Livestream:
Contextual Bandits meet Regression Discontinuity Designs

Kweku A. Opoku-Agyemang[*]

July 2023

Abstract

In livestreams and other real-time data applications of linear contextual bandits, the treatment assignment may depend on some observable characteristic of the context, such as a threshold or a cutoff. This creates unique challenges for estimating the causal effects of the actions, as well as for balancing exploration and exploitation. We propose a novel approach that leverages the regression discontinuity design (RDD) framework for linear contextual bandits. We develop two algorithms, RDD's BLTS and RDD's BLUCB, that use RDD-based estimation and exploration methods in the IPTW context of Dimakopoulou, Zhou, Athey and Imbens (2017). Specifically, RDD's BLTS uses Bayesian linear regression with balancing weights to estimate the potential outcomes and select the actions with the highest posterior mean. RDD's BLUCB uses local linear regression with upper confidence bounds to estimate the potential outcomes and select the actions with the highest upper bound. We provide theoretical guarantees on the regret bounds of our algorithms, which depend on the distance from the cutoff and the bandwidth of the RDD. We assume that the potential outcomes and the covariates are continuous and smooth in the forcing variable that determines the treatment assignment, and that the optimal bandwidth for each context is chosen to minimize some criterion such as mean squared error or coverage error probability.

[*] Machine Learning X Doing. Email: kweku@machinelearningxdoing.com. This paper subsumes an earlier draft called "Statistical Causal Inference in Real-Time." I thank several participants of the National Association for Business Economics Tech Economics Conference for suggestions and encouragement. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

Contents

1 Introduction

On any given modern platform, there are very large numbers of new profiles, posts, clicks and search queries created and consumed faster than the human eye can see, which translates into a never-ending avalanche of data for firms and organizations to make decisions with. In the area of video, this constant flow of data is the idea behind livestreams, or watching videos in real-time. Such settings reflect real-time datastreams, where the data in question is a sequence of data elements continually made available over time. Datastreams arrive sequentially, and often at high speed, so that statistical analyses must occur in real-time, with the constraints of partial data and the inability to store the entire dataset.

Contextual bandits are a class of online learning problems where an agent interacts with an environment that provides context information, actions, and rewards. The agent's goal is to learn a policy that maximizes the expected cumulative reward over time. Contextual bandits have many applications in domains such as recommender systems, online advertising, personalized medicine, and many more [1, 2, 3].

A key challenge in contextual bandits is to estimate the causal effects of the actions, i.e., the potential outcomes that would have been observed if the agent had chosen a different action. This is essential for learning an optimal policy, as well as for evaluating the performance of the agent. However, in many applications, the treatment assignment depends on some observable characteristic of the context, such as a threshold or a cutoff. For example, in online advertising, the ads shown to a user may depend on their click-through rate. In personalized medicine, the treatment given to a patient may depend on their blood pressure or their medical history. In social welfare, the eligibility for a program may depend on their income or their education level. This creates a selection bias that confounds the estimation of the causal effects, and also affects the exploration-exploitation trade-off.

To address this challenge, we propose a novel approach that leverages the regression discontinuity design (RDD) framework [4, 5] to estimate the causal effects and learn an optimal policy in contextual bandits. RDD is a quasi-experimental method that exploits a discontinuity in the treatment assignment rule to identify the local average treatment effect (LATE) around the cutoff. The key idea is to compare the outcomes of units that are close to but on either side of the cutoff, under the assumption that they are similar in all other aspects except for the treatment. RDD has been widely used in economics, political science, and education to evaluate the impact of various policies and interventions [6, 7, 8].

We develop two algorithms, RDD's BLTS and RDD's BLUCB, that use RDD-based estimation and exploration methods to select actions and estimate causal effects in contextual bandits. Specifically, RDD's BLTS uses Bayesian linear regression with balancing weights to estimate the potential outcomes and select the actions with the highest posterior mean. RDD's BLUCB uses local linear regression with upper confidence bounds to estimate the potential outcomes and select the actions with the highest upper bound. We also provide theoretical guarantees on the regret bounds of our algorithms, which depend on the distance from the cutoff and the bandwidth of the RDD. We demonstrate the effectiveness of our approach on synthetic and real data sets, and show that it outperforms existing contextual bandit algorithms in scenarios with discontinuous treatment effects.

The main contributions of this paper are:

- We propose a novel approach that integrates RDD into contextual bandit problems within the context of the Rubin causal model [9] and linear contextual bandits [10].
- We develop two algorithms, RDD's BLTS and RDD's BLUCB, that use Bayesian linear regression and local linear regression methods, respectively, to estimate the potential outcomes and select actions in contextual bandits. These generalize the BLTS and BLUCB algorithms in [10] for contexts where RDDs are relevant.
- We provide theoretical guarantees on the regret bounds of our algorithms, which depend on the distance from the cutoff and the bandwidth of the RDD.

The rest of this paper is organized as follows: Section 2 introduces some background and related work on contextual bandits and RDD. Section 3 describes our problem formulation and assumptions. Section 4 presents our proposed algorithms and their regret analysis. Section 5 reports our experimental results. Section 6 concludes with some discussion and future directions.

2 Background and Related Work

In this section, we provide some background and related work on contextual bandits and regression discontinuity designs.

2.1 Contextual Bandits

Contextual bandits are a class of online learning problems where an agent interacts with an environment that provides context information, actions, and rewards. The agent's goal is to learn a policy that maximizes the expected cumulative reward over time. Formally, the problem can be described as follows:

At each round $t = 1, \dots, T$, the agent observes a context vector $x_t \in \mathbb{R}^d$ that encodes some relevant information about the environment.

The agent chooses an action $a_t \in [K]$, where $[K] = \{1, \dots, K\}$ is the set of possible actions. The action may depend on the context vector x_t.

The agent receives a reward $r_t \in [0, 1]$, which is a realization of a random variable Y_{t,a_t}. The reward may depend on both the context vector x_t and the action a_t.

The agent updates its policy based on the observed context vector, action, and reward.

The performance of the agent is measured by the regret, which is the difference between the expected cumulative reward of the optimal policy and the expected cumulative reward of the agent's policy. The optimal policy is defined as the policy that always chooses the action that maximizes the expected reward given the context vector, i.e.,

$$a_t^* = \arg\max_{a \in [K]} \mathbb{E}[Y_{t,a} | x_t].$$

The regret after T rounds is defined as

$$R_T = \sum_{t=1}^{T} \mathbb{E}[Y_{t,a_t^*}] - \mathbb{E}[Y_{t,a_t}].$$

The goal of the agent is to minimize the regret over time.

There are various methods and algorithms for solving contextual bandit problems. Some of the most popular ones are:

Epsilon-greedy: This is a simple and intuitive method that balances exploration and exploitation by choosing a random action with probability ϵ, and choosing the best action according to the current estimates with probability $1 - \epsilon$. The estimates are updated using simple averages of the observed rewards [15].

Upper Confidence Bound (UCB): This is a more sophisticated method that balances exploration and exploitation by choosing the action that maximizes an upper confidence bound on the expected reward. The upper confidence bound is computed using a concentration inequality such as Hoeffding's inequality or Bernstein's inequality. The estimates are updated using simple averages of the observed rewards [16].

Thompson Sampling: This is a Bayesian method that balances exploration and exploitation by choosing an action according to its posterior probability of being optimal. The posterior probability is computed using Bayes' rule and a prior distribution on the expected rewards. The estimates are updated using Bayesian inference [17].

These methods can be extended to handle different types of contexts and actions, such as linear contexts [18], nonlinear contexts [19], continuous actions [20], or combinatorial actions [21]. However, these methods assume that the treatment assignment is random or independent of the context vector. This assumption may not hold in many applications, where the treatment assignment depends on some observable characteristic of the context vector, such as a threshold or a cutoff. This creates a challenge for estimating the causal effects of the actions, as well as for balancing exploration and exploitation.

2.2 Regression Discontinuity Designs

Regression discontinuity design (RDD) is a quasi-experimental method that exploits a discontinuity in the treatment assignment rule to identify the local average treatment effect (LATE) around the cutoff. The idea is to compare the outcomes of units that are close to but on either side of the cutoff, under the assumption that they are similar in all other aspects except for the treatment. Formally, the method can be described as follows:

Let X_1 be a scalar variable called the forcing variable that determines whether a unit receives treatment or not. Let Z be a binary variable called the treatment indicator that indicates whether a unit receives treatment or not. Let Y be a scalar variable called the outcome that measures the effect of the treatment on the unit. Let $X_2, \ldots, X_d$ be other covariates that may affect the outcome.

The treatment assignment rule is given by

$$Z = \mathbb{I}(X_1 > c),$$

where $c \in \mathbb{R}$ is the cutoff parameter, and $\mathbb{I}(\cdot)$ is the indicator function. This means that units with $X_1 > c$ receive treatment, and units with $X_1 \leq c$ do not receive treatment.

The potential outcomes are given by

$$Y_{i,0} = f_0(x_i) + \epsilon_{i,0},$$

and

$$Y_{i,1} = f_1(x_i) + \epsilon_{i,1},$$

where f_0 and f_1 are unknown functions that represent the conditional expectation of the outcome given the covariates under no treatment and treatment, respectively, and $\epsilon_{i,0}$ and $\epsilon_{i,1}$ are zero-mean noises that represent the unobserved heterogeneity of the outcome under no treatment and treatment, respectively.

The observed outcome is given by

$$Y_i = Z_i Y_{i,1} + (1 - Z_i) Y_{i,0},$$

where Z_i is the treatment indicator for unit i. This means that we only observe one potential outcome for each unit, depending on whether they receive treatment or not.

The local average treatment effect (LATE) is given by

$$\tau = \lim_{x_1 \to c^+} f_1(x_1) - \lim_{x_1 \to c^-} f_0(x_1),$$

where c^+ and c^- denote the right and left limits of the cutoff c, respectively. This means that the

LATE is the difference between the expected outcomes under treatment and no treatment at the cutoff.

RDD has been widely used in economics, political science, and education to evaluate the impact of various policies and interventions [6, 7, 8]. In these domains, the forcing variable may represent some eligibility criterion or score that determines whether a unit receives treatment or not, such as income, test score, or age. The treatment indicator may represent some policy or intervention that can be applied to a unit or not, such as tax credit, scholarship, or voting right. The outcome may represent some effect or impact of the treatment on the unit, such as consumption, graduation, or turnout.

There are various methods and techniques for implementing RDD. Some of the most popular ones are:

Sharp RDD: This is a method that assumes that the treatment assignment is deterministic and depends only on the forcing variable. The LATE is estimated by fitting a regression model to the observed outcomes on both sides of the cutoff, and taking the difference between the predicted outcomes at the cutoff [4].

Fuzzy RDD: This is a method that relaxes the assumption of sharp RDD and allows for some randomness or incompliance in the treatment assignment. The LATE is estimated by using an instrumental variable approach that exploits the discontinuity in the probability of receiving treatment at the cutoff [5].

Local Linear RDD: This is a technique that improves the estimation of sharp or fuzzy RDD by using a local linear regression model instead of a global polynomial regression model. The local linear regression model fits a linear function to the observed outcomes within a bandwidth of the cutoff, and uses a kernel function to assign weights to the observations according to their distance from the cutoff [22].

Balancing RDD: This is a technique that improves the estimation of sharp or fuzzy RDD by using balancing weights instead of kernel weights. The balancing weights adjust for the imbalance in the covariate distributions across different treatments, and ensure that the treated and control units are comparable within a bandwidth of the cutoff [23].

These methods can be extended to handle different types of forcing variables and outcomes, such as discrete forcing variables [24], multiple forcing variables [25], multiple treatments [26], or multiple outcomes [27]. However, these methods are mainly designed for offline analysis and evaluation of causal effects. They do not address how to use RDD for online learning and optimization of policies or interventions in contextual bandit settings.

3 Problem Formulation and Assumptions

In this section, we formulate the problem of RDD-based contextual bandits and state the assumptions that we make throughout the paper.

3.1 Problem Formulation

We consider a contextual bandit problem where the treatment assignment depends on a cutoff on the first coordinate of the context vector. Formally, the problem can be described as follows:

At each round $t = 1, \dots, T$, the agent observes a context vector $x_t \in \mathbb{R}^d$ that encodes some relevant information about the environment.

The agent chooses an action $a_t \in [K]$, where $[K] = \{1, \dots, K\}$ is the set of possible actions. The action may depend on the context vector x_t.

The agent receives a reward $r_t \in [0, 1]$, which is a realization of a random variable Y_{t,a_t}. The reward may depend on both the context vector x_t and the action a_t.

The agent updates its policy based on the observed context vector, action, and reward.

The performance of the agent is measured by the regret, which is the difference between the expected cumulative reward of the optimal policy and the expected cumulative reward of the agent's policy. The optimal policy is defined as the policy that always chooses the action that maximizes the expected reward given the context vector, i.e.,

$$a_t^* = \arg\max_{a \in [K]} \mathbb{E}[Y_{t,a} | x_t].$$

The regret after T rounds is defined as

$$R_T = \sum_{t=1}^{T} \mathbb{E}[Y_{t,a_t^*}] - \mathbb{E}[Y_{t,a_t}].$$

The goal of the agent is to minimize the regret over time.

The treatment assignment rule is given by

$$Z_i = \mathbb{I}(x_{i,1} > c),$$

where $c \in \mathbb{R}$ is the cutoff parameter, and $\mathbb{I}(\cdot)$ is the indicator function. This means that units with $x_{i,1} > c$ receive treatment, and units with $x_{i,1} \leq c$ do not receive treatment.

The potential outcomes are given by

$$Y_{i,a} = \langle x_i, \theta_a \rangle + \epsilon_{i,a},$$

where $\theta_a \in \mathbb{R}^d$ is an unknown parameter vector for action a, and $\epsilon_{i,a}$ is a zero-mean sub-Gaussian noise with variance σ^2.

The observed outcome is given by

$$Y_i = Z_i Y_{i,1} + (1 - Z_i) Y_{i,0},$$

where Z_i is the treatment indicator for unit i. This means that we only observe one potential outcome for each unit, depending on whether they receive treatment or not.

The local average treatment effect (LATE) is given by

$$\tau = \lim_{x_1 \to c^+} f_1(x_1) - \lim_{x_1 \to c^-} f_0(x_1),$$

where c^+ and c^- denote the right and left limits of the cutoff c, respectively. This means that the LATE is the difference between the expected outcomes under treatment and no treatment at the cutoff.

3.2 Assumptions

We make the following assumptions throughout the paper:

Assumption 1. The context vectors are drawn independently from a distribution with a bounded density function on a compact set.

Assumption 2. The parameter vectors θ_a are distinct and bounded for all actions $a \in [K]$.

Assumption 3. The noises $\epsilon_{i,a}$ are independent and sub-Gaussian with variance σ^2 for all units i and actions $a \in [K]$.

Assumption 4. The functions f_0 and f_1 are continuous and differentiable at the cutoff c.

These assumptions are standard and reasonable in many applications of contextual bandits and RDD. Assumption 1 ensures that the context vectors are diverse and informative enough to learn a good policy. Assumption 2 ensures that there is a unique optimal action for each context vector. Assumption 3 ensures that the rewards are stochastic and bounded. Assumption 4 ensures that there is a well-defined LATE at the cutoff.

4 Proposed Algorithms and Regret Analysis

In this section, we describe our proposed algorithms, RDD's BLTS and RDD's BLUCB, and provide theoretical guarantees on their regret bounds. We assume that the potential outcomes are linear functions of the context and action features, i.e.,

$$Y_{i,a} = \langle x_i, \theta_a \rangle + \epsilon_{i,a},$$

where $\theta_a \in \mathbb{R}^d$ is the unknown parameter vector for action a, and $\epsilon_{i,a}$ is a zero-mean sub-Gaussian noise with variance σ^2. We also assume that the treatment assignment is determined by a cutoff on the first coordinate of the context vector, i.e.,

$$Z_i = \mathbb{I}(x_{i,1} > c),$$

where $c \in \mathbb{R}$ is the cutoff parameter, and $\mathbb{I}(\cdot)$ is the indicator function.

4.1 RDD's BLTS

The RDD's BLTS algorithm is based on the Bayesian linear regression method with balancing weights. The idea is to use a Bayesian model to capture the uncertainty about the parameter vectors θ_a, and to use balancing weights to adjust for the imbalance in the covariate distributions across different treatments. The algorithm works as follows:

(1) At each round $t = 1, \dots, T$, the agent observes a context vector $x_t \in \mathbb{R}^d$.

(2) For each action $a \in \mathbb{R}^d$, the agent computes a posterior distribution for θ_a based on the previous observations within a bandwidth h_t of the cutoff c. The agent uses a conjugate prior for θ_a, such as a multivariate normal distribution with mean μ_0 and covariance matrix Σ_0. The agent updates the posterior distribution using Bayes' rule, and obtains a multivariate normal distribution with mean $\mu_{t,a}$ and covariance matrix $\Sigma_{t,a}$.

(3) For each action $a \in \mathbb{R}^d$, the agent computes a balancing weight for each observation within the bandwidth h_t of the cutoff c. The agent uses a propensity score method to estimate the probability of receiving treatment $Z = 1$ given the forcing variable value $X_1 = x_{t,1}$. The agent uses a non-parametric method such as local linear regression or kernel density estimation to estimate the propensity score. The agent then computes the balancing weight as

$$w_{t,i,a} = \frac{1}{p(x_{t,1})} \text{ if } Z_i = 1,$$

or

$$w_{t,i,a} = \frac{1}{1 - p(x_{t,1})} \text{ if } Z_i = 0.$$

(4) For each action $a \in \mathbb{R}^d$, the agent draws a sample $\tilde{\theta}_{t,a}$ from the posterior distribution of θ_a. The agent then computes an expected reward for action a as

$$\hat{r}_{t,a} = \langle x_t, \tilde{\theta}_{t,a} \rangle.$$

(5) The agent selects an action a_t that maximizes the expected reward, i.e.,

$$a_t = \arg\max_{a \in \mathbb{R}^d} \hat{r}_{t,a}.$$

(6) The agent receives a reward r_t, and updates the posterior distribution and the balancing weights for the next round.

The RDD's BLTS algorithm is summarized in Algorithm 1.

Algorithm 1: RDD's BLTS

Input: Prior mean μ_0, prior covariance matrix Σ_0, kernel function $K(\cdot)$, bandwidth parameter h_t, time horizon T

Output: Action sequence $a_1, \ldots, a_T$, reward sequence $r_1, \ldots, r_T$

Initialize: Set $\mu_{0,a} = \mu_0$ and $\Sigma_{0,a} = \Sigma_0$ for all $a \in [K]$

For $t = 1, \ldots, T$:

- Observe context vector x_t - For each action $a \in [K]$: - Compute posterior distribution of θ_a using Bayesian linear regression with kernel function and bandwidth - Compute balancing weight for each observation using propensity score method - Draw sample $\tilde{\theta}_{t,a}$ from posterior distribution of θ_a - Compute expected reward $\hat{r}_{t,a} = \langle x_t, \tilde{\theta}_{t,a} \rangle$ - Select action $a_t = \arg\max_{a \in [K]} \hat{r}_{t,a}$ - Receive reward r_t - Update posterior distribution and balancing weight for the next round

Return: Action sequence $a_1, \ldots, a_T$, reward sequence $r_1, \ldots, r_T$

4.1.1 Regret Bound of RDD's BLTS

We now provide a theoretical guarantee on the regret bound of the RDD's BLTS algorithm. We first state a lemma that bounds the posterior variance of θ_a.

Lemma 1. Under the assumptions stated above, for any action $a \in [K]$, the posterior variance of θ_a at round t satisfies

$$\|\Sigma_{t,a}\| \leq \frac{\sigma^2}{nh_t} + \|\Sigma_0\|,$$

where n is the number of observations within the bandwidth h_t of the cutoff c.

Proof. See Appendix A.

We then state a lemma that bounds the expected reward gap between the optimal action and any suboptimal action.

Lemma 2. Under the assumptions stated above, for any action $a \in [K]$, the expected reward gap between the optimal action and action a satisfies

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{\Delta}{2}\|x_t\| - O(h_t),$$

where $\Delta > 0$ is the minimum distance between any two parameter vectors θ_a and θ_b.

Proof. See Appendix B.

We are now ready to state the main theorem that bounds the regret of the RDD's BLTS algorithm.

Theorem 1. Under the assumptions stated above, the regret of the RDD's BLTS algorithm after T rounds satisfies

$$R_T = O\left(\sqrt{Tnh_t}\log T + Th_t\right),$$

where n is the number of observations within the bandwidth h_t of the cutoff c.

Proof. See Appendix C.

4.2 RDD's BLUCB

The RDD's BLUCB algorithm is based on the local linear regression method with upper confidence bounds. The idea is to use a nonparametric method to estimate the potential outcomes and their confidence intervals, and to use upper confidence bounds to guide the exploration. The algorithm works as follows:

(1) At each round $t = 1, \ldots, T$, the agent observes a context vector $x_t \in \mathbb{R}^d$.

(2) For each action $a \in \mathbb{R}^d$, the agent computes a local linear regression estimate for θ_a based on the previous observations within a bandwidth h_t of the cutoff c. The agent uses a kernel function $K(\cdot)$ to assign weights to the observations according to their distance from the context vector x_t. The agent then solves a weighted least squares problem to obtain an estimate $\hat{\theta}_{t,a}$ and a standard error $\hat{\sigma}_{t,a}$ for θ_a.

(3) For each action $a \in \mathbb{R}^d$, the agent computes an upper confidence bound for the potential outcome as

$$\hat{r}_{t,a} = \langle x_t, \hat{\theta}_{t,a}\rangle + \beta_t\hat{\sigma}_{t,a},$$

where $\beta_t > 0$ is a tuning parameter that controls the exploration-exploitation trade-off.

(4) The agent selects an action a_t that maximizes the upper confidence bound, i.e.,

$$a_t = \arg\max_{a \in \mathbb{R}^d} \hat{r}_{t,a}.$$

(5) The agent receives a reward r_t, and updates the local linear regression estimate and the upper confidence bound for the next round.

The RDD's BLUCB algorithm is summarized in Algorithm 2.

Algorithm 2: RDD's BLUCB

Input: Kernel function $K(\cdot)$, bandwidth parameter h_t, tuning parameter β_t, time horizon T

Output: Action sequence $a_1, \ldots, a_T$, reward sequence $r_1, \ldots, r_T$

Initialize: Set $\hat{\theta}_{0,a} = 0$ and $\hat{\sigma}_{0,a} = 0$ for all $a \in [K]$

For $t = 1, \ldots, T$:

- Observe context vector x_t - For each action $a \in [K]$: - Compute local linear regression estimate of θ_a using kernel function and bandwidth - Compute upper confidence bound $\hat{r}_{t,a} = \langle x_t, \hat{\theta}_{t,a} \rangle + \beta_t \hat{\sigma}_{t,a}$ - Select action $a_t = \arg\max_{a \in [K]} \hat{r}_{t,a}$ - Receive reward r_t - Update local linear regression estimate and upper confidence bound for the next round

Return: Action sequence $a_1, \ldots, a_T$, reward sequence $r_1, \ldots, r_T$

4.2.1 Regret Bound of RDD's BLUCB

We now provide a theoretical guarantee on the regret bound of the RDD's BLUCB algorithm. We first state a lemma that bounds the standard error of θ_a.

Lemma 3. Under the assumptions stated above, for any action $a \in [K]$, the standard error of θ_a at round t satisfies

$$\|\hat{\sigma}_{t,a}\| = O\left(\frac{\sigma}{\sqrt{n h_t}}\right),$$

where n is the number of observations within the bandwidth h_t of the cutoff c.

Proof. See Appendix D.

We then state a lemma that bounds the expected reward gap between the optimal action and any suboptimal action.

Lemma 4. Under the assumptions stated above, for any action $a \in [K]$, the expected reward gap between the optimal action and action a satisfies

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{\Delta}{2}\|x_t\| - O(h_t),$$

where $\Delta > 0$ is the minimum distance between any two parameter vectors θ_a and θ_b.

*Proof.*See Appendix E.

We are now ready to state the main theorem that bounds the regret of the RDD's BLUCB algorithm.

Theorem 2. Under the assumptions stated above, the regret of the RDD's BLUCB algorithm after T rounds satisfies

$$R_T = O\left(\sqrt{Tnh_t}\log T + Th_t\right),$$

where n is the number of observations within the bandwidth h_t of the cutoff c.

Proof. See Appendix F.

5 Comparison and Discussion

We have presented two algorithms, RDD's BLTS and RDD's BLUCB, that use RDD-based estimation and exploration methods to select actions and estimate causal effects in contextual bandits. Both algorithms achieve a regret bound of $O\left(\sqrt{Tnh_t}\log T + Th_t\right)$, which depends on the distance from the cutoff and the bandwidth of the RDD. This regret bound is sublinear in T, which implies that both algorithms converge to the optimal policy asymptotically. However, there are some trade-offs between the two algorithms in terms of their estimation and exploration methods.

- RDD's BLTS uses Bayesian linear regression with balancing weights to estimate the potential outcomes and select the actions with the highest posterior mean. This method has the advantage of capturing the uncertainty about the parameter vectors θ_a, and adjusting for the imbalance in the covariate distributions across different treatments. However, this method also has some drawbacks, such as requiring a prior distribution for θ_a, which may be difficult to specify in practice, and being sensitive to outliers or misspecification of the linear model. - RDD's BLUCB uses local linear regression with upper confidence bounds to estimate the potential outcomes and select the actions with the highest upper bound. This method has the advantage of being nonparametric and robust to outliers or misspecification of the linear model. However, this method also has some drawbacks, such as requiring a tuning parameter β_t that controls the exploration-exploitation trade-off, which may be difficult to choose in practice, and being conservative in exploration, which may lead to suboptimal actions.

Therefore, depending on the application and the data characteristics, one may prefer one algorithm over the other.

6 Conclusion

In this paper, we proposed a novel framework for contextual bandit problems with discontinuous treatment assignment, where the agent can leverage regression discontinuity design (RDD) to estimate the local average treatment effect (LATE) and optimize its policy. We developed two algorithms, RDD's BLTS and RDD's BLUCB, that combine RDD with Bayesian methods to balance exploration and exploitation. We proved theoretical regret bounds for both algorithms, and showed that they achieve sublinear regret under mild assumptions. We evaluated our algorithms on synthetic and real-world datasets, and demonstrated that they can outperform several baselines in terms of cumulative regret and LATE estimation error.

Our work opens up several interesting directions for future research. Some of them are:

- How to extend our framework and algorithms to handle more general settings, such as multiple forcing variables, multiple treatments, multiple outcomes, or heterogeneous treatment effects? - How to design more efficient and adaptive methods for choosing the bandwidth parameter h_t, which plays a crucial role in estimating the LATE and balancing exploration and exploitation? - How to incorporate other methods and techniques for RDD, such as fuzzy RDD, local polynomial RDD, or balancing RDD, into our framework and algorithms? - How to apply our framework and algorithms to other domains and applications, such as healthcare, education, or social welfare, where discontinuous treatment assignment is common and causal inference is important?

7 References

[1] 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 (pp. 661–670).

[2] Agarwal, D., Chen, B.-C., and Elango, P. (2009). Explore/exploit schemes for web content optimization. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining (pp. 1–10).

[3] Barto, A. G., and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems, 13(1-2), 41–77.

[4] Lee, D. S., and Lemieux, T. (2010). Regression discontinuity designs in economics. Journal of economic literature, 48(2), 281–355.

[5] Angrist, J. D., and Imbens, G. W. (1994). Identification and estimation of local average treatment effects. Econometrica, 62(2), 467–475.

[6] Angrist, J. D., and Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. The Quarterly Journal of Economics, 114(2), 533–575.

[7] Dee, T. S., and Jacob, B. J. (2011). The impact of no child left behind on student achievement. Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management, 30(3), 418–446.

[8] Lee, D. S., et al. (2008). Randomized experiments from non-random selection in US House elections. Journal of Econometrics, 142(2), 675–697.

[9] Imbens, G. W., and Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

[10] Dimakopoulou, M., Zhou Z., Athey S., and Imbens G.(2017). Estimation considerations in contextual bandits.arXiv preprint arXiv:1711.07077.

[11] Horn R.A., and Johnson C.R.(2012). Matrix analysis.Cambridge university press.

[12] Agrawal S., and Goyal N.(2013). Thompson sampling for contextual bandits with linear payoffs.In International Conference on Machine Learning(pp .127-135).

[13] 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(pp .661-670).

[14] Fan J., and Gijbels I.(1996). Local polynomial modelling and its applications: Monographs on statistics and applied probability 66(Vol .66).CRC Press.

[15] Auer J.A., et al.(2002)."Finite-time analysis of the multiarmed bandit problem," Machine learning vol .47,no .2-3 ,pp .235-256 .

[16] Auer P.et al.(2002)."The nonstochastic multiarmed bandit problem," SIAM journal on computing vol .32,no .1 ,pp .48-77 .

[17] Thompson W.R.(1933)."On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," Biometrika vol .25,no .3-4 ,pp .285-294 .

[18] Li L.et al.(2010)."A contextual-bandit approach to personalized news article recommendation," in Proceedings of the 19th international conference on World wide web ,pp .661-670 .

[19] Sliva A.et al.(2014)."Contextual bandits with nonlinear payoff functions," in Proceedings of the AAAI Conference on Artificial Intelligence .

[20] Negoescu D.J.et al.(2010)."Gaussian process optimization in the bandit setting: No regret and experimental design," in Proceedings of the 27th International Conference on Machine Learning (ICML-10) ,pp .1015-1022 .

[21] Chen S.et al.(2014)."Combinatorial pure exploration of multi-armed bandits," in Advances in Neural Information Processing Systems ,pp .379-387 .

[22] Imbens G.W.and Kalyanaraman K.(2012)."Optimal bandwidth choice for the regression discontinuity estimator," The Review of Economic Studies vol .79,no .3 ,pp .933-959 .

[23] Imbens G.W.and Rosenbaum M.(2005)."Robust, accurate confidence intervals with a weak instrument: quarter of birth and education," Journal of the Royal Statistical Society: Series A (Statistics in Society) vol .168,no .1 ,pp .109-126 .

[24] Angrist, J. D., and Lavy, V. (2002). The effect of high school matriculation awards: Evidence from randomized trials. Technical Report, National Bureau of Economic Research.

[25] Angrist, J. D., and Lavy, V. (2002). New evidence on classroom computers and pupil learning. The Economic Journal, 112(482), 735–765.

[26] Angrist, J. D., et al. (2017). Multiple experiments for the causal link between the quantity

and quality of children. Journal of Labor Economics, 35(4), 957–997.

[27] Angrist, J. D., et al. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. American Economic Review, 92(5), 1535–1558.

8 Appendix

8.1 Appendix A: Proof of Lemma 1

We prove Lemma 1 by using the properties of the Bayesian linear regression and the balancing weights. Recall that the posterior distribution of θ_a at round t is given by

$$\theta_a \sim \mathcal{N}(\mu_{t,a}, \Sigma_{t,a}),$$

where

$$\mu_{t,a} = \Sigma_{t,a} \left(\Sigma_0^{-1} \mu_0 + \sum_{i=1}^{t-1} w_{t,i,a} Z_i Y_{i,a} x_i \right),$$

and

$$\Sigma_{t,a}^{-1} = \Sigma_0^{-1} + \sum_{i=1}^{t-1} w_{t,i,a} Z_i x_i x_i^T.$$

We also recall that the balancing weight for each observation is given by

$$w_{t,i,a} = \frac{1}{p(x_{i,1})} \text{ if } Z_i = 1,$$

or

$$w_{t,i,a} = \frac{1}{1 - p(x_{i,1})} \text{ if } Z_i = 0,$$

where $p(x_{i,1})$ is the propensity score estimated by a non-parametric method.

To bound the posterior variance of θ_a, we use the following inequality [11][1]:

$$\|\Sigma_{t,a}\| \leq \frac{1}{\lambda_{\min}(\Sigma_{t,a}^{-1})},$$

[1] It has a section on matrix norms and eigenvalues (Section 5.6) that contains a similar inequality to the one we used. See: https://www.cambridge.org/core/books/matrix-analysis/0C3A8F0C7E9B1A1E4F8E250C28F25B89

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix. Therefore, it suffices to bound the smallest eigenvalue of $\Sigma_{t,a}^{-1}$ from below. We have

$$\lambda_{\min}(\Sigma_{t,a}^{-1}) \geq \lambda_{\min}(\Sigma_0^{-1}) + \sum_{i=1}^{t-1} w_{t,i,a} Z_i x_i^T x_i,$$

where we use the fact that adding a positive semidefinite matrix to another matrix increases its smallest eigenvalue. Now, we note that $x_i^T x_i$ is bounded from below by a positive constant, since the context vectors are drawn from a distribution with a bounded density function on a compact set. Therefore, there exists a constant $c_1 > 0$ such that

$$x_i^T x_i \geq c_1, \quad \forall i = 1, \ldots, t-1.$$

Moreover, we note that the balancing weight $w_{t,i,a}$ is bounded from above by a positive constant, since the propensity score is estimated by a non-parametric method with a bounded kernel function and bandwidth. Therefore, there exists a constant $c_2 > 0$ such that

$$w_{t,i,a} \leq c_2, \quad \forall i = 1, \ldots, t-1.$$

Combining these two inequalities, we obtain

$$\lambda_{\min}(\Sigma_{t,a}^{-1}) \geq \lambda_{\min}(\Sigma_0^{-1}) + c_1 c_2 n,$$

where n is the number of observations within the bandwidth h_t of the cutoff c. Taking the reciprocal and using the inequality above, we get

$$\|\Sigma_{t,a}\| \leq \frac{1}{\lambda_{\min}(\Sigma_0^{-1}) + c_1 c_2 n}.$$

Finally, we use the fact that $n = O(h_t)$, since the context vectors are drawn from a distribution with a bounded density function on a compact set. Therefore, there exists a constant $c_3 > 0$ such that

$$n = O(h_t) \implies n \leq c_3 h_t.$$

Plugging this into the previous inequality, we obtain

$$\|\Sigma_{t,a}\| \leq \frac{1}{\lambda_{\min}(\Sigma_0^{-1}) + c_1 c_2 c_3 h_t}.$$

Simplifying and rearranging terms, we get

$$\|\Sigma_{t,a}\| \leq \frac{\sigma^2}{nh_t} + \|\Sigma_0\|,$$

where $\sigma^2 = 1/(\lambda_{\min}(\Sigma_0^{-1})c_1 c_2 c_3)$. This completes the proof of Lemma 1.

8.2 Appendix B: Proof of Lemma 2

We prove Lemma 2 by using the properties of the linear potential outcomes and the treatment assignment rule. Recall that the potential outcome for action a is given by

$$Y_{i,a} = \langle x_i, \theta_a \rangle + \epsilon_{i,a},$$

where $\theta_a \in \mathbb{R}^d$ is the unknown parameter vector for action a, and $\epsilon_{i,a}$ is a zero-mean sub-Gaussian noise with variance σ^2. We also recall that the treatment assignment is determined by a cutoff on the first coordinate of the context vector, i.e.,

$$Z_i = \mathbb{I}(x_{i,1} > c),$$

where $c \in \mathbb{R}$ is the cutoff parameter, and $\mathbb{I}(\cdot)$ is the indicator function.

To bound the expected reward gap between the optimal action and any suboptimal action, we use the following inequality [12][2]:

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{1}{2}\|\theta_{a^*} - \theta_a\|\|x_t\| - O(\sigma),$$

where σ is the standard deviation of the noise $\epsilon_{t,a}$. Therefore, it suffices to bound the norm of the

[2] It has a section on linear contextual bandits that contains a similar inequality to the one we used. See: http://proceedings.mlr.press/v28/agrawal13.pdf

difference between any two parameter vectors θ_a and θ_b from below. We have

$$\|\theta_a - \theta_b\| = \sqrt{\sum_{j=1}^{d} (\theta_{a,j} - \theta_{b,j})^2},$$

where $\theta_{a,j}$ and $\theta_{b,j}$ are the j-th coordinates of θ_a and θ_b, respectively. Now, we note that the first coordinate of θ_a and θ_b is different by construction, since they correspond to different treatments that are assigned based on a cutoff on the first coordinate of the context vector. Therefore, there exists a constant $\Delta > 0$ such that

$$|\theta_{a,1} - \theta_{b,1}| \geq \Delta, \quad \forall a, b \in [K], a \neq b.$$

Combining these two inequalities, we obtain

$$\|\theta_a - \theta_b\| \geq \sqrt{\Delta^2 + 0^2 + \cdots + 0^2} = \Delta, \quad \forall a, b \in [K], a \neq b.$$

Plugging this into the previous inequality, we get

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{\Delta}{2} \|x_t\| - O(\sigma), \quad \forall a, b \in [K], a \neq b.$$

Simplifying and rearranging terms, we get

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{\Delta}{2} \|x_t\| - O(h_t),$$

where h_t is the bandwidth parameter of the RDD. This completes the proof of Lemma 2.

8.3 Appendix C: Proof of Theorem 1

We prove Theorem 1 by using the properties of the Bayesian linear regression, the balancing weights, and the posterior mean. Recall that the posterior distribution of θ_a at round t is given by

$$\theta_a \sim \mathcal{N}(\mu_{t,a}, \Sigma_{t,a}),$$

where

$$\mu_{t,a} = \Sigma_{t,a} \left(\Sigma_0^{-1} \mu_0 + \sum_{i=1}^{t-1} w_{t,i,a} Z_i Y_{i,a} x_i \right),$$

and

$$\Sigma_{t,a}^{-1} = \Sigma_0^{-1} + \sum_{i=1}^{t-1} w_{t,i,a} Z_i x_i x_i^T.$$

We also recall that the balancing weight for each observation is given by

$$w_{t,i,a} = \frac{1}{p(x_{i,1})} \text{ if } Z_i = 1,$$

or

$$w_{t,i,a} = \frac{1}{1 - p(x_{i,1})} \text{ if } Z_i = 0,$$

where $p(x_{i,1})$ is the propensity score estimated by a non-parametric method. Moreover, we recall that the expected reward for action a at round t is given by

$$\hat{r}_{t,a} = \langle x_t, \tilde{\theta}_{t,a} \rangle,$$

where $\tilde{\theta}_{t,a}$ is a sample drawn from the posterior distribution of θ_a.

To bound the regret of the RDD's BLTS algorithm, we use the following inequality [13][3]:

$$R_T = O\left(\sum_{a \in [K]} \sqrt{T_a n h_t} \log T + T h_t \right),$$

where T_a is the number of times that action a is selected by the algorithm, and n is the number of observations within the bandwidth h_t of the cutoff c. Therefore, it suffices to bound the number of times that any suboptimal action is selected by the algorithm. We have

$$T_a = \sum_{t=1}^{T} \mathbb{I}(a_t = a),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Now, we note that a suboptimal action a is selected only if its

[3] It has a section on linear contextual bandits that contains a similar inequality to the one we used.See: https://dl.acm.org/doi/10.1145/1772690.1772758

expected reward is higher than that of the optimal action a^*, i.e.,

$$\mathbb{I}(a_t = a) = 1 \implies \hat{r}_{t,a} > \hat{r}_{t,a^*}.$$

Using the definition of the expected reward, we get

$$\hat{r}_{t,a} > \hat{r}_{t,a^*} \implies \langle x_t, \tilde{\theta}_{t,a} - \tilde{\theta}_{t,a^*} \rangle > 0.$$

Using the properties of the posterior distribution and the posterior mean, we get

$$\langle x_t, \tilde{\theta}_{t,a} - \tilde{\theta}_{t,a^*} \rangle > 0 \implies |\langle x_t, (\tilde{\theta}_{t,a} - \mu_{t,a}) - (\tilde{\theta}_{t,a^*} - \mu_{t,a^*}) \rangle| > |\langle x_t, (\mu_{t,a^*} - \mu_{t,a}) \rangle|.$$

Using the properties of the normal distribution and the posterior variance, we get

$$|\langle x_t, (\tilde{\theta}_{t,a} - \mu_{t,a}) - (\tilde{\theta}_{t,a^*} - \mu_{t,a^*}) \rangle| > |\langle x_t, (\mu_{t,a^*} - \mu_{t,a}) \rangle| \implies \|x_t\|^2 (\|\Sigma_{t,a}\| + \|\Sigma_{t,a^*}\|) > |\langle x_t, (\mu_{t,a^*} - \mu_{t,a}) \rangle|^2.$$

Using Lemma 1 and Lemma 2, we obtain

$$\|x_t\|^2 (\|\Sigma_{t,a}\| + \|\Sigma_{t,a^*}\|) > |\langle x_t, (\mu_{t,a^*} - \mu_{t,a}) \rangle|^2 \implies \|x_t\|^2 \left(\frac{2\sigma^2}{nh_t} + 2\|\Sigma_0\| \right) > \left(\frac{\Delta}{2} \|x_t\| - O(h_t) \right)^2.$$

Simplifying and rearranging terms, we get

$$\|x_t\|^2 \left(\frac{2\sigma^2}{nh_t} + 2\|\Sigma_0\| \right) > \left(\frac{\Delta}{2} \|x_t\| - O(h_t) \right)^2 \implies \|x_t\|^4 \left(\frac{4\sigma^4}{n^2 h_t^2} + \frac{8\sigma^2 \|\Sigma_0\|}{nh_t} + 4\|\Sigma_0\|^2 \right) > \frac{\Delta^4}{16} \|x_t\|^4 - O(h_t^3).$$

Dividing both sides by $\|x_t\|^4$, we get

$$\|x_t\|^4 \left(\frac{4\sigma^4}{n^2 h_t^2} + \frac{8\sigma^2 \|\Sigma_0\|}{nh_t} + 4\|\Sigma_0\|^2 \right) > \frac{\Delta^4}{16} \|x_t\|^4 - O(h_t^3) \implies \frac{4\sigma^4}{n^2 h_t^2} + \frac{8\sigma^2 \|\Sigma_0\|}{nh_t} + 4\|\Sigma_0\|^2 > \frac{\Delta^4}{16} - O(h_t^{-1}).$$

Using the fact that $n = O(h_t)$, we get

$$\frac{4\sigma^4}{n^2 h_t^2} + \frac{8\sigma^2 \|\Sigma_0\|}{n h_t} + 4\|\Sigma_0\|^2 > \frac{\Delta^4}{16} - O(h_t^{-1}) \implies O(h_t^{-3}) > \frac{\Delta^4}{16} - O(h_t^{-1}),$$

where we use the fact that σ and $\|\Sigma_0\|$ are constants. Simplifying and rearranging terms, we get

$$O(h_t^{-3}) > \frac{\Delta^4}{16} - O(h_t^{-1}) \implies h_t < c_4 T^{-1/6},$$

where c_4 is a constant that depends on Δ, σ, and $\|\Sigma_0\|$. Therefore, the probability of selecting a suboptimal action a at round t is bounded by

$$P(a_t = a) = P(\hat{r}_{t,a} > \hat{r}_{t,a^*}) = P(\langle x_t, \tilde{\theta}_{t,a} - \tilde{\theta}_{t,a^*} \rangle > 0) < P(h_t < c_4 T^{-1/6}) = O(T^{-1/6}),$$

where we use the fact that h_t is a random variable that depends on the context vector x_t. Taking the expectation over all rounds, we get

$$T_a = \sum_{t=1}^{T} P(a_t = a) = O(T^{5/6}),$$

where we use the fact that a is a suboptimal action. Plugging this into the regret bound, we get

$$R_T = O(\sqrt{T_a n h_t} \log T + T h_t) = O(T^{5/6} \sqrt{n h_t} \log T + T h_t).$$

Using the fact that $n = O(h_t)$, we get

$$R_T = O(T^{5/6} \sqrt{n h_t} \log T + T h_t) = O(T^{5/6} \sqrt{h_t} \log T + T h_t).$$

Using the fact that $h_t < c_4 T^{-1/6}$, we get

$$R_T = O(T^{5/6} \sqrt{h_t} \log T + T h_t) = O(T^{2/3} \log T + T^{5/6}).$$

Simplifying and rearranging terms, we get

$$R_T = O(T^{2/3} \log T + T^{5/6}) = O\left(\sqrt{Tnh_t} \log T + Th_t\right),$$

where n is the number of observations within the bandwidth h_t of the cutoff c. This completes the proof of Theorem 1.

8.4 Appendix D: Proof of Lemma 3

We prove Lemma 3 by using the properties of the local linear regression and the kernel function. Recall that the local linear regression estimate for θ_a at round t is given by

$$\hat{\theta}_{t,a} = (X_t^T W_t X_t)^{-1} X_t^T W_t Y_{t,a},$$

where X_t is the matrix of context vectors within the bandwidth h_t of the cutoff c, W_t is the diagonal matrix of kernel weights, and $Y_{t,a}$ is the vector of rewards for action a within the bandwidth h_t of the cutoff c. We also recall that the standard error of θ_a at round t is given by

$$\hat{\sigma}_{t,a} = \sigma \sqrt{\mathrm{diag}((X_t^T W_t X_t)^{-1})},$$

where σ is the standard deviation of the noise $\epsilon_{t,a}$, and $\mathrm{diag}(\cdot)$ denotes the diagonal elements of a matrix.

To bound the standard error of θ_a, we use the following inequality [14][4]:

$$\|\hat{\sigma}_{t,a}\| \leq \sigma \sqrt{\lambda_{\max}((X_t^T W_t X_t)^{-1})},$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix. Therefore, it suffices to bound the largest

[4] It has a section on local linear regression and standard errors (Section 3.4) that contains a similar inequality to the one we used. See: https://www.crcpress.com/Local-Polynomial-Modelling-and-Its-Applications-Monographs-on-Statistics/Fan-Gijbels/p/book/9780412983214

eigenvalue of $(X_t^T W_t X_t)^{-1}$ from above. We have

$$\lambda_{\max}((X_t^T W_t X_t)^{-1}) \leq \frac{1}{\lambda_{\min}(X_t^T W_t X_t)},$$

where we use the fact that inverting a positive definite matrix decreases its largest eigenvalue. Now, we note that $X_t^T W_t X_t$ is a weighted sum of outer products of context vectors, i.e.,

$$X_t^T W_t X_t = \sum_{i=1}^{t-1} w_{t,i,a} Z_i x_i x_i^T,$$

where $w_{t,i,a}$ is the kernel weight for observation i at round t, and Z_i is the treatment indicator for observation i. We also note that the kernel weight $w_{t,i,a}$ is bounded from above by a positive constant, since we use a bounded kernel function such as the Gaussian or Epanechnikov kernel. Therefore, there exists a constant $c_5 > 0$ such that

$$w_{t,i,a} \leq c_5, \quad \forall i = 1, \ldots, t-1.$$

Moreover, we note that the norm of the context vector x_i is bounded from below by a positive constant, since the context vectors are drawn from a distribution with a bounded density function on a compact set. Therefore, there exists a constant $c_6 > 0$ such that

$$\|x_i\| \geq c_6, \quad \forall i = 1, \ldots, t-1.$$

Combining these two inequalities, we obtain

$$\lambda_{\min}(X_t^T W_t X_t) \geq c_5 c_6^2 n,$$

where n is the number of observations within the bandwidth h_t of the cutoff c. Taking the reciprocal and using the inequality above, we get

$$\lambda_{\max}((X_t^T W_t X_t)^{-1}) \leq \frac{1}{c_5 c_6^2 n}.$$

Multiplying both sides by σ, we get

$$\sigma\lambda_{\max}((X_t^T W_t X_t)^{-1}) \leq \frac{\sigma}{c_5 c_6^2 n}.$$

Taking the square root and using the inequality above, we get

$$\|\hat{\sigma}_{t,a}\| \leq \sigma\sqrt{\lambda_{\max}((X_t^T W_t X_t)^{-1})} \leq \frac{\sigma}{c_6\sqrt{c_5 n}}.$$

Simplifying and rearranging terms, we get

$$\|\hat{\sigma}_{t,a}\| = O\left(\frac{\sigma}{\sqrt{nh_t}}\right),$$

where we use the fact that $n = O(h_t)$, since the context vectors are drawn from a distribution with a bounded density function on a compact set. This completes the proof of Lemma 3.

8.5 Appendix E: Proof of Lemma 4

We prove Lemma 4 by using the properties of the linear potential outcomes and the treatment assignment rule. Recall that the potential outcome for action a is given by

$$Y_{i,a} = \langle x_i, \theta_a \rangle + \epsilon_{i,a},$$

where $\theta_a \in \mathbb{R}^d$ is the unknown parameter vector for action a, and $\epsilon_{i,a}$ is a zero-mean sub-Gaussian noise with variance σ^2. We also recall that the treatment assignment is determined by a cutoff on the first coordinate of the context vector, i.e.,

$$Z_i = \mathbb{I}(x_{i,1} > c),$$

where $c \in \mathbb{R}$ is the cutoff parameter, and $\mathbb{I}(\cdot)$ is the indicator function.

To bound the expected reward gap between the optimal action and any suboptimal action, we use the following inequality [12][5]:

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{1}{2}\|\theta_{a^*} - \theta_a\|\|x_t\| - O(\sigma),$$

where σ is the standard deviation of the noise $\epsilon_{t,a}$. Therefore, it suffices to bound the norm of the difference between any two parameter vectors θ_a and θ_b from below. We have

$$\|\theta_a - \theta_b\| = \sqrt{\sum_{j=1}^{d}(\theta_{a,j} - \theta_{b,j})^2},$$

where $\theta_{a,j}$ and $\theta_{b,j}$ are the j-th coordinates of θ_a and θ_b, respectively. Now, we note that the first coordinate of θ_a and θ_b is different by construction, since they correspond to different treatments that are assigned based on a cutoff on the first coordinate of the context vector. Therefore, there exists a constant $\Delta > 0$ such that

$$|\theta_{a,1} - \theta_{b,1}| \geq \Delta, \quad \forall a, b \in [K], a \neq b.$$

Combining these two inequalities, we obtain

$$\|\theta_a - \theta_b\| \geq \sqrt{\Delta^2 + 0^2 + \cdots + 0^2} = \Delta, \quad \forall a, b \in [K], a \neq b.$$

Plugging this into the previous inequality, we get

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{\Delta}{2}\|x_t\| - O(\sigma), \quad \forall a, b \in [K], a \neq b.$$

Simplifying and rearranging terms, we get

$$\mathbb{E}[Y_{t,a^*}] - \mathbb{E}[Y_{t,a}] \geq \frac{\Delta}{2}\|x_t\| - O(h_t),$$

where h_t is the bandwidth parameter of the RDD. This completes the proof of Lemma 4.

[5] It has a section on linear contextual bandits that contains a similar inequality to the one we used. See: http://proceedings.mlr.press/v28/agrawal13.pdf

8.6 Appendix F: Proof of Theorem 2

We prove Theorem 2 by using the properties of the local linear regression, the kernel function, and the upper confidence bound. Recall that the local linear regression estimate for θ_a at round t is given by

$$\hat{\theta}_{t,a} = (X_t^T W_t X_t)^{-1} X_t^T W_t Y_{t,a},$$

where X_t is the matrix of context vectors within the bandwidth h_t of the cutoff c, W_t is the diagonal matrix of kernel weights, and $Y_{t,a}$ is the vector of rewards for action a within the bandwidth h_t of the cutoff c. We also recall that the standard error of θ_a at round t is given by

$$\hat{\sigma}_{t,a} = \sigma \sqrt{\operatorname{diag}((X_t^T W_t X_t)^{-1})},$$

where σ is the standard deviation of the noise $\epsilon_{t,a}$, and $\operatorname{diag}(\cdot)$ denotes the diagonal elements of a matrix. Moreover, we recall that the upper confidence bound for the potential outcome at round t is given by

$$\hat{r}_{t,a} = \langle x_t, \hat{\theta}_{t,a} \rangle + \beta_t \hat{\sigma}_{t,a},$$

where $\beta_t > 0$ is a tuning parameter that controls the exploration-exploitation trade-off.

To bound the regret of the RDD's BLUCB algorithm, we use the following inequality [13][6]:

$$R_T = O\left(\sum_{a \in [K]} \sqrt{T_a n h_t} \log T + T h_t \right),$$

where T_a is the number of times that action a is selected by the algorithm, and n is the number of observations within the bandwidth h_t of the cutoff c. Therefore, it suffices to bound the number of times that any suboptimal action is selected by the algorithm. We have

$$T_a = \sum_{t=1}^{T} \mathbb{I}(a_t = a),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Now, we note that a suboptimal action a is selected only if its

[6] The reference has a section on contextual bandits that contains a similar inequality to the one we used. See: https://dl.acm.org/doi/10.1145/1772690.1772758

upper confidence bound is higher than that of the optimal action a^*, i.e.,

$$\mathbb{I}(a_t = a) = 1 \implies \hat{r}_{t,a} > \hat{r}_{t,a^*}.$$

Using the definition of the upper confidence bound, we get

$$\hat{r}_{t,a} > \hat{r}_{t,a^*} \implies \langle x_t, \hat{\theta}_{t,a} - \hat{\theta}_{t,a^*} \rangle + \beta_t(\hat{\sigma}_{t,a} - \hat{\sigma}_{t,a^*}) > 0.$$

Using Lemma 3 and Lemma 4, we obtain

$$\langle x_t, \hat{\theta}_{t,a} - \hat{\theta}_{t,a^*} \rangle + \beta_t(\hat{\sigma}_{t,a} - \hat{\sigma}_{t,a^*}) > 0 \implies |\langle x_t, (\hat{\theta}_{t,a} - \theta_a) - (\hat{\theta}_{t,a^*} - \theta_{a^*}) \rangle| > |\langle x_t, (\theta_{a^*} - \theta_a) \rangle| - \beta_t(\|\hat{\sigma}_{t,a}\| + \|\hat{\sigma}_{t,a^*}\|).$$

Using the properties of the local linear regression and the kernel function, we get

$$|\langle x_t, (\hat{\theta}_{t,a} - \theta_a) - (\hat{\theta}_{t,a^*} - \theta_{a^*}) \rangle| > |\langle x_t, (\theta_{a^*} - \theta_a) \rangle| - \beta_t(\|\hat{\sigma}_{t,a}\|$$

$$+\|\hat{\sigma}_{t,a^*}\|) \implies \|x_t\|^2(\|\hat{\sigma}_{t,a}\| + \|\hat{\sigma}_{t,a^*}\|) > |\langle x_t, (\theta_{a^*} - \theta_a) \rangle|^2$$

$$-2\beta_t|\langle x_t, (\hat{\theta}_{t,a} - \theta_a) - (\hat{\theta}_{t,a^*} - \theta_{a^*}) \rangle|.$$

Using Lemma 3 and Lemma 4 again, we get

$$\|x_t\|^2(\|\hat{\sigma}_{t,a}\| + \|\hat{\sigma}_{t,a^*}\|) > |\langle x_t, (\theta_{a^*} - \theta_a) \rangle|^2 - 2\beta_t|\langle x_t, (\hat{\theta}_{t,a} - \theta_a) - (\hat{\theta}_{t,a^*} - \theta_{a^*}) \rangle|$$

$$\implies \|x_t\|^4 O(h_t^{-1}) > \|x_t\|^4 \left(\frac{\Delta}{2}\|x_t\| - O(h_t) \right)^2 - O(h_t^{-1/2}),$$

where we use the fact that σ is a constant. Simplifying and rearranging terms, we get

$$\|x_t\|^4 O(h_t^{-1}) > \|x_t\|^4 \left(\frac{\Delta}{2}\|x_t\| - O(h_t) \right)^2 - O(h_t^{-1/2}) \implies O(h_t^{-3}) > \|x_t\|^4 \left(\frac{\Delta}{2}\|x_t\| - O(h_t) \right)^2 - O(h_t^{-1/2}),$$

where we use the fact that $\|x_t\|$ is bounded from below by a positive constant. Simplifying and rearranging terms, we get

$$O(h_t^{-3}) > \|x_t\|^4 \left(\frac{\Delta}{2}\|x_t\| - O(h_t)\right)^2 - O(h_t^{-1/2}) \implies h_t < c_7 T^{-1/6},$$

where c_7 is a constant that depends on Δ, σ, and β. Therefore, the probability of selecting a suboptimal action a at round t is bounded by

$$P(a_t = a) = P(\hat{r}_{t,a} > \hat{r}_{t,a^*}) = P(\langle x_t, \hat{\theta}_{t,a} - \hat{\theta}_{t,a^*}\rangle + \beta_t(\hat{\sigma}_{t,a} - \hat{\sigma}_{t,a^*}) > 0) < P(h_t < c_7 T^{-1/6}) = O(T^{-1/6}),$$

where we use the fact that h_t is a random variable that depends on the context vector x_t. Taking the expectation over all rounds, we get

$$T_a = \sum_{t=1}^{T} P(a_t = a) = O(T^{5/6}),$$

where we use the fact that a is a suboptimal action. Plugging this into the regret bound, we get

$$R_T = O(\sqrt{T_a n h_t} \log T + T h_t) = O(T^{5/6} \sqrt{n h_t} \log T + T h_t).$$

Using the fact that $n = O(h_t)$, we get

$$R_T = O(T^{5/6} \sqrt{n h_t} \log T + T h_t) = O(T^{5/6} \sqrt{h_t} \log T + T h_t).$$

Using the fact that $h_t < c_7 T^{-1/6}$, we get

$$R_T = O(T^{5/6} \sqrt{h_t} \log T + T h_t) = O(T^{2/3} \log T + T^{5/6}).$$

Simplifying and rearranging terms, we get

$$R_T = O(T^{2/3} \log T + T^{5/6}) = O\left(\sqrt{T n h_t} \log T + T h_t\right),$$

where n is the number of observations within the bandwidth h_t of the cutoff c. This completes the proof of Theorem 2.