Advanced in Interactive Learning: Alternative Feedback Mechanisms and Adaptive Causal Inference

Ojash Neopane

July 2025
CMU-ML-25-110

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee

Aaditya Ramdas, Co-chair
Aarti Singh, Co-Chair
Alekh Agarwal
Nathan Kallus

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported by: Air Force Research Laboratory awards FA86501925209 and FA87501720212; National Science Foundation awards CCF1563918, DGE1745016 and DGE2140739; and Office of Naval Research award N000142212363.

Keywords: multi-armed bandits, causal inference, adaptive experimentation, transfer learning, preference learning, machine learning

To my parents, Deepak and Arati Neopane and my second set of parents, Shekhar and Shova Sharma.

Abstract

This dissertation addresses fundamental challenges in sequential decision-making and adaptive experimental design, developing theoretically grounded algorithms that achieve significant improvements in both sample complexity and practical performance. We organize our contributions into two main areas: novel sampling mechanisms for learning and adaptive methods for causal inference.

The first area focuses on sampling strategies that improve learning efficiency across different problem settings. We develop transfer learning algorithms for multi-armed bandits that can automatically adapt the degree of knowledge transfer based on observed similarity between source and target tasks, providing theoretical guarantees that gracefully interpolate between perfect transfer and learning from scratch. We also formalize active exploration in preference-based learning as a contextual dueling bandit problem, developing algorithms with polynomial regret bounds using reproducing kernel Hilbert space methods, with applications to reinforcement learning from human feedback and direct preference optimization.

The second area develops adaptive experimental design methods for efficient causal inference. We introduce the Clipped Second Moment Tracking algorithm that achieves exponential improvements in finite-sample regret, reducing dependence from $O(\sqrt{T})$ to $O(\log T)$ while maintaining polynomial dependence on problem parameters. We also develop an Optimistic Policy Tracking approach that leverages the asymptotically optimal Augmented Inverse Probability Weighting estimator through principled optimistic design, demonstrating how techniques from bandit theory can be successfully adapted to causal inference.

Throughout this work, we emphasize the gap between asymptotic and finite-sample performance, developing principled algorithmic approaches that provide both theoretical guarantees and practical improvements. Our contributions advance the state-of-the-art in sequential decision-making by bridging theory and practice across multiple important application domains including clinical trials, online experimentation, and human-AI interaction.

Acknowledgments

First and foremost, I want to express my deepest gratitude to my advisors, Aarti Singh and Aaditya Ramdas. Aaditya, your boundless energy and curiosity have had a lasting impact on me. Your passion for research and for life itself has been deeply inspiring and something I hope to carry with me in all aspects of my life. Aarti, your patience and kindness have been a constant source of reassurance and strength throughout my PhD. Thank you for your support, your steady guidance, and for always believing in me. I am also incredibly grateful to my committee members, Alekh Agarwal and Nathan Kallus, for their thoughtful feedback and support during this journey.

I owe a great deal to the many friends I made during my time at CMU. I am grateful for all of the close friendships that I have made — Anna Bair, Shrey Bagroy, Kartik Chiturri, Theophile Gervet, Rishub Jain, Viraj Mehta, Lauren Parola, Charvi Rastogi, Ourania Siabani, Kushagra Singh, Jacob Tyo, and Minji Yoon — thank you for the late-night discussions, walks around campus, spontaneous celebrations, and for making the grind of graduate school feel lighter and more joyful. Your camaraderie has meant the world to me.

To all the other friends and peers I crossed paths with at CMU — Ben Chugg, Jeremy Cohen, Santiago Cortez, Tomas Gonzales Lara, Diego Martinez Taboada, and Stephanie Milani — thank you for the laughter, the thoughtful conversations, and the encouragement. You helped make this experience what it was.

Finally, to my family: thank you for being my foundation. To my mom and dad, whose love, sacrifices, and unwavering belief in me have made everything possible — I owe you everything. To Bhinaju, Diju Hajur, Shirish Dada, Shreya Didi, and Medha: thank you for your warmth and for always making me feel grounded, no matter how far I wandered. Your support has been a quiet but constant presence throughout this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sequential decision-making under uncertainty is a fundamental challenge that arises across numerous domains, from medical trials and online advertising to robotics and artificial intelligence. At its core, this problem requires balancing exploration — gathering information about unknown aspects of the environment — with exploitation — leveraging current knowledge to make optimal decisions. This dissertation addresses several important instantiations of this challenge, developing principled algorithmic approaches that advance both theoretical understanding and practical performance.

1.1 Overview and Contributions

This dissertation is organized into two main chapters, each containing multiple related contributions.

1.1.1 Chapter 2: Learning from Alternative Feedback

This chapter develops new adaptive algorithms for principled sample efficient learning across different problem settings:

Transfer Learning in Multi-Armed Bandits (MAB) (Section 2.2). We address the challenge of leveraging auxiliary information from related source tasks to improve learning efficiency in new target environments. Our algorithm automatically adapts the degree of transfer based on observed data, providing theoretical guarantees that gracefully interpolate between perfect transfer scenarios and learning from scratch. In doing so, we are able to generalize existing algorithms for Best Arm Identification to much more general settings while still maintaining optimal performance in existing problem settings. Through these generalizations, our work provides improved sample complexity bounds compared to existing algorithms, while enabling applications in new transfer learning settings which have previously not been studied in the MAB literature. This section is based on [1].

Active Exploration for Preference Learning (Section 2.3). In this work, we formalize and investigate the problem of active preference-based learning in contextual bands. Our theoretical contribution focus on developing algorithms with polynomial regret bounds when the underlying reward structure can be described by a function in a Reproducing Kernel Hilbert Space (RKHS). Using the insights developed in this theoretical setting, we propose and experimentally validate a version of this algorithm which works enables applications to reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) for large language model alignment. This section is based on [2].

1.1.2 Chapter 3: Adaptive Causal Inference

This chapter focuses on developing adaptive learning algorithms for problems in Causal Inference. The problem of adaptive Causal Inference is a relatively new field and so our contributions here serve to lay the ground work, focusing on developing sample efficient algorithms with nonasympotic performance guarantees.

Clipped Second Moment Tracking (Section 3.1). In our first work, we propose an algorithm for adaptive experimental design that achieves exponential improvements in finite-sample regret. Prior to this work, the vast majority of the literature on adaptive Average Treatment Effect (ATE) estimation was concerned with designing algorithms with asymptotic optimality guarantees. In contrast, we argue that nonasymptotic guarantees are critical especially because standard applications of causal inference, such as Randomized Control Trials (RCTs) require sample efficiency, which can be obscured by prior asymptotically focused approaches. We analyze existing asymptotically optimal algorithms and demonstrate how appropriate tuning of this algorithms hyperparameters guarantees a regret bound of at most $O(\log T)$. Notably, this is a doubly exponential improvement from existing algorithms with nonasymptotic guarantees: we improve existing regret bounds from $O(\sqrt{T})$ to $O(\log T)$ and also reduce an exponential dependence on critical problem parameters to a polynomial dependence. This section is based on [3].

Optimistic Policy Tracking (Section 3.2). In this work, we develop a new algorithmic design framework for the problem of adaptive ATE estimation by demonstrating how to extend the principle of optimism from the MAB literature on regret minimization to this new setting. We instatiate this framework using the asymptotically optimal Augmented Inverse Probability Weighting (AIPW) estimator and demonstrate how to design new optimistic algorithms. Our algorithm achieves significant theoretical and empirical improvements over prior methods while maintaining strong finite-sample guarantees, demonstrating how optimistic principles from the MAB literature can be successfully applied to causal inference. This section is based on [4].

Chapter 2

Learning with Alternative Feedback Mechanisms

This chapter develops novel sampling strategies that improve learning efficiency across different problem settings. We investigate two complementary forms of alternative feedback: feedback from a different system which we wish to transfer in-order to understand properties of a related system as well as preference-based feedback in the form of pairwise comparisons. Both contributions address the fundamental challenge of how to gather information most efficiently when learning in complex, uncertain environments.

2.1 Introduction and Motivation

As machine learning systems evolve and become increasingly integrated into different aspects of society, our need to efficiently gather information in order to make decisions will grow. Traditionally, interactive learning algorithms assumce access to reward based feedback, where the utility of an action or a decision is defined by a scalar that can be directly compared to other utilities in-order to compare and access different decisions. However, in many real-world scenarios, we often do not have direct access to such reward-based feedback, and need algorithms which can operate with alternative feedback mechansisms. While this issue presents itself as a hurdle, with some additional effort, we can design new algorithms which are able to utilize these alternative feedback mechanisms and not only circumvent the need for standard feedback, but also produce algorithms that can lead to similar downstream decisions with reduced costs and increased sample efficiency. This chapter explores two form of alternative feedback where this is the case.

Transfer Learning. We develop algorithms that propagate *uncertainty* from source tasks to the target, using the transferred posterior variance as a control signal for how aggressively to sample each domain. High estimated similarity lets the algorithm lean on source data; widening discrepancies trigger a shift toward fresh target exploration. The resulting sampling rule reduces unnecessary

data collection and, when run in a pure nontransfer setting, recovers existing optimal algorithms.

Preference Learning. Rewards are often ill-defined or noisy, while it is usually easy to decide which of two outcomes is better. In such cases, algorithms based on pairwise comparisons are the right tool. We develop an active comparison strategy that directs queries to the most informative pairs, and recovers standard optimal behavior under standard reward feedback.

Both approaches illustrate the power of moving beyond naive sampling toward strategies that interrogate problem structure for maximum information gain.

2.2 Best Arm Identification under Additive Transfer Bandits

In many real-world applications of multi-armed bandits, we encounter scenarios where we want to identify the best option in a target domain but can only observe outcomes in related source domains. This work aims to develop new approaches to address these issues by introducing a new problem which intersects the ideas of transfer learning and sequential decision making. At a high-level, the problem we study involves two multi-armed bandit (MAB) instances, which we call the *source* and *target* instances, as well a *transfer function*, which is a known relationship between the two MAB instances. Within this setup we define and consider an appropriately modified variant of the (ϵ, δ)-correct best arm identification (BAI) objective [5, 6].

Motivating Examples. We start off by highlighting various scenarios where the need to transfer knowledge between sequential decision making problems arise:

- **Clinical Trials.** The first scenario we consider is the application of MABs to clinical trials [7]. In this context, the arms can be thought of as the different treatments and we wish to determine which is most effective. A standard practice in this setup is to test treatments on animals before transitioning to clinical trials for humans. Ideally, we wish to identify the optimal treatments for humans by only testing the treatments on animals. Here, we can view the animal trials as the source domain, and human trials as the target domain.
- **Sim-to-Real Transfer in Reinforcement Learning.** A popular paradigm for 'cheap' reinforcement learning is sim-to-real transfer in reinforcement learning [8–10]. In the sim-to-real problem, the objective is to learn a robot's control policy for the real world (target domain) while restricting training to computer simulations (source domain). Currently, in the sim-to-real literature, most algorithms rely on heuristics to learn these control policies – typically by ensuring that a sufficiently diverse set of environments are encountered during training. While some of these heuristics have proven to be successful, our theoretical understanding of this problem remains in its infancy. We believe that studying our proposed problem is a first step

towards gaining a better understanding of how to transfer knowledge in more complicated sequential decision making problems.

- **Rate adaptation in wireless networks.** Rate allocation in wireless networks has been posed as a bandit optimization problem under fixed channel conditions [11, 12]. However, it is important to adapt the rate allocation according to varying channel conditions by transferring rate allocation policies between related channel conditions.

Contributions. The main contributions of this work are

1. A general additive transfer framework that which not only captures the idea of 'transfer' in a sequential decision making framework, but also unifies many existing pure exploration problems under a single framework.
2. A novel algorithm for this framework, Transfer-LUCB, which generalizes the celebrated LUCB algorithm, along with a theoretical analysis providing sample complexity guarantees.
3. Instantiations showing how our general bounds recover and extend known results for problems like TopK identification and thresholding bandits

Related Works

The work most closely resembling ours is a recent line of work on obtaining sample complexity guarantees for Monte Carlo tree search algorithms [13–15]. Specifically, Huang et al. [15] approach this problem by first introducing the more general structured BAI problem. Their structured BAI framework is the same as our transfer BAI framework, however we choose to use a different name to both emphasize that we are transferring knowledge between multiple MAB instances and to avoid confusing the structured BAI problem with the structured MAB framework described in Lattimore and Munos [16] and Gupta et al. [17].

While Huang et al. [15] give a general algorithm for their structured BAI problem, their primary objective was to derive algorithms for the Monte Carlo tree search problem. As such, their assumptions consequently make their algorithm inapplicable to wide range of settings including the simple linear setting discussed in Section 2.2.1. Their Assumption 2(i), which requires the transfer function to be component-wise monotonic, already restricts the applicability of their algorithm to a wide range of problems. However, we can resolve this issue by using our confidence sequence construction given in Section 2.2.2. Their Assumption 2(ii), however, is more troublesome as it requires the confidence sequence of each target arm to be contained in the confidence sequence of at least one source arm. To resolve this, Huang et al. [15] briefly mention a weaker assumption wherein the confidence sequence of each target arm must be contained in a scaled and shifted version of a source arm's confidence sequence — however, this weaker assumption is still inapplicable even in the linear setting. Additionally, as we show in Appendix 2.2.6, the resulting sample complexity for this modified algorithm is significantly worse than the sample complexity of our algorithm. Finally,

we note that the assumptions we make are incomparable to the assumptions made in Huang et al. [15] as neither is more or less general than the other.

The simpler linear setting subsumed by our framework, where the transfer function takes the form $f(\mu) = \mathbf{A}\mu$ also coincides with the Transductive Linear Bandit problem studied in Fiez et al. [18] and Katz et al. [19] when the sampling vectors are the standard basis of $\mathbb{R}^S$. However, it is not clear how to extend the ideas presented in these works to the additive setting since the algorithms strongly utilize the linearity in the problem.

The 'partition identification' problem introduced by Juneja and Krishnasamy [20] is also related to our work. In fact, their framework can be seen as a generalization of the problem studied here. However, in their work, Juneja and Krishnasamy [20] primarily focus on providing lower bounds for variations of the partition identification problem and only briefly discuss an asymptotically optimal algorithm towards the end of their work. Additionally, it is known that Confidence-Interval style algorithms (like the one we propose) outperform their Track-And-Stop style algorithm in so-called moderate-confidence regimes[1] [21]. Moreover, it is not clear that the algorithm they provide is can even implementable in the linear setting because implementing it requires solving a constrained optimization problem over a (possibly) non-convex set. Finally, the analysis in [20] only provides asymptotic guarantees for their algorithm while we provide explicit finite-time guarantees for our algorithm.

2.2.1 Problem Setup

Before introducing the transfer BAI problem, we briefly review the ϵ-BAI problem within the MAB framework. In our notation, we define an n-armed MAB instance to be a set of n tuples $\{(P_i, \mu_i)\}_{i=1}^n$ where $P_i \in \mathcal{P}$ is a probability distribution in some known set $\mathcal{P}$ and $\mu_i \coloneqq \mathbb{E}_{P_i}[X]$ is the mean of P_i. For example, $\mathcal{P}$ could be the set of all sub-Gaussian distributions. In this setup, an algorithm interacts with the MAB instance through a round-based protocol. In each rounds, t, the learner selects an arm $I_t \in \{1, \dots, n\}$, and observes a sample $X_t \sim P_{I_t}$. For the ϵ-BAI problem, the objective is to identify an ϵ-optimal arm $\widehat{a}$ satisfying $\mu_{\widehat{a}} + \epsilon \geq \max_{i \in [n]} \mu_i$, where $[n] = \{1, \dots, n\}$.

This problem is often studied in the so-called *fixed-confidence* setting in which a confidence parameter δ is given and an algorithm is said to be correct if, with probability greater than $1 - \delta$, it stops and returns an ϵ-optimal arm. For any fixed MAB instance, an algorithm's performance is then judged by either a *high-probability* or an *in expectation* upper-bound on the number of samples required to identify an ϵ-optimal arm. *In this work, we will give a high probability bound for a variant of the fixed-confidence setting that naturally arises in our setup.*

[1] By moderate confidence regimes we mean regimes where δ is moderately small, i.e when $\delta \approx .05$ or when it is inverse-polynomial in the number of measurements [21].

Transfer Best Arm Identification

We are now ready to introduce the transfer BAI problem which can be stated as a tuple $(\{\mathcal{S}_i, \mu_i\}_{i=1}^S, \{\mathcal{T}_a, \nu_a\}_{a=1}^T, f)$. Here, $\{\mathcal{S}_i, \mu_i\}_{i=1}^S$ and $\{\mathcal{T}_a, \nu_a\}_{a=1}^T$ are S and T-armed MAB instances which we respectively call the *source* and *target* MAB instances and $f\colon \mathbb{R}^S \mapsto (\mathbb{R}^+)^T$ is a *known* multivariate function which we call the *transfer function*. Here, we have written $\mathbb{R}^+ \coloneqq \mathbb{R} \cup \{\infty, -\infty\}$ to denote the extended real numbers. Specifically, f relates the means of the target and sources arms in the sense that

$$\nu = f(\mu),$$

where $\mu = (\mu_1, \ldots, \mu_S)$ and $\nu = (\nu_1, \ldots, \nu_T)$ refer to the vector of means for the source and target MAB instances.

In this paper we study the special setting in which f is an additive function satisfying

$$\nu_a = f_a(\mu) = \sum_{i=1}^{S} f_{a,i}(\mu_i).$$

Here, and in the rest of this paper, i will always be used to index source arms, and unless otherwise specified, a will be used to index target arms. As we discuss more in Section 2.2.1, this additive setting is already interesting as it captures a large number of existing problems in addition to introducing new problems.

Motivating Examples

To provide more concrete intuition about our algorithm and sample complexity analysis, we will use two running examples: property testing and linear transfer functions.

Property Testing. In the property testing problem we are interested in identifying all arms $i \in [S]$ which satisfy some property $\mu_i \in \mathcal{C}_i \subset \mathbb{R}$. Our additive transfer framework is able to capture this problem. To do so, we first define

$$\mathbb{I}_{\mathcal{C}}(\mu) = \begin{cases} 1 & \mu \in \mathcal{C}, \\ -\infty & \mu \notin \mathcal{C}. \end{cases} \tag{2.1}$$

Then for each set $M \in 2^{[n]}$ we define a target arm whose mean is $\nu_M = \sum_{i \in M} \mathbb{I}_{\mathcal{C}_i}(\mu_i)$. Clearly, the optimal target arm will be a function of all source arms for which $\mu_i \in \mathcal{C}_i$. We note that whenever we refer to the property testing problem, we will index the target arms with M instead of a. Additionally, for the property testing problem, we require $\epsilon = 0$.

Linear Transfer Functions. Another useful special case for contextualizing our results is the setting where the transfer function is a linear transformation of the source means, so that

$$\nu_a = \sum_{i=1}^{S} \mathbf{A}_{a,i}\mu_i.$$

In our proposed framework, we restrict our ability to sample from the target arms, and only consider algorithms which are able to sample from the source arms. We note that studying the problem where we have the ability to sample from both the target and source domains is an interesting problem for future work. Our objective is to develop algorithms which will return an ϵ-optimal target arm with high probability. Formally, we focus on an appropriately modified version of the fixed-confidence setting which we define as follows:

Definition 2.2.1 ((ϵ, δ)-correct)**.** *For any $\epsilon \geq 0$ and $\delta \in (0, 1)$, we say that an algorithm `Alg` is (ϵ, δ)-correct for the transfer BAI problem if, with probability at least $1 - \delta$, and for every problem instance $(\{\mathcal{S}_i, \mu_i\}_{i=1}^{S}, \{\mathcal{T}_a, \nu_a\}_{a=1}^{T}, f)$, `Alg` stops and returns an ϵ-optimal arm $\widehat{a} \in [T]$ satisfying $\nu_{\widehat{a}} + \epsilon \geq \max_{a \in [T]} \nu_a$.*

As is standard with typical BAI algorithms, an algorithm for the transfer BAI problem is comprised of three components: a sampling rule, a stopping rule, and a selection rule. Letting $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ denote the σ-algebra generated by the observations from the source arms up until time t, we have

1. a sampling rule, p_t, which is a $\mathcal{F}_{t-1}$-measurable function which selects the source arms to pull during round t;
2. a stopping rule, τ, which is a $\mathcal{F}_t$-measurable random variable which determines when the algorithm stops;
3. a selection rule, $\widehat{a}$, which is a $\mathcal{F}_\tau$-measurable function which outputs a guess of the optimal target arm a^*.

Assumptions

Before proceeding, we briefly discuss our assumptions. Our first assumption places restrictions on the class of additive transfer functions which our algorithm is able to handle.

Assumption 2.2.2 (Assumptions on f)**.** *We assume that $f_{a,i}$ is continuous at μ_i for all $(a, i) \in [T] \times [S]$.*

We additionally assume that the observations from the source MAB instances are sub-Gaussian.

Assumption 2.2.3 (σ-sub-Gaussian Observations)**.** *We assume that the observations from the source arms are σ-sub-Gaussian so that for any $i \in [S]$ and $\lambda \in \mathbb{R}$ the following holds*

$$\log \mathbb{E}_{X \sim \mathcal{S}_i}[\exp\{\lambda(X - \mu_i)\}] \leq \frac{\lambda^2\sigma^2}{2}. \tag{2.2}$$

This assumption is necessary for the concentration inequalities used in the construction of our LUCB-style algorithm given in the next section. We note that this, with minimal modification, our assumption, algorithm, and the resulting sample complexity analysis can accommodate arbitrary sub-ψ observations through the use of the concentration inequalities given by Howard et al. [22] — in Assumption 2.2.3, we have implicitly set $\psi(\lambda) = \frac{\lambda^2}{2}$. However, to simplify the exposition, we limit the scope of this work to sub-Gaussian observations. Finally, without loss of generality, we assume that the means are ordered in decreasing order so that $\mu_1 \geq \mu_2 \ldots \geq \mu_S$ and $\nu_1 \geq \nu_2 \geq \ldots \geq \nu_T$. We only require the optimal target arm to be unique when $\epsilon = 0$.

Subsumed Settings

Finally, as we alluded in the previous subsection, we now describe how the additive-transfer framework studied here subsumes a range of existing pure exploration problems. In the next section, we instantiate our sample complexity results for some of the problems mentioned below.

TopK Identification. In the TopK problem [23, 24], the objective is to identify the K arms with the largest means. To recover this problem in our formulation, we define the target means as follows. We define a target arm $\mathcal{T}_M$ for each set $M \in 2^{[T]}$ satisfying $|M| = K$. The mean of this target arm is then defined as $\nu_M = \sum_{i \in M} \mu_i$.

Thresholding Bandits. In the Thresholding Bandits problem [25], the objective is to identify the set of arms whose means are greater than some fixed threshold $\underline{\mu} \in \mathbb{R}$. This problem is subsumed by the property testing problem mentioned earlier. To see this, we simply set, for each $i \in [S]$, $\mathcal{C}_i = (\underline{\mu}, \infty)$. Then for every set $M \in 2^{[n]}$ define the mean of target arm $\mathcal{T}_M$ as $\nu_M = \sum_{i \in M} \mathbb{I}_{\mathcal{C}_i}(\mu_i)$.

Combinatorial Pure Exploration. As a final example, we show how our framework generalizes the Combinatorial Pure Exploration problem proposed by Chen et al. [26–29]. This problem is defined by a decision class $\mathcal{M} \subseteq 2^{[S]}$ and the objective is to identify an element $M \in \mathcal{M}$ satisfying $M \in \operatorname{argmax} M \in \mathcal{M} \sum_{i \in M} \mu_i$. It is easy to see that this problem fits into our framework by defining a target mean $\nu_M = \sum_{i \in M} \mu_i$. The Combinatorial Pure Exploration problem additionally subsumes a number of additional problems previously studied in the literature, including the examples discussed above. For more examples of subsumed problems and additional discussions, we refer the reader to the literature on this problem [26–30].

2.2.2 Algorithm

In this section, we present the Transfer LUCB (T-LUCB) algorithm, a variant of the LUCB algorithm [23] used in the fixed-confidence BAI setting. Like the LUCB algorithm, our T-LUCB algorithm is based on constructing confidence sequences which are time-uniform confidence intervals on

the sample means. Before presenting the T-LUCB algorithm, we first discuss the construction of our confidence sequences.

To construct the confidence sequences on the source arms we use standard Hoeffding-like confidence sequences [22, 31] and define the Lower Confidence Bound (LCB), Upper Confidence Bound (UCB), and Confidence Interval (CI) sequences as follows. Recall that I_s denotes the arm that is pulled at time s. We let $N_i(t) = \sum_{s=1}^{t-1} \mathbb{I}I_s = i$ denote the number of times that source arm i has been pulled at the start of round t. Additionally, we let $\widehat{\mu}_t(i) = \frac{1}{N_i(t)} \sum_{s=1}^{t-1} X_s \mathbb{I}I_s = i$ denote the empirical mean of arm i at the beginning of round t. Then, at $t = 0$, we set the lower and upper confidence bounds for source arm i as $\mathcal{L}_{\mathcal{S}}(0, i, \delta) = -\infty$, $\mathcal{U}_{\mathcal{S}}(0, i, \delta) = +\infty$. Next, for $t \geq 1$, we recursively define the confidence sequences as:

$$\mathcal{U}_{\mathcal{S}}(t, i, \delta) \coloneqq \min \left\{ \mathcal{U}_{\mathcal{S}}(t-1, i, \delta), \widehat{\mu}_t(i) + \beta(N_i(t), \delta/(2S)) \right\}, \tag{2.3}$$

$$\mathcal{L}_{\mathcal{S}}(t, i, \delta) \coloneqq \max \left\{ \mathcal{L}_{\mathcal{S}}(t-1, i, \delta), \widehat{\mu}_t(i) - \beta(N_i(t), \delta/(2S)) \right\}, \tag{2.4}$$

$$\mathrm{CI}_{\mathcal{S}}(t, i, \delta) \coloneqq [\mathcal{L}_{\mathcal{S}}(t, i, \delta), \mathcal{U}_S(t, i, \delta)]. \tag{2.5}$$

Here $\beta(\cdot, \cdot)$ is a function which controls the rate at which the confidence intervals shrink. As an example, β can be taken to be the so-called "polynomial stitched boundary" [22, Eq.(6)]:

$$\beta(t, \delta) \coloneqq 1.7 \sqrt{\frac{\sigma^2 \log\log\left(2t\sigma^2\right) + 0.72 \log \frac{5.2}{\delta}}{t}}. \tag{2.6}$$

More generally, for the results given in Section 2.2.3 to hold, β must satisfy the following condition:

$$\mathbb{P}\left\{ \exists t \geq 1 : \mu_i \notin \mathrm{CI}_{\mathcal{S}}(t, i, \delta) \right\} \leq \delta. \tag{2.7}$$

The choice of β in Eq. (2.6) satisfies the above condition.

Next, we use the source arm confidence sequences to construct confidence sequences on the target arms as follows:

$$\mathcal{L}_{\mathcal{T}}(t, a, \delta) \coloneqq \sum_{i=1}^{S} \min_{m_i \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_{a,i}\left(m_i\right), \tag{2.8}$$

$$\mathcal{U}_{\mathcal{T}}(t, a, \delta) \coloneqq \sum_{i=1}^{S} \max_{m_i \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_{a,i}\left(m_i\right), \tag{2.9}$$

$$\mathrm{CI}_{\mathcal{T}}(t, a, \delta) \coloneqq [\mathcal{L}_{\mathcal{T}}(t, a, \delta), \mathcal{U}_{\mathcal{T}}(t, a, \delta)]. \tag{2.10}$$

The intuition for the above construction is as follows. By constructing the source confidence se-

quences as defined in equations Eq. (2.3) and Eq. (2.4), and choosing β to satisfy condition Eq. (2.7), we can control the deviations of the source samples means from the true source means. This in turn implies that the constructed target confidence sequences are well-behaved in the sense that they will contain the target arm means with high probability. This intuition is formalized by Lemma 2.2.13 in the Appendix.

The T-LUCB Algorithm

We are now ready to introduce the T-LUCB algorithm which is stated in Algorithm 1. During each round, the algorithm selects two target arms B_t and C_t with the objective of separating the LCB of B_t from the UCB of C_t. After selecting B_t and C_t, the algorithm samples the source arms I_t and J_t which respectively have the largest contributions to the length of the confidence sequences of B_t and C_t. Formally, we define the following quantity

$$L(i, a, t) = \max_{m \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_{a,i}(m) - \min_{m \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_{a,i}(m), \tag{2.11}$$

which quantifies the amount of uncertainty that source arm i contributes to target arm a. The algorithm stops when the LCB of B_t is greater than the UCB of C_t. Finally the algorithm selects B_t as its guess for the optimal target arm.

Input $\delta > 0, \epsilon \geq 0$, f, σ^2;
Sample each source arm once;
for $t = 1, 2, \ldots$ **do**
  $B_t = \mathrm{argmax}_{a \in [T]} \mathcal{L}_{\mathcal{T}}(t, a, \delta)$;
  $C_t = \mathrm{argmax}_{a \in [T], a \neq B_t} \mathcal{U}_{\mathcal{T}}(t, a, \delta)$;
  if $\mathcal{L}_{\mathcal{T}}(t, B_t, \delta) + \epsilon \geq \mathcal{U}_{\mathcal{T}}(t, C_t, \delta)$ **then**
    return $\widehat{a} = B_t$;
  end
  $I_t = \mathrm{argmax}_{i \in [S]} L(i, B_t, t)$;
  $J_t = \mathrm{argmax}_{i \in [S]} L(i, C_t, t)$;
  Observe $X_{t,1} \sim \mathcal{S}_{I_t}$ and $X_{t,2} \sim \mathcal{S}_{J_t}$;
end

Algorithm 1: Additive Transfer LUCB

2.2.3 Theoretical Analysis

In this section, we analyze the T-LUCB algorithm presented in the previous section. Our first result shows that, regardless of the sampling rule, the stopping rule and selection rule of Algorithm 1 are sufficient to give an (ϵ, δ)-correct algorithm. The proof of this result can be found in the Appendix.

Theorem 2.2.4. *Suppose that β satisfies condition Eq. (2.7). Then, any algorithm which stops*

when there exists an arm $a \in [T]$ *such that*

$$\mathcal{L}_{\mathcal{T}}(t, a, \delta) + \epsilon \geq \mathcal{U}_{\mathcal{T}}(t, a', \delta), \tag{2.12}$$

for all $a' \neq a$*, and selects the arm* $\widehat{a} = a$*, will with probability at least* $1-\delta$*, choose an arm satisfying* $\nu_{\widehat{a}} \geq \nu_1 - \epsilon$*.*

We now shift our attention towards providing a high probability upper bound on the sample complexity of Algorithm 1. To present our function specific upper-bound on the sample complexity we first introduce some additional notation. We remark that due to the generality of our framework our generic sample complexity bound is presented implicitly, and is difficult to immediately interpret. As such, we will present explicit bounds for some instantiations of our problem in the following subsection.

First, we define

$$s_a \coloneqq |\{i : f_{a,i}(x) \neq f_{a,i}(y),\ \forall x, y \in \mathbb{R}\}|, \tag{2.13}$$

which measures the number of source arms which contribute to the uncertainty of a target arm. For the property testing problem, $s_a = |M|$, which is the number of terms in the sum $\sum_{i \in M} \mathbb{I}_{\mathcal{C}_i}(\mu_i)$. For linear transfer functions, $s_a = |\{i : \mathbf{A}_{a,i} \neq 0\}$ which measures the sparsity of the vector $\mathbf{A}_a$.

Next, with a slight abuse of notation, we define the following quantity which has a similar form to equation 2.11

$$L(i, a, t, x) = \max_{m \in [x, x+2\beta(t,\delta)]} f_{a,i}(m) - \min_{m \in [x, x+2\beta(t,\delta)]} f_{a,i}(m). \tag{2.14}$$

This term quantifies how much source arm i contributes to the confidence interval of target arm a when the LCB of source arm i is x. For the property testing problem, we have

$$L(i, M, t, x) = \begin{cases} 0 & \text{if } [x, x + 2\beta(t,\delta)] \subseteq \mathcal{C}_i, i \in M \\ 0 & \text{if } [x, x + 2\beta(t,\delta)] \subseteq \mathcal{C}_i^c, i \in M \\ 0 & \text{if } i \notin M \\ \infty & \text{otherwise} \end{cases}, \tag{2.15}$$

where $\mathcal{C}_i^c$ is the complement $\mathcal{C}_i$ and we have taken the convention that $\infty - \infty = 0$. For linear transfer functions, this quantity is independent of x so that $L(i, a, t, x) = 2|\mathbf{A}_{a,i}|\beta(t, \delta)$.

Having defined this quantity, we are now ready to define an upper bound on the number of times source arm i needs to be sampled in order to determine if target arm a is ϵ-optimal. First, we set

$$\tau_{a,i} = \min \left\{ t \in \mathbb{N} : \sup_{x \in [\mu_i - 2\beta(t,\delta), \mu_i]} L(i, a, t, x) < \frac{\max\{|\bar{\nu}_{1,2} - \nu_a|, \epsilon/2\}}{s_a} \right\}, \tag{2.16}$$

where $\bar{\nu}_{1,2} \coloneqq \frac{\nu_1+\nu_2}{2}$. Then, we define

$$\tau_i = \max_{a\in[T]} \tau_{a,i}, \tag{2.17}$$

which represents the number of times source arm i must be pulled in order to determine which target arms are ϵ-optimal. We are now ready to state our sample complexity result.

Theorem 2.2.5 (Sample Complexity Upper Bound of Algorithm 1)**.** *Let τ denote the stopping time of Algorithm 1. Then with probability at least $1-\delta$, we have that*

$$\tau \leq \sum_{i\in[S]} \tau_i. \tag{2.18}$$

Note that this sample complexity bound is independent of the number of target arms. This fact allows us to recover the sample complexity of some existing problems as we show in the following subsection.

Theorem 2.2.5 implies the following sample complexity result for the property testing problem.

Corollary 2.2.6. *Let τ denote the stopping time of Algorithm 1 for the property testing problem and define*

$$H \coloneqq \sum i = 1S \frac{2}{\Delta^2_{\mathcal{C}_i}(\mu_i)}, \tag{2.19}$$

where

$$\Delta_{\mathcal{C}_i}(\mu_i) = \begin{cases} \inf_{x\in\mathcal{C}_i^c} |x-\mu_i| & \text{if } \mu_i \in \mathcal{C}_i \\ \inf_{x\in\mathcal{C}_i} |x-\mu_i| & \text{if } \mu_i \notin \mathcal{C}_i \end{cases}.$$

Then[2] *with probability at least $1-\delta$,*

$$\tau \leq \tilde{O}\left(H \log\left(\frac{1}{\delta}\right)\right). \tag{2.20}$$

For linear transfer functions, we obtain the following result.

Corollary 2.2.7. *Let τ denote the stopping time of Algorithm 1 for the linear transfer setting and define*

$$H_\epsilon(\mathbf{A},\nu,\mu) \coloneqq \sum i = 1S \max_{a\in[T]} \left\{ \frac{s_a^2 |\mathbf{A}_{a,i}|^2}{\max\left\{|\nu_a - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\}^2} \right\}. \tag{2.21}$$

Then with probability at least $1-\delta$,

$$\tau \leq \tilde{O}\left(H_\epsilon(\mathbf{A},\nu,\mu) \log\left(\frac{1}{\delta}\right)\right). \tag{2.22}$$

[2]We use $\tilde{O}$ to refer to sample complexity results which are correct up to constant and log log factors.

Instantiations of the Sample Complexity Bound

We now proceed to instantiate the sample complexity bound of Theorem 2.2.5 for some previously studied settings. In each of these settings we state an explicit bound which is a direct corollary of the sample complexity bound from Theorem 2.2.5. Proofs of these results can be found in the Appendix.

BAI. To recover the Best Arm Identification problem, we simply set $\nu_a = \mu_i$ so that the mean of each target arm is simply the mean of one of the source arms. First, we set $\overline{\mu} = \frac{\mu_1+\mu_2}{2}$. Then Theorem 2.2.5 implies that

$$\tau \leq \tilde{O}\left(\sum_{i=1}^{S} \frac{1}{(\overline{\mu}-\mu_i)^2}\log(1/\delta)\right).$$

This recovers the sample complexity of the original LUCB algorithm [23].

Thresholding Bandits. Here, Theorem 2.2.5 implies that

$$\tau \leq \tilde{O}\left(\sum_{i\in[S]} \frac{1}{(\mu_i-\underline{\mu})^2}\log(1/\delta)\right),$$

which matches, up to iterated logarithmic factors, the problem's sample complexity lower bound given for the fixed confidence setting [25].

TopK. One example of a Combinatorial Pure Exploration problem is the so-called TopK problem where we wish to identify the K largest means our of S arms. This problem can be recovered in the CPE framework by letting $\mathcal{M}$ to be the all subsets of $\{1,\ldots,S\}$ with cardinality K. To state our sample complexity results in this setup, we first define $\overline{\mu} = \frac{\mu_K+\mu_{K+1}}{2}$. Then, Theorem 2.2.5 implies that

$$\tau \leq \tilde{O}\left(\sum_{i\in[S]} \frac{K^2}{(\mu_i-\bar{\mu})^2}\log(1/\delta)\right).$$

We remark that this sample complexity result is suboptimal by a factor of K^2 [23, 24]. However, we conjecture that this is the price of generality of our framework. We refer the reader to the conclusion for more discussion on this.

2.2.4 Experiments

The theoretical paper focused on establishing the fundamental framework and theoretical guarantees for additive transfer bandits. While the original conference paper did not include experimental validation, the theoretical results provide important insights into the performance characteristics of the T-LUCB algorithm.

The sample complexity bounds demonstrate that:

1. The algorithm recovers classical BAI sample complexity when applied to standard best arm identification
2. For thresholding bandits, the bound matches known lower bounds up to logarithmic factors
3. The framework generalizes to combinatorial pure exploration problems, though with some loss in optimality for specific cases like TopK identification

Future experimental work would validate these theoretical predictions and explore the practical performance of the algorithm across different transfer scenarios, including cases where the transfer function relationship varies in strength and the robustness of the approach when the additive assumption is violated.

The key experimental questions that remain to be addressed include:

- Performance comparison with specialized algorithms for specific subproblems
- Sensitivity to misspecification of the transfer function
- Computational efficiency for large-scale problems
- Real-world applications in domains like clinical trials and reinforcement learning

2.2.5 Conclusion

In this work we presented and analyzed an algorithm for leveraging additive relationships between two MAB instances to identify the best arm in a MAB instance without ever sampling from it. The T-LUCB algorithm provides a principled approach to transfer learning in multi-armed bandits with theoretical guarantees.

Key Contributions Our main contributions include:

1. **General Framework**: We introduced the additive transfer bandit framework that encompasses many existing pure exploration problems while enabling new transfer learning scenarios.
2. **Algorithm Design**: The T-LUCB algorithm extends the classical LUCB approach to handle transfer relationships through carefully constructed confidence sequences for both source and target arms.
3. **Theoretical Analysis**: We provided sample complexity bounds that are independent of the number of target arms and recover known results for special cases while establishing new bounds for the general additive transfer setting.
4. **Problem Unification**: We demonstrated how our framework subsumes important problems including TopK identification, thresholding bandits, and combinatorial pure exploration.

Limitations and Future Directions Several important directions emerge from this work:

- **Optimality.** A first direction for future work would be to investigate if an algorithm for the additive transfer setting can recover the correct sample complexity results for the specialized settings such as the TopK problem. We conjecture that this is not possible. This is because algorithms for these simpler settings either implicitly or explicitly utilize a type of well-ordering property of the problem which does not generally hold for non-linear additive transfer functions. This well-ordering property is made explicit in the work of Gabillon et al. [30], and is implicitly utilized in the work of Fiez et al. [18].
- **Unknown Transfer Functions.** An issue with our proposed framework is that we assume the transfer function is known in advance. Another interesting direction of future research is to study how to alleviate this requirement so that, for example, the transfer function can be learned from historical data. If this approach is taken, it may no longer be possible to identify a ϵ-optimal target arm as the error introduced from estimating the transfer function might lead to a scenario where the true optimal target arm is not the optimal target arm under the approximate transfer function. We believe in this setting a more reasonable criterion to study is the *simple regret* [5] under the assumption that the learned transfer function is close in norm to the true transfer function.
- **Bi-directional Transfer.** Furthermore, in this work we consider the setting where we are unable to sample from the target MAB instance. Another interesting direction would be in developing algorithms which are able to sample from the target MAB instance with the caveat that doing so has some additional cost. This type of setting seems natural as it is often the case that making direct measurements of some system can be significantly more expensive than taking noisier auxiliary measurements of the system. A concrete example of this is in the sim-to-real problem, where collecting observations from the real world is significantly more expensive than collecting observations from a computer simulation. Additionally, the ability to sample the target arm can allow for learning or refining the transfer function on the fly using few transfer queries.

This work establishes the theoretical foundation for transfer learning in multi-armed bandits and opens up numerous avenues for both theoretical and practical extensions.

2.2.6 Detailed Comparison with Prior Work

In this section we provide an in-depth discussion and comparison of our Algorithm 1 and a variant of the Micro-LUCB algorithm which is suitable for linear transfer functions. We first restate their assumptions and demonstrate why the do not hold for our setting. In this assumption, we note that $\leq$ denotes a component wise ordering so $u \leq v$ is equivalent to stating $u_i \leq v_i$ for all i.

Assumption 2.2.8 (Assumption 2 of [15])**.** *The following hold:*

1. *The mapping function f is monotonous with respect to the partial order of vectors: for any $u, v \in \mathbb{R}^S$, $u \leq v$ implies $f(u) \leq f(v)$.*

2. *For any* $u, v \in \mathbb{R}^S$, $u \leq v$, $a \in [T]$, *the set* $D(a, u, v) \coloneqq \{i \in [S] : [f_a(u), f_a(v)] \subset [u_i, v_i]\}$ *is non-empty.*

To see that Assumption 2.2.8 (i) is not satisfied for arbitrary linear transformations, we set some entries of the associated matrix to be negative, then there will exist some a for which f_a is not monotonous. This assumption is used to define the confidence intervals on the target arms, and without it, their proof of correctness does not hold. We modify the assumption to the following which trivially holds true for any function:

Assumption 2.2.9. *The mapping function* f *is monotone with respect to the partial order of vectors: for any* $u, v \in \mathbb{R}^S$, $u \leq v$ *implies* $\min_{u \leq m \leq v} f(m) \leq \max_{u \leq m \leq v} f(m)$.

It can be verified that if our target confidence sequences are constructed as

$$\mathcal{L}_{\mathcal{T}}(t, a, \delta) \coloneqq \min_{m_i \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_a(m), \tag{2.23}$$

$$\mathcal{U}_{\mathcal{T}}(t, a, \delta) \coloneqq \max_{m_i \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_a(m), \tag{2.24}$$

$$\mathrm{CI}_{\mathcal{T}}(t, a, \delta) \coloneqq [\mathcal{L}_{\mathcal{S}}(t, i, \delta), \mathcal{U}_{\mathcal{S}}(t, i, \delta)], \tag{2.25}$$

then the T-LUCB stopping rule and selection rule can be applied to any algorithm to give an (ϵ, δ)-correct algorithm. The proof of this is a simple modification of the proof of Theorem 2.2.4 where we simply replace the construction of the target confidence sequences given in Section 2.2.2 with the construction defined above.

We now switch our attention to Assumption 2.2.8 (ii). In short, Assumption 2.2.8 (ii) requires that for each target arm confidence interval, there exists at least one source arm confidence interval which contains the target arm confidence interval. This assumption is used to determine the set of source arms which should be sampled in the Micro-LUCB algorithm. Indeed, it is integral for the algorithm since, if the assumption is not satisfied, the sampling rule is not well defined. While this assumption is not directly satisfied for the linear setting, [15] mention one avenue for weakening the assumption so that it is satisfied for a larger class of functions. This weaker assumption is as follows:

There exists some $a > 0, b \in \mathbb{R}$ such that for any $u, v \in \mathbb{R}^S$, $u \leq v$, $a \in [T]$, the set $\tilde{D}(a, u, v) = \{i \in [S] : [f_a(u), f_a(v)] \subset [au_i + b, av_i + b]\}$ is non-empty.

However, this assumption also is not well defined as $[f_a(u), f_a(v)]$ is not an interval unless f_a is component-wise monotonically increasing. To fix this, we propose the following assumption:

Assumption 2.2.10 (Modified Assumption 2(ii) of [15])**.** *There exists some* $a_i > 0, b_i \in \mathbb{R}$ *such that for any* $u, v \in \mathbb{R}^S$, $u \leq v$, $a \in [T]$, *the set* $\tilde{D}(a, u, v) = \{i \in [S] : [\min_{u \leq m \leq v} f_a(m), \max_{u \leq m \leq v} f_a(m)] \subset [a_i u_i + b_i, a_i v_i + b_i]\}$ *is non-empty.*

Remark 2.2.11. *This modified assumption is indeed a generalization of the previous assumption, which can be seen by taking* $a = 1, b = 0$.

This assumption then gives rise to a modified version of the Micro-LUCB algorithm which we

state in Algorithm 2.

Sample each source arm once.;
for $t = 1, 2, \ldots$ **do**
  $B_t = \text{argmax}\, a \in [T] \mathcal{L}_{\mathcal{S}}(t, a, \delta)$;
  $C_t = \text{argmax}\, a \in [T], a \neq B_t \mathcal{U}_{\mathcal{T}}(t, a, \delta)$;
  Choose any I_t from $\tilde{D}(B_t, \mathcal{L}_{\mathcal{S}}(t, B_t, \delta), \mathcal{U}_{\mathcal{S}}(t, B_t, \delta))$;
  Choose any J_t from $\tilde{D}(C_t, \mathcal{L}_{\mathcal{S}}(t, C_t, \delta), \mathcal{U}_{\mathcal{S}}(t, C_t, \delta))$;
  Observe $X_{t,1} \sim S_{I_t}$ and $X_{t,2} \sim S_{J_t}$;
  Update $[\mathcal{L}_{\mathcal{S}}(t, I_t, \delta), \mathcal{U}_{\mathcal{S}}(t, I_t, \delta)]$ and $[\mathcal{L}_{\mathcal{S}}(t, J_t, \delta), \mathcal{U}_{\mathcal{S}}(t, J_t, \delta)]$;
  if $\mathcal{L}_{\mathcal{T}}(t+1, B_t, \delta) \geq \mathcal{U}_{\mathcal{T}}(t+1, C_t, \delta)$ **then**
    $\widehat{a} \leftarrow B_t$;
    return $\widehat{a}$;
  end
end

Algorithm 2: Modified Micro-LUCB

It can be shown that only 'diagonal' matrices satisfy the above assumption. We demonstrate this in the case $A \in \mathbb{R}^{2\times 2}_{\geq 0}$ through the following proposition:

Proposition 2.2.12. *Let $A \in \mathbb{R}^{2\times 2}_{\geq 0}$. Suppose A satisfies Assumption 2.2.10, then for $i = 1, 2$, either $A_{i1} = 0$ or $A_{i2} = 0$.*

Proof. Let

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{ij} \geq 0$. Without loss of generality, we assume that $i = 1$ and $A_{11} \neq 0$, and we will demonstrate that this necessarily implies that $A_{12} = 0$. First, under Assumption 2.2.10, we know that

$$b_1 \leq A_{11}u_1 + A_{12}u_2 - a_1u_1, \tag{2.26}$$

$$b_1 \geq A_{11}v_1 + A_{12}v_2 - a_1v_1. \tag{2.27}$$

Suppose we pick v_1 to satisfy

$$v_1 \geq \frac{A_{11}(u_1 - v_1) + A_{12}(u_2, v_2)}{a_1} + u_1.$$

Some straightforward algebra shows that

$$A_{11}u_1 + A_{12}u_2 - a_1u_1 \leq A_{11}v_1 + A_{12}v_2 - a_1v_1.$$

The above inequality then implies that

$$b_1 \leq A_{11}u_1 + A_{12}u_2 - a_1u_1 \leq A_{11}v_1 + A_{12}v_2 - a_1v_1 \leq b_1,$$

which is only possible when

$$A_{11}u_1 + A_{12}u_2 - a_1u_1 = A_{11}v_1 + A_{12}v_2 - a_1v_1. \tag{2.28}$$

To see this is a contradiction, we rearrange equation Eq. (2.28) and observe that the following must hold for all $u \leq v$:

$$A_{12}(v_2 - u_2) = (A_{11} - a_1)(u_1 - v_1).$$

However, this is cannot hold for all $u \leq v$ unless $A_{12} = (A_{11} - a_1) = 0$. This implies that $A_{12} = 0$. Therefore, $A_{12} = 0$, as desired. (The same argument can be repeated to show that if $A_{12} \neq 0$, we must have $A_{11} = 0$). □

2.2.7 Proofs of Results

This section contains the proofs for the results given in Section 2.2.3.

Miscellaneous Results

Our analyses rely on the events that the means of the source and target arms stay within their respective confidence sequences. Formally, we define this 'good event', $\mathcal{E}$ as follows

$$\mathcal{E}_S \coloneqq \bigcap_{t\in\mathbb{N}} \bigcap_{i\in[S]} \left\{\mu_i \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)\right\}, \tag{2.29}$$

$$\mathcal{E}_T \coloneqq \bigcap_{t\in\mathbb{N}} \bigcap_{a\in[T]} \left\{\nu_a \in \mathrm{CI}_{\mathcal{T}}(t,a,\delta)\right\}, \tag{2.30}$$

$$\mathcal{E} \coloneqq \mathcal{E}_S \bigcap \mathcal{E}_T. \tag{2.31}$$

If β is chosen as to satisfy the condition in equation 2.7, then we can show that $\mathcal{E}$ occurs with probability at least than $1-\delta$.

Lemma 2.2.13. *Assume β is chosen to satisfy condition Eq. (2.7) so that*

$$\mathbb{P}\left\{\exists t \geq 1 : \mu_i \notin \mathrm{CI}_{\mathcal{S}}(t,i,\delta)\right\} \leq \delta. \tag{2.32}$$

Then,

$$\mathbb{P}\left\{\mathcal{E}\right\} \geq 1-\delta, \tag{2.33}$$

where $\mathcal{E}$ is defined as in equation Eq. (2.31).

Proof. The condition in equation Eq. (2.7) implies that $\mathbb{P}\left\{\mathcal{E}_S\right\} \geq 1-\delta$. To prove the result, we show that $\mathcal{E}_S$ implies $\mathcal{E}_T$ which directly implies that $\mathbb{P}\left\{\mathcal{E}\right\} = \mathbb{P}\left\{\mathcal{E}_S\right\} \geq 1-\delta$. To see this, we fix $a \in [T]$ and observe that on the event $\mathcal{E}_S$

$$\mathcal{L}_{\mathcal{T}}(t,a,\delta) = \sum_{i\in[S]} \min_{m_i \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_{a,i}(m_i) \leq \sum_{i\in[S]} f_{a,i}(\mu_i) = \nu_a,$$

$$\mathcal{U}_{\mathcal{T}}(t,a,\delta) = \sum_{i\in[S]} \max_{m_i \in \mathrm{CI}_{\mathcal{S}}(t,i,\delta)} f_{a,i}(m_i) \geq \sum_{i\in[S]} f_{a,i}(\mu_i) = \nu_a,$$

so that $\mathcal{L}_{\mathcal{T}}(t,a,\delta) \leq \nu_a \leq \mathcal{U}_{\mathcal{T}}(t,a,\delta)$. Since a is arbitrary, the above result holds for all $a \in [T]$. We have just shown that $\mathcal{E}_S$ implies $\mathcal{E}_T$ so that $\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}_S) \geq 1-\delta$ as desired. □

We now use this result to prove Theorem 2.2.4 which concerns the correctness of Algorithm 1.

Proof of Theorem 2.2.4. We observe that by Lemma 2.2.13, the event $\mathcal{E}$ occurs with probability at least $1-\delta$. In particular, this implies that for each target arm, a, and for every round, t, we have that $\nu_a \in \mathrm{CI}_{\mathcal{T}}(t,a,\delta)$. Suppose that the stopping condition is met and recall that we have set $a = 1$

to be an optimal target arm. Then, if B_t is an optimal target arm, the algorithm clearly returns an ϵ-optimal arm. Next, suppose that B_t is not an optimal target arm. In this case, we observe that

$$\nu_{B_t} + \epsilon \geq \mathcal{L}_{\mathcal{T}}(t, B_t, \delta) \geq \mathcal{U}_{\mathcal{T}}(t, C_t, \delta) \geq \mathcal{U}_{\mathcal{T}}(t, 1, \delta) \geq \nu_1 = \max_{a \in [T]} \nu_a,$$

which implies that B_t is ϵ-optimal and thus proves the correctness of our algorithm, as desired. □

Results for Additive Transfer Functions

For the readers convenience, before presenting the proof of Theorem 2.2.5, we briefly review our notation. We let

$$L(i, a, t, x) \coloneqq \max_{m \in [x, x+2\beta(t,\delta)]} f_{a,i}(m) - \min_{m \in [x, x+2\beta(t,\delta)]} f_{a,i}(m) \tag{2.34}$$

to represent the length of target arm a's confidence interval contributed by source arm i when $\mathcal{L}_{\mathcal{S}}(t, i, \delta) = x$. Next, we define

$$\tau_{a,i} = \min \left\{ t \in \mathbb{N} : \sup_{x \in [\mu_i - 2\beta(t,\delta), \mu_i]} L(i, a, t, x) < \frac{\max\left\{|\bar{\nu}_{1,2} - \nu_a|, \epsilon/2\right\}}{s_a} \right\}, \tag{2.35}$$

and

$$\tau_i = \max_{a \in [T]} \tau_{a,i}. \tag{2.36}$$

Lemma 2.2.14. *Let $(P_t, Q_t) \in \{(B_t, I_t), (C_t, J_t)\}$. On the good event $\mathcal{E}$, if $N_{Q_t}(t) \geq \tau_{P_t,Q_t}$, then*

$$\mathcal{U}_{\mathcal{T}}(t, P_t, \delta) - \mathcal{L}_{\mathcal{T}}(t, P_t, \delta) \leq \max\left\{|\bar{\nu}_{1,2} - \nu_{P_t}|, \epsilon/2\right\}. \tag{2.37}$$

Proof. Since we are on the good event, it must be true that $\mu_{Q_t} \geq \mathcal{L}_{\mathcal{S}}(t, Q_t, \delta) \geq \mu_{Q_t} - 2\beta(N_{Q_t}(t), \delta)$. Therefore, the definition of τ_{P_t,Q_t} implies that if $N_{Q_t}(t) \geq \tau_{P_t,Q_t}$, then

$$\begin{aligned} L(Q_t, P_t, t,) &= \max_{m \in \mathrm{CI}_{\mathcal{S}}(t,Q_t,\delta)} f_{P_t,Q_t}(m) - \min_{m \in \mathrm{CI}_{\mathcal{S}}(t,Q_t,\delta)} f_{P_t,Q_t}(m) \\ &\leq \frac{\max\left\{|\bar{\nu}_{1,2} - \nu_{P_t}|, \epsilon/2\right\}}{s_{P_t}}, \end{aligned}$$

where the inequality follows by the definition of τ_{P_t,Q_t}. Additionally, by the definition of the selection rule, we observe that for all $i \in [S]$,

$$L(Q_t, P_t, t) \geq L(i, P_t, t).$$

Therefore, the following inequalities must hold

$$\begin{aligned}\mathcal{U}_{\mathcal{T}}(t, P_t, \delta) - \mathcal{L}_{\mathcal{T}}(t, P_t, \delta) &= \sum_{i \in [S]} L(i, P_t, t) \\ &\leq \sum_{i \in [S]} L(Q_t, P_t, t) \\ &= s_{P_t} L(Q_t, P_t, t) \\ &\leq \max\left\{|\bar{\nu}_{1,2} - \nu_{P_t}|, \epsilon/2\right\},\end{aligned}$$

which gives us the desired result. □

Lemma 2.2.15. *Recall that $\bar{\nu}_{1,2} = \frac{\nu_1+\nu_2}{2}$. On the good event $\mathcal{E}$ defined in equation Eq. (2.31), if the algorithm has not terminated, then there exists $P_t \in \{B_t, C_t\}$ such that*

$$\max\left\{|\nu_{P_t} - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\} \leq |\operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)|. \tag{2.38}$$

Proof. We will split the proof into two cases which encompass all possible scenarios. The first case is when $|\bar{\nu}_{1,2} - \nu_{P_t}| \geq \frac{\epsilon}{2}$, and the other case is when $\frac{\epsilon}{2} \geq |\bar{\nu}_{1,2} - \nu_{P_t}|$.

Case 1 We start off by showing that $|\operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)| \geq |\bar{\nu}_{1,2} - \nu_{P_t}|$. Here we assume that $|\bar{\nu}_{1,2} - \nu_{P_t}| \geq \frac{\epsilon}{2}$. Suppose for the purpose of contradiction that $\bar{\nu}_{1,2} \notin \operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)$. If this is the case, then one of the following four statements must be true:

1. $\bar{\nu}_{1,2} < \mathcal{L}_{\mathcal{T}}(t, B_t, \delta)$ and $\bar{\nu}_{1,2} < \mathcal{L}_{\mathcal{T}}(t, C_t, \delta)$. However, on $\mathcal{E}$, the only arm which can have a lower confidence bound greater than $\bar{\nu}_{1,2}$ is arm 1.
2. $\bar{\nu}_{1,2} > \mathcal{U}_{\mathcal{T}}(t, B_t, \delta)$ and $\bar{\nu}_{1,2} > \mathcal{U}_{\mathcal{T}}(t, C_t, \delta)$. However, on $\mathcal{E}$, the upper confidence bound of arm 1, and hence the upper confidence bound of B_t, must be greater than $\bar{\nu}_{1,2}$.
3. $\bar{\nu}_{1,2} > \mathcal{U}_{\mathcal{T}}(t, B_t, \delta)$ and $\bar{\nu}_{1,2} < \mathcal{L}_{\mathcal{T}}(t, C_t, \delta)$. However, on $\mathcal{E}$, the upper confidence bound of arm 1, and hence the upper confidence bound of B_t, must be greater than $\bar{\nu}_{1,2}$.
4. $\bar{\nu}_{1,2} < \mathcal{L}_{\mathcal{T}}(t, B_t, \delta)$ and $\bar{\nu}_{1,2} > \mathcal{U}_{\mathcal{T}}(t, C_t, \delta)$. This would imply that he algorithm has terminated, which by assumption, is false.

Therefore, by our initial assumption we observe that there exists $P_t \in \{B_t, C_t\}$ satisfying $\max\left\{|\nu_{P_t} - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\} \leq |\operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)|$.

Case 2 Here we show that exists a $P_t \in \{B_t, C_t\}$ such that $|\operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)| \geq \frac{\epsilon}{2}$. For this case, we assume that $\frac{\epsilon}{2} \geq |\bar{\nu}_{1,2} - \nu_{P_t}|$. By the definition of the stopping rule, we know that

$$\mathcal{L}_{\mathcal{T}}(t, B_t, \delta) < \mathcal{U}_{\mathcal{T}}(t, C_t, \delta) - \epsilon. \tag{2.39}$$

We observe that $|\operatorname{CI}_{\mathcal{T}}(t, B_t, \delta)| + |\operatorname{CI}_{\mathcal{T}}(t, C_t, \delta)| > \mathcal{U}_{\mathcal{T}}(t, C_t, \delta) - \mathcal{L}_{\mathcal{T}}(t, B_t, \delta)$. Then rearranging equation Eq. (2.39) yields

$$\begin{aligned}\epsilon &< \mathcal{U}_{\mathcal{T}}(t, C_t, \delta) - \mathcal{L}_{\mathcal{T}}(t, B_t, \delta) \\ &< |\operatorname{CI}_{\mathcal{T}}(t, B_t, \delta)| + |\operatorname{CI}_{\mathcal{T}}(t, C_t, \delta)|.\end{aligned}$$

Therefore, by our initial assumption we observe that there exists $P_t \in \{B_t, C_t\}$ satisfying $\max\left\{|\nu_{P_t} - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\} \leq |\operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)|$.

We have thus shown that, in both cases, there exists $P_t \in \{B_t, C_t\}$ satisfying $P_t \in \{B_t, C_t\}$ satisfying $\max\left\{|\nu_{P_t} - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\} \leq |\operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)|$, which proves the desired result. □

Lemma 2.2.16. *On the good event, $\mathcal{E}$, if the algorithm has not stopped, then there exists a pair $(P_t, Q_t) \in \{(B_t, I_t), (C_t, J_t)\}$ such that $N_{Q_t}(t) < \tau_{P_t, Q_t}$*

Proof. By Lemma 2.2.15 we know that

$$\begin{aligned}\max\left\{|\nu_a - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\} &\leq |\operatorname{CI}_{\mathcal{T}}(t, P_t, \delta)| \\ &= \sum_{i=1}^{S} L(i, P_t, t).\end{aligned}$$

By applying the pigeonhole principle, we see that there must exist at least one $i' \in [S]$ such that

$$L(i', P_t, t) \geq \frac{\max\left\{|\nu_{P_t} - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\}}{s_{P_t}}.$$

Then, by applying the definition of the selection rule, and the fact that on the good event $\mu_{Q_t} \geq \mathcal{L}_{\mathcal{S}}(t, Q_t, \delta) \geq \mu_{Q_t} - 2\beta(N_{Q_t}(t), \delta)$, we observe that

$$\begin{aligned}\sup_{x \in [\mu_i - 2\beta(N_{Q_t}(t), \delta), \mu_i]} L(Q_t, P_t, t, x) &\geq L(Q_t, P_t, t) \\ &\geq L(i', P_t, t) \\ &\geq \frac{\max\left\{|\nu_{P_t} - \bar{\nu}_{1,2}|, \frac{\epsilon}{2}\right\}}{s_{P_t}}.\end{aligned}$$

This implies that

$$\begin{aligned}N_{Q_t}(t) &\leq \min\left\{t \in \mathbb{N} : \sup_{x \in [\mu_{Q_t} - 2\beta(t,\delta), \mu_{Q_t}]} L(Q_t, P_t, t, x) < \frac{\max\left\{|\bar{\nu}_{1,2} - \nu_{P_t}|, \epsilon/2\right\}}{s_{P_t}}\right\} \\ &= \tau_{P_t, Q_t},\end{aligned}$$

as desired. □

We are now ready to prove Theorem 2.2.5.

Proof of Theorem 2.2.5. We have

$$\begin{aligned}
\tau &= \sum_{t=1} \infty \mathbb{I} t \leq \tau \\
&\leq \sum_{t=1} \infty \mathbb{I} \exists (P_t, Q_t) \in \{(B_t, I_t), (C_t, J_t)\} : N_{Q_t}(t) \leq \tau_{P_t, Q_t} \\
&\leq \sum_{i \in [S]} \sum_{t=1} \infty \mathbb{I} i \in \{I_t, J_t\} \cdot \mathbb{I} N_i(t) \leq \max_{a \in [T]} \tau_{a,i} \\
&= \sum_{i \in [S]} \sum_{t=1} \infty \mathbb{I} i \in \{I_t, J_t\} \cdot \mathbb{I} N_i(t) \leq \tau_i \\
&\leq \sum_{i \in [S]} \tau_i,
\end{aligned}$$

which proves the desired result. □

Proof of Corollary 2.2.6. Suppose $i \in M$ since we otherwise don't need to sample source arm i to determine if M is the optimal target arm. From equation Eq. (2.15) we see that $L(i, \mathcal{C}, t, x) = \infty$ unless the confidence interval for μ_i is a subset of $\mathcal{C}_i$ or $\mathcal{C}_i^c$. We consider two cases.

Case 1. Suppose that $\mu_i \in \mathcal{C}_i$. Then, we require the confidence interval is a subset of $\mathcal{C}_i$. For this to be true, it is easy to see that we require $2\beta(t, \delta) \leq \inf_{x \in \mathcal{C}_i^c} |x - \mu_i| = \Delta_{\mathcal{C}_i}(\mu_i)$.

Case 2. Suppose that $\mu_i \notin \mathcal{C}_i$. Then a similar argument shows that we require $2\beta(t, \delta) \leq \inf_{x \in \mathcal{C}_i} |x - \mu_i| = \Delta_{\mathcal{C}_i}(\mu_i)$.

In conclusion, we see that if $i \in M$, then $\tau_{M,i} = \inf\{t \in \mathbb{N} : \beta(t, \delta) \leq \frac{\Delta_{\mathcal{C}_i}(\mu_i)}{2}\}$. Applying Theorem 16 of [14] gives the desired result. □

Proof of Corollary 2.2.7. We observe that since $L(i, a, t, x) = 2|\mathbf{A}_{a,i}|\beta(t, \delta)$ we have

$$\tau_{a,i} = \min \left\{ t \in \mathbb{N} : \beta(t, \delta) \leq \frac{\max\{|\bar{\nu}_{1,2} - \nu_a|, \epsilon/2\}}{s_a |\mathbf{A}_{a,i}} \right\}.$$

Applying Theorem 16 of [14] and taking the max over target arms gives the desired result

□

2.3 Active Exploration for Preference Learning

2.3.1 Introduction

The alignment of foundation models with user preferences has gained unprecedented importance due to the widespread utilization of large language models (LLMs). The established pipeline for

alignment in LLMs, as outlined in Stiennon et al. [32] and Ouyang et al. [33], comprises two steps given a pretrained LLM. First, in the Supervised Fine-Tuning (SFT) phase, the LLM undergoes fine-tuning via supervised learning with examples demonstrating the desired behavior. In the second step, Reinforcement Learning from Human Feedback (RLHF), a policy generates multiple completions for each conversation prefix (prompt) in a training set; users then give ordinal preferences for the set of completions from a particular prompt. These preferences are used to train a *reward model* via a ranking loss like the Bradley-Terry-Luce model [34]. Finally, the policy is trained, typically via Proximal Policy Optimization [35], to optimize the reward model while not moving too far from the SFT-trained policy. More recent work [36], proposed an alternative to RLHF, Direct preference Optimization (DPO), that enables training the LLM policy directly on preference data without using RL and a proxy reward model.

As LLMs continue to scale and their areas of application broaden, the number of topics on which we need to align increases, as does the overall scale of human-generated training data requirements. Data annotation for preference-based learning is already incurring a considerable cost for companies that train LLMs. This cost is likely to grow alongside the industry. The issue becomes especially acute for LLMs in specialized areas such as safety, health, and scientific problems, where the cost of expert feedback can be substantial.

In this work, we take advantage of the fact that we have control over which prompts and completions we provide to human experts to make efficient use of their efforts. Drawing on recent advancements in active exploration for reinforcement learning [37] and in black-box optimization [38], we introduce a method for assessing the value of collecting preferences on specific datapoints, which is both prospective and task-focused. First, we formalize this setting as a *dueling contextual bandit problem* and design an efficient active exploration algorithm that offers polynomial worst-case sample complexity guarantees regarding the policy's performance. Next, we extend these ideas to the alignment setting in LLMs. We show that choosing data for training LLM policies on expert preferences can be targeted by active learning, leading to efficient use of resources under restrictive budgets. In this paper, we build atop the DPO methodology [36], and develop an acquisition strategy that allows us to actively select preference data based on the DPO training objective. We provide two extensions to our active exploration strategy: the first allows an online learning approach, where data selection and training are based on the model's generations, while the second enables the data selection from offline existing preference data.

We evaluate our methods on four datasets: the Stanford Human Preferences dataset [39], the Anthropic Helpful-Harmless dataset [40], and two additional datasets which we contribute to the literature: Jeopardy! dataset and Haikus dataset. The Jeopardy! dataset is an extension of an existing dataset from the game show Jeopardy!. It is composed of questions and factual answers to evaluate the ability of an alignment method to avoid hallucinations. The Haikus dataset is composed of instruction prompts to write Haikus with specific details and corresponding examples of satisfactory Haikus. We use three LLMs with different sizes—GPT-2 [41], Pythia-2.8B [42],

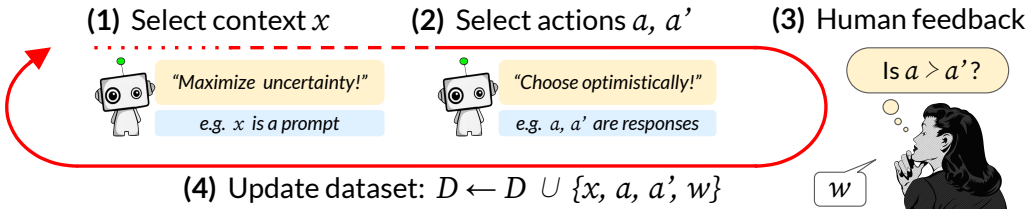


Figure 2.1: Illustration of the active contextual dueling bandit setting, and its application to sample-efficient preference alignment in large language models.

and Llama-3-8B [43]—to showcase a wide range of results and generalization ability. Our main contribution is formalizing the problem of preference data selection as a dueling contextual bandit problem and proposing and analyzing an active exploration algorithm to solve it. We provide a theoretical analysis on the regret bound of our method.

2.3.2 Problem Formulation

In this paper, we consider a dueling variant of what we denote the Active Contextual Dueling Bandit (ACDB) problem introduced in Char et al. [44]. An instance of this problem is defined by a tuple $(\mathcal{X}, \mathcal{A}, f)$ where $\mathcal{X}$ denotes the context space, $\mathcal{A}$ denotes the action space and $f : \mathcal{X} \times \mathcal{A} \times \mathcal{A} \to [0, 1]$ is a preference function so that $f(x, a, a')$ denotes the probability that the action a is preferred to the action a' when the underlying context is x. We also define a domain $\mathcal{D} = \mathcal{X} \times \mathcal{A}$. We will design algorithms that operate under the following interaction protocol, which occurs for T time steps. During each time step $t \in [T]$, the agent selects a context $x_t \in \mathcal{X}$ and a pair of actions $a_t, a'_t \in \mathcal{A}$ and observes a binary random variable $w_t \sim \text{Bernoulli}(f(x_t, a_t, a'_t))$ which equals one if a_t is preferred to a'_t and zero otherwise.

We assume that the preference function has the form

$$f(x, a, a') = \rho\left(r(x, a) - r(x, a')\right), \tag{2.40}$$

where $\rho : \mathbb{R} \to [0, 1]$ is the *link function* and $r : \mathcal{D} \to \mathbb{R}$ is the *unknown* reward function. Common link functions include the logistic function, which leads to the Bradley-Terry-Luce (BTL) model [34] as well as the Gaussian CDF [45].

Our goal is to design algorithms that are able to efficiently identify policies with a small suboptimality gap. We define the suboptimality gap of a learner's policy $\pi : \mathcal{X} \to \mathcal{A}$ as

$$\text{SubOpt}(\pi) = \sup_{x \in \mathcal{X}} \left(\sup_{a \in \mathcal{A}} r(x, a) - r(x, \pi(x)) \right). \tag{2.41}$$

This notion of suboptimality (considered in Char et al. [44] and Li et al. [37]) is stronger than notions that look at the expected suboptimality of the final policy when the contexts are sampled from some known distribution. In this work we also use this suboptimality, which looks at the worst-case context for each policy.

2.3.3 Active Exploration in RKHS

In this section, we describe our first contribution—a theoretically principled algorithm for the ACDB problem—and provide formal guarantees on its performance. To provide the theoretical guarantees, we need to first instantiate our general problem setup by making assumptions on the preference function f (from Eq. (2.40)). In particular, we specify a class of functions that contain the true unknown reward function. This choice is subtle, as we need to balance the trade-off between our function class's expressiveness and theoretical tractability. Motivated by its theoretical popularity and empirical success, we choose the function class to be a Reproducing Kernel Hilbert Space. While this choice of function class is common in the literature, we make a few additional assumptions to more appropriately accommodate our problem setting.

The Contextual Borda Function Before going over our assumptions, we first introduce the *contextual Borda function* f_r, which is core to our algorithm. The contextual Borda function generalizes the Borda function introduced in Xu et al. [38] for dueling-choice optimization, defined as the probability that an action a will be preferred over a random action a' uniformly sampled from the action space. We generalize this definition to the contextual setting as $f_r : \mathcal{D} \to [0, 1]$, where $f_r(x, a) = \mathbb{E}_{a' \sim U(\mathcal{A})} [f(x, a, a')]$ and $U(\mathcal{A})$ is the uniform measure over the action space. It is clear from this definition that f_r and r have the same maximizers.

We now discuss our assumptions. Our first assumption restricts the reward and contextual Borda functions to be 'smooth' in a Reproducing Kernel Hilbert Space (RKHS). Our second assumption relates the reward function to the contextual Borda function.

Assumption 2.3.1. *Let $\kappa : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ denote a positive semi-definite kernel and let $\mathcal{H}_\kappa$ denote its associated RKHS. We assume that $\|r\|_\kappa, \|f_r\|_\kappa \leq B$, where B is a known constant.*

Note that this assumption is stronger than the standard assumption, which only requires that r has a bounded RKHS norm. It is difficult to bound the norm of f_r given a bound on the norm of r due to the generality of our setting, which allows for different link functions. We investigate this issue numerically in Appendix 2.3.9. We find that the norm of the Borda function is almost always smaller than the norm of the reward function for samples drawn from the distribution of basis functions used for experiments in Section 2.3.7.

Assumption 2.3.2. *Let $f_r^*(x) = \max_a f_r(x, a)$ and $r^*(x) = \max_a r(x, a)$. There exists a constant L_1 such that for every $x \in \mathcal{X}$, $a \in \mathcal{A}$ we have $\frac{1}{L_1}(r^*(x) - r(x, a)) \leq f_r^*(x) - f_r(x, a)$.*

This assumption implies that differences in r will cause a similar magnitude of difference in f_r. In fact, when the link function $\rho(\cdot)$ is Lipschitz continuous, it is sufficient for its Lipschitz constant to be at least $1/L_1$ for this condition to hold. We note that this assumption holds for the two most commonly used link functions, the logistic function [34] and the Gaussian CDF [45].

2.3.4 Methods

At a high level, our approach reduces the dueling feedback problem to contextual optimization over a single action via the *contextual Borda function* introduced above. We then apply techniques adapted from recent work on active exploration in reinforcement learning to construct a sampling rule and a policy selection rule, which allow us to output a policy with low suboptimality. Broadly, our sampling rule draws contexts which have maximum uncertainty over the Borda 'value function' and then compares the optimistic action with an action sampled uniformly from the action set.

Estimating the Contextual Borda Function By design, we can estimate the contextual Borda function using preference data $\{x_t, a_t, a'_t, w_t\}$ by selecting x_t, a_t in an arbitrary fashion and sampling a'_t uniformly at random. For low dimensional settings, our algorithm first estimates the contextual Borda function using standard kernelized ridge regression (KRR) [46]. One key feature of KRR is that it provides both an estimate of the contextual Borda function after t observations, $\mu_t(x, a)$, as well as uncertainty quantification of the predictions. Indeed, under Assumptions 2.3.1 and 2.3.2 we can show that $|f_r(x, a) - \mu_t(x, a)| \leq \beta\sigma_t(x, a)$ for an appropriately chosen β and $\sigma_t(x, a)$ (see Lemma 2.3.8).

Selecting Contexts and Actions Our sampling rule builds on top of the one established in Li et al. [37]. Put simply, the rule is to sample the state with the maximum uncertainty over the value function and then act optimistically. We now present our algorithm, which extends these ideas to the dueling setting via the contextual Borda function f_r. For now, we assume that there is a known bonus term $\beta_t^{(r)}$ for all t. We can then define upper and lower confidence bounds $\overline{f_r^t}(x, a) = \mu_t(x, a) + \beta_t^{(r)}\sigma_t(x, a)$ and $\underline{f_r^t}(x, a) = \mu_t(x, a) - \beta_t^{(r)}\sigma_t(x, a)$. Our rule is to select a context

$$x_t \in \operatorname*{argmax}_{x \in \mathcal{X}} \left(\max_{a \in \mathcal{A}} \overline{f_r^t}(x, a) - \max_{a \in \mathcal{A}} \underline{f_r^t}(x, a) \right). \tag{2.42}$$

Here, we are choosing a context that maximizes the difference between the optimistic 'value function' and the pessimistic 'value function' (both of which require a maximization over actions to compute). We then optimistically choose an action

$$a_t \in \operatorname*{argmax}_{a \in \mathcal{A}} \overline{f_r^t}(x_t, a). \tag{2.43}$$

After repeating this process T times, we output a pessimistic policy against the tightest lower bound we can find, which is the maximizer of all our lower bounds through the optimization process. Put formally, we return $\hat{\pi}_T : \mathcal{X} \to \mathcal{A}$ such that

$$\hat{\pi}_T(x) \in \operatorname*{argmax}_{a \in \mathcal{A}} \max_{t \leq T} \underline{f_r^t}(x, a). \tag{2.44}$$

We construct the full active exploration algorithm, AE-Borda, given in Algorithm 3.

Input: kernel function $\kappa(\cdot,\cdot)$, exploration parameters $\beta_t^{(r)}$, number of inital data n_0
Let $D_{n_0} = \{x_i, a_i, a_i', w_i\}_{i=1}^{n_0}$ for x_i, a_i, a_i' drawn uniformly at random.
for $t = n_0 + 1, \ldots, T$ **do**
  Compute $\mu_t(\cdot,\cdot)$, $\sigma_t(\cdot,\cdot)$ using KRR.
  Choose x_t according to Eq. (2.42).
  Choose a_t according to Eq. (2.43), draw $a_t' \sim U(\mathcal{A})$, and draw $w_t \sim \text{Bernoulli}(f(x_t, a_t, a_t'))$.
  Let $D_t = D_{t-1} \cup \{(x_t, a_t, a_t', w_t)\}$.
end for
Output a final policy $\hat{\pi}_T$ according to Eq. (2.44).

Algorithm 3: AE-Borda

2.3.5 Theoretical Analysis

In this section, we provide formal guarantees for our AE-Borda algorithm. Our main result establishes polynomial regret bounds for active exploration in the kernelized dueling bandit setting.

Information-Theoretic Quantities

Our analysis relies on the concept of maximum information gain, which quantifies the information content of our function class.

Definition 2.3.3 (Maximum Information Gain)**.** *The maximum information gain over t rounds, denoted Φ_t, is defined as:*

$$\Phi_t = \max_{A \subset \mathcal{X} \times \mathcal{A} : |A| = t} I(r_A + \epsilon_A; r_A) \tag{2.45}$$

where $r_A = [r(x)]_{x \in A}$, $\epsilon_A \sim \mathcal{N}(0, \eta^2 I)$, and $I(X;Y)$ denotes mutual information.

The maximum information gain captures how much information we can obtain about the reward function from t well-chosen observations. For common kernels, Φ_t grows polynomially in t:

- Linear kernel: $\Phi_t = O(d \log t)$ where d is the dimension
- RBF kernel: $\Phi_t = O((\log t)^{d+1})$ where d is the effective dimension
- Matérn kernel: $\Phi_t = O(t^{d/(2\nu+d)}(\log t)^{1/2})$ where ν is the smoothness parameter

Main Theoretical Result

Our main theoretical contribution is the following regret bound for the AE-Borda algorithm.

Theorem 2.3.4 (Regret Bound for AE-Borda)**.** *Suppose we run Algorithm AE-Borda with confidence parameter:*

$$\beta_t = 2B + \sqrt{2\Phi_t + 1 + \log\left(\frac{2}{\delta}\right)} \tag{2.46}$$

Then, with probability at least $1 - \delta$*, the suboptimality of the returned policy satisfies:*

$$SubOpt(\hat{\pi}_T) \leq O\left(\frac{L_1}{\sqrt{T}}\left(B + \Phi_T\sqrt{\log\frac{1}{\delta}}\right)\right) \tag{2.47}$$

Proof Sketch

The proof follows a confidence-based analysis and relies on several key lemmas:

Confidence Bounds. Under our RKHS assumptions, we establish that the KRR estimates satisfy high-probability confidence bounds:

Lemma 2.3.5 (Confidence Bounds)**.** *With probability at least* $1 - \delta$*, for all* $(x, a) \in \mathcal{D}$ *and all* t*:*

$$|f_r(x, a) - \mu_t(x, a)| \leq \beta_t\sigma_t(x, a) \tag{2.48}$$

Information Gain Bound. The total uncertainty encountered by the algorithm is bounded by the maximum information gain:

Lemma 2.3.6 (Information Gain Bound)**.** *The cumulative uncertainty satisfies:*

$$\sum_{t=1}^{T}\sigma_t^2(x_t, a_t) \leq \Phi_T \tag{2.49}$$

Regret Decomposition. We decompose the suboptimality into bias and variance terms, showing that the active exploration strategy effectively balances exploration and exploitation.

The key insight in the proof is that our context selection rule Eq. (2.42) ensures that we focus our exploration on regions where the uncertainty about the optimal action is highest. This leads to efficient reduction of the confidence regions around the optimal policy.

Implications and Discussion

Our regret bound has several important implications:

1. **Polynomial Sample Complexity:** For common kernels, the bound implies polynomial sample complexity, which is optimal up to logarithmic factors.
2. **Adaptive to Problem Difficulty:** The dependence on Φ_T means the algorithm automatically adapts to the intrinsic difficulty of the problem as measured by the kernel.
3. **Worst-Case Guarantees:** Unlike expected regret bounds, our result provides worst-case guarantees over all possible contexts, making it suitable for safety-critical applications.

The factor L_1 in our bound reflects the relationship between the reward function and the contextual Borda function. While this factor can be problem-dependent, it is typically small for common link functions used in practice.

2.3.6 Extensions to Large Language Models

While our theoretical framework provides strong guarantees in the kernelized setting, extending these ideas to large language models requires addressing several practical challenges. In this section, we discuss how to scale our active exploration principles to the high-dimensional setting of LLMs.

Challenges in the LLM Setting

Extending the AE-Borda method to LLMs faces several limitations:

1. **Unsuitable Action Space:** The contextual Borda function as originally defined requires uniform sampling from the action space. For LLMs, where actions are sequences, most uniformly sampled sequences are trivially distinguishable from natural language.
2. **Batch Training Requirements:** Neural network training proceeds in batches, making it inefficient to label and train on single examples as in the theoretical algorithm.
3. **Limited Uncertainty Estimation:** The uncertainty estimation tools for sequence models are more constrained than those for explicitly kernelized models, especially given memory constraints in training LLMs.

Generalized Contextual Borda Function

To address the first limitation, we propose a generalized contextual Borda function that uses a more meaningful proposal distribution:

Definition 2.3.7 (Generalized Contextual Borda Function)**.** *For a proposal distribution $\pi : \mathcal{X} \to P(\mathcal{A})$, the generalized contextual Borda function is:*

$$f_r^{\pi}(x, a) = \mathbb{E}_{a' \sim \pi(x)} \left[P(w = 1 \mid x, a, a') \right] \tag{2.50}$$

We can recover the original function by setting $\pi(x) = U(\mathcal{A})$. For LLMs, $f_r^{\pi_{SFT}}$ is a natural choice where π_{SFT} is the supervised fine-tuning policy, as it provides meaningful comparison samples.

Direct Preference Optimization Integration

We build upon Direct Preference Optimization (DPO), which avoids training a separate reward model by optimizing the policy directly on preference data. The DPO loss for a policy π_θ with reference policy π_{SFT} is:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{SFT}) = -\mathbb{E}_{(x,a,a',w)\sim\mathcal{D}} \left[\log \sigma \left(\gamma(2w - 1) \left(\log \frac{\pi_\theta(a \mid x)}{\pi_{SFT}(a \mid x)} - \log \frac{\pi_\theta(a' \mid x)}{\pi_{SFT}(a' \mid x)} \right) \right) \right] \tag{2.51}$$

where γ is a hyperparameter controlling the KL penalty strength.

Uncertainty Estimation with Dropout

For uncertainty estimation in the neural network setting, we employ Monte Carlo dropout. For a sequence a consisting of tokens t_i, we have:

$$\log \pi(a \mid x) = \sum_{t_i \in a} \log \pi(t_i \mid x, t_1, \ldots, t_{i-1}) \tag{2.52}$$

We incorporate m dropout masks d_j into the function $\pi(t_i \mid x, t_1, \ldots, t_{i-1}, d_j)$. During inference, Monte Carlo sampling with dropout gives an ensemble with:

$$\mu(t_i \mid x, t_1, \ldots, t_{i-1}) = \frac{1}{m} \sum_{j=1}^{m} \log \pi(t_i \mid x, t_1, \ldots, t_{i-1}, d_j) \tag{2.53}$$

$$\sigma(t_i \mid x, t_1, \ldots, t_{i-1}) = \sqrt{\frac{1}{m-1} \sum_{j=1}^{m} \left(\log \pi(t_i \mid x, t_1, \ldots, t_{i-1}, d_j) - \mu\right)^2} \tag{2.54}$$

The standard deviation serves as an approximation for epistemic uncertainty in a computationally efficient manner.

Active DPO Algorithm

Using the DPO framework and dropout uncertainty estimation, we can define confidence bounds for the generalized Borda function. For upper and lower confidence bounds:

$$\overline{f_r^{\pi_{SFT}}}(x, a) \approx \frac{1}{n} \sum_{i=1}^{N} \frac{1}{1 + \exp\left(\beta \log \frac{\underline{\pi_\theta}(a_i'|x)}{\pi_{SFT}(a_i'|x)} - \beta \log \frac{\overline{\pi_\theta}(a|x)}{\pi_{SFT}(a|x)}\right)} \tag{2.55}$$

$$\underline{f_r^{\pi_{SFT}}}(x, a) \approx \frac{1}{n} \sum_{i=1}^{N} \frac{1}{1 + \exp\left(\beta \log \frac{\overline{\pi_\theta}(a_i'|x)}{\pi_{SFT}(a_i'|x)} - \beta \log \frac{\underline{\pi_\theta}(a|x)}{\pi_{SFT}(a|x)}\right)} \tag{2.56}$$

where $\overline{\pi_\theta}$ and $\underline{\pi_\theta}$ are the upper and lower confidence bounds on the policy probabilities.

Acquisition Function

We define an acquisition function that generalizes our context selection rule to the LLM setting:

$$\alpha(x) = \max_{a \in \mathcal{A}} \overline{f_r^{\pi_{SFT}}}(x, a) - \max_{a \in \mathcal{A}} \underline{f_r^{\pi_{SFT}}}(x, a) \tag{2.57}$$

This acquisition function identifies contexts where there is maximum uncertainty about which action is optimal. In practice, we use this function to select batches of training data by:

1. Sampling a large batch of contexts
2. Evaluating the acquisition function for each context
3. Selecting the top-b contexts with highest acquisition values
4. Collecting preference labels for these contexts
5. Training the policy using DPO on the selected batch

This approach addresses the batch training requirement while maintaining the core principles of active exploration: focusing annotation effort on the most informative examples.

2.3.7 Experimental Validation

While this chapter primarily focuses on the theoretical foundations of active exploration for preference learning, we briefly summarize the experimental validation of our methods to demonstrate their practical effectiveness.

Experimental Setup

The empirical evaluation of our methods spans both synthetic environments that validate our theoretical predictions and real-world applications to large language model alignment. The experiments were designed to answer several key questions:

1. Do our theoretical algorithms achieve the predicted regret bounds in controlled settings?
2. Can the active exploration principles scale effectively to high-dimensional problems like LLM alignment?
3. How much sample efficiency improvement do our methods provide compared to passive baselines?

Kernelized Setting Results

In the kernelized setting, we validated our theoretical predictions using synthetic preference functions with known ground truth. The experiments confirmed that:

- Our AE-Borda algorithm achieves regret bounds consistent with Theorem 2.3.4
- The algorithm effectively identifies high-uncertainty regions for exploration
- The contextual Borda function provides a suitable proxy for optimization in the dueling setting

The empirical regret curves matched the predicted $O(1/\sqrt{T})$ convergence rate, with constants depending on the kernel choice as expected from theory.

Large Language Model Results

For LLM experiments, we evaluated our Active DPO method on multiple datasets and model sizes, including GPT-2, Pythia-2.8B, and Llama-3-8B. Key findings include:

- Nearly 13% relative improvement in performance compared to passive baselines when working with restricted annotation budgets
- Superior performance in avoiding hallucinations on factual question-answering tasks
- Effective scaling across different model architectures and sizes

Practical Impact

The experimental results demonstrate that the theoretical insights from our RKHS analysis successfully transfer to practical applications. The active exploration principles provide significant sample efficiency improvements, which is crucial for real-world deployment where human annotation is expensive.

These results validate our core hypothesis that strategic selection of preference data can substantially improve the efficiency of preference-based learning systems. The theoretical framework provides principled guidance for algorithm design, while the practical extensions make these benefits accessible in high-dimensional settings.

2.3.8 Conclusion

In this chapter, we presented a comprehensive theoretical framework for active exploration in preference-based learning, with applications to reinforcement learning from human feedback and direct preference optimization. Our contributions span both fundamental theory and practical algorithmic innovations.

Summary of Contributions

Our main theoretical contributions include:

1. **Problem Formulation:** We formalized the problem of efficient preference data collection as an Active Contextual Dueling Bandit (ACDB) problem, providing a principled foundation for analyzing active exploration in preference learning.
2. **Algorithm Design:** We developed the AE-Borda algorithm, which reduces the dueling preference problem to contextual optimization via the novel contextual Borda function.
3. **Theoretical Guarantees:** We established polynomial regret bounds for our algorithm in the RKHS setting, showing $O(1/\sqrt{T})$ convergence to optimal policies with high probability.
4. **Practical Extensions:** We extended our theoretical insights to large language models through the Active DPO framework, addressing key challenges in scaling to high-dimensional

sequence models.

Key Insights

Several important insights emerge from our theoretical analysis:

The Power of Active Selection. By strategically choosing which contexts and action pairs to query, we can achieve significantly better sample complexity than passive approaches. This is particularly valuable when human annotation is expensive or time-consuming.

Contextual Borda Function as a Bridge. The contextual Borda function provides an elegant bridge between dueling bandit problems and single-action optimization, enabling the application of well-developed active exploration techniques from reinforcement learning.

Uncertainty-Driven Exploration. Our acquisition functions, based on the difference between optimistic and pessimistic value estimates, provide a principled way to identify the most informative queries. This uncertainty-driven approach naturally balances exploration and exploitation.

Scalability through Approximation. While exact implementation of our theoretical algorithm may not be feasible in high-dimensional settings, the core principles can be preserved through careful approximations, as demonstrated in our LLM extensions.

Implications for Preference-Based Learning

Our work has several important implications for the broader field of preference-based learning:

- **Sample Efficiency:** Active exploration can substantially reduce the amount of human feedback required for effective preference learning, making these approaches more practical for deployment.
- **Theoretical Foundation:** Our regret bounds provide the first polynomial sample complexity guarantees for active exploration in contextual dueling bandits, establishing a theoretical foundation for this important problem class.
- **Practical Algorithms:** The extension to DPO shows how theoretical insights can guide the design of practical algorithms for real-world applications like LLM alignment.

Future Directions

This work opens several promising directions for future research:

1. **Tighter Analysis:** While our bounds are polynomial, there may be room for improvement in the constants and dependence on problem parameters.

2. **Alternative Function Classes:** Exploring active exploration in other function classes beyond RKHS could broaden the applicability of these techniques.
3. **Multi-Armed Settings:** Extending to settings with more than two options per query could capture richer preference structures.
4. **Robustness:** Developing methods that are robust to misspecification of the preference model or adversarial behavior.

The theoretical framework developed in this chapter provides a solid foundation for continued advances in active preference learning, with the potential for significant impact on both fundamental understanding and practical applications.

2.3.9 Proof of Theorem 2.3.4

In this section we will prove our main Theorem, 2.3.4. The overall strategy of the proof is to use our Lipschitz assumption on the link function (more precisely, the relative Lipschitzness of the reward r and the Borda function f_r) in order to go to the Borda function, which we can directly model from data. Then, we use our selection criteria as well as confidence bounds taken from Chowdhury and Gopalan [47] and convergence rates taken from Kandasamy et al. [48] in order to complete the argument. We give these cited results as lemmas in what follows.

In order to attain a particular policy performance with probability $1-\delta$, we must bound the error of the estimates given by our KRR process for a particular confidence level. In order to do so, we adapt the result from Chowdhury and Gopalan [47], Theorem 2.

Lemma 2.3.8. *Let* $\beta_t^{(r)} = 2||f_r||_\kappa + \sqrt{2(\Phi_{t-1}(\mathcal{X}\times\mathcal{A}) + 1 + \log(2/\delta))}$. *Then with probability* $1-\delta$ *we have for all time* t *and any point* $(x,a) \in \mathcal{X}\times\mathcal{A}$,

$$|\mu_{t-1}(x,a) - f_r(x,a)| \leq \beta_t^{(r)}\sigma_{t-1}(x,a).$$

Proof. To prove this result, we will verify that all the conditions from Theorem 2 of Chowdhury and Gopalan [47] hold. Recall Assumption 2.3.1 which states that $\|f_r\|_\kappa \leq B$. Next, we observe that since $a_t' \sim U(\mathcal{A})$ (independent of everything else), we have that $\mathbb{E}[w_t \mid \mathcal{F}_{t-1}] = f_r(x_t,a_t)$, where $\mathcal{F}_t = \rho\left(\{(x_s,a_s,a_s',w_s)\}_{s=1}^t\right)$ is the filtration generated by the past observations. Additionally, since $w_t \in \{0,1\}$ and x_t, a_t are both $\mathcal{F}_{t-1}$ measurable, we see that w_t can be written as

$$w_t = f_r(x_t,a_t) + \eta_t,$$

where η_t is $\mathcal{F}_{t-1}$-conditionally subGaussian. Therefore, we have met all the necessary conditions, and we can apply Theorem 2 of Chowdhury and Gopalan [47] which gives us the desired result. □

This lemma jointly bounds the modeling error over the Borda function for all time t though it introduces a dependence on the RKHS norm of f_r. This dependence is inherited from prior work,

but we empirically study the relationship between the RKHS norm of a particular reward function and that of the associated Borda function in Section 2.3.9. We also adapt a result from Lemma 8 of Kandasamy et al. [48] in order to understand the convergence of our uncertainty function σ_t.

Lemma 2.3.9. *Suppose we have n queries $(q_t)_{t=1}^n$ taken from $\mathcal{X} \times \mathcal{A}$. Then the posterior σ_t satisfies*

$$\sum_{q_t} \sigma_{t-1}^2(q_t) \leq \frac{2}{\log(1+\eta^{-2})} \Phi_n(\mathcal{X} \times \mathcal{A})$$

Proof. In this proof, we condition on the event in Lemma 2.3.8 holding true. Given that occurrence, we can say the following for every $x \in \mathcal{X}$.

$$\begin{aligned}
\max_{a \in \mathcal{A}} r(x,a) - r(x, \hat{\pi}_T(s)) &\overset{\text{Assumption 2.3.2}}{\leq} L_1 \left(\max_{a \in \mathcal{A}} f_r(x,a) - f_r(x, \hat{\pi}_T(x)) \right) && (2.58)\\
&\overset{\text{Lemma 2.3.8}}{\leq} L_1 \left(\max_{a \in \mathcal{A}} f_r(x,a) - \max_{t \in [T]} \underline{f_r^t}(x, \hat{\pi}_T(x)) \right) && (2.59)\\
&\overset{\text{Def. of } \hat{\pi}_T}{=} L_1 \left(\max_{a \in \mathcal{A}} f_r(x,a) - \max_{a \in \mathcal{A}} \max_{t \in [T]} \underline{f_r^t}(x,a) \right) && (2.60)\\
&= L_1 \min_{t \in [T]} \left(\max_{a \in \mathcal{A}} f_r(x,a) - \max_{a \in \mathcal{A}} \underline{f_r^t}(x,a) \right) && (2.61)\\
&\overset{\text{Lemma 2.3.8}}{\leq} L_1 \min_{t \in [T]} \left(\max_{a \in \mathcal{A}} \overline{f_r^t}(x,a) - \max_{a \in \mathcal{A}} \underline{f_r^t}(x,a) \right) && (2.62)\\
&\overset{\text{Def. of } x^t}{\leq} L_1 \min_{t \in [T]} \left(\max_{a \in \mathcal{A}} \overline{f_r^t}(x^t,a) - \max_{a \in \mathcal{A}} \underline{f_r^t}(x^t,a) \right) && (2.63)\\
&\overset{\text{Def. of } a^t}{\leq} L_1 \min_{t \in [T]} \left(\overline{f_r^t}(x^t,a^t) - \underline{f_r^t}(x^t,a^t) \right) && (2.64)\\
&\leq \frac{L_1}{T} \sum_{t=1}^T \left(\overline{f_r^t}(x^t,a^t) - \underline{f_r^t}(x^t,a^t) \right) && (2.65)\\
&= \frac{L_1}{T} \sum_{t=1}^T 2\beta_t^{(r)} \sigma_t(x^t,a^t) && (2.66)\\
&\overset{\beta_t^{(r)} \text{ is increasing}}{\leq} \frac{2L_1 \beta_T^{(r)}}{T} \sqrt{\left(\sum_{t=1}^T \sigma_t(x^t,a^t) \right)^2} && (2.67)\\
&\overset{\text{Cauchy-Schwarz}}{\leq} \frac{2L_1 \beta_T^{(r)}}{T} \sqrt{T \sum_{t=1}^T \sigma_t^2(x^t,a^t)} && (2.68)\\
&\overset{\text{Lemma 2.3.9}}{\leq} \frac{2L_1 \beta_T^{(r)}}{\sqrt{T}} \sqrt{C_1 \Phi_T} && (2.69)\\
&\overset{\text{def of } \beta_T^{(r)}}{=} \frac{2L_1}{\sqrt{T}} (2B + \sqrt{2(\Phi_{t-1} + 1 + \log(2/\delta))}) \sqrt{C_1 \Phi_T} && (2.70)\\
&= O\left(\frac{L_1}{\sqrt{T}} \left(B + \Phi_T \sqrt{\log \frac{1}{\delta}} \right) \right). && (2.71)
\end{aligned}$$

□

RKHS norms of r and f_r

In order to understand the dependence of our estimation bound on the RKHS norm $||f_r||_\kappa$, we ran numerical experiments on sampled reward functions. For a variety of context and action dimensions, we sampled 1000 reward functions as in Section 2.3.7 and numerically approximated their RKHS norms. We also made a Monte-Carlo estimate of the Borda function f_r for each of the reward functions sampled and numerically approximated its RKHS norm. To do this, we uniformly sample 1,000 points x_i from the input space, compute the regularized kernel matrix K for this set x_i, solve the KRR problem $K\alpha = f(x)$ for α. Then we compute the quadratic form $\sqrt{\alpha^T K \alpha}$ as an estimate of the RKHS norm.

In Table 2.1, we present the results of comparing the RKHS norms of 1000 reward functions and their associated Borda functions sampled as in Section 2.3.7. A 'win' was counted when the Borda function had smaller RKHS norm and a 'loss' otherwise. The win margin is the average difference in RKHS norms of the reward and Borda functions, with a positive value when the Borda function was of smaller norm. It is clear here that in general (though not always) the RKHS norm of the Borda function f_r for a particular reward function r is smaller than the RKHS norm of the reward function r itself. This relationship seems to grow stronger as the input dimensionality of the reward function grows larger.

Context Dimension	Action Dimension	Win Rate	Win Margin
0	1	0.16	-6.3
1	1	0.89	5.1
1	3	1	21.4
3	1	1	21.5
3	3	1	38.7
10	10	1	19.6

Table 2.1: Comparison of RKHS norms of reward functions and associated Borda functions

Additional Related Work

In this section, we discuss additional related work, including alternative contextual bandit methods, uncertainty estimation in large language models, and concurrent work on active selection of data in LLMs.

Human feedback in RL and LLMs Here we discuss additional related work on human feedback in reinforcement learning, and more recently, in LLMs. This technique showed significant performance benefits in practice; for example, in the Atari test case [49], where naive deep RL

would have necessitated thousands of hours of gameplay, they accomplished superior performance with just 5,500 or several hours of human queries. More recently, human preference feedback has also been used more recently to improve the performance of LLMs. For example, many recent approaches have demonstrated the effectiveness of using human feedback to enhance LLM stylistic continuation [50], text summarization [32], translation [51], semantic parsing [52], review generation [53], and evidence extraction [54]. In particular, the work by [40] places focus on improving model reliability and robustness by incorporating human feedback to gauge the helpfulness or harmfulness of its responses. However, while effective, the integration of human feedback comes with substantial costs. For example, Stiennon et al. [32] achieved substantial enhancements over baseline methods but required the generation of summaries for 123,169 posts from the TL;DR dataset, a task performed by a large team of labelers from crowdsourcing platforms. This heavy resource requirement is reflected in state-of-the-art work. Ouyang et al. [33] emphasizes RLHF to improve the alignment of the GPT-3 model across aspects such as toxicity, hallucinations, and overall quality. Here, the team enlisted the efforts of 40 labelers and worked with a dataset comprising over 100,000 examples labeled by humans.

Uncertainty Estimation in Large Language Models Estimating the epistemic uncertainty in large language models is still an active area of research and there are few prior works on this topic (focusing specifically on epistemic uncertainty). For example, [55] augment existing models with additional layers to model randomness, and subsequently the uncertainty. However performing uncertainty quantification in a parallelized fashion requires a significant memory overhead. To be more amenable to larger models, we instead use a dropout-augmented model to estimate uncertainty [56].

Concurrent work on active learning in LLMs Concurrently with our work, there has been recent releases of papers related to active data selection for LLMs, which we cover in this section. Note that these papers are predominantly recent and yet unpublished work, released on preprint servers, some of which build on our method and setting. For example, Das et al. [57] builds on our active contextual dueling bandit setting, aiming to develop a method that yields improved theoretical guarantees with reduced assumptions. Zhang et al. [58] proposed a version of DPO using bilevel optimization to optimistically bias towards potentially high-reward responses, though does not use an explicit uncertainty estimate. Xiong et al. [59] develop an an online exploration algorithm as well as a rejection sampling method for offline settings, framing the problem as a reverse-KL regularized contextual bandit problem. Muldrew et al. [60] propose an active learning method for DPO, based on the predictive entropy of LLM predictions as well as uncertainty given by the (implicit) reward model. Xie et al. [61] presents a method that performs DPO with an exploration bonus for improved efficiency. Finally, Hübotter et al. [62] work on a method for active selection of examples for fine-tuning of LLMs using active data selection, for a (single) given prompt at test time.

2.4 Active DPO Using the Reward Function and Offline Data

In this section, we start by proposing another active learning acquisition function based on the reward model. Then we provide a discussion contrasting the real use cases of active learning using online data generated from the policy and the synthetic setting where we can use existing offline benchmarks to evaluate active learning methods.

We propose a new acquisition function that uses the confidence interval of the reward function instead of the generalized Borda function that operates based on the preference model. Using the reward model provides an intuitive solution in RLHF in general and DPO in particular, since the goal is to learn a policy that generates high-reward answers. We can approximate the confidence interval for r ($\overline{r}$ and $\underline{r}$) using the reward expression as the ratio of the policies as defined in the DPO paper. We can compute our upper and lower bounds as

$$\overline{r}(x,a) = \sum_{t_i \in a} \mu(t_i \mid x, t_1, \ldots, t_{i-1}) + \beta\sigma(t_i \mid x, t_1, \ldots, t_{i-1}) - \log \pi_{\text{SFT}}(a \mid x),$$

$$\underline{r}(x,a) = \sum_{t_i \in a} \mu(t_i \mid x, t_1, \ldots, t_{i-1}) - \beta\sigma(t_i \mid x, t_1, \ldots, t_{i-1}) - \log \pi_{\text{SFT}}(a \mid x),$$

for an uncertainty parameter $\beta > 0$. Here, we define an acquisition function as:

$$\alpha(x) = \max_{a \in \mathcal{A}(x)} \overline{r}(x,a) - \max_{a \in \mathcal{A}(x)} \underline{r}(x,a). \tag{2.72}$$

In this equation, $\alpha(x)$ is the uncertainty of the state-value function according to x. In choosing the states where the potential for error in the value achieved is largest, the agent can learn to behave well in those places. This criterion is similar to that in [37] and provides similar guarantees to ours for max-regret in the active contextual bandit setting. In situations like ours where we are using fixed offline datasets, we set $\mathcal{A}(x)$ in Eq. (2.72) to the set of available responses for a particular action; otherwise, we use $\mathcal{A}(x) = \mathcal{A}$.

Chapter 3

Adaptive Causal Inference

This chapter develops adaptive experimental design methods for efficient causal inference. We present two complementary approaches that address different aspects of the adaptive treatment allocation problem: one focuses on achieving exponential improvements in finite-sample regret, while the other leverages the asymptotically optimal AIPW estimator through principled optimistic design. Both contributions bridge classical experimental design with modern algorithmic approaches to sequential decision-making.

3.1 Clipped Second Moment Tracking

3.1.1 Introduction

Randomized Controlled Trials (RCTs) have long been considered the gold standard of evidence in a variety of disciplines, ranging from medicine [63], policy research [64], and economics [65]. In their simplest form, RCTs involve a control arm and a treatment arm, and the objective is to determine if the treatment *causally* outperforms the control. This is typically achieved by fixing a treatment assignment probability (hereafter called an *allocation*), assigning experimental units to an arm, and using the resulting outcomes to estimate the Average Treatment Effect(ATE).

Despite the ubiquity of RCTs, many practitioners have noted that RCTs would benefit from the use of *adaptive* methods—methods in which practitioners vary some aspect of the experiment through the course of the experiment [66–68]. Although there are many reasons for desiring adaptivity, our primary focus is to adaptively select the treatment allocation probability in order to obtain the best possible estimate of the ATE. More concretely, our goal will be to minimize the MSE of our ATE estimate[1] This is the essence of the problem known as Adaptive Neyman Allocation [69] and is the primary focus of this work.

Despite the recent attention given to adaptive approaches, considerable work remains to ensure

[1]In general, one may wish to minimize the mean squared error of the ATE estimate. Since our work focuses on estimation using the unbiased Horvitz-Thompson estimator, this is equivalent to minimizing the variance.

their success in practice. This is because a significant portion of prior work on this topic has focused on developing algorithms with strong asymptotic guarantees. In this asymptotic regime, much is known, such as the semiparametric efficiency bound [70, 71] for non-adaptive approaches, as well as adaptive procedures which asymptotically match the performance of the *best possible* non-adaptive approach [72]. While these results provide a solid foundation, their asymptotic nature overlooks many nuances crucial for practical application. At a high level, prior asymptotic approaches aim to identify the (unknown) variance-minimizing allocation and demonstrate that their allocation converges to this allocation. However, they do not adequately address the challenges of *efficiently learning* this allocation, which is often vital for practical implementation [73].

In order to address these subtleties, we believe a nonasymptotic analysis is required. Unfortunately, such analyses are currently scarce. The only work we are aware of which provides a nonasymptotic analysis is Dai et al. [69] who propose the `ClipOGD` algorithm and show it attains $O(\sqrt{T})$ Neyman regret—a new measure of performance which we formally introduce in Section 3.1.2. Despite offering a promising starting point, this work has several limitations. As we further expand in 3.1.3, `ClipOGD` can demonstrate poor empirical performance; this is explained by the exponential scaling of their bounds with respect to various problem parameters which they treat as constants.

In this paper, we advance the understanding of adaptive estimation procedures for the ATE by providing a finite sample analysis of the Clipped Second Moment Tracking algorithm, a variant of the procedure proposed in [74], tailored for the Horvitz-Thompson estimator. Our analysis meticulously addresses various problem-specific parameters, demonstrating an exponential improvement with respect to problem parameters. We also establish a $O(\log T)$ bound on Neyman regret, representing another significant improvement over `ClipOGD`, although [69] consider the more challenging fixed design setting, while we work in the superpopulation setting defined in Section 3.1.2. Additionally, our finite sample analysis also highlights some aspects of algorithm design that were previously unaddressed.

3.1.2 Problem Setting and Preliminaries

Problem Setup. We consider the following interaction between an algorithm, `Alg`, and a problem instance, ν. At the start of each round t, `Alg` selects a treatment allocation, $\pi_t \in [0, 1]$, based on the history of past observations $\mathcal{H}_{t-1} = \{(\pi_s, A_s, R_s)\}_{s=1}^{t-1}$. Then, the next experimental unit is assigned to either the control ($A_t = 0$) or the treatment ($A_t = 1$) arm by sampling $A_t \sim \textsc{Bernoulli}(\pi_t)$. Following this assignment, an outcome $R_t \in [0, 1]$ is observed, marking the end of the round.

We formalize this interaction protocol as follows. First, we let $\mathcal{F}_t = \sigma(\mathcal{H}_t)$ denote the filtration generated by past observations. Then an algorithm $\texttt{Alg} = (\texttt{Alg}_t)$ is a sequence of $\mathcal{F}_{t-1}$ measurable mappings, $\texttt{Alg}_t : \mathcal{H}_{t-1} \to \mathcal{S}(\{0, 1\})$, where $\mathcal{S}(\mathcal{X})$ is the set of distributions over $\mathcal{X}$. A problem instance $\nu : \{0, 1\} \to \mathcal{S}([0, 1])$ is a probability kernel which maps each arm to a distribution over outcomes which we assume to be bounded in the interval $[0, 1]$. Finally, we let $R_t = \mathbb{I}[A_t =$

$0]R_t(0) + \mathbb{I}[A_t = 1]R_t(1)$, where $\mathbb{I}[\cdot]$ denotes the indicator function, and $R_t(A) \sim \nu(A)$ are called the *potential outcomes*. Within the causal inference literature, this framework is typically referred to as the superpopulation potential outcomes framework [75–77].

Implicit in the above interaction protocol are the following assumptions:

1. *Bounded Observations:* We assume $R_t \in [0, 1]$ almost surely.
2. *Stable Unit Treatment Value Assumption:* We assume that $R_t(A)$ is independent of $R_s(A)$.
3. *Unconfoundedness:* Given the history $\mathcal{H}_{t-1}$, we assume the treatment assignment A_t is independent of the potential outcomes $R_t(0)$ and $R_t(1)$. Formally, $R_t(A) \perp A_t \mid \mathcal{H}_{t-1}$ for $A \in \{0, 1\}$.

While the second and third assumptions are commonplace in the causal inference literature and necessary for identification, the first assumption warrants a brief discussion. We make this assumption so that our methods are compatible with a recent line of work aimed at developing variance-adaptive sequential hypothesis tests [74, 78, 79] where it is currently not known how to construct such tests without assuming bounded observations. However, our analysis and results can be easily modified to accommodate any class of distributions which guarantee concentration of the uncentered second moment. As we will discuss, this differs from existing work which assumes upper and lower bounds on the raw second moments. Indeed, our results don't treat

3.1.3 The `ClipSMT` Algorithm and Results

In this section, we introduce the Clipped Second Moment Tracking(`ClipSMT`) algorithm, state bounds on its Neyman regret, and compare its performance with existing algorithms. *To simplify our presentation and discussions, in this section, we will assume* $\pi_{Ney} \leq \frac{1}{2}$. However, we emphasize our results and analysis can be made to hold for all $\pi_{\text{Ney}} \in (0, 1)$ by flipping the role of the two policies.

The `ClipSMT` Algorithm

We begin by describing the `ClipSMT` algorithm. The idea behind this approach is straightforward: since we do not know the Neyman allocation, we instead choose its empirical counterpart,

$$\tilde{\pi}_t = \frac{\widehat{m}_{t-1}(1)}{\widehat{m}_{t-1}(0) + \widehat{m}_{t-1}(1)}. \tag{3.1}$$

While this approach is appealing, it will not work without modification. This is because $\tilde{\pi}_t$ is overly sensitive to random fluctuations during the early rounds of interaction. As an extreme example, suppose that we select $A_1 = 1, A_2 = 0$ and observe $R_1 = 0, R_2 = 1$. Then, $\tilde{\pi}_t = 0$ for all the subsequent rounds, leading to infinite Neyman regret.

Therefore, we require some form of regularization to guarantee `ClipSMT` is robust to randomness

early in the experiment. To regularize $\tilde{\pi}_t$, we follow Cook et al. [74] and choose the allocation

$$\pi_t = \textsc{Clip}(\tilde{\pi}_t, c_t, 1 - c_t). \tag{3.2}$$

for some clipping sequence c_t. Our subsequent finite sample analysis will show that the setting $c_t = \frac{1}{2}t^{-\frac{1}{3}}$ is the correct choice in a worst-case sense. The complete algorithm can be found in Algorithm 4.

Input: Clipping sequence (c_t)
for each round $t \in \mathbb{N}$ **do**
  Compute $\tilde{\pi}_t = \frac{\widehat{m}_{t-1}(1)}{\widehat{m}_{t-1}(0)+\widehat{m}_{t-1}(1)}$
  Set $\pi_t = \textsc{Clip}(\tilde{\pi}_t, c_t, 1 - c_t)$
  Play $A_t \sim \textsc{Bernoulli}(\pi_t)$ and observe R_t
end for

Algorithm 4: `ClipSMT`

Understanding the Finite Sample Behavior of `ClipSMT`

We now present our finite sample analysis of `ClipSMT`. To begin, we will assume that the clipping sequence has polynomial decay so that $c_t = \frac{1}{2}t^{-\alpha}$ for some $\alpha \in (0, 1)$. We discuss alternative choices for (c_t) in Appendix 3.1.7.

Our analysis splits the behavior of `ClipSMT` into two phases — a clipping phase followed by a concentration phase. In the clipping phase, random fluctuations in R_t will induce large variations in $\tilde{\pi}_t$, leading our algorithm to clip $\tilde{\pi}_t$. The clipping phase ends once we can guarantee that our algorithm will no longer clip the plug-in allocation $\tilde{\pi}_t$, marking the start of the concentration phase, in which we can show that π_t converges to π_{Ney} at a $O\left(t^{-\frac{1}{2}}\right)$ rate.

Our first result characterizes the length of the clipping phase for various choices of α, demonstrating how to select α appropriately.

Lemma 3.1.1. *Assume for simplicity that $\pi_{Ney} \leq \frac{1}{2}$. Suppose we run `ClipSMT` with $c_t = \frac{1}{2}t^{-\alpha}$ for $\alpha \in (0, 1)$. Let $p = \min\left(\alpha, \frac{1-\alpha}{2}\right)$ and define*

$$\tau = \tilde{O}\left(\left[\frac{1}{\pi_{Ney}} + \frac{1}{m_1}\left(\frac{1}{m_0} + \frac{1}{m_1}\right)^{\frac{1}{2}} \log\left(\frac{1}{\delta}\right)\right]^{\frac{1}{p}}\right). \tag{3.3}$$

Then with probability at least $1 - \delta$, for all $t \geq \tau$, we have that $\tilde{\pi}_t = \pi_t$.

Before proceeding we make a few remarks about this result. First, we can show that there exists a problem instance such that the above bound on the length of the clipping phase is tight (modulo some polylogarithmic factors). This implies that without additional knowledge on ν, setting $\alpha = \frac{1}{3}$ minimizes the length of the clipping phase in a worst-case sense. Furthermore, the proceeding results will show that in the concentration phase π_t converges to π_{Ney} at a rate that is independent

of α, thus suggesting that $\alpha = \frac{1}{3}$ is in some sense the correct choice when we don't have additional information about the uncentered second moments. The end of the clipping phase indicates sufficient data collection, mitigating the effects of random fluctuations on π_t, thus marking the start of the concentration phase. In this phase we can show that $\pi_t \in [\pi_{\min}, \pi_{\max}]$, so that $N_t(1) = \Omega(\pi_{\min} \cdot t)$. A simple computation shows that this implies that π_t converges to π_{Ney} at a $O\left((\pi_{\min} \cdot t)^{-\frac{1}{2}}\right)$ rate. While this leads to the correct dependence on t, the scaling with respect to $\pi_{\min}$ is suboptimal—we expect the scaling to be with respect to π_{Ney}. To see why, note that as the interaction progresses, we expect π_t to eventually converge to π_{Ney}. Consequently, we anticipate $N_t(1) = \Theta(\pi_{\text{Ney}} \cdot t)$ which further implies that π_t converges to π_{Ney} at a $O\left((\pi_{\text{Ney}} \cdot t)^{-\frac{1}{2}}\right)$ rate. To remedy this issue, we develop a 'double bounding' technique that uses these initial bounds on π_t and refines them to obtain the correct dependence on π_{Ney}. This gives us the following result which shows that π_t converges to π_{Ney} at the desired rate.

Lemma 3.1.2. *Assume for simplicity that $\pi_{Ney} \leq \frac{1}{2}$. Define*

$$\tau = \tilde{O}\left(\left[\frac{1}{\pi_{Ney}} + \frac{1}{m_1}\left(\frac{1}{m_0} + \frac{1}{m_1}\right)^{\frac{1}{2}} \log\left(\frac{1}{\delta}\right)\right]^3\right). \tag{3.4}$$

Then with probability at least $1 - \delta$, for all $t \geq \tau$, `ClipSMT` guarantees that

$$|\pi_{Ney} - \pi_{t+1}| \leq O\left(\sqrt{\frac{\ell(t,\delta)}{t}}\right) \tag{3.5}$$

where $\ell(t, \delta) = O\left(\log\log t + \log\frac{1}{\delta}\right)$.

The above result shows that following an additional burn-in period after the clipping phase, π_t will converge to π_{Ney} at the desired $O\left((\pi_{\text{Ney}} \cdot t)^{-\frac{1}{2}}\right)$ rate. We also make a remark about the $\sqrt{m_A}$ terms that appear in our bound. These terms appear because of the concentration inequalities we use for m_A. Unfortunately, we can show that this term is asymptotically unavoidable (see Remark 3.1.12 in Appendix 3.1.7).

3.1.4 Theoretical Analysis

Bounding the Neyman Regret

Before stating our bound on the Neyman regret of `ClipSMT`, we first give an alternative expression for the simple Neyman regret and provides insight into our Neyman regret bound.

Lemma 3.1.3. *Fix $\pi_t \in [0, 1]$ and let $\epsilon_t = \pi_t - \pi_{Ney}$. Then we have that*

$$f(\pi_t) - f(\pi_{Ney}) = \Theta\left(\epsilon_t^2\right) \tag{3.6}$$

The proof of this result can be found in Appendix 3.1.7. Surprisingly, this result shows that if

π_t converges to π_{Ney} at a $O\left(t^{-\frac{1}{2}}\right)$ rate, then the simple Neyman regret will shrink at a $O\left(t^{-1}\right)$ rate. Our next result uses this fact in conjunction with the prior bounds on π_t to bound the Neyman regret.

Theorem 3.1.4. *Assume for simplicity that $\pi_{Ney} \leq \frac{1}{2}$. Suppose we run* `ClipSMT` *with $c_t = \frac{1}{2}t^{-\frac{1}{3}}$. Then probability at least $1-\delta$, the Neyman Regret is at most*

$$\tilde{O}\left(\pi_{Ney}^{-1} \cdot \log(T)\right). \tag{3.7}$$

The proof of this result can be found in Appendix 3.1.7. We have just shown that `ClipSMT` obtains *logarithmic* Neyman regret, providing an exponential improvement from the $O\left(\sqrt{T}\right)$ Neyman regret obtained by prior works. As the proceeding discussion highlights, `ClipOGD` works in a more general "design-based" setup. However, it highlights the significant improvements that can be gained in the superpopluation setting considered in this papers.

Comparisons with Prior Work

We continue by comparing our results with past works.

Comparison with Dai et al. [69]. When comparing our Neyman regret bounds to `ClipOGD`, we observe exponential improvements in scaling with respect to π_{Ney} and T.

Starting out with the dependence on π_{Ney}, our bound scales like $O\left(\pi_{\text{Ney}}^{-1}\right)$ while `ClipOGD` scales like $O\left(\exp\left(\pi_{\text{Ney}}^{-1}\right)\right)$. We remark that it is not fully clear if the exponential scaling for `ClipOGD` is a product of the proof technique or is a fundamental drawback of `ClipOGD`. Inspecting the proof in Dai et al. [69], this exponential dependence is introduced to tune the learning rate—if bounds on π_{Ney} are known, `ClipOGD` can be tuned to scale polynomially in π_{Ney}^{-1}. However, even then, not only is the exponent in their polynomial always worse than ours, but it also scales with $\sqrt{\log T}$, while `ClipSMT` does not. Finally, we empirically observe that `ClipOGD` is sensitive to parameter choices. The choices suggested by their analysis can often lead to poor performance (as we demonstrate in Section 3.1.5) indicating that the aforementioned exponential dependence is indeed a fundamental drawback.

Next, we see that our Neyman regret scales like $O(\log T)$ while `ClipOGD` scales like $O(\sqrt{T})$. While this is an exponential improvement, we believe this difference is primarily due to the differences in our problem settings—we consider the superpopulation setting where outcomes are stochastic whereas Dai et al. [69] consider the fixed-design setting where the outcomes are a fixed sequence. In the fixed-design setting, the potential outcomes can be chosen adversarially, including with knowledge of the algorithm, thus increasing the problem's difficulty. The differences between these settings parallels the differences between stochastic and adversarial MABs where we observe similar gaps in regret bounds. In the stochastic bandit setting, the best one can obtain is $O\left(\log T\right)$ problem dependent regret[80]; whereas in the adversarial bandit setting, the best one can hope to do is

$O\left(\sqrt{T}\right)$ minimax regret [81].

Comparison with [74]. As we have mentioned, our algorithm is a variant of the algorithm proposed by Cook et al. [74], tailored to the aHT estimator. The primary difference between our work and Cook et al. [74] is that their focus is asymptotic while ours is nonasymptotic. The asymptotic perspective makes design choices such as the appropriate clipping sequence opaque. In their concluding remarks Cook et al. [74] state that selection of the clipping sequence is an interesting question for future work – our finite sample analysis gives a concrete answer to this question. As an example of the difficulty in choosing the clipping sequence, Cook et al. [74] uses a clipping sequence with exponential decay. Our finite sample analysis indicates that with constant probability, such a clipping sequence will result in an allocation that does not converge to π_{Ney}. Finally, we remark that using a clipping sequence with polynomial decay allows us to slightly generalize their asymptotic results by removing the requirement that bounds on π_{Ney} are known.

3.1.5 Experimental Evaluation

In this section, we experimentally[2] validate our algorithm. Our objective is to compare our algorithm to existing approaches as well as sensible baselines and to understand how well our theoretical characterization of `ClipSMT` aligns with its empirical behavior.

We start by comparing our algorithm to existing approaches and some non-adaptive baselines. In these experiments, we compare `ClipSMT` with `ClipOGD`, the infeasible Neyman Allocation, a balanced allocation with $\pi = \frac{1}{2}$, and a two-stage design we call Explore-then-Commit (EtC). For EtC, we select each treatment arm with equal probability for $T^{\frac{1}{3}}$ rounds, after which we compute the empirical Neyman allocation and use this allocation for the remaining rounds.

We evaluate each approach on nine problem instances, running them for T rounds, where T varies from 1000 to 20000 in increments of 1000. For each fixed value of T, we run `ClipSMT`, `ClipOGD`, and the two-stage design 5000 times to approximate the variance of the resulting ATE estimate. For the Neyman and balanced allocations, we can explicitly compute their variances. Our results show that `ClipSMT` outperforms `ClipOGD` and EtC, and adapts well to difficult problem instances (i.e., when the Neyman allocation deviates from $\frac{1}{2}$). The results of the experiments are displayed in Figure 3.1. Additionally, we perform a more comprehensive simulation of these algorithms in the small sample regime, where we observe similar behavior. The results of this experiment are shown in Figure 3.2.

Next, we validate whether the length of the clipping phase predicted by our theory aligns with the empirical behavior of `ClipSMT`. To do this, we run `ClipSMT` using $c_t = \frac{1}{2}t^{-\alpha}$ for various values of $\alpha \in (0, 1)$. We run `ClipSMT` for each value of α and determine the 0.95 quantile of the clipping phase length based on 5000 simulations. Using these values, we compute the ratio between the

[2]Code for replicating experiments can be found at the following GitHub repo: https://github.com/oneopane/adaptive-ate-estimation.

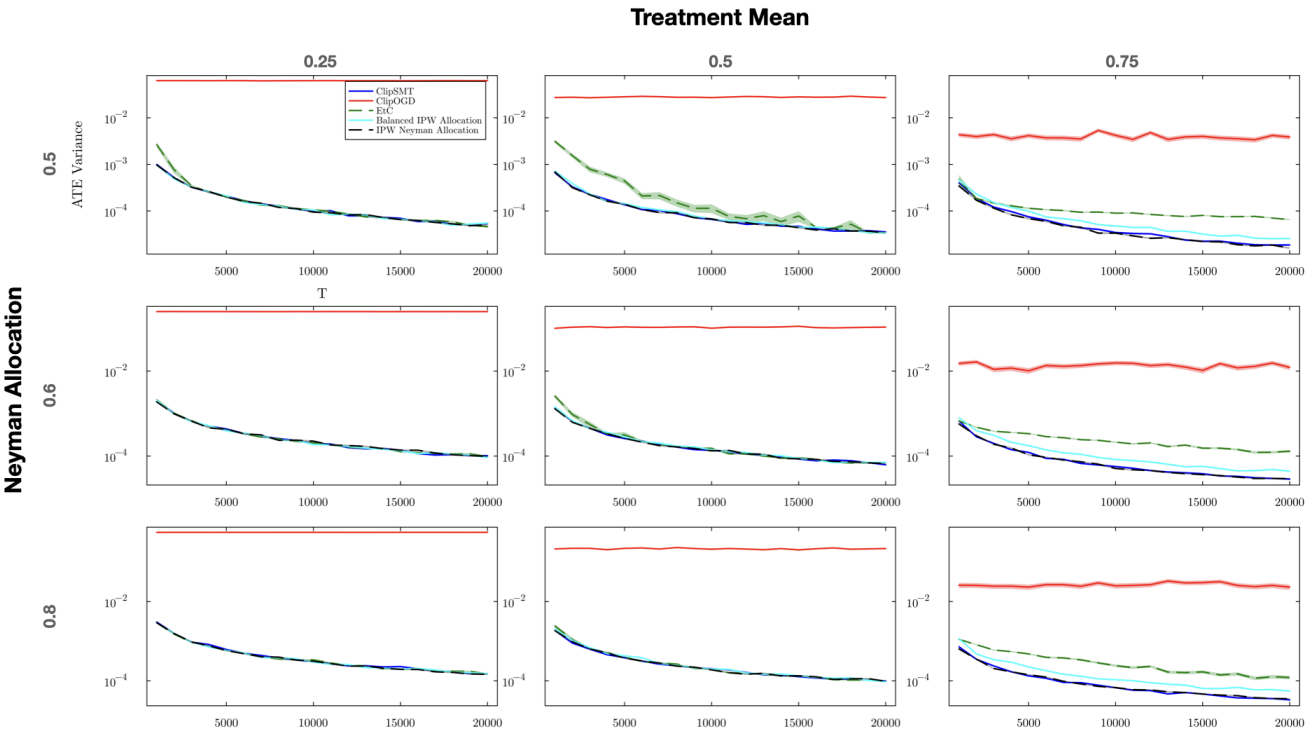


Figure 3.1: Comparison of the performance of `ClipSMT`, `ClipOGD`, Explore-then-Commit (EtC), Neyman allocation, and a balanced allocation with the treatment and control arms following Bernoulli distributions. Individual subplots plot the variance of each design against the number of samples for a fixed problem instance. Each column keeps the treatment mean fixed, and each row keeps the Neyman allocation fixed. Moving to the right increases the treatment mean and moving down increases the Neyman allocation. Overall the performance of `ClipSMT` is always competitive with the performance of the infeasible Neyman allocation and outperforms the other adaptive designs. Furthermore, as the Neyman allocation increases, we see that `ClipSMT` adapts to the increased difficulty while EtC and the balanced design do not. Note that error bars are plotted, however they are narrow due to the large number of simulations performed.

theoretically predicted clipping time to the empirically computed clipping time. The results of this experiment are shown in Figure 3.3 (a).

Inspecting these results, we find that the ratio peaks around $\alpha = \frac{1}{3}$. This behavior is due to a technical difficulty that arises in our proof which we take a brief moment to elucidate. Specifically, upper-bounding the length of the clipping phase involves bounding the quantity $\min\{t : \sum_i t^{p_i} \geq c_1 + c_2 \log\log$ where $c_1 \approx \frac{1}{\pi_{\text{Ney}}^2}$ and $c_2 \approx \frac{1}{\pi_{\text{Ney}}}$. To accomplish this, we compute an upper bound on $\min\{t : t^{\max p_i} \geq c_1 + c_2 \log$ which is also an upper bound on the initial quantity. Noting that $t^{\max p_i} \leq \sum_i t^{p_i}$, it is clear that this step introduces some loosness to our bound. However, we see that as $\pi_{\text{Ney}} \to 0$, we will have $\sum t^{p_i} = \Theta(t^{\max p_i})$ since the growth of $t^{\max p_i}$ will become the dominating term.

Instead, we bound $\min\{t : t^{\max p_i} \geq c_1 + c_2 \log\log t\}$, which provides the correct bound for large values of c_1 and c_2 but is loose for small and moderate values of c_1 and c_2. Therefore, while our bound is tight in the worst case, it has some looseness for specific problems — resolving this issue remains an interesting technical problem for future work.

In order to validate this worst-case optimality, we consider a sequence of problems with Bernoulli

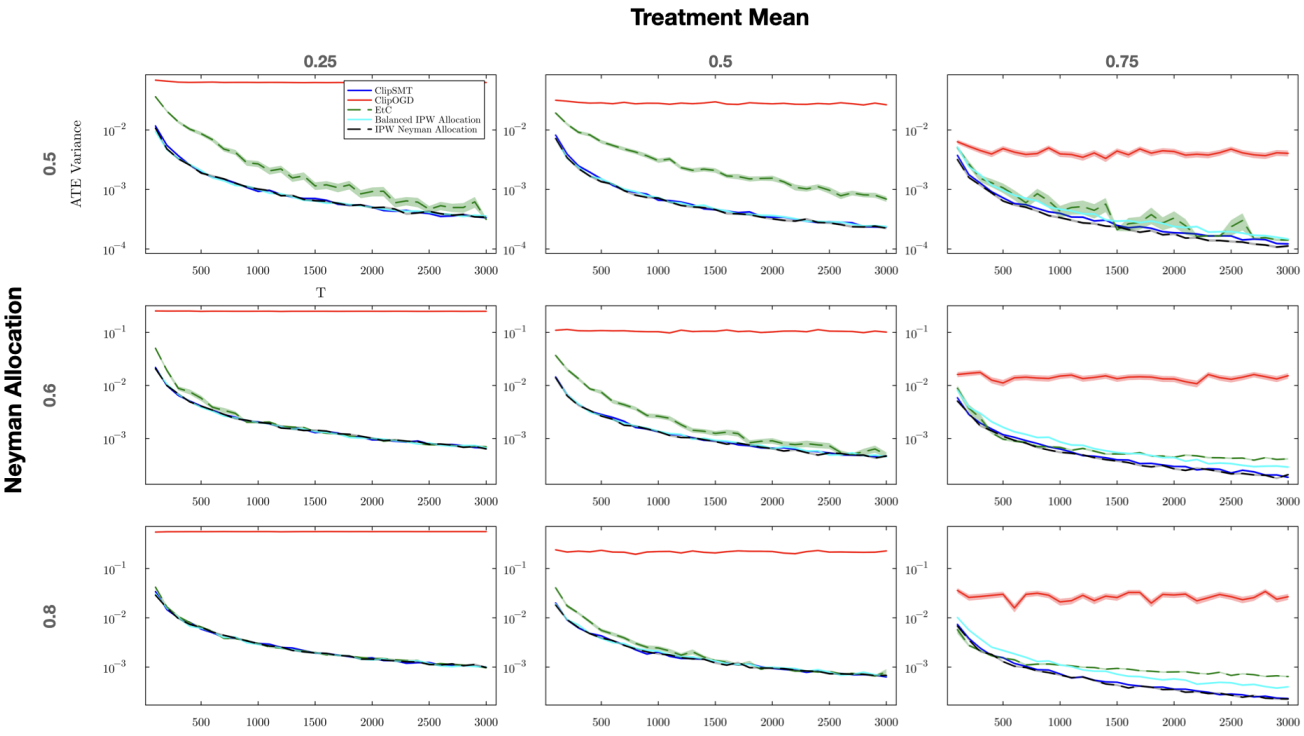


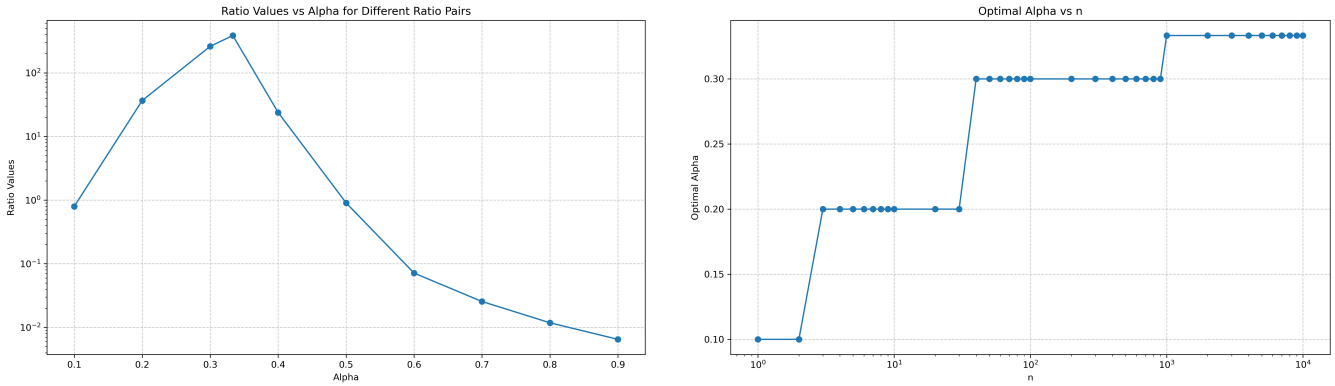
Figure 3.2: Comparison of the performance of `ClipSMT`, `ClipOGD`, Explore-then-Commit (EtC), Neyman allocation, and a balanced allocation with the treatment and control arms following Bernoulli distributions in the small sample regime. Notably, `ClipSMT` is competitive with the Oracle Neyman Allocation even for small sample sizes, indicating its practical utility.

arms $\boldsymbol{\mu}^{(n)} = \left(0.5, \frac{0.5}{n}\right)$. These are chosen to guarantee π_{Ney} converges to 0 which captures the notion of increasing problem difficulty. We run ClipSMT with varying values of α on each problem instance n and compute the median length of the clipping phase over 5000 simulations. For each n, we then determined the value of α, which leads to the shortest clipping phase. The results of this experiment are shown in Figure 3.3 (b), and confirm that setting $\alpha = \frac{1}{3}$ minimizes the length of the clipping phase in the worst case.

3.1.6 Conclusion

In this work, we performed a finite sample analysis of the `ClipSMT` algorithm for adaptive estimation of the ATE. Our analysis clarified several aspects of algorithm design, including how to properly tune the clipping sequence. Furthermore, we demonstrated that our approach achieves exponential improvements in two distinct areas when compared to the only other method with a finite time analysis. Our comprehensive analysis meticulously addressed all problem parameters, providing a clearer and more detailed understanding of the complexity of adaptive ATE estimation.

Several promising directions for future work emerge from our findings. One obvious direction is to extend our analysis to the Augmented Inverse Probability Weighted estimator, which has more desirable properties and is more appropriate when contextual information is available. Additionally,



(a) **(b)**

Figure 3.3: The figure on the left plots the optimal ratio for each problem instance, where the problems get harder as n increases. The figure on the left plots the ratio of the predicted versus empirically computed clipping times. Note that a smaller value implies our theory underestimates the empirical clipping time, implying that the true clipping times are larger.

expanding these results to accommodate larger action spaces and stochastic context-dependent policies warrants further discussion.

3.1.7 Proofs

In this section, we will prove our bound on the Neyman regret of `ClipSMT`.

Preliminaries.

Before we proceed to the analysis, we first introduce some notation and define a 'good event' which we will assume to hold throughout the analysis. We define the following events

$$\mathcal{E}_1(\delta_1) = \bigcap_{t=1}^{\infty} \left\{ N_t(1) \in \left[\sum_{s=1}^{t} \pi_s - \beta_1(t,\delta_1), \sum_{s=1}^{t} \pi_s + \beta_1(t,\delta_1) \right] \right\} \tag{3.8}$$

$$\mathcal{E}_2(\delta_2) = \bigcap_{A\in\{0,1\}} \bigcap_{t=1}^{\infty} \{ m_A \in [\widehat{m}_t(A) - \beta_2(t,\delta), \widehat{m}_t(A) + \beta_2(t,\delta_2)] \} . \tag{3.9}$$

Applying Lemmas 3.1.10 and 3.1.11, using β_1 and β_2 respectively defined in equations Eq. (3.83) and Eq. (3.86) with $\delta_1 = \frac{\delta}{3}, \delta_2 = \frac{2\delta}{3}$, we see that the event $\mathcal{E} = \mathcal{E}_1(\delta_1) \bigcap \mathcal{E}_2(\delta_2)$ occurs with probability at least $1-\delta$. For the remainder of the section, we will assume that this event hold.

Bounding the Neyman Regret (Theorem 3.1.4)

We will bound the cumulative Neyman regret by bounding the simple Neyman regret and then summing over those terms. In order to do so, we will handle the clipping phase and concentration phases separately.

For the clipping phase, Lemma 3.1.5 demonstrates that we can guarantee $\pi_t \in [\pi_{\min}, \pi_{\max}]$ where $\pi_{\min}, \pi_{\max}$ only depend on m_A. This implies that the instantaneous Neyman regret for each round in the clipping phase can be upper bounded by a constant $c(m_0, m_1) = \max_{\pi \in \{\pi_{\min}, \pi_{\max}\}} f(\pi) - f(\pi_{\text{Ney}})$ which only depends on m_A. Furthermore, Lemma 3.1.1 shows that the length of the clipping phase is at most τ so that the cumulative Neyman regret from the clipping phase can be upper bounded as $c(m_0, m_1) \cdot \tau$ which is independent of T.

For the concentration phase, we apply Lemma 3.1.2 which shows that $\epsilon_t \leq \tilde{O}\left(t^{-\frac{1}{2}}\right)$ so that Lemma 3.1.3 implies that the instantaneous Neyman regret for each round of the concentration phase is at most

$$16 \left(\frac{1}{m_0 + m_1}\right)^2 \left(\frac{1}{\sqrt{m_0 (1 - \pi_{\text{Ney}})}} + \frac{1}{\sqrt{m_1 \pi_{\text{Ney}}}}\right)^2 \frac{\ell(t, \delta)}{t}. \tag{3.10}$$

Therefore, we can bound the cumulative Neyman regret during the clipping phase as

$$\begin{aligned}
\sum_{t=\tau+1}^{T} f(\pi_t) - f(\pi_{\text{Ney}}) &\leq 16 \left(\frac{1}{m_0 + m_1}\right)^2 \left(\frac{1}{\sqrt{m_0 (1 - \pi_{\text{Ney}})}} + \frac{1}{\sqrt{m_1 \pi_{\text{Ney}}}}\right)^2 \sum_{t=\tau+1}^{T} \frac{\ell(t, \delta)}{t} && (3.11)\\
&\leq 16 \left(\frac{1}{m_0 + m_1}\right)^2 \left(\frac{1}{\sqrt{m_0 (1 - \pi_{\text{Ney}})}} + \frac{1}{\sqrt{m_1 \pi_{\text{Ney}}}}\right)^2 \sum_{t=1}^{T} \frac{\ell(t, \delta)}{t} && (3.12)\\
&\leq 16 \left(\frac{1}{m_0 + m_1}\right)^2 \left(\frac{1}{\sqrt{m_0 (1 - \pi_{\text{Ney}})}} + \frac{1}{\sqrt{m_1 \pi_{\text{Ney}}}}\right)^2 \ell(T, \delta) \log(T). && (3.13)
\end{aligned}$$

Combining these bounds we see that the Neyman regret can be bounded as

$$c(m_0, m_1) \cdot \tau + 16 \left(\frac{1}{m_0 + m_1}\right)^2 \left(\frac{1}{\sqrt{m_0 (1 - \pi_{\text{Ney}})}} + \frac{1}{\sqrt{m_1 \pi_{\text{Ney}}}}\right)^2 \ell(T, \delta) \log(T) = \tilde{O}(\log(T)), \tag{3.14}$$

which gives the desired result.

Proof. The proof follows from the following series of algebraic manipulations:

$$f(\pi_{\text{Ney}}+\epsilon_t)-f(\pi_{\text{Ney}}) = \frac{m_1^2}{\pi_{\text{Ney}}+\epsilon_t}+\frac{m_0^2}{(1-\pi_{\text{Ney}}-\epsilon_t)}-\frac{m_1^2}{\pi_{\text{Ney}}}+\frac{m_0^2}{(1-\pi_{\text{Ney}})} \tag{3.15}$$

$$= \epsilon_t\left(\frac{m_0^2}{(1-\pi_{\text{Ney}})(1-\pi_{\text{Ney}}-\epsilon_t)}-\frac{m_1^2}{\pi_{\text{Ney}}(\pi_{\text{Ney}}+\epsilon_t)}\right) \tag{3.16}$$

$$\overset{(a)}{=} \epsilon_t\left(\frac{m_0^2}{\left(\frac{m_0}{m_0+m_1}\right)\left(\frac{m_0}{m_0+m_1}-\epsilon_t\right)}-\frac{m_1^2}{\left(\frac{m_1}{m_0+m_1}\right)\left(\frac{m_1}{m_0+m_1}+\epsilon_t\right)}\right) \tag{3.17}$$

$$= \epsilon_t\left(\frac{m_0^2\left(m_0+m_1\right)^2}{m_0^2-m_0\left(m_0+m_1\right)\epsilon_t}-\frac{m_1^2\left(m_0+m_1\right)^2}{m_1^2-m_1\left(m_0+m_1\right)\epsilon_t}\right) \tag{3.18}$$

$$= \epsilon_t\left(\left[\frac{m_0^2\left(m_0+m_1\right)^2}{m_0^2-m_0\left(m_0+m_1\right)\epsilon_t}-\left(m_0+m_1\right)^2\right]\right.$$
$$\left.+\left[\left(m_0+m_1\right)^2-\frac{m_1^2\left(m_0+m_1\right)^2}{m_1^2-m_1\left(m_0+m_1\right)\epsilon_t}\right]\right) \tag{3.19}$$

$$= \epsilon_t^2\left(m_0+m_1\right)^3\left(\frac{1}{m_0-\left(m_0+m_1\right)\epsilon_t}-\frac{1}{m_1-\left(m_0+m_1\right)\epsilon_t}\right), \tag{3.20}$$

where in (a) we have used the fact that $\pi_{\text{Ney}}=\frac{m_1}{m_0+m_1}$. □

Clipping Phase

We now cover various proofs related to the analysis of the clipping phase of our algorithm.

We begin by proving Lemma 3.1.1.

Proof. To begin, we observe that since the function $x,y\mapsto\frac{x}{x+y}$ is monotonic increasing (resp. decreasing) in x (resp. y) we have (on the event $\mathcal{E}$) that

$$\tilde{\pi}_{t+1}\in\left[\frac{m_1-\beta_2(N_t(1),\delta_2)}{m_0+\beta_2(N_t(0),\delta_2)+m_1-\beta_2(N_t(1),\delta_2)},\frac{m_1+\beta_2(N_t(1),\delta_2)}{m_0-\beta_2(N_t(0),\delta_2)+m_1+\beta_2(N_t(1),\delta_2)}\right]. \tag{3.21}$$

We note the above interval is random because $N_t(A)$ is random. In order to construct bounds on $N_t(A)$ we use the fact that $\pi_t\in[c_t,1-c_t]$ so that an integral-sum argument demonstrates

$$\sum_{s=1}^{t}\pi_s\in\left[\frac{1}{2}\cdot\frac{t^{1-\alpha}-1}{1-\alpha},\frac{1}{2}\cdot\frac{t^{1-\alpha}}{1-\alpha}\right]. \tag{3.22}$$

Therefore, on the event $\mathcal{E}$, we obtain

$$\begin{aligned} N_t(1) \in \mathcal{N}(t,\delta_1) &= \left[\frac{1}{2}\cdot\frac{t^{1-\alpha}-1}{1-\alpha} - \beta_1(t,\delta_1), t - \frac{1}{2}\cdot\frac{t^{1-\alpha}}{1-\alpha} + \beta_1(t,\delta_1)\right] \\ &= \left[\frac{1}{2}\cdot\frac{t^{1-\alpha}-1}{1-\alpha} - \sqrt{t\cdot\ell(t,\delta_1)}, t - \frac{1}{2}\cdot\frac{t^{1-\alpha}}{1-\alpha} + \sqrt{t\cdot\ell(t,\delta_1)}\right], \end{aligned} \tag{3.23}$$

where we have set $\ell(t,\delta) = \sqrt{.7225\left(\log\log t + 0.72\log\frac{5.2}{\delta}\right)}$.

Our strategy moving forward will be to use these bounds on $N_t(1)$ to construct a time τ such that for all $t \geq \tau$, we have $\tilde{\pi}_{t+1} \in [c_{t+1}, 1-c_{t+1}]$. We demonstrate how to do so in order to guarantee $\tilde{\pi}_{t+1} \geq c_{t+1}$ as the other case is entirely analogous. Observe that our initial (random) lower bound on $\tilde{\pi}_{t+1}$ together with our bounds on $N_t(1)$ imply that on the event $\mathcal{E}$, we have

$$\begin{aligned} \tilde{\pi}_{t+1} &\geq \min_{n\in\mathcal{N}(t,\delta_1)} \frac{m_1 - \sqrt{\frac{\ell(n,\delta_2)}{m_1\cdot n}}}{m_0 + \sqrt{\frac{\ell(t-n,\delta_2)}{m_0\cdot(t-n)}} + m_1 - \sqrt{\frac{\ell(n,\delta_2)}{m_1\cdot n}}} \\ &\geq \min_{n\in\mathcal{N}(t,\delta_1)} \frac{m_1 - \sqrt{\frac{\ell(t,\delta_2)}{m_1\cdot n}}}{m_0 + \sqrt{\frac{\ell(t,\delta_2)}{m_0\cdot(t-n)}} + m_1 - \sqrt{\frac{\ell(t,\delta_2)}{m_1\cdot n}}}, \end{aligned} \tag{3.24}$$

where the final inequality follows from the monotonic properties of the map $x, y \mapsto \frac{x}{x+y}$. Therefore, our objective is to upper bound the quantity

$$\underline{\tau} = \min\left\{t : \min_{n\in\mathcal{N}(t,\delta_1)} \frac{m_1 - \sqrt{\frac{\ell(t,\delta_2)}{m_1\cdot n}}}{m_0 + \sqrt{\frac{\ell(t,\delta_2)}{m_0\cdot(t-n)}} + m_1 - \sqrt{\frac{\ell(t,\delta_2)}{m_1\cdot n}}} \geq \frac{1}{2}(t+1)^{-\alpha}\right\} \tag{3.25}$$

$$\leq \min\left\{t : \min_{n\in\mathcal{N}(t,\delta_1)} \frac{m_1 - \sqrt{\frac{\ell(t,\delta_2)}{m_1\cdot n}}}{m_0 + \sqrt{\frac{\ell(t,\delta_2)}{m_0\cdot(t-n)}} + m_1 - \sqrt{\frac{\ell(t,\delta_2)}{m_1\cdot n}}} \geq \frac{1}{2}t^{-\alpha}\right\}, \tag{3.26}$$

where the inequality follows from the fact that the LHS in increasing in t and the RHS is decreasing in t. Letting $n^\star$ denote the minimizer of equation Eq. (3.24), by applying Lemma 3.1.13 we observe that

$$n^\star \in \left\{\frac{1}{2}\cdot\frac{t^{1-\alpha}-1}{1-\alpha} - \sqrt{t\cdot\ell(t,\delta_1)}, t - \frac{1}{2}\cdot\frac{t^{1-\alpha}}{1-\alpha} + \sqrt{t\cdot\ell(t,\delta_1)}\right\}. \tag{3.27}$$

Therefore, we can compute an upper bound for each of the two cases so that taking the maximum of these bounds will result in an upper bound on equation Eq. (3.26).

We will demonstrate this for the case $n^\star = \frac{1}{2}\cdot\frac{t^{1-\alpha}}{1-\alpha} - \sqrt{t\cdot\ell(t,\delta_1)}$ since the other case is similar.

After plugging this value of $n^\star$ into equation Eq. (3.26), rearranging terms shows that

$$\min\left\{t : m_1 \geq \frac{1}{2}t^{-\alpha}\left(m_0+m_1\right) + \frac{1}{2}t^{-\frac{2\alpha+1}{2}}\left(\frac{\ell\left(t,\delta_2\right)}{m_0 f\left(t,\delta_1,\alpha\right)}\right)^{\frac{1}{2}} + t^{\frac{\alpha-1}{2}}\left(1-t^{-\alpha}\right)\left(\frac{\ell\left(t,\delta_2\right)}{m_1 g\left(t,\delta_1,\alpha\right)}\right)^{\frac{1}{2}}\right\}, \tag{3.28}$$

where

$$\begin{aligned} f(t,\delta,\alpha) &= 1 + t^{-\frac{1}{2}}\sqrt{\ell(t,\delta)} + \frac{t^{-1}-t^{-\alpha}}{2(1-\alpha)}, \\ g(t,\delta,\alpha) &= \frac{1-t^{\alpha-1}}{2(1-\alpha)} - t^{\frac{2\alpha-1}{2}}\sqrt{\ell(t,\delta)}. \end{aligned}$$

Defining $p = \min\left\{\alpha, \frac{1-\alpha}{2}\right\}$, we can upper bound the RHS of equation Eq. (3.28) with

$$t^{-p}\left((m_0+m_1) + \left(\frac{\ell\left(t,\delta_2\right)}{m_0 f\left(t,\delta_1,p\right)}\right)^{\frac{1}{2}} + \left(\frac{\ell\left(t,\delta_2\right)}{m_1 g\left(t,\delta_1,p\right)}\right)^{\frac{1}{2}}\right). \tag{3.29}$$

Rearranging terms demonstrates that it is sufficient to bound

$$\min\left\{t : t^p \geq \frac{1}{\pi_{\text{Ney}}} + \frac{\sqrt{\ell(t,\delta_2)}}{m_1}\left[\left(\frac{1}{m_0 f\left(t,\delta_1,p\right)}\right)^{\frac{1}{2}} + \left(\frac{1}{m_1 g\left(t,\delta_1,p\right)}\right)^{\frac{1}{2}}\right]\right\}. \tag{3.30}$$

Squaring both sides and applying the inequality $(a+b)^2 \leq a^2+b^2$ twice shows that we can bound

$$\min\left\{t : t^{2p} \geq \frac{2}{\pi_{\text{Ney}}^2} + \frac{4\ell(t,\delta_2)}{m_1^2}\left[\frac{1}{m_0 f\left(t,\delta_1,p\right)} + \frac{1}{m_1 g\left(t,\delta_1,p\right)}\right]\right\}. \tag{3.31}$$

Next, we apply Lemma 3.1.16 and 3.1.18 which show that when

$$t \geq O\left(\max\left\{\left(\frac{1}{1-p}\right)^{\frac{1}{p}}, \left(\log\left(\frac{1}{\delta_1}\right)\right)^{\frac{1}{1-2p}}\right\}\right),$$

we have that $g(t,\delta_1,p), f(t,\delta_1,p) \geq \frac{1}{2}$. Applying Lemma 3.1.15 to equation Eq. (3.31) using the above bounds on g, f demonstrates that

$$\underline{\tau} \leq \underline{\mathfrak{T}}^{\frac{1}{2p}} \tag{3.32}$$

where

$$\underline{\mathfrak{T}} = c_1(\pi_{\text{Ney}}) + c_2(\pi_{\text{Ney}}) \cdot c_3(\pi_{\text{Ney}}) \cdot \log\log c_1(\pi_{\text{Ney}}), \tag{3.33}$$

$$c_1(\pi_{\text{Ney}}) = \frac{2}{\pi_{\text{Ney}}^2} + \frac{4}{m_1^2}\left(\frac{1}{m_0} + \frac{1}{m_1}\right)\log\left(\frac{5.2}{\delta_2}\right), \tag{3.34}$$

$$c_2(\pi_{\text{Ney}}) = \frac{4}{m_1^2}\left(\frac{1}{m_0} + \frac{1}{m_1}\right), \tag{3.35}$$

$$c_3(\pi_{\text{Ney}}) = \frac{\log\log c_1(\pi_{\text{Ney}}) - \log(2p)}{\log\log c_1(\pi_{\text{Ney}})} \cdot \frac{c_1(\pi_{\text{Ney}})\log c_1(\pi_{\text{Ney}})}{c_1(\pi_{\text{Ney}})\log c_1(\pi_{\text{Ney}}) - c_2(\pi_{\text{Ney}})}. \tag{3.36}$$

Repeating the argument for the other choice of $n^\star$ yields the same result.

Finally, we can repeat the above argument for the upper bound on $\tilde{\pi}_{t+1}$ which shows that $\overline{\tau} \leq \overline{\mathfrak{T}}^{\frac{1}{2p}}$, where

$$\overline{\mathfrak{T}} = c_1(\pi_{\text{Ney}}) + c_2(\pi_{\text{Ney}}) \cdot c_3(\pi_{\text{Ney}}) \cdot \log\log c_1(\pi_{\text{Ney}}) \tag{3.37}$$

$$c_1(\pi_{\text{Ney}}) = \frac{2}{(1-\pi_{\text{Ney}})^2} + \frac{4}{m_0^2}\left(\frac{1}{m_0} + \frac{1}{m_1}\right)\log\left(\frac{5.2}{\delta_2}\right) \tag{3.38}$$

$$c_2(\pi_{\text{Ney}}) = \frac{4}{m_0^2}\left(\frac{1}{m_0} + \frac{1}{m_1}\right). \tag{3.39}$$

Letting $\tau = \max\{\underline{\tau}, \overline{\tau}\}$ gives the desired result. □

Concentration Phase

In this section, we will prove Lemma 3.1.2 which we restate for the readers convenience below.

Proof. To begin, we fix $t \geq \tau$ and let $s \in [\tau, t-1]$. Invoking Lemma 3.1.5 implies that on the event $\mathcal{E}$ we have

$$N_s(1) \in \left[\pi_{\min} \cdot s - \sqrt{s\ell(s,\delta_1)}, s - \pi_{\min} \cdot s + \sqrt{s\ell(s,\delta_1)}\right]. \tag{3.40}$$

We will use this to construct a lower bound on π_{s+1} by solving the optimization problem in equation Eq. (3.24) using the interval defined above. Applying Lemma 3.1.13, we can construct a lower bound by considering $N_s(1) \in \left\{\pi_{\min} \cdot s - \sqrt{s\ell(s,\delta_1)}, s - \pi_{\min} \cdot s - \sqrt{s\ell(s,\delta_1)}\right\}$. We demonstrate

this for $N_s(1) = \pi_{\min} \cdot s - \sqrt{s\ell(s,\delta_1)}$. In this case, we have that

$$\pi_{s+1} \geq \frac{m_1 - \sqrt{\frac{\ell(s,\delta_2)}{m_1 N_s(1)}}}{m_0 + \sqrt{\frac{\ell(s,\delta_2)}{m_1(s-N_s(1))}} + m_1 - \sqrt{\frac{\ell(s,\delta_2)}{m_1 N_s(1)}}} \tag{3.41}$$

$$= \pi_{\text{Ney}} \cdot \frac{m_0 + m_1}{m_0 + m_1 + h(\pi_{\min}, s)\sqrt{\frac{\ell(s,\delta_2)}{s}}} - \frac{a_1(\pi_{\min}, s)\sqrt{\frac{\ell(s,\delta_1)}{s}}}{m_0 + m_1 + h(\pi_{\min}, s)\sqrt{\frac{\ell(s,\delta_2)}{s}}} \tag{3.42}$$

$$= \pi_{\text{Ney}} \cdot \frac{\sqrt{\frac{s}{\ell(s,\delta_2)}}(m_0 + m_1)}{\sqrt{\frac{s}{\ell(s,\delta_2)}}(m_0 + m_1) + h(\pi_{\min}, s)} - \frac{a_1(\pi_{\min}, s)}{\sqrt{\frac{s}{\ell(s,\delta_2)}}(m_0 + m_1) + h(\pi_{\min}, s)} \tag{3.43}$$

$$= \underline{\pi}_{s+1}, \tag{3.44}$$

where we have defined

$$a_1(\pi, s) = \sqrt{\frac{1}{m_1\left(\pi - \sqrt{\frac{\ell(s,\delta_1)}{s}}\right)}}, \tag{3.45}$$

$$a_0(\pi, s) = \sqrt{\frac{1}{m_0\left((1-\pi) - \sqrt{\frac{\ell(s,\delta_1)}{s}}\right)}}, \tag{3.46}$$

$$h(\pi, s) = a_0(\pi, s) - a_1(\pi, s). \tag{3.47}$$

Using these bounds, on π_{s+1} we observe that on the event $\mathcal{E}$ we have

$$N_t(1) \geq \sum_{s=1}^{t} \pi_s - \sqrt{t\ell(t,\delta_2)} = \sum_{s=1}^{\tau} \pi_s + \sum_{s=\tau+1}^{t} \pi_s - \sqrt{t\ell(t,\delta_2)} \tag{3.48}$$

$$\geq \frac{\tau^{1-\alpha} - 1}{2(1-\alpha)} + \sum_{s=\tau+1}^{t} \underline{\pi}_s - \sqrt{t\ell(t,\delta_2)} \tag{3.49}$$

We bound $\sum_{s=\tau+1}^{t} \underline{\pi}_s$ using Lemma 3.1.6 so that

$$\begin{aligned} N_t(1) &\geq \pi_{\text{Ney}} \cdot t + \frac{\tau^{1-\alpha} - 1}{2(1-\alpha)} - \pi_{\text{Ney}}(\tau - 1) \\ &\quad - \sqrt{t\ell(t,\delta_2)}\left(2\frac{h(\pi_{\min}) + a_1(\pi_{\min}, \tau)}{m_0 + m_1} + 1\right) \\ &\quad - \frac{a_1(\pi_{\min}, \tau)}{\sqrt{\frac{\tau}{\ell(\tau,\delta_2)}}(m_0 + m_1) + h(\pi_{\min}, \tau)} \\ &= \pi_{\text{Ney}} \cdot t - c, \end{aligned} \tag{3.50}$$

where we have defined $h(\pi_{\min}) = \lim_{t\to\infty} h(\pi_{\min}, t)$. By plugging this value of $N_t(1)$ into equation Eq. (3.24), we obtain

$$\pi_{t+1} \geq \underline{\pi}_{t+1} = \pi_{\text{Ney}} \cdot \frac{\sqrt{\frac{t}{\ell(t,\delta_2)}}\,(m_0+m_1)}{\sqrt{\frac{t}{\ell(t,\delta_2)}}\,(m_0+m_1) + \tilde{h}(\pi_{\text{Ney}}, t)} - \frac{\tilde{a}_1(\pi_{\text{Ney}}, t)}{\sqrt{\frac{t}{\ell(t,\delta_2)}}\,(m_0+m_1) + \tilde{h}(\pi_{\text{Ney}}, t)} \tag{3.51}$$

where we have defined

$$\tilde{a}_0(\pi, t) = \sqrt{\frac{1}{m_0\left((1-\pi) + \frac{c}{t}\right)}} \tag{3.52}$$

$$\tilde{a}_1(\pi, t) = \sqrt{\frac{1}{m_1\left(\pi - \frac{c}{t}\right)}}, \tag{3.53}$$

$$\tilde{h}(\pi, t) = \tilde{a}_0(\pi, t) - \tilde{a}_1(\pi, t) \tag{3.54}$$

Therefore, we have

$$\pi_{\text{Ney}} - \pi_{t+1} \leq \pi_{\text{Ney}} - \underline{\pi}_{t+1} \tag{3.55}$$

$$= \sqrt{\frac{\ell(t,\delta_2)}{t}} \left(\frac{\pi_{\text{Ney}}\tilde{a}_0 + (1-\pi_{\text{Ney}})\tilde{a}_1}{m_0 + m_1 + \sqrt{\frac{\ell(t,\delta_2)}{t}}(\tilde{a}_0 - \tilde{a}_1)} \right) \tag{3.56}$$

$$= \sqrt{\frac{\ell(t,\delta_2)}{t}} \left(\frac{\pi_{\text{Ney}}}{\sqrt{m_0\left((1-\pi_{\text{Ney}}) + \frac{c}{t}\right)}} + \frac{1-\pi_{\text{Ney}}}{\sqrt{m_1\left(\pi_{\text{Ney}} - \frac{c}{t}\right)}} \right) \left(\frac{1}{m_0 + m_1 + \sqrt{\frac{\ell(t,\delta_2)}{t}}(\tilde{a}_0 - \tilde{a}_1)} \right) \tag{3.57}$$

$$\leq 4\sqrt{\frac{\ell(t,\delta)}{t}} \left(\frac{1}{\sqrt{m_0\,(1-\pi_{\text{Ney}})}} + \frac{1}{\sqrt{m_1\pi_{\text{Ney}}}} \right) \left(\frac{1}{m_0+m_1} \right) \tag{3.58}$$

where the final inequality follows from the application of Lemmas 3.1.7 and 3.1.8 which shows that when $t \geq O(\tau)$ we have that $\frac{c}{t} \leq \frac{1}{2}\pi_{\text{Ney}}$. □

Lemma 3.1.5. *Suppose we run* `ClipSMT` *with* $c_t = \frac{1}{2}t^{-\alpha}$ *for some* $\alpha \in (0,1)$ *and let* $p = \min\left(\alpha, \frac{1-\alpha}{2}\right)$. *Then, on the event* $\mathcal{E}$, *for all* $t \geq 1$, *we have that* $\pi_t \in [\pi_{\min}, \pi_{\max}] = \left[\frac{1}{2}\underline{\mathfrak{T}}^{-\frac{\alpha}{2p}}, 1 - \frac{1}{2}\overline{\mathfrak{T}}^{-\frac{\alpha}{2p}}\right]$ *where* $\underline{\mathfrak{T}}, \overline{\mathfrak{T}}$ *are respectively defined in equations Eq. (3.33) and Eq. (3.37).*

Proof. During the clipping phase, we know that $\pi_t \in [c_t, 1-c_t]$. Additionally, once the clipping phase ends, we know that $\tilde{\pi}_t = \pi_t$ so that

$$\pi_{t+1} \in \left[\frac{m_1 - \beta_2(N_t(1), \delta_2)}{m_0 + \beta_2(N_t(0), \delta_2) + m_1 - \beta_2(N_t(1), \delta_2)}, \frac{m_1 + \beta_2(N_t(1), \delta_2)}{m_0 - \beta_2(N_t(0), \delta_2) + m_1 + \beta_2(N_t(1), \delta_2)} \right]. \tag{3.59}$$

It is easy to see that the above bonunds are monotonic in t—the lower bound is monotonically

increasing and the upper bound is monotonically decreasing—which implies that π_t takes its minimum and maximum values at the end of the clipping phase. Therefore, we see that for all $t \geq 1$, we have that

$$1 - \frac{1}{2}\overline{\mathfrak{T}}^{-\frac{\alpha}{2p}} \geq 1 - \frac{1}{2}\tau^{-\alpha} = 1 - c_\tau \geq \pi_t \geq c_\tau = \frac{1}{2}\tau^{-\alpha} \geq \frac{1}{2}\mathfrak{T}^{-\frac{\alpha}{2p}}, \tag{3.60}$$

where the first and last inequality follows from applying Lemma 3.1.1 which shows that $\tau \leq \mathfrak{T}^{\frac{1}{2p}}$. □

Supporting Lemmas

Intermediate Steps

Lemma 3.1.6. *Define*

$$\underline{\pi}_{s+1} = \pi_{Ney} \cdot \frac{\sqrt{\frac{s}{\ell(s,\delta_2)}}\,(m_0+m_1)}{\sqrt{\frac{s}{\ell(s,\delta_2)}}\,(m_0+m_1) + h(\pi_{\min}, s)} - \frac{a_1(\pi_{\min}, s)}{\sqrt{\frac{s}{\ell(s,\delta_2)}}\,(m_0+m_1) + h(\pi_{\min}, s)}. \tag{3.61}$$

Then we have that

$$\sum_{s=\tau+1}^{t} \underline{\pi}_s \geq \pi_{Ney}(t-\tau-1) - 2\sqrt{t\ell(t,\delta_2)}\left(\frac{h(\pi_{\min}) + a_1(\pi_{\min},\tau)}{m_0+m_1}\right) - \frac{h(\pi_{\min},\tau)}{\sqrt{\frac{\tau}{\ell(\tau,\delta_2)}}\,(m_0+m_1) + h(\pi_{\min},\tau)}, \tag{3.62}$$

Proof. We begin by observing

$$\sum_{s=\tau+1}^{t} \underline{\pi}_s = \sum_{s=\tau}^{t-1} \underline{\pi}_{s+1} \tag{3.63}$$

$$= \sum_{s=\tau}^{t-1} \pi_{\text{Ney}} \cdot \frac{\sqrt{\frac{s}{\ell(s,\delta_2)}}\,(m_0+m_1)}{\sqrt{\frac{s}{\ell(s,\delta_2)}}\,(m_0+m_1) + h(\pi_{\min}, s)} - \frac{a_1(\pi_{\min}, s)}{\sqrt{\frac{s}{\ell(s,\delta_2)}}\,(m_0+m_1) + h(\pi_{\min}, s)} \tag{3.64}$$

$$\geq \pi_{\text{Ney}} \cdot \underbrace{\sum_{s=\tau}^{t-1} \frac{\sqrt{\frac{s}{\ell(t,\delta_2)}}\,(m_0+m_1)}{\sqrt{\frac{s}{\ell(t,\delta_2)}}\,(m_0+m_1) + h(\pi_{\min})}}_{\text{Term 1}} - \underbrace{\sum_{s=\tau}^{t-1} \frac{a_1(\pi_{\min},\tau)}{\sqrt{\frac{s}{\ell(t,\delta_2)}}\,(m_0+m_1) + h(\pi_{\min},\tau)}}_{\text{Term 2}}, \tag{3.65}$$

where we have set

$$h(\pi) = \sqrt{\frac{1}{m_0\,(1-\pi)}} - \sqrt{\frac{1}{m_1\pi}}$$

and the inequality follows from the monotonic properties of the map $x, y \mapsto \frac{x}{x+y}$ combined with the fact that $h(\pi, t)$ is increasing in t and $a_1(\pi, t)$ is decreasing in t. From here, we lower bound Term 1 and upper bound Term 2.

To lower bound Term 1, let $c_1 = \frac{m_0+m_1}{\sqrt{\ell(t,\delta_2)}}$ and $c_2 = h(\pi_{\min})$. Then we have

$$\text{Term } 1 = \sum_{s=\tau}^{t-1} \frac{\sqrt{s}c_1}{\sqrt{s}c_1 + c_2} \tag{3.66}$$

$$= (t-\tau-1) - \sum_{s=\tau}^{t} \frac{c_2}{\sqrt{s}c_1 + c_2} \tag{3.67}$$

$$\geq (t-\tau-1) - 2\frac{c_2}{c_1}\sqrt{t} \tag{3.68}$$

$$= (t-\tau-1) - 2\frac{h(\pi_{\min})}{m_0+m_1}\sqrt{t\ell(t,\delta_2)} \tag{3.69}$$

where the inequality follows Lemma 3.1.20.

To upper bound Term 2 we similarly apply Lemma 3.1.20 so that

$$\text{Term } 2 \leq \frac{a_1(\pi_{\min},\tau)}{\sqrt{\frac{\tau}{\ell(\tau,\delta_2)}}(m_0+m_1) + h(\pi_{\min},\tau)} + 2\frac{a_1(\pi_{\min},\tau)\sqrt{t\ell(t,\delta_2)}}{m_0+m_1} \tag{3.70}$$

Combining these bounds shows that

$$\sum_{s=\tau+1}^{t} \underline{\pi}_s \geq \pi_{\text{Ney}}(t-\tau-1) - 2\sqrt{t\ell(t,\delta_2)}\left(\frac{h(\pi_{\min}) + a_1(\pi_{\min},\tau)}{m_0+m_1}\right) - \frac{h(\pi_{\min},\tau)}{\sqrt{\frac{\tau}{\ell(\tau,\delta_2)}}(m_0+m_1) + h(\pi_{\min},\tau)}, \tag{3.71}$$

thus proving the desired result. □

Lemma 3.1.7. *Define*

$$c = \sqrt{t\ell(t,\delta)}\left(2\frac{h(\pi_{\min}) + a_1(\pi_{\min},\tau)}{m_0+m_1} + 1\right) + \frac{a_1(\pi_{\min},\tau)}{\sqrt{\frac{\tau}{\ell(\tau,\delta_2)}}(m_0+m_1) + h(\pi_{\min},\tau)} + \pi_{Ney}(\tau-1) - \frac{\tau^{1-\alpha}-1}{2(1-\alpha)} \tag{3.72}$$

If $t \geq 6\tau$, then $\frac{c}{t} \leq \frac{1}{2}\pi_{Ney}$.

Proof. Note that it is sufficient to bound the first three terms since we are subtracting the fourth term.

For the first term, we observe that when $\tau^{1-2\alpha} \geq 16\ell(\tau,\delta)$, which is satisfied by our definition of τ, we have that $a_1(\pi_{\min},\tau) \leq \sqrt{\frac{2}{m_1\pi_{\min}}}$. Therefore, some algebra shows that

$$2\frac{h(\pi_{\min}) + a_1(\pi_{\min},\tau)}{m_0+m_1} + 1 \leq 2\tau^{\frac{\alpha}{2}}\left(\frac{1}{m_0} + \frac{1}{m_1}\right)\left(\frac{1}{\sqrt{m_0}} + \frac{1}{\sqrt{m_1}}\right) + 1 = c_1 \tag{3.73}$$

Therefore if we want bound this term by $b_1\pi_{\text{Ney}}$, we require $t \geq \frac{c_1^2}{b_1^2}\ell(t,\delta)$. We can apply Lemma 3.1.15 to bound this.

Next, some algebra shows that when $\tau^{1-\alpha} \geq \frac{3}{m_1(m_0+m_1)^2}\ell(\tau,\delta)$, which is satisfied by our definition of τ, we have that

$$\frac{a_1(\pi_{\min},\tau)}{\sqrt{\frac{\tau}{\ell(\tau,\delta_2)}}(m_0+m_1)+h(\pi_{\min},\tau)} \leq 2. \tag{3.74}$$

As such, if we want to bound this term by b_2, we require $t \geq \frac{2}{b_2\pi_{\text{Ney}}}$

Finally, to bound the third term by $b_3\pi_{\text{Ney}}$, we observe that we require $t \geq \frac{\tau-1}{b_2}$.

Setting $b_1 = b_2 = b_3 = \frac{1}{6}$ and $c_1 = \frac{1}{2}$, and using the above results, see that when

$$t \geq \max\left\{144\ell(t,\delta), 6(\tau-1), \frac{12}{\pi_{\text{Ney}}}\right\} \tag{3.75}$$

we have that $\frac{c}{t} \leq \frac{1}{2}\pi_{\text{Ney}}$. □

Lemma 3.1.8. *Suppose $t \geq 6\tau$. Then we have that*

$$\sqrt{\frac{\ell(t,\delta)}{t}}(\tilde{a}_1 - \tilde{a}_0) \leq \frac{1}{2}(m_0+m_1). \tag{3.76}$$

Proof. To being, we see that it is sufficient to find compute an upper bound on the smallest t such that

$$t \geq \ell(t,\delta)\frac{\tilde{a}_1^2}{m_0+m_1}.$$

Next, we apply Lemma 3.1.7 which shows that when $t \geq 6\tau$, we have that $\tilde{a}_1^2 \leq \frac{2}{m_1\pi_{\text{Ney}}}$. Plugging this in and applying Lemma 3.1.15 gives the desired result. □

Useful Tools

Lemma 3.1.9. *We have that*

$$\frac{m_1 - \sqrt{\frac{\ell(t,\delta)}{m_1 N_t(1)}}}{m_0 + \sqrt{\frac{\ell(t,\delta)}{m_0(t-N_t(1))}} + m_1 - \sqrt{\frac{\ell(t,\delta)}{m_1 N_t(1)}}} \tag{3.77}$$

$$= \pi_{Ney}\frac{\sqrt{\frac{t}{\ell(t,\delta)}}(m_0+m_1)}{\sqrt{\frac{t}{\ell(t,\delta)}}(m_0+m_1) + a_0(t,N_t(1)) - a_1(t,N_t(1))} \tag{3.78}$$

$$- \frac{a_1(t,N_t(1))}{\sqrt{\frac{t}{\ell(t,\delta)}}(m_0+m_1) + a_0(t,N_t(1)) - a_1(t,N_t(1))} \tag{3.79}$$

where

$$a_0(t,n) = \sqrt{\frac{1}{m_0\left(1-\frac{n}{t}\right)}}, \tag{3.80}$$

$$a_0(t,n) = \sqrt{\frac{1}{m_1 \cdot \frac{n}{t}}}. \tag{3.81}$$

Concentration Results

Lemma 3.1.10. *Let $X_1, X_2, \ldots$ be a sequence of random variables such that $X_t \sim \text{Bernoulli}(\pi_t)$ where π_t is $\mathcal{F}_{t-1}$ measurable and define $N_t = \sum_{s=1}^{t} X_s$. Then, with probability at least $1-\delta$, the following holds for all $t \in \mathbb{N}$*

$$\left| N_t - \sum_{s=1}^{t} \pi_t \right| \leq \beta_1(t,\delta), \tag{3.82}$$

where

$$\beta_1(t,\delta) = 0.85\sqrt{t\left(\log\log t + 0.72\log\left(\frac{5.2}{\delta}\right)\right)}. \tag{3.83}$$

Proof. Define $M_t^\lambda = \exp\left(\lambda(X-p_t) - \frac{\lambda^2}{8}\right)$. Note that by definition, $X_t \in [0,1]$ almost surely with $\mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] = p_t$ which implies that the following holds for every $\lambda \in \mathbb{R}$

$$\mathbb{E}\left[M_t^\lambda \mid \mathcal{F}_{t-1}\right] \leq 1. \tag{3.84}$$

Therefore, $D_t^\lambda = \prod_{s=1}^{t} M_s^\lambda$ is a test supermartingale and we can apply Theorem 1 from [82] (see equation (11)) to obtain the desired result. □

Lemma 3.1.11. *Let $X_1, X_2, \ldots$ be a sequence of random variables such that $X_t \in [0,1]$, $\mu = \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right]$, and $m^2 = \mathbb{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right]$. Define the empirical second moment as $\widehat{m}_t^2 = \frac{1}{t}\sum_{s=1}^{t} X_s^2$. Then, with probability at least $1-\delta$, the following holds for all $t \in \mathbb{N}$*

$$|\widehat{m}_t - m| \leq \beta_2(t,\delta) \tag{3.85}$$

where

$$\beta_2(t,\delta) = 0.85\sqrt{\frac{\left(\log\log t + 0.72\log\left(\frac{5.2}{\delta}\right)\right)}{m^2 \cdot t}}. \tag{3.86}$$

Proof. To see this, we first observe that

$$|\widehat{m}_t - m| = \frac{|\widehat{m}_t^2 - m^2|}{|\widehat{m}_t + m|} \leq \frac{|\widehat{m}_t^2 - m^2|}{\left|\sqrt{m^2}\right|}.$$

The result then follows by bounding $|\widehat{m}_t^2 - m^2|$ by applying Theorem 1 from [82] (see equation (11)).

□

Remark 3.1.12. *Note that in our above result, the width of the confidence sequences scale like* $O\left(\frac{1}{\sqrt{m^2 \cdot t}}\right)$. *An application of the CLT along with the Delta Method shows that, asymptotically, the scaling with respect to* $\frac{1}{\sqrt{m^2}}$ *is unavoidable.*

Technical Results

Lemma 3.1.13. *Let* $t, \alpha_0, \alpha_1, \gamma_0, \gamma_1 > 0$ *be fixed, and define the function* $f : (0, t) \to \mathbb{R}$ *by*

$$f(x) = \frac{\alpha_1 - \frac{\gamma_1}{\sqrt{x}}}{\alpha_0 + \frac{\gamma_0}{\sqrt{t-x}} + \alpha_1 - \frac{\gamma_1}{\sqrt{x}}}. \tag{3.87}$$

Given an interval $[s, r] \subseteq [1, t]$, *any solution* $x^\star$ *to the optimization problem*

$$\min_{x \in [s,r]} f(x), \tag{3.88}$$

must satisfy $x^\star \in \{s, r\}$.

Proof. Our proof will proceed by demonstrating that one of the preconditions of Lemma 3.1.14 is satisfied, from which the desired result naturally follows. To begin, we let $f'(x) = \frac{d}{dx} f(x)$ denote the derivative of $f(x)$. We compute $f'(x)$ and perform some simplifications to show that

$$\begin{aligned} f'(x) &= -\left(\frac{\left(\frac{\gamma_0}{2(t-x)^{3/2}} + \frac{\gamma_1}{2x^{3/2}}\right)(\alpha_1 - \frac{\gamma_1}{\sqrt{x}})}{(\alpha_0 + \alpha_1 + \frac{\gamma_0}{\sqrt{t-x}} - \frac{\gamma_1}{\sqrt{x}})^2} \right) + \frac{\gamma_1}{2(\alpha_0 + \alpha_1 + \frac{\gamma_0}{\sqrt{t-x}} - \frac{\gamma_1}{\sqrt{x}})x^{3/2}} \\ &= \frac{(\gamma_0\gamma_1 t + \alpha_0\gamma_1 t\sqrt{t-x} - \alpha_0\gamma_1\sqrt{t-x}x - \alpha_1\gamma_0 x^{3/2})}{2(-\gamma_1\sqrt{t-x} + \gamma_0\sqrt{x} + \alpha_0\sqrt{t-x}\sqrt{x} + \alpha_1\sqrt{t-x}\sqrt{x})^2\sqrt{t-x}\sqrt{x}}. \end{aligned} \tag{3.89}$$

Observe that the denominator in Eq. (3.89) is always greater than zero. Therefore, $\text{sign}(f'(x))$ is determined by the numerator which we will now show to be strictly decreasing. The derivative of the numerator in Eq. (3.89) is

$$-\left(\frac{3(\alpha_0\gamma_1 t + \alpha_1\gamma_0\sqrt{t-x}\sqrt{x} - \alpha_0\gamma_1 x)}{2\sqrt{t-x}} \right).$$

From here, we have that by assumption $\alpha_0, \alpha_1, \gamma_0, \gamma_1 > 0$ and $x < t$ imply that the above quantity is strictly negative. Since the derivative of the numerator is strictly negative, we know that the numerator is strictly decreasing. Therefore, our earlier observation, in conjunction with this fact implies that one of the preconditions of Lemma 3.1.14 must hold, thus enabling its application, which in turn implies the desired result. □

The next lemma essentially shows that the minimum of a concave-unimodal function over a

closed interval must occur at one of the boundaries of the interval.

Lemma 3.1.14. *Let $f : \mathcal{D} \to \mathbb{R}$ be any differential function such that its derivative, f', satisfies one of the following conditions:*

1. *$f'(x) > 0$ for all $x \in \mathcal{D}$*
2. *$f'(x) < 0$ for all $x \in \mathcal{D}$*
3. *There exists c such that for all $x < c$, $f'(x) > 0$ and for all $x > c$, $f'(x) < c$.*

Then for any $[a, b] \subset \mathcal{D}$, any solution $x^\star$ to optimization problem,

$$\min_{x \in [a,b]} f(x), \tag{3.90}$$

must satisfy $x^\star \in \{a, b\}$.

Proof. If $f'(x) > 0$ for all $x \in \mathcal{D}$, the function is monotonically increasing and the minimum will occur at $x^\star = a$. If $f'(x) < 0$ for all $x \in \mathcal{D}$, the function is monotonically decreasing and the minimum will occur at $x^\star = b$. For the final case, let c be as defined in the condition and let $\tilde{x}$ denote the minimum of f. If $a < \tilde{x} < c$ then $f(\tilde{x}) - f(a) = \int_a^{\tilde{x}} f'(t)dt > 0$ which is a contradiction. Similarly if $b > \tilde{x} > c$, then $f(b) - f(\tilde{x}) = \int_{\tilde{x}}^b f'(c)dc < 0$ which is also a contradiction. Therefore, for each of the cases, $x^\star$ must satisfy $x^\star \in \{a, b\}$. □

Lemma 3.1.15. *Let $c_1, c_2, p > 0$ such that $\log c_1 > p$ and $c_1 \log c_1 > c_2$ and define*

$$\tau = \min \left\{ t : t^p \geq c_1 + c_2 \log \log(t) \right\}. \tag{3.91}$$

We have that

$$\tau \leq \left(c_1 + c_2 \log \left(\log c_1 \right) \frac{\log \log c_1 - \log(p)}{\log \log c_1} \cdot \frac{c_1 \log c_1}{c_1 \log c_1 - c_2} \right)^{\frac{1}{p}} \tag{3.92}$$

Proof. To prove this, we set

$$t = \left(c_1 + a c_2 \log \log c_1 \right)^{\frac{1}{p}},$$

for some a to be chosen later. Our objective is to show that

$$\log \log \left[\left(c_1 + a c_2 \log \log c_1 \right)^{\frac{1}{p}} \right] \leq a \log \log c_1.$$

To do so, we observe that

$$\begin{aligned}
&\log\left(\log\left((c_1 + ac_2\log\log c_1)^{\frac{1}{p}}\right)\right)\\
&= \log\left(\frac{1}{p}\log\left(c_1 + ac_2\log\log c_1\right)\right)\\
&= \log\left(\frac{1}{p}\left(\log(c_1) + \log\left(1 + \frac{ac_2}{c_1}\log\log c_1\right)\right)\right)\\
&\leq \log\left(\frac{1}{p}\left(\log(c_1) + \frac{ac_2}{c_1}\log\log c_1\right)\right)\\
&= \log\left(\frac{1}{p}\log c_1\right) + \log\left(1 + \frac{ac_2}{c_1\log c_1}\log\log c_1\right)\\
&\leq \log\left(\frac{1}{p}\log c_1\right) + \frac{ac_2}{c_1\log c_1}\log\log c_1,
\end{aligned}$$

where the inequalities follow from applying the inequality $\log(1+x) \leq x$. From here, we set a so the final line above equals $a\log\log c_1$. In particular, by setting

$$a = \frac{\log\log c_1 - \log(p)}{\log\log c_1} \cdot \frac{c_1\log c_1}{c_1\log c_1 - c_2},$$

the above series of inequalities proves that

$$\log\log\left[(c_1 + ac_2\log\log c_1)^{\frac{1}{p}}\right] \leq a\log\log c_1,$$

as desired. □

Lemma 3.1.16. *Fix $\alpha, \delta \in (0,1)$ and consider the function*

$$f(t,\delta,\alpha) = 1 + t^{-\frac{1}{2}}\sqrt{\ell(t,\delta)} + \frac{t^{-1} - t^{-\alpha}}{1-\alpha}.$$

For all $t \geq \left(\frac{2}{1-\alpha}\right)^{\frac{1}{\alpha}}$, we have that $g(t,\delta,\alpha) \geq \frac{1}{2}$.

Proof. First note that

$$1 + t^{-\frac{1}{2}}\sqrt{\ell(t,\delta)} + \frac{t^{-1} - t^{-\alpha}}{1-\alpha} \geq 1 - \frac{t^{-\alpha}}{1-\alpha}.$$

Solving the inequality

$$1 - \frac{t^{-\alpha}}{1-\alpha} \geq \frac{1}{2},$$

for t gives the desired result. □

Corollary 3.1.17. *For $\delta \in (0,1)$, $t \geq 27$ implies that $f(t,\delta,\frac{1}{3}) \geq \frac{1}{2}$.*

Lemma 3.1.18. *Fix $\alpha \in \left(0, \frac{1}{2}\right)$, $\delta \in (0,1)$, and let*

$$g(t, \delta, \alpha) = \frac{1 - t^{\alpha - 1}}{2\left(1 - \alpha\right)} - t^{\frac{2\alpha - 1}{2}} \sqrt{\ell(t, \delta)}.$$

We have that $g(t, \delta, \alpha) \geq \frac{1}{2}$ whenever

$$t \geq \left(c_1 + c_2 \log\left(\log c_1\right) \frac{\log\log c_1 - \log(1 - 2\alpha)}{\log\log c_1} \cdot \frac{c_1 \log c_1}{c_1 \log c_1 - c_2} \right)^{\frac{1}{1 - 2\alpha}},$$

where $c_1 = \frac{2}{\alpha^2} + \frac{8(1-\alpha)^2}{\alpha^2} \log\left(\frac{5.2}{\delta}\right)$ and $c_2 = \frac{8(1-\alpha)^2}{\alpha^2}$.

Proof. To begin, observe that

$$\frac{1 - t^{\alpha - 1}}{2\left(1 - \alpha\right)} - t^{\frac{2\alpha - 1}{2}} \sqrt{\ell(t, \delta)} \geq \frac{1}{2\left(1 - \alpha\right)} - t^{\frac{2\alpha - 1}{2}} \sqrt{\ell(t, \delta)} - \frac{t^{\frac{2\alpha - 1}{2}}}{2\left(1 - \alpha\right)}, \tag{3.93}$$

therefore it is sufficient to bound the quantity

$$\min\left\{ t : \frac{1}{2\left(1 - \alpha\right)} - t^{\frac{2\alpha - 1}{2}} \left(\sqrt{\ell(t, \delta)} + \frac{1}{2\left(1 - \alpha\right)} \right) \geq \frac{1}{2} \right\}. \tag{3.94}$$

Rearranging, we see that this is equivalent to bounding the quantity

$$\min\left\{ t : t^{\frac{1}{2} - \alpha} \geq \frac{2\left(1 - \alpha\right)}{\alpha} \sqrt{\ell(t, \delta)} + \frac{1}{\alpha} \right\}. \tag{3.95}$$

By squaring both sides and applying the inequality $(a + b)^2 \leq 2a^2 + 2b^2$ we see that it is sufficient to bound

$$\min\left\{ t : t^{1 - 2\alpha} \geq \frac{2}{\alpha^2} + \frac{8\left(1 - \alpha\right)^2}{\alpha^2} \ell(t, \delta) \right\}. \tag{3.96}$$

Setting $c_1 = \frac{2}{\alpha^2} + \frac{8(1-\alpha)^2}{\alpha^2} \log\left(\frac{5.2}{\delta}\right)$ and $c_2 = \frac{8(1-\alpha)^2}{\alpha^2}$ we can apply Lemma 3.1.15 to see that whenever

$$t \geq \left(c_1 + c_2 \log\left(\log c_1\right) \frac{\log\log c_1 - \log(1 - 2\alpha)}{\log\log c_1} \cdot \frac{c_1 \log c_1}{c_1 \log c_1 - c_2} \right)^{\frac{1}{1 - 2\alpha}},$$

we have that $g(t, \delta, \alpha) \geq \frac{1}{2}$, as desired. □

Corollary 3.1.19. *For $\delta \in (0,1)$, $t \geq O\left(\log\left(\frac{1}{\delta}\right)^3\right)$ implies that $g\left(t, \delta, \frac{1}{3}\right) \geq \frac{1}{2}$.*

Lemma 3.1.20. *Fix c_1, c_2, c_3, τ, t such that $c_1, c_2 > 0$, $\tau < t$, and $c_2\sqrt{\tau} + c_3 > 0$. Then, we have that*

$$\sum_{s=\tau}^{t} \frac{c_1}{c_2\sqrt{s} + c_3} \leq \frac{c_1}{c_2\sqrt{\tau} + c_3} + \frac{2c_1}{c_2}\left(\sqrt{t} - \sqrt{\tau}\right) \tag{3.97}$$

Proof. Observe that under the stated conditions, we have that $\frac{c_1}{c_2\sqrt{s}+c_3}$ is monotonically decreasing in s. Therefore we can bound

$$\begin{aligned}
\sum_{s=\tau}^{t} \frac{c_1}{c_2\sqrt{s}+c_3} &\leq \frac{c_1}{c_2\sqrt{\tau}+c_3} + \int_{s=\tau}^{t} \frac{c_1}{c_2\sqrt{s}+c_3} ds \\
&\leq \frac{c_1}{c_2\sqrt{\tau}+c_3} + \left(\frac{2c_1}{c_2}\sqrt{s} - \frac{2c_1c_3}{c_2^2}\log\left(c_3+c_2\sqrt{s}\right)\right)\Bigg|_{s=\tau}^{t} \\
&= \frac{c_1}{c_2\sqrt{\tau}+c_3} + \left(\frac{2c_1}{c_2}\sqrt{t} - \frac{2c_1c_3}{c_2^2}\log\left(c_3+c_2\sqrt{t}\right)\right) - \left(\frac{2c_1}{c_2}\sqrt{\tau} - \frac{2c_1c_3}{c_2^2}\log\left(c_3+c_2\sqrt{\tau}\right)\right) \\
&\leq \frac{c_1}{c_2\sqrt{\tau}+c_3} + \frac{2c_1}{c_2}\left(\sqrt{t}-\sqrt{\tau}\right)
\end{aligned}$$

□

Discussion on Clipping Sequences

Recall that our proposed ClipSMT algorithm utilizes clipping sequence with polynomial decay so that $c_t = \frac{1}{2}t^{-\alpha}$ for $\alpha \in (0,1)$. It is natural to wonder if there are other valid choices for the clipping sequence. While there are, the choices of clipping sequences that will work depend on the assumptions that we make.

On one hand, if we do not assume a lower bound on m_A^2, then we must require that $\sum_t c_t$ diverges as $t \to \infty$. To see why, suppose the sum converges, i.e $\lim_{T\to\infty}\sum_{t=1}^{T} c_t = c$. Then, if we choose m_1^2, m_2^2 so that the length of the clipping phase is larger than c, this will ensure that π_t never converges to π_{Ney}. As a concrete example, this implies that in this most general setting, we should not use clipping sequence with exponential decay. However, if we are willing to assume a lower bound on m_0^2, m_1^2, then we can use a similar argument in order to select the rate of decay for a clipping sequence whose sum converges.

3.2 Optimistic Policy Tracking

3.2.1 Introduction

The problem of estimating the average treatment effect (ATE) is central to causal inference and has been extensively studied. We have a precise understanding of the difficulty of this problem in both asymptotic and nonasymptotic regimes. However, our understanding of the challenges associated with *adaptive* ATE estimation remains limited.

Classically, adaptive ATE estimation has been analyzed in an asymptotic setting, where past work has focused on designing adaptive sampling procedures that ensure that the resulting ATE estimate achieves the smallest possible asymptotic variance, that is, the semiparametric efficiency bound. More recently, there has been growing interest in developing algorithms that provide

nonasymptotic performance guarantees. However, these works suffer from certain drawbacks that lead to poor finite sample performance, an issue that we discuss in detail in Sections 3.2.2.

In this work, we take a nonasymptotic perspective on adaptive ATE estimation, focusing on the Augmented Inverse Propensity Weighting (AIPW) estimator. Our finite-sample analysis reveals key aspects of algorithmic design that prior work has overlooked. This enables us to propose a new algorithm with substantially improved theoretical and empirical performance while also simplifying the analysis.

At the heart of our approach is the insight that initially over-sampling arms that should eventually be under-sampled according to the (unknown) optimal allocation can lead to better estimates of the ATE. Interestingly, this idea can be interpreted as an instance of the principle of *optimism*, a well-established algorithmic design paradigm in the literature on regret minimization in multi-armed bandits (MAB) and reinforcement learning. We discuss this connection in more detail in Section 3.2.4.

Contributions. Our main contributions are as follows:

1. We develop and analyze a new algorithm, Optimistic Policy Tracking(OPT), for the adaptive estimation of ATE that enjoys significant theoretical improvements over previous approaches along with a significantly simplified analysis.
2. We perform simulations that demonstrate that our theoretical improvements translate into empirical improvements, especially in the small sample regime, which is critical for applications such as randomized clinical trials.

3.2.2 Related Works

Adaptive experimental design has a long and distinguished history, dating back to the seminal work of Neyman [83] on optimal allocation in experimental studies. Thompson [84] introduced a Bayesian adaptive design, thus laying the foundation for the MAB problem. Thompson's approach of sequential updating beliefs about treatments (or arms) based on observed outcomes is now central in MAB research [85]. However, many problem formulations focus on maximizing cumulative rewards over repeated rounds of exploration-exploitation. In contrast, our objective of ATE estimation differs from the typical MAB focus and raises different forms of exploration trade-offs.

Prior Work

Our work builds on a recent line of work investigating adaptive algorithms aimed at efficiently estimating ATE. Hahn et al. [86] sparked this research direction by proposing a two-stage design, conceptually similar to the Explore-then-Commit algorithms in MAB [87] and showing that it asymptotically attains the minimum-variance semiparametric efficiency bound. Subsequently, Kato et al. [72] introduced a fully adaptive procedure using the *adaptive* AIPW estimator (`A2IPW`), and

showed that it is asymptotically optimal (in the above sense) while also providing improved empirical performance compared to the less adaptive two-stage design. Later, Cook et al. [74] proposed an alternative method called Clipped Standard-Deviation Tracking (`ClipSDT`), which inherits the same asymptotic optimality under milder assumptions, admits modern uncertainty quantification tools [79], and outperforms the earlier approach empirically. In parallel work, Li et al. [88] significantly generalized the two-stage design in Hahn et al. [86], extending its applicability to a broad spectrum of problems, including Markovian and non-Markovian decision processes.

Despite these advances, all of the above approaches focus on characterizing the asymptotic behavior of their approaches, leaving open questions about finite-sample performance of their work. In order to address these questions, Dai et al. [69] takes an initial step toward understanding the nonasymptotic difficulty by introducing the `ClipOGD` algorithm for the fixed-design setting. They introduce and analyze the Neyman regret (in the design-based setting), which is a normalized proxy to the variance of the resulting ATE estimate. Even more recently, Neopane et al. [3] proposed and analyzed the `ClipSMT` algorithm for the superpopulation setting and shows that it enjoys an improved $\log T$ bound on the Neyman regret.

Although these two works take important first steps toward understanding the nonasymptotic difficulty of adaptive ATE estimation, their algorithms rely on the `IPW` estimator which is known to be suboptimal. In fact, these works define the Neyman regret with respect to the minimum variance `IPW` estimator, where the minimization is performed over all possible allocations. In contrast, our definition of the Neyman regret is much stronger as the baseline against which we compete is defined as the minimum attainable variance over *all* pairs of estimators and allocations. Notably, using this stronger definition of regret, the aforementioned approaches obtain linear Neyman regret, where as we are able to design an algorithm which obtains logarithmic Neyman regret.

Related Works

The problem of off-policy evaluation, which generalizes ATE estimation, has been extensively studied in the literature on reinforcement learning [89–91]. Most of the research in this area has focused on offline estimation, leading to precise characterizations of minimax lower bounds along with matching upper bounds [92–95]. Beyond policy evaluation, these methods have been extended to estimate other quantities, such as the cumulative distribution function of rewards [96, 97]. However, there has been limited exploration of adaptive versions of these methods. Some existing work includes Hanna et al. [98], which focuses on off-policy learning, and Konyushova et al. [99], which integrates offline off-policy evaluation techniques with online data acquisition to enhance sample efficiency in policy selection. However, these works are primarily empirical.

A related area of research concerns inference procedures for adaptively collected data. This can be categorized into asymptotic and non-asymptotic approaches. On the asymptotic side, one direction has focused on reweighting estimators and establishing their asymptotic normality [100–102]. Another direction avoids asymptotics, instead leveraging modern advances in martingale

theory to derive nonasymptotic confidence intervals and sequences for adaptively collected data, including estimates of the ATE [22, 79, 103].

3.2.3 Background

Problem Setup We are interested in adaptive estimation of the average treatment effect. During each round, t, the algorithm uses the history of past observations $\mathcal{H}_{t-1} = \{(\pi_s, A_s, Y_s)\}_{s=1}^{t-1}$ to select the probability of treatment allocation π_t. Then, π_t is used to assign the next experimental unit to either the control ($A_t = 0$) or the treatment ($A_t = 1$) by sampling $A_t \sim \text{Bernoulli}(\pi_t)$. Finally, after assigning the experimental unit, we observe the outcome Y_t which marks the end of round t.

We formalize the above interaction protocol as follows. Let $\mathcal{F}_t = \sigma(\mathcal{H}_t)$ denote the filtration generated by the past observations. An algorithm $\mathcal{A} = \{(\pi_t, \hat{\tau}_t)\}_{t=1}^{T}$ is defined as a sequence of $\mathcal{F}_{t-1}$ measurable random elements where $\pi_t \in [0, 1]$ is the treatment allocation probability and $\hat{\tau}_t : (\pi_t, A_t, Y_t) \mapsto \mathbb{R}_{\geq 0}$ which can be thought of as the ATE estimate produced by $\mathcal{A}$ on round t.

We assume that the rewards are generated as $Y_t = \mathbf{1}\{A_t = 1\}Y_t(1) + \mathbf{1}\{A_t = 0\}Y_t(0)$, where $Y_t(a)$ are called the potential outcomes. We assume that the sequence of potential outcomes are jointly distributed according to some probability measure ν (the "environment") that satisfies the following assumptions. The first assumption is that the rewards are unconfounded, which means that, given $\mathcal{F}_{t-1}$, the potential outcomes $Y_t(1), Y_t(0)$ are conditionally independent of the treatment assignment A_t, i.e $Y_t(1), Y_t(0) \perp A_t \mid \mathcal{F}_{t-1}$. The second assumption is that the reward means and variances are conditionally fixed so that for all t, we have $\mathbb{E}_\nu[Y_t(a) \mid \mathcal{F}_{t-1}] = \mu_a$ and $\text{Var}_\nu[Y_t(a) \mid \mathcal{F}_{t-1}] = \sigma_a^2$.

Our objective within this framework is to estimate the ATE τ, which is defined as

$$\tau = \mu_1 - \mu_0.$$

The Adaptive AIPW Estimator. An algorithm for adaptive ATE estimation thus requires us to specify a method to compute the treatment allocation probability π_t as well as the estimate $\hat{\tau}_t$. A natural choice for $\hat{\tau}_t$ is the AIPW estimator, which given some reward estimate $\hat{\mu}$, is defined as

$$\hat{\tau}_t = \frac{g(A_t)}{\mathbb{P}_{\mathcal{A},\nu}[A_t]} (Y_t - \hat{\mu}_{A_t}) + \hat{\tau}[\hat{\mu}], \tag{3.98}$$

where $g(A_t) = \mathbf{1}\{A_t = 1\} - \mathbf{1}\{A_t = 0\}$ and $\hat{\tau}[\hat{\mu}] = \hat{\mu}_1 - \hat{\mu}_0$. However, this estimator isn't well suited to sequential estimation, motivating Kato et al. to propose the Adaptive AIPW (AAIPW) estimator. Specifically, letting $\hat{\mu}_t$ denote any $\mathcal{F}_{t-1}$ measurable function (i.e. a *predictable* reward estimate), they defined

$$\hat{\tau}_t^{AAIPW} = \frac{\mathbf{1}\{A_t = 1\} - \mathbf{1}\{A_t = 0\}}{\pi_t(A_t)} (Y_t - \hat{\mu}_t(A_t)) + \hat{\tau}_t[\hat{\mu}_t]. \tag{3.99}$$

We also choose to use the AAIPW estimator for a few reasons. The first reason is that this

estimator is known to be asymptotically optimal – this is crucial for obtaining sublinear Neyman regret (which we define below). Furthermore, recent advances in sequential analysis have developed tight confidence sequences for the AAIPW, making it a natural choice due to its compatibility with the downstream goals of sequential testing and uncertainty quantification.

Neyman Allocation and Regret We use the mean squared error (MSE) to measure the quality of the estimates produced by our algorithm. However, by itself, the MSE is difficult to interpret because it does not consider the inherent difficulty of the problem. Therefore, we would like to normalize this error with respect to some problem dependent baseline which we now define and motivate. Hahn et al. show that for any fixed allocation, π, the minimum attainable MSE of any estimator is

$$\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}. \tag{3.100}$$

The Neyman allocation π^* is defined as the allocation which minimizes the above variance and a simple calculation shows that

$$\pi^* = \frac{\sigma_1}{\sigma_0 + \sigma_1}. \tag{3.101}$$

Ideally, we would like to design an algorithm whose variance is close to this baseline and in order to understand the rate at which this occurs, we consider the Neyman regret which is defined as

$$R_T = T \cdot (\hat{\tau}_T - \tau)^2 - \left(\frac{\sigma_1^2}{\pi^*} + \frac{\sigma_0^2}{1-\pi^*} \right). \tag{3.102}$$

The Neyman regret is simply the difference in the normalized MSE between the optimal variance and the MSE of the estimate produced by $\mathcal{A}$. This normalization guarantees that the the MSE converges to a constant (rather than 0) so that if $\mathcal{A}$ has sublinear regret, then we are guaranteed that its MSE converges to the optimal MSE.

Using the fact that the AAIPW is unbiased, along with the fact that π_t and $\hat{\mu}_t$ are predictable, we can rewrite the Neyman regret for the AAIPW estimator as

$$R_T = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{A},\nu}[\ell(\pi_t, \hat{\mu}_t)] - \left(\frac{\sigma_1^2}{\pi^*} + \frac{\sigma_0^2}{1-\pi^*} \right), \tag{3.103}$$

where

$$\ell(\pi, \mu) = \sum_{a \in \{0,1\}} \frac{\sigma_a^2}{\pi(a)} + \frac{1-\pi(a)}{\pi(a)} \varepsilon_t(a)^2 \tag{3.104}$$

is the Neyman loss and $\varepsilon_t(a) = \mu_a - \hat{\mu}_t(a)$ is the reward estimation error.

Notation. In what follows, we will let

$$N_t(a) = \sum_{s=1}^{t} \mathbf{1}\{A_s = a\}$$

denote the number of times the action a is selected at the end of round t,

$$\bar{Y}_t(a) = \frac{1}{N_t(a)} \sum_{s=1}^{t} Y_s \mathbf{1}\{A_s = a\}$$

denote the empirical mean after t rounds, and

$$\hat{\sigma}_t^2(a) = \frac{1}{N_t(a)} \sum_{s=1}^{t} \left(Y_s \mathbf{1}\{A_s = a\} - \bar{Y}_t(a)\right)^2$$

denote the empirical variance. We use $\tilde{O}[\cdot]$ to denote asymptotic equivalence up to doubly logarithmic factors.

3.2.4 The Optimistic Policy Tracking Algorithm

In this section, we introduce our Optimistic Policy Tracking (OPT) algorithm. We begin with a discussion of the difficulties of adaptive ATE estimation and the suboptimality of existing approaches. Next, we introduce our algorithm and provide insight into why it resolves the issues of existing approaches. Finally, we conclude with a brief discussion of the algorithmic design principles underlying our algorithm and their relation to ideas in the literature.

The difficulties of adaptive ATE estimation. The primary difficulty of adaptive ATE estimation is in balancing the exploration-exploitation trade-off that arises from adaptive allocation. If we condition on $\mathcal{F}_{t-1}$ some algebra shows that the variance of the AAIPW estimator is

$$\sum_a \frac{\sigma_a^2}{\pi_t(a)} + \frac{1 - \pi_t(a)}{\pi_t(a)} \left(\mu_a - \hat{\mu}_t(a)\right)^2, \tag{3.105}$$

which is minimized by setting $(\pi, \mu) = (\pi^*, \mu)$ where π^* is the Neyman allocation. Since π^* and μ are not known a priori, we need to design an algorithm to adaptively estimate them. However, this is challenging because optimizing the exploration allocation separately for estimating π^* and μ (each requiring a different allocation) results in a procedure with high Neyman regret. As such, designing an algorithm to adaptively balance the exploration of π^* and μ while simultaneously minimizing the Neyman regret becomes a very delicate task.

Insights into improvements. In order to better understand the improvements that can be made, we investigate previous approaches for balancing this trade-off. To simplify the exposition, in this section we assume that $\pi^* \leq \frac{1}{2}$. The primary approach that past works (both asymptotic and nonasymptotic) have utilized is clipping the allocation. In fact, prior algorithms compute the empirical allocation

$$\hat{\pi}_t = \frac{\hat{\sigma}_t(1)}{\hat{\sigma}_t(0) + \hat{\sigma}_t(1)},$$

and plays a clipped version of this estimate

$$\pi_t = \min\{1 - \gamma_t, \max\{\gamma_t, \hat{\pi}_t\}\},$$

for some carefully chosen clipping sequence γ_t satisfying $\gamma_t \to 0$. However, these clipping approaches have some important limitations.

The first limitation is that a clipping approach cannot be fully adaptive to the underlying problem instance because the clipping sequence must be chosen a priori. As such, past works choose γ_t in order to optimize the performance of their algorithm in a worst-case sense, leading to suboptimal Neyman regret for easy problem instances. The second, more pressing issue, is that clipping approaches lead to algorithms which under-exploit which is caused by the asymmetry of the Neyman loss. Practically, the implication is that an algorithms which under-sampled the arm with a smaller probability according to the Neyman allocation must necessarily pay a higher price than the same algorithm which over-sampled the same arm by the same amount.

Optimistic Policy Tracking. Our proposed algorithm, OPT, is designed in order to address these aforementioned issues. Indeed, as we will see, not only does OPT better adapt to the underlying problem instances, it also better handles the exploration-exploitation trade-off when compared to prior works. The algorithm itself if simple and plays the allocation

$$\pi_t = \operatorname*{argmin}_{\pi \in \mathcal{C}_t[\pi^*]} \left| \frac{1}{2} - \pi \right|, \tag{3.106}$$

where $\mathcal{C}_t[\pi^*]$ is a confidence sequence for the Neyman allocation. For reward estimation, we simply use the sample mean $\hat{\mu}_t(a) = \frac{1}{N_{t-1}(a)} \sum_{s=1}^{t-1} Y_s \cdot \mathbf{1}\{A_s = a\}$.

The main difficulty now is in constructing the confidence sequence $\mathcal{C}_t[\pi^*]$. In order to do so, we first construct confidence sequences for the standard deviations of each arm. This constructs a confidence sequence $\mathcal{C}_t[\sigma_a] = [\underline{C}_t(\sigma_a), \overline{C}_t(\sigma_a)]$ whose width scales like $O\left(\sqrt{\frac{\log\log t + \log\frac{1}{\delta}}{t}}\right)$. Using these confidence sequences on σ_a, we can construct a confidence sequence for the Neyman allocation

as follows

$$\mathcal{C}_t[\pi^*] = \left[\frac{\underline{C}_t(\sigma_1)}{\overline{C}_t(\sigma_0) + \underline{C}_t(\sigma_1)}, \frac{\overline{C}_t(\sigma_1)}{\underline{C}_t(\sigma_0) + \overline{C}_t(\sigma_1)}\right]. \tag{3.107}$$

Interpretation as Optimism. We can interpret our algorithm as implementing the celebrated principle of *optimism in the face of uncertainty*. Optimism is an algorithmic design principle which is the basis of many well-known MAB and reinforcement learning algorithms (such as the "upper confidence bound"). Roughly speaking, the principle states that we should act as if the underlying problem instance is the easiest instance, which is feasible according to our past observations. In the regret minimization framework, this means playing the arm which has the largest upper confidence bound. For adaptive ATE estimation, this involves playing the allocation that is closest to $\frac{1}{2}$. This is because the difficulty of a problem is determined by the deviation of the Neyman allocation from $\frac{1}{2}$ – when the Neyman allocation is close to $\frac{1}{2}$, the objectives of exploration and exploitation are aligned. Suppose the Neyman allocation deviates from $\frac{1}{2}$, then as the allocation we play converges to the Neyman allocation, we are necessarily under-sampling one arm and thus *slowing* down our convergence to the Neyman allocation. This intuition is supported by prior results showing that the Neyman regret scales inversely with $|\pi - \frac{1}{2}|$. Therefore, implementing optimism for adaptive ATE estimation involves playing the most feasible allocation (as determined by our past observations) closest to $\frac{1}{2}$ – this is exactly the driving principle behind our OPT algorithm.

Theoretical Analysis

In this section, we build our intuition on the behavior of OPT and conclude by stating our main result which is a bound on the Neyman regret of OPT.

Before we begin, we introduce some additional notation which will make our exposition easier. For any π, we define $\Delta(\pi) = |\frac{1}{2} - \pi|$ and $\underline{\pi} = \min\{\pi, 1-\pi\}$. Additionally, we let $\Delta_\sigma = \sigma_1 - \sigma_0$.

Our analysis splits the behavior of OPT into two phases, an exploration *exploration phase* and the *concentration phase*. We define the exploration phase as the rounds for which $\pi_t = \frac{1}{2}$. During the early stages of interaction, we expect that each arm has been played sufficiently few times so that $\frac{1}{2} \in \mathcal{C}_t[\pi^*]$, and the exploration time T_0 is the length of this phase. Intuitively, during this phase, there is not enough information in our observations to reliably predict π^* and so our best choice is to explore each arm uniformly. Fortunately, the length of this phase is not too long, and our first result bounds the length of this phase in terms of the absolute distance between the standard deviations.

Lemma 3.2.1 (Exploration Phase Length)**.** *Define the exploration time as*

$$T_0 = \min\{t : \pi_t \neq \frac{1}{2}\}. \tag{3.108}$$

Then, with probability at least $1-\delta$*, we have*

$$T_0 = \tilde{O}\left[\Delta_\sigma^{-2} \log \frac{1}{\delta}\right]. \tag{3.109}$$

This result shows that OPT is able to adapt to the difficulty of the underlying problem instance — if the gap between the standard deviations is large, then the exploration phase will be short, and if the gap is small, then the exploration phase will be longer.

Once the exploration phase is over, the algorithm will be able to focus on the concentration phase. In this phase, optimism guarantees $\Delta(\pi_t) < \Delta(\pi^*)$. Therefore, we can control the number of times each arm is played which we can in turn convert to bounds on $|\pi_t - \pi^*|$.

Our next result formalizes this intuition.

Lemma 3.2.2 (Policy Convergence)**.** *With probability at least* $1-\delta$*, we have that*

$$\pi_t - \pi^* = \tilde{O}\left[\sqrt{\frac{\log \frac{1}{\delta}}{\underline{\pi^*} \cdot t}} \cdot \frac{1}{\sigma_0 + \sigma_1}\right]. \tag{3.110}$$

The reason for the appearance of $\underline{\pi^*}$ is due to the convergence of π_t based on the number of times that both arms have been played. If we play one arm too often, then the width of the confidence interval for π^* would depend entirely on the width of the lesser sampled arm.

Our main result combines the above lemmas to provide a bound on the Neyman regret.

Theorem 3.2.3 (Main Result)**.** *With probability at least* $1-\delta$*, the Neyman regret of OPT is upper-bounded as*

$$\tilde{O}\left[\Delta_\sigma^{-2} + \left(\frac{1}{\underline{\pi^*}}\right)^2 \log T\right]. \tag{3.111}$$

The first term above is the per-round Neyman regret during the exploration phase and our bound follows from the fact that the Neyman regret is at most 4 when we play $\pi_t = \frac{1}{2}$. The second term in our bound is the Neyman regret during the concentration phase and follows from the application of Lemma 3.2.2 in conjunction with prior results showing that the Neyman regret scales according to $|\pi^* - \pi_t|^2 \approx \frac{1}{\pi^* \cdot t}$. Since the contribution to the Neyman regret from the reward estimation also scales like $\frac{1}{\pi^* \cdot t}$, taking a sum over these two terms gives us the desired result.

In order to get a better understanding of our result, we consider the behavior of a hypothetical algorithm which plays the optimal Neyman allocation π^* but incurs a loss based on the empirically computed allocation, π_t. A simple calculation shows that π_t converges to π^* at a rate of $\Theta[(\pi^* \cdot t)^{-\frac{1}{2}}]$. This in turn implies that the Neyman regret would be

$$\tilde{O}\left[\left(\frac{1}{\underline{\pi^*}}\right)^2 \log T\right], \tag{3.112}$$

which, modulo the regret from the exploration phase, is the same as the Neyman regret incurred

by OPT. This suggests that our algorithm is correctly adapting to the difficulty of the problem.

Comparison with Prior Work. At first glance, our result appears to be quite similar to the Neyman regret bound from prior work which similarly shows a logarithmic bound on the Neyman regret. However, this is not the case, due to differing definitions of the Neyman regret. In prior work, the Neyman regret is defined with respect to the minimum variance over allocations for the fixed IPW estimator. Our Neyman regret is defined with respect to the minimum attainable variance over any *pair* of estimators and allocations. This means that while our regret bounds share a similar form, the performance of our algorithm is significantly better than the performance of clipping-based algorithms. Concretely, using our definition of the Neyman regret to characterize the performance of clipping algorithms, we see that these algorithms actually have *linear* Neyman regret since the variance of their policies cannot converge to the minimum attainable variance.

Experiments

In this section, we present experiments to evaluate the empirical performance of our algorithm. We compare OPT against the Clipped Standard-Deviation Tracking (ClipSDT) algorithm, as well as two oracle algorithms that follow the Neyman allocation. One of these oracle algorithms sequentially estimates the reward, while the other has access to the true reward.

We do not include results for other clipping-based algorithms, as their variances fail to converge to the oracle variance, consistently leading to significantly worse performance than the other algorithms which obscures the clarity of the plots. This outcome is expected, given that both algorithms incur linear Neyman regret.

We consider 6 problem instances where both arms follows Bernoulli distributions. For each of these problem instances, we fix the treatment mean to be $\frac{1}{2}$ and vary the control mean in order to vary the Neyman allocation. For each of these problems, we run OPT, ClipSDT, and the reward estimation oracle for T ranging from 100 to 2000 and plot the normalized MSE ($T \cdot \text{MSE}$) over multiple simulations. For the oracle baseline, we explicitly compute the MSE.

Our results show that OPT consistently outperforms ClipSDT over all problem instances. The difference between the two becomes negligible for larger values of T which is expected since all algorithms eventually converge to the Neyman allocation and true reward function. However, for smaller sample sizes, we see that OPT provides around a 10-15 percent improvement over ClipSDT. This improvement is due to the reasons given in Section 3.2.4.

The performance of OPT is competitive with the reward estimation oracle for moderate values of π^* and even outperforms the reward estimation oracle on some problem instances. This is because OPT is more exploratory and obtains better reward estimates early on.

Key Findings.

- OPT provides substantial improvements over clipping-based methods, especially in small-

sample regimes critical for applications like clinical trials

- The optimistic approach naturally balances exploration and exploitation without requiring problem-specific hyperparameter tuning
- Performance is competitive with oracle methods that have access to additional information about the problem

3.2.5 Conclusion

This work proposed a new algorithm for adaptive ATE estimation. We identified some key issues with past approaches which limited their performance both empirically and theoretically and demonstrated how to resolve them. Our proposed solution borrows ideas from the literature on regret minimization and showed how to extend some of these ideas to the problem of adaptive ATE estimation. We believe that these ideas will be crucial for developing adaptive algorithms for inference for more complicated settings as well as for related problems like off-policy evaluation.

Key Contributions. Our main contributions are:

1. We developed and analyzed a new algorithm, Optimistic Policy Tracking (OPT), for adaptive estimation of ATE that enjoys significant theoretical improvements over previous approaches along with a significantly simplified analysis.
2. We performed simulations that demonstrate that our theoretical improvements translate into empirical improvements, especially in the small sample regime, which is critical for applications like randomized clinical trials.

Future Directions. We believe there are several compelling directions for future work:

- Extension to settings with covariates and more sophisticated reward estimation using non-parametric regression methods
- Generalization to multiple arms, where the correct extension involves computing a confidence interval around the Neyman allocation and projecting onto the uniform distribution
- Application to more complex interaction protocols such as reinforcement learning settings
- Development of similar optimistic principles for other causal inference problems beyond ATE estimation

The insights developed in this work demonstrate how optimistic design principles from bandit theory can be successfully adapted to causal inference problems involving complex estimators, providing both theoretical guarantees and practical improvements that suggest broader applicability of optimistic design in experimental settings.

3.2.6 Proofs

Preliminaries We will begin by defining our good event. Consider the following events

$$\mathcal{E}_\sigma(\delta) = \bigcap_{a,t\in\mathbb{N}} \left\{ |\widehat{\sigma}_t(a) - \sigma(a)| \leq 4.2\sqrt{\frac{\ell(t,\delta)}{t}} \right\} \tag{3.113}$$

$$\mathcal{E}_N(\delta) = \bigcap_{t\in\mathbb{N}} \left\{ \left| N_t(a) - \sum \pi_t(a) \right| \leq \sqrt{t\ell(t,\delta)} \right\} \tag{3.114}$$

$$\mathcal{E}_r(\delta) = \bigcap_{at\in\mathbb{N}} \left\{ |\widehat{r}_t(a) - r^\star(a)| \leq \sqrt{t\ell(t,\delta)} \right\}. \tag{3.115}$$

Let $\tilde{\delta} = \frac{\delta}{5}$ and define the good event $\mathcal{E}(\tilde{\delta}) = \mathcal{E}_\sigma(\delta) \cap \mathcal{E}_N(\delta) \cap \mathcal{E}_r(\delta)$. Applying Lemma 3.2.5 to control $\mathcal{E}_\sigma(\tilde{\delta})$ and Theorem 1 from [22] to control $\mathcal{E}_N(\tilde{\delta})$, and $\mathcal{E}_r(\tilde{\delta})$ shows that the event $\mathcal{E}(\tilde{\delta})$ occurs with probability at least $1-\delta$. Throughout the remained of this section, we assume the good event holds.

Proof of Theorem 1

We begin by decomposing the Neyman regret

$$\mathfrak{N}_T = \sum_{t=1}^{T} \ell(\pi_t, \widehat{r}_t) \tag{3.116}$$

$$= \sum_{t=1}^{T} \sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} + \frac{1-\pi_t(a)}{\pi_t(a)} \varepsilon_t^2(a) - \frac{\sigma^2(a)}{\pi_{\text{Ney}}[a]} \right) \tag{3.117}$$

$$= \sum_{t=1}^{\mathbf{T}} \sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} + \frac{1-\pi_t(a)}{\pi_t(a)} \varepsilon_t^2(a) - \frac{\sigma^2(a)}{\pi_{\text{Ney}}[a]} \right) + \sum_{t=\mathbf{T}+1}^{T} \sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} + \frac{1-\pi_t(a)}{\pi_t(a)} \varepsilon_t^2(a) - \frac{\sigma^2(a)}{\pi_{\text{Ney}}[a]} \right). \tag{3.118}$$

For the first term, we have that $\pi_t = \frac{1}{2}$, and $\varepsilon_t(a) \leq 1$, so that

$$\sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} + \frac{1-\pi_t(a)}{\pi_t(a)} \varepsilon_t^2(a) - \frac{\sigma^2(a)}{\pi_{\text{Ney}}[a]} \right) \tag{3.119}$$

$$\leq \sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} - \frac{\sigma^2(a)}{\pi_{\text{Ney}}[a]} \right) + 2 \tag{3.120}$$

$$\leq 4, \tag{3.121}$$

to so that the regret from the exploration phase is $4\mathbf{T}$.

For the second term, we have

$$\sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} + \frac{1-\pi_t(a)}{\pi_t(a)} \varepsilon_t^2(a) - \frac{\sigma^2(a)}{\pi_{\text{Ney}}[a]} \right) \tag{3.122}$$

$$= \sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} - \frac{\sigma^2(a)}{\pi_{\text{Ney}}(a)} \right) + \sum_a \left(\frac{1-\pi_t(a)}{\pi_t(a)} \varepsilon_t^2(a) \right) \tag{3.123}$$

$$\tag{3.124}$$

We can bound the first term by applying Lemma 4.3 from [3] in conjunction with so that

$$\sum_a \left(\frac{\sigma^2(a)}{\pi_t(a)} - \frac{\sigma^2(a)}{\pi_{\text{Ney}}(a)} \right) \leq \frac{625}{(\sigma(0)+\sigma(1))^2} \frac{\ell(t,\delta)}{\pi_{\text{Ney}} t} \tag{3.125}$$

In order to bound the second term, we observe that on the good event

$$|r^\star(a) - \widehat{r}_t(a)| \leq \sqrt{\frac{\ell(t,\delta)}{N_t(a)}} \tag{3.126}$$

$$\leq \sqrt{\frac{\ell(t,\delta)}{\pi_{\text{Ney}} t - \sqrt{t\ell(t,\delta)}}} \tag{3.127}$$

$$\leq 2\sqrt{\frac{\ell(t,\delta)}{\pi_{\text{Ney}} t}}, \tag{3.128}$$

where in the last line we have again applied Lemma 4.5 from [3].

Therefore, we have that

$$\sum_a \left(\frac{1-\pi_t(a)}{\pi_t(a)} \varepsilon_t^2(a) \right) \leq \frac{8\ell(t,\delta)}{(\pi_{\text{Ney}})^2 t} \tag{3.129}$$

We can bound the sum of these two terms as $625\frac{\ell(t,\delta)}{(\pi_{\text{Ney}})^2 t}$. The result then follows by summing this over $t < T$ and adding the Neyman regret from the exploration phase.

Proof of Lemma 3.2.1

Proof. Suppose, without loss of generality, that $\pi_{\text{Ney}} < \frac{1}{2}$; in order to obtain results for $\pi_{\text{Ney}} > \frac{1}{2}$, we can simply flip the roles of the treatment and control arms. For the case that $\pi_{\text{Ney}} = \frac{1}{2}$, then OPT will always play π_t.

Since $\pi_{\text{Ney}} < \frac{1}{2}$, bounding $\mathbf{T}$ is equivalent to determining the largest time t such that $\mathcal{U}_[(t)][\pi_{\text{Ney}}] < \frac{1}{2}$, i.e we wish to compute

$$\min \left\{ t : \frac{\sigma(1) + 4.2\sqrt{\frac{\ell(t,\delta)}{N_t(1)}}}{\sigma(0) + \sigma(1) + 4.2\sqrt{\frac{\ell(t,\delta)}{N_t(1)}} - 4.2\sqrt{\frac{\ell(t,\delta)}{N_t(0)}}} < \frac{1}{2} \right\} \tag{3.130}$$

Using the fact that $\pi_t = \frac{1}{2}$ for all $t < \mathbf{T}$, can control

$$N_t(a) \in \left[\frac{t}{2} \pm 1.7\sqrt{t\ell(t,\delta)}\right]. \tag{3.131}$$

Plugging this into equation Eq. (3.130) and rearranging shows that we need to bound

$$\min\left\{t : \frac{\ell(t,\delta)}{t\left(\frac{1}{2} - 1.7\sqrt{\frac{\ell(t,\delta)}{t}}\right)} < \frac{\Delta_{(\sigma)}^2}{18}\right\}. \tag{3.132}$$

Applying Lemma B.10 from [3] shows that whenever $t \geq \widetilde{\mathcal{O}}\left(\log(\frac{1}{\delta})\right)$, we have that $1.7\sqrt{\frac{\ell(t,\delta)}{t}} < \frac{1}{4}$ so that we need to bound

$$\min\left\{t : t > \frac{64}{\Delta_{(\sigma)}^2}\ell(t,\delta)\right\}. \tag{3.133}$$

Another application of Lemma B.10 shows that this quantity is bounded by

$$\frac{64}{\Delta_{(\sigma)}^2}\log\frac{5.2}{\delta} + \frac{64}{\Delta_{(\sigma)}^2}\log\log\frac{64}{\Delta_{(\sigma)}^2} \tag{3.134}$$

which gives us the desired result.

□

Proof of Lemma 3.2.2

Lemma 3.2.4. *Let $t \geq \mathbf{T}$. Then, with probability at least $1-\delta$, we have that*

$$\pi_{t+1} - \pi_{Ney} \leq \frac{25}{\sigma(0) + \sigma(1)}\sqrt{\frac{\ell(t,\delta)}{\pi_{Ney}t}}. \tag{3.135}$$

Proof. WLOG we assume $\pi_{\text{Ney}} < \frac{1}{2}$ so that $\underline{\pi_{\text{Ney}}} = \pi_{\text{Ney}}$. First note that $s \geq \mathbf{T}$, we have that

$$\pi_{t+1} \in \left[\pi_{\text{Ney}}, \frac{\sigma(1) + Z_{1,t}}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}}\right] \tag{3.136}$$

$$= \left[\pi_{\text{Ney}}, \pi_{\text{Ney}}\frac{\sigma(0) + \sigma(1)}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}} + \frac{Z_{1,t}}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}}\right] \tag{3.137}$$

$$\subset \left[\pi_{\text{Ney}}, \frac{1}{2}\right], \tag{3.138}$$

where we have defined

$$Z_t(a) = 4.2\sqrt{\frac{\ell(t,\delta)}{N_t(a)}},$$

and equation Eq. (3.138) follows from the definition of the $\mathbf{T}$.

Since $\pi_t \in \left[\pi_{\text{Ney}}, \frac{1}{2}\right]$, we know that $1 - \pi_t \in \left[\frac{1}{2}, 1 - \pi_{\text{Ney}}\right]$ which we use to control the number of times each arm is played.

$$N_t(1) \geq \pi_{\text{Ney}} \cdot t - \sqrt{t\ell(t,\delta)} \tag{3.139}$$

$$N_t(0) \geq \frac{t}{2} - \sqrt{t\ell(t,\delta)}. \tag{3.140}$$

Plugging these values into the upper bound in equation Eq. (3.137), some algebra shows that

$$\pi_{t+1} - \pi_{\text{Ney}} = \pi_{\text{Ney}} \frac{\sigma(0) + \sigma(1)}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}} + \frac{Z_{1,t}}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}} - \pi_{\text{Ney}} \tag{3.141}$$

$$= \pi_{\text{Ney}} \cdot \frac{Z_0(t) - Z_1(t)}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}} + \frac{Z_{1,t}}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}} \tag{3.142}$$

$$\leq \frac{Z_{0,t}}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}} + \frac{Z_{1,t}}{\sigma(0) + \sigma(1) + Z_{1,t} - Z_{0,t}} \tag{3.143}$$

$$\leq 8.4\sqrt{\frac{\ell(t,\delta)}{\pi_{\text{Ney}} t - \sqrt{t\ell(t,\delta)}}} \cdot \left(\frac{1}{\sigma(0) + \sigma(1) - Z_{0,t}}\right). \tag{3.144}$$

Applying Lemma B.10 from Neopane et al. [3], we have that when $t = \widetilde{\mathcal{O}}\left(\left(\frac{1}{\pi_{\text{Ney}}}\right)^2 \log \frac{1}{\delta}\right)$, we have that $\pi_{\text{Ney}} t - \sqrt{t\ell(t,\delta)} \geq \frac{1}{2}\pi_{\text{Ney}} t$. Next, since $t \geq \mathbf{T}$, we have that

$$Z_{0,t} = 4.2\sqrt{\frac{\ell(t,\delta)}{t}} \tag{3.145}$$

$$\leq \frac{\Delta_{(\sigma)}}{8}. \tag{3.146}$$

Therefore,

$$\sigma(0) + \sigma(1) + Z_{1,t} \geq \sigma(0) + \sigma(1) - \frac{\Delta_{(\sigma)}}{8} \tag{3.147}$$

$$= \sigma(0) + \sigma(1) - \frac{\sigma(0) - \sigma(1)}{8} \tag{3.148}$$

$$\geq \frac{\sigma(0) + \sigma(1)}{2}. \tag{3.149}$$

Combining these results, we have that

$$\pi_{t+1} - \pi_{\text{Ney}} \leq \frac{25}{\sigma(0) + \sigma(1)} \sqrt{\frac{\ell(t,\delta)}{\pi_{\text{Ney}} t}}, \tag{3.150}$$

which proves the desired result. □

Concentration Results

The proof of this lemma is based on a similar proof found in [104] and extends the results to hold in the sequential setting.

Lemma 3.2.5. *Let (X_t) be a $[0,1]$-valued stochastic process defined on some filtration $(\mathcal{F}_t)$ satisfying $\mu = \mathbb{E}_{t-1}[X_t]$ and $\sigma^2 = \mathbb{V}_{t-1}[X_t]$. Define*

$$\mu_t = \frac{1}{t}\sum_{t=1}^{t} X_t \tag{3.151}$$

$$\widehat{\sigma}_t^2 = \frac{1}{t}\sum_{s=1}^{t} (X_t - \mu_t)^2 . \tag{3.152}$$

Then, with probability at least $1-\delta$, for all $t \geq 2$ we have that

$$\sigma \in \left[\widehat{\sigma}_t - 1.7\sqrt{\frac{\ell(t,\delta)}{t}}, \widehat{\sigma}_t + 4.2\sqrt{\frac{\ell(t,\delta)}{t}}\right]. \tag{3.153}$$

Proof. Define $Y_t = (X_t - \mu)^2 - \sigma^2$, and $S_t = \sum_{i=1}^{t} Y_t$. Letting $\mathcal{V} = \mathbb{V}_{t-1}[Y_t]$, we apply Theorem 1 from [22] which gives us the following time-uniform Bernstein inequality (see Table 3 in the Appendix). Applying a union bound, we have with probability at least $1-\delta$, for all $t \in \mathbb{N}$, that

$$|\mu_t - \mu| \leq 1.7\sigma\sqrt{\frac{\ell\left(t, \frac{\delta}{4}\right)}{t}} + 1.7\frac{\ell\left(t, \frac{\delta}{4}\right)}{t}, \tag{3.154}$$

$$|Y_t| \leq 1.7\sqrt{\frac{\mathcal{V}\ell\left(t, \frac{\delta}{4}\right)}{t}} + 1.7\frac{\ell\left(t, \frac{\delta}{4}\right)}{4t} \tag{3.155}$$

$$\leq 1.7\sigma\sqrt{\frac{\ell\left(t, \frac{\delta}{4}\right)}{t}} + 1.7\frac{\ell\left(t, \frac{\delta}{4}\right)}{t}, \tag{3.156}$$

where we set $\ell(t,\delta) = \log\log 2t + 0.72\log\frac{5.2}{\delta}$ and the last inequality follows from the fact that

$\mathcal{V} < \sigma^2$. Letting $\mu_t = \frac{1}{t}\sum_{s=1}^{t} X_s$ some algebra demonstrates that

$$\begin{aligned}
S_t &= \sum_{i=1}^{t}(X_i - \mu)^2 - \sigma^2 \\
&= \sum_{i=1}^{t}\left[((X_i - \mu_t) - (\mu_t - \mu))^2 - \sigma^2\right] \\
&= \sum_{i=1}^{t}\left[(X_i - \mu_t)^2 + 2(X_i - \mu_t)(\mu_t - \mu) + (\mu_t - \mu)^2 - \sigma^2\right] \\
&= t\sigma_t^2 + 2(\mu_t - \mu)\sum_{i=1}^{t}(X_i - \mu_t) + t(\mu_t - \mu)^2 - t\sigma^2 \\
&= t\sigma_t^2 + 0 + t(\mu_t - \mu)^2 - t\sigma^2 \\
&= t(\sigma_t^2 - \sigma^2 + (\mu_t - \mu)^2),
\end{aligned}$$

which implies

$$\left(\sigma_t^2 - \sigma^2\right) = \frac{1}{t}\sum_{s=1}^{t} Y_s - (\mu_t - \mu)^2 \leq \frac{1}{t}\sum_{s=1}^{t} Y_s. \tag{3.157}$$

Letting $L = \frac{\ell(t,\delta)}{t}$, and applying the bounds in equations Eq. (3.154) and 3.156, some algebra shows that

$$\sigma^2 + 1.7\sigma\sqrt{L} + 1.7L - \sigma_t^2 \geq 0. \tag{3.158}$$

Completing the square and rearranging shows that

$$\sigma \geq \sqrt{\sigma_t^2 + (1.7^2 - 1.7)\,L} - 1.7\sqrt{L} \tag{3.159}$$

$$\geq \sigma_t - 1.7\sqrt{L}. \tag{3.160}$$

Repeating the same argument with $-Y_t$ shows that

$$\sigma \leq \sigma_t + 4.2\sqrt{L}. \tag{3.161}$$

Combining these bounds we have with probability at least $1 - \delta$, for all $t > 2$

$$\sigma \in \left[\sigma_t - 1.7\sqrt{\frac{\ell(t,\delta)}{t}}, \sigma_t + 4.2\sqrt{\frac{\ell(t,\delta)}{t}}\right]. \tag{3.162}$$

□

Misc. Results

Lemma 3.2.6. *For any* $\texttt{Alg}$*, we have that*

$$\mathbb{V}_{\texttt{Alg},\nu}\left[\widehat{\Delta}_T\right] = \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}_{\texttt{Alg},\nu}\left[\texttt{AIPW}_t\right] \tag{3.163}$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}_{\texttt{Alg},\nu}\left[\sum_a \frac{\sigma^2(a)}{\pi_t(a)} + \left(\frac{1-\pi_t(a)}{\pi_t(a)}\right)\varepsilon_{t-1}^2(a)\right] \tag{3.164}$$

Proof. Leting $z_t = \texttt{AIPW}_t - \Delta$, we have

$$\mathbb{V}_{\texttt{Alg},\nu}\left[\widehat{\Delta}_T\right] = \frac{1}{T^2}\mathbb{E}\left[\left(\sum_{t=1}^{T} z_t\right)^2\right] \tag{3.165}$$

$$= \frac{1}{T^2}\left(\sum_{t=1}^{T}\mathbb{E}\left[z_t^2\right] + \sum_{t=1}^{T}\sum_{s=1}^{t=1}\mathbb{E}\left[z_t \cdot z_s\right]\right) \tag{3.166}$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\left[z_t^2\right] \tag{3.167}$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}\left[\texttt{AIPW}_t\right]. \tag{3.168}$$

Then applying the law of total variance shows that $\mathbb{V}_{\texttt{Alg},\nu}\left[\texttt{AIPW}_t\right] = \mathbb{E}_{\texttt{Alg},\nu}\left[\mathbb{V}\left[\texttt{AIPW}_t \mid \mathcal{F}_{t-1}\right]\right]$ since $\mathbb{V}\left[\mathbb{E}\left[\texttt{AIPW}_t \mid \mathcal{F}_{t-1}\right]\right] = 0$. Computing the conditional variance, we obtain

$$\mathbb{V}_{\texttt{Alg},\nu}\left[\texttt{AIPW}_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{\texttt{Alg},\nu}\left[\left(\texttt{AIPW}_t - \Delta\right)^2 \mid \mathcal{F}_{t-1}\right] \tag{3.169}$$

$$= \mathbb{E}_{\texttt{Alg},\nu}\left[\left(w_t\left(\delta_t + \varepsilon_{t-1}\right) + \widehat{\Delta}_{t-1}^{(\texttt{r})} - \Delta\right)^2 \mid \mathcal{F}_{t-1}\right] \tag{3.170}$$

$$= \mathbb{E}_{\pi_t}\left[w_t^2\left(\sigma^2 + \varepsilon_{t-1}^2\right) - \left(\Delta - \widehat{\Delta}_{t-1}^{(\texttt{r})}\right)^2\right] \tag{3.171}$$

$$= \sum_a \frac{\left(\sigma^2(a) + \varepsilon_{t-1}^2(a)\right)}{\pi_t(a)} - \left(\varepsilon_{t-1}(1) - \varepsilon_{t-1}(0)\right)^2 \tag{3.172}$$

$$= \sum_a \left[\frac{\sigma^2(a)}{\pi_t(a)} + \left(\frac{1}{\pi_t(a)} - 1\right)\cdot\varepsilon_{t-1}^2(a)\right] + 2\varepsilon_{t-1}(1)\cdot\varepsilon_{t-1}(0) \tag{3.173}$$

$$= \sum_a \left[\frac{\sigma^2(a)}{\pi_t(a)} + \left(\frac{1-\pi_t(a)}{\pi_t(a)}\right)\cdot\varepsilon_{t-1}^2(a)\right] + 2\varepsilon_{t-1}(1)\cdot\varepsilon_{t-1}(0). \tag{3.174}$$

Therefore, we have

$$\mathbb{V}_{\mathtt{Alg},\nu}\left[\widehat{\Delta}_T\right] = \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_a\left(\frac{\sigma^2(a)}{\pi_t(a)} + \left(\frac{1-\pi_t(a)}{\pi_t(a)}\right)\cdot\varepsilon_{t-1}^2(a)\right) + 2\varepsilon_{t-1}(1)\cdot\varepsilon_{t-1}(0)\right] \tag{3.175}$$

$$= \mathbb{E}_{\mathtt{Alg},\nu}\left[\sum_a \frac{\sigma^2(a)}{\pi_t(a)} + \left(\frac{1-\pi_t(a)}{\pi_t(a)}\right)\cdot\varepsilon_{t-1}^2(a)\right], \tag{3.176}$$

where the second inequality follows from the fact that $\varepsilon_t^2(a)$ are uncorrelated. □

Chapter 4

Conclusion

This dissertation has presented a unified approach to sequential decision-making and adaptive experimental design, developing theoretically grounded algorithms that achieve significant improvements in both sample complexity and practical performance. Through contributions spanning transfer learning, causal inference, and preference-based learning, we have advanced the state-of-the-art by bridging the gap between theory and practice.

4.1 Summary of Contributions

Our work has made several key contributions across multiple domains:

Transfer Learning in Multi-Armed Bandits (Chapter 2.2). We developed algorithms that can effectively leverage auxiliary information from related source tasks while maintaining robustness against negative transfer. Our approach provides theoretical guarantees that gracefully interpolate between perfect transfer scenarios and learning from scratch.

Adaptive Experimental Design (Chapters 3.1 and 3.2). We introduced two complementary approaches for improving the efficiency of Average Treatment Effect estimation. The ClipSMT algorithm achieves exponential improvements in regret from $O(\sqrt{T})$ to $O(\log T)$, while the Optimistic Policy Tracking method leverages the AIPW estimator through principled optimistic design.

Active Preference Learning (Chapter 2.3). We formalized active exploration in preference-based learning as a contextual dueling bandit problem, developing algorithms with polynomial regret bounds and practical extensions to RLHF and DPO for large language models.

4.2 Unifying Insights

Several key insights emerge from our work that extend beyond the specific technical contributions:

The Importance of Finite-Sample Analysis. A consistent theme throughout this dissertation is the significant gap between asymptotic optimality and finite-sample performance. Our work demonstrates that algorithms designed with finite-sample considerations often achieve dramatically better practical performance while maintaining strong theoretical guarantees.

Optimism as a Design Principle. The principle of optimism, well-established in bandit theory, proves remarkably effective across diverse domains. Our work shows how optimistic design can be adapted to causal inference and preference learning, suggesting broader applicability of this algorithmic paradigm.

Adaptive Algorithms for Complex Estimators. While much prior work focuses on simple estimators for tractability, our research demonstrates that adaptive algorithms can effectively leverage more sophisticated estimators like AIPW while maintaining theoretical guarantees and improving practical performance.

4.3 Directions for Future Work

Our contributions open several promising directions for future research:

Multi-Task Learning Extensions. The transfer learning framework developed in Chapter 2.2 could be extended to more complex multi-task scenarios, including hierarchical task relationships and continual learning settings.

High-Dimensional Causal Inference. The adaptive experimental design methods could be extended to high-dimensional settings with many treatments or covariates, potentially leveraging modern techniques from high-dimensional statistics.

Foundation Model Alignment. The active preference learning framework provides a foundation for more sophisticated approaches to aligning large language models and other foundation models with human values and preferences.

Robust Algorithm Design. Future work could explore how to make adaptive algorithms more robust to model misspecification and distribution shift, building on the robustness insights from our transfer learning work.

4.4 Broader Impact

The algorithms and insights developed in this dissertation have potential applications across numerous domains where sequential decision-making and adaptive experimentation are crucial. From improving the efficiency of clinical trials to enabling more effective human-AI interaction, our contributions provide practical tools for addressing real-world challenges while maintaining theoretical rigor.

The emphasis on finite-sample performance is particularly important for applications where data is expensive or limited, such as medical research, where our adaptive experimental design methods could reduce the cost and duration of clinical trials while improving statistical power.

4.5 Final Remarks

Sequential decision-making under uncertainty remains a rich and challenging area with countless opportunities for impactful research. This dissertation has advanced our understanding by developing principled approaches that bridge theory and practice, but many important questions remain open. We hope that the insights and techniques developed here will inspire future work that continues to push the boundaries of what is possible in adaptive algorithm design.

The journey from theoretical insights to practical algorithms is often long and challenging, but the potential impact makes this effort worthwhile. As we continue to develop more sophisticated AI systems and face increasingly complex decision-making challenges, the need for principled, adaptive approaches will only continue to grow.

Chapter 5

Bibliography

[1] Ojash Neopane, Aaditya Ramdas, and Aarti Singh. Best arm identification under additive transfer bandits. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pages 464–470. IEEE, 2021. Cited on page 1.

[2] Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint arXiv:2312.00267*, 2023. Cited on page 2.

[3] Ojash Neopane, Aaditya Ramdas, and Aarti Singh. Logarithmic Neyman regret for adaptive estimation of the average treatment effect. *AISTATS*, 2025. Cited on pages 2, 68, 78, 79, and 80.

[4] Ojash Neopane, Aaditya Ramdas, and Aarti Singh. Optimistic algorithms for adaptive estimation of the average treatment effect. In *Forty-second International Conference on Machine Learning*, 2025. Cited on page 2.

[5] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009. Cited on pages 4 and 16.

[6] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer, 2010. Cited on page 4.

[7] Abbas Kazerouni and Lawrence M Wein. Best arm identification in generalized linear bandits. *arXiv preprint arXiv:1905.08224*, 2019. Cited on page 4.

[8] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, and Kurt Konolige. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018. Cited on page 4.

[9] Andrei A Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *Conference on Robot*

Learning, pages 262–270, 2017. Not cited.

[10] Fereshteh Sadeghi and Sergey Levine. CAD2RL: Real single-image flight without a single real image. In *Proceedings of Robotics: Science and Systems XIII*, 2017. Cited on page 4.

[11] R. Combes, J. Ok, A. Proutiere, D. Yun, and Y. Yi. Optimal rate sampling in 802.11 systems: Theory, design, and implementation. *IEEE Transactions on Mobile Computing*, 18(5): 1145–1158, 2018. Cited on page 5.

[12] H. Qi, Z. Hu, X. Wen, and Z. Lu. Rate adaptation with thompson sampling in 802.11 ac wlan. *IEEE Communications Letters*, 23(10):1888–1892, 2019. Cited on page 5.

[13] Aurélien Garivier, Emilie Kaufmann, and Wouter M Koolen. Maximin action identification: a new bandit framework for games. In *Conference on Learning Theory*, pages 1028–1050, 2016. Cited on page 5.

[14] Emilie Kaufmann and Wouter M Koolen. Monte-carlo tree search by best arm identification. In *Advances in Neural Information Processing Systems*, pages 4897–4906, 2017. Cited on page 24.

[15] Ruitong Huang, Mohammad M Ajallooeian, Csaba Szepesvári, and Martin Müller. Structured best arm identification with fixed confidence. In *International Conference on Algorithmic Learning Theory*, pages 593–616, 2017. Cited on pages 5, 6, 16, and 17.

[16] Tor Lattimore and Rémi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558, 2014. Cited on page 5.

[17] Samarth Gupta, Gauri Joshi, and Osman Yağan. Exploiting correlation in finite-armed structured bandits. *arXiv preprint arXiv:1810.08164*, 2018. Cited on page 5.

[18] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, pages 10666–10676, 2019. Cited on pages 6 and 16.

[19] Julian Katz-Samuels, Lalit Jain, Zohar Karnin, and Kevin Jamieson. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *arXiv preprint arXiv:2006.11685*, 2020. Cited on page 6.

[20] Sandeep Juneja and Subhashini Krishnasamy. Sample complexity of partition identification using multi-armed bandits. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019. Cited on page 6.

[21] Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. *arXiv preprint arXiv:1702.05186*, 2017. Cited on page 6.

[22] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018. Cited

on pages 9, 10, 69, 77, and 81.

[23] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, volume 12, pages 655–662, 2012. Cited on pages 9 and 14.

[24] Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251, 2013. Cited on pages 9 and 14.

[25] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1690–1698, 2016. Cited on pages 9 and 14.

[26] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014. Cited on page 9.

[27] Lijie Chen, Anupam Gupta, and Jian Li. Pure exploration of multi-armed bandit under matroid constraints. In *Conference on Learning Theory*, pages 647–669, 2016. Not cited.

[28] Lijie Chen, Anupam Gupta, Jian Li, Mingda Qiao, and Ruosong Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pages 482–534, 2017. Not cited.

[29] Tongyi Cao and Akshay Krishnamurthy. Disagreement-based combinatorial pure exploration: Sample complexity bounds and an efficient algorithm. In *Conference on Learning Theory*, pages 558–588, 2019. Cited on page 9.

[30] Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, Ronald Ortner, and Peter Bartlett. Improved learning complexity in combinatorial pure exploration bandits. In *Artificial Intelligence and Statistics*, pages 1004–1012, 2016. Cited on pages 9 and 16.

[31] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016. Cited on page 10.

[32] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback. 2020. Cited on pages 25 and 39.

[33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. Cited on pages 25 and 39.

[34] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the

method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. Cited on pages 25, 26, and 27.

[35] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. Cited on page 25.

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. Cited on page 25.

[37] Xiang Li, Viraj Mehta, Johannes Kirschner, Ian Char, Willie Neiswanger, Jeff Schneider, Andreas Krause, and Ilija Bogunovic. Near-optimal policy identification in active reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3OR2tbtnYC-. Cited on pages 25, 26, 28, and 40.

[38] Yichong Xu, Aparna Joshi, Aarti Singh, and Artur Dubrawski. Zeroth order non-convex optimization with dueling-choice bandits. In *Conference on Uncertainty in Artificial Intelligence*, pages 899–908. PMLR, 2020. Cited on pages 25 and 27.

[39] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 17–23 Jul 2022. Cited on page 25.

[40] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. Cited on pages 25 and 39.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. Cited on page 25.

[42] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023. Cited on page 25.

[43] AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Cited on page 26.

[44] Ian Char, Youngseog Chung, Willie Neiswanger, Kirthevasan Kandasamy, Andrew Oakleigh Nelson, Mark Boyer, Egemen Kolemen, and Jeff Schneider. Offline contextual bayesian optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Gar-

nett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7876acb66640bad41f1e1371ef30c180-Paper.pdf. Cited on page 26.

[45] Louis L Thurstone. The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384, 1927. Cited on pages 26 and 27.

[46] Carl Edward Rasmussen, Christopher KI Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006. Cited on page 28.

[47] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017. Cited on page 36.

[48] Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, and Barnabas Poczos. Multi-fidelity gaussian process bandit optimisation. *Journal of Artificial Intelligence Research*, 66:151–196, 2019. Cited on pages 36 and 37.

[49] Volodymyr Mnih. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. Cited on page 38.

[50] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. Cited on page 39.

[51] J. Kreutzer, S. Khadivi, E. Matusov, and S Riezler. Can neural machine translation be improved with user feedback? 2018. Cited on page 39.

[52] C. Lawrence and S. Riezler. Improving a neural semantic parser by counterfactual learning from human bandit feedback. 2018. Cited on page 39.

[53] W. S. Cho, P. Zhang, Y. Zhang, X. Li, M. Galley, C. Brockett, M. Wang, and J. Gao. Towards coherent and cohesive long-form text generation. 2018. Cited on page 39.

[54] E. Perez, S. Karamcheti, R. Fergus, J. Weston, D. Kiela, and K Cho. Finding generalizable evidence by learning to convince q&a models. 2019. Cited on page 39.

[55] Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-tuning language models via epistemic neural networks. *arXiv preprint arXiv:2211.01568*, 2022. Cited on page 39.

[56] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. Cited on page 39.

[57] Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably sample efficient rlhf via active preference optimization. *arXiv preprint arXiv:2402.10500*, 2024. Cited on page 39.

[58] Shenao Zhang, Donghan Yu, Hiteshi Sharma, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024. Cited on page 39.

[59] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024. Cited on page 39.

[60] William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024. Cited on page 39.

[61] Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024. Cited on page 39.

[62] Jonas Hübotter, Sascha Bongni, Ido Hakimi, and Andreas Krause. Efficiently learning at test-time: Active fine-tuning of llms. *arXiv preprint arXiv:2410.08020*, 2024. Cited on page 39.

[63] Sally Hollis and Fiona Campbell. What is meant by intention to treat analysis? survey of published randomised controlled trials. *The BMJ*, 319(7211):670–674, 1999. Cited on page 41.

[64] Coady Wing, Kosali Simon, and Ricardo A Bello-Gomez. Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, 39: 453–469, 2018. Cited on page 41.

[65] Abhijit Vinayak Banerjee, Esther Duflo, and Michael Kremer. The influence of randomized controlled trials on development economics research and on development policy. *The State of Economics, The State of the World*, pages 482–488, 2016. Cited on page 41.

[66] Shein-Chung Chow, Mark Chang, and Annpey Pong. Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*, 15(4):575–591, 2005. Cited on page 41.

[67] Shein-Chung Chow and Mark Chang. Adaptive design methods in clinical trials–a review. *Orphanet Journal of Rare Diseases*, 3:1–13, 2008. Not cited.

[68] US Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics: guidance for industry. *Rockville: Food and Drug Administration*, page 2020, 2019. Cited on page 41.

[69] Jessica Dai, Paula Gradu, and Christopher Harshaw. Clip-OGD: An experimental design for adaptive Neyman allocation in sequential experiments. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on pages 41, 42, 46, and 68.

[70] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993. Cited on page 42.

[71] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21:167:1–167:63, 2019. Cited on page 42.

[72] Masahiro Kato, Takuya Ishihara, Junya Honda, and Yusuke Narita. Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020. Cited on pages 42 and 67.

[73] Andrew J Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR, 2023. Cited on page 42.

[74] Thomas Cook, Alan Mishler, and Aaditya Ramdas. Semiparametric efficient inference in adaptive experiments. In *Causal Learning and Reasoning*, pages 1033–1064. PMLR, 2024. Cited on pages 42, 43, 44, 47, and 68.

[75] Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Annals of Agricultural Sciences*, pages 1–51, 1923. Cited on page 43.

[76] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. Not cited.

[77] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Rress, 2015. Cited on page 43.

[78] Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2021. Cited on page 43.

[79] Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/JMS Journal of Data Science*, 2022. Cited on pages 43, 68, and 69.

[80] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2002. Cited on page 46.

[81] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995. Cited on page 47.

[82] Steven R. Howard, Aaditya Ramdas, Jon D. McAuliffe, and Jasjeet S. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 2021. Cited

on page 61.

[83] Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:123–150, 1934. Cited on page 67.

[84] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933. Cited on page 67.

[85] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. Cited on page 67.

[86] Jinyong Hahn, Keisuke Hirano, and Dean S. Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29:108 – 96, 2009. Cited on pages 67 and 68.

[87] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016. Cited on page 67.

[88] Ting Li, Chengchun Shi, Jianing Wang, Fan Zhou, et al. Optimal treatment allocation for efficient policy evaluation in sequential decision making. *Advances in Neural Information Processing Systems*, 36, 2024. Cited on page 68.

[89] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011. Cited on page 68.

[90] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011. Not cited.

[91] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016. Cited on page 68.

[92] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015. Cited on page 68.

[93] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017. Not cited.

[94] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020. Not cited.

[95] Cong Ma, Banghua Zhu, Jiantao Jiao, and Martin J Wainwright. Minimax off-policy eval-

uation for multi-armed bandits. *IEEE Transactions on Information Theory*, 68:5314–5339, 2021. Cited on page 68.

[96] Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34: 23714–23726, 2021. Cited on page 68.

[97] Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment for markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 5022–5050. PMLR, 2022. Cited on page 68.

[98] Josiah P. Hanna, Philip S. Thomas, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. In *International Conference on Machine Learning*, 2017. Cited on page 68.

[99] Ksenia Konyushova, Yutian Chen, Thomas Paine, Caglar Gulcehre, Cosmin Paduraru, Daniel J. Mankowitz, Misha Denil, and Nando de Freitas. Active offline policy selection. *Advances in Neural Information Processing Systems*, 34:24631–24644, 2021. Cited on page 68.

[100] Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15):e2014602118, 2021. Cited on page 68.

[101] Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. *Advances in Neural Information Processing Systems*, 33:9818–9829, 2020. Not cited.

[102] Kelly Zhang, Lucas Janson, and Susan Murphy. Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34:7460–7471, 2021. Cited on page 68.

[103] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B Methodological*, 2023. Cited on page 69.

[104] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Use of variance estimation in the multi-armed bandit problem. 2006. Cited on page 81.