The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Nonparametric Stochastic Contextual Bandits

Melody Y. Guan*
Stanford University
450 Serra Mall
Stanford, California 94305
mguan@stanford.edu

Heinrich Jiang*
Google
1600 Amphitheatre Pwky
Mountain View, California 94043
heinrich.jiang@gmail.com

Abstract

We analyze the K-armed bandit problem where the reward for each arm is a noisy realization based on an observed context under mild nonparametric assumptions. We attain tight results for top-arm identification and a sublinear regret of $\widetilde{O}\left(T^{\frac{1+D}{2+D}}\right)$, where D is the context dimension, for a modified UCB algorithm that is simple to implement. We then give global intrinsic dimension dependent and ambient dimension independent regret bounds. We also discuss recovering topological structures within the context space based on expected bandit performance and provide an extension to infinite-armed contextual bandits. Finally, we experimentally show the improvement of our algorithm over existing approaches for both simulated tasks and MNIST image classification.

Introduction

Multi-armed bandits (MABs) are an important sequential optimization problem introduced by Robbins (1985). These models have extensively been used in a wide variety of fields related to statistics and machine learning.

The classical MAB consists of K arms where at each point in time the learner can sample (or pull) one of them and observe a reward. Then various objectives can be established, such as finding the best arm (Top-Arm Identification) or minimizing some regret over time.

For contextual bandits (also referred to as bandits with side information or covariates), the learner has access to a context on which the payoffs depend. Then, based on the observations, we aim to determine the best policy (or context-to-arm mapping) and to optimize some notion of regret.

Most approaches to stochastic contextual bandits make strong assumptions on the payoffs. A popular approach models the mean reward for each arm as being linear in the context space (Chu et al. 2011; Li et al. 2010). However, this is rarely the case in real data. In this paper, we take a more general approach and allow the reward functions to be non-linear and of arbitrary shape.

Using recent developments in nonparametric statistics (Jiang 2017b), we show that with simple and easily implementable techniques, we can construct bandit algorithms which can learn over the entire context space with strong guarantees, despite the difficulty that arises with allowing a wide variety of reward functions. While this is not the first work which blends nonparametric statistics with bandits, we are the first to show simple and practical methods while still maintaining strong theoretical guarantees.

We reanalyze the uniform and upper confidence bound sampling strategies and demonstrate what nonparametric approaches can offer to contextual bandit learning. No other technique can adapt to the inherently difficult and complex real world reward functions while allowing such a strong theoretical understanding of the underlying algorithms.

While nonparametric models are powerful in their ability to learn arbitrary functions free of distributional assumptions, a major weakness is the curse of dimensionality. In order to have any theoretical guarantees, they require an exponential-in-dimension number of samples. However, when the data lies on an unknown low-dimensional structure such as a manifold, we show that our algorithms can converge as if the data was on a lower dimension and not in the potentially much large ambient dimension. Another striking fact is that no preprocessing of the data is required. This is of practical importance because modern data has increasingly more features but the underlying degrees of freedom often remain small.

We then discuss recovering geometric structures in the context space based on bandit performance. Specifically, we recover the connected components of the context space in which a particular bandit is the top-arm. Although learning a context-to-arm mapping gives us the estimated top-arm at each point in the context space, this alone does not tell the space's topological structure, such as the number and shapes of connected components. We recover these structures with uniform consistency guarantees with mild assumptions, where the shapes and relative positions of the components can be arbitrary and the number of such components is recovered automatically.

We then provide an extension to infinite-armed bandits and conclude with empirical results from simulations and image classification on the MNIST dataset.

Setup

Suppose there are K bandit arms indexed in $[K]$. At each time-step t, the learner observes a context $x_t \in \mathbb{R}^D$

*Equal Contribution.

where x_t is drawn i.i.d. from a context density p_X with compact support $\mathcal{X}$ bounded below away from zero (e.g. $\inf_{x\in\mathcal{X}} p_X(x) \geq p_0$ for some $p_0 \geq 0$). Then the learner chooses an arm $I_t \in [K]$ and observes reward

$$r_t = f_{I_t}(x_t) + \xi_t$$

where ξ_t is drawn according to white noise random variable ξ and $f_i : \mathcal{X} \to \mathbb{R}$ is the i-th arm's mean reward. We make the following assumptions.

Assumption 1. *(Lipschitz Mean Reward) There exists L such that $|f_i(x) - f_i(x')| \leq L|x - x'|$ for all $x, x' \in \mathcal{X}$ and $i \in [K]$.*

Assumption 2. *(Sub-Gaussian White noise) ξ satisfies $E[\xi] = 0$ and is sub-Gaussian with parameter σ^2 (i.e. $E[\exp(\lambda\xi)] \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$).*

We require the finite-sample strong uniform consistency result (Theorem 1) for k-NN regression defined as fellows:

Definition 1 (k-NN)**.** *Let the k-NN radius of $x \in \mathcal{X}$ be $r_k(x) := \inf\{r : |B(x,r) \cap X| \geq k\}$ where $B(x,r) := \{x \in \mathcal{X} : |x - x'| \leq r\}$ and the k-NN set of $x \in \mathcal{X}$ be $N_k(x) := X \cap B(x, r_k(x))$. Then for $x \in \mathcal{X}$,*

$$\widehat{f}_{k\text{-NN}}(x) := \frac{1}{|N_k(x)|} \sum_{i=1}^{n} y_i \cdot 1\left[x_i \in N_k(x)\right].$$

Theorem 1. *(Rate for k-NN (Jiang 2017b)) Let $\delta > 0$. There exists N_0 and universal constant C such that if $n \geq N_0$ and $k = \lfloor n^{2/(2+D)} \rfloor$, then with probability at least $1-\delta$,*

$$\sup_{x\in\mathcal{X}} |f(x) - \widehat{f}_{k\text{-NN}}(x)| \leq C\sqrt{\log n \log(1/\delta)} \cdot n^{-1/(2+D)}.$$

It will be implicitly understood from here on that $\widehat{f}_i$ denotes the k-NN regression estimate of f_i under the settings of Theorem 1.

Top-Arm Identification

Algorithm 1 Uniform Sampling

1: Parameters: T, total number of time steps.
2: For each arm i of the K arms:
3: For each time step $t \in [\frac{(i-1)T}{K} + 1, \frac{iT}{K}]$:
4: Pull arm $I_t := i$.
5: Define $\widehat{f}_i : \mathcal{X} \to \mathbb{R}$ to be the k-NN regression estimator from the sampled context and reward observations for each $i \in [K]$.

Definition 2. *(ϵ-optimal arm) Arm i is be ϵ-optimal at context $x \in \mathcal{X}$ if $\max_{j\in[K]} f_j(x) - f_i(x) \leq \epsilon$.*

Following we show a uniform (over context) result about ϵ-optimal arm recovery:

Theorem 2. *(ϵ-optimal arm recovery) Let $\delta > 0$. For Algorithm 1, with probability at least $1 - \delta/K$, if*

$$T \geq K \max\Bigg\{N_0, \log\left(\frac{C\sqrt{\log(1/\delta)}}{\epsilon}\right) \cdot \frac{(2+D)(2C)^{2+D}\log(1/\delta)^{1+D/2}}{\epsilon^{2+D}}\Bigg\},$$

then $\hat{\pi}(x) := \text{argmax}_{i\in[K]} \hat{f}_i(x)$ is ϵ-optimal at context x uniformly for all $x \in \mathcal{X}$.

Remark 1. *This result shows that with $\widetilde{O}(\epsilon^{-(2+D)})$ samples, we can determine an ϵ-approximate best arm. Known lower bounds in nonparametric regression stipulate that we need $\Omega(\epsilon^{-(2+D)})$ to identify differences between functions of size ϵ so our result matches lower bounds up to logarithmic factors.*

Proof. By Theorem 1, it follows that based on the choice of T, each arm has at least enough time such that $\sup_{x\in\mathcal{X}} |\widehat{f}_i(x) - f_i(x)| \leq \epsilon/2$. Thus, we have $\forall x \in \mathcal{X}$, defining $\pi(x) = \max_{j\in[K]} f_j(x)$,

$$f_{\pi(x)}(x) - f_{\hat{\pi}(x)}(x) \leq \widehat{f}_{\pi(x)}(x) - \widehat{f}_{\hat{\pi}}(x) + \epsilon \leq \epsilon,$$

as desired. □

Regret Analysis For UCB Strategy

Define $T_i(t)$ to be the number of times arm i was pulled by time t.

Algorithm 2 Upper Confidence Bound (UCB)

1: Parameters: M_0, M_1, δ, T.
2: Define $\sigma(n) = M_1\sqrt{\log n(\log(nK/\delta))} \cdot n^{-1/(2+D)}$.
3: Pull each of the K arms M_0 times.
4: For each round $t = KM_0, KM_0 + 1, \ldots, T$:
5: Pull $I_t := \text{argmax}_{i\in[K]} \widehat{f}_i(t) + \sigma(T_i(t-1))$.

We use the following notion of regret.

$$R_T = \sum_{t=1}^{T} \left[\max_i f_i(x_t) - f_{I_t}(x_t)\right].$$

Remark 2. *Note that this notion of regret is different from those studied in classical MABs as well as other works in nonparametric contextual bandits. Usually the expected form $E[R_T]$ is bounded. Here, our regret analysis is not under this expectation and hence is a stronger notion of regret.*

Theorem 3. *Let $\delta > 0$. Suppose that $M_0 \geq N_0$ and $M_1 > C$ in Algorithm 2. Then we have that with probability at least $1 - \delta$,*

$$R_T \leq M_1 2\frac{1+D}{2+D} K\sqrt{\log T(\log(TK/\delta)} \cdot T^{\frac{1+D}{2+D}} + KM_0 \max_i ||f_i||_\infty.$$

Remark 3. *This shows a sub-linear regret of $\widetilde{O}(T^{\frac{1+D}{2+D}})$.*

Proof. Denote $\widehat{f}_{i,T_i(t)}$ to be the k-NN regression estimate of f_i at time t. Letting $C_0 = KM_0 \max_i ||f_i||_\infty$, we have by

Theorem 1

$$
\begin{aligned}
R_T &\leq \sum_{i=1}^{T} \sigma(T_{\hat{\pi}(x_t)}(t-1)) + C_0 \leq K \sum_{i=1}^{T} \sigma(i) + C_0 \\
&= M_1 K \sqrt{\log T(\log(TK/\delta)} \sum_{t=1}^{T} t^{-1/(2+D)} + C_0 \\
&\leq M_1 K \sqrt{\log T(\log(TK/\delta)} \int_{t=0}^{T} (1+t)^{-1/(2+D)} dt \\
&\quad + C_0 \\
&\leq M_1 2 \frac{1+D}{2+D} K \sqrt{\log T(\log(TK/\delta)} \cdot T^{\frac{1+D}{2+D}} + C_0.
\end{aligned}
$$

The first inequality holds because the confidence bound of a sub-optimal arm must be higher than that of the optimal at x_t in order for that arm to be chosen and the regret at that time-step is bounded by the confidence bound. The second inequality holds because of the following simple combinatorial argument. Each time a suboptimal arm is chosen, its count increments, or otherwise there is no regret incurred. □

Contextual Bandits on Manifolds

Assumption 3. *(Manifold Assumption) p_X and the family of f_i are supported on M, where:*

- *M is a d-dimensional smooth compact Riemannian manifold without boundary embedded in compact subset $\mathcal{X} \subseteq \mathbb{R}^D$.*
- *The volume of M is bounded above by a constant.*
- *M has condition number $1/\tau$, which controls the curvature and prevents self-intersection.*

Let p_X be the density of $\mathcal{P}$ with respect to the uniform measure on M.

Theorem 4. *(Manifold Rate for k-NN (Jiang 2017b)) Let $\delta > 0$. There exists N_0 and universal constant C such that if $n \geq N_0$ and $k = \lfloor n^{2/(2+d)} \rfloor$, then with probability at least $1 - \delta$,*

$$\sup_{x \in \mathcal{X}} |f(x) - f_k(x)| \leq C \sqrt{\log n \log(1/\delta)} \cdot n^{-1/(2+d)}.$$

Then, simply by using Theorem 4 instead of Theorem 1, we automatically enjoy faster rates for Theorems 2 and 3.

Theorem 5. *(ϵ-optimal arm recovery on manifolds) Let $\delta > 0$. For Algorithm 1, with probability at least $1 - \delta/K$, if*

$$
\begin{aligned}
T \geq K \max \Bigg\{ & N_0, \\
& \log\left(\frac{C\sqrt{\log(1/\delta)}}{\epsilon}\right) \cdot \frac{(2+D)(2C)^{2+d} \log(1/\delta)^{1+D/2}}{\epsilon^{2+d}} \Bigg\},
\end{aligned}
$$

then $\hat{\pi}(x) := \text{argmax}_{i \in [K]} \widehat{f}_i(x)$ is ϵ-optimal at context x uniformly for all $x \in \mathcal{X}$.

Remark 4. *Now the sample complexity is $\widetilde{O}(\epsilon^{2+d})$ instead of $\widetilde{O}(\epsilon^{2+D})$.*

Theorem 6. *(UCB Regret Analysis on Manifolds) Let $\delta > 0$. Suppose that $M_0 \geq N_0$ and $M_1 > C$ in Algorithm 2. Then we have that with probability at least $1 - \delta$,*

$$
\begin{aligned}
R_T \leq & M_1 2 \frac{1+d}{2+d} K \sqrt{\log T(\log(TK/\delta)} \cdot T^{\frac{1+d}{2+d}} \\
& + K M_0 \max_i ||f_i||_\infty.
\end{aligned}
$$

Topological Analysis

In this section, we discuss how topological features about the bandit arms can be recovered. This is similar to recovering the Hartigan notion of clusters as level-sets of the density functions from a finite sample (Chaudhuri and Dasgupta 2010; Jiang 2017a), but here, we find similar structures in the reward functions based on noisy observations of them. We give procedures which can estimate with consistency guarantees the following structure: maximal connected regions in $\mathcal{X}$ where a particular arm is the top-arm.

From the uniform sampling strategy earlier, we obtained estimated policy $\hat{\pi}$ which is δ-optimal uniformly in $\mathcal{X}$ with high probability. Although this is already powerful in giving us the mapping between context space and the corresponding top-arm, it does not immediately tell us the topological features of this mapping. In this subsection, we discuss how to recover the connected components of $\{x \in \mathcal{X} : r_i(x) = \max_{j \in [K]} r_j(x)\}$, the region where arm i is the top-arm.

We give the following simple procedure.

Algorithm 3 Recovering Regions where i-th arm is top arm.

1: Given: Bandit arm i and $R > 0$.
2: Pull each of the K arms T/K times.
3: Let G be the graph with vertices $\{x_t : t \in [T], \widehat{f}_i(x_t) = \max_{j \in [K]} \widehat{f}_j(x_t)\}$ and edges between vertices whose euclidean distance is at most R.
4: **return** The connected components of G.

We now give a consistency result for Algorithm 3.

First, we require the following regularity assumption, which ensures that there are no full-dimensional regions where the top-arm is not unique. This ensures that it is possible to unambiguously recover the regions where a particular arm is top.

Assumption 4. *The region in $\mathcal{X}$ where the top-arm is not unique has measure 0, and for each arm i, the region $\mathcal{X}_i$ where it is unique can be partitioned into full-dimensional connected components.*

Our rates will be in terms of the Hausdorff distance.

Definition 3.

$$d_H(A, B) = \inf\{\epsilon \geq 0 : A \subseteq B \oplus \epsilon, B \subseteq A \oplus \epsilon\},$$

where $A \oplus r := \{x \in \mathcal{X} : \inf_{a \in A} d(x, a) \leq r\}$.

Theorem 7. *Suppose that $\mathcal{X}_i := \{x \in \mathcal{X} : f_i(x) = \max_{j \in [K]} f_j(x)\}$. Let $C_1, ..., C_l$ be the maximal connected components of $\mathcal{X}_i$. Define the following minimum distance between two connected components.*

$$R_0 := \min_{p \neq q} \inf_{x \in C_p, y \in C_q} d(x, y).$$

Also define the following minimum separation in the reward functions

$$D_0 := \inf_{x \notin \mathcal{X}_i \oplus R_0/4} \max_{j \in [K]} f_j(x) - f_i(x).$$

Then the following holds simultaneously for all $C_1, ..., C_l$. Let Algorithm 3 with setting $0 < R < R_0/4$ return $\widehat{C}_1, ..., \widehat{C}_{\hat{l}}$. Then for n sufficiently large, $\hat{l} = l$ and there exists permutation γ of $[l]$ such that

$$d_H(C_j, C_{\gamma(j)}) \leq \xi(n)$$

for some ξ that satisfies $\xi(n) \to 0$ as $n \to \infty$.

Proof. We first show that no two connected components can appear in the same returned component in Algorithm 3. We choose n sufficiently large such that in light of Theorem 1, we have

$$\sup_{x \in \mathcal{X}} \max_{j \in [K]} \widehat{f}_j(x) \leq \frac{D_0}{3}.$$

. Then, uniformly for any $x \notin \mathcal{X}_i \oplus R_0/4$, we have

$$\widehat{f}_i(x) \leq f_i(x) + \frac{D_0}{3} \leq \max_{j \in [K]} f_j(x) - \frac{2D_0}{3}$$
$$\leq \max_{j \in [K]} \widehat{f}_j(x) - \frac{D_0}{3} < \max_{j \in [K]} \widehat{f}_j(x).$$

Thus, $\mathcal{X}_i \oplus R_0/4$ is disjoint from the returned points. Since $R < R_0/4$, it follows that no two connected components points will appear in the same returned connected component from Algorithm 3.

Next, we show that for each connected component C_p, there exists $\widehat{C}_q$ for some $q \in [\hat{l}]$ such that $d_H(\widehat{C}_q, C_p) \to 0$. It suffices to show that for each $r > 0$, we have that for n sufficiently large, $d_H(\widehat{C}_q, C_p) < r$. There are thus two directions to show, that $\widehat{C}_p \subseteq C_q \oplus r$ and $C_q \subseteq \widehat{C}_p \oplus r$. To show the first, define

$$D_1 := \inf_{x \in (C_q \oplus r) \setminus (C_q \oplus (r/2))} \max_{j \in [K]} f_j(x) - f_i(x).$$

Then choose n sufficiently large such that in light of Theorem 1, we have

$$\sup_{x \in \mathcal{X}} \max_{j \in [K]} |\widehat{f}_j(x) - f_j(x)| \leq \frac{D_1}{3}.$$

. Then we have for all $x \in \widehat{C}_p$, if $x \neq C_q \oplus r/2$, then

$$\widehat{f}_i(x) \leq f_i(x) + \frac{D_1}{3} \leq \max_{j \in [K]} f_j(x) - \frac{2D_1}{3} < \max_{j \in [K]} \widehat{f}_j(x),$$

thus, $x \in C_q \oplus r/2 \subseteq C_q \oplus r$. The other direction follows from a similar argument.

All that remains is to show that such points appear in in the same connected component in the graph computed by Algorithm 3. This follows from uniform concentration bounds on balls (e.g. Chaudhuri and Dasgupta (2010)). □

Infinite-Armed Bandits

In this section, we consider the setting where the action space $\mathcal{A}$ is no longer a finite set of bandits, but a compact subset of $\mathbb{R}^{D'}$ for some $D' > 0$.

We given analogous results for the uniform sampling top-arm identification and regret bounds for UCB-type strategy.

Definition 4. *(Mean Reward function)*

$$f : \mathcal{X} \times \mathcal{A} \to \mathbb{R},$$

where $f(x, a)$ is the expected reward of action $a \in \mathcal{A}$ at context $x \in \mathcal{X}$.

Assumption 5. *(Lipschitz Reward) There exists $L > 0$ such that for all $x, x' \in \mathcal{X}$ and $a, a' \in \mathcal{A}$, $|f(x, a) - f(x', a')| \leq L|(x, a) - (x', a')|$, where (x, a) represents the $(D + D')$-dimensional concatenation of x and a.*

Then at each time t, the learner chooses arm $a_t \in \mathcal{A}$ and observes context $x_t \in \mathcal{X}$ and a stochastic reward

$$R_T = f(x_t, a_t) + \xi_t,$$

where $\xi_1, ...$ are i.i.d. white noise with mean 0 and variance σ^2.

Algorithm 4 Infinite-Armed Uniform Sampling

1: Parameters: T, total number of time steps.
2: For $t = 1, ..., T$:
3: Pull I_t, sampled uniformly from $\mathcal{A}$.
4: Observe context x_t and reward R_t.
5: Define $\hat{f}$ to be the k-NN regression estimate from samples $(a_1, R_1), ..., (a_T, R_T)$ with setting $k = \lfloor n^{2/(2+D+D')} \rfloor$.

Definition 5. *(ϵ-optimal arm) Define arm $a \in \mathcal{A}$ to be ϵ-optimal at context $x \in \mathcal{X}$ if $\sup_{a' \in \mathcal{A}} f(x, a') - f(x, a) \leq \epsilon$.*

Following is a uniform (over context and action space) result about ϵ-optimal arm recovery:

Theorem 8. *(ϵ-optimal arm recovery) There exists constant $\tilde{C}_1, \tilde{C}_2$ such that the following holds. Let $\delta > 0$. For Algorithm 4, with probability at least $1 - \delta$, we have that for*

$$T \geq \tilde{C}_1 \log\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right) \frac{\log(1/\delta)^{1+(D+D')/2}}{\epsilon^{D+D'+2}} + \tilde{C}_2,$$

arm $\hat{\pi}(x) := \text{argmax}_{a \in \mathcal{A}} \hat{f}(x)$ is ϵ-optimal at context x uniformly for all $x \in \mathcal{X}$.

Proof. By Theorem 1, it follows that based on the choice of T, there is enough time spent on pulling each arm such that $\sup_{a \in \mathcal{A}, x \in \mathcal{X}} |\hat{f}(x, a) - f(x, a)| \leq \epsilon/2$. Thus, we have $\forall x \in \mathcal{X}$, defining $\pi(x) = \text{argmax}_{a \in \mathcal{A}} f(x, a)$,

$$f(x, \pi(x)) - f(x, \hat{\pi}(x))$$
$$\leq \frac{\epsilon}{2} + \hat{f}(x, \pi(x)) + \frac{\epsilon}{2} - \hat{f}(x, \hat{\pi}(x)) \leq \epsilon,$$

as desired. □

Algorithm 5 Infinite-Armed Upper Confidence Bound (UCB)

1: Parameters: M, M_1, T
2: Define $\sigma(n) = M_1 n^{-1/(2+D+D')}$.
3: For $t = 1, ..., M$:
4: Sample a_t uniformly from $\mathcal{A}$.
5: Observe context x_t and reward R_t.
6: For $t = M+1, ..., T$:
7: Choose $I_t := \text{argmax}_{a\in\mathcal{A}} \widehat{f}(x_t, a) + \sigma(t)$.

Finally, using the notion of regret

$$R_T = \sum_{t=1}^{T} \big[\sup_{a\in\mathcal{A}} f(x_t, a) - f(x_t, a_t) \big],$$

we give the following result. The proof idea is similar to that of Theorem 3 and is omitted here.

Theorem 9. *There exists $\tilde{C}_1$ and $\tilde{C}_2$ such that the following holds. Let $\delta > 0$. Suppose that M and M_1 are chosen sufficiently large in Algorithm 5 depending on f and σ. Then we have that with probability at least $1-\delta$,*

$$R_T \leq \tilde{C}_1 \sqrt{\log T(\log(T/\delta)} \cdot T^{\frac{1+D+D'}{2+D+D'}} + \tilde{C}_2$$

Remark 5. *This shows a sub-linear regret of $\widetilde{O}(T^{\frac{1+D+D'}{2+D+D'}})$.*

Related Works

Canonical works for the standard bandit problem are Lai and Robbins (1985); Berry and Fristedt (1985); Gittins (1989); Auer et al. (2002); Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012).

Work in contextual bandits can be roughly classified into adversarial and stochastic approaches. Much of the former, initiated by Auer et al. (2002), assumes that there is an adversarial game between nature and the learner where, based on a context seen by both players, nature generates rewards for each arm at the same time the learner chooses an arm. Solutions typically involve game theoretical methods. In the stochastic approach, one assumes that the rewards for the arms are generated by a context-dependent distribution.

Approaches to modeling the arm rewards as a function of context are most commonly parametric. One of the most popular is that of linear payoffs, studied under a minimax framework (Goldenshluger and Zeevi 2009; 2013), with UCB-type algorithms (Chu et al. 2011; Li et al. 2010; Auer et al. 2002), or with Thompson sampling (Agrawal and Goyal 2013).

However, it is often the case that the dependency between the payoffs and the contexts are complex and therefore difficult to capture with models such as linear payoffs, many of which requiring strong assumptions on the data. To alleviate this, we can go beyond parametric modeling and blend nonparametric statistics with contextual bandits. Despite the advantage of learning much more general context-payoff dependencies, this line of work has received far less attention.

To the best of our knowledge, the first such work appeared in Yang and Zhu (2002), who used histogram, k-NN, and kernel methods and showed asymptotic convergence rates. Rigollet and Zeevi (2010); Perchet and Rigollet (2013) then combined histogram-type binning techniques in nonparametric statistics to obtain strong regret guarantees for contextual bandits with optimality guarantees.

Lu, Pál, and Pál (2009) study an interesting setting where the reward depends on a Lipschitz measure which is jointly in the context and the action space. They provide upper and lower regret bounds based on a covering argument and give results in terms of the packing dimension. This is highly related to the infinite-armed bandit setting in the present work; we provide similar regret guarantees but with a simple and practical procedure.

More recently, Qian and Yang (2016b); Qian and Yang (2016a) use the strong uniform consistency properties of kernel smoothing regression to establish regret guarantees.

Langford and Zhang (2008); Dudik et al. (2011) alternatively impose neither linear nor smoothness assumptions on the mean reward function. The former propose a modification of an ϵ-greedy policy and showed that expected regret converges to 0 while the latter considers a finite class of policies.

In this paper, using recent finite-sample results about k-NN regression established in Jiang (2017b), we show that using the simple k-NN regression is an effective alternative approach. Moreover, unlike many other nonparametric techniques, k-NN *adapts* to a lower intrinsic dimension (Kpotufe 2011) and thus we show that our regret bounds can adapt to a lower intrinsic dimension automatically and perform as if we were operating in that lower dimensional space.

Experiments

Simulations

We consider three two-arm bandit scenarios in the two-dimensional unit square, where $p_{\mathcal{X}}$ is uniform. We set arm $i \in \{1, 2\}$ to be top in region R_i respectively. Figure 1 illustrates the regions for the different scenarios.

- **Scenario 1 (Quintic Function):** We define two regions above and below a quintic function:
- **Scenario 2 (Smiley):** We use two circles and a semicircle to demarcate the regions in a "smiley face" pattern.
- **Scenario 3 (Bullseye):** We define the regions using the alternating regions of four concentric circles centered in the support.

The true reward functions of the two arms are as follows.

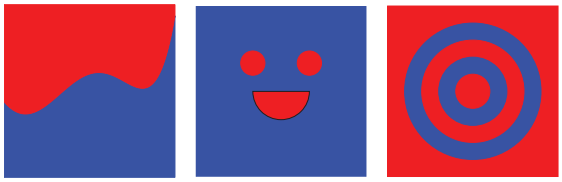
$$f_i(x) = \begin{cases} 1, & x \in R_i \\ 0.5, & x \in R_{j\neq i} \end{cases}$$

The learner observes the rewards with white noise random variable $\xi \sim \mathcal{N}(\mu = 0, \sigma = 0.5)$.

We compare the performance of k-NN regression (nonparametric) and Ridge regression at top-arm identification and regret minimization in the three scenarios. Mirroring our theoretical discussion, we use uniform sampling for top-arm identification and UCB strategy for regret analysis. Note that Ridge regression with UCB is the LinUCB algorithm.

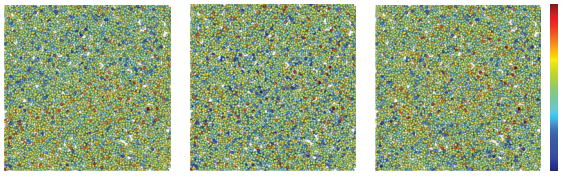
Table 1: Top-arm identification and regret results from Ridge and k-NN regressors. Each model was tuned individually and optimal hyperparameters are shown. k-NN performs better on both metrics for all three scenarios.

	Quintic Function		**Smiley**		**Bullseye**	
	Ridge	kNN	Ridge	kNN	Ridge	kNN
Top-Arm Test Error from Uniform Sampling	0.065	**0.002**	0.080	**0.000**	0.335	**0.005**
Number of samples	500k	500k	2k	5000k	100k	500k
Number of neighbors	-	100	-	50	-	20
Test Regret from UCB sampling	0.0315	**0.001**	0.0375	**0.0135**	0.161	**0.004**
Number of samples	1k	500k	5k	1000k	50k	1000k
Number of neighbors	-	100	-	20	-	100



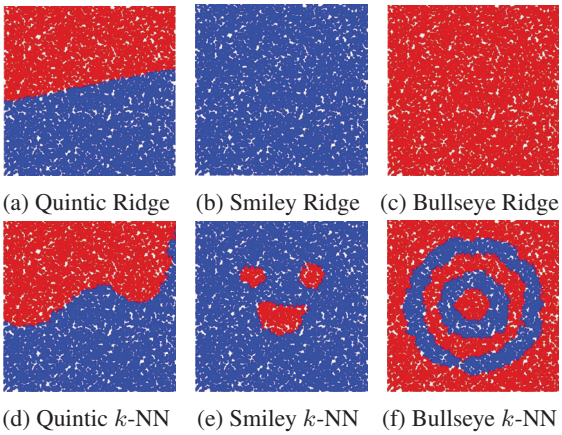
(a) Quintic (b) Smiley (c) Bullseye

Figure 1: Top-arm boundaries. Red and blue regions correspond to where top-arm is arm 1 and 2 respectively.



(a) Quintic (b) Smiley (c) Bullseye

Figure 2: Observed reward density plots from 10k uniform samples illustrating pseudo-randomness of training data. In the colormap (right) warmer colors correspond to higher values, normalized on the range of the observed rewards.



(a) Quintic Ridge (b) Smiley Ridge (c) Bullseye Ridge

(d) Quintic k-NN (e) Smiley k-NN (f) Bullseye k-NN

Figure 3: Test results on top-arm identification using Ridge regression and 25-NN regression. Contexts are labeled in red and blue if arms 1 and 2 are estimated to be top respectively.

Qualitative Analysis We first qualitatively show that k-NN regression can successfully model the bandits whereas the linear method cannot. The difficulty of the task is illustrated by Figure 2, which plots 10k uniformly sampled samples from each scenario with a colormap. We can see that a human would have a hard time recovering the regions where each arm is top due to the randomness in the observed rewards. This randomness is considerable as we set $\sigma = 0.5$ to be the same as $|f_i(x \in R_i) - f_i(x \notin R_i)|$.

We fix the number of training samples N to 10k and the number of nearest neighbors to $k = 25$. We evaluate on 10k random test samples. Figure 3 shows that k-NN regression does an excellent job of reproducing the region boundaries. Ridge regression does a poor job in the Quintic Function case, making a linear approximation to the quintic curve, and completely fails in the Smiley and Bullseye Cases, simply choosing the arm whose top-arm region is larger.

Quantitative Analysis We report numerical results and optimal hyperparameters in Table 1. We tuned other hyperparameters using grid search on a validation set of size 1k using grid search and we evaluate performance of our models on a test set of size 1k. We use the UCB strategy in Auer et al. (2002) (a simplified version of UCB by Agrawal and Goyal (2013)). We found that a confidence level of 0.1 worked well for all settings. We see that k-NN significantly outperforms Ridge regression for both top-arm identification and regret minimization in all three scenarios (Table 1).

Image Classification Experiments

We extend our experiments to image classification of the canonical MNIST dataset, which consists of 60k training images and 10k test images of isolated, normalized, handwritten digits. The task is to classify each 28×28 image into one of ten classes. We reframe this as a contextual MAB problem by treating the classes as arms and the images as the contexts. Note that for every context, the payoff of all arms are known: 1 if the class is the true label and 0 otherwise. We compare k-NN and Ridge regressions at regret minimization using the UCB strategy. As before we use the UCB strategy in Auer et al. (2002) and fix the confidence level to 0.1. We do not employ any data augmentation.

We obtain test regret of 17.5% from LinUCB with $\alpha = 5$, where α is the coefficient of L2 regularization, and significantly lower test regret of 5.8% from 4-NNUCB. Figure

4 shows that k-NN regression maintains lower regret than Ridge regression over a range of values of k and α. We note that Ridge regression working well for relatively large values of α itself suggests that it is a poor model for the task.

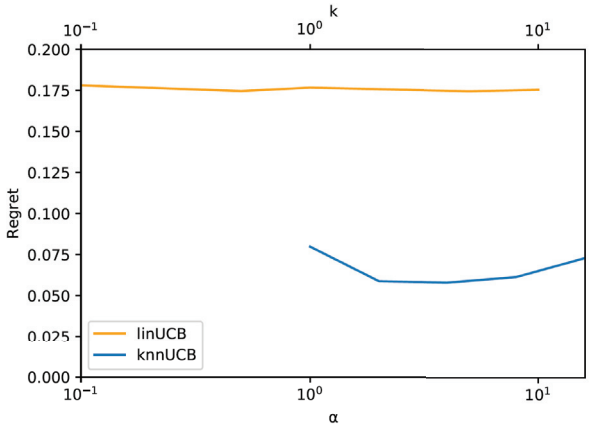


Figure 4: Test results on regret minimization for MNIST image classification over varied values of α (for LinUCB) and k (for k-NNUCB). The nonparametric approach achieves significantly lower regret over a range of hyperparameters.

Conclusion

For the multi-armed bandit setting, we use nonparametric regression to attain tight results for top-arm identification and a sublinear regret of $\widetilde{O}(T^{\frac{1+D}{2+D}})$, where D is the dimension of the context. We also show that if the underlying context space has a lower intrinsic dimension $d < D$, then our algorithm automatically adapts to the lower dimension and attains a faster rate of $\widetilde{O}(T^{\frac{1+d}{2+d}})$. We also provide a procedure for recovering the maximal connected regions in a support where a particular arm is the top-arm and provide a consistency analysis. We then give a natural extension to infinite-armed contextual bandits. Our simulations confirm that our method is able to learn in the contextual setting with arbitrary decision boundaries, even in the presence of significant noise, and our experiments on classification of MNIST images demonstrate superior performance of our method over LinUCB on a real world task.

References

Agrawal, S., and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135.

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

Berry, D. A., and Fristedt, B. 1985. *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*, volume 12. Springer.

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

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

Chaudhuri, K., and Dasgupta, S. 2010. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, 343–351.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. E. 2011. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 208–214.

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

Gittins, J. 1989. Multi-armed bandit allocation indices. wiley-interscience series in systems and optimization.

Goldenshluger, A., and Zeevi, A. 2009. Woodroofes one-armed bandit problem revisited. *The Annals of Applied Probability* 19(4):1603–1633.

Goldenshluger, A., and Zeevi, A. 2013. A linear response bandit problem. *Stochastic Systems* 3(1):230–261.

Jiang, H. 2017a. Density level set estimation on manifolds with dbscan. *arXiv preprint arXiv:1703.03503*.

Jiang, H. 2017b. Rates of uniform consistency for k-nn regression. *arXiv preprint arXiv:1707.06261*.

Kpotufe, S. 2011. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, 729–737.

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

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

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670. ACM.

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

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

Qian, W., and Yang, Y. 2016a. Kernel estimation and model combination in a bandit problem with covariates. *Journal of Machine Learning Research*.

Qian, W., and Yang, Y. 2016b. Randomized allocation with arm elimination in a bandit problem with covariates. *Electronic Journal of Statistics* 10(1):242–270.

Rigollet, P., and Zeevi, A. 2010. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*.

Robbins, H. 1985. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*. Springer. 169–177.

Yang, Y., and Zhu, D. 2002. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics* 30(1):100–121.