^{d_x}$,

$$B_k = \left\{ x \in \mathcal{X} : \frac{k_l - 1}{K} \le x_l \le \frac{k_l}{K}, l = 1, \dots, d_x \right\}. \tag{2}$$

The algorithm thus keeps track of K^{d_x} independent learning sub-problems. When a context is generated in B_k at some epoch t, the exact location of X_t is no longer used as long as the knowledge of B_k is preserved. The sub-problem k then proceeds with one more epoch while the other sub-problems remain the same. Therefore, the sub-problem associated with B_k is equivalent to a classic continuum-armed bandit problem without contextual information.

One can clearly see the trade-off in choosing a proper granularity of discretization, represented by K. When K is too small, the algorithm aggregates too much information into a bin and loses accuracy; when K is too large, then there are too many bins and the learning horizon is short for each sub-problem. Later, we will choose an optimal K to balance the trade-off and obtain small regret.

Stochastic Approximation (SA). To solve the sub-problem in a particular bin $B \in \mathcal{B}_K$, the algorithm treats all contextual information equally as long as a context falls into B. In this case, the oracle for the context-free problem can obtain the average payoff

$$f_B(y) = \mathbb{E}\left[f(X,y)|X \in B\right] \tag{3}$$

with the following optimal solution

$$y^*(B) = \arg\max_{y\in\mathcal{Y}} f_B(y). \tag{4}$$

We develops an algorithm based on stochastic approximation (see Kushner and Yin [2003], Chau and Fu [2015] for a comprehensive review) to find $y^*(B)$. The basic idea is demonstrated below: Suppose at epoch t a context X_t is generated in bin B. A decision y_t is chosen and a random reward $f(X_t, y_t)$ is observed. After a number of epochs, another context $X_{t'}$ is observed in B for some $t' > t$. If the gradient information of $f_B(y)$ at y_t is known, then stochastic approximation could be used to determine $y_{t'}$. In particular,

$$y_{t'} = y_t - a_t \nabla f_B(y_t).$$

For a properly chosen step size a_t, one can show that y_t converges to the optimal solution $y^*(B)$.

However, there are two pitfalls of when applying SA directly. First, $y_{t'}$ may be outside the domain $\mathcal{Y}$. This issue can be addressed by projecting $y_{t'}$ back to $\mathcal{Y}$, denoted by the operator $\Pi_{\mathcal{Y}}$. Second, the function f is not known to the decision maker, not to mention the gradient ∇f_B. Thus, we need to estimate the gradient from noisy payoffs Z_t. We use the Kiefer-Wolfowitz (KW) algorithm [Kiefer et al., 1952] as a subroutine. After applying y_t and observing Z_t, the decision maker explores the neighborhood of y_t and uses a finite-difference method to estimate the gradient. More precisely, suppose the contexts generated at $t_{d_y} > t_{d_y-1} > \cdots > t_1 > t$ fall into the same bin B. Our algorithm applies decision $y_t + ce_i$ in epoch t_i, where e_i is the unit vector with the ith entry equal to 1 and 0 elsewhere. The step size c will be specified later. The payoff Z_{t_i} can thus be regarded as an estimator for $f_B(y_t + ce_i)$. After epoch t_{d_y}, the KW algorithm suggests an estimator for $\nabla f_B(y_t)$

$$\hat{\nabla} f_B(y_t) = \frac{1}{c}\left[Z_{t_1} - Z_t, Z_{t_1} - Z_t, \ldots, Z_{t_{d_y}} - Z_t\right]^T \tag{5}$$

After d_y more contexts generated in B, the algorithm finally moves along the direction of the estimated gradient. Suppose in epoch $t' > t_{d_y}$, the context $X_{t'} \in B$. Then the decision $y_{t'}$ is chosen according to

$$y_{t'} = \Pi_{\mathcal{Y}}\left(y_t - a_t \hat{\nabla} f_B(y_t)\right).$$

Another d_y contexts need to be observed in B in order to estimate $\nabla f_B(y_{t'})$. The algorithm associated with a single bin is demonstrated in Algorithm 1. To simplify the notation, we only focus on the epochs when the contexts generated are in the same bin and re-order the index by $t = 1, 2, 3, \ldots$.

Next we combine the two components for the contextual continuum-armed bandit problem. From the above description, we keep track of K^{d_x} instances of Algorithm 1 and updates the number of epochs in each bin separately. The detailed steps are demonstrated in Algorithm 2.

Remark. *Technically speaking, the finite difference $\tilde{y}_n + c_n e_m$ in Step 8 of Algorithm 1 may be outside the domain $\mathcal{Y}$ and the algorithm needs to adjust for that. In that case, we let $y_t \to \tilde{y}_n - c_n e_m$ which must be inside the domain $\mathcal{Y}$. And then replace the corresponding difference $Z_{t-i} - Z_{t-d_y-i}$ in $G(y_{n-1})$ by its opposite. After that, all the following analysis remains the same.*

Algorithm 1 Online KWSA algorithm in one bin

Input: pre-defined step size sequences a_n, c_n
Initialize: $\tilde{y}_0 \in \mathcal{Y}$
$y_1 \leftarrow \tilde{y}_0$; observe Z_1
for $t = 2, 3, \ldots$ **do**
  $m \leftarrow (t-1) \mod d_y + 1$
  $n \leftarrow (t-1-m)/(d_y+1)$ {In epoch t, the algorithm has estimated the gradient n times}
  if $m \neq 0$ **then**
    $y_t \leftarrow \tilde{y}_n + c_n e_m$ {Use finite difference to estimate the gradient}
  else
    $G(y_{n-1}) \leftarrow \frac{1}{c_{n-1}} \left[Z_{t-1} - Z_{t-d_y-1}, Z_{t-2} - Z_{t-d_y-1}, \ldots, Z_{t-d_y} - Z_{t-d_y-1}\right]^T$
    $\tilde{y}_n \leftarrow \Pi_{\mathcal{Y}} \left(\tilde{y}_{n-1} + a_{n-1} G(y_{n-1})\right)$ {Apply KWSA}
    $y_t \leftarrow \tilde{y}_n$
  end if
  Observe Z_t
end for

Algorithm 2 KWSA with binning

Input: T, M_1
Tunable parameters: K {K is the number of bins}
Partition the context space into $\mathcal{B}_K$ as in (2)
for $t = 1, 2, \ldots, T$ **do**
  Observe context X_t
  $B \leftarrow \{B \in \mathcal{B}_K : X_t \in B\}$ {Determine the bin which X_t belongs to}
  Apply Algorithm 1 associated with bin B
end for

4 Analysis of the Regret

In this section, we provide a roadmap of the analysis. We first bound the regret incurred in a single bin.

Proposition 1. *Let $a_n = an^{-1}$ and $c_n = \delta n^{-\frac{1}{4}}$, where $1/(4M_1) < a < 1/(2M_1)$ and $\delta > 0$. Then the regret of Algorithm 1 in bin B satisfies*

$$\mathbb{E}\left[f_B(y^*(B)) - f_B(y_t)\right] \leq \frac{\sqrt{d_y} Q(d_y)}{\sqrt{t}} \tag{6}$$

where Q is a linear function of d_y whose coefficients are independent of t. Specifically,

$$Q(d_y) = M_5 \delta^2 + M_5 \max\left\{\mathbb{E}\left[\|\tilde{y}_0 - y^*(B)\|_2^2\right], \left(\frac{2\delta M_5 + \sqrt{4\delta^2 M_5^2 + 8 d_y a^2 M_3^2 (4aM_1 - 1)/\delta^2}}{4\delta M_5 - 1}\right)^2\right\}$$

Proposition 1 uses the standard convergence results of KWSA. However, we need to analyze the property of $f_B(y)$ in (3) carefully. In particular, the assumptions imposed in Section 2.1 are for the function $f(x, y)$. First, we translate the assumptions in Section 2.1 of $f(x, y)$ to obtain other crucial properties, including Lipschitz continuity, Lipschitz-continuous gradient and negative-definite Hessian matrix. Second, we prove the interchangeability of expectation and differentiation of $f(x, y)$ to make sure the properties are extended to $f_B(y)$. Third, with the regulairty conditions of $f_B(y)$, we apply the asymptotic analysis in the stochastic approximation literature to derive finite time bound of Algorithm 1. More precisely, the left-hand side of (6) can be bounded by $\|y^*(B) - y_t\|_2^2$, because of bounded eigenvalues of the Hessian matrix. Then we bound $\|y^*(B) - y_t\|_2^2$ by a decreasing sequence with convergence rate $t^{-\frac{1}{2}}$.

Remark. *According to Proposition 1, when there is not contextual information, Algorithm 1 achieves a bound $O(d_y^{3/2}\sqrt{T})$ for continuum-armed bandits. The problem is studied before in the literature and we compare to their results below. Cope [2009] shows a similar regret bound $O(\sqrt{T})$*

asymptotically. We relax the thrice-continuously-differentiable assumption in their paper and provide a finite-sample bound. Bubeck et al. [2011] find that if the smoothness parameter of the payoff functions around the maxima were known, then the near-optimal regret $\tilde{O}(\sqrt{T})$ could be achieved, independent of the dimension of the arm space. We do not require a certain degree of smoothness for the payoff function; instead, global convexity/concavity is imposed. In practice, knowing whether the unknown payoff function is convex seems to be more reasonable than knowing how smooth the function is. In a similar setting, Agarwal et al. [2013] propose an algorithm whose high-probability regret bound is $\tilde{O}(poly(d)\sqrt{T})$. We eliminate the logarithmic terms in the bound and obtain a bound in expectation.

From Proposition 1, we know that the regret incurred in one bin is bounded by $O(d_y^{3/2}\sqrt{T})$ if there are T epochs to learn in that bin. This is because summing up $1/\sqrt{t}$ leads to

$$2(\sqrt{T}-1) = \int_1^T \frac{1}{\sqrt{t}}dt \le \sum_{t=1}^{T} \frac{1}{\sqrt{t}}dt = 1 + \sum_{t=2}^{T} \frac{1}{\sqrt{t}} \le 1 + \int_1^T \frac{1}{\sqrt{t}}dt = 2\sqrt{T}-1.$$

To analyze the regret incurred by Algorithm 2, we need to aggregate the regret incurred in all the bins in the partition. Moreover, the optimal solution $y^*(B)$ for the context-free problem in a bin is still not as good as the oracle $y^*(X_t)$. We expect to bound $f(X_t, y^*(X_t)) - f(X_t, y^*(B))$ by the size the bin and the continuity of $f(x, y)$. We choose $K = O(T^{\frac{1}{d_x+2\alpha}})$ in Algorithm 2.

Theorem 1. *For any function f satisfying the assumptions in Section 2.1, the regret by Algorithm 2 is bounded by*

$$R(T) \le C d_y^{\frac{\alpha(d_x+6)}{2(d_x+2\alpha)}} T^{\frac{d_x+\alpha}{d_x+2\alpha}} \tag{7}$$

for a constant C that is independent of d_x, d_y, T.

For the most common case of Lipschitz functions, we let $\alpha = 1$ and the regret bound becomes

$$\tilde{O}(d_y^{\frac{(d_x+6)}{2(d_x+2)}} T^{\frac{d_x+1}{d_x+2}}).$$

It recovers the regret bound in Lipschitz bandit [Slivkins, 2014] with $d_y = 0$. Also notice the fact that when $d_x \ge 2$, $d_y^{\frac{(d_x+6)}{2(d_x+2)}} \le d_y$. Therefore, no matter how large the dimension of the decision variable d_y is, the dependence of the regret on d_y is at most linear. Compared to the exponential dependence (i.e., $\tilde{O}(T^{\frac{d_x+d_y+1}{d_x+d_y+2}})$) in the literature, the mild dependence makes our algorithm more suitable for problems with high-dimensional decision variables. The significantly improved regret comes at the cost of a more restrictive form of $f(x, y)$, in particular, it has to be globally concave in y. The additional assumption is still reasonable in various applications. Moreover, the algorithm eliminates the logarithmic terms commonly seen in the literature.

The main steps of the proof are described below. First, the regret incurred in epoch t, $\mathbb{E}[f^*(X_t) - f(X_t, y_t)]$ is bounded by a constant multiplied by the mean square error $\mathbb{E}\left[\|y_t - y^*(X_t)\|_2^2\right]$, because of the global convexity. The distance between y_t and $y^*(X_t)$ is further bounded by $\|y_t - y^*(B)\|_2$ and $\|y^*(B) - y^*(X_t)\|_2$, where the bound of the first term is implied by Proposition 1. For the second term, the discretization error incurred by binning, is bounded by the diameter of bin B. Second, after deriving the regret incurred in one epoch, the bound of total regret could be obtained by summing up the regret in all bins. The worst case in terms of the regret is when the covariates are generated evenly in the bins, and the best case is when the covariates are generated in a single bin. Therefore, suppose each of the K^{d_x} bins observes T/K^{d_x} covariates and we can compute the aggregate regret for this worst case. Finally, we need to minimize the regret over the number of bins K. The tunable parameter K can be regarded as a balance between exploration and exploitation. When the bin is large, there is enough observations in the bin that the selected arm is close to its optimum. However, due to the large diameter of the bin, its optimum may be far away from the optimum respect to one specific covariate in the bin. On the other hand, when the bin is small, the distance between the optimum of the bin and optimum of one covariate is quite small, but the arms chosen by Algorithm 1, may be far from the bin's optimum. To balance the trade-off and obtain smallest regret, the number of bins is chosen to be $K = O(T^{\frac{1}{d_x+2\alpha}})$.

Remark. *Another type of stochastic approximation is referred to as the Robbins-Monro [Robbins and Monro, 1951] algorithm. Different from KW, RM algorithm requires an oracle to* *return the unbiased estimators for the gradient of $f_B(y)$ for any given y. The unbiased estimator can be used to replace $\hat{\nabla} f_B(y_{t-1})$ in (5). As a result, the convergence rate of RM is better than KW and the regret bound in Theorem 1 can be further improved. However, the information of an unbiased estimator for the gradient is a somewhat unrealistic scenario, and thus we do not present the regret for the RM algorithm in this paper.*

5 Conclusion and Future Work

In this paper, we study the continuum-armed bandit problem under contextual information, where the context space and arms space are both continuous. After assuming the curvature of payoff functions, strong convexity and second-order smoothness, we propose a novel method combining stochastic approximation with bining partition framework and obtain a much better regret than existing literature. Surprisingly, our method achieves a dimension-free result.

In the future, we will investigate how to reduce the effect incurred by context space. In other words, whether the curvature conditions corresponding to covariates can be utilized to improve the regret bound. If so, another nonparametric estimation framework other than binning partition is required to solve the curse of dimensionality problem. It's an open problem that how to incorporate nonparametric statistic learning methods in reducing the growth rate of the regret with respect to the dimension of context space.

6 Appendix

6.1 Proof of Proposition 1

At a high level the proof first extends assumptions in section 2 to more properties of mean payoff function f in Lemma 1. Then we prove theses assumptions and properties are also satisfied by the payoff function f_B in Lemma 3. Finally, Proposition 1 can be proved using these properties.

Lemma 1 (Properties of $f(x, y)$)**.** *According to assumptions 1-3, we have that,*

(1) *$f(x, y)$ is Lipschitz continuous in y with a constant M_4, i.e. $|f(x, y_1) - f(x, y_2)| \leq M_4 \|y_1 - y_2\|_2$.*

(2) *$\frac{\partial}{\partial y} f(x, y)$ is Lipschitz continuous in y with a constant M_5, i.e. $\|\frac{\partial}{\partial y} f(x, y_1) - \frac{\partial}{\partial y} f(x, y_2)\| \leq M_5 \|y_1 - y_2\|_2$.*

(3) *For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the Hessian matrix $\frac{\partial^2}{\partial y^2} f(x, y)$ is negative definite and all the eigenvalues are in the interval $[-M_5, -M_1]$, i.e. $-M_5 I \preceq \frac{\partial^2}{\partial y^2} f(x, y) \preceq -M_1 I$.*

(4) *For every $x \in \mathcal{X}$, the function $f(x, y)$ has a unique maximizer $y^*(x) \in int(\mathcal{Y})$, i.e. there exists a unique $y^*(x) = \arg\max_{y \in \mathcal{Y}} f(x, y)$ and $\frac{\partial}{\partial y} f(x, y^*(x)) = 0$.*

Proof of Lemma 1.

(1) Since $\frac{\partial}{\partial y} f(x, \cdot)$ is a continuous function on a convex set $\mathcal{Y}$, there exists a constant M_1 s.t $\|\frac{\partial}{\partial y} f(x, \cdot)\|_2 \leq M_1$. Then by generalized mean value theorem (Theorem 9.19 in Rudin et al. [1964]), $|f(x, y_1) - f(x, y_2)| \leq M_1 \|y_1 - y_2\|_2$ for all $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$.

(2) Since $\frac{\partial^2}{\partial y^2} f(x, \cdot)$ is a continuous function on a convex set $\mathcal{Y}$, there exists a constant M_5 s.t $\|\frac{\partial^2}{\partial y^2} f(x, \cdot)\|_2 \leq M_5$. Then by generalized mean value theorem, $\|\frac{\partial}{\partial y} f(x, y_1) - \frac{\partial}{\partial y} f(x, y_2)\|_2 \leq M_5 \|y_1 - y_2\|_2$ for all $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$.

(3) Notice that $f(x, y)$ is a strongly concave function, for every $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$,

$$f(x, y_1) \leq f(x, y_2) + (y_1 - y_2)^T \frac{\partial}{\partial y} f(x, y_2) - \frac{1}{2} M_1 \|y_1 - y_2\|_2^2$$

By second order Taylor expansion, there exists a $y_0 = \beta y_1 + (1-\beta)y_2,\ \beta \in [0,1]$ such that,

$$f(x,y_1) = f(x,y_2) + (y_1-y_2)^T \frac{\partial}{\partial y} f(x,y_0) + \frac{1}{2}(y_1-y_2)^T \frac{\partial^2}{\partial y^2} f(x,y_0)(y_1-y_2)$$

Thus,

$$\frac{1}{2}(y_1-y_2)^T \frac{\partial^2}{\partial y^2} f(x,y_0)(y_1-y_2) \leq -\frac{1}{2}M_1\|y_1-y_2\|_2^2$$

and

$$\frac{\partial^2}{\partial y^2} f(x,y_0) \preceq -M_1 I$$

Again by twice continuous differentiability of $f_y(x,\cdot)$, we have

$$\frac{\partial^2}{\partial y^2} f(x,y_0) \to \frac{\partial^2}{\partial y^2} f(x,y)\ as\ y_0 \to y$$

So for all $y \in \mathcal{Y}$,

$$\frac{\partial^2}{\partial y^2} f(x,y) \preceq -M_1 I$$

Recall that $\frac{\partial^2}{\partial y^2} f(x,\cdot)$ is a continuous function on a convex set $\mathcal{Y}$ and $\|\frac{\partial^2}{\partial y^2} f(x,\cdot)\|_2 \leq M_5$. Then for any $y \in \mathcal{Y}$,

$$-M_5 I \preceq \frac{d^2}{dy^2} f_B(y) \preceq -M_1 I,\ M_1 \leq \|\frac{d^2}{dy^2} f_B(y)\|_2 \leq M_5$$

(4) Assuming there exists $y_1 \neq y_2$ satisfying

$$f(x,y_1) = f(x,y_2) = \max_{y\in\mathcal{Y}} f(x,y)$$

By the strong concavity of $f(x,y)$,

$$2f(x,\frac{y_1+y_2}{2}) \geq f(x,y_1) + f(x,y_2) = 2\max_{y\in\mathcal{Y}} f(x,y)$$

which contradicts with the definition of $y^*(x)$. Thus there only exists one maximizer $y^*(x)$. Next, if $\frac{\partial}{\partial y} f(x,y^*(x)) \neq 0$, we can find a small step l such that $y' = y^*(x) + l\frac{\partial}{\partial y} f(x,y^*(x)) \in \mathcal{Y}$. Then $f(x,y') > f(x,y^*(x))$ which contradicts with the definition of $y^*(x)$. So $\frac{\partial}{\partial y} f(x,y^*(x)) = 0$.

Before proving the smoothness and convexity conditions of f_B, we first provide condition for interchangeability of expectation and derivative in Lemma 2.

Lemma 2 (Pathwise Method)**.** *Assume for every x, $f(x,\cdot)$ is differentiable on $y \in \mathcal{Y}$ and Lipschitz continuous with constant M_1. Then $\frac{\partial}{\partial y}\mathbb{E}_X[f(X,y)] = \mathbb{E}_X[\frac{\partial}{\partial y} f(X,y)]$.*

Proof of Lemma 2.

Considering one certain dimension y_i, if we can prove $\frac{\partial}{\partial y_i}\mathbb{E}_X[f(X,y)] = \mathbb{E}_X[\frac{\partial}{\partial y_i} f(X,y)]$, then Lemma 2 is obvious.

$$\begin{aligned}
\frac{\partial}{\partial y_i}\mathbb{E}_X[f(X,y)] &= \lim_{h\to 0}\frac{\mathbb{E}_X[f(X,y+he_i)] - \mathbb{E}_X[f(X,y)]}{h} \\
&= \lim_{h\to 0}\mathbb{E}_X\left[\frac{f(X,y+he_i) - f(X,y)}{h}\right] \\
&= \mathbb{E}_X\left[\lim_{h\to 0}\frac{f(X,y+he_i) - f(X,y)}{h}\right] \\
&= \mathbb{E}_X\left[\frac{\partial}{\partial y_i} f(X,y)\right]
\end{aligned}$$

In the first equality, e_i denotes the unit vector with the i-th entry 1. The third equality is supported by dominated convergence theorem, where $\dfrac{|f(X,y+he_i) - f(X,y)|}{h} \leq M_5$.

Lemma 3 (Optimal arm in hypercube)**.** *For any hypercube $B \subset \mathcal{X}$, including a singleton $B = \{x\}$, define the function $f_B(y) \triangleq \mathbb{E}\left[f(X,y)|X \in B\right]$. According to assumptions 1-3 ,we have that for any B,*

(1) *$f_B(y)$ is twice continuously differentiable on the convex set $y \in \mathcal{Y}$. Additionally, expectation and gradient are exchangeable, i.e. $\nabla f_B(y) = \mathbb{E}[\frac{\partial}{\partial y} f(X,y)|X \in B]$ and $\frac{d^2}{dy^2} f_B(y) = \mathbb{E}\left[\frac{\partial^2}{\partial y^2} f(X,y)|X \in B\right]$.*

(2) *$f_B(y)$ is strongly concave in y with the same constant of $f(x,y)$, i.e. $f_B(y_1) \leq f_B(y_2) + (y_1 - y_2)^T \nabla f_B(y_2) - \frac{1}{2} M_1 \|y_1 - y_2\|_2^2$ for all $y_1, y_2 \in \mathcal{Y}$*

(3) *$f_B(y)$ maintains Lipschitz continuous property of $f(x,y)$ with the same constant, i.e. $|f_B(y_1) - f_B(y_2)| \leq M_4 \|y_1 - y_2\|_2$ for all $y_1, y_2 \in \mathcal{Y}$*

(4) *$f_B(y)$ maintains Lipschitz gradient property of $f(x,y)$ with the same constant, i.e. $\|\nabla f_B(y_1) - \nabla f_B(y_2)\|_2 \leq M_5 \|y_1 - y_2\|_2$ for all $y_1, y_2 \in \mathcal{Y}$.*

(5) *For all $y \in \mathcal{Y}$, the Hessian matrix of $f_B(y)$ is negative definite and all the eigenvalues are in the interval $[-M_5, -M_1]$, i.e. $-M_5 I \preceq \frac{d^2}{dy^2} f_B(y) \preceq -M_1 I$.*

(6) *The function $f_B(y)$ has a unique maximizer $y^*(B) \in int(\mathcal{Y})$, i.e. there exists a unique $y^*(B) = \arg\max_{y \in int(\mathcal{Y})} \mathbb{E}\left[f(X,y)|X \in B\right]$, and $\nabla f_B(y^*(B)) = 0$.*

Proof of Lemma 3.

(1) Since $f(\cdot, y)$ are differentiable and $f(\cdot, y)$ are Lipschitz continuous with constant M_1, according to pathwise method, $\nabla f_B(y) = \frac{\partial}{\partial y} \mathbb{E}[f(X,y)|X \in B] = \mathbb{E}[\frac{\partial}{\partial y} f(X,y)|X \in B]$. $\nabla f_B(y)$ exists for the reason that $\frac{\partial}{\partial y} f(x,y)$ exists for any x. Similarly, as $\frac{\partial}{\partial y} f(\cdot, y)$ are differentiable and Lipschitz continuous with constant M_5, we have

$$\begin{aligned}
\frac{d^2}{dy^2} f_B(y) &= \frac{\partial^2}{\partial y^2} \mathbb{E}\left[f(X,y)|X \in B\right] \\
&= \frac{\partial}{\partial y} \mathbb{E}\left[\frac{\partial}{\partial y} f(X,y)|X \in B\right] \\
&= \mathbb{E}\left[\frac{\partial^2}{\partial y^2} f(X,y)|X \in B\right]
\end{aligned}$$

The continuity of $\frac{d^2}{dy^2} f_B(y)$ is the consequence of $\frac{\partial^2}{\partial y^2} f(x,y)$ continuous in y.

(2) Strong concavity is obvious because expectation operator maintains linear relationship.

$$\begin{aligned}
f_B(y_1) &= \mathbb{E}[f(X,y_1)|X \in B] \\
&\leq \mathbb{E}\left[f(X,y_2) + (y_1 - y_2)^T \frac{\partial}{\partial y} f(X,y_2) - \frac{1}{2} M_1 \|y_1 - y_2\|_2^2 \Big| X \in B\right] \\
&= f_B(y_2) + (y_1 - y_2)^T \nabla f_B(y_2) - \frac{1}{2} M_1 \|y_1 - y_2\|_2^2
\end{aligned}$$

(3)

$$\begin{aligned}
|f_B(y_1) - f_B(y_2)| = |\mathbb{E}\left[f(X,y_1)|X \in B\right] - \mathbb{E}\left[f(X,y_2)|X \in B\right]| &\leq \mathbb{E}\left[|f(X,y_1) - f(X,y_2)| \Big| X \in B\right] \\
&\leq \mathbb{E}\left[M_4 \|y_1 - y_2\|_2\right] = M_4 \|y_1 - y_2\|_2
\end{aligned}$$

(4)

$$\begin{aligned}
\|\nabla f_B(y_1) - \nabla f_B(y_2)\|_2 &= \left\| \mathbb{E}\left[\frac{\partial}{\partial y} f(X,y_1)|X \in B\right] - \mathbb{E}\left[\frac{\partial}{\partial y} f(X,y_2)|X \in B\right] \right\|_2 \\
&\leq \mathbb{E}\left[\left\| \frac{\partial}{\partial y} f(X,y_1) - \frac{\partial}{\partial y} f(X,y_2) \right\|_2 |X \in B\right] \\
&\leq \mathbb{E}\left[M_5 \|y_1 - y_2\|_2\right] \\
&\leq M_5 \|y_1 - y_2\|_2
\end{aligned}$$

(5) Since $f_B(y)$ has the twice continuous differentiable and strong concave property of $f(x, y)$, this property can be proved following the same way of Lemma 1(3).

(6) Assuming there exist $y_1 \neq y_2$ satisfying

$$f_B(y_1) = f_B(y_2) = \max_{y \in \mathcal{Y}} \mathbb{E}\left[f(X, y)|X \in B\right]$$

By the strong concavity of $f(x, y)$,

$$2f(x, \frac{y_1 + y_2}{2}) > f(x, y_1) + f(x, y_2)$$

Thus,

$$\begin{aligned} 2f_B\left(\frac{y_1 + y_2}{2}\right) &= 2\mathbb{E}\left[f(X, \frac{y_1 + y_2}{2})|X \in B\right] \\ &> \mathbb{E}\left[f(X, y_1)|X \in B\right] + \mathbb{E}\left[f(X, y_2)|X \in B\right] = 2f_B(y^*(B)) \end{aligned}$$

which contradicts with the definition of $y^*(B)$. Thus there is only one maximizer $y^*(B)$. To prove the second part $\nabla f_B(y^*(B)) = 0$, we define that $y^*(B) = \left[y_1^*(B), \cdots, y_{d_y}^*(B)\right]^T$ and $y^*(x) = \left[y_1^*(x), \cdots, y_{d_y}^*(x)\right]^T$. Recall that $\mathcal{Y}$ is a convex set $[0, 1]^{d_y}$, the lower bounds and upper bounds in each dimension are denoted by $[y_1^l, y_1^h], \cdots, [y_{d_y}^l, y_{d_y}^h]$. In each dimension i, we define the minimum distance between $y_i^*(x)$ and y_i^l as $\delta_i^l = \inf_{x \in B}\left|y_i^*(x) - y_i^l\right|$. Similarly, $\delta_i^h = \inf_{x \in B}\left|y_i^*(x) - y_i^h\right|$. $y^*(x) \in int(\mathcal{Y})$, then $\delta_i^l \geq 0$ and $\delta_i^h \geq 0$. Considering that $f(x, y)$ is a strongly concave function in y, $\dfrac{\partial f(x, y)}{\partial y_i} \geq 0$ for $y_i \in \left[y_i^l, y_i^l + \delta_i^l\right]$ and $x \in B$ and $\dfrac{\partial f(x, y)}{\partial y_i} \leq 0$ for $y_i \in \left[y_i^h - \delta_i^h, y_i^h\right]$. So $\dfrac{\partial f_B(y)}{\partial y_i} \geq 0$ for $y_i \in \left[y_i^l, y_i^l + \delta_i^l\right]$ and $\dfrac{\partial f_B(y)}{\partial y_i} \leq 0$ for $y_i \in \left[y_i^h - \delta_i^h, y_i^h\right]$. Recall that $\dfrac{\partial f_B(y)}{\partial y_i}$ is a continuous function. Then, there exists $y_i^* \in \left[y_i^l, y_i^h\right]$ such that $\dfrac{\partial f_B(y)}{\partial y_i}\Big|_{y_i = y_i^*} = 0$.

Lemma 4. *Suppose the sequence b_n satisfies*

$$b_{n+1} \leq \left(1 - \frac{\alpha}{n}\right) b_n + \beta n^{-\frac{5}{4}} \sqrt{b_n} + \omega n^{-\frac{3}{2}}$$

Let $\lambda = \max\{b_1, \lambda_0\}$, where

$$\lambda_0 = \left(\frac{\beta + \sqrt{\beta^2 + 2\omega(2\alpha - 1)}}{2\alpha - 1}\right)^2$$

Then, by induction, we have $b_n \leq \lambda n^{-\frac{1}{2}}$.

Proof of Lemma 4.

We prove by induction. It is easy to see that it hold for $n = 1$. For any $n = 1, 2, \cdots$, suppose that $b_n \leq \lambda n^{-\frac{1}{2}}$. Because $1 - \alpha/n > 0$ due to $\alpha < 1$,

$$\begin{aligned} b_{n+1} &\leq \left(1 - \frac{\alpha}{n}\right) \lambda n^{-\frac{1}{2}} + \beta\sqrt{\lambda} n^{-\frac{3}{2}} + \omega n^{-\frac{3}{2}} \\ &= \lambda n^{-\frac{1}{2}} - \left(\alpha\lambda - \beta\sqrt{\lambda} - \omega\right) n^{-\frac{3}{2}} \\ &= \lambda n^{-\frac{1}{2}} - \frac{\lambda}{2} n^{-\frac{3}{2}} - \frac{1}{2}\left[(2\alpha - 1)\lambda - 2\beta\sqrt{\lambda} - 2\omega\right] n^{-\frac{3}{2}} \\ &\leq \lambda\left(n^{-\frac{1}{2}} - \frac{1}{2} n^{-\frac{3}{2}}\right) \end{aligned}$$

where the last inequality follows form the definition of λ_0 that satisfies $(2\alpha-1)z-2\beta\sqrt{z}-2\omega\leq 0$ for any $z\geq\lambda_0$. Let $g(x)=x^{-\frac{1}{2}}$. Then, $g'(x)=-\frac{1}{2}x^{-\frac{3}{2}}$. Notice that $g(x)$ is convex. Then,

$$g(x')-g(x)\geq g'(x)(x'-x)$$

Then,

$$(n+1)^{-\frac{1}{2}}-n^{-\frac{1}{2}}=g(n+1)-g(n)\geq g'(n)=-\frac{1}{2}n^{-\frac{3}{2}}$$

Therefore,

$$n^{-\frac{1}{2}}-\frac{1}{2}n^{-\frac{3}{2}}\leq(n+1)^{-\frac{1}{2}}$$

Then, we have $b_{n+1}\leq\lambda(n+1)^{-\frac{1}{2}}$. This concludes the induction proof. □

Proof of Proposition 1. According to the dimension of arm space, the time epochs is divided into periods: period 1=$\{1,\ldots,d_y\}$, period 2=$\{d_y+1,\ldots,2d_y\}$ last period $=\lfloor T/(d_y+1)\rfloor(d_y+1),\ldots,T$. In each period, the gradient is estimated by finite-difference exactly once at the first epoch. Let n denote the number of period, which is also the gradient estimation times. Then let $b_n:=\mathbb{E}\left(\|\tilde{y}_n-y^*(B)\|_2^2\right)$. Notice that $\Pi_{\mathcal{Y}}(y^*(B))=y^*(B)$ and $\|\Pi_{\mathcal{Y}}(\tilde{y}_{n+1})-\Pi_{\mathcal{Y}}(\tilde{y}_n)\|_2\leq\|\tilde{y}_{n+1}-\tilde{y}_n\|_2$, then

$$\begin{aligned}b_{n+1}&=\mathbb{E}\left\{\|\Pi_{\mathcal{Y}}(\tilde{y}_n+a_nG(\tilde{y}_n))-y^*(B)\|_2^2\right\}\\&=\mathbb{E}\left\{\|\Pi_{\mathcal{Y}}(\tilde{y}_n+a_nG(\tilde{y}_n))-\Pi_{\mathcal{Y}}(y^*(B))\|_2^2\right\}\\&\leq\mathbb{E}\left\{\|\tilde{y}_n+a_nG(\tilde{y}_n)-y^*(B)\|_2^2\right\}\\&=b_n+a_n^2\mathbb{E}\left(\|G(\tilde{y}_n)\|_2^2\right)+2a_n\mathbb{E}\left[G(\tilde{y}_n)^T(\tilde{y}_n-y^*(B))\right]\end{aligned}\tag{8}$$

Let $g(\tilde{y}_n)=\frac{1}{c_n}\left([f(X_{n+1},y_n+c_ne_1)-f(X_n,\tilde{y}_n)],\cdots,[f(X_{n+d},y_n+c_ne_{d_y})-f(X_n,\tilde{y}_n)]\right)^T$ and $g_B(\tilde{y}_n)=\frac{1}{c_n}\left([f_B(y_n+c_ne_1)-f_B(\tilde{y}_n)],\cdots,[f_B(y_n+c_ne_{d_y})-f_B(\tilde{y}_n)]\right)^T$. Then,

$$\begin{aligned}\mathbb{E}\left[G(\tilde{y}_n)^T(\tilde{y}_n-y^*(B))\right]&=\mathbb{E}\left\{\mathbb{E}\left[G(\tilde{y}_n)^T(\tilde{y}_n-y^*(B))|\tilde{y}_n,X_n,X_{n+1},\ldots,X_{n+d_y}\in B\right]\right\}\\&=\mathbb{E}\left\{\mathbb{E}\left[g(\tilde{y}_n)^T|X_n,X_{n+1},\ldots,X_{n+d_y}\in B\right](\tilde{y}_n-y^*(B))\right\}\\&=\mathbb{E}\left[g_B(\tilde{y}_n)^T(\tilde{y}_n-y^*(B))\right]\\&=\mathbb{E}\left[\nabla f_B(\tilde{y}_n)^T(\tilde{y}_n-y^*(B))\right]+\mathbb{E}\left[(g_B(\tilde{y}_n)-\nabla f_B(\tilde{y}_n))^T(\tilde{y}_n-y^*(B))\right]\end{aligned}\tag{9}$$

The first equality follows from tower law, second from the definition of $F(x,y,\xi)$ and third from the definition of $f_B(y)$. According to strong concavity property of $f_B(y)$ (Lemma 3(2)), we have

$$f_B(\tilde{y}_n)\leq f_B(y^*(B))+(\tilde{y}_n-y^*(B))^T\nabla f_B(y^*(B))-\frac{1}{2}M_1\|\tilde{y}_n-y^*(B)\|_2^2$$
$$f_B(y^*(B))\leq f_B(\tilde{y}_n)+(y^*(B)-\tilde{y}_n)^T\nabla f_B(\tilde{y}_n)-\frac{1}{2}M_1\|\tilde{y}_n-y^*(B)\|_2^2$$

Add them together,

$$(\tilde{y}_n-y^*(B))^T(\nabla f_B(\tilde{y}_n)-\nabla f_B(y^*(B)))\leq-M_1\|\tilde{y}_n-y^*(B)\|_2^2$$

Note that strong concavity of $f_B(y)$ implies that maximizer $y^*(B)$ is unique. By optimality of $y^*(B)$, we have

$$(\tilde{y}_n-y^*(B))^T\nabla f_B(y^*(B))\leq 0$$

which together with last equation implies that $(\tilde{y}_n-y^*(B))^T\nabla f_B(\tilde{y}_n)\leq-M_1\|\tilde{y}_n-y^*(B)\|_2^2$. Taking expectation of both sides,

$$\mathbb{E}\left[\nabla f_B(\tilde{y}_n)^T(\tilde{y}_n-y^*(B))\right]\leq-M_1b_n\tag{10}$$

Then by Lemma 3(5),

$$\frac{1}{2}M_5c_n\mathbf{1}\geq g_B(\tilde{y}_n)-\nabla f_B(\tilde{y}_n)=\frac{1}{2}c_n\left(e_1^T\frac{d^2}{dy^2}f_B(\eta_1)e_1,\cdots,e_d^T\frac{d^2}{dy^2}f_B(\eta_d)e_d\right)\geq-\frac{1}{2}M_5c_n\mathbf{1}$$

Then by Cauthy-Schwarz inequality,

$$\mathbb{E}\left[(g_B(\tilde{y}_n) - \nabla f_B(\tilde{y}_n))^T(\tilde{y}_n - y^*(B))\right] \le \frac{1}{2}M_5 c_n \mathbb{E}\left(\|\tilde{y}_n - y^*(B)\|_2\right) \le \frac{1}{2}M_5 c_n \sqrt{b_n}$$

Thus,

$$\mathbb{E}\left[G(\tilde{y}_n)^T(\tilde{y}_n - y^*(B))\right] \le -M_1 b_n + \frac{1}{2}M_5 c_n \sqrt{b_n} \tag{11}$$

Furthermore, by Assumption 5,

$$\mathbb{E}\left(\|G(\tilde{y}_n)\|_2^2\right) \le \frac{4d_y M_3^2}{c_n^2} \tag{12}$$

Therefore, by equations (8),(9),(10),(11),(12),

$$b_{n+1} \le (1 - 2a_n M_1)b_n + a_n c_n M_5 \sqrt{b_n} + \frac{4d_y M_3^2 a_n^2}{c_n^2}$$

Suppose that $a_n = an^{-1}$ and $c_n = \delta n^{-\frac{1}{4}}$ with $1/(4M_1) < a < 1/(2M_1)$ and $\delta > 0$, we have

$$b_{n+1} \le (1 - \frac{2ac}{n})b_n + a\delta M_5 n^{-\frac{5}{4}}\sqrt{b_n} + \frac{4d_y a^2 M_3^2}{\delta^2}n^{-\frac{3}{2}}$$

Let $\alpha = 2aM_1, \beta = \alpha\delta M_5, \omega = \frac{4d_y a^2 M_3^2}{\delta^2}$, by induction (see Lemma 4), there exists $\lambda > 0$ such that

$$b_n \le \lambda n^{-\frac{1}{2}}$$

In each period, $d_y + 1$ arms, $\tilde{y}_n, \tilde{y}_{n+1}, \ldots, y_{n+d_y}$ required to be implemented. Notice that, by Lemma 3(6), $\nabla f_B(y^*(B)) = 0$ and Lemma 3(5), $\|\frac{d^2}{dy^2}f_B(y)\|_2 \le M_5$. Then by Taylor expansion,

$$f_B(y^*(B)) - f_B(\tilde{y}_n) \le \frac{1}{2}M_5\|\tilde{y}_n - y^*(B)\|_2^2$$

Taking expectation of both sides,

$$\mathbb{E}[f_B(y^*(B)) - f_B(\tilde{y}_n)] \le \frac{1}{2}M_5\mathbb{E}\left[\|\tilde{y}_n - y^*(B)\|_2^2\right] = \frac{1}{2}M_5\lambda n^{-\frac{1}{2}}$$

and for $i = 1, \ldots, d$,

$$\begin{aligned}
\mathbb{E}[f_B(y^*(B)) - f_B(\tilde{y}_n + c_n e_i)] &\le \frac{1}{2}M_5\mathbb{E}\left[\|\tilde{y}_n + c_n e_i - y^*(B)\|_2^2\right] \\
&\le M_5\mathbb{E}\left(\|\tilde{y}_n - y^*(B)\|_2^2 + c_n^2\right) \\
&\le M_5(\lambda + \delta^2)n^{-\frac{1}{2}}
\end{aligned}$$

Recall that λ is an affine function of d_y, then there exist a function $Q(d_y)$ such that

$$\mathbb{E}[f_B(y^*(B)) - f_B(y_t)] \le \frac{\sqrt{d_y}Q(d_y)}{\sqrt{t}}$$

for every epoch t. $\square$

6.2 Proof of Theorem 1.

We first prove a continuity result of maximizers in Lemma 5. It gives an error bound for the distance between the maximizer in one bin and optimum for a covariate. The error will disappear as the diameter of B shrinks to zero.

Lemma 5 (Hölder continuous of $y^*(x)$ and $y^*(B)$)**.** *For a hypercube $B \subset \mathcal{Y}$, the diameter of arms space $\mathcal{Y}$ is $\sqrt{d_y}$ and let d_B be the diameter of bin B. By Assumptions 1-4, there exists a uniform constant $M_6 > 0$ such that $\|y^*(B) - y^*(x)\|_2 \le M_6 d_B^{\alpha/2} = M_6 d_y^{\alpha/4} d_x^{\alpha/4} K^{-\alpha/2}$ for any $x \in B$.*

Proof of Lemma 5. From twice continuously differentiable and strongly concavity of function f, we have that

$$\frac{M_1}{2}\|y^*(x)-y^*(B)\|_2^2 \le f(x,y^*(x)) - f(x,y^*(B)) \le \frac{M_5}{2}\|y^*(x)-y^*(B)\|_2^2$$
$$\frac{M_1}{2}\|y^*(x)-y^*(B)\|_2^2 \le f_B(y^*(B)) - f_B(y^*(x)) \le \frac{M_5}{2}\|y^*(x)-y^*(B)\|_2^2$$

Then add them together,

$$\begin{aligned}
M_1\|y^*(x)-y^*(B)\|_2^2 &\le f(x,y^*(x)) - f_B(y^*(x)) + f_B(y^*(B)) - f(x,y^*(B)) \\
&= \mathbb{E}\left[f(x,y^*(x)) - f(X,y^*(x))\Big| X\in B\right] \\
&\quad + \mathbb{E}\left[f(X,y^*(B)) - f(x,y^*(B))\Big| X\in B\right] \\
&\le 2\mathbb{E}\left[M_2\|x-X\|_2^\alpha | X\in B\right] \\
&\le 2M_2 d_B^\alpha
\end{aligned}$$

where $d_B = \sqrt{d_y d_x}/K$. So we get the conclusion,

$$\|y^*(x)-y^*(B)\|_2 \le M_6 d_B^{\alpha/2} = M_6 d_y^{\alpha/4} d_x^{\alpha/4} K^{-\alpha/2}$$

Proof of Theorem 1. According to the algorithm, $\mathcal{B}_K$ denotes the partition formed by the bins. The regret associated with X_t can be counted by bins $B\in\mathcal{B}_K$ into which X_t falls. Therefore,

$$R(T) = \mathbb{E}\left[\sum_{t=1}^T (f^*(X_t) - f(X_t,\pi_t))\right] = \mathbb{E}\left[\sum_{t=1}^T \sum_{B\in\mathcal{B}_K} (f^*(X_t) - f(X_t,\pi_t))\, \mathbb{I}_{\{X_t\in B\}}\right]$$

According to Lemma 1(3), we have $f^*(X_t) - f(X_t,y_t) \le \frac{1}{2}M_5\|y_t - y^*(X_t)\|_2^2$. The distance between y_t and $y^*(X_t)$ can be bounded by $\|y_t - y^*(B)\|_2$ and $\|y^*(B) - y^*(X_t)\|_2$, where the first term is bounded by the error bound of stochastic approximation proved in Proposition 1 and the second term is bounded by the diameter of bin B using Lemma 5. Therefore,

$$\begin{aligned}
\mathbb{E}\left[(f^*(X_t) - f^*(X_t,y_t))\,\mathbb{I}_{\{X_t\in B\}}\right] &\le \frac{1}{2}M_5\mathbb{E}\left[\|y_t - y^*(X_t)\|_2^2 \mathbb{I}_{\{X_t\in B\}}\right] \\
&\le M_5\left\{\mathbb{E}\left[\|y_t-y^*(B)\|_2^2\right] + \mathbb{E}\left[\|y^*(B)-y^*(X_t)\|_2^2\right]\right\} \\
&\le M_5\mathbb{E}\left[\|y_t-y^*(B)\|_2^2\right] + M_5\left(M_6 d_y^{\alpha/4} d_x^{\alpha/4} K^{-\alpha/2}\right)^2 \\
&\le M_5\left[\frac{Q(d_y)}{\sqrt{t_B/(d_y+1)}}\right] + M_5 M_6^2 d_y^{\alpha/2} d_x^{\alpha/2} K^{-\alpha}
\end{aligned}$$

where $Q(d_y)$ is the function defined in Proposition 1 and t_B denotes observation times in B. After deriving the regret bound in one period, we can sum them together and obtain the bound of total regret.

Since B_K forms a partition of the covariate space and X always falls into one of the bins. The worst case is that all the covariates are uniformly distributed in the whole covariate space, for the regret incurred in each bin increases in the reciprocal order as observations increasing. Thus the total regret is less than the accumulative regret in K^{d_x} bins when covariates occur T/K^{d_x} times in each bin.

$$R(T) \le K^{d_x}(d_y+1)M_5 \sum_{t=1}^{\lceil T/[(d_y+1)K^{d_x}]\rceil} \left[Q_{d_y}/\sqrt{t}\right] + TM_5M_6^2 d_y^{\alpha/2} d_x^{\alpha/2} K^{-\alpha}$$

By the summation of series, $\sum_{t=1}^n 1/\sqrt{t} \le 1 + 2\sqrt{n}$, we have

$$\frac{R(T)}{M_5} \le K^{d_x}(d_y+1)Q_{d_y}\sqrt{1 + \frac{T}{(d_y+1)K^{d_x}}} + TM_5M_6^2 d_y^{\alpha/2} d_x^{\alpha/2} K^{-\alpha}$$

Using the Cauchy inequality,

$$\frac{R(T)}{M_5} \le K^{d_x}(d_y+1)Q_{d_y} + \sqrt{d_y+1}K^{d_x/2}Q_{d_y}\sqrt{T} + TM_5M_6^2 d_y^{\alpha/2} d_x^{\alpha/2} K^{-\alpha}$$

We separate in 2 cases to design the K that minimizes the total regret.

(1) If $T \ge K^{d_x}$, then

$$\frac{R(T)}{M_5} \le b_0 d_y^{3/2} K^{d_x/2}\sqrt{T} + b_1 d_y^{\alpha/2} d_x^{\alpha/2} K^{-\alpha} T$$

Hence, to minimize the total regret, by the choice of $K = \left(d_x^{\alpha-2} d_y^{\alpha-3} T\right)^{\frac{1}{d_x+2\alpha}}$ (satisfying $T \ge K^{d_x}$),

$$R(T) \le c_2 d_x^{\frac{(\alpha-2)d_x}{2(d_x+2\alpha)}} d_y^{\frac{\alpha(d_x+6)}{2(d_x+2\alpha)}} T^{\frac{d_x+\alpha}{d_x+2\alpha}}$$

(2) If $T \le K^{d_x}$, then

$$\frac{R(T)}{M_5} \le b_3 d_y^2 T + b_4 d_x^{\alpha/2} K^{-\alpha} T$$

Hence,

$$R(T) \le b_5 T$$

Hence, we choose $K = \left(d_x^{\alpha-2} d_y^{\alpha-3} T\right)^{\frac{1}{d_x+2\alpha}}$ and complete the proof of Theorem 1. $\square$

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

A. Agarwal, M. Dudík, S. Kale, J. Langford, and R. E. Schapire. Contextual bandit learning with predictable rewards. *arXiv preprint arXiv:1202.1334*, 2012.

A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

R. Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33 (6):1926–1951, 1995.

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

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

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

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.

O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

O. Besbes and A. Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 61(4):723–739, 2015.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

S. Bubeck, N. 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.

M. Chau and M. C. Fu. An overview of stochastic approximation. In *Handbook of Simulation Optimization*, pages 149–178. Springer, 2015.

N. Chen and G. Gallego. Nonparametric learning and optimization with covariates. *Working paper*, 2018.

Y. Chen and C. Shi. Network revenue management with online inverse batch gradient descent method. *Working paper*, 2019.

E. W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control*, 54(6):1243–1253, 2009.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, 2008.

A. den Boer. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

A. V. den Boer and B. Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management science*, 60(3):770–783, 2013.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

J. Kiefer, J. Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

R. Kleinberg and A. Slivkins. Sharp dichotomies for regret minimization in metric spaces. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 827–846. Society for Industrial and Applied Mathematics, 2010.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM, 2008.

R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.

H. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

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

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

T. Lu, D. Pál, and M. Pál. Showing relevant ads via context multi-armed bandits. Technical report, Tech. rep, 2009.

O.-A. Maillard and R. Munos. Online learning in adversarial lipschitz environments. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 305–320. Springer, 2010.

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

S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski. Bandits for taxonomies: A model-based approach. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 216–227. SIAM, 2007.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Annals of Statistics*, 41 (2):693–721, 2013.

P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. *COLT 2010*, page 54, 2010.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

W. Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

S. Shalev-Shwartz and N. Srebro. Svm optimization: inverse dependence on training set size. In *Proceedings of the 25th international conference on Machine learning*, pages 928–935. ACM, 2008.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.

A. Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

C.-C. Wang, S. R. Kulkarni, and H. V. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

M. Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.