THE UNIVERSITY OF CHICAGO

ALIGNING AI SYSTEM WITH HUMAN PREFERENCE: PERSONALIZATION, ROBUSTIFICATION, AND EVALUATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
YUWEI CHENG

CHICAGO, ILLINOIS
AUGUST 2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to express my deepest gratitude to my advisor, Professor Haifeng Xu. Thank you for your unwavering support and encouragement whenever I faced difficulties. Even after three years, I still vividly remember the time when I told you how discouraged I felt about my slow research progress. You wrote me a long email explaining why we call it "research" — because most of the time, a researcher must search, and search again, before finally finding an answer.

Your passion for understanding questions and exploring new ideas has always inspired me. I deeply admire your positive mindset and your perspective toward "failure." In fact, you often do not even view something as a failure at all, but rather as new knowledge about why a certain approach does not work. This way of thinking has continuously encouraged me and given me a new perspective on research and life. Your spirit of diving deeply into problems — making sure we fully understand why something works or does not work until everything becomes crystal clear — will remain a guiding principle in my future work.

Beyond your academic guidance, what I have learned most from you is how to live as a person. Your outlook on life has left a profound impact on me. I will always remember your reminder that people are not comparable: we are high-dimensional vectors without any natural ranking or ordering in such a complex space. Every individual possesses their own unique values, and it is important to discover what truly matters to us and bring our focus back to our own lives. I will carry this lesson with me deeply in the years ahead.

Second, I would like to express my sincere gratitude to my committee members.

Thank you, Professor Chao Gao, for the insightful discussions after my theoretical preliminary examination. Your intuition regarding distributions, concentration bounds, and the distinctions between sub-Gaussian and sub-exponential behavior has been deeply inspiring — especially your emphasis on understanding what a distribution truly describes, why certain quantities need to be controlled, what should be controlled, and how to control them.

Your overall philosophy of using intuition to guide proofs has profoundly influenced the way I approach theoretical research and develop mathematical arguments.

Thank you, Professor Cong Ma, for hosting me as a reading student and for training me in the fundamental skills of how to read, understand, decompose, and present complex proofs. Your guidance strengthened my mathematical maturity and taught me how to approach technical ideas with greater clarity and structure.

Thank you, Professor Yuxin Chen, for agreeing to serve on my committee and for asking thoughtful and insightful questions during both my proposal and defense. Your questions consistently provided unique perspectives and encouraged me to think about my work from broader and deeper angles.

Third, I would like to express my special thanks to my amazing collaborators, SigmaLab members, and friends: Zifeng Zhao, Ermin Wei, Fan Yao, Ganghua Wang, Weiyi Tian, Qia Wang, Bingji Yi, Qiyuan Liu, and Xuefeng Liu.

Thank you all for your support through both the highs and the lows. I feel incredibly fortunate and honored to have had the opportunity to collaborate with such talented and inspiring people to produce meaningful and high-quality research together. Beyond academic collaboration, your encouragement, friendship, and companionship made this journey far less lonely and much more joyful. I will always cherish the memories of working, learning, and growing alongside all of you.

I would also like to express my sincere gratitude to my parents, Huiqin Nie and Baoping Cheng. Thank you for always supporting me and for bearing with my bad temper, especially during the times when I was stuck in the middle of difficult proofs or when my interviews did not go well. Thank you for all your care and companionship — for cooking delicious meals, for calling me countless times to remind me not to stay up too late, and for staying by my side whenever I felt sad or discouraged. Your unconditional love and support have carried me through many difficult moments in my life. I feel incredibly fortunate and grateful to be

your daughter, and your love will always remain with me wherever I go.

Last but not least, I want to thank myself. Thank you for never giving up. No matter how many times you failed, you always tried to find a way forward through those failures. The resilience forged during my PhD journey — enduring hardship after hardship while remaining steadfast — will give me the courage to face whatever storms may come in the future. As the saying goes, "Plum blossoms bloom before chrysanthemums, and there is no need for comparison; just as in life, everyone has their own season." I hope that in the years ahead, I will continue to find myself, stay true to myself, and live a life filled with peace, stability, and joy.

When the mind is pure, lotus flowers bloom everywhere.

ABSTRACT

This dissertation studies the problem of aligning artificial intelligence systems with human preferences through three fundamentally connected dimensions: personalization, robustification, and evaluation. Modern AI systems increasingly operate in interactive environments where they must adapt to diverse users, learn from imperfect feedback, and continuously assess the quality and alignment of their own behaviors. These challenges are intrinsically sequential: effective personalization requires learning from human interactions; such interaction-driven systems must remain reliable under noisy or adversarial feedback, motivating robustification; and understanding whether these systems truly align with human preferences ultimately requires principled evaluation frameworks. Addressing this pipeline demands both rigorous modeling and learning algorithms with strong theoretical guarantees.

First, from the perspective of personalization, this dissertation develops personalized decision-making frameworks based on contextual reinforcement learning, with applications such as auto-bidding systems in digital advertising. A central challenge in these environments is learning from delayed and cumulative user responses while adapting to heterogeneous preferences across individuals and contexts. We propose learning algorithms that effectively balance exploration and exploitation under delayed feedback and state-dependent effects, achieving near-optimal theoretical guarantees and strong empirical performance. These results demonstrate how AI systems can adapt their behaviors to individual users in complex interactive environments.

However, personalization necessarily relies on feedback collected from human interactions, which is often noisy, inconsistent, strategic, or even adversarial. This motivates the second part of the dissertation on robustification. We investigate learning under imperfect and corrupted human feedback and establish an intrinsic efficiency–robustness trade-off showing that algorithms with faster statistical efficiency are generally more vulnerable to adversarial or corrupted observations. This vulnerability becomes even more pronounced in

multi-agent systems, where the manipulation of a single agent can substantially alter global interaction dynamics and collective behavior. Our results reveal how localized attacks may propagate system-wide and motivate the design of defense-aware and corruption-resilient learning algorithms.

Finally, once AI systems become adaptive and robust, a fundamental question remains: how should we evaluate whether their behaviors are truly informative, diverse, and aligned with human preferences? To address this challenge, the final part of this dissertation studies information conveyance in the generation space of large language models and other generative AI systems. We introduce principled information-theoretic measures and statistically consistent estimators, grounded in a tree-based view of language generation, to characterize how uncertainty and semantic diversity are distributed across model outputs. Our analysis provides new insights into how fine-tuning reshapes generation uncertainty into more informative, semantically meaningful, and human-aligned behaviors, rather than merely reducing diversity. Together, these contributions establish a unified perspective on aligning AI systems with human preferences through adaptive personalization, robust learning under imperfect feedback, and principled evaluation of generative behaviors.

CHAPTER 1

INTRODUCTION

Artificial intelligence systems are increasingly deployed in high-stakes, interactive environments where decisions are shaped by human feedback, strategic interactions, and long-term consequences. From recommendation systems and online advertising to multi-agent platforms and human-in-the-loop AI alignment, modern learning systems must adapt to diverse users while operating under imperfect observations, delayed outcomes, and strategic or adversarial behavior. In such environments, the central challenge is no longer merely optimizing prediction accuracy, but aligning artificial intelligence systems with human preferences in realistic and dynamic settings. Achieving this goal requires a principled understanding of how learning algorithms behave when classical assumptions—such as clean feedback, independent observations, and immediate rewards—are violated.

To address these challenges, this dissertation studies the design and analysis of learning algorithms in complex interactive environments through three fundamentally connected perspectives: personalization, robustification, and evaluation. These directions follow an intrinsic logical progression. First, AI systems must learn to personalize decisions and adapt to heterogeneous users in sequential environments. However, personalization necessarily depends on learning from human interactions, where feedback is often noisy, inconsistent, strategic, or adversarial, motivating the need for robustification. Finally, once systems become adaptive and robust, a fundamental question remains: how can we evaluate whether their behaviors are truly informative, reliable, and aligned with human preferences? This dissertation develops theoretical foundations and algorithmic tools across all three stages, contributing toward the broader goal of building reliable, adaptive, and trustworthy AI systems.

1.1 Personalized Decision-Making under Delayed Rewards

We begin with personalization, a defining feature of modern AI systems that interact continuously with heterogeneous users. In applications such as online advertising, recommendation systems, and adaptive decision-making platforms, effective learning requires understanding how actions influence users over time. These effects are often delayed, cumulative, and user-specific. For example, an advertisement may not trigger an immediate conversion but may gradually influence future user behavior through repeated exposure, reinforcement, or fatigue.

To model these challenges, we formulate ad bidding as a contextual Markov decision process (CMDP) with delayed Poisson rewards. This framework simultaneously captures delayed and long-term effects of actions, cumulative impacts such as reinforcement or fatigue, and heterogeneity across users.

To tackle the resulting estimation and decision-making difficulties, we propose a two-stage maximum likelihood estimation framework combined with data-splitting techniques that control error propagation across stages. Building on this framework, we design reinforcement learning algorithms with near-optimal regret guarantees of $\widetilde{\mathcal{O}}(dH^2\sqrt{T})$, where d is the contextual dimension, H is the horizon, and T is the number of users. These results demonstrate that explicitly modeling delayed, cumulative, and heterogeneous treatment effects can substantially improve personalization quality in interactive learning systems.

1.2 Learning from Imperfect Human Feedback

While personalization enables AI systems to adapt to users, such adaptation fundamentally relies on human feedback. In practice, however, human responses are rarely clean or stationary. Feedback may be noisy, inconsistent, strategic, or evolve over time due to cognitive biases, limited information, or learning effects. Ignoring these imperfections can lead to

misleading conclusions and unstable learning dynamics.

To address this challenge, we study the framework of *Learning from Imperfect Human Feedback* (LIHF), modeling the problem as a continuous-action dueling bandit under structured corruption whose magnitude decays over time. This captures the realistic phenomenon that humans gradually improve and provide more reliable feedback through repeated interactions.

Within this setting, we establish a regret lower bound of $\Omega\big(\max\{\sqrt{T}, T^{\rho}\}\big)$, revealing the intrinsic difficulty of learning under imperfect feedback. To match this lower bound, we develop the *Robustified Stochastic Mirror Descent for Imperfect Dueling* (RoSMID) algorithm, which achieves near-optimal regret $\widetilde{\mathcal{O}}\big(\max\{\sqrt{T}, T^{\rho}\}\big)$. Beyond this specific algorithm, we introduce a broader analytical framework for studying gradient-based methods under corruption, enabling applications across a wide range of preference-learning and interactive optimization problems.

1.3 Robustness in Multi-Agent Learning

The challenge of imperfect feedback becomes even more severe in multi-agent systems, where multiple learning agents interact strategically and influence one another's behavior. Such environments naturally arise in markets, auctions, recommendation platforms, and distributed decision-making systems. In these settings, local perturbations may propagate through the interaction dynamics and produce large-scale system-wide effects.

A central question we study is: how robust are multi-agent learning algorithms to adversarial manipulation? Surprisingly, we show that even minimal adversarial power can have dramatic consequences. In particular, poisoning the observations of a single agent is sufficient to destabilize the entire system and drive the learning dynamics toward outcomes far from the desired Nash equilibrium.

Our analysis reveals a fundamental *efficiency–robustness trade-off* in multi-agent learn-

ing. Algorithms with faster convergence rates are inherently more vulnerable to adversarial perturbations and can be manipulated using only a sublinear corruption budget. Conversely, algorithms with stronger robustness guarantees must necessarily sacrifice efficiency. These results expose fundamental limitations of existing learning dynamics and provide guidance for designing defense-aware algorithms that balance speed, robustness, and resilience in strategic environments.

1.4 Information Conveyance in Fine-Tuned Language Models

Once AI systems become adaptive and robust, a final challenge remains: how should we evaluate whether these systems are truly aligned with human preferences? Existing evaluation frameworks for large language models often focus on task accuracy or token-level uncertainty while overlooking how information is distributed across the entire generation process. In particular, output length acts as a major confounding factor, making it difficult to distinguish whether fine-tuning genuinely improves information conveyance or merely reduces diversity.

To address this issue, we introduce *Canopy Entropy* ($\mathrm{CE}^\star$), an information-theoretic measure that views language generation from a tree-based perspective, where the "canopy" represents the space of all possible rollouts. Under this view, $\mathrm{CE}^\star$ naturally characterizes the effective size and structure of the generation space. Unlike token-level uncertainty measures, it jointly captures uncertainty in both the output length N and generated sequence $Y_{1:N}$ through the total Shannon entropy $H(N, Y_{1:N} \mid X)$, where X denotes the prompt.

This formulation further yields interpretable quantities such as the length–uncertainty correlation $\rho(N, r_N)$, where r_N denotes the entropy rate. This correlation measures information conveyance efficiency by characterizing whether longer outputs remain informative on a per-token basis. Empirically, across tasks and model families, we find that fine-tuned models consistently exhibit stronger positive correlation $\rho(N, r_N)$, even when total entropy decreases. Moreover, after controlling for task, prompt, model family, and output length,

we find that fine-tuning substantially strengthens the relationship between entropy rate and semantic diversity. These findings suggest that aligned models do not simply reduce uncertainty; rather, they reorganize uncertainty into more informative, semantically meaningful, and human-aligned generations.

1.5 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 studies personalized decision-making under delayed rewards and develops contextual reinforcement learning algorithms for heterogeneous users. Chapter 3 introduces the framework of learning from imperfect human feedback and presents the RoSMID algorithm with near-optimal guarantees. Chapter 4 studies robustness in multi-agent learning and establishes the efficiency–robustness trade-off under adversarial poisoning. Chapter 5 develops information-theoretic tools for evaluating information conveyance in fine-tuned language models through the lens of Canopy Entropy. Finally, Chapter 6 concludes with broader implications and future research directions.

In summary, this dissertation develops a unified perspective on aligning AI systems with human preferences through adaptive personalization, robust learning under imperfect feedback, and principled evaluation of generative behaviors. By studying learning under delayed rewards, corrupted observations, strategic interactions, and uncertainty in generation, this work advances the theoretical and algorithmic foundations of reliable, adaptive, and human-centered artificial intelligence systems.

CHAPTER 2

LEARNING PERSONALIZED AD IMPACT VIA CONTEXTUAL REINFORCEMENT LEARNING UNDER DELAYED REWARDS

2.1 Introduction

E-commerce is expanding rapidly worldwide, with online sales expected to constitute 23% of total retail by 2027, supported by a 14.4% annual growth [ITA, 2025]. This growth has made digital advertising inevitable, empowering tech giants like Google, Meta, Microsoft, and Amazon, which leverage automated auctions to connect advertisers with customers. Auto-bidding, where platforms handle bid placement on behalf of advertisers, has grown significantly because of its simplified interaction for the advertisers and improved performance thanks to real-time optimization [Aggarwal et al., 2024, Google Ads Support, 2025, Facebook Ads, 2025, Microsoft Ads, 2025, Amazon Ads, 2025]. Given its importance, it is crucial to develop effective ad bidding strategies.

Developing effective bidding strategies requires accurately understanding advertising impacts. Psychological studies have long shown that advertising has a delayed and long-term impact on consumer beliefs and attitudes, ultimately shaping purchasing behavior over time [Vakratsas and Ambler, 1999, Lewis and Wong, 2022, Sakalauskas and Kriksciuniene, 2024]. Advertising effectiveness varies significantly across individuals. Observational studies reveal substantial heterogeneity based on demographics, platform usage patterns, and census data [Liu et al., 2019, Gordon et al., 2019]. These insights recommend personalized e-commerce advertising, suggesting platforms should leverage customer data to target high-value users and optimize bidding strategies tailored to diverse behavioral profiles [Sakalauskas and Kriksciuniene, 2024]. Additionally, the impact of ads also depends on their cumulative effect: while repeated exposure can strengthen brand recognition, it may also lead to ad fatigue—highlighting a subtle trade-off that is often overlooked in "learning to bid" literature

[Pechmann and Stewart, 1988, Lane, 2000, You et al., 2015, Bell et al., 2022, Guo and Jiang, 2024].

Limitation in recent work. However, these insights are not fully reflected in current algorithmic designs. The problem of "learning to bid" has been widely studied, with most prior work modeling it as a (contextual) bandit problem that assumes immediate rewards, such as click-through rates, which prioritizes short-term customer engagement [Weed et al., 2016, De Haan et al., 2016, Ren et al., 2017, Feng et al., 2018, 2023, Han et al., 2024a, Zhang and Luo, 2024b]. This approach, however, neglects the delayed and cumulative effects of advertising on consumption, potentially leading to incentive misalignment [Deng et al., 2022]. This misalignment is further exacerbated by the rise of auto-bidding systems, where platforms automatically manage bidding decisions on behalf of advertisers. While platforms typically optimize for engagement-based metrics, advertisers ultimately care about long-term revenue growth via production conversion [1]. Recent work has therefore begun to emphasize metrics like target return on ad spend as a practical alternative [Wang et al., 2019, Aggarwal et al., 2024] and design bidding strategies which account for delayed and cumulative effects of ads. Recently, Badanidiyuru et al. [2023] models the long-term causal impact of ad impressions using Markov Decision Process with mixed and delayed Poisson rewards. However, this approach assumes homogeneous treatment effects across users, overlooking the importance of personalization, a crucial factor in advertising effectiveness. To the best of our knowledge, no theoretical work jointly addresses all three aspects—delayed effects, cumulative impacts, and heterogeneity—in modeling ad effectiveness and designing bidding algorithms, potentially due to the modeling complexity and difficulty in estimation.

Our contribution. To address this gap, motivated by the initial proposal of Contextual Markov Decision Process (CMDP) [Hallak et al., 2015] to model customer behavior during website interactions, we model auto-bidding as CMDP with delayed Poisson rewards—using

1. We defer a more detailed discussion of the "learning-to-bid" literature to Appendix A.1.1

context to capture personalized ad impacts and states to capture ads cumulative effects. For effective estimation, rather than fitting all model parameters simultaneously using a single giant likelihood function, we introduce a novel data-splitting strategy and develop a two-stage maximum likelihood estimator that ensures controlled estimation error in the presence of delayed impacts. Based on this efficient online estimation oracle, we design a reinforcement learning algorithm to solve the CMDP with near-optimal regret of $\tilde{\mathcal{O}}(dH^2\sqrt{T})$, where d is the contextual dimension, H is the number of rounds, and T is the number of customers. Finally, we perform simulation studies which validate our theoretical findings.

2.2 Modeling Personalized Ad Impact by CMDP with Delayed Rewards

Notation. For a positive integer T, we denote $[T] = \{1, 2, \ldots, T\}$. We use standard asymptotic notations, including $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$, as well as their counterparts $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Omega}(\cdot)$, and $\tilde{\Theta}(\cdot)$ to suppress logarithmic factors. The symbol e represents the base of the natural logarithm. The Mahalanobis norm is defined as $\|\boldsymbol{x}\|_\Sigma = \sqrt{\boldsymbol{x}^\top \Sigma \boldsymbol{x}}$. $\|\cdot\|_2$ represents L_2 norm. For vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we use $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and $\boldsymbol{x}^\top \boldsymbol{y}$ interchangeably to denote their inner product.

To address the complexities of advertisement impacts on product conversions, we model online ad bidding as a Contextual Markov Decision Process (CMDP). This framework captures the three key impacts discussed in Section 2.1 and allows transitions and rewards to depend on context $\mathbf{x}_t$, personalized information of customer t, supporting dynamic and personalized bidding. While incorporating $\boldsymbol{x}_t$ directly into the state is possible, it greatly enlarges the state space and complicates learning [Levy and Mansour, 2023]. Instead, we adopt a CMDP formulation—common in user-driven applications—that keeps the state compact and treats context as auxiliary information [Hallak et al., 2015], preserving both efficiency and personalization. Our analysis focuses on ad platforms that bid on behalf of advertisers.

We use "ad platform" and "learner" interchangeably.

Mathematically, a CMDP is defined as the tuple $(\mathcal{X}, \mathcal{S}, \mathcal{A}, \mathcal{M})$, where $\mathcal{X} \subseteq \mathbb{R}^d$ represents the contextual feature space, $\mathcal{S}$ denotes the state space capturing customer ad exposure history, and $\mathcal{A}$ is the action space for ad bidding strategies. The mapping $\mathcal{M}$ assigns each context $\mathbf{x}$ to a Markov Decision Process (MDP), $\mathcal{M}(\mathbf{x}) = (\mathcal{S}, \mathcal{A}, \mathcal{P}^{\mathbf{x}}, \mathcal{R}^{\mathbf{x}}, S_1, H)$, where the state transition probability $\mathcal{P}^{\mathbf{x}}$ and reward function $\mathcal{R}^{\mathbf{x}}$ depend on $\mathbf{x}$. S_1 is the initial state. H represents the maximum number of customer-learner interactions with details in the context of online bidding as below.

Definition 2.2.1 (State)**.** *The state of customer t at round h, denoted as $S_h^t = [S_{h,1}^t, S_{h,2}^t]$, encodes information about the two most recent ad exposures. Specifically, let $G_{h,1}^t$ represent the round of the most recent ad impression, we set $S_{h,1}^t = h - G_{h,1}^t$, which captures the time elapsed since that impression. Similarly, let $G_{h,2}^t$ denote the round of the second-to-last ad impression, we set $S_{h,2}^t = G_{h,1}^t - G_{h,2}^t$, representing the time interval between the two most recent impressions. $S_{h,1}^t \in \mathcal{H}_1 = \{\infty, 1, 2, \ldots, H-1\}$ and $S_{h,2}^t \in \mathcal{H}_2 = \{-\infty, 1, 2, \ldots, H-2\}$. The interpretation of ∞ and $-\infty$ is deferred to Remark 2.2.1.*

Assumption 1 (Observation)**.** *The expected product conversion rate μ_h^t at round h for customer t with context $\boldsymbol{x}_t$ follows*

$$\mu_h^t = \begin{cases} d_{S_{h,1}^t} \boldsymbol{x}_t^\top \boldsymbol{\theta}_{S_{h,2}^t} & \text{if } \boldsymbol{o}_h^t = 0 \\ \boldsymbol{x}_t^\top \boldsymbol{\theta}_{S_{h,1}^t} & \text{otherwise.} \end{cases}$$

The observed product conversion $\boldsymbol{y}_h^t$ follows a Poisson distribution, i.e. $\boldsymbol{y}_h^t \sim Poi(\mu_h^t)$. $d_{S_{h,1}^t}$ represents the delayed advertisements impact, $S_{h,1}^t \in \mathcal{H}_1 = \{\infty, 1, 2, \ldots, H-1\}$. $\boldsymbol{o}_h^t$ indicates the bidding outcome for customer t at round h. $\boldsymbol{o}_h^t = 1$ if the bid is won and $\boldsymbol{o}_h^t = 0$ otherwise.

Remark 2.2.1. *If customer t has no prior ad exposure, the state is $S_h^t = [\infty, -\infty]$. Specif-*

ically, when no advertisement has been successfully displayed, product conversion $\boldsymbol{y}_h^t \sim Poi(\mu_h^t)$*, where* $\mu_h^t = d_\infty \boldsymbol{x}_t^\top \boldsymbol{\theta}_{-\infty}$*, Here,* $\boldsymbol{x}_t^\top \boldsymbol{\theta}_{-\infty}$ *represents the natural demand in the absence of ads, which may vary with context* $\boldsymbol{x}_t$ *to capture factors such as income, preferences, tastes, and seasonality [Manandhar, 2018]. We set* $d_\infty = 1$ *to avoid identifiability issue. If the learner wins the bid,* $\boldsymbol{y}_h^t \sim Poi(\mu_h^t)$ *with* $\mu_h^t = \mathbf{x}_t^\top \boldsymbol{\theta}_\infty$*, where* $\boldsymbol{\theta}_\infty$ *captures the effect of first-time ad exposure. It is possible to generalize the linear modeling of the expected product conversion rate to a more flexible form, i.e.,* $\mu_h^t = h_{S_{h,1}^t}(x_t)$*. For example,* $h_{S_{h,1}^t}$ *may be modeled as a state-dependent neural network, though this would come at the cost of sacrificing theoretical guarantees. We refer readers to Appendix A.1.2 for a detailed discussion.*

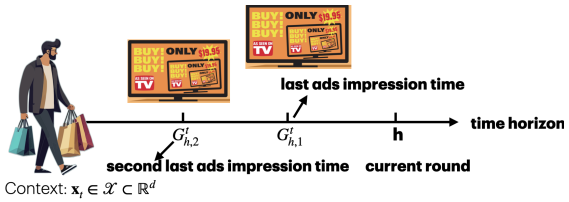

Figure 2.1: Illustration

As illustrated in Figure 2.1, if no new advertisement is displayed at round h, the impact of the previous ad shown at round $G_{h,1}^t$ carries over, with a delayed and long-term effect governed by $d_{h-G_{h,1}^t}$, depending on the time interval $h - G_{h,1}^t$. This delayed factor allows ad effects to span multiple rounds, enabling delayed conversion peaks, as d_l is not restricted to be less than 1. This results in an expected product conversion of $d_{h-G_{h,1}^t} \boldsymbol{x}_t^\top \boldsymbol{\theta}_{G_{h,1}^t - G_{h,2}^t}$ (Assumption 1), where $\boldsymbol{\theta}_{G_{h,1}^t - G_{h,2}^t}$ captures the impact of the ad shown at $G_{h,1}^t$, which affected by its most recent predecessor at $G_{h,2}^t$. In contrast, if a new ad is successfully displayed at round h, its effect overrides the previous one, but itself is influenced by the most recent prior impression at round $G_{h,1}^t$. The resulting impact is modeled by $\boldsymbol{\theta}_{h-G_{h,1}^t}$, leading to an expected conversion of $\boldsymbol{x}_t^\top \boldsymbol{\theta}_{h-G_{h,1}^t}$.

A natural question arises: why does the cumulative advertising impact depend only on the most recent display, rather than the entire history of ad exposures? This modeling choice is motivated by the recency effect, a well-documented cognitive bias in psychology and marketing [Murphy et al., 2006, Chatfield, 2016, Phillips-Wren et al., 2019]. It suggests that individuals tend to place greater weight on recent experiences when making decisions. This

phenomenon is especially relevant in advertising, where exposure timing significantly affects consumer behavior [Chatfield, 2016], and underlies practical approaches like Google's attribution models [Google Ads Attribution, 2025] which uses last-touch heuristics that prioritize recent ad exposures. In fact, our state formulation is very flexible and readily accommodates an enlarged state space. It supports a broad class of CMDP formulations for $\boldsymbol{\theta}_S$, where S encodes domain knowledge about ad effects, and remains compatible with our proposed learning algorithm. For example, if we believe that all past ad exposures contribute to customer behavior, we can easily extend the state to include $S_h^t = [S_{h,1}^t, S_{h,2}^t, n_h^t]$, where n_h^t is the total number of successful ad displays in the past. We refer readers to Appendix A.1.3 for a detailed discussion of the flexibility of our state formulation.

The bidding outcome $\boldsymbol{o}_h^t$ depends on both learner's bid $\mathbf{a}_h^t$ and the competitors' bids. The bidding space is given by $\mathbf{a}_h^t \in [0, \boldsymbol{B}_{\mathcal{A}}]$, where $\boldsymbol{B}_{\mathcal{A}}$ is the maximum allowable bid. $\mathbf{a}_h^t = 0$ indicates that the learner opts out of the auction. At each round, the learner submits $\mathbf{a}_h^t$ and wins if its bid exceeds all competitors' bid. The probability of winning by $\mathbf{a}_h^t$ is shown below.

Definition 2.2.2 (Transition)**.** *For a given bid amount $\mathbf{a}_h^t$, the probability of winning the auction, denoted as $\mathbb{P}(\boldsymbol{o}_h^t = 1)$, is modeled by $\mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t)$, where $\mathcal{F}_h : [0, \boldsymbol{B}_{\mathcal{A}}] \times \mathcal{X} \to [0, 1]$ represents the cumulative distribution function (CDF) of the Highest Other Bids (HOB).*

Remark 2.2.2. *Given the current state $S_h^t = [S_{h,1}^t, S_{h,2}^t]$ and bid $\mathbf{a}_h^t$, the next state transitions according to $\mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t)$. With probability $\mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t)$, the bid is successful, and the next state is $S_{h+1}^t = [1, S_{h,1}^t]$. Otherwise, the bid is lost, and the state transitions to $S_{h+1}^t = [S_{h,1}^t + 1, S_{h,2}^t]$*[2].

This CDF, $\mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t)$, of HOB captures the competitive bidding environment by modeling both context $\boldsymbol{x}_t$ and time h, allowing for round-to-round variation from changing

2. Due to the special meaning of ∞ and $-\infty$, we define $h - \infty = \infty$, and $h - \infty - (-\infty) = \infty$.

competitors and context-driven bid adjustments. In line with standard practice, we consider a full information feedback setting, where the realized HOB m_h^t is observed by the learner regardless of the bidding outcome [Cesa-Bianchi et al., 2014, Feng et al., 2018, Zhang et al., 2022, Badanidiyuru et al., 2023]. This setting is practical since bidders can always access the "minimum-bid-to-win" feedback [Google Developers, 2024]. Also, ad platforms inherently know realized bids for their auto-bidding algorithms. We assume m_h^t follows a lognormal distribution, a common modeling choice in the literature for advertisement auctions [Laffont et al., 1995, Wilson, 1998, Smith et al., 2003, Skitmore, 2008, Ballesteros-Pérez and Skitmore, 2017] (Assumption 2), resulting the probability of winning with a bid $\mathbf{a}_h^t$ as $\mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t) = \Phi((\log(\mathbf{a}_h^t) - \langle \boldsymbol{x}_t, \boldsymbol{\beta}_h \rangle)/\sigma_h)$, where Φ denotes the CDF of the standard normal distribution, $\boldsymbol{\beta}_h \in \mathbb{R}^d$ represents the highest willingness to pay for displaying ads from other competitors, capturing the competitiveness of the environment, while σ_h reflects the associated variability.

Assumption 2 (Lognormal Distribution of HOB)**.** $\log(m_h^t) \sim \mathcal{N}(\langle \boldsymbol{x}_t, \boldsymbol{\beta}_h \rangle, \sigma_h^2)$.

In a second-price auction, the format we study, the winning bidder pays the second-highest bid, with payment given by $p_h(\mathbf{a}_h^t, \boldsymbol{x}_t) = \mathbf{a}_h^t - \frac{1}{\mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t)} \int_0^{\mathbf{a}_h^t} \mathcal{F}_h(v, \boldsymbol{x}_t) dv$. This format is standard in both industry and research [Cooper and Fang, 2008, Weed et al., 2016, Zhao and Chen, 2019], and our analysis extends to other single-item auctions without entry fees, such as first-price auctions. To bid effectively, the learner must balance the probability of winning, the incurred payment, and the expected product conversion. Without loss of generality, we normalize the value of each unit of product conversion to $\nu = 1$, leading to the following definition of the expected reward.

$$R_h^t\left(S_h^t, \mathbf{a}_h^t, \boldsymbol{x}_t\right) = d_{S_{h,1}^t} \langle \boldsymbol{\theta}_{S_{h,2}^t}, \boldsymbol{x}_t \rangle (1 - \mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t)) + (\langle \boldsymbol{\theta}_{S_{h,1}^t} \boldsymbol{x}_t \rangle - p_h(\mathbf{a}_h^t, \boldsymbol{x}_t)) \mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t).$$

By these formulation, in the context of online bidding,

the tuple $(\mathcal{X},\ \mathcal{S},\ \mathcal{A},\ \mathcal{P}^{\boldsymbol{x}_t}, \{R_h^t\left(S_h^t, \mathbf{a}_h^t, \boldsymbol{x}_t\right)\}_{h=1}^H, S_1^t, H)$ is a CMDP (Fact 2.2.1, Appendix A.4.1). $\mathcal{P}^{\mathbf{x}_t}$ is the joint probability distribution of states S_h^t, induced by $\{\mathcal{F}_h(\cdot, \boldsymbol{x}_t)\}_{h\in[H]}$. We direct reader to Appendix A.1.4 for an illustration example for modeling online bidding as a CMDP.

Fact 2.2.1. *$(\mathcal{X}, \mathcal{S}, \mathcal{A}, \mathcal{P}^{\boldsymbol{x}_t}, \{R_h^t\left(S_h^t, \mathbf{a}_h^t, \boldsymbol{x}_t\right)\}_{h=1}^H, S_1^t, H)$ is a CMDP.*

2.2.1 Learning goal: regret minimization

The learner interacts with T customers over H rounds, receiving context information $\boldsymbol{x}_t$ for customer t. The learner begins without prior knowledge of the product conversion parameters $\Theta = \{\boldsymbol{\theta}_l\}_{l\in\mathcal{H}} \cup \{d_l\}_{l\in\mathcal{H}_1}$[3] nor the transition $\mathcal{F}^{\boldsymbol{x}_t} := \{\mathcal{F}_h(\cdot, \boldsymbol{x}_t)\}_{h=1}^H$. We assume Θ and transition parameters $\{\boldsymbol{\beta}_h\}_{h\in[H]} \cup \{\sigma_h\}_{h\in[H]}$ are bounded. The context space $\mathcal{X}$ is also bounded, with no distributional assumptions on the arrival of $\mathbf{x}_t$. $\boldsymbol{b}$ is a small positive constant, which ensures that displaying an ad always yields a non-zero (expected) purchase quantity (Assumption 3).

Assumption 3. *There exists positive constants $\boldsymbol{b}, \boldsymbol{B}_x,\ \boldsymbol{B}_\theta, \boldsymbol{B}_d, \boldsymbol{B}_{\boldsymbol{\beta}}, \bar{\sigma}$ such that $\|\boldsymbol{x}_t\|_2 \leq \boldsymbol{B}_x$, and $\boldsymbol{\theta}_l^\top \boldsymbol{x}_t \geq \boldsymbol{b}$, $\forall t \in [T], \forall l \in \mathcal{H}$, $\|\boldsymbol{\theta}_l\|_2 \leq \boldsymbol{B}_\theta, \forall l \in \mathcal{H}$, $d_l \in [0, \boldsymbol{B}_d], \forall l \in \mathcal{H}_1$, and $\|\boldsymbol{\beta}_h\|_2 \leq \boldsymbol{B}_{\boldsymbol{\beta}}, \sigma_h \leq \bar{\sigma}, \forall h \in [H]$.*

The learner's objective is to minimize the cumulative regret over T customers with regret defined as:

$$\mathrm{Reg}_T := \sum_{t=1}^{T} \mathrm{OPT}\left(\Theta, \mathbf{x}_t, \mathcal{F}^{\mathbf{x}_t}\right) - \mathbb{E}_{\pi_1,\ldots,\pi_T\sim\mathcal{G}}\left[\sum_{t=1}^{T} R\left(\pi_t; \mathbf{x}_t, \Theta, \mathcal{F}^{\mathbf{x}_t}\right)\right]. \tag{2.1}$$

The expectation is taken over the policy $\pi_t : \mathcal{S} \times \mathcal{X} \to [0, \boldsymbol{B}_{\mathcal{A}}]$ generated by the algorithm $\mathcal{G}$ employed by the learner. We define $\mathrm{OPT}(\Theta, \mathbf{x}_t, \mathcal{F}^{\mathbf{x}_t})$ as the optimal expected

3. $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$

utility achievable for a customer with contextual features $\mathbf{x}_t$ over H rounds, under the true parameters Θ and the distribution $\mathcal{F}^{\mathbf{x}_t}$. The reward collected by the policy π_t over H rounds for customer t, denoted as $R(\pi_t; \mathbf{x}_t, \Theta, \mathcal{F}^{\mathbf{x}_t})$, denoted by $R\left(\pi_t; \mathbf{x}_t, \Theta, \mathcal{F}^{\mathbf{x}_t}\right) = \mathbb{E}_{S_h^t \sim \mathcal{P}^{\mathbf{x}_t}}[\sum_{h=1}^{H} R_h^t\left(S_h^t, \pi_t(S_h^t, \boldsymbol{x}_t), \mathbf{x}_t\right)]$.

2.3 Algorithm Design

In this section, we introduce design principles for solving CMDPs with delayed Poisson rewards. The proposed algorithm (Algorithm 1) consists of three main stages: exploration, exploitation, and estimation, with the estimation stage playing a central role.

The estimation stage contains three key components. First, ridge regression [Abbasi-Yadkori et al., 2011] combined with two-stage variance estimators is used to estimate the transition dynamics $\mathcal{F}^{\boldsymbol{x}_t}$, handling potential adversarial arrivals of $\boldsymbol{x}_t$. Second, a variant of online Newton estimator [Xue et al., 2024] is employed to estimate the instant advertisement effects $\{\boldsymbol{\theta}_l\}_{l\in\mathcal{H}}$. Third, a novel two-stage maximum likelihood estimator (TS-MLE) is developed to estimated the delayed impacts $\{d_l\}_{l\in\mathcal{H}_1}$.

The key challenge in online estimation is that a naive approach—simultaneously estimating and updating all parameters by maximizing a joint log-likelihood $\mathcal{L}(\Theta)$—is infeasible due to two main issues. First, the log-likelihood function $\mathcal{L}(\Theta)$ is non-concave in Θ, making it difficult to ensure a unique solution. Second, the score equations $\nabla_\Theta \mathcal{L}(\Theta) = 0$ lack closed-form solutions, rendering direct analysis intractable. These challenges motivate the development of the estimator, TS-MLE, which efficiently estimates delayed effects while maintaining computational tractability.

The core idea of the TS-MLE is to divide the estimation into manageable steps. Intuitively, if the estimation error, $\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\|_2$, is small, $\hat{\boldsymbol{\theta}}_l$ can then be treated as a close approximation of the true parameter $\boldsymbol{\theta}_l$. Based on this approximation, a maximum likelihood estimator $\hat{d}_l$ can be constructed, ensuring that the estimation error for d_l remains

small. This approach is feasible because the conditional log-likelihood $\mathcal{L}(\{d_l\}_{l\in\mathcal{H}_1}|\{\hat{\boldsymbol{\theta}}_l\}_{l\in\mathcal{H}})$ is concave in $\{d_l\}_{l\in\mathcal{H}_1}$ (Eqn. (2.2)), and both its gradient $\nabla_{\{d_l\}_{l\in\mathcal{H}_1}}\mathcal{L}$ (Eqn. (A.3)) and the solution to the corresponding score equation $\nabla_{\{d_l\}_{l\in\mathcal{H}_1}}\mathcal{L}=0$ (Eqn. (2.3)) are straightforward to compute.

To control the estimation error of $\hat{d}_l$, it is crucial to ensure that this error depends only on the error in $\hat{\boldsymbol{\theta}}_l$, not vice versa. This prevents feedback loops between the estimation errors of $\hat{d}_l$ and $\hat{\boldsymbol{\theta}}_l$, which would otherwise complicate mathematical analysis and make the estimation process intractable. To achieve this, we employ a carefully designed data-splitting strategy, where separate data subsets are allocated exclusively for estimating $\boldsymbol{\theta}_l$ and d_l, ensuring their errors remain independent.

To achieve this separation, we introduce two datasets, $\mathbf{W}_{t,l}$ and $\mathbf{D}_{t,l}$, which share a common structure but serve distinct purposes (Def. 2.3.1). Both datasets focus on round h for customer t where the most recent bid win occurred l rounds before the current round ($S^t_{h,1}=l$). The key difference lies in the parameter being estimated. $\mathbf{W}_{t,l}$ includes rounds that observations $\{\boldsymbol{y}^t_h\}_{h\in\mathbf{W}_{t,l}}$ with mean $\langle\boldsymbol{\theta}_l,\boldsymbol{x}_t\rangle$. Thus this datasets is used to estimate θ_l, as they depend solely on $\boldsymbol{\theta}_l$. $\mathbf{D}_{t,l}$ contains rounds that observations $\{\boldsymbol{y}^t_h\}_{h\in\mathbf{D}_{t,l}}$ with mean $d_l\langle\boldsymbol{\theta}_{S^t_{h,2}},\boldsymbol{x}_t\rangle$. This datasets is used to estimate d_l because they depend solely on the unknown parameter d_l when $\{\hat{\boldsymbol{\theta}}_l\}_{l\in\mathcal{H}}$ are given. This ensures that the estimation error of d_l depends on $\boldsymbol{\theta}_l$, while the estimation error of $\boldsymbol{\theta}_l$ remains independent of d_l.

Definition 2.3.1. *For customer t, we define two datasets: $\mathbf{W}_{t,l}=\{h|S^t_{h,1}=l,\boldsymbol{o}^t_h=1\}$ and $\mathbf{D}_{t,l}=\{h|S^t_{h,1}=l,\boldsymbol{o}^t_h=0\}$, for $l\in\mathcal{H}\setminus[-\infty]$. We define $\mathbf{W}_{t,-\infty}=\{h|S^t_{t,1}=\infty,\boldsymbol{o}^t_h=0\}$ for the estimation of $\theta_{-\infty}$. The collection of observations across the first t customers are $\mathbf{W}^t_l=\{\mathbf{W}_{s,l}\}^t_{s=1}$ and $\mathbf{D}^t_l=\{\mathbf{D}_{s,l}\}^t_{s=1}$. The size of $\mathbf{D}^t_l$ is given by $N_{t,l}=|\mathbf{D}^t_l|$.*

Taken this together, the log-likelihood of d_l up to the first t customers as:

$$\mathcal{L}\left(\boldsymbol{y}, d_l; \hat{\boldsymbol{\theta}}_l^s\right) = \sum_{s=1}^{t} \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{y}_h^s \log\left(d_l \boldsymbol{x}_s^\top \hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s\right) - d_l \boldsymbol{x}_s^\top \hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s. \tag{2.2}$$

Differentiating the log-likelihood with respect to d_l, setting it to zero, and solving for d_l, results in

$$\hat{d}_l^t = \frac{\sum_{s=1}^{t} \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{y}_h^s}{\sum_{s=1}^{t} \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s, \boldsymbol{x}_s \rangle}. \tag{2.3}$$

In the online setting, $\hat{d}_l^t$ estimates d_l as of customer t, analogous to $\hat{\boldsymbol{\theta}}_l^t$. Instead of using the most recent estimates $\{\hat{\boldsymbol{\theta}}_{S_{h,2}^s}^t\}_{s=1}^t$ as the first-stage input for estimating $\hat{d}_l^t$, we leverage $\{\hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s\}_{s=1}^t$, a progressively updated estimate of the advertisement's impact across customer arrivals. This novel design is motivated by the observation that the estimation error of $\hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s$ instead of $\hat{\boldsymbol{\theta}}_{S_{h,2}^s}^t$ in the direction of $\boldsymbol{x}_s$ is well-controlled. Further details, including the confidence region $\mathcal{D}_l^t$ (line 2 of Algorithm 2), are provided in Theorem 2.4.1 and its proof.

Algorithm 1: Online Contextual RL with Delayed Poisson Reward

Input: $d, T, H, \boldsymbol{b}, \boldsymbol{B}_x, \boldsymbol{B}_\theta, \boldsymbol{B}_d, \boldsymbol{B}_\mathcal{A}$, δ, γ, Γ, $\underline{n_l}$

for $t = 1$ **to** T **do**

  Obtain the context $\boldsymbol{x}_t$ for the arriving customer t;

  if $t \le (H+1)\underline{n_l}$ **then**

    `// Exploration`

    Compute $l = \left\lfloor \frac{t}{\underline{n_l}} \right\rfloor$;

    if $l = H$ **then**

      Set $\mathbf{a}_h^t = 0,\ \forall h \in [H]$;

    else

      Set $\mathbf{a}_1^t = \mathbf{a}_{l+1}^t = \boldsymbol{B}_\mathcal{A}$, and $\mathbf{a}_h^t = 0$ for $\forall h \neq 1, l+1$;

    for $h = 1$ **to** H **do**

      Observe $\boldsymbol{y}_h^t$, m_h^t; Update $\hat{\boldsymbol{\beta}}_h^t$ by Eqn. (2.5) and $\hat{\sigma}_h^t$ by Eqn. (2.6);

  else

    `// Exploitation`

    Update $\pi_t = \arg\max_\pi \max_{\tilde{\theta} \in \mathcal{C}_{t-1}} R(\pi; \tilde{\theta}, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t})$;

    for $h = 1$ **to** H **do**

      Observe S_h^t; Take action $\mathbf{a}_h^t = \pi_t(S_h^t, \boldsymbol{x}_t)$; Observe $\boldsymbol{o}_h^t$, m_h^t, and $\boldsymbol{y}_h^t$; Update $\hat{\boldsymbol{\beta}}_h^t$ by Eqn. (2.5) and $\hat{\sigma}_h^t$ by Eqn. (2.6);

  `// Estimation`

  for $l \in \mathcal{H}$ **do**

    Update dataset $\mathbf{W}_l^t, \mathbf{D}_l^t$ by Def. 2.3.1;

    if $\mathbf{W}_{t,l}$ *is nonempty* **then**

      Compute $\hat{\boldsymbol{\theta}}_l^t$ and $\mathcal{C}_l^t$ by Algo. 4;

    else

      Set $\hat{\boldsymbol{\theta}}_l^t = \hat{\boldsymbol{\theta}}_l^{t-1}$ and $\mathcal{C}_l^t = \mathcal{C}_l^{t-1}$;

    if $\mathbf{D}_{t,l}$ *is nonempty* **then**

      Compute $\hat{d}_l^t$ and $\mathcal{D}_l^t$ by Algo. 2;

    else

      Set $\hat{d}_l^t = \hat{d}_l^{t-1}$ and $\mathcal{D}_l^t = \mathcal{D}_l^{t-1}$;

  Set $\mathcal{C}_t = \{\Theta \mid \{\boldsymbol{\theta}_l \in \mathcal{C}_l^t\} \cap \{d_l \in \mathcal{D}_l^t\},\ \forall l \in \mathcal{H}\}$;

Remark 2.3.1 (Key Tuning Parameters for Algorithm 3)**.** *The input parameters d, T, and H are structural components of the CMDP. The quantities $\boldsymbol{b}, \boldsymbol{B}_x, \boldsymbol{B}_\theta, \boldsymbol{B}_d, \boldsymbol{B}_\mathcal{A}$ are defined in Assumption 3. The tail probability δ determines the confidence region for $\hat{\boldsymbol{\beta}}_l$ and $\hat{d}_l$ and serves as an input for Algorithm 4 and Algorithm 2. The truncation threshold Γ, used to handle heavy-tailed distributions, is defined in Eq. (A.1) and appears in Algorithm 4. The parameter γ, defined in Lemma 2.3.1, serves as a weighted estimation error bound and is used in Algorithm 2. The exploration stage guarantees a minimum number of observations to ensure reliable estimation. Specifically, the threshold $\underline{n}_l$ is given by $\underline{n}_l := \lceil \frac{32\log(HT)}{e\boldsymbol{B}_d\boldsymbol{B}_x\boldsymbol{B}_\theta\boldsymbol{b}^2} \rceil$, which guarantees sufficient observations in $\mathbf{W}_l^t$ and $\mathbf{D}_l^t$, ensuring $|\mathbf{W}_l^t| \geq \underline{n}_l$ and $N_{t,l} = |\mathbf{D}_l^t| \geq \underline{n}_l$ for all $l \in \mathcal{H}$ in all subsequent episodes after the exploration stage.*

The estimation of $\hat{d}_l^t$ depends on accurately estimating $\hat{\boldsymbol{\theta}}_l^t$. Using observations $\boldsymbol{y}_h^t$ from $\mathbf{W}_l^t$, the problem reduces to efficiently estimating $\boldsymbol{\theta}_l$ from Poisson data. To achieve this, we adopt the Confidence Region with Truncated Mean approach (see Algorithm 4 and Appendix A.2), introduced by Xue et al. [2024]. This method utilizes a variant of the online Newton estimator, given by:

$$\hat{\boldsymbol{\theta}}_l^t = \arg\min_{\|\boldsymbol{\theta}\|_2 \leq \boldsymbol{B}_\theta} \{\tfrac{1}{2}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_l^{t-1}\|_{\mathbf{V}_l^t}^2 + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_l^{t-1})^\top \nabla \tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1})\}, \tag{2.4}$$

$\mathbf{V}_l^t = \mathbf{V}_l^{t-1} + \frac{1}{2}\boldsymbol{x}_s\boldsymbol{x}_s^\top$, with $\mathbf{V}_l^0$ initialized as the identity matrix $\mathbf{I}_d$. $\nabla \tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1}) = (-\tilde{\boldsymbol{y}}_h^t + (\mathbf{x}_t)^\top \hat{\boldsymbol{\theta}}_l^{t-1})\mathbf{x}_t$. The truncated observation $\tilde{\boldsymbol{y}}_h^t$ is defined as $\tilde{\boldsymbol{y}}_h^t = \boldsymbol{y}_h^t \mathbb{I}_{\|\mathbf{x}_t\|_{(\mathbf{V}_l^t)^{-1}}|\boldsymbol{y}_h^t| \leq \Gamma}$.

This truncation mitigates the impact of extreme values in $\boldsymbol{y}_h^t$ by setting outliers to zero, a technique originally designed for generalized linear bandit problems with heavy-tailed data.

By Lemma 2.3.1, the weighted estimation error $\|\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_l^t\|_{\mathcal{V}_l^t}^2$ remains well-controlled with high probability.

Lemma 2.3.1 (Theorem 1 in Xue et al. [2024])**.** *Given $l \in \mathcal{H}$, with probability at least $1-\delta$, $\hat{\boldsymbol{\theta}}_l^t$ defined in Eqn. (2.4) satisfies $\|\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_l^t\|_{\mathcal{V}_l^t}^2 \leq \gamma, \forall t \geq 0$, where $\gamma = 896d\boldsymbol{B}_x\boldsymbol{B}_\theta(1 +$*

$\boldsymbol{B}_x\boldsymbol{B}_\theta)\log\left(\frac{4T}{\delta}\right)\log\left(1+\frac{T}{2d}\right)+2\boldsymbol{B}_x^2\boldsymbol{B}_\theta^2+48d\boldsymbol{B}_x\boldsymbol{B}_\theta\log\left(1+\frac{T}{2d}\right)$.

To estimate the transition dynamics $\mathcal{F}_h$, we estimate $\boldsymbol{\beta}_h$ by Eqn. (2.5). Without loss of generality, we set $\lambda = 1$.

$$\hat{\boldsymbol{\beta}}_h^t = \left(\sum_{s=1}^{t}\boldsymbol{x}_s\boldsymbol{x}_s^\top + \lambda\mathbb{I}_d\right)^{-1}\left(\sum_{s=1}^{t}\boldsymbol{x}_s\log(m_h^s)\right). \tag{2.5}$$

The variability σ_h, similar to $\hat{d}_l^t$, is estimated based on the first-stage estimators $\{\hat{\boldsymbol{\beta}}_h^s\}_{s=1}^t$ (Eqn. (2.6)). We use $\{\hat{\boldsymbol{\beta}}_h^s\}_{s=1}^t$ instead of $\hat{\boldsymbol{\beta}}_h^t$ to control over estimation error in the direction of $\boldsymbol{x}_s$ (details in Appendix A.3.2).

$$\hat{\sigma}_h^t = \sqrt{\frac{1}{t}\sum_{s=1}^{t}\left(\log(m_h^s) - \boldsymbol{x}_s^\top\hat{\boldsymbol{\beta}}_h^s\right)^2}. \tag{2.6}$$

In addition to estimation, the exploration period aims to gather sufficient observations for d_l to ensure quadratic tail decay of Poisson (Remark 2.3.1).

After exploration, the algorithm enters exploitation, selecting actions via the greedy policy π_t to maximize rewards within the confidence region. π_t is computed via dynamic programming with time complexity $\mathrm{Poly}(H, |\mathcal{S}|, \boldsymbol{B}_\mathcal{A}/\epsilon)$ when discretizing the bidding space for an ϵ-optimal solution (detail in Appendix A.1.5). By balancing exploration and exploitation, Algorithm 3 achieves near-optimal performance, formally proven in the next section.

Algorithm 2: Two-Stage Maximum Likelihood Estimation

Input: datasets $\mathbf{D}_l^t, \{\boldsymbol{x}_s\}_{s=1}^t, \{\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}\}_{s=1}^t$; parameters $\boldsymbol{b}, \boldsymbol{B}_x, \boldsymbol{B}_\theta, \boldsymbol{B}_d, d, T, H, \delta, \gamma$

1 Update estimator $\hat{d}_l^t$ by Eqn. (2.3);

2 Compute

$$\mathcal{D}_l^t = \left\{d_l \in [0, \boldsymbol{B}_d] \mid |\hat{d}_l^t - d_l| \le \frac{4H\boldsymbol{B}_d\sqrt{d\log\left(1+\frac{T}{2d}\right)\gamma} + \sqrt{2e\boldsymbol{B}_d\boldsymbol{B}_x\boldsymbol{B}_\theta\log(2/\delta)}}{\boldsymbol{b}\sqrt{N_{t,l}}}\right\}$$

Output: $(\hat{d}_l^t, \mathcal{D}_l^t)$

2.4 Analysis of Near-Optimal Regret Bound

This section demonstrates the efficiency of the TS-MLE (Theorem 2.4.1) and analyzes the near-optimal performance of the proposed Algorithm 3 (Theorem 2.4.2).

Theorem 2.4.1 (Confidence Region for $\hat{d}_l^t$)**.** *Let $\delta \geq \frac{1}{T^4 H}$ and $N_{t,l} \geq \underline{n_l}$. With probability at least $1-\delta$, the estimation error $|\hat{d}_l^t - d_l|$, with $\hat{d}_l^t$ defined in Eqn. (2.3), is bounded by:*

$$\left|\hat{d}_l^t - d_l\right| \leq \frac{4H\boldsymbol{B}_d\sqrt{d\log\left(1+\frac{T}{2d}\right)\gamma} + \sqrt{2e\boldsymbol{B}_d\boldsymbol{B}_x\boldsymbol{B}_\theta\log\left(\frac{2}{\delta}\right)}}{\boldsymbol{b}\sqrt{N_{t,l}}}. \tag{2.7}$$

γ is as defined in Lemma 2.3.1 and $\underline{n_l}$ defined in Remark 2.3.1.

Theorem 2.4.1 shows that the estimation error of TS-MLE is bounded by $\tilde{\mathcal{O}}(1/\sqrt{N_{t,l}})$, where $N_{t,l}$ (Def. 2.3.1) is the total number of observations used to estimate d_l. This result demonstrates the efficiency of the estimator, achieving a near-optimal parametric convergence rate [Rao, 1992].

Proving the near-optimal convergence when data are collected from a complex CMDP with delayed observations requires precise understanding of the sources of estimation error and refined analysis to control them. As discussed in Section 2.3, the core idea for estimating d_l is to use observations that are "purified" with respect to d_l. In particular, we construct $\hat{d}_l^t$ using only observations $\{\{\boldsymbol{y}_h^s\}_{h\in\mathbf{D}_{s,l}}\}_{s=1}^t$, where $\boldsymbol{y}_h^s \sim \text{Poi}(d_l\langle\boldsymbol{\theta}_{\mathcal{S}_{h,2}^s}, \mathbf{x}_s\rangle)$. Then, to analyze how the estimation error of TS-MLE depends on the first-stage estimators $\hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s$, we decompose $|\hat{d}_l^t - d_l|$ into two parts: the error by the randomness in Poisson observations (Term A in Eqn. 2.8) and the cumulative estimation error from the first-stage estimators (Term B in Eqn. 2.8). η_h^s in Term A has sub-exponential distributions, specifically, $\text{SubE}(ed_l\,\mathbf{x}_s^\top\boldsymbol{\theta}_{S_{h,2}^s}, 1)$ (Lemma A.4.1), since $\boldsymbol{y}_h^s = d_l\mathbf{x}_s^\top\boldsymbol{\theta}_{S_{h,2}^s} + \eta_h^s$. To control the heavy tail of Term A, we show that $\mathbb{P}(\text{Term A} > \epsilon) \leq \delta$, where $\epsilon = \sqrt{\frac{2e\boldsymbol{B}_d\boldsymbol{B}_x\boldsymbol{B}_\theta}{\boldsymbol{b}^2 N_{t,l}}\log(\frac{2}{\delta})}$. The exploration phase of Algorithm 3 ensures $N_{t,l} \geq \underline{n_l}$, providing sufficient observations to make ϵ small enough to have

faster convergence.

$$|\hat{d}_l^t - d_l| \le \underbrace{\left| \frac{\sum_{s=1}^t \sum_{h\in\mathbf{D}_{s,l}} \eta_h^s}{\sum_{s=1}^t \sum_{h\in\mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s\rangle} \right|}_{\text{Term A}} + \underbrace{\frac{\boldsymbol{B}_d}{N_{t,l}\boldsymbol{b}} \left| \sum_{s=1}^t \sum_{h\in\mathbf{D}_{s,l}} \boldsymbol{x}_s^\top \left(\boldsymbol{\theta}_{S^s_{h,2}} - \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} \right) \right|}_{\text{Term B}}. \tag{2.8}$$

Controlling Term B is essential for bounding $|\hat{d}_l^t - d_l|$. Unlike traditional second-stage estimators, which typically plug in the most recent estimates $\hat{\boldsymbol{\theta}}^t_{S_{h,2}}$, we instead use $\hat{\boldsymbol{\theta}}^s_{S_{h,2}}$, as its estimation error is better controlled in the direction of $\mathbf{x}_s$. Specifically, let $r_h^s = |(\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} - \boldsymbol{\theta}_{S^s_{h,2}})^\top \mathbf{x}_s|$ denote this directional estimation error. By the Cauchy–Schwarz inequality, we have $r_h^s \le \|\hat{\boldsymbol{\theta}}^s_{l'} - \boldsymbol{\theta}_{l'}\|_{\mathbf{V}^s_{l'}} \cdot \|\mathbf{x}_s\|_{(\mathbf{V}^s_{l'})^{-1}}$, where $l' = S^s_{h,2}$. Letting $\gamma^s_{l'} = \|\hat{\boldsymbol{\theta}}^s_{l'} - \boldsymbol{\theta}_{l'}\|^2_{\mathbf{V}^s_{l'}}$, Lemma 2.3.1 guarantees that $\gamma^s_{l'} \le \gamma$ holds with high probability for all s and l'.

To analyze the growth the directional estimation error r_h^s for each $l \in \mathcal{H}$ over time, we apply recounting techniques. Define the dataset $\mathcal{F}^l_{s,l'} := \{h | S^s_{h,1} = l, S^s_{h,2} = l', \boldsymbol{o}^s_h = 0\}$, which partitions $\mathbf{D}_{s,l}$ by $S^s_{h,2}$, the time interval between the two recent ads impression. This partitioning ensures $\sum_{l'\in\mathcal{H}_2} |\mathcal{F}^l_{s,l'}| = |\mathbf{D}_{s,l}|$. We then bound the total directional error as: $\sum_{s=1}^t \sum_{h\in\mathbf{D}_{s,l}} |(\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} - \boldsymbol{\theta}_{S^s_{h,2}})^\top \mathbf{x}_s| \le \sqrt{N_{t,l}} \sqrt{\sum_{s=1}^t \sum_{l'=1}^H \sum_{h\in\mathcal{F}^l_{s,l'}} (r_h^s)^2}$. Let $n_{s,l'} = |\mathcal{F}^l_{s,l'}|$, where $n_{s,l'} \le H$. The total count across all partitions satisfies $\sum_{s=1}^t \sum_{l'\in\mathcal{H}_2} n_{s,l'} = N_{t,l}$. We further bound the error as $\sqrt{N_{t,l}} \sqrt{\sum_{l'=1}^H \gamma \sum_{s=1}^t n_{s,l'} \min(\|\boldsymbol{x}_s\|_{(\mathcal{V}^s_{l'})^{-1}}, 1)}$. Applying the Elliptical Potential Lemma [Abbasi-Yadkori et al., 2011], we obtain $2H\sqrt{d}\sqrt{N_{t,l} \log\left(1 + \frac{T}{2d}\right) \gamma}$. Finally, by applying a union bound to account for the simultaneous occurrence of Term A and Term B, we establish Theorem 2.4.1 (details in Appendix A.3.1).

Building on the efficiency of TS-MLE, we now show that the regret of Algorithm 3 is nearly optimal, as established in Theorem 2.4.2.

Theorem 2.4.2 (Nearly Optimal Regret). *For any $\delta \ge \frac{6}{T^3}$, with probability at least $1-\delta$,* Reg_T *incurred by Algorithm 3 is $\mathcal{O}(dH^2\sqrt{T\log(\frac{TH}{\delta})}\log(1+\frac{T}{2d}))$.*

Theorem 2.4.2 shows that, with high probability, Algorithm 3 achieves a regret bound of

$\tilde{\mathcal{O}}(\sqrt{T})$. Given the $\Omega(\sqrt{T})$ lower bound for CMDPs with even known transitions [Levy and Mansour, 2023], this confirms that Algorithm 3 is nearly optimal in T. As discussed earlier, the key challenge in designing an efficient reinforcement learning algorithm for CMDPs is developing an online estimation oracle that handles long-term Poisson rewards under potentially adversarial arrivals of contextual $\boldsymbol{x}_t$, without relying on distributional assumptions. This improves upon prior work [Modi and Tewari, 2020, Levy and Mansour, 2022, 2023, Levy et al., 2023], which is limited to instantaneous rewards and cannot capture long-term effects in CMDPs, or addresses long-term rewards but fails to account for the heterogeneous effects of $\boldsymbol{x}_t$ [Badanidiyuru et al., 2023]. Even with TS-MLE, establishing a high-probability regret upper bound for Algorithm 3 is nontrivial and requires careful analysis. We provide a proof sketch below, with detailed arguments in Appendix A.3.

The proof begins by decomposing Reg_T into four terms. In Eqn. (2.9), Term (i) equals $\sum_{t=\tau}^{T} \mathrm{OPT}(\Theta, \boldsymbol{x}_t, \mathcal{F}^{\boldsymbol{x}_t}) - \sum_{t=\tau}^{T} \mathrm{OPT}(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t})$, and Term (iv) $= \mathbb{E}(\sum_{t=\tau}^{T} R(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) - R(\pi_t; \boldsymbol{x}_t, \Theta, \mathcal{F}^{\boldsymbol{x}_t}))$, capturing the cumulative regret due to transition error, arising from the discrepancy between $\mathcal{F}^{\boldsymbol{x}_t}$ and $\hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}$ [4].

Term (iii) is $\sum_{t=\tau}^{T} \mathrm{OPT}(\mathcal{C}_{t-1}, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) - \mathbb{E}(\sum_{t=\tau}^{T} R(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}))$, which accounts for decision error resulting from imprecise estimates of the model parameters $\Theta = \{\boldsymbol{\theta}_l\}_{l \in \mathcal{H}} \cup \{d_l\}_{l \in \mathcal{H}_1}$. Thus, accurate estimators of $\mathcal{F}^{\boldsymbol{x}_t}$ and Θ lead to low cumulative regret, as shown jointly by Lemma 2.4.1 and 2.4.2.

$$\mathrm{Reg}_T \leq \Theta(\log T) + \text{Term (i)} + \text{Term (ii)} + \text{Term (iii)} + \text{Term (iv)} \tag{2.9}$$

Lemma 2.4.1 (Term (i) Upper Bound)**.** *For $\delta \geq 1/T^4$, with probability at least $1-\delta$, Term (i) in Eqn. (2.9) is bounded above by $\mathcal{O}(dH^2\sqrt{T}\sqrt{\log((1+T)/\delta)\log(1+T/d)})$.*

Lemma 2.4.1 provides a high-probability upper bound for Term (i). The proof firstly uses the simulation lemma [Kearns and Singh, 2002] to bound Term (i) by the cumulative

4. We have $\tau = H\underline{n_l} + 1$ and regret incurred in the exploration phase is $\Theta(\log T)$.

estimation error in the transition dynamics, $\sum_{t=1}^{T} \sup_b \sup_h |\hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) - \mathcal{F}_h(b, \boldsymbol{x}_t)|$, which depends on $|\hat{\boldsymbol{\beta}}_h^t - \boldsymbol{\beta}_h|$ and $|(\hat{\sigma}_h^t)^2 - \sigma_h^2|$. A high-probability bound is then derived for these estimations errors, with $|(\hat{\sigma}_h^t)^2 - \sigma_h^2|$ shown to follow a sub-exponential distribution. To ensure quadratic tail decay, we set $\delta \geq 1/T^4$. A similar analysis applies to Term (iv), which is bounded by $\mathcal{O}(dH^2\sqrt{T}\sqrt{\log((1+T)/\delta)\log(1+T/d)})$ with probability at least $1 - 1/T^4$. Details are in Appendix A.3.2.

Lemma 2.4.2 (Term (iii) Upper Bound)**.** *With probability at least* $1 - 2\delta$*, where* $\delta \geq \frac{1}{T^3}$*, Term (iii) in Eqn.* (2.9) *is bounded by* $\mathcal{O}(H^2\sqrt{dT\log(4TH/\delta)}\log(1+T/2d))$.

Lemma 2.4.2 provides a high-probability bound for Term (iii), which can be decomposed into two parts: the cumulative estimation error of $\hat{\boldsymbol{\theta}}_l^t$, which is $\sum_{t=\tau}^{T} \mathbb{E}(\sum_{h=1}^{H} |\langle \tilde{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} + \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t\rangle|) + \sum_{t=\tau}^{T} \mathbb{E}(\sum_{h=1}^{H} |\langle \tilde{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1} - \hat{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1} + \hat{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,2}^t}, \boldsymbol{x}_t\rangle|)$ and the cumulative estimation error of $\hat{d}_l$, specifically, $\sum_{t=\tau}^{T} \mathbb{E}(\sum_{h=1}^{H} |\tilde{d}_{S_{h,1}^t}^{t-1} - \hat{d}_{S_{h,1}^t}^{t-1} + \hat{d}_{S_{h,1}^t}^{t-1} - d_{S_{h,1}^t}|)$. Here, $\tilde{\Theta}_{t-1} := \arg\max_{\Theta \in \mathcal{C}_{t-1}} R(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t})$. Applying Lemma 2.3.1 and Theorem 2.4.1, together with a union bound, yields the high-probability bound for Term (iii) (Appendix A.3.3).

Meanwhile, in Eqn. (2.9), Term (ii) $= \sum_{t=\tau}^{T} \mathrm{OPT}(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) - \sum_{t=\tau}^{T} \mathrm{OPT}(\mathcal{C}_{t-1}, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t})$, measures the gap in regret between a chosen point and the optimal point within a feasible set. Lemma 2.4.3 proves $\Theta \in \mathcal{C}_t$ for all $t \geq \tau$ with high probability, ensuring that Term (ii) can be negative with high probability. Combining these results, we conclude that, with probability at least $1 - \frac{6}{T^3}$, the overall regret Reg_T is $\mathcal{O}(dH^2\sqrt{T\log(\frac{TH}{\delta})}\log(1+\frac{T}{2d}))$.

Lemma 2.4.3 (Confidence Region of Θ)**.** *With probability at least* $1 - \delta$ *with* $\delta \geq \frac{2}{T^3}$*,* $\Theta \in \mathcal{C}_t, \forall t \geq \tau$.

2.5 Experiments

We conducted experiments, which leads to two key findings: first, our algorithm achieves near-optimal regret scaling of $\sqrt{T}$, and second it consistently outperforms several strong baselines, including *aggressive bidding*, *passive bidding*, and *random bidding* strategies, with definition shown below.

Environment Setup. We simulate a second-price auction with horizon $H = 3$ (a realistic setting since advertisers typically show ads only a few times per user), context dimension $d = 2$, and $T = 20{,}000$ sequential customers. The first 2400 rounds are used for exploration, and the remaining rounds for exploitation. Context vectors $\boldsymbol{x}_t \in \mathbb{R}^2$, as well as parameters $d_l, \boldsymbol{\beta}_h, \sigma_h$, are sampled elementwise from $|\mathcal{N}(0,1)| + 0.1$ to ensure positivity, while $\boldsymbol{\theta}_l \sim 5|\mathcal{N}(0,1)| + 0.1$. Under this configuration, each ad impression yields an average instantaneous reward roughly five times its cost (i.e., the highest other bid). The reward also decays rapidly over time, making *aggressive bidding*—always winning the impression—close to optimal and therefore a competitive baseline.

Estimators. We estimate four sets of parameters: $\boldsymbol{\theta}_l$ via the online Newton method (Algorithm 4) with truncation threshold $\Gamma = 100{,}000$, bound $\boldsymbol{B}_\theta = 10$, zero initialization, and $V_0 = I$. We estimate delay impact d_l via the two-stage MLE (Eq. (2.3)) using $\mathbf{D}_{t,l}$, $\hat{\boldsymbol{\theta}}$, and $\boldsymbol{x}_t$; $\boldsymbol{\beta}$ by ridge regression (Eq. (2.5), $\lambda = 1.0$); and σ via empirical variance (Eq. (2.6)).

Connection Between Bid, Outcome, and Reward. In our setting, the bidder's action is the bid amount, but due to the second-price auction format, the reward depends only on the outcome–whether the bid exceeds the HOB-not the bid itself. Winning results in paying the HOB; losing incurs no cost. This decoupling allows optimal rewards to be computed from outcomes rather than bids. We leverage this to evaluate both oracle and policy-based rewards. With full knowledge of true parameters, we enumerate all $2^H = 8$ possible win/loss outcomes and define the oracle reward as the maximum achievable value. With estimated parameters, we compute expected rewards for all outcome sequences, select

the best, and evaluate it in simulation. Regret is then the cumulative difference between oracle and realized rewards, reflecting both estimation and decision errors.

Baseline Policies and Regret Comparison. We compare our algorithm against three fixed policies: *Aggressive bidding*: Always bids above HOB ($\boldsymbol{o}_t = [1, 1, 1]$). *Random bidding*: Samples $\boldsymbol{o}_t$ uniformly from all 8 outcomes. *Passive bidding*: Rarely bids or wins ($\boldsymbol{o}_t = [0, 0, 0]$). All baselines are evaluated using the same oracle-based regret metric. Note that *aggressive bidding*, while competitive, still incurs $O(T)$ regret since it is non-adaptive.

We ran all algorithms over 20 independent trials. Table 2.1 reports the mean cumulative regret ($\pm$ 0.5 standard deviation) and the fitted regret order (via log–log regression). Our algorithm achieves an estimated regret order of 0.37, while all baselines exhibit linear regret, validating our theoretical results and demonstrating the substantial advantage of adaptive learning in ad bidding.

t	**Algorithm 1**	**Aggressive Bid**	**Random Bid**	**Passive Bid**
500	1770 [1174, 2379]	228 [121, 336]	1920 [1791, 2049]	4630 [3391, 5869]
5000	8465 [5585, 11345]	2373 [1243, 3502]	19285 [17873, 20697]	46179 [33850, 58508]
10000	8484 [5606, 11363]	4741 [2477, 7005]	38399 [35573, 41226]	92468 [67735, 117201]
15000	8505 [5628, 11382]	7142 [3724, 10561]	57651 [53452, 61849]	138652 [101576, 175729]
20000	**8525 [5649, 11401]**	9568 [4991, 14146]	76945 [71390, 82500]	184873 [135414, 234332]
T^{α}	**0.37**	**1.0**	**1.0**	**1.0**

Table 2.1: Average cumulative regret ($\pm$ 0.5 standard deviation) and fitted regret order (T^{α}).

2.6 Discussion and Limitation

In this paper, we model ad bidding as a CMDP, presenting a unified framework to address delayed impacts, cumulative effects, and individual personalization. We designed a near-optimal algorithm using a novel two-stage maximum likelihood estimator to effectively handle delayed rewards. However, several important directions remain open for future research. One promising extension involves incorporating budget constraints, a critical practical consideration in real-world advertising. Developing constrained optimization algorithms for CMDP

with delayed rewards is highly non-trivial and will be explored in future work.Although our work validates the theoretical insights through experiments, implementing and evaluating the proposed algorithms in real-world settings would provide valuable evidence of its practical applicability. In particular, such experiments could help assess two modeling choices: the adequacy of modeling the expected product conversion rate using linear functions, and the robustness of focusing only on very recent ad exposures due to recency bias. However, given the lack of publicly available online advertising and bidding datasets, we leave this empirical validation to future work as an important step toward bridging the gap between theory and practice.

CHAPTER 3

LEARNING FROM IMPERFECT HUMAN FEEDBACK

3.1 Introduction

Many real-world problems, such as personalized recommendation [Yue and Joachims, 2009b, Immorlica et al., 2020, Yao et al., 2022a] and fine-tuning generative models [Bai et al., 2022, Casper et al., 2023, Han et al., 2024b], require learning from human feedback. Expressing preferences as numerical values or concrete functions is generally challenging for humans. Therefore, an approach that has achieved significant real-world success is to learn humans' preferences by eliciting their comparative feedback between two options [Bai et al., 2022]. A natural theoretical framework capturing such learning task is the seminal dueling bandits framework introduced by Yue and Joachims [2009b]. The dueling bandit problem features a sequential online learning problem, during which a pair of actions are selected at each round and only their comparison result is revealed to the learner. The comparison outcome between two actions is modeled using a utility function of actions, alongside a link function that determines the probability of each action winning based on their utility difference. This modeling approach is termed as utility-based dueling bandit and achieves great success in generating summaries closely aligned with human preferences [Stiennon et al., 2020]. However, human feedback is often *imperfect*, a crucial factor that has been largely overlooked in previous studies of dueling bandit. As we are all aware of, humans are not always rational [Posner, 1997] neither perfectly know our preferences [Pu et al., 2012]. A powerful framework that can capture this problem is robust bandit learning under adversarial corruption [Bogunovic et al., 2020, Saha and Gaillard, 2022, Di et al., 2024]. However, assuming fully adversarial feedback from a regular human seems over-pessimistic at the first glance, hence might make our algorithm development overly conservative. Indeed, ample behavioral studies show that humans often refine their preferences through interactions with the system,

navigating a complex tradeoff between exploration and exploitation [Cohen et al., 2007, Wilson et al., 2014]. Furthermore, their response errors may exhibit systematic patterns that depend on their past consumption history, rather than merely random noise. Though the sources of imperfections vary a lot in human feedback, one common finding in these studies is that *the feedback's imperfection tends to decrease over time*, as humans interact more with the system [Cohen et al., 2007, Wilson et al., 2014, Immorlica et al., 2020, Yao et al., 2022a]. This premise is the core motivation of this work, which studies dueling bandit learning under corruption that can be arbitrary but have decaying scales.

Our Model and Contributions. We cast Learning from Imperfect Human Feedback (LIHF) as a concave-utility continuous-action dueling bandit learning problem under adversarial utility corruption yet with decaying scale, upper bounded by $\mathcal{O}(t^{\rho-1})$ at round t, for some $\rho \in [0, 1]$ (hence the total amount of attack within time T is $\mathcal{O}(T^{\rho})$). A fundamental research question that motivates our study is whether LIHF is fundamentally easier than learning under arbitrarily corrupted feedback.

Our first main technical result hints on a potentially negative answer to the above question. We prove a strong regret lower bound of $\Omega(d \max\{\sqrt{T}, T^{\rho}\})$ under LIHF, where d is the action's dimension. This bound holds even when the decaying rate of imperfection index ρ is known. This lower bound has the same order as recent lower bounds of dueling bandits under (the harder) arbitrary corruption [Agarwal et al., 2021, Di et al., 2024], thus hinting that learning from imperfect human feedback may be no easier than learning from arbitrary adversarial corruption. Notably, however, we use the phrase "hint", while not an affirmative claim, because our setting with continuous action space and concave utility function is different from the K-armed bandits setting of Agarwal et al. [2021] and linear utility function case of Di et al. [2024]. Therefore, neither their lower bounds nor their proof techniques are directly applicable to our setting with corruption to strictly concave utilities; see more discussions and comparisons below.

Our second main result is an efficient algorithm with regret upper bound $\tilde{\mathcal{O}}(d \max\{\sqrt{T}, T^{\rho}\})$ that matches the above lower bound, up to logarithmic terms, in the same setting (i.e., decaying corruption level $t^{\rho-1}$ with known ρ). Unlike previous corruption-robust dueling bandit algorithms based on robust K-armed bandits [Agarwal et al., 2021] or robustified LinUCB [Di et al., 2024], our algorithm is a gradient-based algorithm that is built upon the recent Noisy Comparison-based Stochastic Mirror Descent (NC-SMD) of Kumagai [2017]. This is natural for our setup with continuous actions and strictly concave utility. While our algorithm itself can be viewed a natural generalization of Kumagai [2017], its analysis is highly non-trivial and much more complex, due to the necessity of accounting for corruption.

Our main technical contribution is a new framework for analyzing the regret of gradient-based algorithms under arbitrary corruption. This framework introduces a novel regret decomposition lemma for dueling bandits, which upper bounds total regret by combining decision regret and feedback error due to corruption, based on quantifying bias in gradient estimation. It yields a tight analysis of our new algorithm and implies regret bounds for popular existing algorithms in (even fully) adversarial corruption settings, such as the dueling bandit gradient descent (DBGD, Yue and Joachims [2009b]).

These upper bounds are looser than our lower bound for the LIHF setting. It is an intriguing open question to understand whether these upper bounds under arbitrary corruption are tight, or the lower bound under arbitrary corruption may be stronger than our lower bound under LIHF. Finally, we conduct simulations which validate our theory.

Comparisons with Related Works. Since multiple recent works [Agarwal et al., 2021, Saha and Gaillard, 2022, Di et al., 2024] study dueling bandits under arbitrarily adversarial corruption , it is worthwhile to discuss the key difference between these works and ours. The first key difference is the setting, which leads to fundamentally different algorithms hence analysis. Specifically, our setting with continuous action space and a completely unknown concave utility function is motivated by addressing problems from modern recommender sys-

tems where the action (i.e., contents) space is extremely large, not enumerable, thus often embedded as high-dimensional continuous vectors [Chen et al., 2019, 2021a]. This setting and our techniques are crucially different from earlier works. Agarwal et al. [2021], Saha and Gaillard [2022] consider the K-arm dueling bandit setting and utilizes the reduction of Ailon et al. [2014] from dueling bandits to multi-K-armed bandits (MAB); [Di et al., 2024] considers linear utility setting where the unknowns is a parameter vector θ, and utilizes techniques from linear contextual bandits based on the optimism principle. The methods all rely on parameter estimation which are not applicable to our setting with arbitrary concave utilities. Moreover, they often need to enumerate all actions to identify the best one under optimism, which are not computationally efficient for our setting with continuous action space. This is why our algorithm is gradient-based and falls within the family of stochastic mirror descent. To the best of our knowledge, this is the first time a corruption-robust algorithm is developed for dueling bandits with continuous actions and strictly concave utilities, which is also why it is necessary for us to develop a new analysis framework. Additionally, we also note that dueling bandits with strictly concave utility and linear utility are generally not comparable, both in terms of their problem difficulties and methodologies. A strictly more general situation is the concave utility case (no need to be strongly concave), however the best upper bound so far for general concave utility is $\mathcal{O}(T^{3/4})$ [Flaxman et al., 2005, Yue and Joachims, 2009b], which are considerably worse than the $\mathcal{O}(T^{1/2})$ for linear and for strongly concave utility.

The second main difference between our work and previous is the more restricted class of adversarial environments. Our model is motivated by imperfect human feedback whereas previous studies are motivated by arbitrarily adversarial adversaries.

Finally, a more subtle difference is that the corruption in our model is on utility with motivations from learning from imperfect human feedback, whereas corruption in all previous settings directly flip the comparison outcome. These two differences lead to a very different

proof techniques of our lower bound. Our proof is more involved because the adversary has limited capability to alter the comparison outcomes. Specifically, the corruption are decaying; moreover, a constant amount of utility corruption only slightly shifts the outcome probability whereas in previous settings each corruption completely changes the outcome.

3.2 The Problem of Learning from Imperfect Human Feedback

Notation. For a positive integer T, we use $[T]$ to denote $\{1, 2, \ldots, T\}$. We use standard asymptotic notations including $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$. We use $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Omega}(\cdot)$, $\tilde{\Theta}(\cdot)$ to hide logarithmic factors. We use $\|\cdot\|_2$ to define L_2 norm, $\|\cdot\|_\infty$ to define infinity norm, and $\lambda_{\max}(\cdot)$ to denote maximum eigenvalue.

Motivated by alignment of machine learning models with human preferences [Bai et al., 2022], we study learning from comparative human feedback within the seminal dueling bandit framework [Yue and Joachims, 2009b] but account for "imperfections" in human feedback, as described below.

Basics of Continuous Dueling Bandits. We consider the dueling bandit framework with continuous *action* set $\mathcal{A} \subset \mathbb{R}^d$ [Yue and Joachims, 2009b]. At each round $t \in [T]$, the learner chooses two actions a_t, a_t' to present to some human agent, henceforth denoted as the "user". The user receives *utility* $\mu(a_t), \mu(a_t')$ from the two actions respectively, and will pick one of these actions following a *link* function $\sigma(\cdot)$. In the absence of corruption, the user selects action a_t with probability $\sigma(\mu(a_t) - \mu(a_t'))$ and chooses action a_t' otherwise. Since the comparison outcome has less errors compared to the exact utility value while still carries useful information about the underlying utility function μ, this form of human feedback has been widely used for learning human preferences [Ailon et al., 2014, Maystre and Grossglauser, 2017, Bengs et al., 2021a, Bai et al., 2022]. Following standard assumptions in this field [Yue and Joachims, 2009b, Kumagai, 2017], we also assume

1. the action set $\mathcal{A}$ is a convex compact set that contains the origin, has non-empty

interior, and is contained in a d-dimensional ball of radius R;

2. the utility function $\mu : \mathcal{A} \to \mathbb{R}$ is strongly concave and twice-differentiable. The following constants are useful for our algorithm analysis: μ is L^{μ}-Lipschitz, α-strongly concave,[1] κ-smooth, bounded, $R^{\mu} := \sup_{a,a' \in \mathcal{A}} \mu(a) - \mu(a')$, and $a^* := \arg\max_{a \in \mathcal{A}} \mu(a)$ is the unique optimal within $\mathcal{A}$;

3. the link function $\sigma : \mathbb{R} \to [0, 1]$ is smooth, rotation-symmetric (i.e. $\sigma(x) = 1 - \sigma(-x)$), and concave for $\mathbb{R}^+$. Since R^{μ} is finite, there exists positive l_1^{σ}, L_1^{σ} and R_2^{σ} such that $l_1^{\sigma} \leq \sigma' \leq L_1^{\sigma}$ on $[-R^{\mu}, R^{\mu}]$, and the second derivative is bounded above by R_2^{σ} and L_2^{σ}-Lipschitz on $[-R^{\mu}, R^{\mu}]$. For a more detailed discussion on the generality of the modeling assumptions, please refer to Appendix B.2.

A broad range of natural link functions satisfy our assumptions, including the standard logistic distribution function, the cumulative standard Gaussian distribution function, and the linear function. This distinguishes our work from some prior studies, working with concrete link function formats such as logistic function [Maystre and Grossglauser, 2017, Saha, 2021b, Xu et al., 2024] or the linear function [Ailon et al., 2014, Chen and Frazier, 2017, Zimmert and Seldin, 2018, Lin and Lu, 2018].

Modeling Imperfect Human Feedback as a Restricted Form of Adversarial Corruption. To incorporate imperfect human feedback, we study a generalization of the dueling bandit framework above, by introducing a *corruption* term, $c(a, a')$, to the utility difference. Formally, let $a \succ a'$ represent the event where the user chooses a over a'. The probability of this event in the presence of corruption $c(a, a')$ is denoted by $\hat{\mathbb{P}}(a \succ a')$, expressed as $\hat{\mathbb{P}}(a \succ a') := \sigma(\mu(a) - \mu(a') + c(a, a'))$. We denote the user's preferential feedback under corruption as $\hat{\mathcal{F}}(a, a')$, referred to as *imperfect dueling feedback*. Mathematically, $\hat{\mathcal{F}}(a, a')$ follows a binomial distribution with mean $\hat{\mathbb{P}}(a \succ a')$, as expressed by the following

1. μ is α-strongly concave for some $\alpha > 0$ if $\mu(x) \geq \mu(y) + \langle \nabla \mu(x), x - y \rangle + \frac{\alpha}{2} \|y - x\|_2^2$ for any $x, y \in \mathcal{A}$.

equation.

$$\hat{\mathcal{F}}(a, a') := \begin{cases} 1 \text{ w.p. } \hat{\mathbb{P}}(a \succ a') \\ 0 \text{ w.p. } 1 - \hat{\mathbb{P}}(a \succ a') \end{cases}$$

The "hat" notation in $\hat{\mathcal{F}}$ and $\hat{\mathbb{P}}$ is to emphasize the presence of corruption. When there is no corruption, we use $\mathbb{P}(a \succ a') := \sigma(\mu(a) - \mu(a'))$ to represent the probability of the event $a \succ a'$, and $\mathcal{F}(a, a')$ to denote the corresponding dueling feedback.

We remark that the term $c(a, a')$ above is often called "strong corruption" in the literature of adversarial corruption because it can be introduced after observing actions a and a' and the corruption function, $c(\cdot)$, can depend on μ and past actions [Bogunovic et al., 2020, He et al., 2022, Di et al., 2024]. The only difference between our model and the above works on adversarial corruption is a natural restriction to the scale of the corrupted term $c(a, a')$. Our motivation is that human feedback, while imperfect, should not be arbitrarily adversarial and tends to improve as human agent interact with the learner. This motivates our following definition of *imperfect human feedback*.

Definition 3.2.1 (ρ-Imperfect Human Feedback)**.** *The human feedback is said to be ρ-imperfect for some $\rho \in [0, 1]$ if there exists a constant C_κ such that corruption $c_t(a_t, a_t')$ satisfies $|c_t(a_t, a_t')| \leq C_\kappa t^{\rho-1}, \forall t \in [T]$.*[2]

The definition shows that ρ-imperfect human feedback is mathematically equivalent to a restricted form of strong adversarial corruption to user utility functions. The total corruption is bounded by $\sum_{t=1}^{T} |c_t(a_t, a_t')| \leq C_\kappa T^\rho$, which is a key parameter influencing the intrinsic difficulty of the learning problem. We thus cast the Learning from Imperfect Human Feedback (LIHF) problem as dueling bandits under strong corruption but with decaying scales. For clarity, we use "arbitrary adversarial corruption" as corruption without decaying constraints and "ρ-Imperfect Human Feedback" as corruption with decaying constraints,

2. All our results naturally applies to $\rho < 0$, which is a significantly easier situation since accumulated corruption is a constant in that case. Therefore, we will not explicitly consider it in this paper.

as described in Definition 3.2.1. Through this modeling, we aim to: first create a more optimistic framework for LIHF compared to arbitrary adversarial corruption, and second, explore LIHF's intrinsic difficulty to achieve more efficient learning guarantees than those possible under arbitrary adversarial corruption. We direct readers to Appendix B.1 for a more comprehensive discussion on challenges in developing robust algorithms for continuous dueling bandits with generally concave utility under arbitrary adversarial corruption.

Several points are worth noting about Definition 3.2.1. While we are not the first to model imperfect human feedback, we are the first to apply it to dueling bandits with strictly concave utilities. Recent works, motivated by recommendation systems, have explored learning user preferences from imperfect feedback [Immorlica et al., 2020, Yao et al., 2022a, Wang et al., 2023]. These studies assume that users do not know their true expected reward θ_i for each choice i (i.e., arm) and instead behave based on an estimated reward $\hat{\theta}_{t,i}$ at time t. The utility difference $|\theta_i - \hat{\theta}_{t,i}| = c_t$ is modeled as a decreasing function of time, $t^{\rho-1}$, indicating that users improve their preference estimates over time. By viewing the human utility $\mu(a)$ as the average reward, our modeling of ρ-Imperfect Human Feedback shares the same spirit; it captures imperfect yet gradually improving human feedback (e.g., due to inaccuracies in utility perception or initial ignorance of true preferences). As one of our motivations, when generative models like ChatGPT learn to create personalized content, the user could undergo a process of preference refinement during interactions with the model. Additionally, our model differs from recent corruption-robust dueling bandit studies [Komiyama et al., 2015, Saha and Gaillard, 2022, Di et al., 2024], which corrupt realized outcomes (e.g., flipping $a' \succ a$ to $a \succ a'$) and are motivated by adversaries like malicious users or fraudulent clicks [Deshpande and Montanari, 2013, Lykouris et al., 2018]. In contrast, our model focuses on utility corruption caused by imperfect perception of true utility, not outcome manipulation.

Learning goal: regret minimization by Learning from ρ-Imperfect Human Feedback (ρ-LIHF). The goal of the learner is to optimize her sequential actions to

minimize the following dueling regret in the ρ-LIHF problem:

$$\mathrm{Reg}_T := \mathbb{E}\left(\sum_{t=1}^{T} \sigma(\mu(a^*) - \mu(a_t)) + \sigma(\mu(a^*) - \mu(a_t')) - 1\right). \tag{3.1}$$

This regret measure has been extensively studied in prior literature [Yue and Joachims, 2009b, Komiyama et al., 2015, Kumagai, 2017, Saha and Gaillard, 2022]. Other regret measures, such as the cumulative difference between optimal and obtained utility, are also considered. We discuss the equivalence of various regret measures in Lemma B.4.1.

3.3 The Intrinsic Limit of LIHF

In this section, we study the intrinsic limit of learning from ρ-Imperfect Human Feedback. We are particularly interested in understanding whether learning from this restricted version of corrupted feedback is fundamentally "easier" than learning from arbitrarily adversarial corruption. Somewhat surprisingly, the answer seems to be *no*, as illustrated by the following lower bound result.

Theorem 3.3.1. *There exists a ρ-Imperfect Human Feedback (see Definition 3.2.1), strongly concave utility function μ, and link function σ such that any learner has to suffer $\mathrm{Reg}_T \geq \Omega\left(d \max\{\sqrt{T}, T^\rho\}\right)$, even* with the knowledge of ρ.

Notably, similar lower bound results have appeared in recent studies on adversarial corruption in bandits [Bogunovic et al., 2022], including dueling bandits [Agarwal et al., 2021, Di et al., 2024], showing the same lower bound $\Omega(d \max\{\sqrt{T}, T^\rho\})$ (though they often use $C := \Theta(T^\rho)$ to denote total corruption budget). However, a few key distinctions are worth noting. First, Theorem 3.3.1 presents a stronger result, showing that even when corruption decays over time, this structure (which defines our LIHF problem) does not make learning easier. Second, our corruption targets utility values, which is quite different from the corruption of comparison outcomes studied in [Agarwal et al., 2021, Di et al., 2024]. In their case,

corruption can fully flip comparison outcomes, which is critical for promoting sub-optimal arms in their lower bound proofs. However, such full control of the outcome is infeasible in our corruption of only the utility, which only slightly shifts the comparison probabilities since the scale of the corruption is also upper bounded by $\mathcal{O}(t^{\rho-1})$.

The aforementioned difference from previous works [Agarwal et al., 2021, Di et al., 2024] also render our proof of Theorem 3.3.1 fundamentally different from (and much more involved than) their proofs. Specifically, due to limited and diminishing corruption power, the corruption in our LIHF problem are insufficient to completely "mask" the instance as one with a different sub-optimal arm, which is the key strategy in previous lower bound proofs. Thus, our proof has to leverage information-theoretic lower bounds of statistical distributions to understand how small-scale corruption at each round could influence the overall function estimation errors. Core of our proof is the follow lower bound result for a different, and intuitively easier, problem: the standard bandit setup with direct reward feedback shown in Lemma 3.3.1 below. We then convert this lower bound for direct reward feedback to dueling feedback through a linear link function, i.e. $\sigma(x) = \frac{1+x}{2}$.

Lemma 3.3.1 (Lower Bound under Direct Reward Feedback)**.** *Consider bandit learning with direct reward feedback, where reward* $r(a) := \mu(a) + \epsilon$, ϵ *follows standard normal distribution. The action space* $\mathcal{A}$ *is contained in a* d*-dimensional unit ball, and* μ *is* $\mu_\theta(a) := \theta^\top a - \frac{1}{2}\|a\|_2^2$, *with* $\theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1$. *Under* ρ*-Imperfect Human Feedback (Definition 3.2.1), for any* T, $d \leq \frac{1}{C_\kappa} T^{1-\rho}$, *and for any learner, there exists a* θ *such that under direct reward feedback,* even with the knowledge of ρ[3], *has to suffer regret* $Reg_T := \mathbb{E}\{\sum_{t=1}^T \mu_\theta(a^*) - \mu_\theta(a_t)\} \geq \frac{d}{4} C_\kappa T^\rho$.

Proof Sketch of Lemma 3.3.1. We aim to show that there exists a corruption strategy such that the regret incurred by any learner with direct reward feedback is no less than $\Omega(dT^\rho)$. The formulation of such a corruption strategy hinges on the observation that the regret

3. We only consider $\rho \geq \frac{1}{2}$. If $\rho < \frac{1}{2}$, the lower bound degenerates to $\Omega(d\sqrt{T})$ [Kumagai, 2017].

incurred by the learner is bounded below by the sum of the Kullback-Leibler (KL) divergence between the distribution of the corrupted reward feedback $\hat{v}_t$ conditioned on different values of θ. Specifically, let θ be uniformly drawn from $\{-\beta, \beta\}^d$, for some constant $\beta > 0$. Given an action a_t and index i, conditioned on the value of the i-th coordinate of θ, θ_i, if $\theta_i > 0$, then $\hat{v}_t$ is

$$\hat{v}_t = \mu_\theta(a_t) + c_t(a_t|\theta_i > 0) + \xi_{a_t} = \left(-\frac{1}{2}\|a_t\|^2 + \sum_{j \neq i} \theta_j a_{t,j} \right) + \beta a_{t,i} + c_t(a_t|\theta_i > 0) + \xi.$$

Conditioned on $\theta_i < 0$, the corrupted reward feedback $\hat{v}'_t$ is

$$\hat{v}'_t = \mu_\theta(a_t) + c_t(a_t|\theta_i < 0) + \xi_{a_t} = \left(-\frac{1}{2}\|a_t\|^2 + \sum_{j \neq i} \theta_j a_{t,j} \right) - \beta a_{t,i} + c_t(a_t|\theta_i < 0) + \xi.$$

The noise ξ follows standard normal distribution. Consider the following corruption strategy which sets $c_t(a_t|\theta_i > 0) := -\beta a_{t,i}$ and $c_t(a_t|\theta_i < 0) := \beta a_{t,i}$ to minimizes the KL divergence between $\hat{v}_t$ and $\hat{v}'_t$. Together with 1-strongly concavity of μ_θ, we can establish $\mathbb{E}\left\{\sum_{t=1}^T \mu_\theta(a^*) - \mu_\theta(a_t)\right\} \geq \frac{1}{2}\max\left\{\sum_{t=1}^T (\frac{C_\kappa T^{\rho-1}}{\beta} - \beta)^2, \frac{dT\beta^2}{2}\right\}$. Since $d \leq \frac{1}{C_\kappa} T^{1-\rho}$, we can set $\beta := \sqrt{C_\kappa} T^{\frac{\rho-1}{2}}$ to satisfy $\|\theta\|_2 \leq 1$ and to optimize the aforementioned lower bound. Then we obtain $\mathbb{E}\left\{\sum_{t=1}^T \mu_\theta(a^*) - \mu_\theta(a_t)\right\} \geq dC_\kappa T^\rho/4$. Notice that the corruption level each round is bounded by $\beta\|a_t\|_\infty$. In the scenario when the radius of $\mathcal{A}$ is upper bounded by β, the corruption budget each round is bounded by $C_\kappa t^{\rho-1}$. It implies that the corruption is ρ-Imperfect Human Feedback, completing the proof. □

Theorem 3.3.1 can be proved by converting the lower bound in Lemma 3.3.1 for direct reward feedback to dueling feedback (see Lemma B.3.1). Before concluding this section, we note that the strongly concave utility function in the lower bound of Theorem 3.3.1 is not necessary. In particular, Proposition 1 below shows that similar lower bound applies to linear utilities as well. This result happens to resolve an open problem posed by Yao et al.

[2022a] which studies a similar learning from imperfect user problem with linear utility and dueling feedback (they termed it the "learning from a learning user" problem), develops a no-regret learning algorithm for that setting, but left the lower bound as an open problem. Proposition 1 implies that their upper bound's dependence on T is tight. We direct readers to Appendix B.3.1 and B.3.2 for proof details of Lemma 3.3.1 and Proposition 1.

Proposition 1. *There exists an LIHF instance with ρ-Imperfect Human Feedback (Definition 3.2.1), linear user utility μ, and link function σ, such that any learner has to suffer $\mathrm{Reg}_T \geq \Omega(d \max\{\sqrt{T}, T^{\rho}\})$, even with the knowledge of ρ.*

3.4 An Efficient and Tight ρ-LIHF Algorithm

In this section, we present a learning algorithm with a regret upper bound that matches the $\Omega(d \max\{\sqrt{T}, T^{\rho}\})$ lower bound, up to a logarithmic term, for the ρ-LIHF setting discussed in Section 3.3. Our algorithm, **Ro**bustified **S**tochastic **M**irror Descent for **I**mperfect **D**ueling (RoSMID), outlined in Algorithm 3, generalizes the Noisy Comparison-Based Stochastic Mirror Descent (NC-SMD) from Kumagai [2017]. RoSMID employs a corruption-aware learning rate, $\eta_\rho := \sqrt{\log T}/(dT^{\max\{0.5,\rho\}})$, to slow down the learning when the imperfection level ρ is large, with NC-SMD being the special case with $\rho = 0$. Although RoSMID may initially seem like a pretty natural extension of NC-SMD, the choice of optimal η_ρ is in fact through careful derivation from a novel framework for analyzing continuous bandits under utility corruption. We conclude by demonstrating the applicability of this new framework to analyzing other gradient-based algorithms.

Algorithm 3: **Ro**bustified **S**tochastic **M**irror Descent for **I**mperfect **D**ueling (**RoSMID**)

Input: Learning rate $\eta_\rho := \sqrt{\log T}/\left(dT^{\max\{0.5,\rho\}}\right)$, ν-self-concordant function $\mathcal{R}$ [a], time horizon T, tuning parameters $\lambda \leq \frac{l_1^\sigma \alpha}{2}$, $\phi \geq ((L_1^\sigma)^3 L_2^\sigma/\lambda)^2$

Initialize $a_1 = \arg\min_{a\in A} \mathcal{R}(a)$

for $t \in [T]$ **do**

  Update concordant function $\mathcal{R}_t(a) = \mathcal{R}(a) + \frac{\lambda\eta_\rho}{2}\sum_{i=1}^t \|a - a_i\|^2 + \phi\|a\|^2$

  Sample direction u_t uniformly at random from a *unit sphere* $\mathbb{S}^d$

  Obtain corrupted feedback $\hat{\mathcal{F}}(a_t', a_t) \in \{0, 1\}$, for a_t and $a_t' = a_t + \nabla^2\mathcal{R}_t(a_t)^{-1/2}u_t$

  Compute the corrupted gradient: $\hat{g}_t = \hat{\mathcal{F}}(a_t', a_t)d \cdot \nabla^2\mathcal{R}_t(a_t)^{1/2} \cdot u_t$

  Set $a_{t+1} = \nabla\mathcal{R}_t^{-1}(\nabla\mathcal{R}_t(a_t) - \eta_\rho\hat{g}_t)$

Output: a_{T+1}

a. We direct reader to Definition B.4.1 for formal definition of self-concordancy.

Theorem 3.4.1. *RoSMID satisfies* $Reg_T \leq \tilde{\mathcal{O}}(d\max\{\sqrt{T}, T^\rho\})$ *for any* ρ*-LIHF problem.*

The main technical contribution underlying this theorem is a novel framework for analyzing continuous dueling bandits with *arbitrary strongly concave*, yet possibly imperfect or corrupted utilities, which we believe has independent value. This framework not only enables a tight regret upper bound analysis, as shown in Theorem 3.4.1, but also allows us to easily derive regret guarantees for other dueling bandit algorithms under adversarial corruption, as demonstrated in Subsection 3.4.1. While recent works have explored corruption-robust regret analysis for dueling bandits with K arms [Saha and Gaillard, 2022] or linear utilities [Di et al., 2024], neither their algorithms nor analysis can be applied to our setting with arbitrary concave utilities in continuous action space.

The full proof of Theorem 3.4.1 is quite involved and is deferred to Appendix B.4. Here, we outline the main ideas, divided into four key steps. Step 1-3 are applicable for analyzing

continuous dueling bandits under any form of corruption, whereas last Step 4 is the only step that hinges on the decaying corruption level assumptions of ρ-LIHF and is also the most involved and novel step.

Proof Sketch. The proof of Theorem 3.4.1 is divided into four major steps.

Step 1: quantifying bias of gradient estimation in ρ-LIHF

Our proof starts from quantifying the bias of the gradient $\hat{g}_t$ caused by ρ-Imperfect Human Feedback, as shown in the following Lemma 3.4.1.

Lemma 3.4.1 (Corrupted Gradients Estimation)**.** *The gradient $\hat{g}_t$ in Line 6 of Algorithm 3 satisfies*

$$\mathbb{E}\left(\hat{g}_t | a_t\right) = \nabla|_{a=a_t} \bar{P}_t(a) - \mathbb{E}\left(b_t | a_t\right).$$

where $\bar{P}_t(a) := \mathbb{E}_{x\in\mathbb{B}}\left[\mathbb{P}(a_t \succ a + \nabla^2 R_t(a_t)^{-\frac{1}{2}} x)\right]$ (with $\mathbb{B}$ as the unit ball) is the smoothed probability that a_t is preferred over any action a,
and $b_t = d\left(\mathbb{P}(a_t \succ a_t') - \hat{\mathbb{P}}(a_t \succ a_t')\right) \nabla^2 \mathcal{R}_t(a_t)^{\frac{1}{2}} u_t$.

Step 2: decomposing regret for dueling bandits into regret of decision and feedback error

Core to our proof is a "regret decomposion" lemma below that upper bounds the regret by two parts: regret of sub-optimal decision making under uncertainty, referred as regret of decision (first term in (3.2)), and regret due to corruption, referred as feedback error (second term in (3.2)).

Lemma 3.4.2 (Regret Decomposition for Dueling Bandits)**.** *The Reg_T of Algorithm 3 satisfies:*

$$Reg_T \le \underbrace{\frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^\mu L_1^\sigma R}_{\textit{Regret of Decision}} + \underbrace{2\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top (a_t - a_T^*)\right\}}_{\textit{Feedback Error}}, \tag{3.2}$$

where $C_0 := (2\nu + 4L_1^\sigma\kappa + 4R_2^\sigma (L^\mu)^2 + L^\mu L_1^\sigma \kappa)/\lambda$, and $a_T^* := \frac{1}{T}a_1 + (1-\frac{1}{T})a^*$.

Similar regret decomposition has been shown to be useful for reinforcement learning which has direct reward feedback [Foster et al., 2023]. However, to the best of our knowledge, Lemma 3.4.2 is the first to exhibit such a decomposition for dueling feedback, which contains much sparser information than direct reward feedback as in classic RL or online learning. Hence the lemma may be of independent interest for future research on dueling bandits under corruption or erroneous feedback.

Step 3: upper bounding the regret from feedback error

Bounding the regret of decision term in (3.2) is standard. Thus, this step develops a novel upper bound for bounding the regret from feedback error, as shown below.

Lemma 3.4.3 (Feedback Error)**.** *The feedback error term in* (3.2) *can be bounded as follows:*

$$\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top (a_t - a_T^*)\right\} \leq C_1 d\sqrt{\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^{T}\sqrt{t}|c_t(a_t, a_t')|\sqrt{\mu(a^*) - \mu(a_t)}\right\} + C_2 dT^\rho + 2d\sqrt{\lambda}\sqrt{\eta_\rho T},$$

where $C_1 := \sqrt{\frac{2\lambda(L^\sigma)^2}{\alpha}}$, *and* $C_2 := 2L^\sigma R\sqrt{\sup_{a\in\mathcal{A}}\lambda_{\max}\left(\nabla^2\mathcal{R}(a)\right) + 2\phi}$.

To our knowledge, such a bound is new for dueling bandits. The less common term in the bound is $\sqrt{\mu(a^*) - \mu(a_t)}$. It comes from the strong concavity of the utility function μ, which implies $\langle b_t, a_t - a^*\rangle \leq \|b_t\|_2\|a_t - a^*\|_2 \leq \|b_t\|_2\sqrt{\frac{2}{\alpha}(\mu(a^*) - \mu(a_t))}$.

Step 4: The *iterative regret refinement* analysis and resultant optimal learning rate

Our last and also the most intriguing step is to pin down the regret upper bound. Given the bounds from previous steps, the only remaining term to upper bound is $\mathbb{E}\left\{\sum_{t=1}^{T}\sqrt{t}|c_t(a_t, a_t')|\sqrt{\mu(a^*) - \mu(a_t)}\right\}$ in Lemma 3.4.3. We can upper bound $\sqrt{t}|c_t(a_t, a_t')|$ by $C_\kappa t^{\rho-\frac{1}{2}}$ under ρ-Imperfect Human Feedback assumption (this is the point where we start to need the decaying structure in Definition 3.2.1). This is a tricky question, since this term involving optimal action a^* closely connects to the regret Reg_T itself (see (3.1)). This is

precisely the motivation for our analysis approach of "iterative regret refinement". Concretely, it turns out that a good upper bound for the term $A := \mathbb{E}\left(\sum_{t=1}^{T} \mu(a^*) - \mu(a_t)\right)$ can be leveraged to derive a good upper bound for term $B := \mathbb{E}\left\{\sum_{t=1}^{T} t^{\rho-\frac{1}{2}}\sqrt{\mu(a^*) - \mu(a_t)}\right\}$, which can then be leveraged to derive a good upper bound for the regret Reg_T, which can be used to refine the original upper bound of term A, at which point we can re-apply the above refinement process.

Two key factors are essential for the iterative regret refinement mentioned above to result in a tight bound. The first step is to derive a good upper bound for term B using term A, which is shown in Lemma 3.4.4. The second step is to prove that this iterative refinement consistently improves the regret bound, eventually reaching a limit. This is demonstrated by Lemma 3.4.5. The choice of learning rate $\eta_\rho = \frac{\sqrt{\log T}}{dT^{\max\{1/2,\rho\}}}$ is crucial in the second step; in a sense, it is the only choice that converges to the tight regret upper bound $\tilde{\mathcal{O}}\left(d\max\{\sqrt{T}, T^\rho\}\right)$ in the limit, rather than exploding the bound.

Lemma 3.4.4. *If* $A = \mathbb{E}\left(\sum_{t=1}^{T}[\mu(a^*) - \mu(a_t)]\right) \leq \mathcal{O}(T^{\frac{1}{2}+\psi})$ *for some* $\psi \in [0, \frac{1}{2})$*, then we must have* $B = \mathbb{E}\left(\sum_{t=1}^{T} t^{\rho-\frac{1}{2}}\sqrt{\mu(a^*) - \mu(a_t)}\right) \leq \mathcal{O}(T^{-\frac{1}{4}+\frac{\psi}{2}+\rho})$.

To prove Lemma 3.4.4, we first bound term B by $\sum_{t=1}^{T} t^{\rho-1/2}\sqrt{\mathbb{E}[\mu(a^*) - \mu(a_t)]}$ using the concavity of the square root function. The key novelty of our proof is to use the Abel's equation [Williams, 1963] to re-arrange the term $\sum_{t=1}^{T} t^{\rho-1/2}\sqrt{\mathbb{E}[\mu(a^*) - \mu(a_t)]}$ and establish its connection with another term $C = \sum_{t=1}^{T}\sqrt{\mathbb{E}[\mu(a^*) - \mu(a_t)]}$. We finally connect term C with the term A in the lemma statement, by proving $\sum_{t=1}^{T}\sqrt{\mathbb{E}[\mu(a^*) - \mu(a_t)]} \leq \mathcal{O}(T^{\frac{3}{4}+\frac{\psi}{2}})$ if $\mathbb{E}\left(\sum_{t=1}^{T}\mu(a^*) - \mu(a_t)\right) \leq \mathcal{O}(T^{\frac{1}{2}+\psi})$ via concavity of square root. Together with the decay scale $t^{\rho-1}$, which allows $t^{\rho-1} - (t+1)^{\rho-1} \leq t^{\rho-2}$, we obtain the tight bound $\mathcal{O}(T^{-\frac{1}{4}+\frac{\psi}{2}+\rho})$.[4]

Armed with Lemma 3.4.4, we can use it to prove the induction claim (Lemma 3.4.5). The challenge in this proof lies in identifying the constant K, which must hold across all

4. If corruption c_t was arbitrary, then $t^{\rho-1/2}$ is $\sqrt{t}|c_t|$, and the bound could deteriorate to $\mathcal{O}(T^{\frac{2\rho}{3}})$ in the first T^ρ rounds.

iterations, k, of the analysis to ensure the bound remains stable and does not explode.

Lemma 3.4.5 (Induction Claim). *Consider* $T > \sqrt{2L^{\mu}L_1^{\sigma}R}$ *and* $\eta_\rho = \frac{\sqrt{\log T}}{dT^{\max\{1/2,\rho\}}}$. *If* $Reg_T \leq 144dK\sqrt{\log T}T^{\rho+\frac{\frac{3}{2}-\rho}{2^k}}$ *for some integer* k, *then we must also have* $Reg_T \leq 144dK\sqrt{\log T}T^{\rho+\frac{\frac{3}{2}-\rho}{2^{k+1}}}$. K *is a instance-dependent constant, where* $K := \max\left\{C_0, \frac{C_2C_\kappa}{\sqrt{\log T}}, (L^{\sigma}C_\kappa)^2, 2\sqrt{\lambda}RC_\kappa L^{\sigma}\right\}$, C_0 *and* C_2 *defined in Lemma 3.4.2 and 3.4.3 respectively.*

□

3.4.1 Additional Applications of Our Regret Analysis Framework

To showcase useful applications of the new regret analysis framework above for proving Theorem 3.4.1, in this subsection we employ the framework to study continuous dueling bandits under *arbitrary* and *agnostic* corruption — i.e., remove decaying constraints on c_t and ρ is unknown to the learner. We analyze natural variants of RoSIMD, which apply to strongly concave utilities, and the well-known algorithm Dueling Bandit Gradient Descent (DBGD) [Yue and Joachims, 2009b], which apply to general concave utilities. The proofs of these results are direct applications of the analysis from Steps 1-3 outlined above (without Step 4, since it is the only one requiring knowledge of ρ and decaying corruption). Additionally, we present these results to offer a set of principled benchmark algorithms for the experiments in Section 3.5. One thing to highlight is that our robustified algorithms retain the same computational complexity as the originals, ensuring added robustness without compromising efficiency. For further details, see Appendix B.7.1.

Proposition 2 (Efficiency-Robustness Tradeoff in RoSMID). *If the total corruption level satisfies* $\sum_{t=1}^{T}|c_t(a_t, a_t')| \leq \mathcal{O}(T^{\rho})$, *then for any* $\alpha \in [\frac{1}{2}, 1)$, *if the utility function is strongly concave, for a sufficiently large round* T, *by choosing learning rate* $\eta = \frac{\sqrt{\log T}}{2d}T^{-\alpha}$

1. *(Robustness) RoSMID incurs* $Reg_T \leq \tilde{\mathcal{O}}(dT^{\alpha} + \sqrt{d}T^{\frac{1}{2}(1-\alpha)+\rho} + dT^{\rho})$.

2. *(Efficiency) There exists a strongly concave utility function μ and a link function σ such that Reg_T suffered by RoSMID is at least $\Omega\left(T^{\alpha}\right)$ in scenario without corruption.*

Proposition 3 (Efficiency-Robustness Tradeoff in DBGD)**.** *If the total corruption level satisfies $\sum_{t=1}^{T}|c_t(a_t, a_t')| \leq \mathcal{O}(T^{\rho})$, then for any $\alpha \in (0, \frac{1}{4}]$, if utility function μ is generally concave, for a sufficiently large round T, choosing $\gamma = \frac{R}{\sqrt{T}}$ and $\delta = \frac{\sqrt{2Rd}}{\sqrt{13L^{\sigma}L^{\mu}T^{\alpha}}}$ for DBGD*

1. *(Robustness) Reg_T incurred by DBGD satisfies $Reg_T \leq \mathcal{O}(\sqrt{d}T^{1-\alpha} + \sqrt{d}T^{\alpha+\rho})$.*

2. *(Efficiency) There exists a linear utility function μ and a link function σ such that Reg_T suffered by DBGD is at least $\Omega\left(T^{1-\alpha}\right)$ in scenario without corruption.*

Both propositions illustrate the tradeoff between *learning efficiency* and *robustness* to adversarial corruption, achieved through their tunable learning rate $\eta_\rho, \gamma, \delta$ (learning rate in DBGD is $\frac{\gamma\delta}{d}$; we direct reader to Algorithm 6 for detail). Specifically, greater tolerance for agnostic corruption (i.e., larger α in Proposition 2) results in worse regret in non-corrupt settings. A similar robustness-accuracy trade-off has been studied in classification problems with adversarial examples, both empirically [Tsipras et al., 2018] and theoretically [Zhang et al., 2019]. However, to our knowledge, this is the first formal analysis of such a trade-off in online learning. While expanding the confidence region for UCB-type algorithms to handle agnostic corruption in linear contextual dueling bandits increases tolerance for corruption but worsens regret—sharing a similar idea of a trade-off [Di et al., 2023]—it's unclear if this approach necessarily reduces efficiency. In contrast, our work demonstrates that adjusting learning rates in gradient-based algorithms to enhance robustness against agnostic corruption inevitably decreases efficiency when no corruption is present. In particular, the efficiency statements in both propositions can be interpreted as algorithm-dependent lower bounds, concretely demonstrating that for certain problem instances, increasing robustness necessarily decreases the learning efficiency of the algorithm. In fact, the "efficiency-robustness tradeoff" is inherent under unknown adversarial attacks. Lemma B.5.1 in the appendix

shows that algorithms that learn faster tend to be more vulnerable to unknown corruption. However, the nature of this tradeoff varies across different algorithms. It is a very intriguing and fundamental open question to study what the optimal tradeoff is under unknown adversarial corruption and how to achieve it.

3.5 Experiments

In this section, we validate our theoretical analysis of RoSMID and DBGD under ρ-Imperfect Human Feedback through simulations. We also compare their performance against two baseline algorithms: Doubler and the heuristic algorithm, Sparring, proposed by Ailon et al. [2014]. For Sparring, we use bandit gradient descent [Flaxman et al., 2005] as the black-box bandit algorithm. We direct readers to Appendix B.7 for implementation details.

Experiment Setup. We consider a standard experiment setup, which adopts a strongly concave utility $\mu_\theta(a) := \theta^\top a - \frac{1}{2}\|a\|_2^2$, and a logistic link function $\sigma(x) = \frac{1}{1+\exp(-x)}$. We choose $d = 5$, and $T = 10^5$. Our action space $\mathcal{A}$ is a d-dimensional ball with radius $R = 10$. The preference parameter θ is randomly sampled from the surface of $\mathcal{A}$. In our problem setting, the optimal action $a^* = \theta$ and $\mu_\theta(a^*) = 50$. We simulate ρ-Imperfect Human Feedback for $\rho \in [0.5, 1]$.

Results and Discussion. Figure 3.1 presents a log-log plot of regret versus iteration, t, which the order of regret in T is represented by the slope of the line. Each line represents the average over five trials with different random seeds, and each line corresponds to a distinct value of ρ. The shaded region indicates $\pm$ one standard deviation. In the legend, "o" denotes the estimated order of regret. We estimate it using least squares on the last 1% of the data. Figure 3.1 supports Theorem 3.4.1, as for $\rho \in [0.5, 1)$, the estimated slope is less

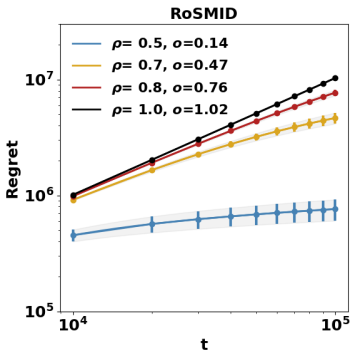

Figure 3.1: Robustness of RoSMID for ρ-LIHF

than ρ, suggesting that the choice of $\eta_\rho = \frac{\sqrt{\log T}}{dT^{\max\{1/2,\rho\}}}$ for RoSMID results in a regret bound of at most $\tilde{\mathcal{O}}(T^\rho)$.

Figure 3.2 compares the performance of our proposed algorithms, variants of DBGD and RoSMID, with two baseline algorithms, Sparring and Doubler, in the setting when c_t is arbitrary and agnostic. We set the parameter $\alpha = 1/4$ for DBGD and $\alpha = 1/2$ for RoSMID. When ρ is unknown to the algorithms, these parameter choices enable DBGD and RoSMID to tolerate $\mathcal{O}(T^{3/4})$ levels of unknown corruption, since the slope estimates are less than 1, implying sublinear regret, aligning with our theoretical predictions. Experimentally, RoSMID demonstrates the best performance under strongly concave utility functions. Sparring performs comparably to RoSMID, despite being a heuristic approach without theoretical guarantees. In contrast, Doubler exhibits the worst performance, likely because its theoretical guarantees are only applicable to linear link functions, which differs from our experimental setup. We direct readers to Appendix B.7 for experiment results with other choice of α for efficiency-robustness tradeoff.

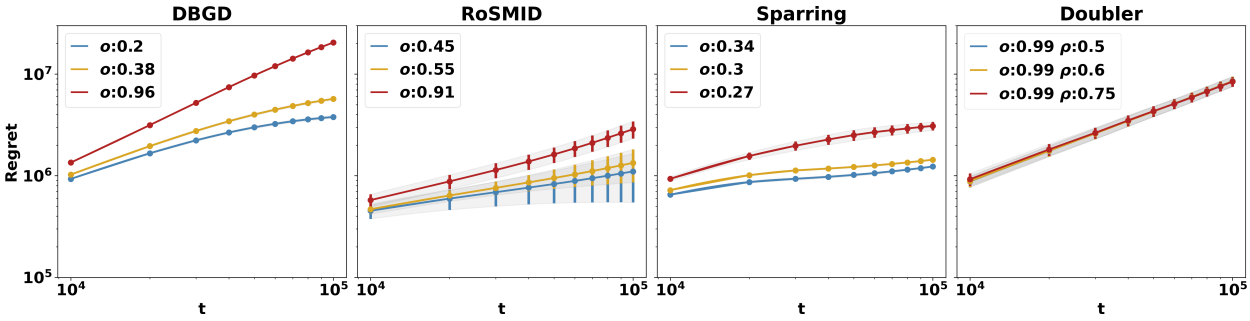

Figure 3.2: For each algorithm, we tested its performance under ρ-Imperfect Human feedback with $\rho = 0.5, 0.6, 0.75$. For each ρ, we presented a line plot of the average regret over five simulations, accompanied by $\pm$ one standard deviation shown by the shaded region. In the legend, o denotes the estimated line slope, calculated using least squares on the last 1% of the data.

3.6 Conclusion

In conclusion, this paper enhances the understanding of continuous dueling bandits by addressing ρ-Imperfect Human Feedback with concave utilities. We established fundamental regret lower bounds, introduced the RoSMID algorithm with nearly optimal performance, and developed a novel regret analysis framework that highlights the inherent efficiency-robustness trade-off in gradient-based algorithms. Our experimental results corroborate the theoretical predictions. A promising future direction is the development of an algorithm that achieves $\tilde{O}(d\sqrt{T} + dT^{\rho})$ regret for continuous dueling bandits with strongly concave utilities under arbitrary adversarial corruption. Additionally, incorporating contextual information to better model real-world scenarios could further enhance the applicability and robustness of these algorithms. We left these directions for future research.

CHAPTER 4

SINGLE-AGENT POISONING ATTACKS SUFFICE TO RUIN MULTI-AGENT LEARNING

4.1 Introduction

In recent years, multi-agent learning (MAL) systems have become increasingly prevalent, finding applications in diverse fields such as autonomous systems, distributed optimization, and economic markets [Leo et al., 2014, Shalev-Shwartz et al., 2016, Jin et al., 2018, Qiu et al., 2021, Zhou et al., 2021, Wu et al., 2022]. These systems, characterized by agents acting independently to maximize their own utilities in a non-cooperative environment, offer significant potential but also introduce unique vulnerabilities to potential adversarial attacks. Consider, for example, autonomous vehicles that communicate to coordinate traffic flow or financial trading algorithms that interact to optimize individual profits [Sharif and Marijan, 2022, Ataiefard and Hemmati, 2023, Shah et al., 2024]. In both cases, each agent (vehicle or trader) aims to maximize its own utility without direct collaboration. However, these systems can be highly susceptible to manipulation, as a single compromised or malicious agent can significantly influence the overall outcome. While it may be difficult for an adversary to directly attack or control multiple agents, targeting a single vulnerable or insider agent can be more feasible [Lin et al., 2020]. This raises an intriguing question of whether an adversary can still create a disruptive outcome by simply poisoning the utility of one agent.

This scenario—where a single agent in an MAL system is targeted for a utility poisoning attack—presents a new and critical challenge. Prior research has primarily focused on two related but distinct areas: adversarial attacks on individual agents in online learning environments [Bogunovic et al., 2021a, Jun et al., 2018, Garcelon et al., 2020], and many-agent attacks on MAL systems [McMahan et al., 2024, Guo et al., 2021, Wu et al., 2023, Ma et al., 2021, Zhang et al., 2023, Liu and Lai, 2023]. However, the impact of single-agent attacks

specifically in the multi-agent learning context remains underexplored. In this work, we address this gap by investigating the robustness of MAL dynamics against single-agent utility poisoning attacks. We show that even when one agent is compromised, the system can be disrupted, revealing a stark vulnerability that has been overlooked in prior studies.

To study the effect of single-agent poisoning in multi-agent systems, we use NE-finding dynamics in monotone games as our benchmark. Monotone games [Rosen, 1965], which include well-known models like Cournot competition [Cournot, 1838], Kelly auctions [Kelly et al., 1998], and Tullock contests [Tullock, 2008], represent a broad class of multi-agent strategic systems. In these games, agents' utility functions are concave, and many well-established MAL dynamics converge to their unique NE [Even-Dar et al., 2009, Farina et al., 2022, Bravo et al., 2018, Ba et al., 2024b, Tatarenko and Kamgarpour, 2020, Cai and Zheng, 2023], making them ideal benchmarks for studying the effects of single-agent poisoning on system-wide behavior.

Our theoretical results demonstrate that single-agent poisoning attacks can drive the dynamics away from the original NE, using an imperceptible, sublinear budget with respect to the time horizon. This holds for all MAL algorithms in this setting. We also investigate the robustness of two representatives of such MAL algorithms [Ba et al., 2024b, Bravo et al., 2018] and show that by sacrificing convergence speed, one can increase the system's tolerance to corruption, revealing a fundamental trade-off between efficiency and robustness.

Our contributions can be summarized as follows: (1) We propose a utility poisoning strategy that can mislead any MAL dynamics in strongly monotone games to converge to a new NE with a sublinear corruption budget, even when only targeting a single agent. (2) We analyze the robustness of MAL algorithms against general utility poisoning attacks, showing that adjusting the learning rate introduces inherent robustness. (3) Our findings uncover an efficiency-robustness trade-off in MAL dynamics, where faster-converging dynamics are more vulnerable to utility poisoning. To the best of our knowledge, this is the first work to

reveal this trade-off in multi-agent learning systems.

Related Work. Adversarial attacks on multi-agent learning—often termed "steering" or "policy teaching"—have been extensively studied under various assumptions about the attacker's knowledge, objectives, and strategies. For instance, Liu and Lai [2023] show that in certain Markov games, neither action poisoning nor reward poisoning alone can be both efficient and successful, even with complete knowledge of the environment ("white-box" setting), highlighting challenges in designing effective attacks. Similarly, Wu et al. [2023] investigate reward poisoning attacks on offline multi-agent reinforcement learners, demonstrating how adversaries can manipulate outcomes in offline settings.

Another line of research focuses on designing incentives to guide multi-agent dynamics toward a desirable equilibrium, known as "equilibrium steering" [Zhang et al., 2023, Canyakmaz et al., 2024, Huang et al., 2024]. Ma et al. [2021] consider game redesign where a designer, operating in a white-box setting, aims to induce players to choose targeted actions in a normal-form game using a sublinear cost, assuming the players employ no-regret learning. The most relevant work to ours is Zhang et al. [2023], which shows that a sublinear total incentive can induce a predetermined equilibrium against no-regret agents. However, their approach can only steer the game to an existing equilibrium and cannot achieve a non-equilibrium outcome without incurring a cost of $\Omega(T)$.

4.2 Preliminaries

Our study focuses on strategic games with a finite number of agents $[n] = \{1, \cdots, n\}$ and continuous action sets $\{\mathcal{X}_i\}_{i=1}^n$ associated with agents $i \in [n]$. During play, each agent i chooses an action (i.e., a pure strategy) $\boldsymbol{x}_i \in \mathcal{X}_i \subset \mathbb{R}^{d_i}$, where d_i is the dimension of action space of agent i, and forms the joint action profile $\boldsymbol{x} = (\boldsymbol{x}_i; \boldsymbol{x}_{-i}) \equiv (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$. We denote $d := \max_{i\in[n]} d_i$. Let $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$ be the game's strategy space, and each agent's reward is determined by a utility function $u_i : \mathcal{X} \to \mathbb{R}$. Without loss of generality we

assume the range of each u_i is a bounded region $[0, 1]$. We do not require agents know their utility functions and only assumes that they can observe the point-wise feedback $u_i(\boldsymbol{x})$, also known as *bandit* feedback in online learning literature. Such a game is denoted by a tuple $\mathcal{G} \equiv \mathcal{G}(n, \mathcal{X}, \{u_i\}_{i=1}^n)$. The commonly adopted solution concept for such strategic games is Nash equilibrium (NE), an action profile where no agent can deviate unilaterally to improve her utility. The formal definition of NE is given by the following:

Definition 4.2.1. *A joint strategy profile $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_n^*)$ forms a Nash equilibrium (NE) of a game $\mathcal{G}$, if for every agent i, $\boldsymbol{x}_i^*$ is a best response to the opponents' strategies $\boldsymbol{x}_{-i}^*$; formally,*

$$u_i(\boldsymbol{x}_i^*, \boldsymbol{x}_{-i}^*) \geq u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}^*) \quad \textit{for every } \boldsymbol{x}_i \in \mathcal{X}_i. \tag{4.1}$$

4.2.1 Monotone Games

Much of this paper's theoretical development focuses on an important class of games called *strictly monotone games* (strictly MG). This focus is due to at least two reasons. First, from computational perspective, strictly monotone games are one of the most general classes of continuous games that admit efficient algorithms for NE; it includes most popular continuous games such as Cournot competition [Cournot, 1838], Tullock contest [Tullock, 2008, He et al., 2024] and its variants that model content creation competitions [Yao et al., 2024a,b,c]. Additionally, our framework encompasses all strictly convex-concave zero-sum games and any game with a strictly concave potential. Second, strictly monotone games turn out to admit fast converging (to NE) multi-agent learning algorithms [Bravo et al., 2018]. Both conditions are important for our theoretical analysis.

The concept of "monotonicity", first introduced by Rosen [1965], refers to games where each agent has a concave utility function, along with a global condition known as diagonal strict concavity (DSC). A stronger version of the DSC condition defines a sub-class known as strongly monotone games. Formal definitions of these are provided below.

Definition 4.2.2 (Strictly/strongly monotone games)**.** *A continuous multi-agent game* $\mathcal{G}(n, \{\mathcal{X}_i\}_{i=1}^n, \{u_i\}_{i=1}^n)$ *is strictly monotone if for any* $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n), \boldsymbol{x}' = (\boldsymbol{x}'_1, \cdots, \boldsymbol{x}'_n) \in \mathcal{X}$*, it satisfies the diagonal strict concavity (DSC) property:*

$$\sum_{i=1}^{n} (\boldsymbol{x}'_i - \boldsymbol{x}_i)^\top (v_i(\boldsymbol{x}'_i, \boldsymbol{x}'_{-i}) - v_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})) < 0, \tag{4.2}$$

where $v_i(\boldsymbol{x}) = \nabla_{\boldsymbol{x}_i} u_i(\boldsymbol{x})$ *is the first-order derivative of* u_i *with respect to* x_i*. In addition, the game* $\mathcal{G}$ *is* β*-strongly monotone if* $\sum_{i=1}^n (\boldsymbol{x}'_i - \boldsymbol{x}_i)^\top (v_i(\boldsymbol{x}'_i, \boldsymbol{x}'_{-i}) - v_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})) \leq -\beta \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2$.

When discussing strongly monotone game in this paper, we treat β as a small constant and sometimes omit it in bounds unless there is a need to emphasize the dependence on β. Notably, the DSC property readily implies the strict concavity of u_k with respect to $\boldsymbol{x}_k$ for any fixed $\boldsymbol{x}_{-k}$. This can be observed by taking $\boldsymbol{x}' = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{k-1}, \boldsymbol{x}'_k, \boldsymbol{x}_{k+1}, \cdots, \boldsymbol{x}_n)$ in (4.2). As a result, the following best response mapping for each agent-k is well-defined[1]:

$$\text{Best response mapping:} \qquad BR_k(\boldsymbol{x}_{-k}) \triangleq \arg\max_{\boldsymbol{x}_k \in \mathcal{X}_k} u_k(\boldsymbol{x}_k, \boldsymbol{x}_{-k}). \tag{4.3}$$

Moreover, the following notion of game Hessian matrix will be useful throughout the paper:

$$\text{Game Hessian:} \qquad H_{\mathcal{G}} = [\nabla_j \nabla_i u_i(\boldsymbol{x})]_{i,j}. \tag{4.4}$$

The following smoothness property of agents' utility functions will be useful for our analysis.

Definition 4.2.3 (Second-order L-smooth utilities)**.** *Given any game* $\mathcal{G}$*, we say an agent* k*'s utility function* u_k *is second-order* L*-smooth if there exists a feasible set* $\mathcal{F} \subseteq \mathbb{R}^d$ *such that for any* $\boldsymbol{\delta} \in \mathcal{F}$*, the following function* $h_k(\cdot, \boldsymbol{\delta}) : \mathbb{R}^{nd} \to \mathbb{R}^d$ *defined as*

$$h_k(\boldsymbol{x}; \boldsymbol{\delta}) \triangleq \frac{v_k(\boldsymbol{x}_k + \boldsymbol{\delta}, \boldsymbol{x}_{-k}) - v_k(\boldsymbol{x}_k, \boldsymbol{x}_{-k})}{\|\boldsymbol{\delta}\|}, \tag{4.5}$$

1. Because the maximizer of a strictly concave function always exists and is unique.

is L-Lipschitz continuous; that is, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$*, it holds that*

$$\|h_k(\boldsymbol{x}; \boldsymbol{\delta}) - h_k(\boldsymbol{x}'; \boldsymbol{\delta})\| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|. \tag{4.6}$$

Second-order smoothness requires that, for a given agent k, while holding other agents' strategies fixed, the marginal change in k's utility with respect to her own strategy $\boldsymbol{x}_k$ is Lipschitz continuous with respect to the joint strategies $\boldsymbol{x}$ of all agents. This is a relatively mild assumption, as it essentially demands the smoothness of the second-order derivatives of the agent's utility. To see this connection, by letting $\boldsymbol{\delta} \to \mathbf{0}$, we have $h_k(\boldsymbol{x}; \boldsymbol{\delta}) \to \frac{\partial^2 u_k}{\partial \boldsymbol{x}_k^2} \cdot \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}$. Hence, in this case Condition (4.6) becomes the Lipschitz continuity on the Hessian matrix $\frac{\partial^2 u_k}{\partial \boldsymbol{x}_k^2}$ of h_k, which is why it is called second-order smoothness.

A few remarks are worth mentioning. First, our attack poison's only a single agent and its success is guaranteed so long as this single agent's utility is second-order smooth. Second, the smooth parameter L in Definition 4.2.3 turns out to affect how "controllable" the attack outcome is. Our results will show that clever adversaries would tend to poison an agent with small L parameter. Third, many well-known strongly monotone games exhibit this smoothness property. For example, the n-agent Cournot competition is second-order 0-smooth, where n-agent Tullock contest is second-order $\mathcal{O}(\sqrt{n})$-smooth (see Appendix C.1 for full descriptions of these games and proofs of these claims in Proposition 5 and 6).

4.2.2 The (α, p)-Multi-Agent Learning Dynamics under Bandit Feedback

Monotone games are known to have a unique Nash equilibrium, hence a substantial body of recent work develops equilibrium-converging dynamics, including [Bravo et al., 2018, Tatarenko and Kamgarpour, 2020, Cai and Zheng, 2023, Ba et al., 2024b]. These studies demonstrate that in a strictly monotone game if each agent independently follows such an algorithm, the joint strategies of all agents converge in last iterate to the NE at a polyno-

mial rate. Among these, we are particularly interested in algorithms that operate in a bandit feedback environment, where agents do not have access to their utility functions neither utility gradients, and can only observe the utilities (possibly noisy) resultant from the actions they take at that round. This feedback setting is challenging yet realistic, as it reflects un-coupled learning situations where agents are unaware of their opponents' existence and simply optimize their own utilities selfishly. The following notion is crucial to our analysis.

Definition 4.2.4 ((α, p)-MAL Dynamics)**.** *We say a (possibly randomized) multi-agent learning (MAL) dynamics $\mathcal{A} = (\mathcal{A}_1, \cdots, \mathcal{A}_n)$ for an n-agent game $\mathcal{G}$ with bandit feedback is an (α, p)-MAL dynamics if when each agent i uses algorithm $\mathcal{A}_i$ to determine her strategy $x_i^{(t)}$ at time t, then their joint strategy sequence $\{\boldsymbol{x}^{(t)}\}_{t=1}^{\infty}$ converges to a Nash Equilibrium $\boldsymbol{x}^*$ in the following sense: there exists positive constants C, T_0 such that for any $t > T_0$ we have*

$$\mathbb{E}[\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2^p] \leq C^p \cdot t^{-p\alpha}. \tag{4.7}$$

In other words, an (α, p)-MAL dynamics leads to last iterate convergence to some NE of the game $\mathcal{G}$ in the sense that the p-th power of the strategy profile difference, evaluated under the L_2 norm, has a polynomial convergence rate $\alpha > 0$. In the following section, we shall see that the α, p parameter in (α, p)-MAL algorithms fundamentally affects the algorithm's robustness to adversarial corruptions. In the literature of multi-agent learning for monotone games, various (α, p)-MAL dynamics have been developed. For instance, the **M**ulti-**A**gent **M**irror **D**escent with Bandit Feedback (MAMD) proposed by Bravo et al. [2018] leads to a $(\frac{1}{6}, 2)$-MAL algorithm, while the Multi-Agent **M**irror **D**escent with **S**elf-**C**oncordant **B**arrier Bandit Learning (MD-SCB) introduced in Ba et al. [2024b] leads to a $(\frac{1}{4}, 2)$-MAL dynamics. Notably, while these designs often let every agent use the same learning algorithm, neither our definition of the (α, p)-MAL dynamics nor our designed attack later require this restriction — all we assume is that the learning dynamics converge.

4.3 Vulnerability of MAL to Utility Poisoning in MG

In this section, we examine the vulnerability of multi-agent learning (MAL) dynamics. Our main result is the design of a single-agent utility poisoning attack, which is provably successful to any (α, p)-MAL algorithm for β-strongly monotone games. Our results also quantitatively characterize how the attack costs and outcome depend on parameters β, α, p. The risk of having a poisoned agent within a community is very realistic, hence our attack raises serious concerns regarding naively applying MAL dynamics to setting with potential adversaries. We conclude this section by showing how the attack can lead to worse outcomes when the adversary can poison even more agents.

4.3.1 Poisoning a Single agent Suffices to Steer the Equilibrium Away

We design an explicit implementation of such a utility poisoning attack, referred to as the **S**ingle-agent **U**tility **S**hifting **A**ttack (SUSA). SUSA first selects a victim agent k to poison, and computes a corrupted utility function $\tilde{u}_k$ for this agent. Then during MAL process, at each round SUSA simply shift the realized utility of agent k from u_k to the $\tilde{u}_k$ with attacking cost $c = |\tilde{u}_k - u_k|$ (hence the name "utility shifting"). The formal definition of SUSA is provided below.

Definition 4.3.1 (Single-agent Utility Shifting Attack)**.** *A Single-agent Utility Shifting Attack (SUSA) against a strategic game instance $\mathcal{G}$ is specified by a tuple $(k, \boldsymbol{\delta}, \Delta)$, which is implemented by the following two stages:*

1. ***[Preparation Stage]*** *Pre-compute a corrupted utility function for "victim agent" k:*

$$\tilde{u}_k(\boldsymbol{x}_k, \boldsymbol{x}_{-k}) = u_k(\boldsymbol{x}_k + \boldsymbol{\delta}, \boldsymbol{x}_{-k}) + \Delta(\boldsymbol{x}_{-k}, \boldsymbol{\delta}), \tag{4.8}$$

where $\Delta : \mathcal{X} \to \mathbb{R}$ is a function depending on the joint strategy $\boldsymbol{x}_{-k}$ and $\boldsymbol{\delta} \in \mathcal{X}_k$[2]*.*

2. Rigorously, we need to extend the domain $\mathcal{X}_k$ to ensure $u_k(\boldsymbol{x}_k + \boldsymbol{\delta}, \boldsymbol{x}_{-k})$ is well-defined. In fact, any

2. ***[Attacking Stage]*** *During the execution of an MAL algorithm at round t, add a corruption*

$$c_t = \tilde{u}_k(\boldsymbol{x}_k^{(t)}, \boldsymbol{x}_{-k}^{(t)}) - u_k(\boldsymbol{x}_k^{(t)}, \boldsymbol{x}_{-k}^{(t)})$$

to agent-k's utility observation.

In addition, we denote the corrupted game instance as $\tilde{\mathcal{G}}(k, \boldsymbol{\delta}, \Delta)$, in which the k-th agent's utility function is replaced with $\tilde{u}_k$.

A few remarks are worth clarifying. First, c_t is introduced after observing the current round action. This type of corruption is often referred as "strong" corruption in the literature [Liu and Shroff, 2019, Jun et al., 2018, Bogunovic et al., 2021b]. Second, SUSA is applicable to any MAL learning dynamics in any games with continuous utility functions, though below we only show that the success of such attacks can be guaranteed for strongly monotone games under certain conditions. Third, SUSA only poisons agent k's utility observations and does not interfere with her actions, though the poisoning amount depends on an action shifting term $\boldsymbol{\delta}$. Essentially, it makes k believe her utilities are drawn from a modified function $\tilde{u}_k$.

Notably, SUSA does *not* attack any other agents, but will nevertheless steer their equilibrium behaviors through influencing the victim k's behaviors. Since agents do not know their true utility functions but only observe the bandit feedback about utilities, this form of corruption not only disrupts the sequential strategy updates of the victim agent k, but also influences the dynamical behaviors of all other agents involved.

The following Lemma 4.3.1 reveals a nice property of SUSA. That is, if the target game $\mathcal{G}$ is strongly monotone, SUSA preserves the strongly monotone property as long as the l_2-norm of the deviation $\|\boldsymbol{\delta}\|_2$ is upper bounded by a constant depending on the game parameters.

Lemma 4.3.1. *Consider any β-strongly monotone game $\mathcal{G}(n, \{\mathcal{X}_i\}_{i=1}^n, \{u_i\}_{i=1}^n)$. If u_k is second-order L-smooth, then under SUSA the corrupted game $\tilde{G}(k, \boldsymbol{\delta}, \Delta)$ with $\|\boldsymbol{\delta}\|_2 < \beta/L$*

concave extension would suffice, and it does not affect the validity of our theoretical results.

remains strongly monotone.

Lemma 4.3.1 indicates that the corrupted game preserves strong monotonicity as long as the norm of the shifting offset $\boldsymbol{\delta}$ is reasonably bounded. This is useful because MAL dynamics retains convergence in the corrupted game (still strongly monotone). As mentioned in Section 4.2.1, Cournot competition is $\mathcal{O}(1)$-strongly monotone and 0-smooth, so a SUSA using any $\boldsymbol{\delta}$ preserves its monotonicity. In contrast, the Tullock contest is $\mathcal{O}(1)$-strongly monotone but $\mathcal{O}(\sqrt{n})$-smooth, hence the adversary is limited to using $\|\boldsymbol{\delta}\| < \mathcal{O}(1/\sqrt{n})$. This means that as the number of agents increases, it becomes more difficult for SUSA to preserve the monotone property. The proof of Lemma 4.3.1 can be found in Appendix C.2.1.

Our following main theorem of this section, which formally characterizes SUSA's capability.

Theorem 4.3.1. *Let $\mathcal{G}(n, \{\mathcal{X}_i\}_{i=1}^n, \{u_i\}_{i=1}^n)$ be any β-strongly monotone game (with unique NE $\boldsymbol{x}^*$) and $\mathcal{A}$ be any (α, p)-MAL dynamics. Suppose a victim agent-k's utility function u_k is second-order L-smooth. Consider SUSA$(k, \boldsymbol{\delta}, \Delta)$ constructed as*

$$\Delta(\boldsymbol{x}_{-k}, \boldsymbol{\delta}) = -u_k(BR_k(\boldsymbol{x}_{-k}), \boldsymbol{x}_{-k}) + u_k(BR_k(\boldsymbol{x}_{-k}) - \boldsymbol{\delta}, \boldsymbol{x}_{-k}) \tag{4.9}$$

where $\boldsymbol{\delta}$ is any vector satisfying $\|\boldsymbol{\delta}\|_2 < \beta/L$ and $BR_k(\boldsymbol{x}_{-k}) \triangleq \arg\max_{\boldsymbol{x}_k \in \mathcal{X}_k} u_k(\boldsymbol{x}_k, \boldsymbol{x}_{-k})$ is the best response mapping of agent-k. Then the resulting dynamics induced by $\mathcal{A}$ and SUSA$(k, \boldsymbol{\delta}, \Delta)$ converges to some $\tilde{\boldsymbol{x}}^*$, such that

1. ***[Attack Success]*** *The deviation to the original NE satisfies*

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*\|_2 \geq \beta \|\boldsymbol{\delta}\|_2 / \sup\{\rho(H_{\mathcal{G}}(\boldsymbol{x})) : \boldsymbol{x} \in [\boldsymbol{x}^*, \tilde{\boldsymbol{x}}^*]\}, \tag{4.10}$$

where $\rho(H_{\mathcal{G}}(\boldsymbol{x}))$ represents the spectral norm of the game's Hessian $H_{\mathcal{G}}$ at $\boldsymbol{x}$, and $[\boldsymbol{x}^*, \tilde{\boldsymbol{x}}^*]$ denotes the segment (with a slight abuse of notation) $\lambda\tilde{\boldsymbol{x}}^* + (1-\lambda)\boldsymbol{x}^*, \lambda \in [0, 1]$.

2. ***[Sublinear Attack Costs]*** *The expected total budget satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T}|c_t|\right] \leq C_0 \cdot T^{1-\frac{p\alpha}{p+1}}, \tag{4.11}$$

where the constant $C_0 = CL_1(4L_2+5)+2$, L_1 *is the Lipschitz constant of all* u_i *w.r.t.* $\boldsymbol{x}_i$ *and* L_2 *is the Lipschitz constant of victim agent* k*'s best response* BR_k *w.r.t.* $\boldsymbol{x}_{-k}$.

As described in (4.10) and (4.11), the success of a SUSA relies on achieving two goals simultaneously: (1) steering the convergence to a joint strategy profile that is at least a constant distance away from the original NE, and (2) ensuring that the total budget required is $o(T)$. Theorem 4.3.1 guarantees both even when the adversary only targets a single agent.

One might wonder, beyond the NE shifting distance guarantee (4.10), whether an adversary can have any control over the shifting direction, especially if it aims to influence specific agents' strategies. Our answer is affirmative: although rigorously characterizing the direction $\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*$ is challenging, we can approximate it as $H_{\mathcal{G}}^{-1}[:,k][\nabla_{kk}u_k]\boldsymbol{\delta}$ (see Appendix C.2.3), where $H_{\mathcal{G}}^{-1}[:,k]$ is the k-th block column of the inverse of $H_{\mathcal{G}}$, and $\nabla_{kk}u_k$ is the Hessian of u_k w.r.t. $\boldsymbol{x}_k$. This suggests that if an adversary possesses some additional global knowledge about the game's Hessian $H_{\mathcal{G}}$, it can further steer the NE deviation in a desired direction.

To get a better sense of the attack success guarantee, we derive constants in Theorem 4.3.1 for two examples as follows. The proof can be found in Proposition 5 and 6 in Appendix C.1.

Remark 4.3.1. *For* n*-person Cournot competition, the corresponding parameters* L_0, L_1, L_2, β *and* $H_{\mathcal{G}}$ *specified in Theorem 4.3.1 satisfy*

$$L_0 = 0, L_1 = \mathcal{O}(1), L_2 = \mathcal{O}(\sqrt{n}), \beta = \mathcal{O}(1), \sup_{\boldsymbol{x}\in\mathcal{X}}\{\rho(H_{\mathcal{G}}(\boldsymbol{x}))\} = \mathcal{O}(n),$$

therefore, for an arbitrary $\boldsymbol{\delta}$*, SUSA can induce an NE shift* $\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*\|_2 \geq \mathcal{O}(n^{-1})$*. For* n*-person Tullock contest, we have*

$$L_0 = \mathcal{O}(\sqrt{n}), L_1 = \mathcal{O}(1), L_2 = \mathcal{O}(\sqrt{n}), \beta = \mathcal{O}(1/\sqrt{n}), \sup_{\boldsymbol{x}\in\mathcal{X}}\{\rho(H_{\mathcal{G}}(\boldsymbol{x}))\} = \mathcal{O}(n),$$

and thus for $\|\boldsymbol{\delta}\| < \mathcal{O}(n^{-1})$*, SUSA can induce* $\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*\|_2 \geq \mathcal{O}(n^{-\frac{5}{2}})$*. Both results are obtained within a total budget* $\mathbb{E}\left[\sum_{t=1}^{T}|c_t|\right] \leq \mathcal{O}(\sqrt{n}T^{1-\frac{p\alpha}{p+\alpha}})$*.*

We observe that while some of the constants in the bounds are $\mathcal{O}(1)$, some of them inevitably depend on n and the specific guarantees vary across different game structures. In our two examples, Cournot games are clearly more vulnerable to attacks compared to Tullock games.

One might also wonder the induced deviation from the original NE may be too small when n gets large. As we will discuss in Section 4.3.2, the adversary can improve the deviation in (4.10) by removing the factor $\sup\{\rho(H_{\mathcal{G}})\}$ and may even force the NE deviation to an arbitrary direction if allowed to poison multiple agents.

Next, we discuss the implications of the necessary requirements in Theorem 4.3.1, which provide insight into why SUSA can be successful. The first condition demands that the l_2-norm of the deviation $\|\boldsymbol{\delta}\|$, must be bounded by a constant. This is to ensure that the corrupted game remains strongly monotone and thus allows $\mathcal{A}$ to converge to $\tilde{\boldsymbol{x}}^*$, the unique NE of the corrupted game, as indicated by Lemma 4.3.1. The second condition provides an explicit construction of the offset function, which is essential for maintaining a sub-linear total budget. In fact, the rationale behind the design of Δ is to satisfy that for the target agent-k and any $\boldsymbol{x}_{-k} \in \mathcal{X}_{-k}$, $\max_{\boldsymbol{x}_k\in\mathcal{X}_k}\tilde{u}_k(\boldsymbol{x}_k, \boldsymbol{x}_{-k}) = u_k(\tilde{BR}_k(\boldsymbol{x}_{-k}), \boldsymbol{x}_{-k})$, where $\tilde{BR}_k(\boldsymbol{x}_{-k}) = \arg\max_{\boldsymbol{x}\in\mathcal{X}_k}\tilde{u}_k(\boldsymbol{x}, \boldsymbol{x}_{-k}) = BR_k(\boldsymbol{x}_{-k}) - \boldsymbol{\delta}$ is the best response function for agent-k under the corrupted utility u_k. In other words, regardless of the opponents' strategies, the best response of any agent under their corrupted utility $\tilde{u}_k$ aligns with the

intersection of $\tilde{u}_k$ and the original utility u_k. This means that as an agent gradually learns the best response to their opponents' strategies–particularly when converging to $\tilde{\boldsymbol{x}}^*$–the adversary's budget decreases to zero, ultimately resulting in a total budget that is sub-linear in T. The proof for (4.10) requires establishing a connection between the Jacobian of best response mapping and the game's Hessian and applying Lagrange mean-value theorem for vector-valued functions [Hall and Newell, 1979]. The proof for (4.11) follows from a careful analysis of the expected budget $\mathbb{E}[|c_t|]$ at each round. In fact, we can show that the same convergence rate applies to the corrupted game $\tilde{G}$ and as the joint strategy approaches $\tilde{\boldsymbol{x}}$ at rate $t^{-\alpha}$, $\mathbb{E}[|c_t|]$ decreases at rate $t^{-\frac{p\alpha}{p+1}}$. For detailed proof, please refer to Appendix C.2.2.

4.3.2 Potential Power of Poisoning Many Agents

Theorem 4.3.1 demonstrates the significant impact of attacking even a single agent. A natural follow-up question is whether an adversary can gain additional power by poisoning multiple agents. While this is not the primary focus of our work—since it is unclear how realistic or feasible it would be for an adversary to poison many agents—we provide preliminary evidence that such an adversary would indeed have significantly more influence. We illustrate this through an example in the Cournot competition, as detailed below.

Proposition 4. *Under the same assumptions stated in Theorem 4.3.1, consider an adversary performing SUSA$(k, \boldsymbol{\delta}_k, \Delta_k)$ against all agents $k \in [n]$ in an n-person Cournot competition. Then, the required attacking budget for each agent is still bounded as (4.11), and the shifted NE under attack satisfies $\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^* = H_{\mathcal{G}}^{-1} D\boldsymbol{\delta}$, where $H_{\mathcal{G}}$ is the game Hessian defined in (4.4), and D is a diagonal block matrix with the i-th diagonal block being the Hessian of u_i w.r.t. $\boldsymbol{x}_i$. As a result,*

1. *picking $\boldsymbol{\delta} = D^{-1} H_{\mathcal{G}} \boldsymbol{v}$ can induce an arbitrary NE deviation direction $\boldsymbol{v}$,*
2. *picking $\boldsymbol{\delta}$ aligning with the direction associated with the largest eigenvalue of $H_{\mathcal{G}}^{-1} D$*

induces the largest possible NE deviation distance $\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*\|$*, which is at least* $\|\boldsymbol{\delta}\|$*.*

Although limited to Cournot games, Proposition 4 offers a preview of the additional advantages of attacking multiple agents, namely more refined control over both the magnitude and direction of NE deviation. Compared to Remark 4.3.1, attacking multiple agents increases the magnitude of the NE deviation from $\mathcal{O}(1/n)$ to $\mathcal{O}(1)$, and allows for arbitrary directional deviation. We are able to derive a closed-form solution of $\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*$ thanks to the quadratic utility function in Cournot games which yields a constant game Hessian and a linear best response mapping. For more general game structures, we can show that similar results hold, albeit in an approximate sense, as a strongly convex function can always be approximated by a quadratic form, and the best response mapping can be linearized near $\boldsymbol{x}^*$. However, a rigorous exploration of this is beyond the scope of our current work.

While it may be unrealistic to assume that all agents are vulnerable to attack, the proposition suggests that we can still expect a guarantee between these extremes, when only a subset of agents is subject to poisoning.

As we show in the detailed proof in Appendix C.2.2, if there is a flexibility to design $\boldsymbol{\delta}_i$ for all agents, the adversary can pick any $\boldsymbol{\delta}$ that aligns with the eigenvector associated with the smallest eigenvalue of $H_{\mathcal{G}}(\boldsymbol{x}^*)$ so that we can improve the term $\sup_{\boldsymbol{x}}\{\rho(H_{\mathcal{G}})\}$ in (4.10) to $\inf_{\boldsymbol{x}}\{\rho(H_{\mathcal{G}})\}$. Since for both Cournot and Tullock games, $\inf\{\rho(H_{\mathcal{G}})\}$ is $\mathcal{O}(1)$ and $\sup\{\rho(H_{\mathcal{G}})\} = \mathcal{O}(n)$, such flexibility enables the adversary to amplify the induced deviation by a factor of n in both cases. Additionally, attacking multiple agents can also induce deviation in arbitrary directions, regardless of the game structure. These additional guarantees are achieved within the same order of sublinear budget. While it may be unrealistic to assume that all agents are vulnerable to attack, the proposition suggests that we can still expect a guarantee between these extremes, when only a subset of agents is subject to poisoning.

Returning from the detour on the discussion of multi-agent attacks, we now focus back on our main setting of single-agent attacks.

Interestingly, we will demonstrate that simply tuning the learning rate can effectively balance these two objectives.

4.3.3 Additional Guarantees on Attack Success

In many scenarios, an adversary's primary objective is not merely to steer agents toward a new equilibrium $\tilde{\boldsymbol{x}}^*$. Instead, a more robust and practical goal is to reduce a specific target metric, $W(\boldsymbol{x})$ (e.g., social welfare), which depends on the PNE action profile $\boldsymbol{x}^*$ of all agents. In this section, we provide an additional guarantee of attack success for adversaries aiming to manipulate the target metric $W(\boldsymbol{x})$. Our Theorem 4.3.2 demonstrates that, irrespective of the game structure, for almost any function $W(\boldsymbol{x})$, the adversary can successfully deploy a SUSA to decrease the value of W at the newly induced PNE under attack.

Theorem 4.3.2 (Attackable Target Welfare Metrics)**.** *Consider any β-strongly monotone game*
$\mathcal{G}(n, \{\mathcal{X}_i\}_{i=1}^n, \{u_i\}_{i=1}^n)$ *with a (unique) PNE $\boldsymbol{x}^*$. Assume that for each i, u_i is second-order L-smooth and $\boldsymbol{x}^* \notin \partial\mathcal{X}$. Then for any first-order differentiable function $W(\boldsymbol{x})$ satisfying that*

$$\frac{dW}{d\boldsymbol{x}^*} \cdot H_{\mathcal{G}}^{-1}(\boldsymbol{x}^*) \cdot [\nabla_{ii} u_i(\boldsymbol{x}^*)]_{i=1}^n \neq \mathbf{0}, \tag{4.12}$$

there exists SUSA$(k, \boldsymbol{\delta}_k, \Delta)$ such that it strictly decreases W at its induced corrupted PNE, i.e.,

$$W(\tilde{\boldsymbol{x}}^*) < W(\boldsymbol{x}^*). \tag{4.13}$$

Here $\frac{dW}{d\boldsymbol{x}^*} = \frac{dW}{d\boldsymbol{x}}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*} \in \mathbb{R}^{1\times nd}$ denotes the derivative of W w.r.t. $\boldsymbol{x}$ at $\boldsymbol{x}^*$, $H_{\mathcal{G}}(\boldsymbol{x}^*) \in \mathbb{R}^{nd\times nd}$ is $\mathcal{G}$'s Hessian at $\boldsymbol{x}^*$, and $[\nabla_{ii} u_i(\boldsymbol{x}^*)]_{i=1}^n = \left[\frac{\partial^2 u_i(\boldsymbol{x})}{\partial \boldsymbol{x}_i^2}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*}\right]_{i\in[n]}^\top \in \mathbb{R}^{nd\times d}$ denotes the concatenation the Hessians of u_i w.r.t. $\boldsymbol{x}_i$ at $\boldsymbol{x} = \boldsymbol{x}^*$.

The proof of Theorem 4.3.2 is straightforward, as we can show when $\boldsymbol{x}^*$ is an interior point of $\mathcal{X}$, W is differentiable w.r.t. $\boldsymbol{\delta}_k$, so that one can always manipulate W as long

as its derivative w.r.t. $\boldsymbol{\delta}_k$ is non-zero at $\mathbf{0}$. The detailed proof can be found in Appendix **??**. Theorem 4.3.2 characterizes all manipulable metrics W—those for which the gradient at the original PNE $\boldsymbol{x}^*$ does not lie within the null space of $H_{\mathcal{G}}^{-1}(\boldsymbol{x}^*) \cdot [\nabla_{ii} u_i(\boldsymbol{x}^*)]_{i=1}^n$. This condition imposes at most d degrees of freedom on $\frac{dW}{d\boldsymbol{x}^*}$, leading to the conclusion that almost all target metrics W are manipulable.

The following corollary illustrates the extent to which the adversary can expect to decrease the target metric W by selecting an arbitrary (k, δ_k).

Corollary 1. *Under the same assumptions stated in Theorem 4.3.2, for any sufficiently small $\epsilon > 0$ and $\hat{\boldsymbol{\delta}}$ such that $\frac{dW}{d\boldsymbol{x}^*} \cdot H_{\mathcal{G}}^{-1}(\boldsymbol{x}^*)[:, k] \cdot \nabla_{kk} u_k(\boldsymbol{x}^*) \cdot \hat{\boldsymbol{\delta}} > 0$, $SUSA(k, \epsilon\hat{\boldsymbol{\delta}}, \Delta)$ induces*

$$W(\tilde{\boldsymbol{x}}^*) \leq W(\boldsymbol{x}^*) - \frac{\epsilon}{2} \left(\frac{dW}{d\boldsymbol{x}^*} \right) \cdot H_{\mathcal{G}}^{-1}(\boldsymbol{x}^*)[:, k] \cdot \nabla_{kk} u_k(\boldsymbol{x}^*) \cdot \hat{\boldsymbol{\delta}}. \tag{4.14}$$

Although Corollary 1 only provides a way to secure a small amount of decrease in W, an adversary can implement a sequence of SUSA actions to accumulate a more substantial reduction in W. Such a process still remains nearly imperceptible, as Theorem 4.3.1 ensures that the cost of each attack is negligible. However, a key challenge lies in accurately measuring the reduction in W, as this requires the knowledge of the inverse of the game's Hessian. Nevertheless, identifying an effective attack direction is relatively simple: (4.14) implies that if $\boldsymbol{\delta}$ cannot guarantee a decrease in W, then $-\boldsymbol{\delta}$ certainly can.

4.4 Intrinsic Trade-Off Between Efficiency and Robustness

Our main result in previous sections (i.e. Theorem 4.3.1) reveals an intriguing trade-off between the efficiency and robustness of MAL dynamics: the faster an algorithm converges (i.e., a larger α), the smaller corruption needed for inducing an NE shift. For simplicity, we name such outcome as NE Shifting Attack (NSA). While such a trade-off is well-documented in single-agent online learning [Cheng et al., 2024], we are the first to identify it in the

multi-agent context. In this section, we explore whether sacrificing the convergence rate of specific MAL dynamics can improve robustness against utility poisoning attacks. We focus on two dynamics: MD-SCB [Ba et al., 2024b], the fastest-converging state-of-the-art MAL algorithm, and MAMD [Bravo et al., 2018], the foundational MAL algorithm for strongly concave games with bandit feedback.

Essentially, our key insight is that the robustness of these dynamics can be improved (i.e., become more resilient to NSA) by simply adjusting the learning rate. Theorem 4.4.1 shows how the learning rate of MD-SCB (see Appendix C.3.1 for details) influences its robustness against general utility poisoning attacks. Originally proposed by Ba et al. [2024b], MD-SCB achieves an optimal convergence rate of $\mathcal{O}(d/\sqrt{T})$ without adversarial attacks, if choosing the learning rate $\eta_t \propto t^{-\frac{1}{2}}$ At each round t, we consider a general adversary who selects a set of agents $\tilde{\mathcal{I}}_t$ to attack: any agent $\tilde{i} \in \tilde{\mathcal{I}}_t$ will receive an altered utility observation $\tilde{\mu}_{\tilde{i}}^{(t)} \leftarrow \mu_{\tilde{i}}(\boldsymbol{x}_t) + c_{t,\tilde{i}}(\boldsymbol{x}_t)$. Theorem 4.4.1 highlights how the algorithm's convergence under such a generic utility poisoning attack depends on both the learning rate η_t and the total attack budget ρ.

Theorem 4.4.1. *Consider an adversary that attacks a set of agents $\tilde{\mathcal{I}}_t$ at round t with a total budget satisfying $\sum_{j=1}^{t}\sum_{\tilde{i}\in\tilde{\mathcal{I}}_j}|c_{j,\tilde{i}}| \leq \mathcal{O}(t^{\rho})$ for all t and some $\rho \in [0,1]$. By choosing a learning rate sequence $\eta_t = \frac{1}{2d}t^{-\phi}$, for sufficiently large T the last-iterate convergence rate of MD-SCB satisfies $\mathbb{E}\left[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^*\|_2^2\right] \leq \max\left\{\mathcal{O}(dT^{\phi-1}), \mathcal{O}(dT^{-\phi}), \mathcal{O}(dT^{\rho-\frac{1}{2}(\phi+1)})\right\}$. Specifically, $\mathbb{E}\left[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^*\|_2^2\right] \leq \mathcal{O}\left(dT^{\frac{2(\rho-1)}{3}}\right)$ can be achieved by choosing $\phi = \frac{2}{3}(\rho+\frac{1}{2})$.*

Theorem 4.4.1 delivers a strong message: for a given sublinear total budget $\mathcal{O}(T^{\rho})$, even as $\rho \to 1$, we can always adjust the learning rate decay to recover last-iterate convergence, making MD-SCB absolutely resilient to NSA[3], although at the cost of potentially slower convergence speed. The proof of Theorem 4.4.1 quantifies the bias introduced by attacks in the utility observations of the gradient estimator $\tilde{\boldsymbol{v}}_i^{(t)}$, and tracks the propagation of this

3. As NSA necessarily leads to divergence, last-iterate convergence must imply resilience to NSA.

bias throughout the convergence analysis (see Appendix C.3). We will later use a similar approach to establish an analogous result for Algorithm 9 as well.

Theorem 4.4.1 provides a potential defense strategy for an algorithm designer, assuming they can estimate the upper bound of the total corruption. To clarify the inherent trade-off between efficiency and robustness in MAL dynamics, we compare this result with our findings in Section 4.3 from a unified perspective. For this comparison, let ADV represent all possible utility poisoning strategies, and define the following quantity:

$$\begin{aligned}\rho(\mathcal{C},\mathcal{A}) &= \inf\{\rho\in[0,1]|\mathcal{C} \text{ with budget } \mathcal{O}(T^\rho) \textit{ always} \text{ achieve NSA against } \mathcal{A}.\},\\ &= \sup\{\rho\in[0,1]|\mathcal{C} \text{ with budget } \mathcal{O}(T^\rho) \textit{ cannot} \text{ achieve NSA against } \mathcal{A}.\},\end{aligned}$$

where $\mathcal{C}\in ADV$ represents a generic utility poisoning adversary. Now, a fundamental question from both the attacker and defender's perspectives is to quantify the following two functions (with $p=2$ fixed):

$$\rho_-(\alpha)=\inf_{\mathcal{C}\in ADV}\sup_{\mathcal{A}\in MAL(\alpha,2)}\rho(\mathcal{C},\mathcal{A}),\quad \rho_+(\alpha)=\sup_{\mathcal{A}\in MAL(\alpha,2)}\inf_{\mathcal{C}\in ADV}\rho(\mathcal{C},\mathcal{A}). \tag{4.15}$$

Function $\rho_-(\alpha)$ represents the optimal strategy of a utility poisoning adversary: for any MAL dynamics known to enjoy convergence rate α, $\rho_-(\alpha)$ denotes the minimum budget level ρ required for NSA. In contrast, $\rho_+(\alpha)$ describes the optimal defense from the perspective of the algorithm designer: given a required convergence rate α, $\rho_+(\alpha)$ identifies the most robust algorithm that can withstand NSA of level ρ from any utility poisoning adversary. The well-known max-min inequality also implies the following basic fact.

Fact 4.4.2. *$\rho_-(\alpha)\geq\rho_+(\alpha)$ for any $\alpha\in[0,1/4]$.*

Note that we restrict $\alpha\in[0,\frac{1}{4}]$ because $\alpha=\frac{1}{4}$ is proven to be the optimal convergence rate under $p=2$ in the non-corrupted setting [Ba et al., 2024b]. In general, one should expect strict inequality in this context (i.e., $\rho_-(\alpha)>\rho_+(\alpha)$), unless $\rho(\mathcal{C},\mathcal{A})$ has particular

special properties. For instance, if $\rho(\mathcal{C}, \mathcal{A})$ is convex in $\mathcal{C}$ and concave in $\mathcal{A}$, the equality must hold, as seen in the well-known minimax theorem from v. Neumann [1928], Sion [1958]. In our problem, however, both $\mathcal{C}, \mathcal{A}$ represent algorithms and it is unclear how ρ changes w.r.t. $\mathcal{C}$ and $\mathcal{A}$. Thus, an intriguing open question remains: does the equality hold in Fact 4.4.2? In other words, does the adversary gain a strict advantage (i.e., $\rho_-(\alpha) > \rho_+(\alpha)$) by first observing the dynamics $\mathcal{A}$ and then designing the attack $\mathcal{C}$?

To better understand the strength of the adversary, our main results in Theorems 4.3.1 and 4.4.1 provide insights by establishing upper and lower bounds of $\rho_-(\alpha), \rho_+(\alpha)$, respectively, as summarized in Corollary 2.

Corollary 2. *Given the definitions of $\rho_-(\alpha)$ and $\rho_+(\alpha)$ in (4.15), for some particular MAL algorithm $\mathcal{A}_0$ and adversary $\mathcal{C}_0 \in ADV$, we further define*

$$\rho(\mathcal{A}_0(\alpha)) = \inf_{\mathcal{C} \in ADV} \rho(\mathcal{C}, \mathcal{A} = \mathcal{A}_0(\alpha)), \quad \rho(\alpha, \mathcal{C}_0) = \sup_{\mathcal{A} \in MAL(\alpha, 2)} \rho(\mathcal{C} = \mathcal{C}_0, \mathcal{A}),$$

where $\mathcal{A}_0(\alpha)$ denotes algorithm $\mathcal{A}_0$ with a set of hyper-parameters that guarantees $MAL(\alpha, 2)$ convergence. Then for any $\alpha \in [0, \frac{1}{4}]$, it holds that

$$1 - \alpha \leq \rho(\textit{MD-SCB}(\alpha)) \leq \rho_+(\alpha) \leq \rho_-(\alpha) \leq \rho(\alpha; \textit{SUSA}) \leq 1 - \frac{2\alpha}{3}. \tag{4.16}$$

We defer the proof of Corollary 2 to Appendix C.4.1 as it is straightforward: the three middle inequalities follow directly from the definitions of ρ_-, ρ_+ and Fact 4.4.2. The two outer inequalities are derived from a restatement of Theorem 4.3.1 and 4.4.1. In fact, Theorem 4.3.1 quantifies the capability of a particular adversary (i.e., SUSA) and thus establish the upper bounds of $\rho(\alpha; \text{SUSA})$. Theorem 4.4.1, on the other hand, characterizes the robustness of a particular algorithm (i.e. MD-SCB), thus gives the lower bounds of $\rho(\text{MD-SCB}(\alpha))$.

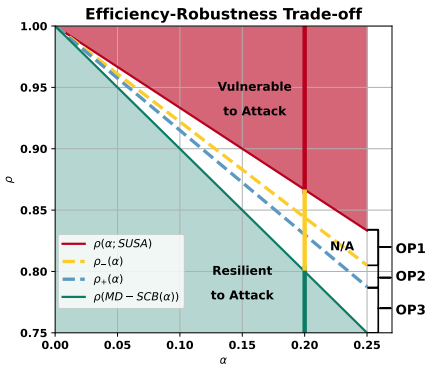

Figure 4.1: Illustration of the four quantities in Corollary 2, revealing the intrinsic efficiency-robustness trade-off under NSA.

Although the bounds in Corollary 2 are not tight, they highlight an intrinsic trade-off between efficiency and robustness: MAL dynamics with higher convergence rates become more vulnerable—they can withstand less corruption, and adversaries need a smaller budget to attack them. Figure 4.1 illustrates this trade-off. The yellow dashed line represents $\rho_-(\alpha)$; above it, a budget of $\mathcal{O}(T^\rho)$ suffices for a powerful adversary to induce NSA in any $MAL(\alpha, 2)$ dynamics. The red region, a subset of this area, corresponds to the actual budget required for NSA using our specific attack strategy, SUSA. Conversely, the blue dashed line represents $\rho_+(\alpha)$; below it, the most robust $MAL(\alpha, 2)$ dynamics can withstand corruption of $\mathcal{O}(T^\rho)$ against any adversary. For a specific MAL dynamics like MD-SCB, the green region shows the budget within which it remains resilient to NSA. We only present $\rho \in [0.75, 1]$ because the MAL dynamics with the best convergence rate in such a setting ($\alpha = 0.25$)[Ba et al., 2024b] can withstand a corruption level of $\rho = 0.75$ (see Theorem 4.4.1). The three gaps in Figure 4.1 correspond to three open problems: OP1. what is the most budget-efficient adversary for any MAL dynamics; OP2. does an adversary have a strict advantage by observing the dynamics $\mathcal{A}$ first and then design the attack $\mathcal{C}$; and OP3. what is the most robust MAL dynamics resilient to any adversary? We left these for future research.

Such an efficiency-robustness trade-off extends to more general contexts. For example, the MAL algorithm MAMD (detailed in Algorithm 9 in Appendix C.3.4) exhibits a similar phenomenon. We establish a corresponding result in analogous to Theorem 4.4.1 in Proposition 7 in Appendix C.3.4, showing how to adjust MAMD's learning rate to enhance its robustness. Moreover, this trade-off applies to a broader notion of robustness beyond just resilience to NSA attacks. In Appendix C.4, we present similar results for a stronger form

of robustness that not only ensures resilience to NE shifting but also prevents any possible dampening of the convergence speed.

4.5 Experiments

In this section, we validate our theoretical results from Theorem 4.3.1, Theorem 4.3.2, and Theorem 4.4.1 by implementing SUSA against MD-SCB on n-person Cournot games, in which each player-i has a utility function $u_i(x_i, x_{-i}) = x_i \left(a - b\sum_{j=1}^{n} x_j\right) - c_i x_i$. The formal definition is given in Appendix C.1. We also direct readers to Appendix C.5 for additional results for attacks against another algorithm MAMD (see Algorithm 9). We run simulations on two types of Cournot game instances. In the first type (homogeneous), all players share the same marginal cost c_i, resulting in a symmetric game. In the second type (heterogeneous), players have varying costs. The results for the homogeneous case are presented in the main paper, while those for the heterogeneous case are included in Appendix C.5.2, as they basically convey the same message.

***Homogeneous* n-person Cournot Game Experimental Setup:** We set $a = 10, b = 0.05$, and we consider the cost $c_i = 1$ for all agents $i \in [n]$. The action space is set to $\mathcal{X}_i = [0, 50]$ and the unique NE of such games can be verified as $\boldsymbol{x}_i^* = \frac{180}{n+1}, i \in [n]$. We let each agent run MD-SCB with the game size $n \in \{10, 50, 100\}$. The learning rate schedule of MD-SCB is set to $(\eta_t \propto t^{-\phi}, \phi = 0.5, 0.7, 0.9)$, corresponding to convergence rates $\alpha = 0.25, 0.15, 0.05, p = 2$ when no attack is present. We implement SUSA against agent 1, with a fixed $\delta = -10.0$. For each game instance specified by n and the attacked algorithm specified by the convergence rate α, we compare the dynamics' behavior both without attack and under SUSA, and report the actual total budget used.

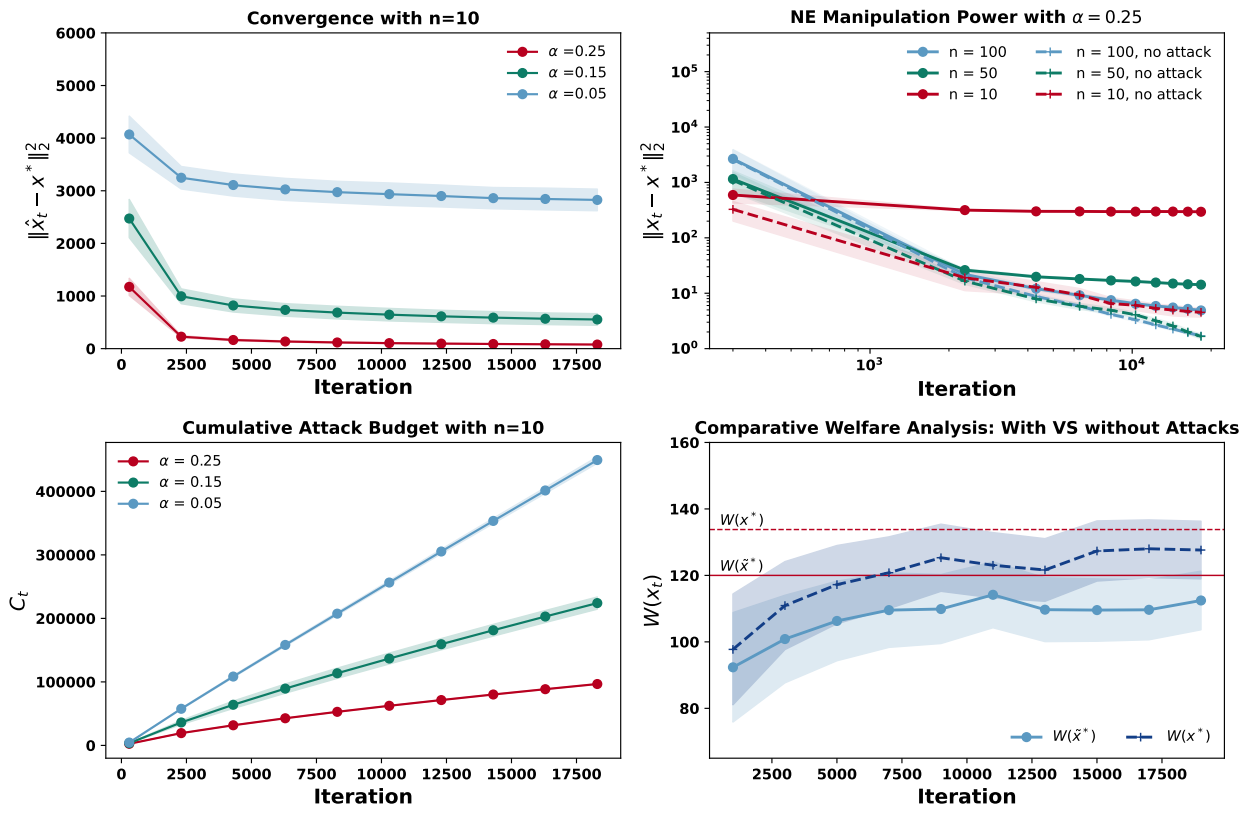

Figure 4.2: This figure displays results for *Homogeneous* n-person Cournot Game. Top left: square error of MD-SCB with varying convergence rates. Top right: square error of MD-SCB on different sizes of game instances against SUSA. Bottom left: cumulative attack budget used by SUSA against MD-SCB with varying convergence rates. Bottom right: social welfare $W(\boldsymbol{x}_t)$ under SUSA and without attack for $n = 10$. The red dashed line denote the social welfare under NE without attack, $W(x^*)$, while the red solid line denote the social welfare under manipulated NE under SUSA, $W(\tilde{x}^*)$. We plot 0.1-σ region from 20 independent simulations for social welfare (the bottom right figure) and 0.5-σ region for the rest figures.

Result: The top left panel of Figure 4.2 plots the convergence curve of the L_2 square error for $n = 10$, serving as a sanity check to confirm that different learning rate schedules induce different convergence rates for MD-SCB in absence of an adversary. The top right panel shows the outcome of SUSA against MD-SCB with $\alpha = 0.25$, with the y-axis representing L_2 distance between $\boldsymbol{x}_t$ and NE. The dashed lines display the convergence curve without the adversary, while the solid lines represent the dynamics under SUSA, with different colors indicating results for various game sizes n. As shown, for different values of n, the attacked

dynamics exhibit divergence, as the solid lines saturate and stop decreasing, indicating that the dynamics are being steered to converge to a new point. Additionally, we observe that the induced NE deviation decreases with respect to n, which aligns with our theoretical predictions in Theorem 4.3.1 and Remark 4.3.1. The bottom left panel shows the cumulative attack budget $C_t = \sum_{\tau=1}^{t} |c_\tau|$ at each time step t against MD-SCB with different values of α. As the results indicate, although all exhibit a sublinear trend, a faster-converging dynamics ($\alpha = 0.25$) requires a smaller attack budget, corroborating our findings in Theorem 4.3.1 and 4.4.1. The bottom right panel plots the social welfare $W(\boldsymbol{x}_t)$ without attack (indicted by the dashed line) and under SUSA(indicted by the solid line). The panel shows that SUSA with $\delta = -10$ can lead the learning dynamic to a new NE with a lower social welfare, aligning with the result in Theorem 4.3.2. These observations also hold for Cournot games with *heterogeneous* costs, and the corresponding results are presented in Appendix C.5.2.

4.6 Discussion

In addition to the three open problems regarding the fundamental limits of attack efficiency and defense robustness raised in Section 4.4, we discuss more potential limitations and future directions here. One limitation of our setting is that SUSA assumes the adversary has full knowledge of the victim agent's utility function—a potentially strong assumption in some cases. An important future direction is to broaden the applicability and practicality of SUSA from the current white-box setting to a black-box or grey-box setting, where the adversary has no or limited knowledge of the victim agent's utility. Another key question we did not address is the societal impact of utility poisoning attacks, particularly its potential for "steering for social good": as although SUSA could be used by adversaries to induce a harmful NE, it also opens up opportunities for a benevolent social planner to guide the system toward a better social outcome, simply by incentivizing a single agent. These intriguing open problems offer valuable avenues for future exploration.

CHAPTER 5

INFORMATION CONVEYANCE IN FINE-TUNED MODELS

5.1 Introduction

Large language models (LLMs) have achieved strong performance across tasks such as natural language understanding [Brown et al., 2020, Hendrycks et al., 2020], code generation [Chen et al., 2021b, Roziere et al., 2023], and mathematical reasoning [Ouyang et al., 2022, Lu et al., 2025]. A key driver of this success is fine-tuning, including instruction tuning [Wei et al., 2021, Chung et al., 2024] and alignment with human preferences [Ouyang et al., 2022, Rafailov et al., 2023]. While fine-tuning substantially improves helpfulness and task performance, a growing body of work suggests that it may also reduce output diversity and compresses the effective generation space of language models [Wang et al., 2025, Yang et al., 2025b, Lake et al., 2025, West and Potts, 2025]. Prior studies report that base models exhibit greater randomness and creativity, whereas aligned models produce more concentrated and similar responses across samples [West and Potts, 2025]. Similar conclusions have been drawn using branching factor [Yang et al., 2025b], lexical diversity [Lake et al., 2025], and token-level uncertainty measures [Agarwal et al., 2025, Wang et al., 2026], reinforcing the view that fine-tuning makes language models more deterministic and less diverse.

However, existing analyses suffer from an important limitation–they do not control for output length, despite length being a major confounder in diversity measurement [Shaib et al., 2024]. Classical metrics such as type-token ratio are biased by text length, often making shorter responses appear more diverse [Shaib et al., 2024]. Such metrics that ignore generation length cannot distinguish genuinely diverse outputs from statistical artifacts induced by shorter outputs. This limitation is especially problematic in aligned language models because fine-tuning changes not only what models generate, but also how long they generate. Base models often produce shorter responses, while instruction-tuned models generate longer and more structured outputs [Singhal et al., 2023]. As illustrated in Figure 5.2,

base models typically exhibit decreasing entropy rates over generation, implying that longer continuations become increasingly predictable and redundant. In contrast, fine-tuned models tend to maintain more stable entropy rates across longer trajectories, suggesting that additional tokens remain informative rather than collapsing into low-information continuation.

More importantly, prior work largely treats diversity as a static property of outputs, without studying how uncertainty evolves along generation trajectories. To the best of our knowledge, no existing work analyzes the relationship between generation length and entropy rate (see Definition 5.3.3)—that is, whether longer generations become more informative or increasingly redundant. Yet this interaction is crucial, since fine-tuning learns not only *what* to say, but also *when* to elaborate and *when* to stop. As a result, output length itself becomes part of the learned generation strategy, making diversity comparisons incomplete without accounting for how uncertainty is allocated across trajectories.

Meanwhile, recent studies suggest a more nuanced picture than simple diversity reduction. Fine-tuning can increase certain forms of diversity, such as semantic diversity in code generation, even while reducing lexical variation [Shypula et al.]. These findings suggest that fine-tuning may reorganize uncertainty and generation structure, rather than uniformly suppress diversity, leaving the relationship between uncertainty allocation, semantic variation, and generation trajectories still poorly understood.

Our contribution. In this work, we revisit diversity in language models through a principled information-theoretic lens. We propose *Canopy Entropy* (CE*), an information-theoretic measure that views language generation from a tree perspective. Just as a biological canopy represents the total spread and reach of a tree's branches, our metric views the "canopy" as the set of all potential leaves—or finished rollouts—available to the model. This allows CE* to naturally quantify the effective volume of the generation space by accounting for both how *widely* the model branches and how *deeply* it explores. CE* is introduced to jointly capture output length and content uncertainty, and we show that it admits an exact

characterization as the total Shannon entropy $H(N, Y_{1:N} \mid X)$, where X denotes the prompt, N is the generated length, and $Y_{1:N}$ is the generated sequence. Building on this formulation, we derive interpretable metrics and introduce a length–uncertainty correlation term that measures how uncertainty is allocated across generation trajectories (see section 5.3). Since $\mathrm{CE}^\star$ exhausts the entire space of rollouts, a key challenge is to estimate it efficiently. Towards that end, we develop Monte Carlo estimators and establish their consistency (see section 5.4).

Empirically, we evaluate multiple model families across sentence completion, mathematical reasoning, coding, and story generation, while explicitly controlling for output length. Our results reveal three consistent findings: **(i)** fine-tuning generally reduces token-level and trajectory-level uncertainty, although the magnitude of reduction is highly task-dependent; **(ii)** fine-tuning systematically shifts $\rho(N, r_N)$ from negative toward positive, indicating that longer generations become more informative rather than being redundant as often in base models; and **(iii)** entropy rate becomes a substantially stronger predictor of semantic diversity after fine-tuning, suggesting that aligned models utilize uncertainty more efficiently and translate it more effectively into meaningful semantic variation (see section 5.5). Overall, our results suggest that fine-tuning does not simply reduce uncertainty or shrink the generation space. Rather, it restructures how uncertainty is allocated across generation trajectories, producing outputs that are more organized, semantically meaningful, and information-efficient.

5.2 Related work

Length as a fundamental confounder in diversity measurement. Recent work [Singhal et al., 2023] shows that RLHF and instruction tuning systematically increase LLM output length, since longer responses are often positively correlated with human preference scores. This makes output length a critical confounder in diversity evaluation. Metrics that favor shorter generations may therefore unfairly penalize instruct models simply because they produce longer responses. For example, n-gram diversity metrics [Li et al., 2016] naturally decrease with sequence length, as longer texts inevitably reuse a finite vocabulary, reducing

the ratio of unique n-grams even when the underlying content remains semantically diverse. Embedding-based semantic diversity measures [Tevet and Berant, 2021] suffer from a related issue: longer generation trajectories must be compressed into fixed-dimensional representations, often through mean pooling, which can dilute semantic variation. Moreover, many existing diversity metrics rely on heuristic sampling procedures and lack a unified probabilistic foundation [West and Potts, 2025]. Consequently, current approaches cannot rigorously distinguish whether the increased verbosity of instruct models reflects genuine information gain or merely redundant continuation, motivating the need for a principled length-aware and model-intrinsic measure of generation diversity.

From local branching to global canopy. Conceptualizing autoregressive generation as a stochastic, path-dependent branching tree, Branching Factor (BF) [Yang et al., 2025b] was introduced to quantify the probability concentration induced by model alignment. However, their framework lacks a formal heuristic and focuses primarily on the *width* of the generation tree (branching) while ignoring its *depth* (length) and the coupling between the two. CE* addresses this gap by providing a model-intrinsic characterization of the generation space by jointly modeling content uncertainty and stochastic stopping behavior. By deriving the BF as a normalized version of CE*, we provide a principled theoretical foundation for its use. Crucially, our length-uncertainty correlation demonstrates that while fine-tuning reduces the absolute BF, it restructures the generation tree to sustain information density across longer paths while translating uncertainty into semantic variation more efficiently. To the best of our knowledge, this is the first work to provide a unified measure of the global LLM generation space that accounts for both the local branching uncertainty and the stochasticity of trajectory length, while also studying the redistribution of uncertainty across generation trajectories.

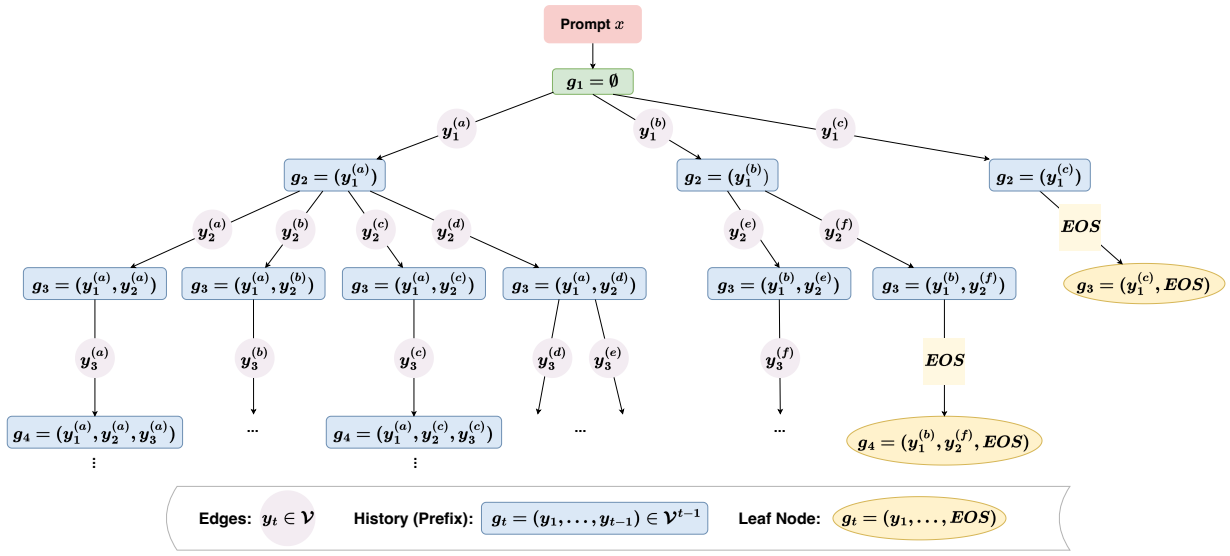

Figure 5.1: Path-dependent generation tree induced by autoregressive LLMs.

5.3 Methodology

5.3.1 Measuring the size of LLM generation space from a tree perspective

Autoregressive generation as a stochastic tree. Autoregressive language models induce a stochastic generation process that can naturally be represented as a rooted, path-dependent tree. Let $X \sim p_X$ denote a prompt sampled from a prompt distribution over a space $\mathcal{X}$, and let $\mathcal{V}$ denote the vocabulary. Conditioned on $X = x$ and a decoding policy (e.g., temperature, top-k, or nucleus sampling), generation proceeds sequentially by sampling tokens from a conditional distribution. Formally, let $Y_1, Y_2, \ldots,$ denote the autoregressive token process generated by the model. For each step $t \geq 1$, define the generation history as $G_t := (Y_1, \ldots, Y_{t-1})$, with $G_1 := \emptyset$. The next-step random variable is denoted by $Z_t \in \mathcal{Z} := \mathcal{V} \cup \{\texttt{EOS}\}$, where EOS denotes termination. The conditional distribution $q(Z_t \mid G_t, X)$ governs the next-step generation process.[1] Under this construction, each node of the generation tree (see Figure 5.1) corresponds to a history G_t. Each outgoing edge corresponds to a

1. This formulation is equivalent to a two-stage stochastic decision process Specifically, the model first makes a binary decision between STOP (terminate) and CONT (continue). Conditional on CONT, it then samples the next token from the vocabulary $\mathcal{V}$. We provide a formal proof of this equivalence in subsection D.1.1.

possible realization of Z_t. A complete generated sequence therefore corresponds to a root-to-leaf trajectory. A key feature of this construction is its path dependence. The next-step distribution q depends on the entire generation history G_t, not merely the most recent token. Thus, two trajectories ending with the same token may evolve very differently if they arise from different histories.

Width and depth of the generation tree. The size of the generation tree is determined by two complementary factors. First, at each history G_t, the spread of $q(\cdot \mid G_t, X)$ determines the number of plausible continuations, capturing the local branching behavior (*width*) of the tree. Second, generation proceeds for a random number of steps before termination, inducing variability in trajectory depth. We define the stopping time $N := \inf\{t \geq 1 : Z_t = \texttt{EOS}\}$, which represents the generation length (*depth*). Under this formulation, the stochastic process jointly models both continuation uncertainty and stopping behavior, together determining the overall size of the generation space.

A unified measure of generation space. To quantify local uncertainty, we use Shannon entropy [Shannon, 1948b]. For each generation history G_t, define $\widetilde{H}(G_t \mid X) := H(Z_t \mid G_t, X) = -\sum_{z \in \mathcal{Z}} q(z \mid G_t, X) \log q(z \mid G_t, X)$. This quantity measures uncertainty in the next generation step, including both continuation-token selection and termination. To jointly account for local branching diversity and random trajectory length, we define the following quantity.

Definition 5.3.1 (Canopy Entropy)**.** *We define* Canopy Entropy *as* $\mathrm{CE}^\star := \mathbb{E}\left[\sum_{t=1}^{N} \widetilde{H}(G_t \mid X)\right]$. *Equivalently,* $\mathrm{CE}^\star = \mathbb{E}_X\left[\mathbb{E}\left[\sum_{t=1}^{N(X)} H(Z_t \mid G_t, X = x) \,\middle|\, X\right]\right]$. *Conditioned on the prompt* X*, length* $N(X)$ *and histories* $\{G_t\}_{t=1}^{N(X)}$ *are random variables induced during the generation process.*

The quantity $\mathrm{CE}^\star$ aggregates uncertainty along generation trajectories and across prompts. A single generated sequence corresponds to a specific traversal from the tree root (prompt X) to a leaf (`EOS` termination). The summation $\sum_{t=1}^{N(X)}$ thus represents the total uncertainty

accumulated along that entire path. By taking the expectation over all trajectories, $\mathrm{CE}^\star$ captures the aggregate "canopy" of the model's output distribution across all possible roll-outs. As such, it provides a unified measure of generation space by accounting for both *local branching width* (the diversity of choices at each node) and *trajectory depth* (the variability in generation length).

$\mathrm{CE}^\star$ as the joint entropy of length and content. The above construction is not only intuitive from a tree perspective, but also admits an exact information-theoretic characterization. In particular, the cumulative uncertainty along a trajectory can be interpreted as the total entropy of the generation process. For ease of reference, we define *total uncertainty*: $H(N, Y_{1:N}|X)$, *length uncertainty*: $H(N|X)$, and *content uncertainty*: $H(Y_{1:N} \mid N, X)$. We now formalize the connection.

Theorem 5.3.1 (Equivalence to total uncertainty)**.** $\mathrm{CE}^\star = H(N, Y_{1:N} \mid X)$.

Theorem 5.3.1 (see subsection D.1.2 for a detailed proof) shows that the Canopy Entropy coincides with the total uncertainty of the generation process. By the chain rule of entropy, $H(N, Y_{1:N} \mid X) = H(N \mid X) + H(Y_{1:N} \mid N, X)$, which decomposes the total uncertainty into two components. The *length uncertainty* $H(N \mid X)$ captures how unpredictable the stopping time is, while the *content uncertainty* $H(Y_{1:N} \mid N, X)$ captures how diverse the generated sequences are once the length is fixed. This decomposition aligns naturally with the tree perspective: length uncertainty corresponds to variability in the depth of trajectories, while content uncertainty corresponds to the diversity of paths of a given length. We can interpret $\exp(\mathrm{CE}^\star)$ as the effective number of plausible full sequences, analogous to how perplexity represents the effective branching at the token level.

5.3.2 Token-level and trajectory-level normalization of Canopy Entropy

While $\mathrm{CE}^\star = H(N, Y_{1:N} \mid X)$ provides a complete characterization of generation space, it aggregates uncertainty across both sequence length and content variability, making direct

interpretation difficult. To obtain more interpretable quantities, we introduce two normalized measures derived naturally from $\text{CE}^\star$, which are Generation Perplexity (GenPPL) and Branching Factor (BF) [Yang et al., 2025b].

GenPPL and BF capture uncertainty from complementary perspectives. GenPPL (see Definition 5.3.2) is *token-centric*, measuring average uncertainty per generation step, while BF (see Definition 5.3.3) is *trajectory-centric*, capturing diversity at the sequence level. Together, they provide a more complete view of generation behavior across tasks. In particular, structured tasks such as mathematical reasoning or coding are better characterized by token-level uncertainty (GenPPL), whereas open-ended tasks like story generation are better captured by trajectory-level diversity (BF). Our framework provides a principled formulation of GenPPL as a generative analogue of perplexity [Jelinek et al., 1977] for variable-length processes, and offers a unified information-theoretic interpretation of BF, initially proposed by Yang et.al [Yang et al., 2025b], showing that both arise naturally as normalized forms of Canopy Entropy.[2]

Definition 5.3.2 (GenPPL). *Define* GenPPL $:= \exp\left(\frac{\sum_{X\sim p_X} H(N,Y_{1:N}|X=x)}{\sum_{X\sim p_X} \mathbb{E}[N|X=x]}\right)$. *GenPPL captures uncertainty from a* token-centric *perspective. It measures the effective number of plausible continuations maintained at a typical generation step, weighting each token equally across all trajectories.*

Definition 5.3.3 (BF). *Define* BF $:= \exp\left(\mathbb{E}_X \mathbb{E}_N[r_N \mid X = x]\right)$, *where* $r_N := \frac{H(Y_{1:N}|N=n,X=x)}{n}$ *is the entropy rate (i.e., per-token content uncertainty).*[3] *BF captures uncertainty from a* trajectory-centric *perspective. It measures the effective branching behavior of a typical generated trajectory by assigning equal weight to each sequence regardless of length.*

2. See Appendix D.2.1 for guidance on when to use BF or GenPPL, and Appendix D.2.2 for a comparison with perplexity. For further intuition, we refer readers to subsection D.2.3 for a concrete example illustrating the distinction between GenPPL and BF.

3. In information theory, the entropy rate quantifies the average amount of information produced per generated token by a stochastic process, and can be interpreted as a measure of information density or non-redundancy.

5.3.3 How entropy evolves with generation length

Length–uncertainty correlation. While aggregate measures such as Canopy Entropy, GenPPL, and BF quantify the overall magnitude of uncertainty, they do not capture how uncertainty evolves across generation rollouts or how efficiently uncertainty is translated into meaningful variation as generation length increases. In particular, existing diversity metrics [Tevet and Berant, 2021, Li et al., 2016, Friedman and Dieng, 2022] cannot distinguish whether longer generations remain information-rich or become progressively redundant [Singhal et al., 2023, Shaib et al., 2024]. Understanding this interaction is crucial for characterizing the *information-conveying efficiency* of language models—namely, how much uncertainty is maintained per token as generation progresses.

To study this phenomenon, we introduce the prompt-controlled length–uncertainty correlation $\rho_{\mathrm{pc}}(N, r_N)$ (see Definition 5.3.4), which measures how entropy rate varies with generation length. A positive correlation indicates that longer sequences remain informative per token, while a negative correlation indicates that longer outputs become increasingly predictable. This quantity serves as a trajectory-level characterization of uncertainty allocation and is later used in section 5.5 to compare base and fine-tuned models across tasks and model families.

Definition 5.3.4 (Prompt-controlled length–uncertainty correlation)**.** *For a fixed prompt X, define the prompt-conditional correlation* $\rho_X := \rho(N, r_N \mid X = x) = \frac{\mathrm{Cov}(N, r_N \mid X=x)}{\sqrt{\mathrm{Var}(N \mid X=x)\,\mathrm{Var}(r_N \mid X=x)}}$. *Since correlation is bounded and nonlinear, we aggregate correlations on the Fisher-z scale.*[4] *Then the prompt-controlled correlation is* $\rho_{\mathrm{pc}} := \tanh\left(\frac{\mathbb{E}_{X\sim p_X}[w_X \operatorname{atanh}(\rho_X)]}{\mathbb{E}_{X\sim p_X}[w_X]}\right)$, *where* $w_X \geq 0$ *is a prompt-specific reliability weight.*

4. This transformation maps correlation coefficients from $[-1, 1]$ to $\mathbb{R}$ and stabilizes their variance, making them approximately normally distributed with variance $1/(n_p - 3)$. n_p is the number of rollouts associated with each prompt p. Directly averaging correlations would lead to biased estimates due to the nonlinear and bounded nature of correlation coefficients. In addition, we report Spearman's rank correlation and Kendall's τ to capture monotonic dependence beyond linear effects. We refer the reader to subsection D.1.6 for the corresponding prompt-controlled estimation and aggregation procedures.

5.4 Consistent estimation under finite sampling with length constraint

In the previous section, we establish that $\mathrm{CE}^\star = H(N, Y_{1:N} \mid X)$ along with its normalized variants. However, it is not directly computable in practice. Generation is constrained by a finite token budget (requiring a maximum length), and the expectation is taken over exponentially many trajectories [Hu et al., 2023, Yang et al., 2025b]. These challenges motivate a tractable estimation approach. We therefore introduce a Monte Carlo estimator under a maximum length constraint $T_{\max}$, and analyze its consistency.

For each prompt X, let $N(X)$ be the generation stopping time. Given a maximum length $T_{\max}$, define the truncated stopping time $N_{\max}(X) := \min\{N(X), T_{\max}\}$. The corresponding truncated functional is $\mathrm{CE}^\star_{\max} =: \mathbb{E}\left[\sum_{t=1}^{N_{\max}(X)} \widetilde{H}(G_t \mid X)\right]$. Now suppose $X_1, \ldots, X_P$ are sampled i.i.d. from p_X, the prompt distribution. For each prompt X_p, we generate M independent rollouts. For the i-th rollout, define the truncated cumulative local entropy by $S^{(p,i)}_{\max} := \sum_{t=1}^{N^{(p,i)}_{\max}} \widetilde{H}(G^{(p,i)}_t \mid X_p)$, which measures the total branching uncertainty along the generated trajectory up to truncation. We then define the Monte Carlo estimator $\widehat{\mathrm{CE}}^\star_{\max,P,M} := \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M} S^{(p,i)}_{\max}$ (see algorithm 10) with convergence behavior in Theorem 5.4.1.

Theorem 5.4.1 (Consistency and non-asymptotic error bounds under prompt distribution)**.** *Assume there exists a deterministic function $g : \mathbb{N} \to [0, \infty)$, with $g(T) \to 0$ as $T \to \infty$, such that for every $T_{\max} \in \mathbb{N}$, $0 \le \mathrm{CE}^\star - \mathrm{CE}^\star_{\max} \le g(T_{\max})$. Let $\nu^2_{\max} := \mathbb{E}\left[N_{\max}(X)^2\right]$. If $T_{\max} = T_{\max}(P, M) \to \infty$ and $\nu_{\max}\sqrt{\frac{1}{P} + \frac{1}{PM}} \to 0$, then $\widehat{\mathrm{CE}}^\star_{\max,P,M}$ converges to $\mathrm{CE}^\star$ in probability when $P, M \to \infty$. Moreover, $\left|\widehat{\mathrm{CE}}^\star_{\max,P,M} - \mathrm{CE}^\star\right| = O_{\mathbb{P}}\left(\nu_{\max}\sqrt{\frac{1}{P} + \frac{1}{PM}} + g(T_{\max})\right)$.*

Remark 5.4.1 (Connecting $g(T_{\max})$ to the prompt-conditional tail of N)**.** *Define the trun-*

cation bias

$$B_{T_{\max}} := \mathrm{CE}^\star - \mathrm{CE}^\star_{\max} = \mathbb{E}_{X\sim p_X}\left[\mathbb{E}\left[\sum_{t=T_{\max}+1}^{N(X)} \widetilde{H}(G_t \mid X)\mathbf{1}\{N(X) > T_{\max}\}\middle| X\right]\right] \quad (5.1)$$

We have $B_{T_{\max}} \le H_{\max}\mathbb{E}_{X\sim p_X}\left[\mathbb{E}\left[(N(X) - T_{\max})_+|X\right]\right]$, *since* $\widetilde{H}(G_t \mid X) \le H_{\max}$ *(the entropy upper bound). By the tail-sum formula,* $\mathbb{E}\left[(N(X) - T_{\max})_+|X = x\right] = \sum_{t=T_{\max}+1}^{\infty} \mathbb{P}(N(x) \ge t \mid X = x)$. *Therefore, take* $g(T_{\max}) = H_{\max}\mathbb{E}_{X\sim p_X}\left[\sum_{t=T_{\max}+1}^{\infty} \mathbb{P}(N(X) \ge t \mid X)\right]$, $H_{\max} := \log|\mathcal{Z}|$, *showing that the truncation bias is determined by the prompt-averaged tail behavior of the stopping time. We refer readers to subsection D.1.4 for examples of prompt-averaged bias decay rates.*

In Theorem 5.4.1, we show that the estimator $\widehat{\mathrm{CE}}^\star_{\max,P,M}$ is consistent, with estimation error naturally decomposing into a Monte Carlo error term and a truncation bias term. The Monte Carlo error is of order $O_{\mathbb{P}}\left(\frac{1}{\sqrt{P}} + \frac{1}{\sqrt{PM}}\right)$, arising from sampling P prompts and M rollouts per prompt. The second component is a truncation bias $g(T_{\max})$, which captures the bias introduced by truncating trajectories at length $T_{\max}$ instead of allowing full generations. Empirically, we observe that the generation length distribution of LLMs typically exhibits exponential decay across tasks and model architectures (see Figure 5.3). This implies that long trajectories are increasingly rare, so the truncation bias $g(T_{\max})$ decays rapidly with $T_{\max}$.

We defer the proof of Theorem 5.4.1 to Appendix D.1.5. The same Monte Carlo framework extends to GenPPL, BF, and prompt-controlled correlation. Their consistency follows by the continuous mapping theorem [Mann and Wald, 1943] together with mild integrability and truncation-bias conditions, which are empirically supported by the observed low truncation rates and rapidly decaying active-rollout distributions (see Appendix D.1.6 for detailed proofs and discussions).

5.5 Empirical studies

In this section, we present our experimental results, organized into two complementary parts. First, we examine the global uncertainty structure of the LLM generation space using the metrics introduced in section 5.3, including CE*, GenPPL, BF, and length–uncertainty correlation. Second, building on this analysis, we investigate how uncertainty is translated into semantic variation through a regression framework that links entropy rate to semantic diversity. We evaluate three model families—LLaMA-3.1-8B [Grattafiori et al., 2024], Qwen3-8B [Yang et al., 2025a], and Gemma-3-12B [Team et al., 2025]—testing both their pre-trained (base) and instruction-tuned (instruct) versions. We evaluate these models on four tasks spanning a spectrum from highly structured settings, including mathematical reasoning [Cobbe et al., 2021], coding [Zhuo et al., 2024], and sentence completion Zellers et al. [2019], to more open-ended creative story generation [Ismayilzada et al., 2025]. For each task, we sample $P = 100$ prompts, and for each sampled prompt X_p, we generate $M = 100$ independent rollouts using stochastic decoding with temperature equals 1 and top-$k = 100$ (renormalized next-token probabilities),[5] where each rollout produces a sequence $Y_{1:N}$ with stopping time N up to a maximum of 4000 new tokens, i.e. $T_{\max} = 4000$ (see subsection D.3.1 for detailed description). To ensure consistent evaluation across model families, we disable the thinking mode in Qwen3-8B Instruct.

5.5.1 Global effects of fine-tuning on uncertainty

Across all model families and tasks, we observe three consistent patterns.

Reduction in token-level and trajectory-level normalized uncertainty. Instruction fine-tuning consistently reduces both token-level and trajectory-level normalized uncertainty, as evidenced by systematic declines in GenPPL and BF across nearly all models

5. Specifically, the step entropy $\tilde{H}(G_t|X)$ is computed as $-\sum_{z\in\mathcal{Z}_{100}} q(z|G_t, X)\log q(z|G_t, X)$, where $\mathcal{Z}_{100}$ is the set of the top-100 tokens (including `EOS` if present) and q is the renormalized distribution over the sampled subset.

Table 5.1: Task-dependent changes in generation-space uncertainty after instruction tuning across different model families. We report three complementary uncertainty measures: GenPPL, BF, and CE* (see subsection D.1.6). For each task and model family, we compare the base and instruct variants and report the relative percentage change after instruction tuning. Standard errors (SE) are estimated via prompt-cluster bootstrapping (2K iterations, resampled with replacement) as in subsection D.3.4. Blue cells indicate reductions in uncertainty after instruction tuning, while red cells indicate increases. Darker color intensity corresponds to larger magnitude changes.

Task	Type	GenPPL			BF			CE*		
		Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B
Math	Base	1.23 ± 0.0052	3.57 ± 0.041	4.56 ± 0.052	1.23 ± 0.0064	5.31 ± 0.12	4.98 ± 0.055	55.7 ± 2.0	458 ± 18	487 ± 11
	Instruct	1.17 ± 0.0067	1.74 ± 0.063	1.17 ± 0.0098	1.16 ± 0.0057	1.57 ± 0.024	1.15 ± 0.0054	40.5 ± 2.5	98.7 ± 9.1	59.1 ± 5.1
	% Change	-4.44 ± 0.47%	-51.2 ± 2.0%	-74.4 ± 0.34%	-5.76 ± 0.38%	-70.5 ± 0.73%	-76.9 ± 0.24%	-27.3 ± 3.3%	-78.4 ± 2.2%	-87.9 ± 1.1%
Coding	Base	1.68 ± 0.017	3.16 ± 0.055	2.60 ± 0.041	2.08 ± 0.034	4.96 ± 0.057	3.24 ± 0.038	207 ± 4.8	326 ± 13	492 ± 25
	Instruct	1.23 ± 0.0053	1.31 ± 0.0088	1.05 ± 0.0039	1.21 ± 0.0061	1.31 ± 0.0081	1.05 ± 0.0021	81.2 ± 4.0	135 ± 4.3	17.7 ± 1.6
	% Change	-27.1 ± 0.79%	-58.4 ± 0.74%	-59.7 ± 0.64%	-41.9 ± 0.99%	-73.7 ± 0.35%	-67.8 ± 0.38%	-60.7 ± 1.6%	-58.6 ± 1.7%	-96.4 ± 0.39%
Sentence Completion	Base	4.77 ± 0.16	6.73 ± 0.12	8.81 ± 0.11	4.20 ± 0.11	7.80 ± 0.15	9.49 ± 0.11	826 ± 55	969 ± 40	1084 ± 36
	Instruct	1.68 ± 0.025	2.56 ± 0.069	1.80 ± 0.019	1.55 ± 0.018	2.75 ± 0.045	1.74 ± 0.024	112 ± 11	201 ± 11	515 ± 29
	% Change	-64.9 ± 1.1%	-62.0 ± 1.3%	-79.6 ± 0.30%	-63.0 ± 0.88%	-64.8 ± 0.83%	-81.6 ± 0.30%	-86.5 ± 1.0%	-79.3 ± 1.3%	-52.5 ± 2.8%
Story Generation	Base	3.57 ± 0.11	7.40 ± 0.071	7.98 ± 0.062	4.03 ± 0.092	8.24 ± 0.073	8.68 ± 0.070	445 ± 20	1320 ± 41	692 ± 23
	Instruct	2.28 ± 0.037	3.24 ± 0.092	2.22 ± 0.023	2.11 ± 0.039	3.37 ± 0.083	2.17 ± 0.025	454 ± 28	543 ± 22	925 ± 40
	% Change	-36.0 ± 1.3%	-56.2 ± 1.4%	-72.2 ± 0.28%	-47.6 ± 1.3%	-59.1 ± 1.2%	-75.0 ± 0.28%	1.92 ± 3.5%	-58.9 ± 1.6%	33.7 ± 8.3%

and tasks (see Table 5.1). Nevertheless, the magnitude of this reduction varies substantially across settings. For example, Qwen3-8B on mathematical reasoning exhibits only a mild decrease of roughly 5%, whereas many other model–task combinations show much larger reductions, often in the range of 60%–80%. These results suggest that fine-tuning contracts the effective generation space at both the token and trajectory levels, producing outputs that are more concentrated and deterministic. At the same time, the heterogeneous magnitude of the reduction suggests that alignment does not impose a uniform compression of uncertainty. Instead, the extent of contraction depends on the underlying task structure and model behavior, a phenomenon we discuss in the following paragraph. Overall, our findings are consistent with prior observations that alignment reduces diversity and increases probability concentration [Wang et al., 2025, Yang et al., 2025b, Lake et al., 2025, West and Potts, 2025], although the shrinkage we observe is less extreme than the order-of-magnitude reductions reported in the previous study Yang et al. [2025b].

Task-dependent behavior of total entropy. The effect of fine-tuning on total uncertainty CE* varies substantially across tasks. In structured domains such as math, coding,

and sentence completion, $CE^{\star}$ decreases significantly after fine-tuning. In contrast, story generation exhibits a different pattern: Qwen3-8B shows a slight increase of roughly 2%, while Gemma-3-12B shows a much larger increase of about 34% (see Table 5.1). This suggests that fine-tuning does not uniformly reduce uncertainty, but instead behaves differently in constrained versus open-ended generation settings.

For deterministic tasks such as math and coding, the valid generation space is naturally constrained, so fine-tuning acts as a pruning force that collapses the model's generation space into a small set of correct trajectories. Consequently, both stopping-time variance and local token-level entropy are significantly reduced. By contrast, on creative tasks, fine-tuned models exhibit a notable rightward shift in sequence length distributions (see Figure 5.3), reflecting a tendency toward longer and more elaborative generations. In the same domain, the corresponding base models frequently fall into "completion traps," treating prompts as short text continuation tasks rather than directly answering the underlying questions (see subsection D.3.2).

Although instruction tuning reduces local branching diversity, it also increases stopping-time uncertainty, and sustains generation over much deeper trajectories. As illustrated in Figure 5.2, while the base model has higher local entropy rates (*width*), it lacks enough trajectory length (*depth*) to accumulate a large Canopy Entropy ($CE^{\star}$). Since $CE^{\star}$ acts as a measure of the total *volume* of the generation tree by accumulating uncertainty across the entire rollout, including stopping behavior, the increase in trajectory depth can mathematically outweigh the reduction in local branching.

These findings suggest that fine-tuning changes not only the magnitude of uncertainty, but also how uncertainty is distributed across the trajectory, particularly through its interaction with generation length and stopping behavior.

Systematic improvement in length–uncertainty correlation. To better understand how uncertainty is distributed across trajectories, we analyze the relationship between

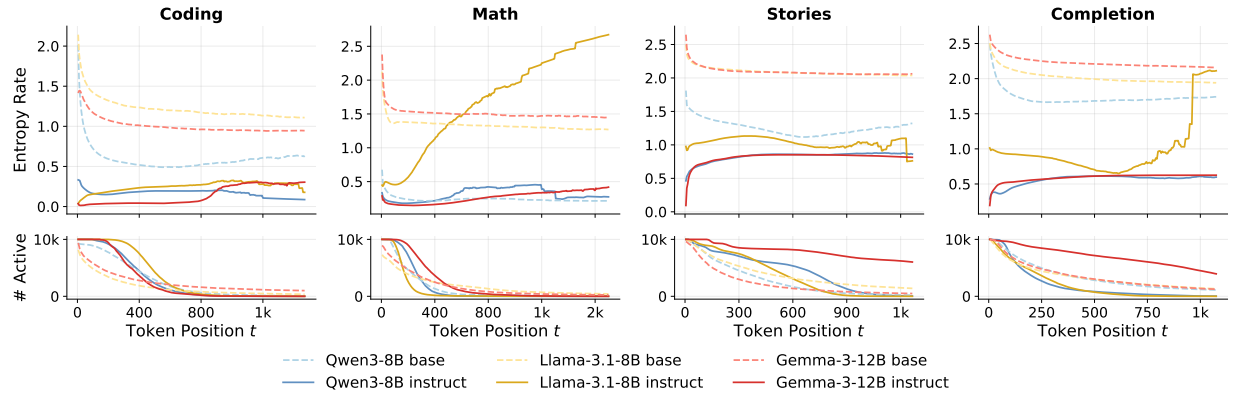

Figure 5.2: Running entropy rate vs. token position. We plot mean running entropy rate $\bar{R}_{\leq t} = \frac{1}{t}\sum_{s=1}^{t} H(Z_s \mid X, G_s)$ averaged across active rollouts at position t, with the corresponding active rollout count beneath. Dashed curves denote base models and solid curves denote fine-tuned models.

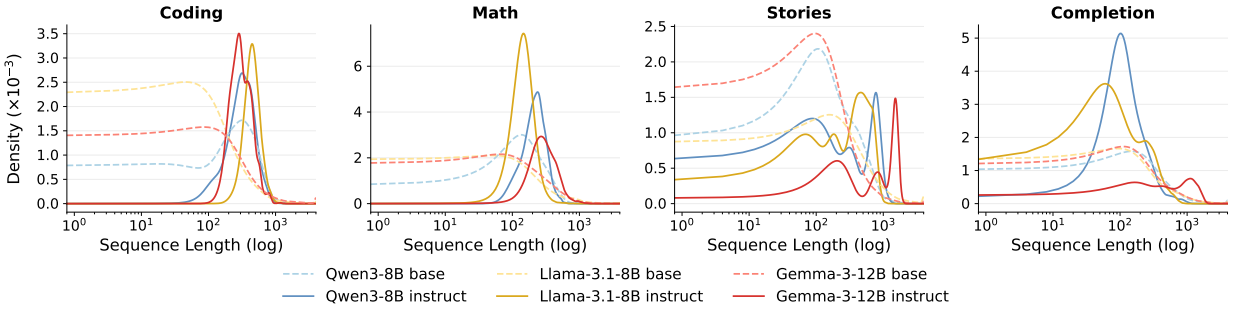

Figure 5.3: Kernel density estimates of generated sequence lengths across task domains with sequence length shown on a logarithmic scale.

output length N and entropy rate r_N through a decomposition of CE* (details in subsection D.3.3). Empirically, fine-tuning consistently shifts the correlation (see Definition 5.3.4) between N and r_N from negative toward positive values in structured tasks, while making the correlation less negative in open-ended tasks such as story generation (see Table 5.2).

This reveals a fundamental change in generation behavior. In base models, longer outputs are often associated with lower entropy rates, meaning later tokens become increasingly redundant and contribute little new information. In contrast, fine-tuned models exhibit much stronger positive length–uncertainty correlation: longer generations tend to sustain higher entropy rates and remain informative throughout the trajectory (see Figure 5.2).

The distinction is especially important in the presence of length confounding. As shown in Figure 5.3, fine-tuned models generally produce longer outputs than their base counter-

Table 5.2: Base–instruct comparison of the length–uncertainty Pearson correlation, measured by $\rho_{\text{pc}}(N, r_N)$ (see Definition 5.3.4), for the Qwen3-8B, LLaMA-3.1-8B, and Gemma-3-12B model families. Results based on Spearman correlation and Kendall's τ are provided in Table D.1 and exhibit same trends. Each entry reports the estimate ± standard error, where the absolute change is computed as the instruct value minus the corresponding base value.

Metric	Model	MATH			CODING			SENTENCE COMPLETION			STORY GENERATION		
		Base	Instruct	Abs. Change	Base	Instruct	Abs. Change	Base	Instruct	Abs. Change	Base	Instruct	Abs. Change
$\rho(N, r_N)$	Qwen	-0.018 ± 0.016	0.087 ± 0.037	0.11 ± 0.042	-0.41 ± 0.014	0.61 ± 0.034	1.0 ± 0.037	0.062 ± 0.017	0.52 ± 0.024	0.46 ± 0.027	-0.27 ± 0.017	0.16 ± 0.021	0.43 ± 0.027
	Llama	-0.27 ± 0.016	0.38 ± 0.040	0.65 ± 0.044	-0.38 ± 0.011	0.29 ± 0.021	0.67 ± 0.021	-0.26 ± 0.015	-0.20 ± 0.024	0.053 ± 0.028	-0.24 ± 0.012	-0.10 ± 0.020	0.14 ± 0.022
	Gemma	-0.13 ± 0.014	0.19 ± 0.040	0.32 ± 0.043	-0.27 ± 0.012	0.39 ± 0.035	0.67 ± 0.037	-0.23 ± 0.015	0.12 ± 0.038	0.36 ± 0.041	-0.17 ± 0.012	-0.049 ± 0.027	0.12 ± 0.032

parts, consistent with observations in prior work [Singhal et al., 2023]. Under conventional diversity metrics that do not explicitly account for length [Tevet and Berant, 2021, Li et al., 2016], these longer generations can artificially appear less diverse due to normalization effects or repetition accumulation, potentially biasing diversity comparisons [Friedman and Dieng, 2022, Shaib et al., 2024]. The positive length–uncertainty correlation observed here demonstrates that increased length in fine-tuned models reflects sustained information content rather than simple redundancy.

Overall, these results suggest that fine-tuning improves not merely the quantity of uncertainty, but also its organization and efficiency. Rather than uniformly suppressing diversity, fine-tuning restructures uncertainty to better align with task requirements—reducing unnecessary randomness while preserving informative variation across longer trajectories.

5.5.2 Fine-tuning more effectively converts token uncertainty into semantic diversity

After establishing how fine-tuning reshapes entropy allocation, we next investigate how these changes in entropy rate translate into semantic diversity. Specifically, we study how generation uncertainty is converted into semantic variation across sampled trajectories.

For each task–prompt–model tuple (t, p, m), we measure semantic diversity by computing the average pairwise cosine distance between generated outputs:
$D_{tpm} = \frac{2}{M(M-1)} \sum_{i<j} d(e_{t,p,m,i}, e_{t,p,m,j})$, where $e_{t,p,m,i}$ denotes the embedding of the i-th

Monte Carlo rollout, prepended with the retrieval prefix (`"search_document:"`) to ensure alignment with the encoder's training objective for modernbert-embed-large [Warner et al., 2024]. Here $d(\cdot, \cdot)$ denotes cosine distance. Since cosine distance is bounded in $[0, 1]$, we employ a Beta mixed-effects regression [Figueroa-Zúñiga et al., 2013], which naturally accommodates bounded responses while allowing flexible modeling of dispersion and hierarchical variation.[6]

Formally, we model $D_{tpm} \sim \text{Beta}\big(\mu_{tpm}\phi_{tpm}, (1-\mu_{tpm})\phi_{tpm}\big)$, where $\mu_{tpm} \in [0, 1]$ denotes the conditional mean and $\phi_{tpm} > 0$ is the precision parameter. The conditional mean model is specified as $\text{logit}(\mu_{tpm}) = \alpha + \beta_1 R_{tpm} + \beta_2 \mathbf{1}\{m = \text{FT}\} + f\left(\text{invN}_{tpm}\right) + \sum_k \tau_k \mathbf{1}\{t = k\} + \beta_3 (R_{tpm} \cdot \mathbf{1}\{m = \text{FT}\}) + g\left(\text{invN}_{tpm}\right) \mathbf{1}\{m = \text{FT}\} + u_p + v_m$, where R_{tpm} is the entropy rate, invN_{tpm} is the inverse sequence length, and $f(\cdot)$, $g(\cdot)$ are natural spline functions corresponding to nonlinear length effects and their interaction with instruction tuning. The coefficients τ_k represent fixed task effects. To account for hierarchical dependence, we include random intercepts $u_p \sim \mathcal{N}(0, \sigma_p^2), v_m \sim \mathcal{N}(0, \sigma_m^2)$, for prompt identity and model family, respectively. To capture heteroscedasticity, we also model the precision parameter ϕ_{tpm} as a function of task, model variant, entropy rate, sequence length, and model family. Residual diagnostics (see Figure D.4) indicate good calibration of the mean structure (Kolmogorov-Smirnov Test [Massey Jr, 1951]: $p = 0.55$) and no significant dispersion issues ($p = 0.068$) and no outliers ($p = 0.16$).

Main results. First, entropy rate exhibits a strong and highly significant positive association with semantic diversity. The coefficient on entropy rate is large and significant ($\beta = 0.708$, $p < 2 \times 10^{-16}$), indicating that higher per-token uncertainty leads to substantially greater semantic variation across generated outputs. More importantly, the interaction between entropy rate and fine-tuning is strongly positive ($\eta = 1.183$, $p < 2 \times 10^{-16}$). Under the logit link, this implies a substantial amplification of the slope: $\partial \,\text{logit}(\mu)/\partial R =$

6. In contrast, Gaussian linear models are inappropriate for bounded responses and exhibit clear heteroscedasticity and residual misspecification in diagnostic analyses.

Table 5.3: Beta GLMM with dispersion modeling and interaction effects. For completeness, we direct readers to subsection D.3.5 for the full conditional mean and dispersion regression tables, along with additional model specifications and regression details.

Conditional model	**Estimate**	**Std. Error**	p**-value**
Intercept	-0.880	0.070	$< 2 \times 10^{-16}$
R	0.708	0.033	$< 2 \times 10^{-16}$
Instruct	-0.377	0.086	1.23×10^{-5}
$\text{ns}(\text{invN}, 4)_1$	0.152	0.041	1.99×10^{-4}
$R\times$ Instruct	1.183	0.046	$< 2 \times 10^{-16}$

0.708 (base), $0.708 + 1.183 = 1.891$ (fine-tuned). Thus, fine-tuning nearly triples the sensitivity of semantic diversity to entropy rate. This indicates that uncertainty in fine-tuned models is significantly more *semantically productive*: small increases in token-level uncertainty translate into much larger increases in trajectory-level diversity.

At the same time, the main effect of fine-tuning is negative ($\tau = -0.377$, $p < 2 \times 10^{-5}$), indicating that at fixed entropy rate and length, fine-tuned models produce less baseline semantic diversity. This is consistent with alignment compressing the generation space and enforcing more structured outputs. However, the strong positive interaction shows that the remaining uncertainty is used more efficiently. We observe a nonlinear and generally negative relationship between semantic diversity and sequence length, making length control essential in diversity analysis (see Table 5.3).

5.6 Conclusion and limitations

In this work, we revisit diversity in large language models through a trajectory-level information-theoretic perspective. By introducing Canopy Entropy and length–uncertainty coupling, we show that fine-tuning does not simply reduce uncertainty or shrink the generation space. Instead, aligned models reorganize uncertainty more efficiently across generation trajectories, producing longer generations that remain more semantically informative. Our results highlight the importance of modeling how uncertainty evolves throughout the rollout process, beyond aggregate diversity metrics alone.

Our work also has several limitations. Entropy-based measures characterize uncertainty allocation, but do not directly capture factuality or human preference quality. Our framework is descriptive rather than causal: although we observe systematic changes in length–uncertainty coupling after fine-tuning, the underlying training mechanisms remain an important direction for future work.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This dissertation studies the problem of aligning artificial intelligence systems with human preferences through three interconnected perspectives: personalization, robustification, and evaluation. At their core, these directions are fundamentally statistical problems: how to collect informative data from complex interactive environments, how to perform reliable inference under uncertainty and imperfect observations, how to model heterogeneous effects across users and contexts, and how to rigorously evaluate whether learned systems behave in ways that are truly aligned with human objectives. These challenges have become increasingly important in the era of large language models and foundation models, where AI systems interact with humans at unprecedented scale, operate under highly heterogeneous feedback distributions, and continuously adapt through large-scale data-driven learning.

A central theme throughout this dissertation is that modern AI systems can no longer be understood through static prediction settings alone. Instead, they operate in sequential, interactive, and strategic environments where data collection itself depends on the learning algorithm, where observations are delayed or corrupted, and where human preferences vary substantially across individuals and contexts. Consequently, statistical inference, uncertainty quantification, and evaluation become deeply intertwined with decision-making and interaction dynamics.

From the perspective of personalization, this dissertation develops contextual reinforcement learning frameworks for learning under delayed, cumulative, and heterogeneous treatment effects. By modeling ad bidding as a contextual Markov decision process with delayed Poisson rewards, we provide a unified framework for understanding how AI systems can adapt to heterogeneous users over time. The proposed two-stage estimation procedure highlights the importance of statistically principled inference in sequential environments, particularly when delayed rewards induce complex dependencies between observations and latent

states. More broadly, these results emphasize that personalization fundamentally requires understanding heterogeneous effects and long-term interactions, which are central themes in modern statistics and causal inference. Future work may further explore richer nonlinear effect models, constrained decision-making under budgets, and scalable inference procedures for large-scale interactive systems.

From the perspective of robustification, this dissertation studies how imperfect or adversarial data fundamentally limits reliable inference and learning. In the setting of imperfect human feedback, we establish regret lower bounds and develop near-optimal algorithms that reveal intrinsic trade-offs between statistical efficiency and robustness. In multi-agent environments, we further show that even localized perturbations can propagate globally through strategic interactions, causing large-scale failures in collective learning dynamics. These findings highlight a core statistical challenge in modern AI systems: learning algorithms must not only extract signal efficiently from data, but also remain stable under corrupted, biased, or strategically manipulated observations. As large language models increasingly rely on human-generated feedback and online interaction data, understanding robustness under imperfect supervision becomes even more critical. Future directions include black-box adversarial settings, adaptive defense mechanisms, and statistically principled methods for robust inference under strategic data contamination.

Finally, from the perspective of evaluation, this dissertation develops information-theoretic tools for understanding uncertainty and semantic diversity in large language model generation. By introducing Canopy Entropy and length–uncertainty coupling, we propose a trajectory-level statistical framework for quantifying how uncertainty evolves throughout the generation process. Our findings suggest that alignment and fine-tuning do not merely reduce uncertainty, but instead reorganize uncertainty into more informative and semantically meaningful generations. More broadly, this work highlights the growing importance of statistical evaluation methodologies in the era of generative AI. As language models be-

come increasingly capable and widely deployed, evaluation itself becomes a central statistical problem: how to measure uncertainty, diversity, calibration, informativeness, and alignment in extremely high-dimensional generation spaces. While entropy-based measures provide one useful perspective, many important questions remain open, including the relationship between uncertainty and factuality, the causal mechanisms underlying alignment training, and the development of evaluation frameworks that better reflect human preferences and downstream societal impact.

Taken together, this dissertation argues that the future of AI alignment is fundamentally tied to the future of statistics. As AI systems become increasingly interactive, personalized, and generative, statistical challenges surrounding data collection, inference under uncertainty, heterogeneous effects, robustness, and evaluation will only become more central. We hope that the theoretical frameworks and methodologies developed in this dissertation contribute toward a broader statistical foundation for building reliable, adaptive, and human-centered artificial intelligence systems in the era of large language models.

APPENDIX A

APPENDIX TO PERSONALIZED ADVERTISEMENT IMPACT

A.1 Additional Discussion

A.1.1 Related Work

Motivated by online advertising auctions, Weed et al. [2016] study repeated Vickrey auctions where goods with unknown values (v_t) are sold sequentially. Bidders receive (potentially noisy) feedback about a good's value only after purchasing it—analogous to observing product conversions after displaying an ad. They formulate this problem as online learning with bandit feedback and propose a minimax-optimal bidding strategy for bidding (b_t). However, a key limitation of their model is the neglect of long-term advertising effects: winning a bid at time t may influence future observations and outcomes, such as conversions at time $t+1$. Building on this line of work, Feng et al. [2018] examines learning to bid without knowing one's value in a similar bandit setting and develops the WIN-EXP algorithm, which achieves sublinear regret ($O(\sqrt{T|O|\log(B)})$). Yet, like the earlier study, it assumes that the effect of winning a bid is instantaneous, ignoring possible intertemporal dependencies.

In a related direction, Han et al. [2024a] investigate online learning in repeated first-price auctions, where a bidder observes only the winning bid after each auction and must adaptively learn to maximize cumulative payoff. Facing censored feedback—since winning bidders cannot observe their competitors' bids—they design the first algorithm achieving near-optimal ($O(\sqrt{T})$) regret by exploiting structural properties of first-price auctions, including their feedback and payoff functions. They formulate the problem as partially ordered contextual bandits, incorporating graph feedback across actions, cross-learning across contexts, and a partial order over private values. Nonetheless, their framework similarly assumes independence between the value process (v_t) and past bidding outcomes, thereby excluding long-term effects of winning bids.

Extending to contextual settings, Zhang and Luo [2024b] study online learning in contextual pay-per-click auctions. At each of T rounds, the learner observes contextual information and a set of candidate ads, estimates their click-through rates (CTRs), and conducts a second-price pay-per-click auction. The objective is to minimize regret relative to an oracle that makes perfect CTR predictions. They show that a $\sqrt{T}$-regret rate is achievable (albeit with computational inefficiency) and provably optimal, as the problem reduces to the classical multi-armed bandit setting.

While these studies advance our understanding of strategic bidding under uncertainty, they share a common limitation: none account for the reinforcement or long-term impact of successful ad displays. This omission contrasts with empirical evidence such as De Haan et al. [2016], who demonstrate that online advertisements exert persistent effects on purchase conversion in a multi-channel attribution framework—though their dataset is not publicly available.

More recently, Badanidiyuru et al. [2021] address this gap by modeling the long-term causal impact of ad impressions using a Markov Decision Process (MDP) with mixed and delayed Poisson rewards. However, their approach assumes homogeneous treatment effects across users, overlooking the crucial role of personalization in determining advertising effectiveness. To address these limitations, our work jointly models the long-term, reinforcing, and personalized effects of advertisements, hoping to bridge the gap between sequential decision-making and individualized ad impact estimation.

A.1.2 On the Potential to Extend the Linearity Assumption

Briefly recap of our model (Assumption 1), if we won the bid $\boldsymbol{o}_h^t = 1$, the average product conversion rate $\mu_h^t = h_{S_{h,1}^t}(\boldsymbol{x}_t)$, where $\boldsymbol{x}_t$ is the received feature of customer, and $S_{h,1}^t$ is the state which captures the time elapsed since most recent ads impression, $h_{S_{h,1}^t}$ is the state dependent neural network which we want to estimate. If we lose the bid $\boldsymbol{o}_h^t = 0$, the average

product conversion rate $\mu_h^t = d_{S_{h,1}^t} h_{S_{h,2}^t}(\boldsymbol{x}_t)$, where $d_{S_{h,1}^t}$ is the delayed impact, and $S_{h,2}^t$ is the time interval between the two most recent impressions (Def 2.2.1).

To incorporate the estimation of the state-dependent neural network h_l (analogous to $\boldsymbol{\theta}_l$) into our algorithm, we could do follows. Everything in Algorithm 1 remains unchanged, i.e. the exploration, exploitation, data-splitting strategy, and estimation, exception the following modifications.

- First, we replace the estimation of $\boldsymbol{\theta}_l$ (Eqn. (2.4)) by estimating the neural network h_l by the following procedure (similar to Algorithm 1 in Jia et al. [2022b]).
  - We approximate $h(x)$ by $f(x,\theta) = \sqrt{m} W_L \phi(W_{L-1}\phi(\dots\phi(W_1 x)))$, where m is the network width, and $\phi(x) = ReLU(x)$, L is the neural network depth.
  - We initialize $\theta_0 = (\text{vec}(W_1)\dots\text{vec}(W_L))$ by random sample entry from $\mathcal{N}(0, 4/m)$ and do online updating by gradient descent with perturbed reward using observed product conversion $\boldsymbol{y}_h^t$ by follows
    - Generated observation perturbation $\gamma_{s\in[t]}^t \sim \text{Poi}(\lambda)$, then generate binomial random variable $\mathbb{I}_{s\in[t]}$ to add or subtract the noise from the observation

$$\min_{\theta} L(\theta) = \sum_{s=1}^{t} (f(\boldsymbol{x}_t, \theta) - (\boldsymbol{y}_h^s + \mathbb{I}_s \gamma_s^t))^2/2 + m\beta\|\theta - \theta_0\|^2/2$$

  - Then based on this first stage estimates of h, plugging in to estimate delayed impact factor d by $\frac{\sum_{s=1}^t \sum_{h\in\mathbf{D}_{s,l}} \boldsymbol{y}_h^s}{\sum_{s=1}^t \sum_{h\in\mathbf{D}_{s,l}} f_{S_{h,2}^s}(\boldsymbol{x}_s, \hat{\boldsymbol{\theta}})}$ (analogous to Eqn. (2.3))
- Since we do not have theoretical guarantees for h estimation (no confidence interval), we cannot do upper confidence bound (Line 10 of our Algorithm 1). Instead, we might act greedily based on our point estimate. Even though Algorithm in Jia et al. [2022b] has theoretical guarantees, the key difference is that they assume noise of their observation is sub-Gaussian, while in our case, it is not realistic to assume that the observation

$\boldsymbol{y}_h^t$ has sub-Gaussian noise, since product conversion has been known to have heavy tail for a long period of time. It is still a challenging open problem of developing theoretical-guaranteed neural contextual bandits with heavy tail observations.

However, our proposed framework leads to near-optimal results if we can parameterize $h_l(\boldsymbol{x}_t) = \boldsymbol{\theta}_l^\top \phi(\boldsymbol{x}_t)$, where ϕ can be the neural net based embedding or other known feature mapping, then our algorithm remains near optimal, satisfying $\sqrt{T}$ regret bound in this case. This construction is very common in academia and industry as we mentioned from foundational studies such as Jin et al. [2020] to the recent ones [He et al., 2022]. In addition, in real world, companies might construct these feature mapping using complicated methods but retain the overall linear structure for the ease of scalability and interpretability. For example, the expected conversion rate could be a linear function of the expected webpage stay time and the predicted total spending but the expected webpage stay time and the predicted total spending are deep neural networks of the observed customer feature $\boldsymbol{x}_t$.

To summarize, our proposed framework leads to near-optimal results if there exists an online estimation oracle with theoretical guarantee for the estimation of h and the estimation oracle employed in our paper based on our knowledge is state-of-the-art. Our algorithm is possible to extend to a neural network if we drop the theoretical concerns. Developing theoretical guarantee for neural bandits under heavy tail observation could be a promising future direction.

A.1.3 Flexibility in State Space Modeling

If we believe that all past ad exposures contribute to customer behavior, we can enrich the state variable to $S_h^t = [S_{h,1}^t, S_{h,2}^t, n_h^t]$, where n_h^t denotes the total number of ads the user has seen prior to round h. Under this formulation, the expected conversion from showing an ad at round h is given by $\langle \boldsymbol{\theta}_{S_{h,1}^t, n_h^t}, \boldsymbol{x}_t \rangle$, with other modeling assumptions unchanged. In other words, the effect of the current ad depends not only on the time since the last impression,

$h - G^t_{h,1}$ (as shown in Figure 2.1), but also on the cumulative number of ad exposures the user has experienced so far.

Accordingly, Assumption 1 as follows:

1. When $\boldsymbol{o}^t_h = 0$, no ad is shown at round h, and the effect of recent ads at $G^t_{h,1}$ carries over, given by $\langle \boldsymbol{\theta}_{S^t_{h,2},(n^t_h-1)\vee 0}, \boldsymbol{x}_t \rangle \, d_{S^t_{h,1}}$, where $a \vee b$ denotes $\max(a, b)$.

2. When $\boldsymbol{o}^t_h = 1$, an ad is shown at round h, and the impact of the current ad is modeled as $\langle \boldsymbol{\theta}_{S^t_{h,1},n^t_h}, \boldsymbol{x}_t \rangle$, influenced by all the past ads exposure history.

We emphasize that, our core algorithmic design, including the data-splitting strategy, the proposed two-stage estimator, and the exploration-exploitation algorithm, readily extends to this new model with the enriched state representation. In addition, our theoretical results continue to hold under this new model, with a new optimal regret bound of order $\widetilde{O}(dH^3\sqrt{T})$, in comparison to $\widetilde{O}(dH^2\sqrt{T})$ under the previous model. We provide supporting evidence below.

Now the enriched states is $S^t_h = [S^t_{h,1}, S^t_{h,2}, n^t_h]$. In terms of state transition, if $\boldsymbol{o}^t_h = 1$, $S^t_{h+1} = [1, S^t_{h,1}, n^t_h + 1]$; else, $S^t_{h+1} = [S^t_{h,1} + 1, S^t_{h,2}, n^t_h]$. Θ now becomes $\{\boldsymbol{\theta}_{l,n}\}_{l,n\in\mathcal{H}} \cup \{d_l\}_{l\in\mathcal{H}_1}$ with $\mathcal{H} = \{(l, n) : l \in \mathcal{H}_1 \cup \mathcal{H}_2, n < l\}$, $\mathcal{H}_1 = \{\infty, 1, 2, \ldots, H-1\}$ and $\mathcal{H}_2 = \{-\infty, 1, 2, \ldots, H-2\}$.

For data spiting strategy in Def 2.3.1, $\mathbf{D}_{t,l}$ does not change and $\mathbf{W}_{t,l}$ will extend to $\mathbf{W}_{t,l,n} = \{h | S^t_{h,1} = l, n^t_h = n, \boldsymbol{o}^t_h = 1\}, \forall l \in [\infty, 1, 2, \ldots, H], \mathbf{W}_{t,-\infty,n} = \{h | S^t_{h,1} = -\infty, n^t_h = n, \boldsymbol{o}^t_h = 0\}$. Also, $\mathbf{W}^t_l$ will extend to $\mathbf{W}^t_{l,n} = \{\mathbf{W}_{s,l,n}\}^t_{s=1}$. The estimation of $\boldsymbol{\theta}_{l,n}$ using observation in $\mathbf{W}^t_{l,n}$ still follows online newton estimator with Eqn. (2.4). The TS-MLE now becomes

$$\hat{d}^t_l = \frac{\sum^t_{s=1} \sum_{h\in\mathbf{D}_{s,l}} \boldsymbol{y}^s_h}{\sum^t_{s=1} \sum_{h\in\mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2},n^s_h-1\vee 0}, \boldsymbol{x}_s \rangle}.$$

Plugging in these new estimators in to the regret decomposition (Eqn. ((2.9)), Term (i) and Term (iv) does not change since new states did not affect transition error. For the estimation error, Term (iii), we just need an additional for loop to account the effect of n_h^t, which makes the new regret upper bound now $\tilde{\mathcal{O}}(dH^3\sqrt{T})$.

The above example suggests that our state formulation is very flexible and readily accommodates an enlarged state space. It supports a broad class of CMDP formulations for $\boldsymbol{\theta}_S$, where S encodes domain knowledge about ad effects, and remains compatible with our proposed learning algorithm.

A.1.4 Illustrative Example

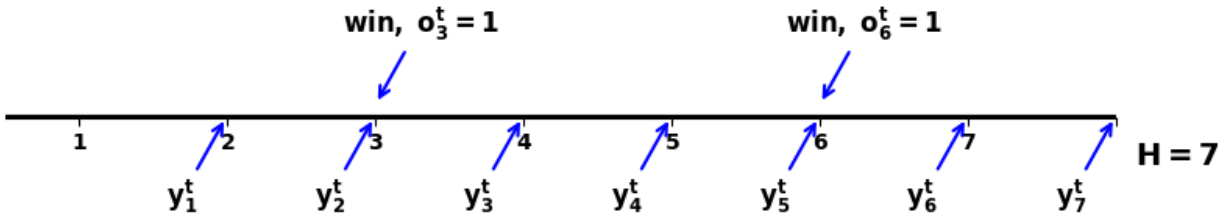

Figure A.1: Illustration of Learner-Customer Interaction in CMDP

Figure A.1 presents a simple example of a customer, with contextual $\boldsymbol{x}_t$ and no prior ad exposure, interacting with the learner $\mathcal{L}$ over $H = 7$ rounds. At the start of each round, $\mathcal{L}$ observes the state S_h^t, determines the bid amount $\mathbf{a}_h^t = \pi_t(S_h^t, \boldsymbol{x}_t)$ based on policy π_t, observes HOB m_h^t, and receives the bidding outcome $\boldsymbol{o}_h^t$ along with the reward $R_h^t(S_h^t, \mathbf{a}_h^t, \boldsymbol{x}_t)$. At the end of each round, $\mathcal{L}$ observes the total product conversion $\boldsymbol{y}_h^t$ over the current round. As shown in Figure A.1, the customer's first ad impression occurs at $h = 3$, resulting in $S_1^t = S_2^t = S_3^t = [-\infty, \infty]$ and $\mu_1^t = \mu_2^t = \langle \boldsymbol{\theta}_{-\infty}, \boldsymbol{x}_t \rangle$ (μ_h^t defined in Def.1). The first ads exposure updates the expected conversion to $\mu_3^t = \langle \boldsymbol{\theta}_{\infty}, \boldsymbol{x}_t \rangle$, capturing the immediate impact of ad exposure. The second ad impression occurs at $h = 6$. Between $h = 4$ and $h = 6$, the state evolves as $S_4^t = [1, \infty], S_5^t = [2, \infty], S_6^t = [3, \infty]$, with expected conversions

$\mu_4^t = d_1\langle \boldsymbol{\theta}_\infty, \boldsymbol{x}_t\rangle$ and $\mu_5^t = d_2\langle \boldsymbol{\theta}_\infty, \boldsymbol{x}_t\rangle$, reflecting the delayed and long-term impact of the ad displayed at $h = 3$. $\mu_6^t = \langle \boldsymbol{\theta}_3, \boldsymbol{x}_t\rangle$, capturing the effect of repeated ad exposure. Finally, at $h = 7$, $S_7^t = [1, 3]$ and $\mu_7^t = d_1\langle \boldsymbol{\theta}_3, \boldsymbol{x}_t\rangle$, demonstrating the long-term effects of the second ad impression.

A.1.5 Computational Complexity

For each episode of Algorithm 1, the computational complexities for updating $\hat{\boldsymbol{\theta}}_l^t$, $\hat{\boldsymbol{\beta}}_l^t$, and $\hat{d}_l^t$ are $\mathcal{O}(d^2)$ [Xue et al., 2024], $\mathcal{O}(d^2)$, and $\mathcal{O}(d)$, respectively, for each $l \in \mathcal{H}$. Notably, for a fixed policy π and transition $\mathcal{F}^{\boldsymbol{x}_t}$, $R(\pi, \Theta, \mathcal{F}^{\boldsymbol{x}_t})$ is nondecreasing in d_l and $\boldsymbol{\theta}_l$, simplifying the computation of $\max_{\tilde{\Theta}\in\mathcal{C}_{t-1}} R(\pi; \tilde{\Theta}, \mathcal{F}_{t-1}^{\boldsymbol{x}_t})$. The corresponding maximizer are $\tilde{d}_l^t = \hat{d}_l^t + \frac{4H\boldsymbol{B}_d\sqrt{d\log\left(1+\frac{T}{2d}\right)\gamma}+\sqrt{2e\boldsymbol{B}_d\boldsymbol{B}_x\boldsymbol{B}_\theta\log(2/\delta)}}{\boldsymbol{b}\sqrt{N_{t,l}}}$ and $\tilde{\boldsymbol{\theta}}_l^t = \hat{\boldsymbol{\theta}}_l^t + \frac{\sqrt{\gamma}(\mathcal{V}_l^t)^{-1}\boldsymbol{x}_t}{\|\boldsymbol{x}_t\|_{(\mathcal{V}_l^t)^{-1}}}$, obtained via constrained linear optimization, where $\hat{\boldsymbol{\theta}}_l^t$ is from Eqn. (2.4) and γ is from Lemma 2.3.1. The greedy policy π is then computed via dynamic programming with time complexity $\mathrm{Poly}(H, |\mathcal{S}|, \boldsymbol{B}_{\mathcal{A}}/\epsilon)$ when discretizing the bidding space for an ϵ-optimal solution [Agarwal et al., 2019].

A.2 Details of Omitted Algorithm

In this section, we detail the algorithm used to estimate $\boldsymbol{\theta}_l$ for $l \in \mathcal{H}$. Originally introduced by Xue et al. [2024], the Confidence Region with Truncated Mean (CRTM) algorithm (Algorithm 4) serves as an efficient estimator for generalized bandits with heavy-tailed rewards. In their framework, the stochastic observations $\boldsymbol{y}_t$ follow a generalized linear model:

$$\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{x}_t) = \exp\left(\frac{\boldsymbol{y}_t\boldsymbol{x}_t^\top\boldsymbol{\theta}_l - m(\boldsymbol{x}_t^\top\boldsymbol{\theta}_l)}{g(\tau)} + h(\boldsymbol{y}_t, \tau)\right),$$

where $\boldsymbol{\theta}_l$ represents the inherent parameters, $\tau > 0$ is a known scale parameter, and $g(\cdot)$ and $h(\cdot)$ are normalizers. The conditional expectation of $\boldsymbol{y}_t$ is given by:

$$\mathbb{E}(\boldsymbol{y}_t|\boldsymbol{x}_t) = m'(\boldsymbol{x}_t^\top \boldsymbol{\theta}_l),$$

where $m'(\cdot)$ is the link function, denoted as $\mu(\cdot) = m'(\cdot)$. Thus, $\boldsymbol{y}_t$ can be expressed as:

$$\boldsymbol{y}_t = \mu(\boldsymbol{x}_t^\top \boldsymbol{\theta}_l) + \eta_t,$$

where η_t is a random noise term satisfying $\mathbb{E}(\eta_t|\mathcal{G}_{t-1}) = 0$.

Here, $\mathcal{G}_{t-1} = \{\boldsymbol{x}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t\}$ denotes the σ-filtration up to time $t-1$. CRTM assumes that the link function $\mu(\cdot)$ is L-Lipschitz, continuously differentiable, and has a first derivative $\mu'(\cdot)$ bounded below by a positive constant κ, i.e., $\mu'(z) \geq \kappa$. Additionally, CRTM assumes that the observations $\boldsymbol{y}_t$ have bounded moments. Specifically, there exist positive constants u and $0 < \epsilon \leq 1$ such that:

$$\mathbb{E}(|\boldsymbol{y}_t|^{1+\epsilon} \mid \mathcal{G}_{t-1}) \leq u.$$

In our setting, where $\boldsymbol{y}_t$ follows a Poisson distribution, we have $\epsilon = 1$, $u = \boldsymbol{B}_x\boldsymbol{B}_\theta(1 + \boldsymbol{B}_x\boldsymbol{B}_\theta)$, $\mu(\cdot)$ as a linear function, and $\kappa = 1$. Based on the specific parameters in our setting, the CRTM algorithm (Algorithm 4) is adapted as follows:

Algorithm 4: Confidence Region with Truncated Mean [Xue et al., 2024]

Input: $\mathbf{W}_{t,l}, \boldsymbol{x}_t, \mathcal{V}_l^{t-1}, \hat{\boldsymbol{\theta}}_l^{t-1}, \Gamma, \gamma$

for $\boldsymbol{y}_h^t \in \mathbf{W}_{t,l}$ **do**

Truncate payoff: $\tilde{\boldsymbol{y}}_h^t = \boldsymbol{y}_h^t \mathbb{I}_{\|\boldsymbol{x}_t\|_{(\mathcal{V}_l^t)^{-1}} |\boldsymbol{y}_h^t| \leq \Gamma}$;

Compute gradient: $\nabla \tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1}) = (-\tilde{\boldsymbol{y}}_h^t + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_l^{t-1})\boldsymbol{x}_t$;

Update $\mathcal{V}_l^t = \mathcal{V}_l^{t-1} + \frac{1}{2}\boldsymbol{x}_t \boldsymbol{x}_t^\top$;

Update estimator:

$$\hat{\boldsymbol{\theta}}_l^t = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_l^{t-1}\|_{\mathcal{V}_l^t}^2}{2} + \langle \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_l^{t-1}, \nabla \tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1}) \rangle$$

Construct confidence region:

$$\mathcal{C}_l^t = \left\{ \boldsymbol{\theta}_l \in \mathbb{R}^d \mid \|\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_l^t\|_{\mathcal{V}_l^t}^2 \leq \gamma \right\}$$

Output: $(\hat{\boldsymbol{\theta}}_l^t, \mathcal{C}_l^t)$

In particular, the truncation threshold Γ is defined by Eqn. (A.1).

$$\Gamma = 2\sqrt{\boldsymbol{B}_x \boldsymbol{B}_\theta (1 + \boldsymbol{B}_x \boldsymbol{B}_\theta) \log\left(\frac{4T}{\delta}\right) d \log\left(1 + \frac{T}{2d}\right)} \tag{A.1}$$

Moreover, the bound for the weighted estimation error, $\|\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_l^t\|_{\mathcal{V}_l^t}^2$, denoted by γ, is given by:

$$\gamma = 896 d \boldsymbol{B}_x \boldsymbol{B}_\theta (1 + \boldsymbol{B}_x \boldsymbol{B}_\theta) \log\left(\frac{4T}{\delta}\right) \log\left(1 + \frac{T}{2d}\right) + 2\boldsymbol{B}_x^2 \boldsymbol{B}_\theta^2 + 48 d \boldsymbol{B}_x \boldsymbol{B}_\theta \log\left(1 + \frac{T}{2d}\right).$$

Running Algorithm 4 leads to the following lemma.

Lemma (Lemma 2.3.1 restated)**.** *Given $l \in \mathcal{H}$, with probability at least $1 - \delta$, $\hat{\boldsymbol{\theta}}_l^t$ defined in*

Eqn. (2.4) *satisfies* $\|\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_l^t\|_{\mathcal{V}_l^t}^2 \leq \gamma, \forall t \geq 0$.

A.3 Proof of Theorem 2.4.2

Theorem (Theorem 2.4.2 restated)**.** *For any* $\delta \geq \frac{6}{T^3}$*, with probability at least* $1-\delta$*,* Reg_T *incurred by Algorithm 3 is* $\mathcal{O}(dH^2\sqrt{T\log(\frac{TH}{\delta})}\log(1+\frac{T}{2d}))$.

Proof. In this section, we present the detailed proof of Theorem 2.4.2. We begin by decomposing Reg_T (defined in Eqn. (2.1)) into four terms—Term (i), Term (ii), Term (iii), and Term (iv)—as shown in Eqn. (2.9). $\Theta(\log T)$ comes from regret incurred in the exploration phase with $\tau = H\underline{n_l} + 1$.

$$
\begin{aligned}
\mathrm{Reg}_T =& \sum_{t=1}^{T} \mathrm{OPT}\left(\Theta, \boldsymbol{x}_t, \mathcal{F}^{\boldsymbol{x}_t}\right) - \mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left(\sum_{t=1}^{T} R\left(\pi_t; \boldsymbol{x}_t, \Theta, \mathcal{F}^{\boldsymbol{x}_t}\right)\right) \\
\leq& \Theta(\log T) + \underbrace{\sum_{t=\tau}^{T} \mathrm{OPT}\left(\Theta, \boldsymbol{x}_t, \mathcal{F}^{\boldsymbol{x}_t}\right) - \sum_{t=\tau}^{T} \mathrm{OPT}\left(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)}_{\text{Term (i)}} + \\
& \underbrace{\sum_{t=\tau}^{T} \mathrm{OPT}\left(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right) - \sum_{t=\tau}^{T} \mathrm{OPT}\left(\mathcal{C}_{t-1}, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)}_{\text{Term (ii)}} + \\
& \underbrace{\sum_{t=\tau}^{T} \mathrm{OPT}\left(\mathcal{C}_{t-1}, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right) - \mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left(\sum_{t=\tau}^{T} R\left(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)\right)}_{\text{Term (iii)}} + \\
& \underbrace{\mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left(\sum_{t=\tau}^{T} R\left(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)\right) - \mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left(\sum_{t=\tau}^{T} R\left(\pi_t; \boldsymbol{x}_t, \Theta, \mathcal{F}^{\boldsymbol{x}_t}\right)\right)}_{\text{Term (iv)}}
\end{aligned}
$$

At a high level, Terms (i) and (iv) account for the cumulative regret from transition error, i.e., the discrepancy between $\mathcal{F}^{\boldsymbol{x}_t}$ and $\hat{\mathcal{F}}^{\boldsymbol{x}_t}$. Term (iii) represents the decision error arising from imprecise estimates of the model parameters $\Theta = \{\boldsymbol{\theta}_l\}_{l\in\mathcal{H}} \cup \{d_l\}_{l\in[H]}$. Thus, more accurate

estimators of $\mathcal{F}^{\boldsymbol{x}_t}$ and Θ yield lower cumulative regret. Meanwhile, Term (ii) measures the gap in regret between a chosen point and the optimal point within a feasible set. Our goal is to show $\Theta \in \mathcal{C}_t$ for all $t \geq \tau$ (after the exploration phase) with high probability, ensuring that Term (ii) can be negative with high probability. This is proved in Lemma 2.4.3. By Lemma 2.4.1, with probability at least $1 - \frac{1}{T^4}$, Term (i) in Eqn. (2.9) is bounded by $\widetilde{\mathcal{O}}\big(dH^2\sqrt{T}\big)$. A similar argument implies that Term (iv) is also bounded by $\widetilde{\mathcal{O}}\big(dH^2\sqrt{T}\big)$ with probability at least $1 - \frac{1}{T^4}$. Further, Lemma 2.4.2 shows that, with probability at least $1 - \frac{2}{T^3}$, Term (iii) is bounded by $\widetilde{\mathcal{O}}\big(H^2\sqrt{dT}\big)$. Combining these results, we conclude that, with probability at least $1 - \frac{6}{T^3}$, the overall regret Reg_T is $\widetilde{\mathcal{O}}\big(dH^2\sqrt{T}\big)$. □

A.3.1 Proof for Theorem 2.4.1

Theorem (Theorem 2.4.1 restated)**.** *Let $\delta \geq \frac{1}{T^4 H}$ and $N_{t,l} \geq \underline{n_l}$. With probability at least $1-\delta$, the estimation error $\big|\hat{d}_l^t - d_l\big|$, with $\hat{d}_l^t$ defined in Eqn. (2.3), is bounded by:*

$$\big|\hat{d}_l^t - d_l\big| \leq \frac{4H\boldsymbol{B}_d\sqrt{d\log\big(1+\frac{T}{2d}\big)\gamma} + \sqrt{2e\boldsymbol{B}_d\boldsymbol{B}_x\boldsymbol{B}_\theta\log\big(\frac{2}{\delta}\big)}}{\boldsymbol{b}\sqrt{N_{t,l}}}.$$

γ is as defined in Lemma 2.3.1 and $\underline{n_l}$ defined in Remark 2.3.1.

Proof. The key idea in estimating d_l is to isolate observations that are "purified" with respect to d_l. Define

$$\mathbf{D}_{t,l} \;:=\; \{\, h \;\mid\; \mathcal{S}_{h,1}^t = l,\; \boldsymbol{o}_h^t = 0\},$$

which collects the episodes for user t where: The second-to-last bid winning is l episodes before the most recent bid winning ($\mathcal{S}_{h,1}^t = l$); The current bid is lost ($\boldsymbol{o}_h^t = 0$), meaning no new advertisement is shown. Under these conditions, the product-conversion observations $\{\boldsymbol{y}_h^t\}_{h\in\mathbf{D}_{t,l}}$ satisfy

$$\boldsymbol{y}_h^t \;\sim\; \mathrm{Poi}\big(d_l\,\langle \boldsymbol{\theta}_{\mathcal{S}_{h,2}^t},\, \mathbf{x}_t\rangle\big).$$

Because these observations are purified within the same l that defines d_l, they provide reliable information for estimating the long-term advertising impact parameter d_l. To estimate the long-term advertising impact parameter d_l, we adopt a two-stage procedure. First, we construct the estimates $\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}$ for all $s \in [t]$. Then, we substitute these estimates into the negative log-likelihood for d_l (see Eqn. (A.2)), yielding our two-stage estimator for d_l.

$$\mathcal{L}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}) = \sum_{s=1}^{t} \sum_{h \in \mathbf{D}_{s,l}} d_l \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle - \boldsymbol{y}^s_h \log \left(d_l \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle \right). \tag{A.2}$$

Taking the derivative of the negative log-likelihood with respect to d_l yields:

$$\nabla_{d_l} \mathcal{L}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}) = \sum_{s=1}^{t} \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle - \frac{\boldsymbol{y}^s_h}{d_l} = 0 \tag{A.3}$$

This expression guides the next step of solving for d_l. By solving Eqn. (A.3), we obtain the estimator $\hat{d}^t_l$ for d_l as follows:

$$\hat{d}^t_l = \frac{\sum_{s=1}^{t} \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{y}^s_h}{\sum_{s=1}^{t} \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle}.$$

Here, $\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}$ denotes the updated estimate of the advertisement's impact for user s. Next, we bound the estimation error of $\hat{d}^t_l$. We express $\boldsymbol{y}^s_h$ as

$$\boldsymbol{y}^s_h = d_l \, \mathbf{x}_s^\top \boldsymbol{\theta}_{S^s_{h,2}} + \eta^s_h,$$

where η^s_h is sub-exponential with parameters $(\nu^2, \alpha) = \left(e d_l \, \mathbf{x}_s^\top \boldsymbol{\theta}_{S^s_{h,2}}, 1 \right)$ (see Lemma A.4.1). Based on this formulation, the estimation error $\left| \hat{d}^t_l - d_l \right|$ can be decomposed as follows.

$$
\begin{aligned}
|\hat{d}_l^t - d_l| &= \left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{y}_h^s}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} - d_l \right| \\
&= \left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \left(d_l \boldsymbol{x}_s^\top \boldsymbol{\theta}_{S^s_{h,2}} + \eta_h^s \right)}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} - d_l \right| \\
&= \left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \eta_h^s}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} + d_l + \frac{d_l \sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{x}_s^\top \left(\boldsymbol{\theta}_{S^s_{h,2}} - \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} \right)}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} - d_l \right| \\
&\leq \left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \eta_h^s}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} \right| + d_l \left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{x}_s^\top \left(\boldsymbol{\theta}_{S^s_{h,2}} - \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} \right)}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} \right| \\
&\leq \underbrace{\left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \eta_h^s}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} \right|}_{\text{Term (i)}} + \underbrace{\frac{\boldsymbol{B}_d}{N_{t,l} \boldsymbol{b}} \left| \sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{x}_s^\top \left(\boldsymbol{\theta}_{S^s_{h,2}} - \hat{\boldsymbol{\theta}}_{S^s_{h,2}} \right) \right|}_{\text{Term (ii)}}
\end{aligned}
\tag{A.4}
$$

To upper bound the estimation error (see Eqn. (A.4)), it suffices to bound Term (i) and Term (ii) in Eqn. (A.4) respectively. We can use sub-exponential concentration bound for Term (i), shown as follows.

$$
\begin{aligned}
\mathbb{P}\left(\left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{\eta}_h^s}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}_{S^s_{h,2}}, \boldsymbol{x}_s \rangle} \right| > \epsilon \right) &\leq \mathbb{P}\left(\left| \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \boldsymbol{\eta}_h^s}{N_{t,l} \boldsymbol{b}} \right| > \epsilon \right), \boldsymbol{\eta}_h^s \sim \text{SubE}\left(e d_l \boldsymbol{x}_s^\top \boldsymbol{\theta}_{S^s_{h,2}}, 1 \right) \\
&\leq \mathbb{P}\left(\frac{1}{\boldsymbol{b}} \left| \frac{1}{N_{t,l}} \sum_{i=1}^{N_{t,l}} X_i \right| > \epsilon \right), X_i \sim \text{SubE}(e \boldsymbol{B}_d \boldsymbol{B}_x \boldsymbol{B}_\theta, 1) \\
&\leq \delta.
\end{aligned}
\tag{A.5}
$$

We derive Eqn. (A.5) by applying the tail bounds for sub-exponential random variables (Lemmas A.4.4, A.4.5, and A.4.6) and setting

$$\epsilon \;=\; \sqrt{\frac{2\,e\,\boldsymbol{B}_d\,\boldsymbol{B}_x\,\boldsymbol{B}_\theta}{\boldsymbol{b}^2\,N_{t,l}}\;\log\!\big(\tfrac{2}{\delta}\big)}.$$

We also use the accelerated convergence rate for sub-exponential variables under the conditions $N_{t,l} \;>\; \frac{32\,\log(HT)}{e\,\boldsymbol{B}_d\,\boldsymbol{B}_x\,\boldsymbol{B}_\theta\,\boldsymbol{b}^2}$ and $\delta \;\geq\; \frac{1}{2HT^4}$, which ensure $0 \;\leq\; \epsilon \;\leq\; e\,\boldsymbol{B}_d\,\boldsymbol{B}_x\,\boldsymbol{B}_\theta$. Therefore, with probability at least $1-\delta$, we have Term (i) $\leq \sqrt{\frac{2\,e\,\boldsymbol{B}_d\,\boldsymbol{B}_x\,\boldsymbol{B}_\theta}{\boldsymbol{b}^2\,N_{t,l}}\;\log\!\big(\frac{2}{\delta}\big)}$. Then it remains to bound Term (ii) in Eqn. (A.4). Define the dataset $\mathcal{F}^l_{s,l'} := \{\, h \mid S^s_{h,1} = l,\; S^s_{h,2} = l',\; \boldsymbol{o}^s_h = 0\}$. Observe that

$$\sum_{l'=1}^{H} |\mathcal{F}^l_{s,l'}| \;=\; |\mathbf{D}_{s,l}|,$$

meaning $\mathcal{F}^l_{s,l'}$ partitions $\mathbf{D}_{s,l}$ according to $S^s_{h,2}$, the second-to-last winning state. Let $n_{s,l'} := |\mathcal{F}^l_{s,l'}|$. Then

$$\sum_{s=1}^{t}\sum_{l'=1}^{H} n_{s,l'} \;=\; N_{t,l}.$$

Since $\|\theta_l\|_2 \leq \boldsymbol{B}_\theta$ for all l, and $\|\mathbf{x}_s\|_2 \leq \boldsymbol{B}_x$ for each $s \in [T]$, define

$$r^s_h \;=\; \Big|\big(\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} - \boldsymbol{\theta}_{S^s_{h,2}}\big)^\top \mathbf{x}_s\Big| \;\leq\; \|\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} - \boldsymbol{\theta}_{S^s_{h,2}}\|_{\mathbf{V}^s_{l'}}\;\|\mathbf{x}_s\|_{(\mathbf{V}^s_{l'})^{-1}},$$

where $l' = S^s_{h,2}$. For convenience, let $\sqrt{\gamma^s_{l'}} \;=\; \|\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}} - \boldsymbol{\theta}_{S^s_{h,2}}\|_{\mathbf{V}^s_{l'}}$. Then, with probability at least $1-\delta$, we have

$$
\begin{aligned}
\sum_{s=1}^{t}\sum_{h\in\mathbf{D}_{s,l}}\left|\left(\hat{\boldsymbol{\theta}}^s_{S^s_{h,2}}-\boldsymbol{\theta}_{S^s_{h,2}}\right)^\top \boldsymbol{x}_s\right| =& \sum_{s=1}^{t}\sum_{h\in\mathbf{D}_{s,l}} r^s_h \\
\leq& \sqrt{N_{t,l}}\sqrt{\sum_{s=1}^{t}\sum_{h\in\mathbf{D}_{s,l}}(r^s_h)^2} \\
=& \sqrt{N_{t,l}}\sqrt{\sum_{s=1}^{t}\sum_{l'=1}^{H}\sum_{h\in\mathcal{F}^l_{s,l'}}(r^s_h)^2} \\
=& \sqrt{N_{t,l}}\sqrt{\sum_{s=1}^{t}\sum_{l'=1}^{H} n_{s,l'}\gamma\min\left(\|\boldsymbol{x}_s\|_{(\mathcal{V}^s_{l'})^{-1}},1\right)} \\
=& \sqrt{N_{t,l}}\sqrt{\sum_{l'=1}^{H}\sum_{s=1}^{t} n_{s,l'}\gamma\min\left(\|\boldsymbol{x}_s\|_{(\mathcal{V}^s_{l'})^{-1}},1\right)} \\
\leq& \sqrt{N_{t,l}}\sqrt{\sum_{l'=1}^{H}\gamma\sum_{s=1}^{t} n_{s,l'}\min\left(\|\boldsymbol{x}_s\|_{(\mathcal{V}^s_{l'})^{-1}},1\right)} \\
\leq& \sqrt{N_{t,l}H}\sqrt{\sum_{l'=1}^{H}\gamma\sum_{s=1}^{t}\min\left(\|\boldsymbol{x}_s\|_{(\mathcal{V}^s_{l'})^{-1}},1\right)} \\
\leq& H\sqrt{N_{t,l}}\sqrt{\gamma\max_{l'\in[H]}\sum_{s=1}^{t}\min\left(\|\boldsymbol{x}_s\|_{(\mathcal{V}^s_{l'})^{-1}},1\right)} \\
\leq& 2H\sqrt{d}\sqrt{N_{t,l}\log\left(1+\frac{T}{2d}\right)\gamma}.
\end{aligned}
\tag{A.6}
$$

Therefore, with probability at least $1-\delta$, $\delta \geq \frac{1}{HT^4}$, we have

$$
|\hat{d}^t_l - d_l| \leq \frac{4H\boldsymbol{B}_d\sqrt{d\log\left(1+\frac{T}{2d}\right)\gamma}+\sqrt{2e\boldsymbol{B}_d\boldsymbol{B}_x\boldsymbol{B}_\theta\log(2/\delta)}}{\boldsymbol{b}}\frac{1}{\sqrt{N_{t,l}}}.
$$

□

A.3.2 Proof of Lemma 2.4.1

Lemma (Lemma 2.4.1 restated). *For $\delta \geq 1/T^4$, with probability at least $1-\delta$, Term (i) in Eqn. (2.9) is bounded above by $\mathcal{O}(dH^2\sqrt{T}\sqrt{\log((1+T)/\delta)\log(1+T/d)})$.*

Proof. In essence, the proof of Lemma 2.4.1 relies on two main steps. First, we show that Term (i) can be bounded by the cumulative estimation error in transition, expressed as

$$\sum_{t=1}^{T} \sup_b \sup_h \left| \hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) - \mathcal{F}_h(b, \boldsymbol{x}_t) \right|.$$

Second, we demonstrate that, when $\hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t)$ is estimated using Algorithm 3, this cumulative estimation error remains small. Combined, these two steps yield the desired upper bound on Term (i).

Step 1: Upper bound Term (i) by the cumulative estimation error in transition.

Term (i) captures the regret induced by transition error, as shown in Eqn. (A.7). The final inequality in Eqn. (A.7) holds because $R(\pi_t^*; \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) - \mathrm{OPT}\big(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\big) \leq 0$, due to the definition of $\mathrm{OPT}\big(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\big)$. We then apply the Simulation Lemma (Lemma A.3.1) to bound $R(\pi_t^*; \Theta, \mathcal{F}^{\boldsymbol{x}_t}) - R(\pi_t^*; \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t})$.

$$\begin{aligned}
\text{Term (i)} &= \sum_{t=\tau}^{T} \mathrm{OPT}\left(\Theta, \boldsymbol{x}_t, \mathcal{F}^{\boldsymbol{x}_t}\right) - \sum_{t=\tau}^{T} \mathrm{OPT}\left(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right) \\
&= \sum_{t=\tau}^{T} R(\pi_t^*; \Theta, \mathcal{F}^{\boldsymbol{x}_t}) - R(\pi_t^*; \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) + R(\pi_t^*; \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) - \mathrm{OPT}\left(\Theta, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right) \\
&\leq \sum_{t=\tau}^{T} R(\pi_t^*; \Theta, \mathcal{F}^{\boldsymbol{x}_t}) - R(\pi_t^*; \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}).
\end{aligned} \tag{A.7}$$

Lemma A.3.1 (Simulation Lemma [Kearns and Singh, 2002]). *For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, if*

$\sum_{s' \in \mathcal{S}} \Big| \mathbb{P}(s' \mid s, a) \;-\; \mathbb{P}'(s' \mid s, a) \Big| \;\leq\; \epsilon_1 \; and \, \forall h \in [H], \; \big| r_h(s,a) \;-\; r'_h(s,a) \big| \;\leq\; \epsilon_2, then$

$$\Big| \mathbb{E}_{\{s_h, a_h\}_{h=1}^{H} \sim \mathcal{M}, \pi} \Big[\sum_{h \in [H]} r_h(s_h, a_h) \Big] - \mathbb{E}_{\{s_h, a_h\}_{h=1}^{H} \sim \mathcal{M}', \pi} \Big[\sum_{h \in [H]} r'_h(s_h, a_h) \Big] \Big| \leq \frac{H(H-1)\,\epsilon_1}{2} + H\,\epsilon_2.$$

To find ϵ_1 in Lemma A.3.1, we compare two MDPs:

$$\mathcal{M}_t, \text{ induced by policy } \pi_t^*, \; \theta, \text{ and HOB distribution } \mathcal{F}^{\boldsymbol{x}_t},$$

and

$$\hat{\mathcal{M}}_t, \text{ induced by policy } \pi_t^*, \; \theta, \text{ and HOB distribution } \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}.$$

Given state $s = (s_1, s_2)$, we have:

$$\begin{aligned}
\sum_{s' \in S} \Big| \mathbb{P}_h(s'|s,a) - \hat{\mathbb{P}}_h(s'|s,a) \Big| &= \Big| \mathbb{P}_h(s' = (s_1+1, s_2)|s,a) - \hat{\mathbb{P}}_h(s' = (s_1+1, s_2)|s,a) \Big| \\
&\quad + \Big| \mathbb{P}_h(s' = (1, s_1)|s,a) - \hat{\mathbb{P}}_h(s' = (1, s_1)|s,a) \Big| \\
&= 2 \Big| \hat{\mathcal{F}}_h^{t-1}(a, \boldsymbol{x}_t) - \mathcal{F}_h(a, \boldsymbol{x}_t) \Big| \\
&\leq 2 \sup_b \sup_h \Big| \hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) - \mathcal{F}_h(b, \boldsymbol{x}_t) \Big|.
\end{aligned} \tag{A.8}$$

Define

$$\epsilon_t \;=\; \sup_b \sup_h \Big| \hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) \;-\; \mathcal{F}_h(b, \boldsymbol{x}_t) \Big|.$$

From Eqn. (A.8), it follows that the parameter ϵ_1 in Lemma A.3.1 can be taken as $2\epsilon_t$. In the following, we bound the difference between R_h^t and $R_h^{t'}$ to find ϵ_2 in Lemma A.3.1. By definition, $R_h^t(S_h^t, A_h^t, \boldsymbol{x}_t) = d_{S_{h,1}^t} \langle \boldsymbol{\theta}_{S_{h,2}^t}, \boldsymbol{x}_t \rangle \big(1 - \mathcal{F}_h(A_h^t, \boldsymbol{x}_t)\big) + \big(\langle \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t \rangle - p_h(A_h^t, \boldsymbol{x}_t) \big) \mathcal{F}_h(A_h^t, \boldsymbol{x}_t)$. Here, $p_h(A_h^t, \boldsymbol{x}_t)$ is the second highest price conditioned on winning. Since $p_h^{\boldsymbol{x}_t}(b) = b - \frac{1}{\mathcal{F}_h(b, \boldsymbol{x}_t)} \int_0^b \mathcal{F}_h(v, \boldsymbol{x}_t)\, dv$,

we obtain $\hat{p}_h(A_h^t, \boldsymbol{x}_t)\,\hat{\mathcal{F}}_h^{t-1}(A_h^t, \boldsymbol{x}_t) - p_h(A_h^t, \boldsymbol{x}_t)\,\mathcal{F}_h(A_h^t, \boldsymbol{x}_t) = \int_0^{A_h^t} \mathcal{F}_h(v, \boldsymbol{x}_t)\,dv - \int_0^{A_h^t} \hat{\mathcal{F}}_h^{t-1}(v, \boldsymbol{x}_t)\,dv$. In addition, we can show

$$\begin{aligned}\int_0^b f(v)dv - \int_0^b \hat{f}(v)dv &\leq \int_0^b \sup_x |f(x) - \hat{f}(x)|dv \\ &= \sup_x |f(x) - \hat{f}(x)| \int_0^b 1dv \\ &= b \sup_x |f(x) - \hat{f}(x)|.\end{aligned}$$

Therefore, we can bound the reward difference $|R_h^t(s,a) - R_h^{t'}(s,a)|$ by Eqn. (A.9).

$$\begin{aligned}\left|R_h^t - R_h^{\prime t}\right| &\leq \left(d_{S_{h,1}^t}\langle \boldsymbol{\theta}_{S_{h,2}^t}, \boldsymbol{x}_t\rangle + \langle \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t\rangle + \boldsymbol{B}_{\mathcal{A}}\right) \sup_{b\in[0,\boldsymbol{B}_{\mathcal{A}}]} \sup_h \left|\hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) - \mathcal{F}_h(b, \boldsymbol{x}_t)\right| \\ &\leq ((\boldsymbol{B}_d + 1)\boldsymbol{B}_x\boldsymbol{B}_\theta + \boldsymbol{B}_{\mathcal{A}}) \sup_{b\in[0,\boldsymbol{B}_{\mathcal{A}}]} \sup_h \left|\hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) - \mathcal{F}_h(b, \boldsymbol{x}_t)\right| \\ &= ((\boldsymbol{B}_d + 1)\boldsymbol{B}_x\boldsymbol{B}_\theta + \boldsymbol{B}_{\mathcal{A}})\epsilon_t,\end{aligned} \tag{A.9}$$

Combining Eqn. (A.8) and Eqn. (A.9), Lemma A.3.1 implies

$$\begin{aligned}R(\pi_t^*; \Theta, \mathcal{F}^{\boldsymbol{x}_t}) - R(\pi_t^*; \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) &\leq H(H-1)\epsilon_t + H((\boldsymbol{B}_d + 1)\boldsymbol{B}_x\boldsymbol{B}_\theta + \boldsymbol{B}_{\mathcal{A}})\epsilon_t \\ &= H^2\epsilon_t + H\left((\boldsymbol{B}_d + 1)\boldsymbol{B}_x\boldsymbol{B}_\theta + \boldsymbol{B}_{\mathcal{A}} - 1\right)\epsilon_t.\end{aligned} \tag{A.10}$$

Therefore, by Eqn. (A.10), bounding Term (i) reduces to bounding

$$\sum_{t=\tau}^{T}\Big(H^2\,\epsilon_t \;+\; H\Big((\boldsymbol{B}_d + 1)\,\boldsymbol{B}_x\,\boldsymbol{B}_\theta \;+\; \boldsymbol{B}_{\mathcal{A}} \;-\; 1\Big)\epsilon_t\Big).$$

Step 2: Construct high probability bound for cumulative estimation error in transition

The next step is to derive a high-probability bound for $\sum_{t=\tau}^{T}\epsilon_t$. By Assumption 2, the

Highest Other Bids (HOB) distribution satisfies $\log(m_h^t) \sim \mathcal{N}\big(\langle \boldsymbol{x}_t, \boldsymbol{\beta}_h \rangle, \sigma_h^2\big)$, suggesting that the transition estimation error takes the form

$$\left|\hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) - \mathcal{F}_h(b, \boldsymbol{x}_t)\right| = \left|\Phi\left(\frac{\log(b) - \langle \boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_h^{t-1}\rangle}{\hat{\sigma}_h^{t-1}}\right) - \Phi\left(\frac{\log(b) - \langle \boldsymbol{x}_t, \boldsymbol{\beta}_h\rangle}{\sigma_h}\right)\right|. \quad \text{(A.11)}$$

Using a first-order Taylor expansion of the standard normal CDF $\Phi(\cdot)$, we obtain $\big|\Phi(a) - \Phi(a+\Delta)\big| \leq |\Delta|\phi(a)$, ϕ is the standard normal pdf. This inequality allows us to further bound the estimation error by $\left|\hat{\mathcal{F}}_h^{t-1}(b, \boldsymbol{x}_t) - \mathcal{F}_h(b, \boldsymbol{x}_t)\right| \leq |\Delta_h^t|\phi\left(\frac{\log(b) - \langle \boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_h^{t-1}\rangle}{\hat{\sigma}_h^{t-1}}\right)$, where $|\Delta_h^t|$ is defined and bounded by follows.

$$\begin{aligned}
|\Delta_h^t| &= \left|\frac{\log(b) - \langle \boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_h^{t-1}\rangle}{\hat{\sigma}_h^{t-1}} - \frac{\log(b) - \langle \boldsymbol{x}_t, \boldsymbol{\beta}_h\rangle}{\sigma_h}\right| \\
&= \left|\frac{\log(b) - \langle \boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_h^{t-1}\rangle}{\hat{\sigma}_h^{t-1}} - \frac{\log(b) - \langle \boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_h^{t-1}\rangle}{\sigma_h} + \frac{\log(b) - \langle \boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_h^{t-1}\rangle}{\sigma_h} - \frac{\log(b) - \langle \boldsymbol{x}_t, \boldsymbol{\beta}_h\rangle}{\sigma_h}\right| \\
&\leq \frac{1}{\sigma_h}\left|\boldsymbol{x}_t^\top\left(\hat{\boldsymbol{\beta}}_h^{t-1} - \boldsymbol{\beta}_h\right)\right| + \left|\log(b) - \langle \boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_h^{t-1}\rangle\right|\left|\frac{1}{\hat{\sigma}_h^{t-1}} - \frac{1}{\sigma_h}\right| \\
&\leq \frac{1}{\sigma_h}\left|\boldsymbol{x}_t^\top\left(\hat{\boldsymbol{\beta}}_h^{t-1} - \boldsymbol{\beta}_h\right)\right| + \frac{\big(\log(\boldsymbol{B}_{\mathcal{A}}) + \boldsymbol{B}_x\boldsymbol{B}_\beta\big)\left(\sigma_h + \hat{\sigma}_h^{t-1}\right)}{\left(\sigma_h\hat{\sigma}_h^{t-1}\right)^3}\left|\sigma_h^2 - \left(\hat{\sigma}_h^{t-1}\right)^2\right|.
\end{aligned} \quad \text{(A.12)}$$

Note that we estimate the variance σ_h^2 using second-stage empirical estimates. Concretely,

$$\big(\hat{\sigma}_h^{t-1}\big)^2 \;=\; \frac{1}{t-1}\sum_{s=1}^{t-1}\Big(\log(m_h^s) - \boldsymbol{x}_s^\top\hat{\boldsymbol{\beta}}_h^s\Big)^2.$$

Meanwhile, by definition,

$$\sigma_h^2 \;=\; \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{E}\Big[\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\Big]^2.$$

To complete the argument, it suffices to show that these two quantities are close with high probability.

$$\begin{aligned}
\left|\sigma_h^2 - \left(\hat{\sigma}_h^{t-1}\right)^2\right| &= \left|\frac{1}{t-1}\sum_{s=1}^{t-1}\left[\left(\log(m_h^s) - \boldsymbol{x}_s^\top\hat{\boldsymbol{\beta}}_h^s\right)^2 - \mathbb{E}\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\right)^2\right]\right| \\
&= \left|\frac{1}{t-1}\sum_{s=1}^{t-1}\left[\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h + \boldsymbol{x}_s^\top\boldsymbol{\beta}_h - \boldsymbol{x}_s^\top\hat{\boldsymbol{\beta}}_h^s\right)^2 - \mathbb{E}\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\right)^2\right]\right| \\
&= \left|\frac{1}{t-1}\sum_{s=1}^{t-1}\left[\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\right)^2 - \mathbb{E}\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\right)^2\right]\right| \\
&\quad + \left|\frac{1}{t-1}\sum_{s=1}^{t-1}\left(\boldsymbol{x}_s^\top\boldsymbol{\beta}_h - \boldsymbol{x}_s^\top\hat{\boldsymbol{\beta}}_h^s\right)^2\right| \\
&\quad + 2\left|\frac{1}{t-1}\sum_{s=1}^{t-1}\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\right)\left(\boldsymbol{x}_s^\top\boldsymbol{\beta}_h - \boldsymbol{x}_s^\top\hat{\boldsymbol{\beta}}_h^s\right)\right|.
\end{aligned}$$

Using Eqns. (A.11), (A.12), and (2.6), we derive the following equations:

$$\begin{aligned}
\sum_{t=1}^{T}\epsilon_t \le\ & \underbrace{\sum_{t=\tau}^{T}\sup_{h\in H}\frac{1}{\sigma_h}\left|\boldsymbol{x}_t^\top\left(\hat{\boldsymbol{\beta}}_h^{t-1} - \boldsymbol{\beta}_h\right)\right|}_{\text{Term A}} \\
& + 2\left(\log(\boldsymbol{B}_{\mathcal{A}}) + \boldsymbol{B}_x\boldsymbol{B}_\theta\right)^2\underbrace{\sum_{t=1}^{T}\sup_{h\in H}C_h\left|\frac{1}{t-1}\sum_{s=1}^{t-1}\boldsymbol{x}_s^\top\left(\hat{\boldsymbol{\beta}}_h^s - \boldsymbol{\beta}_h\right)\right|}_{\text{Term B}} \\
& + \left(\log(\boldsymbol{B}_{\mathcal{A}}) + \boldsymbol{B}_x\boldsymbol{B}_\theta\right)\underbrace{\sum_{t=1}^{T}\sup_{h\in H}\left|\frac{1}{t-1}\sum_{s=1}^{t-1}\left(\boldsymbol{x}_s^\top\boldsymbol{\beta}_h - \boldsymbol{x}_s^\top\hat{\boldsymbol{\beta}}_h^s\right)^2\right|}_{\text{Term C}} \\
& + \left(\log(\boldsymbol{B}_{\mathcal{A}}) + \boldsymbol{B}_x\boldsymbol{B}_\theta\right)\underbrace{\sum_{t=1}^{T}\left|\frac{1}{t-1}\sum_{s=1}^{t-1}\left[\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\right)^2 - \mathbb{E}\left(\log(m_h^s) - \boldsymbol{x}_s^\top\boldsymbol{\beta}_h\right)^2\right]\right|}_{\text{Term D}},
\end{aligned} \tag{A.13}$$

where $C_h = \frac{\sigma_h + \hat{\sigma}_h^{t-1}}{(\sigma_h\hat{\sigma}_h^{t-1})^3} \le \frac{2\bar{\sigma}}{\underline{\sigma}^6}$, by Assumption 3. In what follows, we bound each term one by

one. To bound Term A, we show that for all h, with probability at least $1-\delta$, the following holds:

$$\begin{aligned}
\sum_{t=1}^{T}\left|\boldsymbol{x}_t^\top\left(\hat{\boldsymbol{\beta}}_h^{t-1}-\boldsymbol{\beta}_h\right)\right| &\leq \sqrt{T\sum_{t=1}^{T}\left|\boldsymbol{x}_t^\top\left(\hat{\boldsymbol{\beta}}_h^{t-1}-\boldsymbol{\beta}_h\right)\right|^2} \\
&\leq \sqrt{T\sum_{t=1}^{T}\min\left\{2\boldsymbol{B}_\beta\boldsymbol{B}_x, \|\boldsymbol{x}_t\|^2_{\left(\Sigma_h^{t-1}\right)^{-1}}\|\hat{\boldsymbol{\beta}}_h^{t-1}-\boldsymbol{\beta}_h\|^2_{\Sigma_h^{t-1}}\right\}} \\
&\leq \sqrt{T\sum_{t=1}^{T}\|\hat{\boldsymbol{\beta}}_h^{t-1}-\boldsymbol{\beta}_h\|^2_{\Sigma_h^{t-1}}\min\left\{\frac{2\boldsymbol{B}_\beta\boldsymbol{B}_x}{\|\hat{\boldsymbol{\beta}}_h^{t-1}-\boldsymbol{\beta}_h\|^2_{\Sigma_h^{t-1}}}, \|\boldsymbol{x}_t\|^2_{\left(\Sigma_h^{t-1}\right)^{-1}}\right\}} \\
&\leq \left(\sigma_h^2\sqrt{d\log\left(\frac{1+T\boldsymbol{B}_x^2/\lambda}{\delta}\right)}+\sqrt{\lambda}\boldsymbol{B}_\beta\right)\sqrt{T\sum_{t=1}^{T}\min\left\{1, \|\boldsymbol{x}_t\|^2_{\left(\Sigma_h^{t-1}\right)^{-1}}\right\}}, \\
&\quad \text{by Lemma A.4.2} \\
&\leq \left(\sigma_h^2\sqrt{d\log\left(\frac{1+T\boldsymbol{B}_x^2/\lambda}{\delta}\right)}+\sqrt{\lambda}\boldsymbol{B}_\beta\right)\sqrt{Td\log\left(1+\frac{T\boldsymbol{B}_x}{\lambda d}\right)} \\
&\leq d\sqrt{T}\sup_{h\in H}\left(\sigma_h^2\sqrt{\log\left(\frac{1+T\boldsymbol{B}_x^2/\lambda}{\delta}\right)}+\sqrt{\lambda}\boldsymbol{B}_\beta\right)\sqrt{\log\left(1+\frac{T\boldsymbol{B}_x}{\lambda d}\right)}.
\end{aligned}$$

Therefore,

$$\sum_{t=1}^{T}\sup_{h\in H}\frac{1}{\sigma_h}|\boldsymbol{x}_t^\top(\hat{\boldsymbol{\beta}}_h^{t-1}-\boldsymbol{\beta}_h)| \leq d\sqrt{T}\Big(\frac{\bar{\sigma}^2}{\underline{\sigma}}\sqrt{\log\Big(\frac{1+\frac{T\boldsymbol{B}_x^2}{\lambda}}{\delta}\Big)}+\sqrt{\lambda}\,\boldsymbol{B}_\beta\Big)\sqrt{\log\Big(1+\frac{T\boldsymbol{B}_x}{\lambda\, d}\Big)}.$$

This shows that Term C is on the order of $\mathcal{O}(\log T)$. For Term B, a similar argument implies that, with probability at least $1-\delta$,

$$\Big|\frac{1}{t-1}\sum_{s=1}^{t-1}\boldsymbol{x}_s^\top(\hat{\boldsymbol{\beta}}_h^s-\boldsymbol{\beta}_h)\Big| \;\leq\; \frac{d}{\sqrt{t-1}}\Big(\frac{\bar{\sigma}^2}{\underline{\sigma}}\sqrt{\log\Big(\frac{1+\frac{T\boldsymbol{B}_x^2}{\lambda}}{\delta}\Big)}\;+\;\sqrt{\lambda}\,\boldsymbol{B}_\beta\Big)\sqrt{\log\Big(1+\frac{T\boldsymbol{B}_x}{\lambda\, d}\Big)}.$$

Therefore, we conclude the following bound.

$$\sum_{t=\tau}^{T} \sup_{h\in H} C_h \left| \frac{1}{t-1} \sum_{s=1}^{t-1} \boldsymbol{x}_s^\top \left(\hat{\boldsymbol{\beta}}_h^s - \boldsymbol{\beta}_h \right) \right| \leq \frac{2\bar{\sigma}}{\underline{\sigma}^6} d\sqrt{T} \left(\frac{\bar{\sigma}^2}{\underline{\sigma}} \sqrt{\log\left(\frac{1 + T\boldsymbol{B}_x^2/\lambda}{\delta} \right)} + \sqrt{\lambda} \boldsymbol{B}_\beta \right) \times \sqrt{\log\left(1 + \frac{T\boldsymbol{B}_x}{\lambda d} \right)}.$$

It remains to bound Term D. First, observe that

$$\left[\left(\log(m_h^s) - \boldsymbol{x}_s^\top \boldsymbol{\beta}_h \right)^2 - \mathbb{E}\left(\log(m_h^s) - \boldsymbol{x}_s^\top \boldsymbol{\beta}_h \right)^2 \right] = \sigma_h^2 \left(\frac{\left(\log(m_h^s) - \boldsymbol{x}_s^\top \boldsymbol{\beta}_h \right)^2}{\sigma_h^2} - 1 \right) = \sigma_h^2 \left(\chi_1^2 - 1 \right).$$

Since $\chi_1^2 \sim \text{SubE}(4, 4)$, by Lemma A.3.2 and a union bound argument, we can show that, with probability at least $1 - \delta$, Term D is bounded above by

$$\sqrt{T} \sqrt{8 \log\left(\frac{2T}{\delta} \right)} + \mathcal{O}\left(\log(T) \right).$$

The $\mathcal{O}\left(\log(T) \right)$ term arises from "burning in" the first $64 \log(T)$ which guarantees $\delta \geq 1/4T^4$ and ensure quadratic decay of sub-exponential bound.

Lemma A.3.2 (Example 2.11 in Wainwright [2019b]). *A chi-squared (χ^2) random variable with n degrees of freedom, denoted by $Y \sim \chi_n^2$, can be represented as $Y = \sum_{k=1}^n Z_k^2$, where $Z_k \sim N(0,1)$ are i.i.d. variables. Since each Z_k^2 is sub-exponential with parameters $(\nu^2, \alpha) = (2, 4)$, we have $\mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^n Z_k^2 - 1 \right| \geq \sqrt{\frac{8}{n} \log \frac{2}{\delta}} \right) \leq \delta$, as long as $\delta \geq 2 \exp\left(-\frac{n}{8} \right)$.*

Bringing all these results together and applying union bound, we conclude that there exists a constant C, depending on $\boldsymbol{B}_x, \boldsymbol{B}_\theta, \boldsymbol{B}_\beta, \bar{\sigma}, \underline{\sigma}, \boldsymbol{B}_d$, and $\boldsymbol{B}_{\mathcal{A}}$, such that with probability at least $1 - \delta$ (where $\delta \geq \frac{1}{T^4}$), we have

$$\text{Term (i)} \leq C\, d\, H^2 \sqrt{T} \sqrt{\log\left(\frac{1+T}{\lambda\, \delta} \right) \log\left(1 + \frac{T}{\lambda\, d} \right)}.$$

By choosing $\lambda = 1$, we proves Lemma 2.4.1, we states that for $\delta \geq 1/T^4$, with probability at least $1-\delta$, Term (i) in Eqn. (2.9) is bounded above by $\mathcal{O}(dH^2\sqrt{T}\sqrt{\log((1+T)/\delta)\log(1+T/d)})$.

□

A.3.3 Proof for Lemma 2.4.2

Lemma (Lemma 2.4.2 restated)**.** *With probability at least* $1-2\delta$*, where* $\delta \geq \frac{1}{T^3}$*, Term (iii) in Eqn. (2.9) is bounded by* $\mathcal{O}\Big(H^2\sqrt{dT\log\big(\frac{4TH}{\delta}\big)}\,\log\big(1+\frac{T}{2d}\big)\Big)$.

Proof. Term (iii) reflects the portion of regret attributable to inaccuracies in estimating the true parameter Θ. Let $\tilde{\Theta}_{t-1} := \arg\max_{\Theta\in\mathcal{C}_{t-1}} R(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t})$. By definition,

$$R(\pi_t; \boldsymbol{x}_t, \tilde{\Theta}_{t-1}, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}) = \mathrm{OPT}\big(\mathcal{C}_{t-1}, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\big).$$

By Eqn. (A.14), we demonstrate that Term (iii) in (2.9) can be split into two parts corresponding to the cumulative estimation error of $\hat{\boldsymbol{\theta}}_l^t$ (see Terms 1 and 2 in (A.14)) and the cumulative estimation error of $\hat{d}_l$ (see Term 3 in (A.14)). Recall that $\hat{\boldsymbol{\theta}}_l^t$ is a variant of the online Newton estimator (see Algorithm 4), defined by

$$\hat{\boldsymbol{\theta}}_l^t = \arg\min_{\|\boldsymbol{\theta}\|_2\in\boldsymbol{B}_\theta}\Big\{\tfrac{1}{2}\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_l^{t-1}\|_{\mathbf{V}_l^t}^2 + \langle\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_l^{t-1}, \nabla\tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1})\rangle\Big\},$$

where $\nabla\tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1}) = (-\tilde{\boldsymbol{y}}_h^t + \mathbf{x}_t^\top\hat{\boldsymbol{\theta}}_l^{t-1})\mathbf{x}_t$ and $\mathbf{V}_l^t = \mathbf{V}_l^{t-1} + \frac{1}{2}\mathbf{x}_t\mathbf{x}_t^\top$. $\tilde{\boldsymbol{y}}_h^t$ is the truncated observation of $\boldsymbol{y}_h^t$. Because $\boldsymbol{y}_h^t$ follows a Poisson distribution and is linear in $\boldsymbol{\theta}_l$, it is a special case of a heavy-tailed distribution with bounded second moment. Consequently, Lemma 2.3.1 can be used to construct a high-confidence bound on $\hat{\boldsymbol{\theta}}_l^t$, which in turn controls Terms 1 and 2 in Eqn. (A.14).

$$
\begin{aligned}
\text{Term (iii)} &= \sum_{t=\tau}^{T} \mathrm{OPT}\left(\mathcal{C}_{t-1}, \boldsymbol{x}_t, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right) - \mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left(\sum_{t=\tau}^{T} R\left(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)\right) \\
&= \mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left[\sum_{t=\tau}^{T} R\left(\pi_t; \boldsymbol{x}_t, \tilde{\Theta}_{t-1}, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right) - R\left(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)\right] \\
&= \mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left[\sum_{t=\tau}^{T} \mathbb{E}_{S_h^t\sim\hat{P}_{t-1}^{\boldsymbol{x}_t}}\left(\sum_{h=1}^{H} R_h^t(S_h^t, A_h^t, \boldsymbol{x}_t)\right) - \sum_{t=\tau}^{T} R\left(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)\right] \\
&= \mathbb{E}(\sum_{t=\tau}^{T}\mathbb{E}(\sum_{h=1}^{H} \tilde{d}_{S_{h,1}^t}^{t-1}\langle\tilde{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1}, \boldsymbol{x}_t\rangle(1-\hat{\mathcal{F}}_h^{t-1}(A_h^t, \boldsymbol{x}_t)) \\
&+ (\langle\tilde{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1}, \boldsymbol{x}_t\rangle - \hat{p}_h(A_h^t, \boldsymbol{x}_{t-1}))\hat{\mathcal{F}}_h^{t-1}(A_h^t, \boldsymbol{x}_t))) \\
&- \mathbb{E}_{\pi_1,\pi_2,\ldots,\pi_T\sim\mathcal{G}}\left(\sum_{t=\tau}^{T} R\left(\pi_t; \boldsymbol{x}_t, \Theta, \hat{\mathcal{F}}_{t-1}^{\boldsymbol{x}_t}\right)\right), \text{where } A_h^t = \pi_t(S_h^t, \boldsymbol{x}_t) \\
&\leq \sum_{t=\tau}^{T} \mathbb{E}_{S_h^t\sim\hat{P}_{t-1}^{\boldsymbol{x}_t}}\left(\sum_{h=1}^{H}\left|\langle\tilde{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t\rangle\right| + \left|d_{S_{h,1}^t}\langle\boldsymbol{\theta}_{S_{h,2}^t}, \boldsymbol{x}_t\rangle - \tilde{d}_{S_{h,1}^t}^{t-1}\langle\tilde{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1}, \boldsymbol{x}_t\rangle\right|\right) \\
&\leq \sum_{t=\tau}^{T} \mathbb{E}_{S_h^t\sim\hat{P}_{t-1}^{\boldsymbol{x}_t}}(\sum_{h=1}^{H}\{\left|\langle\tilde{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t\rangle\right| + \boldsymbol{B}_{\mathcal{A}}\left|\langle\tilde{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,2}^t}, \boldsymbol{x}_t\rangle\right| \\
&+ \left|\tilde{d}_{S_{h,1}^t}^{t-1} - d_{S_{h,1}^t}\right|\boldsymbol{B}_\theta\boldsymbol{B}_x\}) \\
&= \underbrace{\sum_{t=\tau}^{T}\mathbb{E}\left(\sum_{h=1}^{H}\left|\langle\tilde{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} + \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t\rangle\right|\right)}_{\text{Term 1}} \\
&+ \boldsymbol{B}_{\mathcal{A}}\underbrace{\sum_{t=\tau}^{T}\mathbb{E}\left(\sum_{h=1}^{H}\left|\langle\tilde{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1} - \hat{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1} + \hat{\boldsymbol{\theta}}_{S_{h,2}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,2}^t}, \boldsymbol{x}_t\rangle\right|\right)}_{\text{Term 2}} + \\
&+ \boldsymbol{B}_\theta\boldsymbol{B}_x\underbrace{\sum_{t=\tau}^{T}\mathbb{E}\left(\sum_{h=1}^{H}\left|\tilde{d}_{S_{h,1}^t}^{t-1} - \hat{d}_{S_{h,1}^t}^{t-1} + \hat{d}_{S_{h,1}^t}^{t-1} - d_{S_{h,1}^t}\right|\right)}_{\text{Term 3}}.
\end{aligned}
\tag{A.14}
$$

To bound Term 1 in Eqn. (A.14), we first introduce n_t^l, where

$$n_t^l \;:=\; \big|\{\, h \in [H] : S_{h,1}^t = l\}\big|; n_t^l \;\leq\; H.$$

Then we have the following inequality:

$$\sum_{t=\tau}^{T} \mathbb{E}\left(\sum_{h=1}^{H}\left|\langle \tilde{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} + \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t\rangle\right|\right) \tag{A.15}$$

$$= \mathbb{E}\left(\sum_{t=\tau}^{T}\sum_{h=1}^{H}\left|\langle \tilde{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} + \hat{\boldsymbol{\theta}}_{S_{h,1}^t}^{t-1} - \boldsymbol{\theta}_{S_{h,1}^t}, \boldsymbol{x}_t\rangle\right|\right)$$

$$\leq 2\mathbb{E}\left(\sum_{t=\tau}^{T}\sum_{l\in\mathcal{H}} n_t^l \sqrt{\gamma_l^{t-1}}\|\boldsymbol{x}_t\|_{(\mathcal{V}_l^{t-1})^{-1}}\right) \tag{A.16}$$

$$\leq 2H\sum_{t=\tau}^{T}\sum_{l\in\mathcal{H}} \sqrt{\gamma_l^{t-1}}\|\boldsymbol{x}_t\|_{(\mathcal{V}_l^{t-1})^{-1}} \tag{A.17}$$

$$\leq 2H\sum_{l\in\mathcal{H}} \sqrt{\sum_{t=\tau}^{T}\gamma_l^{t-1}}\sqrt{\sum_{t=\tau}^{T}\|\boldsymbol{x}_t\|^2_{(\mathcal{V}_l^{t-1})^{-1}}} \tag{A.18}$$

$$\leq \sqrt{d}H^2\mathcal{O}\left(\sqrt{T\log\left(\frac{4TH}{\delta}\right)}\log\left(1+\frac{T}{2d}\right)\right) \tag{A.19}$$

Let us denote $\sqrt{\gamma_{l'}^s} := \big\|\hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s - \boldsymbol{\theta}_{S_{h,2}^s}\big\|_{\mathcal{V}_{l'}^s}$. We obtain Eqn. (A.16) by rearranging terms in the expression involving $\boldsymbol{\theta}_l$ with same l. We use the fact that $n_t^l \leq H$ to get Eqn. (A.17). Next, we apply the Cauchy–Schwarz inequality to derive Eqn. (A.18). By Lemma 2.3.1, we know that, with probability at least $1-\delta$, $\gamma_{l'}^s \leq \gamma$ for all $t \geq 1$, with γ defined in Lemma 2.3.1. In addition, we use Lemma 11 of Abbasi-Yadkori et al. [2011] to bound $\sum_{t=1}^{T}\big\|\boldsymbol{x}_t\big\|^2_{(\mathcal{V}_l^t)^{-1}}$, yielding

$$\sum_{t=1}^{T}\|\boldsymbol{x}_t\|^2_{(\mathcal{V}_l^t)^{-1}} \;\leq\; 4\,\log\!\Big(\tfrac{\det(\mathcal{V}_{T+1})}{\det(\mathcal{V}_1)}\Big) \;\leq\; 4\,d\,\log\!\Big(1+\tfrac{T}{2\,d}\Big).$$

Combining these results and applying the union bound gives us Eqn. (A.19). Similarly, by an

analogous argument, we can show that, with probability at least $1-\delta$, Term 2 in Eqn. (A.14) satisfies a similar bound, as shown below.

$$\sum_{t=\tau}^{T}\mathbb{E}\left(\sum_{h=1}^{H}\left|\langle\tilde{\boldsymbol{\theta}}^{t-1}_{S^t_{h,2}}-\hat{\boldsymbol{\theta}}^{t-1}_{S^t_{h,2}}+\hat{\boldsymbol{\theta}}^{t-1}_{S^t_{h,2}}-\boldsymbol{\theta}_{S^t_{h,2}},\boldsymbol{x}_t\rangle\right|\right)\leq\sqrt{d}H^2\mathcal{O}\left(\sqrt{T\log\left(\frac{4TH}{\delta}\right)}\log\left(1+\frac{T}{2d}\right)\right).$$

The principal difficulty in designing Algorithm 3 lies in estimating $\hat{d}^t_l$, which captures the long-term impact of advertising on product conversion. Since d_l always appears alongside $\boldsymbol{\theta}_{l'}$ in the model $\boldsymbol{y}^t_h\sim\text{Poi}(d_l\,\boldsymbol{\theta}_{l'}^\top\mathbf{x}_t)$, the main technical challenge is establishing a high-probability bound on $|\hat{d}^t_l-d_l|$ that leverages the estimation error of $\hat{\boldsymbol{\theta}}^t_{l'}$. Theorem 2.4.1 summarizes our results in this direction. By Theorem 2.4.1, we know that given $t\geq H\underline{n_l}$, and $l\in\mathcal{H}_1$, with probability $1-\delta,\delta\geq\frac{1}{T^4H}$, $d_l\in\mathcal{D}^t_l$, where

$$\mathcal{D}^t_l=\left\{d\in[0,\boldsymbol{B}_d]\mid|\hat{d}^t_l-d|\leq\frac{4H\,\boldsymbol{B}_d\sqrt{d\,\log\!\left(1+\frac{T}{2d}\right)\gamma}+\sqrt{2\,e\,\boldsymbol{B}_d\,\boldsymbol{B}_x\,\boldsymbol{B}_\theta\,\log\!\left(\frac{2}{\delta}\right)}}{\boldsymbol{b}}\frac{1}{\sqrt{N_{t,l}}}\right\}.$$

Therefore, by applying the union bound to make results valid $\forall t\in[T]$ and $t\geq H\underline{n_l}$, $\forall l\in\mathcal{H}_1$, we have that for $\delta\geq\frac{1}{T^3}$, we conclude that with probability at least $1-\delta$, Term (3) in Eqn. (A.14) can be bounded above by

$$\begin{aligned}&\sum_{t=\tau}^{T}\mathbb{E}\left(\sum_{h=1}^{H}\left|\tilde{d}^{t-1}_{S^t_{h,1}}-\hat{d}^{t-1}_{S^t_{h,1}}+\hat{d}^{t-1}_{S^t_{h,1}}-d_{S^t_{h,1}}\right|\right)\\&\leq\frac{8H\,\boldsymbol{B}_d\sqrt{d\,\log\!\left(1+\frac{T}{2d}\right)\gamma}+\sqrt{2\,e\,\boldsymbol{B}_d\,\boldsymbol{B}_x\,\boldsymbol{B}_\theta\,\log\!\left(\frac{2}{\delta}\right)}}{\boldsymbol{b}}\mathbb{E}\left(\sum_{l=1}^{H}\sum_{t=1}^{N_{T,l}}\frac{1}{\sqrt{t}}\right)+\mathcal{O}(\log(HT))\qquad\text{(A.20)}\\&\leq\mathcal{O}\left(H^2\sqrt{dT}\log\left(1+\frac{T}{2d}\right)\sqrt{\log\left(\frac{4HT}{\delta}\right)}\log\left(\frac{2H}{\delta}\right)\right)\qquad\text{(A.21)}\end{aligned}$$

Equation (A.20) follows by gathering all terms involving the same l of d_l. Next, we obtain Equation (A.21) by noting that $\sum_{l=1}^{H} N_{T,l} = T\,H$. Combining the above results, we conclude that, with probability at least $1-2\delta$ (where $\delta \geq \frac{1}{T^3}$), Term (iii) in Eqn. (2.9) is bounded by

$$\mathcal{O}\Big(H^2\sqrt{d\,T\,\log\big(\tfrac{4\,T\,H}{\delta}\big)}\;\log\big(1+\tfrac{T}{2\,d}\big)\Big).$$

□

A.3.4 Proof of Lemma 2.4.3

Lemma (Lemma 2.4.3 restated)**.** *With probability at least* $1-\delta$ *with* $\delta \geq \frac{2}{T^3}$, $\Theta \in \mathcal{C}_t, \forall t \geq \tau$.

Proof.

$$\begin{aligned}
\mathbb{P}(\Theta \in \mathcal{C}_t, \forall t \geq \tau) &= \mathbb{P}(\{\forall t \geq \tau, \forall l \in \mathcal{H}, \boldsymbol{\theta}_l \in \mathcal{C}_l^t\} \cap \{\forall t \geq \tau, \forall l \in \mathcal{H}_1, d_l \in \mathcal{D}_l^t\}) \\
&= 1 - \mathbb{P}\left\{\Big(\{\forall t \geq \tau, \forall l \in \mathcal{H}, \boldsymbol{\theta}_l \in \mathcal{C}_l^t\} \cap \{\forall t \geq \tau, \forall l \in \mathcal{H}_1, d_l \in \mathcal{D}_l^t\}\Big)^c\right\} \\
&= 1 - \mathbb{P}\left\{\Big(\{\forall t \geq \tau, \forall l \in \mathcal{H}, \boldsymbol{\theta}_l \in \mathcal{C}_l^t\}\Big)^c \cup \Big(\{\forall t \geq \tau, \forall l \in \mathcal{H}_1, d_l \in \mathcal{D}_l^t\}\Big)^c\right\} \\
&\geq 1 - \mathbb{P}\left(\Big(\{\forall t \geq \tau, \forall l \in \mathcal{H}, \boldsymbol{\theta}_l \in \mathcal{C}_l^t\}\Big)^c\right) - \mathbb{P}\left(\Big(\{\forall t \geq \tau, \forall l \in \mathcal{H}_1, d_l \in \mathcal{D}_l^t\}\Big)^c\right)
\end{aligned}$$

By Lemma 2.3.1, we have

$$\mathbb{P}\left(\Big(\{\forall t \geq \tau, \forall l \in \mathcal{H}, \boldsymbol{\theta}_l \in \mathcal{C}_l^t\}\Big)^c\right) \leq \sum_{l \in \mathcal{H}} \mathbb{P}(\exists t \geq \tau \text{ s.t. } \boldsymbol{\theta}_l \notin \mathcal{C}_l^t) \leq \delta.$$

Moreover, by Theorem 2.4.1, with $\delta \geq \frac{1}{HT^4}$ we have

$$\mathbb{P}\left(\Big(\{\forall t \geq \tau, \forall l \in \mathcal{H}, d_l \in \mathcal{D}_l^t\}\Big)^c\right) \leq \sum_{l \in \mathcal{H}_1} \sum_{t \in [T]} \mathbb{P}(d_l^t \notin \mathbf{D}_l^t) \leq HT\delta.$$

Combining these results together, we have with probability at least $1-\delta$ with $\delta \geq \frac{2}{T^3}$, $\Theta \in \mathcal{C}_t, \forall t$. □

A.4 Technical Lemmas

Lemma A.4.1 (Sub-Exponentiality of Poisson Random Variable)**.** *If $X \sim \text{Poi}(\mu)$, then the centered random variable $X - \mu$ is* $\text{SubE}(e\mu, 1)$. *Equivalently, there exist constants $\nu^2 = e\mu$ and $\alpha = 1$ such that, for all $|t| < 1/\alpha = 1$, $\mathbb{E}\big[e^{\,t\,(X-\mu)}\big] \;\le\; \exp\Big(\nu^2\,t^2\Big) \;=\; \exp\big(e\,\mu\,t^2\big)$.*

Proof. Below is the proof showing that a Poisson(μ) random variable is sub-exponential with parameters $\big(e\mu,\, 1\big)$. The proof hinges on bounding the moment-generating function (MGF) of the centered random variable $X - \mu$. Recall that if $X \sim \text{Poi}(\mu)$, its moment-generating function (MGF) is

$$\mathbb{E}\big[e^{tX}\big] \;=\; \exp\big(\mu(e^t - 1)\big).$$

For the centered random variable $X - \mu$,

$$\mathbb{E}\big[e^{t(X-\mu)}\big] \;=\; e^{-t\mu}\,\mathbb{E}\big[e^{tX}\big] \;=\; \exp\big(\mu\big(e^t - 1 - t\big)\big).$$

We claim that, for all $|t| \le 1$,

$$e^t - 1 - t \;\le\; e\,t^2.$$

Indeed, expanding e^t in its Taylor series about $t = 0$,

$$e^t \;=\; 1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \ldots \quad\Longrightarrow\quad e^t - 1 - t \;=\; \frac{t^2}{2} + \frac{t^3}{6} + \ldots$$

For $|t| \le 1$, this sum of higher-order terms is bounded above by a constant times t^2. A convenient choice is e, giving

$$e^t - 1 - t \;\le\; e\,t^2.$$

Combining the two steps, for $|t| \le 1$,

$$\mathbb{E}\big[e^{t(X-\mu)}\big] \;=\; \exp\big(\mu\big(e^t - 1 - t\big)\big) \;\le\; \exp\big(\mu \cdot e\,t^2\big).$$

Therefore, $X - \mu$ satisfies

$$\mathbb{E}\big[e^{t(X-\mu)}\big] \;\le\; \exp\big(e\,\mu\,t^2\big) \quad \text{for all } |t| < 1.$$

By definition, this means $X - \mu$ is $\mathrm{SubE}\big(e\mu, 1\big)$. □

Lemma A.4.2 (Theorem 2 in Abbasi-Yadkori et al. [2011]). *Assume the same in Theorem A.4.3, let $\Sigma_0 = \lambda \mathbb{I}_d, \lambda > 0$. Define $\boldsymbol{y}_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \eta_t$, with η_t defined in Lemma A.4.3, and assume that $\|\boldsymbol{\beta}\|_2 \le \boldsymbol{B}_\beta$, $\|\boldsymbol{x}_t\|_2 \le \boldsymbol{B}_x$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t > 0$, β lies in the set*

$$\mathcal{C}_t = \left\{ \boldsymbol{\beta} \in \mathbb{R}^d : \|\hat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}\|_{\Sigma_t} \le \sigma^2 \sqrt{d \log\left(\frac{1 + t\boldsymbol{B}_x^2/\lambda}{\delta}\right)} + \sqrt{\lambda}\boldsymbol{B}_\beta \right\},$$

where $\hat{\boldsymbol{\beta}}^t := \left(\sum_{s=1}^t \boldsymbol{x}_s \boldsymbol{x}_s^\top + \lambda \mathbb{I}_d\right)^{-1} \left(\sum_{s=1}^t \boldsymbol{x}_s \boldsymbol{y}_s\right)$.

Lemma A.4.3 (Theorem 1 in Abbasi-Yadkori et al. [2011]). *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally σ^2-sub-Gaussian for some $\sigma^2 \ge 0$. Let $\{\boldsymbol{x}_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{x}_t$ is $\mathcal{F}_{t-1}$ measurable. Assume that Σ_0 is a $d \times d$ positive definite matrix. For any $t \ge 0$, define $\Sigma_t = \Sigma_{t-1} + \boldsymbol{x}_t \boldsymbol{x}_s^\top$ and $S_t = \sum_{s=1}^t \eta_s \boldsymbol{x}_s$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \ge 0$, we have $\|S_t\|^2_{\Sigma_t^{-1}} \le 2\sigma^4 \log\left(\sqrt{\frac{\det \Sigma_t}{\det \Sigma_0}}/\delta\right)$.*

Lemma A.4.4 (Sub-exponential tail bound). *Suppose X is sub-exponential with parameters (ν^2, α). Then*

$$\mathbb{P}(|X - \mu| > t) \le \begin{cases} 2\exp(-\frac{t^2}{2\nu^2}), & \text{if } 0 \le t \le \frac{\nu^2}{\alpha} \\ 2\exp(-\frac{t}{2\alpha}). & \text{for } t > \frac{\nu^2}{\alpha} \end{cases}$$

Lemma A.4.5 (Lemma 3 in Hao et al. [2021]). *Consider a random variable $X_i \sim SE(\nu^2, \alpha)$ and β is a non-zero scalar, then $\beta X_i \sim SE(\beta^2\nu^2, |\beta|\alpha)$*

Lemma A.4.6 (Lemma 4 in Hao et al. [2021]). *Consider independent random variables* $X_i \sim SE(\nu_i^2, \alpha_i)$ *for* $i = 1, \ldots, n$, *then* $X = \sum_{i=1}^n X_i$ *follows* $SE\left(\sum_{i=1}^n \nu_i^2, \max_i \alpha_i\right)$.

A.4.1 Proof of Fact 2.2.1

Proof. In this section, we prove that the tuple $(\mathcal{X}, \mathcal{S}, \mathcal{A}, \mathcal{P}^{\boldsymbol{x}_t}, \{R_h^t(S_h^t, \mathbf{a}_h^t, \boldsymbol{x}_t)\}_{h=1}^H, s_1, H)$ constitutes a Contextual Markov decision process (CMDP) by demonstrating that it satisfies the Markov property.

If we lose the bid, then we have

$$\begin{aligned}
\mathbb{P}\Big(R_{h+1}^t &= d_{S_{h+1,1}^t} \langle \boldsymbol{\theta}_{S_{h+1,2}^t}, \boldsymbol{x}_t \rangle (1 - \mathcal{F}_{h+1}(\mathbf{a}_{h+1}^t, \boldsymbol{x}_t)) \\
&\quad + \left(\langle \boldsymbol{\theta}_{S_{h+1,1}^t} \boldsymbol{x}_t \rangle - p_{h+1}(\mathbf{a}_{h+1}^t, \boldsymbol{x}_t) \right) \mathcal{F}_{h+1}(\mathbf{a}_{h+1}^t, \boldsymbol{x}_t), \\
S_{h+1}^t &= (S_{h,1}^t + 1, S_{h,2}^t) | S_0^t, \mathbf{a}_0^t, R_1^t, \ldots, S_h^t, \mathbf{a}_h^t \Big) = 1 - \mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t).
\end{aligned}$$

If we win the bid, then we have

$$\begin{aligned}
\mathbb{P}\Big(R_{h+1}^t &= d_{S_{h+1,1}^t} \langle \boldsymbol{\theta}_{S_{h+1,2}^t}, \boldsymbol{x}_t \rangle (1 - \mathcal{F}_{h+1}(\mathbf{a}_{h+1}^t, \boldsymbol{x}_t)) \\
&\quad + \left(\langle \boldsymbol{\theta}_{S_{h+1,1}^t} \boldsymbol{x}_t \rangle - p_{h+1}(\mathbf{a}_{h+1}^t, \boldsymbol{x}_t) \right) \mathcal{F}_{h+1}(\mathbf{a}_{h+1}^t, \boldsymbol{x}_t), \\
S_{h+1}^t &= (1, S_{h,1}^t) | S_0^t, \mathbf{a}_0^t, R_1^t, \ldots, S_h^t, \mathbf{a}_h^t \Big) = \mathcal{F}_h(\mathbf{a}_h^t, \boldsymbol{x}_t).
\end{aligned}$$

From the above equation, we notice that

$$\mathbb{P}\left(R_{h+1}^t, S_{h+1}^t | S_0^t, \mathbf{a}_0^t, \boldsymbol{r}_1^t, \ldots, S_h^t, \mathbf{a}_h^t\right) = \mathbb{P}\left(R_{h+1}^t, S_{h+1}^t | S_h^t, \mathbf{a}_h^t\right).$$

Therefore, the Markov property holds and the model is a CMDP. □

APPENDIX B

APPENDIX TO LEARNING FROM IMPERFECT HUMAN FEEDBACK

B.1 Challenges

In the following, we discuss challenges in developing robust algorithms for continuous dueling bandits with generally concave utility under corruption induced by strong adversary. There is a line of work that studies bandits under corruption by a weak adversary [Lykouris et al., 2018, Gupta et al., 2019, Agarwal et al., 2021, Zimmert and Seldin, 2021, Saha and Gaillard, 2022, Wei et al., 2022], where the corruption is introduced before observing the agent's actions. This research develops efficient algorithms that achieve sublinear regret without requiring prior knowledge of the total corruption. Specifically, for corruption induced by a weak adversary, Wei et al. [2022] adopts the idea of model selection and introduces a robust framework called Corruption-Robustness through Balancing and Elimination (COBE). This framework allows an algorithm designed for a known corruption level to be adapted for scenarios with an unknown corruption level. Notably, in the case of weak adversaries, Zimmert and Seldin [2021], Saha and Gaillard [2022] develop "best-of-both-world' algorithms, which achieve optimal performance under both begin and corrupted settings.

However, for corruption induced by a strong adversary—where the corruption occurs after observing the agent's actions (the scenario considered in our setting)—it may not be possible to develop a best-of-both-world algorithm. This limitation arises because, for any algorithm that does not know the total corruption, there exists a problem instance where the algorithm suffers linear regret [Bogunovic et al., 2020, Kang et al., 2024]. To achieve sublinear regret in such settings, it is crucial for the algorithm to have precise knowledge of the total corruption or at least an accurate estimation of its upper bound.

For cases with known corruption, nearly optimal algorithms are proposed for various

settings, including stochastic linear bandits in discrete action spaces [Bogunovic et al., 2020], Gaussian process bandits [Bogunovic et al., 2022], linear contextual bandits [He et al., 2022], Lipschitz bandits [Kang et al., 2024], and linear contextual dueling bandits [Di et al., 2024]. These approaches typically rely on phase elimination techniques to eliminate suboptimal arms or Upper Confidence Bound (UCB)-based algorithms to pull optimal arms with high probability. The success of these methods heavily depends on constructing precise utility estimates to guide decision-making.

However, for learning from nonlinear utility, constructing utility estimates under dueling feedback is a very challenging problem, as only preferential (directional) information is available. This limited information motivates the exploration of a new approach: robust gradient-based algorithms. Zimmert and Seldin [2021] and Saha and Gaillard [2022] utilize robust stochastic mirror descent algorithms to update the probability distribution for sampling K arms. However, it is difficult to efficiently generalize this design to continuous action spaces, as it requires transforming the probability vector into a probability density function and efficiently managing its updates and approximations. While Yue and Joachims [2009b] and Kumagai [2017] exploits dueling bandit gradient descent and stochastic mirror descent and achieves sublinear regret for continuous dueling bandits, their performance under corruption remains unexplored. It is a nontrivial question and still remains an interesting open problem of developing robust and computationally efficient algorithms for continuous dueling bandits with generally concave utility functions under corruption induced by strong adversaries.

B.2 Discussion on the Generality of Modeling Assumptions

The assumptions of a continuous action space are standard in the online convex optimization literature [Flaxman et al., 2004, Yue and Joachims, 2009b, Shamir, 2013, Kumagai, 2017]. Moreover, they are particularly natural for today's preference learning for recommendation

systems since languages, texts, videos are mostly processed as embedding vectors which are continuous by nature. We assumed strong concave utility functions. This is standard due to at least two reasons. First, there is already a rich literature studying online optimization of strongly convex or concave functions, and our work subscribes to that rich literature [Shamir, 2013, Kumagai, 2017, Wan et al., 2022, Saha et al., 2022]. Second, strongly concave utilities functions capture essential properties like diminishing marginal returns and risk-averse behavior. These characteristics align closely with real-world scenarios in economics, behavioral studies, and decision-making processes. Strongly concave functions are frequently utilized in economic literature for modeling utility functions such as the "indirect utility function" [Montrucchio, 1987, Sorger, 1995, Venditti, 1997a], which represents the maximum utility a consumer can achieve based on a specific income level and set of prices. Our assumption on the link function aligns with those in Kumagai [2017] but is slightly stronger than those in Yue and Joachims [2009b]. Nevertheless, it encompasses many natural link functions, including the sigmoid function, the cumulative Gaussian distribution function, and the linear function, more general than merely sigmoid link functions [Maystre and Grossglauser, 2017, Saha, 2021b, Xu et al., 2024] or linear link function [Ailon et al., 2014, Chen and Frazier, 2017, Zimmert and Seldin, 2018, Lin and Lu, 2018], as assumed in some earlier works.

B.3 Proofs for Theorem 3.3.1

Theorem (Theorem 3.3.1 restated)**.** *There exists a ρ-Imperfect Human Feedback (see Definition 3.2.1), strongly concave utility function μ, and link function σ such that any learner has to suffer $Reg_T \geq \Omega\left(d \max\{\sqrt{T}, T^{\rho}\}\right)$, even* with the knowledge of ρ.

Proof. Before showing the regret lower bound proof for Theorem 3.3.1, let's first establish the following Lemma B.3.1, which builds the connection between regret incurred under bandit reward feedback and dueling feedback. This proof is inspired by Theorem 6 in Yao et al. [2022a]. To prove Lemma B.3.1, in addition to Reg_T, we introduce a new metric to measure

the performance of algorithm under dueling feedback, which we coin functional regret $\mathrm{Reg}_T^{\mathrm{FO}}$, defined as follows.

$$\mathrm{Reg}_T^{\mathrm{FO}} := \mathbb{E}\left\{\sum_{t=1}^{T} \mu(a^*) - \mu(a_t) + \mu(a^*) - \mu(a_t')\right\}. \tag{B.1}$$

Lemma B.3.1. *Given a utility function μ, if the regret lower bound for algorithm with bandit reward feedback is $\overline{Reg}$, then any learner with dueling feedback has to suffer regret $Reg_T^{FO} \geq 2\overline{Reg}$.*

Proof. We prove our claim by contradiction. Let $(a_{0,t}, a_{1,t})$ be the pair of recommendation at round t. Suppose $\mathrm{Reg}_T^{\mathrm{FO}} < 2\overline{\mathrm{Reg}}$. As a result, at least one of the following inequality must hold:

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^{T} \mu(a^*) - \mu(a_{0,t})\right\} &\leq \overline{\mathrm{Reg}}; \\
\mathbb{E}\left\{\sum_{t=1}^{T} \mu(a^*) - \mu(a_{1,t})\right\} &\leq \overline{\mathrm{Reg}}.
\end{aligned} \tag{B.2}$$

Consider a principal who can observe the interaction between a user and the learner $\mathcal{L}$, then we can construct two algorithms $\mathcal{L}_0$ and $\mathcal{L}_1$ as follows.

Algorithm 5: Algorithm $\mathcal{L}_i$

Input: the time horizon T

for $t \leq T$ **do**

- Call the learner $\mathcal{L}$ to generate two candidates $(a_{0,t}, a_{1,t})$.
- Present $(a_{0,t}, a_{1,t})$ to user and received the feedback.
- Return user feedback to the learner $\mathcal{L}$ and update $\mathcal{L}$.

Output: the sequential decision $\{a_{i,t}\}_{t=1}^{T}$

From (B.2), we know that at least one of $\{\mathcal{L}_0, \mathcal{L}_1\}$ achieves an expected regret lower than $\overline{\mathrm{Reg}}$. However, we know all algorithm with bandit feedback with utility function μ

has to occur regret at least $\overline{\text{Reg}}$, contradiction. Therefore, Lemma B.3.1 must hold, which completes the proof. □

Next, we prove Theorem 3.3.1. First, consider the case $\rho \leq 0.5$. Given that the utility function is strongly concave, by applying Lemma B.3.2, we have $\overline{\text{Reg}} = 0.02 \min\{T, d\sqrt{T}\}$. From Lemma B.3.1, we also obtain $\text{Reg}_T^{\text{FO}} \geq 0.04 \min\{T, d\sqrt{T}\}$. Using the linear link function $\sigma(x) = \frac{1}{2} + \frac{1}{2}x$, it follows that $\text{Reg}_T \geq 0.02 \min\{T, d\sqrt{T}\} \geq \Omega(d \max\{\sqrt{T}, T^{\rho}\})$, completing the proof.

Lemma B.3.2 (Theorem 6 in [Shamir, 2013]). *Let the number of rounds T be fixed. Then for any learner, there exists a quadratic function of the form $\mu_\theta(a) := \theta^\top a - \frac{1}{2}\|a\|^2$ which is minimized at θ and $\|\theta\|_2 \leq 0.5$ such that*

$$\mathbb{E}\left(\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\right) \geq 0.02 \min\left\{T, d\sqrt{T}\right\}.$$

Now, let's focus on the scenario when $\rho > 0.5$. In essence, we want to extend Lemma B.3.2 to the scenario in presence of adversarial corruption induced by ρ-Imperfect Human Feedback. If we can show the regret lower bound can be generalized to $\Omega(dT^{\rho})$, then applying the same technique which uses Lemma B.3.1 to connect regret suffered under bandit reward feedback and regret suffered under dueling feedback and choose linear link function to connect Reg_T^{FO} and Reg_T, we will get the desired regret lower bound. The extension of Lemma B.3.2 is proven at section B.3.1.

B.3.1 Proof for Lemma 3.3.1

Lemma (Lemma 3.3.1 restated). *Consider bandit learning with direct reward feedback, where reward $r(a) := \mu(a) + \epsilon$, ϵ follows standard normal distribution. The action space $\mathcal{A}$ is contained in a d-dimensional unit ball, and μ is $\mu_\theta(a) := \theta^\top a - \frac{1}{2}\|a\|_2^2$, with $\theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1$. Under ρ-Imperfect Human Feedback (Definition 3.2.1), for any T, $d \leq \frac{1}{C_\kappa} T^{1-\rho}$, and for*

any learner, there exists a θ such that under direct reward feedback, even with the knowledge of ρ, *has to suffer regret* $Reg_T := \mathbb{E}\{\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\} \geq \frac{d}{4} C_\kappa T^\rho$.

Proof. To prove Lemma 3.3.1, we aim to prove that there exists a corruption strategy such that the regret incurred by any learner with *imperfect dueling feedback* is not less than $\Omega(dT^\rho)$. The formulation of such a corruption strategy hinges on the observation that when the utility function μ_θ adopts the quadratic form parameterized by θ, specifically $\mu_\theta(a) = \theta^\top a - \frac{1}{2}\|a\|_2^2$, which is both smooth and strongly concave, then the regret incurred by the learner is bounded below by the sum of the Kullback-Leibler (KL) divergence between the distribution of the *corrupted reward* feedback obtained at round t, denoted $\hat{v}_t, t \in [T]$, conditioned on different possible values of θ. (see Lemma B.3.4). Assume that θ is uniformly drawn from $\{-\beta, \beta\}^d$, given an action a_t, conditioned on $\theta_i > 0$, the corrupted reward feedback $\hat{v}_t$ is

$$\hat{v}_t = \mu_\theta(a_t) + c_t(a_t|\theta_i > 0) + \xi = \underbrace{\left(-\frac{1}{2}\|a_t\|^2 + \sum_{j \neq i} \theta_j a_{t,j}\right) + \beta a_{t,i} + c_t(a_t|\theta_i > 0)}_{\mu_1} + \xi. \tag{B.3}$$

Conditioned on $\theta_i < 0$, the corrupted reward feedback $\hat{v}_t$ is

$$\hat{v}_t = \mu_\theta(a_t) + c_t(a_t|\theta_i < 0) + \xi = \underbrace{\left(-\frac{1}{2}\|a_t\|^2 + \sum_{j \neq i} \theta_j a_{t,j}\right) - \beta a_{t,i} + c_t(a_t|\theta_i < 0)}_{\mu_2} + \xi. \tag{B.4}$$

We use $c_t(a_t|\theta_i > 0)$ and $c_t(a_t|\theta_i < 0)$ to empathize the fact that the magnitude and sign of corruption can be dependent on θ. ξ follows a standard Gaussian distribution. Therefore, $\hat{v}_t$ in Equation B.3 follows $\mathrm{N}(\mu_1, 1)$. $\hat{v}_t$ in Equation B.4 follows $\mathrm{N}(\mu_2, 1)$. By using Lemma

B.3.3, we have

$$\mathrm{D}_{\mathrm{KL}}(\mathrm{N}(\mu_1, 1)||\mathrm{N}(\mu_2, 1)) = (\mu_1 - \mu_2)^2.$$

Lemma B.3.3 (KL Divergence for Normal Distribution)**.** *Let $N(\mu, \sigma^2)$ represent a Gaussian distribution variable with mean μ and variance σ^2. Then*

$$D_{KL}(N(\mu_1, \sigma^2)||N(\mu_2, \sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

To optimize the lower bound in Lemma B.3.4, the adversary selects $c_t(a_t|\theta_i > 0)$ and $c_t(a_t|\theta_i < 0)$ to minimize the difference between μ_1 and μ_2. Consider the following corruption strategy: when $\theta_i > 0$, set $c_t(a_t|\theta_i > 0) = -\beta a_{t,i}$; when $\theta_i < 0$, set $c_t(a_t|\theta_i < 0) = \beta a_{t,i}$. This strategy ensures that $\mu_1 = \mu_2$.

Next step is to determine the value of β. Notice that the corruption budget is $C_\kappa T^\rho$. To execute the corruption strategy described above, it requires

$$\sum_{t=1}^{T} |c_t(a_t)| \leq \beta \sum_{t=1}^{T} \|a_t\|_\infty \leq C_\kappa T^\rho. \tag{B.5}$$

Therefore, by exploiting the fact that μ_θ is 1-strongly concave, we establish another lower bound for the regret by

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\right\} &\geq \frac{1}{2}\mathbb{E}\left\{\sum_{t=1}^{T} \|a_t - \theta\|_2^2\right\} \\
&\geq \frac{1}{2}\mathbb{E}\left\{\sum_{t=1}^{T} (\|a_t\|_\infty - \beta)^2\right\} \\
&\geq \frac{1}{2}\sum_{t=1}^{T} (C_\kappa T^{\rho-1}/\beta - \beta)^2.
\end{aligned}$$

Since $|a_{t,i} - \theta_i| \geq ||a_{t,i}| - \beta|$ for all i, we obtain the second inequality. By utilizing the corruption constraint (B.5), we derive the last inequality, which is minimized when $\|a_t\|_\infty = \frac{C_\kappa T^{\rho-1}}{\beta}$ for all t. Together with Lemma B.3.4, we have:

$$\mathbb{E}\left\{\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\right\} \geq \frac{1}{2}\max\left\{\sum_{t=1}^{T}(C_\kappa T^{\rho-1}/\beta - \beta)^2, \frac{dT\beta^2}{2}\right\}. \tag{B.6}$$

We select β to optimize the lower bound in Equation (C.53). Since $d \leq \frac{1}{C_\kappa}T^{1-\rho}$, we can set $\beta = \sqrt{C_\kappa}T^{\frac{\rho}{2}-\frac{1}{2}}$ to satisfy $\|\theta\|_2 \leq 1$. Then we obtain $\mathbb{E}\left\{\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\right\} \geq dC_\kappa T^\rho/4$. Notice that the corruption level each round is bounded by $\beta\|a_t\|_\infty$. Consider the scenario when the radius of the *action* space $\mathcal{A}$ is upper bounded by β, then the corruption budget each round is bounded by $C_\kappa t^{-1+\rho}$, which satisfies the definition of ρ-Imperfect Human Feedback, which completes the proof. □

Lemma B.3.4. *Let's consider the utility function $\mu_\theta(a) := \theta^\top a - \frac{1}{2}\|a\|_2^2$. Let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_T$ be a sequence of* corrupted reward feedback *obtained by a learner. Then there exists a $\theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1$, uniformly drawn from $\{-\beta, \beta\}^d$, such that the regret suffered by the learner is*

$$\mathbb{E}\left\{\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\right\} \geq \frac{dT\beta^2}{4}\left(1 - \sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}\mathbb{D}_{t,i}}\right).$$

$\mathbb{D}_{t,i} := \sup_{\{\theta_j\}_{j\neq i}} D_{KL}\left(\mathbb{P}(\hat{v}_t|\theta_i > 0, \{\theta_j\}_{j\neq i}, \{\hat{v}_l\}_{l=1}^{t-l})||\mathbb{P}(\hat{v}_t|\theta_i < 0, \{\theta_j\}_{j\neq i}, \{\hat{v}_l\}_{l=1}^{t-l})\right)$. *$D_{KL}$ is the KL divergence between two distributions.*

Proof. Using the similar argument in Shamir [2013], assume that the learner is deterministic: a_t is a deterministic function of the realized corrupted reward feedback $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_{t-1}$ at $a_1, a_2, \ldots, a_{t-1}$. This assumption is without loss of generality, since any random learners can be seen as a randomization over deterministic learning algorithms. Thus a lower bound which

holds uniformly for any deterministic learner would also hold over a randomization. To lower bound (B.7), we use Lemma B.3.5, which relates this to the question of how informative are the query values (as measured by Kullback-Leibler divergence) for determining the sign of θ's coordinates. Intuitively, the more similar the query values are, the smaller is the KL divergence and the harder it is to distinguish the true sign of θ_i, leading to a larger lower bound. In addition, we are facing a powerful adversary who has the complete knowledge of the problem and is able to add corruption on the query value to make they are even more similar, which resulting a even smaller KL divergence, consequently, an even larger lower bound. Let $\bar{a}_T := \frac{1}{T}\sum_{t=1}^{T} a_t$ represent the average action, we have

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*) - \mu_\theta(a_t)\right\} &= T\mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T}\mu_\theta(a^*) - \mu_\theta(a_t)\right\} \\
&\geq T\mathbb{E}\left(\mu_\theta(a^*) - \mu_\theta(\bar{a}_T)\right) \\
&\geq T\mathbb{E}\left(\frac{1}{2}\|\bar{a}_T - \theta\|^2\right) \\
&= T\mathbb{E}\left(\frac{1}{2}\sum_{i=1}^{d}(\bar{a}_i - \theta_i)^2\right) \\
&\geq \mathbb{E}\left(\frac{\beta^2 T}{2}\sum_{i=1}^{d}\mathbb{I}_{\bar{a}_i\theta_i<0}\right) \\
&\geq \frac{dT\beta^2}{4}\left(1 - \sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}\mathbb{D}_{t,i}}\right).
\end{aligned} \tag{B.7}$$

We get the second inequality by using the fact that μ_θ is 1-strongly concave. We get the last inequality by using Lemma B.3.5, which completes the proof.

Lemma B.3.5 (Lemma 4 in [Shamir, 2013]). *Let θ be a random vector, none of those coordinates is supported on 0. Let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_T$ be a sequence of values obtained by a deterministic learner returning a point $\bar{a}_T$ (so that the action a_t is a deterministic function of $\hat{v}_1, \ldots, \hat{v}_{t-1}$*

and $\bar{a}_T$ is a deterministic function of $\hat{v}_1, \ldots, \hat{v}_T$). Then we have

$$\mathbb{E}\left(\sum_{i=1}^{d} \mathbb{I}_{\bar{a}_i \theta_i}\right) \geq \frac{d}{2}\left(1 - \sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T} \mathbb{D}_{t,i}}\right),$$

where $\mathbb{D}_{t,i} = \sup_{\{\theta_j\}_{j\neq i}} D_{KL}\left(\mathbb{P}(\hat{v}_t|\theta_i > 0, \{\theta_j\}_{j\neq i}, \{\hat{v}_l\}_{l=1}^{t-l}) || \mathbb{P}(\hat{v}_t|\theta_i < 0, \{\theta_j\}_{j\neq i}, \{\hat{v}_l\}_{l=1}^{t-l})\right)$, D_{KL} represents the KL divergence between two distributions.

□

□

B.3.2 Proof for Proposition 1:

Proposition (Proposition 1 restated)**.** *There exists an LIHF instance with ρ-Imperfect Human Feedback (Definition 3.2.1), linear user utility μ, and link function σ, such that any learner has to suffer $Reg_T \geq \Omega(d\max\{\sqrt{T}, T^{\rho}\})$,* even with the knowledge of ρ.

Proof. The proof structure is very similar to the lower bound proof in Theorem 3.3.1. We start by discussing the value of ρ. Consider the scenario when $\rho \leq 0.5$. Applying Lemma B.3.6 and B.3.1 together with linear link function, we have $\mathrm{Reg}_T \geq \Omega(d\max\{\sqrt{T}, T^{\rho}\})$, which completes the proof.

Lemma B.3.6 ([Dani et al., 2008])**.** *Let $\mathcal{A} = [-1, 1]^d$ and $\Theta = [-T^{-\frac{1}{2}}, T^{-\frac{1}{2}}]^d$. Consider the linear reward function $r_t = \theta^{\top} A_t + \epsilon_t$, ϵ_t is independent standard Gaussian noise. Then for any learner, there exists a vector $\theta \in \Theta$ such that*

$$Reg_T(\mathcal{A}, \theta) \geq \frac{\exp(-2)}{8} d\sqrt{T}.$$

Now, let's focus on the scenario when $\rho > 0.5$ and the essence is to extend Lemma B.3.6 to the scenario in presence of ρ-Imperfect Human Feedback, which is shown in Lemma B.3.7

Lemma B.3.7. *Assume that the* action *space* $\mathcal{A} \subset [-1,1]^d$. *Given* ρ*-Imperfect Human Feedback (Definition 3.2.1), for stochastic linear bandit, there exists a* $\theta \in [-C_\kappa T^{\rho-1}, C_\kappa T^{\rho-1}]^d$ *such that any learner even with the knowledge of* ρ *has to suffer regret no less than* $\frac{d}{8} C_\kappa T^\rho$.

Proof. The proof extends Lemma B.3.6 to a scenario in presence of corruption. We want to construct a parameter family and a corruption strategy such that for all algorithm, it will occur at least $\Omega(dT^\rho)$ regret. Consider the action set $\mathcal{A} \in [-1,1]^d$ and $\Theta = \{-\beta, \beta\}^d$. For any learner, we can lower bound its regret by

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{A}, \theta) &= \mathbb{E}_\theta \left[\sum_{t=1}^{T} \sum_{i=1}^{d} (\mathrm{sign}(\theta_i) - a_{ti}) \theta_i \right] \\
&\geq \beta \sum_{i=1}^{d} \mathbb{E}_\theta \left[\sum_{t=1}^{T} \mathbb{I}\{\mathrm{sign}(a_{ti}) \neq \mathrm{sign}(\theta_i)\} \right] \\
&\geq \frac{T\beta}{2} \sum_{i=1}^{d} \mathbb{P}_\theta \left(\sum_{t=1}^{T} \mathbb{I}\{\mathrm{sign}(a_{ti}) \neq \mathrm{sign}(\theta_i)\} \geq \frac{T}{2} \right)
\end{aligned}$$

Let's denote

$$p_{\theta_i} = \mathbb{P}_\theta \left(\sum_{t=1}^{T} \mathbb{I}\{\mathrm{sign}(a_{ti}) \neq \mathrm{sign}(\theta_i)\} \geq \frac{T}{2} \right)$$

Let $i \in [d]$ and $\theta \in \Theta$ be fixed, and let $\theta'_j = \theta_j$, for $j \neq i$ and $\theta'_i = -\theta_i$. Then using Lemma B.3.8, we have

$$p_{\theta_i} + p_{\theta'_i} \geq \frac{1}{2} \exp\left(-\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{D}_{\mathrm{KL}}(P_{a_t}, P'_{a_t}) \right] \right).$$

P_{a_t} is the distribution of corrupted reward observed by the learner after playing action a_t when reward parameter is θ. Similarly, P'_{a_t} is the distribution of corrupted observed by the learner after playing action a_t when the reward parameter is θ'.

Lemma B.3.8 (Bretagnolle-Huber inequality)**.** *Let* $\mathbb{P}$ *and* $\mathbb{Q}$ *be probability measures on the same measurable space* $(\Omega, \mathcal{F})$. *Let* $A \in \mathcal{F}$ *be any arbitrary event and* A^c *is the complement*

of A. Then we have

$$\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2} \exp\left(-D_{KL}(\mathbb{P}, \mathbb{Q})\right). \tag{B.8}$$

In presence of adversarial corruption, for θ, the corrupted reward is

$$\hat{r}_t = \sum_{j \neq i} a_{tj}\theta_j + a_{ti}\theta_i + c_t(a_t|\theta) + \epsilon_t. \tag{B.9}$$

For θ', we have

$$\hat{r}_t = \sum_{j \neq i} a_{tj}\theta_j - a_{ti}\theta_i + c_t(a_t|\theta') + \epsilon_t. \tag{B.10}$$

Consider the corruption strategy such that $c_t(a_t|\theta) = -a_{t,i}\theta_i$ in (B.9) and $c_t(a_t|\theta') = a_{t,i}\theta_i$ in (B.10). This is achievable since we assume the adversary has complete knowledge of the problem instance. By doing so, the corrupted reward $\hat{r}_t$ are from the same distribution regardless whether the preference parameter is θ or θ'. This is because θ and θ' only differs in the i-th coordinate with the magnitude β, and this difference could be masked by the corruption, resulting in $\mathbb{E}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(P_{a_t}, P'_{a_t})\right] = 0$, which makes it indistinguishable by the learner. Notice that executing such a corruption strategy requires total corruption budget

$$\sum_{t=1}^{T} |c_t(a_t)| \leq \beta \sum_{t=1}^{T} \|a_t\|_\infty \leq C_\kappa T^\rho. \tag{B.11}$$

Choose $\beta = C_\kappa T^{\rho-1}$, Equation (B.11) holds. Therefore, we have

$$p_{\theta_i} + p_{\theta'_i} \geq \frac{1}{2}.$$

Applying an "averaging hammer" over all $\theta \in \Theta$, which satisfies $|\Theta| = 2^d$, we get

$$\sum_{\theta \in \Theta} \frac{1}{|\Theta|} \sum_{i=1}^{d} p_{\theta_i} = \frac{1}{|\Theta|} \sum_{i=1}^{d} \sum_{\theta \in \Theta} p_{\theta_i} \geq \frac{d}{4}.$$

This implies that there exists a $\theta \in \Theta$ such that $\sum_{i=1}^{d} p_{\theta_i} \geq \frac{d}{4}$. Therefore we have

$$\mathrm{Reg}_T(\mathcal{A}, \theta) \geq \frac{dC_\kappa}{8} T^\rho,$$

which completes the proof. We want to highlight that the corruption level each round $|c_t(a_t)|$ is bounded by $\beta \|a_t\|_\infty \leq C_\kappa T^{\rho-1} \leq C_\kappa t^{\rho-1}, \forall t$, satisfying definition of ρ-Imperfect Human Feedback. □

□

B.4 Proof for Theorem 3.4.1

Theorem (Theorem 3.4.1 restated)**.** *RoSMID satisfies* $Reg_T \leq \tilde{\mathcal{O}}(d \max\{\sqrt{T}, T^\rho\})$ *for any* ρ*-LIHF problem.*

Proof. At the beginning of the proof, we formally define self-concordance (see Definition B.4.1). The input, ν-self-concordant function $\mathcal{R}$, serves as a regularizor in RoSMID (see Algorithm 3).

Definition B.4.1. ***(Self-concordance.)*** *A function* $\mathcal{R} : int(\mathcal{A}) \to \mathbb{R}$ *is self-concordant if it satisfies*

1. $\mathcal{R}$ *is three times continuously differentiable, convex, and approaches infinity along any sequence of points approaching the boundary of* $int(\mathcal{A})$.
2. *For every* $h \in \mathbb{R}^d$ *and* $x \in int(\mathcal{A})$, $|\nabla^3 \mathcal{R}(x)[h,h,h]| \leq 2\left(h^\top \nabla^2 \mathcal{R}(x) h\right)^{\frac{3}{2}}$ *holds, where* $|\nabla^3 \mathcal{R}(x)[h,h,h] := \frac{\partial^3 \mathcal{R}}{\partial t_1 \partial t_2 \partial t_3}(x + t_1 h + t_2 h + t_3 h)|_{t_1=t_2=t_3=0}$.

In addition to these two conditions, if for every $h \in \mathbb{R}^d$ *and* $x \in int(\mathcal{A})$, $|\nabla \mathcal{R}(x)^\top h \leq \nu^{\frac{1}{2}} (h^\top \nabla^2 \mathcal{R}(x) h)^{\frac{1}{2}}|$ *for a positive real number* ν, $\mathcal{R}$ *is* ν*-self-concordant.*

Additionally, we remind the reader of the standard assumptions [Yue and Joachims, 2009b, Kumagai, 2017] used in our analysis.

1. the action set $\mathcal{A}$ is a convex compact set that contains the origin, has non-empty interior, and is contained in a d-dimensional ball of radius R;

2. the utility function $\mu : \mathcal{A} \to \mathbb{R}$ is strongly concave and twice-differentiable. The following constants are useful for our algorithm analysis: μ is L^{μ}-Lipschitz, α-strongly concave, κ-smooth, bounded $R^{\mu} := \sup_{a,a'} \mu(a) - \mu(a')$, and $a^* := \arg\max_{a \in \mathcal{A}} \mu(a)$ is the unique optimal within $\mathcal{A}$;

3. the link function $\sigma : \mathbb{R} \to [0, 1]$ is smooth, rotation-symmetric (i.e. $\sigma(x) = 1 - \sigma(-x)$), and concave for any $x \geq 0$. For the ease of analysis: let l_1^{σ} [resp. L_1^{σ}] denote the lower [resp. upper] bound of the first-order derivative of σ. σ is L^{σ}-Lipschitz and its second-order derivative σ'' is L_2^{σ}-Lipschitz and upper bounded by R_2^{σ}.

Under these assumption, Lemma B.4.1 implies Reg_T defined in (3.1) and $\mathrm{Reg}_T^{\mathrm{FO}}$ defined in (B.1) has the same order.

Lemma B.4.1 (Lemma 12 in Kumagai [2017]). *With Reg_T^{FO} defined in (B.1), we have*

$$\frac{Reg_T}{L_1^{\sigma}} \leq Reg_T^{FO} \leq \frac{Reg_T}{l_1^{\sigma}}. \tag{B.12}$$

In the following, we show the proof. The main proof can be divided into four key steps. Step 1 is natural, which quantifies the bias in gradient estimation by RoSMID due imperfect feedback from corrupted utilities. Building on this, Step 2 develops a regret decomposition lemma for dueling bandits under corruption, decomposing the regret into two components: regret from sub-optimal decisions and regret from feedback error. Techniques from previous analyses of continuous dueling bandits handle the sub-optimal decision regret, thus in Step 3, we focus on deriving new techniques to bound the regret from feedback error, a unique

challenge in our setting. We identify a tight upper bound for this error. Finally, Step 4, the only step relying on the decaying corruption level in ρ-imperfect human feedback, uses this assumption, along with the carefully chosen learning rate η_ρ and an iterative regret refinement process, to ensure the tightness of the analysis in Step 3, thus completing the proof.

Step 1: quantifying bias of gradient estimation in ρ-LIHF

Our proof starts from quantifying the bias of the gradient $\hat{g}_t$ caused by ρ-imperfect human feedback, as shown in the following Lemma 3.4.1.

Lemma (Lemma 3.4.1 restated)**.** *The gradient $\hat{g}_t$ in Line 6 of Algorithm 3 satisfies*

$$\mathbb{E}\left(\hat{g}_t | a_t\right) = \nabla|_{a=a_t} \bar{P}_t(a) - \mathbb{E}\left(b_t | a_t\right).$$

where $\bar{P}_t(a) := \mathbb{E}_{x\in\mathbb{B}}\left[\mathbb{P}(a_t \succ a + \nabla^2 R_t(a_t)^{-\frac{1}{2}} x)\right]$ (with $\mathbb{B}$ as the unit ball) is the smoothed probability that a_t is preferred over any action a, and $b_t = d\left(\mathbb{P}(a_t \succ a_t') - \hat{\mathbb{P}}(a_t \succ a_t')\right)\nabla^2 \mathcal{R}_t(a_t)^{\frac{1}{2}} u_t$.

Proof. If we do not have corruption (i.e. the probability for observing noisy comparative feedback $\mathcal{F}(a, a') = 1$ is based on true cost difference $\mu(a) - \mu(a')$), then the uncorrupted gradient $g_t := \mathcal{F}(a_t', a_t) d \nabla^2 \mathcal{R}_t(a_t)^{1/2} u_t$ should be an unbiased estimate of $\nabla \bar{P}_t(a_t)$, i.e.

$$\mathbb{E}(g_t | a_t) = \nabla \bar{P}_t(a_t).$$

We use $\nabla \bar{P}_t(a_t)$ as a shorthand of $\nabla|_{a=a_t} \bar{P}_t(a)$. We use $P_t(a)$ as a shorthand of $P_t(a) := \sigma(\mu(a_t) - \mu(a))$ equivalent to $\mathbb{P}(a_t \succ a)$, and $\hat{P}_t(a) := \sigma(\mu(a_t) - \mu(a) + c_t(a_t, a_t'))$, equivalent to $\hat{\mathbb{P}}(a_t \succ a)$ under our modelling assumption. The proof is similar to Lemma 2.1 in Flaxman

et al. [2005]. Using the Law of total expectation, we have

$$\begin{aligned}
\mathbb{E}(g_t|a_t) &= \mathbb{E}_{u_t}[\mathbb{E}(g_t|a_t, u_t)] \\
&= \mathbb{E}_{u_t}\left(d\mathbb{E}(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t, u_t)\right) \\
&= d\mathbb{E}(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t) \\
&= \nabla\mathbb{E}_{x\in\mathbb{B}}\left(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x|a_t)\right) \\
&= \nabla\bar{P}_t(a_t).
\end{aligned}$$

We get the second inequality by using the definition of $\mathcal{F}(a'_t, a_t)$. We use the Stroke's Theorem to get the second last equality. The gradient which we get to perform gradient descent $\hat{g}_t$ is corrupted. If we let $a'_t = a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t$, we have

$$\begin{aligned}
\mathbb{E}(\hat{g}_t|a_t) &= \mathbb{E}_{u_t}(\mathbb{E}(\hat{g}_t|a_t, u_t)) \\
&= \mathbb{E}_{u_t}\left(d\mathbb{E}(\hat{P}_t(a'_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t, u_t)\right) \\
&= d\mathbb{E}\left(\hat{P}_t(a'_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right) \\
&= d\mathbb{E}\left(\left[P_t(a'_t) + \hat{P}_t(a'_t) - P_t(a'_t)\right]\nabla^2 R_t(a_t)^{\frac{1}{2}}u_t|a_t\right) \\
&= \nabla\mathbb{E}_{x\in\mathbb{B}}\left(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x|a_t)\right) + d\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right] \\
&= \nabla\bar{P}_t(a_t) + d\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right] \\
&= \mathbb{E}(g_t|a_t) + d\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right].
\end{aligned}$$

We get the second last equality by using the fact that g_t is unbiased, which we proved earlier. If we defined b_t as

$$b_t := d\left(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t) - \hat{P}_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t,$$

we get

$$\mathbb{E}(g_t|a_t) = \mathbb{E}(\hat{g}_t|a_t) + \mathbb{E}\left[b_t|a_t\right].$$

□

Step 2: decomposing regret for dueling bandits into regret of decision and feedback error

Core to our proof is a "regret decomposion" lemma below that upper bounds the regret by two parts: regret of sub-optimal decision making under uncertainty, referred as regret of decision (first term in (3.2)), and feedback error (second term in (3.2)).

Lemma (Lemma 3.4.2 restated)**.** *The Reg_T of Algorithm 3 satisfies:*

$$Reg_T \leq \underbrace{\frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^\mu L_1^\sigma R}_{\textit{Regret of Decision}} + \underbrace{2\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top (a_t - a_T^*)\right\}}_{\textit{Feedback Error}},$$

where $C_0 := (2\nu + 4L_1^\sigma\kappa + 4R_2^\sigma(L^\mu)^2 + L^\mu L_1^\sigma\kappa)/\lambda$, *and* $a_T^* := \frac{1}{T}a_1 + (1-\frac{1}{T})a^*$.

Proof. The cornerstone of the analysis below is to separate the impact of b_t from the total regret.

$$\begin{aligned}
\mathrm{Reg}_T &\leq 2\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t) - P_t(a_T^*)\right] + \frac{L^\mu L_1^\sigma\kappa}{\lambda\eta_\rho} + 2L^\mu L_1^\sigma R \\
&\leq 2\left(\mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} + \mathbb{E}\left\{\sum_{t=1}^{T}(P_t(a_t) - \bar{P}_t(a_t))\right\} + \mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_T^*) - P_t(a_T^*))\right\}\right) + \\
&\frac{L^\mu L_1^\sigma\kappa}{\lambda\eta_\rho} + 2L^\mu L_1^\sigma R \\
&\leq 2\mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} + \frac{4L_1^\sigma\kappa + 4R_2^\sigma(L^\mu)^2 + L^\mu L_1^\sigma\kappa}{\lambda\eta_\rho}\log T + 2L^\mu L_1^\sigma R.
\end{aligned}$$

We get the first inequality by using Lemma B.4.2. We get the last inequality by because

$$\bar{P}_t(a) - P_t(a) \leq \frac{L_1^\sigma \kappa + R_2^\sigma (L^\mu)^2}{2} \|\nabla^2 \mathcal{R}_t(a_t)^{-\frac{1}{2}} u_t\|^2 \leq \frac{L_1^\sigma \kappa + R_2^\sigma (L^\mu)^2}{\lambda \eta_\rho t},$$

and

$$\mathbb{E}\left\{\sum_{t=1}^{T}(P_t(a_t) - P_t(a_T^*))\right\} \leq \mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} + 2\frac{L_1^\sigma \kappa + R_2^\sigma (L^\mu)^2}{\lambda \eta_\rho} \log T.$$

Lemma B.4.2 (Lemma 5 in Kumagai [2017])**.**

$$Reg_T \leq 2\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t) - P_t(a_T^*)\right] + \frac{L^\mu L_1^\sigma \kappa}{\lambda \eta_\rho} + 2L^\mu L_1^\sigma R.$$

Then it remains to bound $\mathbb{E}\{\sum_{t=1}^{T} \bar{P}_t(a_t) - \bar{P}_t(a_T^*)\}$. And we have the following.

$$\begin{aligned}
\mathbb{E}\{\bar{P}_t(a_t) - \bar{P}_t(a_T^*)\} &\leq \mathbb{E}\left[\nabla \hat{P}_t^\top(a_t)(a_t - a_T^*) - \frac{L_1^\sigma \alpha}{4}\|a_t - a_T^*\|^2\right] \\
&= \mathbb{E}\left[g_t^\top (a_t - a_T^*) - \frac{L_1^\sigma \alpha}{4}\|a_t - a_T^*\|^2\right] \\
&= \mathbb{E}\left[(\hat{g}_t + b_t)^\top (a_t - a_T^*) - \frac{L_1^\sigma \alpha}{4}\|a_t - a_T^*\|^2\right] \\
&= \mathbb{E}\left[\hat{g}_t^\top (a_t - a_T^*) - \frac{L_1^\sigma \alpha}{4}\|a_t - a_T^*\|^2\right] + \mathbb{E}\{b_t^\top (a_t - a_T^*)\},
\end{aligned}$$

where the first and second equality is resulted from Lemma 3.4.1. Using the definition of a_{t+1} in Algorithm 3, we have

$$\nabla \mathcal{R}_t(a_{t+1}) - \nabla \mathcal{R}_t(a_t) = \eta_\rho \hat{g}_t.$$

Therefore we have

$$\begin{aligned}\mathbb{E}\left[\hat{g}_t^\top(a_t - a_T^*) - \frac{L_1^\sigma \alpha}{4}\|a_t - a_T^*\|^2\right] &= \frac{1}{\eta_\rho}\mathbb{E}\left[(\nabla\mathcal{R}_t(a_{t+1}) - \nabla\mathcal{R}_t(a_t))^\top(a_t - a_T^*) - \frac{L_1^\sigma \alpha\eta_\rho}{4}\|a_t - a_T^*\|^2\right] \\ &= \frac{1}{\eta_\rho}\mathbb{E}\left[D_{\mathcal{R}_t}(a_T^*, a_t) + D_{\mathcal{R}}(a_t, a_{t+1}) - D_{\mathcal{R}_t}(a_T^*, a_{t+1})\right] \\ &\quad - \mathbb{E}\left[\frac{L_1^\sigma \alpha\eta_\rho}{4}\|a_t - a_T^*\|^2\right].\end{aligned}$$

$D_{\mathcal{R}}(a, b)$ is the Bregman divergence associated with $\mathcal{R}$, defined by

$$D_{\mathcal{R}}(a, b) = \mathcal{R}(a) - \mathcal{R}(b) - \nabla\mathcal{R}(b)^\top(a - b).$$

Then we can get (B.13) by using Lemma B.4.3, B.4.4, B.4.5.

$$\mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} \leq \frac{\nu\log(T)}{\eta_\rho} + 4d^2\eta_\rho T + \mathbb{E}\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\}. \tag{B.13}$$

Lemma B.4.3 (Lemma 10 in Kumagai [2017]). *Let $\mathcal{R}_t^*(a) = \sup_{x\in\mathbb{R}^d} x^\top a - \mathcal{R}_t(a)$ denote the Frenchel dual of $\mathcal{R}_t$. Then we have*

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\left[\hat{g}_t^\top(a_t - a_T^*) - \frac{L_1^\sigma \alpha}{4}\|a_t - a_T^*\|^2\right] \leq \frac{1}{\eta_\rho}\left(\mathcal{R}(a_T^*) - \mathcal{R}(a_1)\right) \\ &+ \frac{1}{\eta_\rho}\left(\mathbb{E}\left[\sum_{t=1}^{T} D_{\mathcal{R}_t^*}(\nabla\mathcal{R}_t(a_t) - \eta_\rho\hat{g}_t, \nabla\mathcal{R}_t(a_t))\right]\right).\end{aligned}$$

Lemma B.4.4 (Lemma 11 in Kumagai [2017]). *When $\eta_\rho \leq \frac{1}{2d}$, we have*

$$D_{\mathcal{R}_t^*}(\nabla\mathcal{R}_t(a_t) - \eta_\rho\hat{g}_t, \nabla\mathcal{R}_t(a_t)) \leq 4d^2\eta_\rho^2.$$

Lemma B.4.5 (Lemma 4 in Hazan and Levy [2014]). $\mathcal{R}(a_T^*) - \mathcal{R}(a_1) \leq \nu\log(T)$.

If we let $C_0 := (2\nu + 4L_1^\sigma\kappa + 4R_2^\sigma(L^\mu)^2 + L^\mu L_1^\sigma\kappa)/\lambda$, we have

$$\begin{aligned}
\mathrm{Reg}_T &\le \frac{2\nu\log(T)}{\eta_\rho} + 8d^2\eta_\rho T \\
&\quad + 2\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\right\} + \frac{4L_1^\sigma\kappa + 4R_2^\sigma(L^\mu)^2 + L^\mu L_1^\sigma\kappa}{\lambda\eta_\rho}\log T + 2L^\mu L_1^\sigma R \\
&= \underbrace{\frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^\mu L_1^\sigma R}_{\textit{Regret of Decision}} + \underbrace{2\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\right\}}_{\textit{Feedback Error}},
\end{aligned}$$

which completes the proof. □

Step 3: upper bounding the regret from feedback error

Bounding the regret of decision term in (3.2) is standard. Thus, this step develops a novel upper bound for bounding the regret from feedback error, as shown below.

Lemma (Lemma 3.4.3 restated)**.** *The feedback error term in* (3.2) *can be bounded as follows:*

$$\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\right\} \le C_1 d\sqrt{\eta_\rho}\,\mathbb{E}\left\{\sum_{t=1}^{T}\sqrt{t}|c_t(a_t, a_t')|\sqrt{\mu(a^*) - \mu(a_t)}\right\} + C_2 dT^\rho + 2d\sqrt{\lambda}\sqrt{\eta_\rho T},$$

where $C_1 := \sqrt{\frac{2\lambda(L^\sigma)^2}{\alpha}}$, *and* $C_2 := 2L^\sigma R\sqrt{\sup_{a\in\mathcal{A}}\lambda_{\max}\left(\nabla^2\mathcal{R}(a)\right) + 2\phi}$.

Proof. In the following, we will analyze the expected cumulative sum of the bias b_t on the total regret, which is $\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\right\}$. By Cauchy-Schwartz inequality, $b_t^\top(a_t - a_T^*) \le \|b_t\|_2\|(a_t - a_T^*)\|_2$. We can bound the $\|b_t\|_2$ by the following.

$$\begin{aligned}
b_t^\top b_t &= d^2\left(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t) - \hat{P}_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\right)^2 u_t^\top\nabla^2\mathcal{R}_t(a_t)u_t \\
&\le \left(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t) - \hat{P}_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\right)^2 d^2\lambda_{\max}(\nabla^2\mathcal{R}_t(a_t)).
\end{aligned}$$

We get the first equality by Lemma 3.4.1. We get the inequality by using the fact $\|u_t\|_2 = 1$.

If we let $a'_t = a_t + \nabla^2 \mathcal{R}_t(a_t)^{-\frac{1}{2}} u_t$, then we have

$$\begin{aligned} |P_t(a'_t) - \hat{P}_t(a'_t)| &= |\sigma(f(a_t) - f(a'_t)) - \sigma(f(a_t) - f(a'_t) + c_t(a_t, a'_t))| \\ &\leq \min\left(2, L^\sigma |c_t(a_t, a'_t)|\right). \end{aligned}$$

We get the first equality by using the definition of $P_t(a_t)$ and $\hat{P}_t(a_t)$. We get the first inequality by using the Lipschitz property of σ. Therefore, we have

$$\|b_t\|_2 \leq d\sqrt{\lambda_{\max}(\nabla^2 \mathcal{R}_t(a_t))} \min\left\{2, L^\sigma |c_t(a_t, a'_t)|\right\}.$$

If we use $\lambda_R^* := \sup_{a \in \mathcal{A}} \lambda_{\max}\left(\nabla^2 \mathcal{R}(a)\right)$, then we have

$$\lambda_{\max}(\nabla^2 \mathcal{R}_t(a_t)) = \lambda_{\max}(\nabla^2 \mathcal{R}(a_t)) + \lambda \eta_\rho t + 2\phi \leq \lambda_R^* + 2\phi + \lambda \eta_\rho t.$$

Therefore we have

$$\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top (a_t - a_T^*)\right\} \leq \mathbb{E}\left\{\sum_{t=1}^{T} \|b_t\|_2 \|a_t - a_T^*\|_2\right\}$$

$$\leq \mathbb{E}\left\{\sum_{t=1}^{T} (d\sqrt{\lambda_{\max}(\nabla^2 \mathcal{R}_t(a_t))} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right)) \|a_t - a_T^*\|_2\right\}$$

$$\leq \mathbb{E}\left\{\sum_{t=1}^{T} (d\sqrt{\lambda_R^* + 2\phi + \lambda\eta_\rho t} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right)) \|a_t - a_T^*\|_2\right\}$$

$$\leq \mathbb{E}\left\{\sum_{t=1}^{T} (d\sqrt{\lambda_R^* + 2\phi} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right) \|a_t - a_T^*\|_2\right\} \tag{B.14}$$

$$+ \mathbb{E}\left\{\sum_{t=1}^{T} (d\sqrt{\lambda\eta_\rho t} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right) \|a_t - a_T^*\|_2\right\}$$

$$\leq 2RdL^\sigma \sqrt{\lambda_R^* + 2\phi} T^\rho + d\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^{T} \sqrt{t} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right) \|a_t - a_T^*\|_2\right\} \tag{B.15}$$

$$\leq 2RdL^\sigma \sqrt{\lambda_R^* + 2\phi} T^\rho + 2d\sqrt{\lambda\eta_\rho T} + d\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^{T} \sqrt{t} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right) \|a_t - a^*\|_2\right\} \tag{B.16}$$

$$\leq 2RdL^\sigma \sqrt{\lambda_R^* + 2\phi} T^\rho + 2d\sqrt{\lambda\eta_\rho T} \tag{B.17}$$

$$+ d\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^{T} \sqrt{t} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right) \min\left(2R, \sqrt{\frac{2}{\alpha}\left(\mu(a^*) - \mu(a_t)\right)}\right)\right\}. \tag{B.18}$$

We get (B.15) by using the fact that $\|a_t - a_T^*\|_2 \leq 2R$, $\sum_{t=1}^{T} |c_t(a_t, a_t')| \leq T^\rho$. We get (B.16) the definition of a_T^*, and the fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Using the α-strong convexity of μ, we have $\|a_t - a_*\|_2 \leq \min\left(2R, \sqrt{\frac{2}{\alpha}\left(\mu(a^*) - \mu(a_t)\right)}\right)$ we get the last inequality, which completes the proof. □

Step 4: The *iterative regret refinement* analysis and resultant optimal learning rate

Lemma (Lemma 3.4.4 restated)**.** *If* $A = \mathbb{E}\left(\sum_{t=1}^{T}[\mu(a^*) - \mu(a_t)]\right) \leq \mathcal{O}(T^{\frac{1}{2}+\psi})$ *for some* $\psi \in [0, \frac{1}{2})$*, then we must have* $B = \mathbb{E}\left(\sum_{t=1}^{T} t^{\rho-\frac{1}{2}}\sqrt{\mu(a^*) - \mu(a_t)}\right) \leq \mathcal{O}(T^{-\frac{1}{4}+\frac{\psi}{2}+\rho})$.

Proof. To prove Lemma 3.4.4, it is equivalent to prove the following Lemma B.4.6. If Lemma B.4.6 holds, let $n_t = \sqrt{\mathbb{E}(\mu(a^*) - \mu(a_t))}$, $m_t = \min\left(2, L^{\sigma}|c_t(a_t, a_t')|\right)$, proves Lemma 3.4.4.

Lemma B.4.6. *Consider* $\sum_{t=1}^{T} n_t^2 \leq C' T^{\frac{1}{2}+\alpha}$*, with constants* $\alpha \geq 0$*,* $C' > 0$*. In addition,* n_t *is uniformly upper bounded by a constant* K*. Specifically, we have* $0 \leq n_t \leq K, \forall t$*.* $m_t \leq c_k t^{\rho-1}$ *with constant* $c_k \geq 0$*. Then we have* $\sum_{t=1}^{T}\sqrt{t} m_t n_t \leq 5\sqrt{C'} c_k T^{-\frac{1}{4}+\frac{\alpha}{2}+\rho}$.

Proof. Using Abel's equality (Lemma B.4.7), we have

$$\sum_{t=1}^{T}\sqrt{t} m_t n_t \leq \sqrt{T}\left(\sum_{t=1}^{T} m_t n_t\right)$$

Since we have $\sum_{t=1}^{T} n_t^2 \leq C' T^{\frac{1}{2}+\alpha}$ and $0 \leq n_t \leq K, \forall t$, we have $\sum_{t=1}^{T} n_t \leq \sqrt{C'} T^{\frac{3}{4}+\frac{\alpha}{2}}, \forall t$. Moreover, for $t \geq 1$, $t^{-1+\rho} - (t+1)^{-1+\rho} \leq t^{\rho-2}$ holds for all $t \geq 1$. This is because

$$f(t) = t^{-1+\rho}\left((1+\frac{1}{t})^{-1+\rho} + \frac{1}{t} - 1\right).$$

is decreasing over $t \geq 1$ and $\lim_{t\to\infty} f(t) = 0$. Consequently, applying Abel's Summation

Equation again, we have

$$\begin{aligned}
\sum_{t=1}^{T} m_t n_t &= \left(\sum_{t=1}^{T} n_t\right) m_T + \sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i\right) (m_t - m_{t+1}) \\
&\leq \sqrt{C'} c_k T^{\frac{3}{4}+\frac{\alpha}{2}} T^{-1+\rho} + \sqrt{C'} c_k \sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i\right) \left(t^{-1+\rho} - (t+1)^{-1+\rho}\right) \\
&\leq \sqrt{C'} c_k T^{\frac{3}{4}+\frac{\alpha}{2}} T^{-1+\rho} + \sqrt{C'} c_k \sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i\right) t^{-2+\rho} \\
&\leq \sqrt{C'} c_k T^{\frac{3}{4}+\frac{\alpha}{2}} T^{-1+\rho} + \sqrt{C'} c_k \sum_{t=1}^{T-1} t^{\frac{3}{4}+\frac{\alpha}{2}} t^{-2+\rho} \\
&\leq 5\sqrt{C'} c_k T^{-\frac{1}{4}+\frac{\alpha}{2}+\rho}.
\end{aligned}$$

Since $0 \leq n_t \leq K$, $\sum_{i=1}^{t} n_i$ is increasing and the increasing rate is upper bound by t. Together with the constraint $\sum_{t=1}^{T} n_t \leq \sqrt{C'} T^{\frac{3}{4}+\frac{\alpha}{2}}$, $\sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i\right) t^{-\frac{1}{2}+\rho}$ is optimized when $\sum_{i=1}^{t} n_i$ increasing in the speed of $\sqrt{C'} t^{\frac{3}{4}+\frac{\alpha}{2}}$. Because of this, the second last inequality holds.

Lemma B.4.7. ***(Abel's Summation Equation [Williams, 1963]).*** *For any numbers a_k, b_k, we have*

$$\sum_{k=1}^{n} a_k b_k = \left(\sum_{k=1}^{n} b_k\right) a_n + \sum_{k=1}^{n-1} \left(\sum_{i=1}^{k} b_i\right) (a_k - a_{k+1}).$$

□

□

After establishing Lemma 3.4.4, we can use it to prove the induction claim (Lemma 3.4.5). The challenge in this proof lies in identifying the constant K, which must hold across all iterations of the analysis to ensure the bound remains stable and does not diverge. This requires careful refinement of the analysis and calculations.

Lemma (Lemma 3.4.5 restated). *Consider* $T > \sqrt{2L^{\mu}L_1^{\sigma}R}$ *and* $\eta_\rho = \frac{\sqrt{\log T}}{dT^{\max\{1/2,\rho\}}}$. *If* $Reg_T \leq 144dK\sqrt{\log T}T^{\rho+\frac{\frac{3}{2}-\rho}{2^k}}$ *for some integer* k, *then we must also have* $Reg_T \leq 144dK\sqrt{\log T}T^{\rho+\frac{\frac{3}{2}-\rho}{2^{k+1}}}$. K *is a instance-dependent constant,* *where* $K := \max\left\{C_0, \frac{C_2C_\kappa}{\sqrt{\log T}}, (L^{\sigma}T_\kappa^{\rho})^2, 2\sqrt{\lambda}RC_\kappa L^{\sigma}\right\}$, C_0 *and* C_2 *defined in Lemma 3.4.2 and 3.4.3 respectively,*

Proof. In the following, we denote

$$m_t := \min\left(2, L^{\sigma}|c_t(a_t, a_t')|\right), n_t := \min\left(2R, \sqrt{\frac{2}{\alpha}\left(\mu(a^*) - \mu(a_t)\right)}\right)$$

Continue from (B.18), we have the following.

The Base Case: When $k = 1$, let $C_3 = C_2dT^{\rho} + 2d\sqrt{\lambda\eta_\rho T}$, we have

$$\begin{aligned}
\mathrm{Reg}_T &\leq \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^{\mu}L_1^{\sigma}R + 2C_3 + 2d\sqrt{\lambda\eta_\rho}\mathbb{E}\left(\sum_{t=1}^{T}\sqrt{t}m_t n_t\right). && \text{(B.19)}\\
&\leq \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^{\mu}L_1^{\sigma}R + 2C_3 + 4dR\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^{T}\sqrt{t}m_t\right\} && \text{(B.20)}\\
&\leq \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^{\mu}L_1^{\sigma}R + 2C_3 + 4dR\sqrt{\lambda\eta_\rho T}\left(\sum_{t=1}^{T}m_t\right) && \text{(B.21)}\\
&\leq \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^{\mu}L_1^{\sigma}R + 2C_3 + 4dL^{\sigma}R\sqrt{\lambda\eta_\rho}C_\kappa T^{\rho}. && \text{(B.22)}
\end{aligned}$$

We get (B.19) according to Lemma 3.4.2 and 3.4.3. We get (B.20) because of $\mathbb{E}(n_t) \leq 2R$. We get (B.21) by Abel's Summation Equation (see Lemma B.4.7). We get (B.22) because of corruption budget. Choosing the learning rate η_ρ according to Theorem 3.4.1, we obtain $\mathrm{Reg}_T \leq 144\sqrt{\log T}KdT^{\frac{\rho+1}{2}}$. This confirms the validity of the claim when $k = 1$.

The Induction Argument: Let's assume that the claim holds true for a general step k.

In the following, we will demonstrate that the claim also holds true for step $k+1$. Because of Equation (B.12) and the induction claim, we obtain $\mathrm{Reg}_T^{\mathrm{FO}} \leq \frac{144}{l_1^\sigma} dK\sqrt{\log T}T^{\rho+\frac{3/2-\rho}{2^k}}$. This suggests that

$$\sum_{t=1}^{T} \mathbb{E}\left(\mu(a^*) - \mu(a_t)\right) \leq \frac{144}{l_1^\sigma} dK\sqrt{\log T}T^{\rho+\frac{3/2-\rho}{2^k}}.$$

By Jensen's inequality, we have

$$\mathbb{E}(n_t) = \min\left\{2, \sqrt{\frac{2}{\alpha}}\mathbb{E}\left(\sqrt{\mu(a^*) - \mu(a_t)}\right)\right\} \leq \sqrt{\frac{2}{\alpha}}\sqrt{\mathbb{E}\left(\mu(a^*) - \mu(a_t)\right)}.$$

Therefore we have

$$\mathbb{E}\left(\sum_{t=1}^{T}\sqrt{t}m_t n_t\right) \leq \sqrt{\frac{2}{\alpha}}\sum_{t=1}^{T}\sqrt{t}m_t\sqrt{\mathbb{E}\left(\mu(a^*) - \mu(a_t)\right)}.$$

By the definition of ρ-Imperfect Human Feedback, we have $|c_t(a_t, a_t')|$ is upper bounded by $C_k t^{-1+\rho}$.

Therefore, using Lemma 3.4.4 we can upper bound $\sqrt{\lambda}\sum_{t=1}^{T}\sqrt{t}m_t n_t$ by

$$\sqrt{\lambda}\sum_{t=1}^{T}\sqrt{t}m_t n_t \leq 60L^\sigma\sqrt{\frac{2\lambda}{\alpha l_1^\sigma}}(\log T)^{0.25}C_\kappa\sqrt{dK}T^{\frac{3}{2}\rho+\frac{3/2-\rho}{2^{k+1}}}$$

Since $\lambda \leq \alpha l_1^\sigma/2$, choosing $\eta_\rho = \frac{1}{d}\frac{\sqrt{\log T}}{T^\rho}$, we have

$$\begin{aligned}\mathrm{Reg}_T &\leq 2d\left(C_0 + \frac{C_2}{\sqrt{\log T}} + 60C_\kappa\sqrt{K}T^{\frac{3/2-\rho}{2^{k+1}}}\right)\sqrt{\log T}T^\rho + 12Kd\sqrt{\log T}T^\rho \\ &\leq 144dK\sqrt{\log T}T^{\rho+\frac{3/2-\rho}{2^{k+1}}}.\end{aligned}$$

The claim holds true for step $k+1$. Therefore, it implies that the claim holds true for all k.

Repeating this process infinitely many times, we obtain

$$\mathrm{Reg}_T \leq \mathcal{O}\left(d\sqrt{\log T}T^{\rho}\right),$$

which completes the proof when $\rho \in [0.5, 1]$. □

□

B.5 Proof for Proposition 2

Proposition (Proposition 2 restated)**.** *If the total corruption level satisfies* $\sum_{t=1}^{T} |c_t(a_t, a_t')| \leq \mathcal{O}(T^{\rho})$*, then for any* $\alpha \in [\frac{1}{2}, 1)$*, if the utility function is strongly concave, for sufficiently large round* T*, by choosing learning rate* $\eta_\rho := \frac{\sqrt{\log T}}{2d} T^{-\alpha}$

1. *(Robustness) RoSMID incurs* $\mathrm{Reg}_T \leq \tilde{\mathcal{O}}(dT^{\alpha} + \sqrt{d}T^{\frac{1}{2}(1-\alpha)+\rho} + dT^{\rho})$.
2. *(Efficiency) There exists a strongly concave utility function* μ *and a link function* σ *such that* Reg_T *suffered by RoSMID is at least* $\Omega\left(T^{\alpha}\right)$ *in scenario without corruption.*

B.5.1 Proof for Robustness Statement in Proposition 2

Proof. From Lemma 3.4.3, we know

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^{T} b_t^{\top}(a_t - a_T^*)\right\} &\leq 2RdL^{\sigma}\sqrt{\lambda_R^* + 2\phi}T^{\rho} + d\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^{T}\sqrt{t}\min\left(2, L^{\sigma}|c_t(a_t, a_t')|\right)\|a_t - a_T^*\|_2\right\} \\
&\leq 2RdL^{\sigma}\sqrt{\lambda_R^* + 2\phi}T^{\rho} + 2Rd\sqrt{\lambda\eta_\rho}L^{\sigma}\sum_{t=1}^{T}\sqrt{t}|c_t(a_t, a_t')| \\
&\leq 2RdL^{\sigma}\sqrt{\lambda_R^* + 2\phi}T^{\rho} + 2Rd\sqrt{\lambda\eta_\rho T}L^{\sigma}T^{\rho}.
\end{aligned}$$

We get the second inequality by using the fact that the diameter of the action space of R. We get the last inequality by using Abel Summation Equation (see Lemma B.4.7). Therefore,

by Lemma 3.4.2, we have

$$\mathrm{Reg}_T \leq \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2L^\mu L_1^\sigma R + 4RdL^\sigma\sqrt{\lambda_R^* + 2\phi}T^\rho + 4Rd\sqrt{\lambda\eta_\rho T}L^\sigma T^\rho.$$

Choosing $\eta_\rho = \frac{\sqrt{\log T}}{2d}T^{-\alpha}, \alpha \in [0.5, 1)$, we have

$$\begin{aligned}\mathrm{Reg}_T &\leq 2dC_0\sqrt{\log T}T^\alpha + 4d\sqrt{\log T}T^{1-\alpha} + 4RdL^\sigma\sqrt{\lambda_R^* + 2\phi}T^\rho \\ &\quad + 4R\sqrt{d}L^\sigma(\log T)^{0.25}T^{\frac{1}{2}(1-\alpha)}T^\rho + 2L^\mu L_1^\sigma R \\ &\leq O\left(d\sqrt{\log T}T^\alpha + \sqrt{d}(\log T)^{\frac{1}{4}}T^{\frac{1}{2}(1-\alpha)}T^\rho + dT^\rho\right),\end{aligned}$$

which completes the proof of robustness statement. □

B.5.2 Proof for Efficiency Statement in Proposition 2

Proof. The robustness statement in Proposition 2 implies that RoSMID could afford agnostic corruption level $\mathcal{O}(T^{\frac{1+\alpha}{2}})$. If Lemma B.5.1 holds, this implies that there exists a hard instance which makes RoSMID incur regret in order of $\Omega(T^\alpha)$ in non-corrupt setting. We can prove this by contradiction. Assume the efficiency statement in Proposition 2 is not true, which implies that for all μ, and σ, there exists an $\epsilon > 0$ such that $\mathrm{Reg}_T = c_0 T^{\alpha-\epsilon}$, for some constant $c_0 > 0$, we assume ϵ to be the least possible. In this scenario, when we consider a linear link function $\sigma(x) = \frac{1+x}{2}$, Lemma B.5.1 implies that an agnostic corruption budget $2\sqrt{c_0}T^{\frac{1+\alpha-\epsilon}{2}}$ suffices to induce linear regret, which contradicts the robustness statement in Proposition 2.

Lemma B.5.1. *For any algorithm $\mathcal{G}$ incurs regret $Reg_T \leq c_0 T^\alpha$, for some constant $c_0 > 0$, on learning from bandit with reward feedback for strongly concave utilities μ in non-corrupted setting, then there exists an instance μ such that $\mathcal{G}$ will suffer linear regret under agnostic corruption with budget $\sqrt{c_0}T^{\frac{1+\alpha}{2}}$.*

Proof. Consider the scenario when $d = 2$, the action space $\mathcal{A} := \{(a_1, a_2) : a_1 \geq 0, a_2 \geq 0, a_1 + a_2 \leq 1\}$, the utility function $\mu_\theta(a) = \langle \theta, a \rangle - \frac{1}{2}\|a\|_2^2$ with $\theta := [1, 0]$. In this scenario, the optimal action is $a^* := [1, 0]$. Assume that the adversary want to "fool" the algorithm $\mathcal{G}$ to make it believe the $\theta := [1, 1]$, with optimal arm $a^* = [\frac{1}{2}, \frac{1}{2}]$, each round, the corruption it should introduce is $c_t(a_t) = a_{t2}$, where a_{ti} is the i-th coordinate value of action a_t. To make reward indistinguishable from these two environments, the total corruption budget which the adversary has to pay is $\sum_{t=1}^T |a_{t2}|$. We know that $\mathcal{G}$ incurs regret $\text{Reg}_T := \sum_{t=1}^T \langle \theta, a^* - a_t \rangle - \frac{1}{2}\|a^*\|_2^2 + \frac{1}{2}\|a_t\|_2^2 = \frac{1}{2} - a_{t1} + \frac{1}{2}a_{t1}^2 + \frac{1}{2}a_{t2}^2 \leq c_0 T^\alpha$. Given this constraint, the maximum possible corruption paid by the adversary to make $\theta = [1, 1]$ and $\theta = [1, 0]$ indistinguishable is $\max \sum_{t=1}^T |a_{t2}| = \sqrt{c_0} T^{\frac{\alpha+1}{2}}$. Therefore, for any agnostic corruption no less than $\sqrt{c_0} T^{\frac{\alpha+1}{2}}$, $\mathcal{G}$ can not distinguish $\theta = [1, 0]$ and $\theta = [1, 1]$ and has to incur linear regret in either of these environments. □

□

B.6 Proof for Proposition 3:

In this section, we present the regret upper for dueling bandit gradient descent Yue and Joachims [2009b] in presence of *arbitrary* and *agnostic* corruption. Before presenting the proof, we make several remarks. First, in Yue and Joachims [2009b], the utility function μ is assumed to be strictly concave to ensure the uniqueness of the maximizer. However, in scenarios where the maximizer is unique within the *action* set for a concave utility function μ, the requirement for "strictness" can be relaxed. Second, we highlight the notation differences. We use a to denote actions while Yue and Joachims [2009b] uses w. In our proof, we define $P_t(a) := \sigma(\mu(a_t) - \mu(a))$ to represent the probability of the event $a_t \succ a$. While in Yue and Joachims [2009b], it is denoted as $c_t(w)$. Just to highlight, its corrupted version $\hat{P}_t(a) := \sigma(\mu(a_t) - \mu(a) + c_t(a_t, a))$ is newly introduced in our paper. Moreover, $\bar{P}_t(a) := \mathbb{E}_{x \in \mathbb{B}}[P_t(\mathcal{P}_\mathcal{A}(a + \delta x))]$ is denoted as $\hat{\epsilon}_t(w) + \frac{1}{2}$ in Yue and Joachims [2009b]. The definition

of $\mathcal{F}\left(\mathcal{P}_{\mathcal{A}}\left(a_t+\delta u_t\right), a_t\right)$ is same as $X_t\left(\mathcal{P}_W(w_t+\delta u_t)\right)$ defined in Yue and Joachims [2009b], except different notation. We remark $\mathcal{P}_{\mathcal{A}}(a)$ represent the projection of action a on $\mathcal{A}$.

Algorithm 6: Dueling Bandit Gradient Descent (DBGD) [Yue and Joachims, 2009b]

Input: Exploration rate δ, exploitation rate γ, initial action $a_1 = \mathbf{0} \in \mathcal{A}$.

for $t \in [T]$ **do**

- Sample unit vector u_t uniformly and set $a_t' = \mathcal{P}_{\mathcal{A}}\left(a_t + \delta u_t\right)$.
- Present action pair (a_t, a_t') to user and receive corrupted dueling feedback $\hat{\mathcal{F}}(a_t', a_t)$.
- Compute gradient $\hat{g}_t = -\frac{d}{\delta}\hat{\mathcal{F}}(a_t', a_t)u_t$.
- Set learning rate $\eta = \frac{\gamma\delta}{d}$ and update $a_{t+1} = \mathcal{P}_{\mathcal{A}}(a_t - \eta\hat{g}_t)$.

Proposition (Proposition 3 restated)**.** *If the total corruption level satisfies* $\sum_{t=1}^{T}|c_t(a_t, a_t')| \leq \mathcal{O}(T^{\rho})$*, then for any* $\alpha \in (0, \frac{1}{4}]$*, if utility function* μ *is generally concave, for a sufficiently large round* T*, choosing* $\gamma := \frac{R}{\sqrt{T}}$ *and* $\delta := \frac{\sqrt{2Rd}}{\sqrt{13L^{\sigma}L^{\mu}T^{\alpha}}}$ *for Algorithm 6*

1. *(Robustness)* Reg_T *incurred by Algorithm 6 satisfies* $Reg_T \leq \mathcal{O}(\sqrt{d}T^{1-\alpha} + \sqrt{d}T^{\alpha+\rho})$.
2. *(Efficiency) There exists a linear utility function* μ *and a link function* σ *such that* Reg_T *suffered by Algorithm 6 is at least* $\Omega\left(T^{1-\alpha}\right)$ *in scenario without corruption.*

B.6.1 Proof for Robustness Statement in Proposition 3

Proof. The proof of Theorem 3 builds on the proof of Yue and Joachims [2009b], where we extend it to a setting in presence of agnostic corruption. The proof follows Step 1-3 introduced in Theorem 3.4.1. Similarly, in Step 1, we decompose bias in the gradient caused by corruption. Notice that without corruption, DBGD performs expected gradient ascent, where the gradient g_t is defined as follows

$$g_t = -\frac{d}{\delta}\mathcal{F}\left(\mathcal{P}_{\mathcal{A}}\left(a_t+\delta u_t\right), a_t\right)u_t.$$

In the presence of adversarial corruption, the corrupted gradient is

$$\hat{g}_t = -\frac{d}{\delta}\hat{\mathcal{F}}\left(\mathcal{P}_{\mathcal{A}}\left(a_t + \delta u_t\right), a_t\right) u_t.$$

Therefore, we quantify the bias $\mathbb{E}(b_t|a_t)$ in the following lemma.

Lemma B.6.1. *Let $P_t(a)$ represent the likelihood of action a_t being preferred over a, where $P_t(a) := \sigma(\mu(a_t) - \mu(a))$. Likewise, the corrupted version is defined as $\hat{P}_t(a) = \sigma(\mu(a_t) - \mu(a) + c_t(a_t, a))$. The smoothed version of P_t over $\mathcal{A}$ is $\bar{P}_t(a) := \mathbb{E}_{x\in\mathbb{B}}\left[P_t(\mathcal{P}_{\mathcal{A}}(a + \delta x))\right]$. Let $a'_t = \mathcal{P}_{\mathcal{A}}(a_t + \delta u_t)$, where u_t is uniformly sampled from $\mathbb{S}$ and $\hat{g}_t = -\frac{d}{\delta}\hat{\mathcal{F}}\left(a'_t, a_t\right) u_t$. Let $b_t := \frac{d}{\delta}\left(\hat{P}_t(a'_t) - P_t(a'_t)\right) u_t$. We have $\mathbb{E}(\hat{g}_t|a_t) = \mathbb{E}(\nabla\bar{P}_t(a_t)|a_t) - \mathbb{E}(b_t|a_t)$.*

Proof. **(Proof for Lemma B.6.1).** Since $a'_t := \mathcal{P}_{\mathcal{A}}(a_t + \delta u_t)$, we have

$$\begin{aligned}
\mathbb{E}(\hat{g}_t|a_t) &= \mathbb{E}_{u_t}(\mathbb{E}(\hat{g}_t|a_t, u_t)) \\
&= -\frac{d}{\delta}\mathbb{E}_{u_t}\left(\mathbb{E}(\hat{\mathcal{F}}(a'_t, a_t)u_t|a_t, u_t)\right) \\
&= -\frac{d}{\delta}\mathbb{E}\left(\hat{P}_t(a'_t)u_t|a_t\right) \\
&= -\frac{d}{\delta}\mathbb{E}\left(\left[P_t(a'_t) + \hat{P}_t(a'_t) - P_t(a'_t)\right] u_t|a_t\right) \\
&= \nabla\mathbb{E}_{x\in\mathbb{B}}\left(P_t(\mathcal{P}_{\mathcal{A}}(a_t + \delta x))|a_t\right) - \frac{d}{\delta}\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right) u_t|a_t\right] \qquad \text{(B.23)} \\
&= \nabla\bar{P}_t(a_t) - \frac{d}{\delta}\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right) u_t|a_t\right] \\
&= \mathbb{E}(g_t|a_t) - \frac{d}{\delta}\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right) u_t|a_t\right]. \qquad \text{(B.24)}
\end{aligned}$$

We get the equation (B.23) by using Lemma B.6.2. We get equation (B.24) by using Lemma B.6.3.

Lemma B.6.2 (Lemma 2 in Yue and Joachims [2009b])**.** *Fix $\delta > 0$, over random unit vector u, we have*

$$\mathbb{E}[P_t(\mathcal{P}_{\mathcal{A}}(a + \delta u))u] = \frac{\delta}{d}\nabla\bar{P}_t(a).$$

Lemma B.6.3 (Lemma 1 in Yue and Joachims [2009b]). $\mathbb{E}_{\mathcal{F},u}[\mathcal{F}(a'_t, a_t)u] = -\mathbb{E}_u[P_t(a'_t)u]$.

□

In Step 2, we decompose regret under dueling bandits into two parts, regret of decision and feedback error.

Lemma B.6.4 (Regret Decomposition for DBGD). *Define* $\lambda := \frac{L^\sigma}{L^\sigma - \delta L^\mu L_2}$, *$L_2$ is the Lipschitz constant for σ'*, $\gamma = R/\sqrt{T}$, *for b_t defined in Lemma B.6.1, we have*

$$Reg_T \leq \underbrace{\lambda\left(\frac{2Rd\sqrt{T}}{\delta} + 13\delta L^\sigma L^\mu T\right)}_{\textit{Regret of Decision}} + \underbrace{2\lambda\sum_{t=1}^{T}\mathbb{E}\left(b_t^\top(a_t - a^*)\right)}_{\textit{Feedback Error}}.$$

Proof. Because of Lemma B.6.5, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\bar{P}_t(a_t) - \bar{P}_t(a^*)\right] &\leq \sum_{t=1}^{T}\mathbb{E}\left(\lambda\nabla\bar{P}_t(a_t)(a_t - a^*) + (3+\lambda)\delta L^\sigma L^\mu\right) \\
&= \lambda\sum_{t=1}^{T}\mathbb{E}(\mathbb{E}(g_t|a_t)^\top(a_t - a^*)) + (3+\lambda)\delta L^\sigma L^\mu T \\
&= \lambda\sum_{t=1}^{T}\mathbb{E}(\mathbb{E}(\hat{g}_t + b_t|a_t)^\top(a_t - a^*)) + (3+\lambda)\delta L^\sigma L^\mu T \\
&= \lambda\sum_{t=1}^{T}\mathbb{E}(\hat{g}_t^\top(a_t - a^*)) + \lambda\sum_{t=1}^{T}\mathbb{E}(b_t^\top(a_t - a^*)) + (3+\lambda)\delta L^\sigma L^\mu T
\end{aligned}$$

Lemma B.6.5 (Lemma 4 in Yue and Joachims [2009b]). *Fix $\delta \in (0, \frac{L^\sigma}{L^\mu L_2})$, for λ defined in Lemma B.6.4, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\bar{P}_t(a_t) - \bar{P}_t(a^*)\right] \leq \sum_{t=1}^{T}\mathbb{E}\left(\lambda\nabla\bar{P}_t(a_t)^\top(a_t - a^*) + (3+\lambda)\delta L^\sigma L^\mu\right).$$

Since $a_{t+1} = \mathcal{P}_{\mathcal{A}}(a_t - \eta\hat{g}_t)$, and $\eta = \frac{\gamma\delta}{d}$, by applying the telescoping sum and because

$a_1 = 0$, $\|g_t\|_2 \leq G, \forall t$, we have

$$\lambda \sum_{t=1}^{T} \mathbb{E}\left(\hat{g}_t^\top (a_t - a^*)\right) \leq \lambda \left(\frac{R^2}{2\eta} + \frac{T\eta G^2}{2}\right).$$

□

Noticing that $\|b_t\|_2 \leq \frac{d}{\delta} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right)$. Moreover, $\|a_t - a^*\|_2 \leq 2R$. Therefore, in Step 3, we can control the feedback error by the following.

Lemma B.6.6. $\sum_{t=1}^{T} \mathbb{E}\left(b_t^\top (a_t - a^*)\right) \leq 2RdL^\sigma T^\rho/\delta$.

Proof. **(Proof for Lemma B.6.6).**

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}(b_t(a_t - a^*)) &\leq 2R\frac{d}{\delta} \sum_{t=1}^{T} \min\left(2, L^\sigma |c_t(a_t, a_t')|\right) \\
&\leq 2R\frac{dL^\sigma}{\delta} \sum_{t=1}^{T} |c_t(a_t, a_t')| \\
&= 2R\frac{dL^\sigma T^\rho}{\delta}.
\end{aligned}$$

□

Lemma B.6.7 (Lemma 3 in Yue and Joachims [2009b]). $Reg_T \leq 2\mathbb{E}\left[\sum_{t=1}^{T} \bar{P}_t(a_t) - \bar{P}_t(a^*)\right] + 5\delta L^\sigma L^\mu T$

Set $G = \frac{d}{\delta}$, and by Lemma B.6.7, we have

$$\mathrm{Reg}_T \leq \lambda \left(\frac{2Rd\sqrt{T}}{\delta} + 13\delta L^\sigma L^\mu T\right) + 4R\frac{\lambda d L^\sigma T^\rho}{\delta}.$$

Therefore, by choosing $\delta := \frac{\sqrt{2Rd}}{\sqrt{13L^\sigma L^\mu T^\alpha}}, \gamma := \frac{R}{\sqrt{T}}$, and $T > \left(\frac{\sqrt{2Rd}L_\mu L_2}{\sqrt{13L^\sigma L^\mu} L^\sigma}\right)^4$, we have

$$\mathrm{Reg}_T \leq 2\lambda_T \sqrt{26RdL^\sigma L^\mu} T^{1-\alpha} + 2L^\sigma \sqrt{26RdL^\sigma L^\mu} T^{\alpha+\rho}.$$

$\lambda_T = \frac{L^\sigma\sqrt{13L^\sigma L^\mu}T^\alpha}{L^\sigma\sqrt{13L^\sigma L^\mu}T^\alpha - L^\mu L_2\sqrt{2Rd}}$, completes the proof. □

B.6.2 Proof for Efficiency Statement in Proposition 3

Proof. The robustness statement in Proposition 3 implies that DBGD could afford agnostic corruption level $\mathcal{O}(T^{1-\alpha})$. This implies that there exists a hard instance which makes DBGD attain regret in order of $\Omega(T^{1-\alpha})$ in non-corrupt setting, formally described as follows. To start a proof by contradiction, we assume that the efficiency statement in Proposition 3 is false. Specifically, for all problem instance μ, σ, there exist a $\epsilon > 0$, such that $\text{Reg}_T \leq \mathcal{O}(T^{1-\alpha-\epsilon})$. We assume ϵ to be the least possible. In particular, there exists a pair of instance μ, σ such that $\text{Reg}_T = c_0 T^{1-\alpha-\epsilon}$, c_0 is a positive constant. In other words, it says that there exists a pair of instance μ, σ such that DBGD suffers regret in $\Theta(T^{1-\alpha-\epsilon})$.

Consider the following problem instance. $\mu(a) = \theta^\top a$, specifically μ is linear function. Let $d = 2$ with action set $\mathcal{A}_1 := \{(a_1, a_2) : a_1 \geq 0, a_2 \geq 0, \frac{1}{2}a_1 + a_2 - \frac{1}{4} \leq 0\}$ with $\theta = [\frac{1}{2}, \frac{1}{2}]$. The optimal arm is $a_1 = [\frac{1}{2}, 0]$, which is unique. Let the link function $\sigma(x) = \frac{1}{2} + \frac{1}{2}x$. It is easy to verify that it is rotational symmetric and Lipschitz. Denote Reg_T as the regret occurred by DBGD on this problem instance. We know $\text{Reg}_T \leq \mathcal{O}(T^{1-\alpha-\epsilon})$. Consequently, it implies that a_2 is at most proposed $8c_0 T^{1-\alpha-\epsilon}$ times. This is because $\text{Reg}_T \leq c_0 T^{1-\alpha-\epsilon}$ and if the proposed action pair (a_t, a_t') including a_2 at round t, it occurs regret at least $\frac{1}{8}$.

Now consider a different problem instance with the same μ, σ but different action set, which is $\mathcal{A}_2 := \{(a_1, a_2) : a_1 \geq 0, a_2 \geq 0, \frac{3}{2}a_1 + a_2 - \frac{3}{4} \leq 0\}$. The optimal arm in this action set $a_2 = [0, \frac{3}{4}]$, which is also unique. Proposing arm $a_1 = [\frac{1}{2}, 0]$ occurs at least $\frac{1}{8}$ regret. Consider an adversary that pulls the utility of a_2 from $\frac{3}{8}$ down to $\frac{1}{8}$ whenever it is pulled at a cost of $c_t(A_t) = \frac{1}{4}$. It is equivalent to say that every time when the proposed arm w falls in the region $\mathcal{W} = \{a_1 \geq 0, a_2 \geq 0, \frac{1}{2}a_1 + a_2 - \frac{1}{4} \geq 0, \frac{3}{2}a_1 + a_2 - \frac{3}{4} \leq 0\}$, we assume that the utility is sampled from $\theta^\top \mathcal{P}_{\mathcal{A}_1}(w)$ and $|c_t(w)| = |\theta^\top \mathcal{P}_{\mathcal{A}_1}(w) - \theta^\top w| \leq \frac{1}{4}$. Choosing the total corruption budget as $2c_0 T^{1-\alpha-\epsilon}$, the adversary can afford to corrupt $8c_0 T^{1-\alpha-\epsilon}$ times.

This means that as long as the adversary is corrupting, the utility observed in the first problem instance is exactly the same as the utility observed in the second problem instance, in which we pull a_2 as most $8c_0T^{1-\alpha-\epsilon}$. However, a_2 is the optimal arm in the second problem instance, which implies that the regret occurred on the second problem instance is at least $T - 8c_0T^{1-\alpha-\epsilon}$, which is linear in T. This contradicts the robustness statement in Proposition 3, which says that when agnostic corruption is $2c_0T^{1-\alpha-\epsilon}$, the regret upper bound is $O(T^{1-\epsilon})$, which is sublinear. To reconcile the conflicts, we must have the efficiency statement in Proposition 3 true to make the agnostic corruption budget at least $\Omega(T^{1-\alpha})$ to make the parallel world argument valid. □

B.7 Additional Experiments

In this section, we list all the experiment details. At the beginning, we introduce the regret order fitting method, and baseline algorithms for comparison.

Fitted Order of Regret. We introduce the methodology that we use to compute the order of Reg_T in terms of the number of iteration, T. To fit the order of the regret, we first convert it into log scale. Then we input the last 1% of data, run linear regression, and use ordinary least squares to estimate the slope of the line, which is the *fitted order* of Reg_T.

Baseline Algorithms. We consider three baseline algorithms for comparison, Doubler [Ailon et al., 2014] and Sparring [Ailon et al., 2014]

Doubler is the first approach that transforms a dueling bandit problem into a standard multi-armed bandit (MAB) problem. It operates in epochs of exponentially increasing length: in each epoch, the left arm is sampled from a fixed distribution, and the right arm is selected using a MAB algorithm to minimize regret against the left arm. The feedback received by the MAB algorithm is the number of wins the right arm achieves compared to the left arm. Under linear link assumption, Doubler has been proven to experience regret as the same order as underlying MAB algorithm. For continuous *action* space and general concave utility, we

choose Bandit Gradient Descent (BGD, [Flaxman et al., 2004]), with regret $\mathcal{O}(T^{3/4})$, as the underlying MAB algorithm.

Sparring initializes two MAB instances and lets them compete against each other. It is a heuristic improvement over Doubler. Although it does not come with a regret upper bound guarantee, it is reported to enjoy better performance compared to Doubler [Ailon et al., 2014]. We also choose BGD as the underlying MAB algorithm.

To validate efficiency-robustness trade-off in Proposition 2 and Proposition 3, we try different α values. In the first row of Figure B.1, we consider $\alpha = 0.05$ for DBGD and $\alpha = 0.9$ for RoSMID. We consider $\rho \in [0.5, 0.95]$. According to the theoretical prediction, both DBGD and RoSMID can tolerate at most $\mathcal{O}(T^{0.95})$ agnostic corruption, which aligns with experiment results. This is because the estimated order of regret is less than 1. In the second row of Figure B.1, we consider $\alpha = 0.1$ for DBGD and $\alpha = 0.8$ for RoSMID. According to the theoretical prediction, both DBGD and RoSMID can tolerate at most $\mathcal{O}(T^{0.9})$ agnostic corruption. From the figure, we can see they have smaller fitted order of regret when $\rho = 0.5$ while at a cost of tolerating smaller magnitude of agnostic corruption (it has linear regret when $\rho = 0.95$), which reveals intrinsic tradeoff between efficiency and robustness.

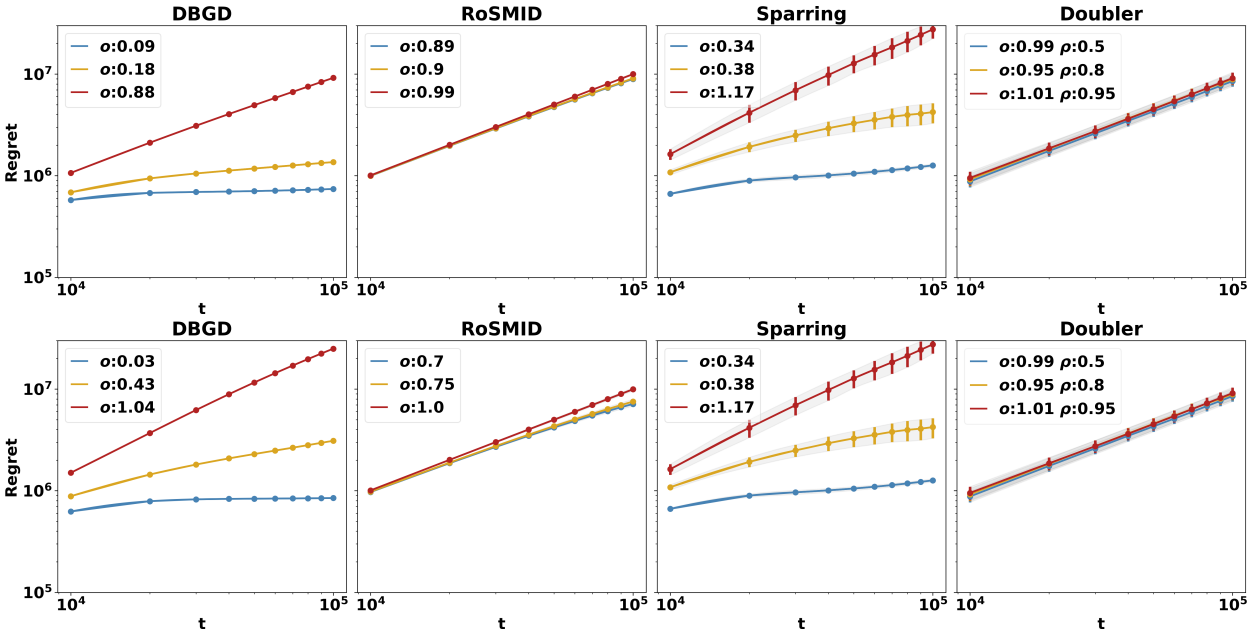

Figure B.1: In the first row, we consider $\alpha = 0.05$ for DBGD and $\alpha = 0.9$ for RoSMID. In the second row, we consider $\alpha = 0.1$ for DBGD and $\alpha = 0.8$ for RoSMID. Given the α, for each algorithm, we tested its performance under ρ-Imperfect Human feedback with $\rho = 0.5, 0.8, 0.95$. For each ρ, we presented a line plot of the average regret over five simulations, accompanied by $\pm$ one standard deviation shown by the shaded region. In the legend, o denotes the estimated line slope, calculated using least squares on the last 1% of the data.

B.7.1 Computational Complexity

Firstly, we emphasize that our robustified algorithm retains the same computational complexity as the originals. This indicates that incorporating robustness does not compromise computational efficiency, ensuring the practicality and scalability of the enhanced algorithm for real-world applications. Specifically, RoSMID has a running time of $O(d^3T)$, where the d^3 term arises from the singular value decomposition. Similarly, the robustified DBGD algorithm runs in $O(dT)$, assuming that $P_A(a_t + \delta u_t)$ can be efficiently computed through matrix transformations.

As our work is the first to address corruption-robust learning from a continuous-action strictly concave utility, there is no directly comparable literature for this exact setting. However, we provide relevant comparisons for reference. RCDB [Di et al., 2024], designed

for corruption-robust learning in a linear contextual dueling bandit framework, has a running time of $O(K^2d^2T)$, where K is the number of actions. Similarly, Versatile-DB [Saha and Gaillard, 2022], which addresses dueling bandit problems in a multi-armed bandit setting under adversarial attacks, has a running time of $O(K^2T)$, where K represents the number of arms. K^2 comes from solving $\arg\max_{\boldsymbol{x}\in S}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle - f(\boldsymbol{x}))$ using second order Newton methods and f is a convex function.

APPENDIX C

APPENDIX TO SINGLE-AGENT POISONING ATTACKS

C.1 Examples of Second-Order Smooth Monotone Games

In this section, we provide two explicit examples of second-order L-smooth and β-strongly monotone games introduced in Definition 4.2.3. These examples offer a concrete sense of what we can expect regarding the outcome of a utility poisoning attack in such games. The two cases we will discuss are the n-person Cournot Competition Cournot [1838] and the Tullock contest Tullock [2008] (a.k.a. resource allocation auctions with costs).

Cournot competition is an economic model describing a number of firms independently competing on the amount of output they will produce. Each agent-i's pure strategy x_i is the quantity of product, and the payoff is determined by the marginal return of a unit of product which decreases w.r.t. the opponents' total production $(a - b\sum_{i=1}^{n})$, and a marginal cost c_i. The utility of agent-i is given by

$$u_i(x_i, x_{-i}) = x_i \left(a - b \sum_{j=1}^{n} x_j \right) - c_i x_i. \tag{C.1}$$

Tullock contest involves n agents competing for a unit amount of prize in a "winner-takes-all" framework. Each agent i's pure strategy is to choose an effort level $x_i \in [0, 1]$, and the winning probability is proportional to each agent's effort level. The prize is awarded to the contestant with the highest relative effort, and the payoff of each agent is the expected reward minus a cost given by some function of the invested effort. Specifically, the payoff functions of agent-i is given by

$$u_i(x_i, x_{-i}) = \frac{x_i}{a + \sum_{j=1}^{n} x_j} - c_i(x_i), \tag{C.2}$$

where $a > 0$ models the exogenous factor that affects the outcome of the contest, and

$c_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is an increasing cost function.

According to Rosen [1965], a sufficient condition to verify whether a game is β-strongly monotone (a.k.a. β-diagonal strict concavity, i.e., β-DSC) can be summarized in the following Lemma C.1.1.

Lemma C.1.1. *A sufficient condition for a game $\mathcal{G}(n, \{\mathcal{X}_i\}, \{u_i\})$ to be β-DSC is that*

1. *Each $\mathcal{X}_i$ is compact and convex, and $u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})$ is concave in $\boldsymbol{x}_i$ and convex in $\boldsymbol{x}_{-i}$.*
2. *$\mathcal{G}$'s negative symmetric game Hessian defined as $-H(\boldsymbol{x}) = -\frac{1}{2}\left(H_{\mathcal{G}}(\boldsymbol{x}) + H_{\mathcal{G}}^{\top}(\boldsymbol{x})\right)$ is negative definite and its smallest eigenvalue is at least β.*

And to verify the second-order L-smooth condition, we only need to directly apply Definition 4.2.3. The following two propositions formalize our claims.

Proposition 5. *An n-person Cournot competition $\mathcal{G}$ with payoff functions defined by (C.1) satisfies:*

1. *each agent's utility function u_i is $(a+b+c_i)$-Lipschitz in x_i;*
2. *u_k is $\frac{\sqrt{n-1}}{2}$-Lipschitz in $\boldsymbol{x}_{-k}$;*
3. *$\mathcal{G}$ is $2b$-strongly monotone;*
4. *the spectral norm of the $\mathcal{G}$'s Hessian (defined in (4.4)) satisfies $\rho(H_{\mathcal{G}}) \leq (n+1)b$;*
5. *all agents' utility functions are second-order 0-smooth;*

Proof. Because u_i is continuous in x_i, an Lipschitz constant can be $\max_{\boldsymbol{x}\in\mathcal{X}} |\nabla_{x_i} u_i(x_i, \boldsymbol{x}_{-i})| \leq |a - b\sum_{j=1}^{n} x_j - bx_i - c_i|$. Since $a - b\sum_{j=1}^{n} x_j \geq 0$ and $bx_i \geq 0, c_i \geq 0$, we have $|a - b\sum_{j=1}^{n} x_j - bx_i - c_i| \leq a + b + c_i$.

The best response mapping has a closed form $BR_k(\boldsymbol{x}_{-k}) = \frac{a-b\sum_{j\neq k} x_j - c_k}{2b}$ and we can verify

$$|BR_k(\boldsymbol{x}_{-k}) - BR_k(\boldsymbol{x}'_{-k})| = \frac{1}{2b}\left|b\sum_{j\neq k}(x'_j - x_j)\right| \leq \frac{\sqrt{n-1}}{2}\|\boldsymbol{x}_{-k} - \boldsymbol{x}'_{-k}\|.$$

Therefore, u_k is $\frac{\sqrt{n-1}}{2}$-Lipschitz in $\boldsymbol{x}_{-k}$.

For strongly monotonicity, simple calculation shows the Hessian of n-person Cournot game has the following form

$$-H_{ij}(\boldsymbol{x}) = b(1+\delta_{ij}), \tag{C.3}$$

where $\delta_{ij} = \mathbb{I}[i=j]$. We can easily verify that $H = b(I + \mathbf{1}\mathbf{1}^\top)$ and the smallest and largest eigenvalues of H are $2b$ and $(n+1)b$. Therefore, the game is $2b$-strongly monotone, and $\rho(H_{\mathcal{G}}) = \rho(H) = (n+1)b$.

To verify second-order smoothness, we directly compute function $h_i(\boldsymbol{x};\delta)$ for agent-i from (4.5), which gives

$$\begin{aligned}&\frac{v_i(x_i, \boldsymbol{x}_{-i}) - v_i(x_i+\delta_i, \boldsymbol{x}_{-i})}{\delta}\\ =&\frac{(a - b\sum_{j=1}^n x_j - c_i - bx_i) - (a - b\sum_{j=1}^n x_j - b\delta - c_i - bx_i - b\delta)}{\delta} = 2b,\end{aligned}$$

which means for any $\boldsymbol{x}\in\mathcal{X}$ and $\delta\neq 0$, $h_i(\boldsymbol{x};\boldsymbol{\delta})$ is a constant function. Therefore, the game is second-order smooth for any $L\geq 0$.

□

Proposition 6. *Consider an n-person Tullock contest $\mathcal{G}$ with payoff functions defined by (C.2) in which each agent's cost function c_i is β_i-strongly convex with its r-th order derivative bounded in $[-M_r, M_r], 1\leq r\leq 3$. Then $\mathcal{G}$ satisfies:*

1. *each agent's utility function u_i is $\left(\frac{1}{a}+M_1\right)$-Lipschitz in x_i;*

2. *u_k is $\sqrt{n-1}\left(\frac{2}{a^3\beta_k}+\frac{1}{a^2\beta_k}\right)$-Lipschitz in $\boldsymbol{x}_{-k}$;*

3. *$\mathcal{G}$ is $\left(\min_{i\in[n]}\{\beta_i\}+\frac{a}{(a+n)^3}\right)$-strongly monotone;*

4. *the spectral norm of the $\mathcal{G}$'s Hessian (defined in (4.4)) satisfies $\rho(H_{\mathcal{G}}) \leq \frac{n+1}{a^2} + \max_{i\in[n]}\{\beta_i\}$;*

5. *all agents' utility functions are second-order $\left(\frac{4\sqrt{n}}{a^3}+\frac{6\sqrt{n}}{a^4}+\frac{2}{a^3}+M_3\right)$-smooth;*

Proof. We prove these claims one by one.

1. Because u_i is continuous in x_i, an Lipschitz constant can be

$$\begin{aligned}\max_{\boldsymbol{x}\in\mathcal{X}}|\nabla_{x_i}u_i(x_i,\boldsymbol{x}_{-i})| &= \max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{a+\sum_{j\neq i}x_j}{(a+\sum_{j=1}^n x_j)^2}-\nabla c_i\right| \\ &\leq \frac{1}{|a+\sum_{j=1}^n x_j|}+|\nabla c_i| \leq \frac{1}{a}+M_1.\end{aligned}$$

2. From the implicit function theorem,

$$\nabla_{\boldsymbol{x}_{-k}}BR_k(\boldsymbol{x}_{-k}) = \frac{\partial \arg\max_t u_k(t,\boldsymbol{x}_{-k})}{\partial \boldsymbol{x}_{-k}} = -\left(\frac{\partial^2 u_k(x_k,\boldsymbol{x}_{-k})}{\partial x_k^2}\right)^{-1}\cdot\frac{\partial^2 u_k(x_k,\boldsymbol{x}_{-k})}{\partial x_k\partial \boldsymbol{x}_{-k}}. \tag{C.4}$$

From Lagrange mean value theorem for multi-variable functions, there exists some $\boldsymbol{y}_{-k}\in\mathcal{X}_{-k}$ such that

$$BR_k(\boldsymbol{x}_{-k})-BR_k(\boldsymbol{x}'_{-k}) = \nabla^\top_{\boldsymbol{x}_{-k}}BR_k(\boldsymbol{y}_{-k})(\boldsymbol{x}_{-k}-\boldsymbol{x}'_{-k}).$$

As a result, a Lipschitz constant for function $BR_k(\boldsymbol{x}_{-k})$ can be $\max_{\boldsymbol{y}_{-k}\in\mathcal{X}_{-k}}|\nabla_{\boldsymbol{x}_{-k}}BR_k(\boldsymbol{y}_{-k})|$.

Direct calculation shows that for any i, j,

$$\nabla_{ii} u_i(\boldsymbol{x}) = -2\left(a + \sum_{j\neq i} x_j\right)\left(a + \sum_{i=1}^n x_i\right)^{-3} - \nabla^2 c_i(x_i), \tag{C.5}$$

$$\nabla_j \nabla_i u_i(\boldsymbol{x}) = \left(\sum_{i=1}^n x_i - 2\sum_{j\neq i}^n x_j - a\right)\left(a + \sum_{i=1}^n x_i\right)^{-3}. \tag{C.6}$$

Since u_k is β_k-strongly concave, we have $|\nabla_{ii} u_i(x_i, \boldsymbol{x}_{-i})| \geq \beta_i$, and $|\nabla_j \nabla_i u_i(\boldsymbol{x})| \leq \frac{2}{a^3} + \frac{1}{a^2}$. Substitute them into (C.4) we obtain

$$|\nabla_{\boldsymbol{x}_{-k}} BR_k(\boldsymbol{x}_{-k})| \leq \sqrt{n-1}\left(\frac{2}{a^3\beta_k} + \frac{1}{a^2\beta_k}\right).$$

3. To see why $\mathcal{G}$ is $\left(\min_{i\in[n]}\{\beta_i\} + \frac{a}{(a+n)^3}\right)$-strongly monotone, we need to show $-H_{\mathcal{G}}$ is positive definite and then pin down its smallest eigenvalue. From (C.5) and (C.6) we can derive

$$\begin{aligned} -H_{\mathcal{G}} &= \left(a + \sum_{i=1}^n x_i\right)^{-3} \cdot \begin{bmatrix} 2(a + \sum_{i\neq 1} x_i) & a + \sum_{i\notin\{1,2\}} x_i & \dots \\ a + \sum_{i\notin\{1,2\}} x_i & 2(a + \sum_{i\neq 2} x_i) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} + \begin{bmatrix} \nabla^2 c_1(x_1) & 0 & \dots \\ 0 & \nabla^2 c_2(x_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \\ &\triangleq \left(a + \sum_{i=1}^n x_i\right)^{-3} \cdot M(a, \boldsymbol{x}) + \operatorname{diag}(\nabla^2 c_1(x_1), \cdots, \nabla^2 c_n(x_n)). \end{aligned} \tag{C.7}$$

For any $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathbb{R}^n$, we have

$$
\begin{aligned}
-\boldsymbol{y}^\top M(a,\boldsymbol{x})\boldsymbol{y} &= 2\sum_{i=1}^{n} y_i^2 \left(a + \sum_{j\neq i} x_j\right) + 2\sum_{i<j} y_i y_j \left(a + \sum_{k\notin\{i,j\}} x_k\right) \\
&= \sum_{i=1}^{n} y_i^2 \sum_{j\neq i} x_j + \sum_{i=1}^{n} x_i \left(\sum_{j\neq i} y_j\right)^2 + a\left(\sum_{i=1}^{n} y_i^2 + \left(\sum_{i=1}^{n} y_i\right)^2\right) \qquad \text{(C.8)} \\
&\geq a\|\boldsymbol{y}\|_2^2.
\end{aligned}
$$

Therefore, $\lambda_{\min}(-M(a,\boldsymbol{x})) \geq a$ and

$$
\lambda_{\min}(-H_{\mathcal{G}}) \geq \frac{a}{(a+\sum_{i=1}^{n} x_i)^3} + \beta_0 \geq \frac{a}{(a+n)^3} + \beta_0, \qquad \text{(C.9)}
$$

where $\beta_0 = \min_{i\in[n]}\{\beta_i\}$.

4. $\mathcal{G}$ is second-order $\left(\frac{4\sqrt{n}}{a^3} + \frac{3\sqrt{n}}{a^4}\right)$-smooth:

Direct calculation shows that

$$
\begin{aligned}
h_i(\boldsymbol{x};\delta) &= \frac{v_i(x_i,\boldsymbol{x}_{-i}) - v_i(x_i+\delta,\boldsymbol{x}_{-i})}{\delta} \\
&= \frac{1}{(A+\delta)^2} + \frac{1}{A(A+\delta)} - \frac{x_i}{A(A+\delta)^2} - \frac{x_i}{A^2(A+\delta)} + \frac{\nabla c_i(x_i+\delta) - \nabla c_i(x_i)}{\delta},
\end{aligned}
$$

where $A = a + \sum_{i=1}^{n} x_i$. From the mean value theorem for vector-valued functions Hall and Newell [1979], there exists $\boldsymbol{y} = \boldsymbol{x}' + \lambda(\boldsymbol{x} - \boldsymbol{x}')$ such that

$$
\|h_i(\boldsymbol{x};\delta) - h_i(\boldsymbol{x}';\delta)\|_2 \leq \|J(\boldsymbol{y})\| \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|_2,
$$

where $J(\boldsymbol{y})$ is the Jacobian matrix of h_i at $\boldsymbol{y}$, and $\|\cdot\|$ denotes the spectral norm of a

matrix, defined as

$$\|M\| = \sup_{\|\boldsymbol{x}\|_2=1} \|Mx\|_2 = \sqrt{\sigma_{\max}(M^T M)}. \tag{C.10}$$

Therefore, the smallest constant L such that $\mathcal{G}$ is second-order smooth is

$$\|J(\boldsymbol{y})\| \leq \|J_0\| + \|J_1\| + \|J_2\| + \|J_3\| + \|J_4\|, \tag{C.11}$$

where J_0, J_1, J_2, J_3, J_4 are the Jacobian matrices of functions $f(\boldsymbol{x}) = \frac{1}{(A+\delta)^2}, f(\boldsymbol{x}) = \frac{1}{A(A+\delta)}, f(\boldsymbol{x}) = \frac{x_i}{A(A+\delta)^2}, f(\boldsymbol{x}) = \frac{x_i}{A^2(A+\delta)}, f(\boldsymbol{x}) = \frac{\nabla c_i(x_i+\delta) - \nabla c_i(x_i)}{\delta}$. For a single-valued functions f, $J(f) = \nabla f$ and thus $\|J(f)\| = \|\nabla f^\top \nabla f\|$. From straightforward calculation we obtain

$$\begin{aligned}
\|J_0\| &= \frac{2}{(A+\delta)^3}\|\boldsymbol{x}\| \leq \frac{2\sqrt{n}}{a^3}, \\
\|J_1\| &= \frac{2A+\delta}{A^2(A+\delta)^2}\|\boldsymbol{x}\| \leq \frac{2\sqrt{n}}{a^3}, \\
\|J_2\| &= \left\| -x_i \frac{3A^2 + 4A\delta + 2\delta^2}{A^2(A+\delta)^4}\boldsymbol{x} + \frac{1}{A(A+\delta)^2}\boldsymbol{e}_i \right\| \leq \frac{3\sqrt{n}}{a^4} + \frac{1}{a^3}, \\
\|J_3\| &= \left\| -x_i \frac{3A^2 + 2A\delta}{A^4(A+\delta)^2}\boldsymbol{x} + \frac{1}{A^2(A+\delta)}\boldsymbol{e}_i \right\| \leq \frac{3\sqrt{n}}{a^4} + \frac{1}{a^3}, \\
\|J_4\| &\leq \|\nabla^3 c_i\| \leq M_3.
\end{aligned}$$

Hence, the game is second order $\left(\frac{4\sqrt{n}}{a^3} + \frac{6\sqrt{n}}{a^4} + \frac{2}{a^3} + M_3\right)$-smooth.

5. Upperbound of $\rho(H_{\mathcal{G}})$.

From (C.8) it holds that

$$
\begin{aligned}
-\boldsymbol{y}^\top M(a,\boldsymbol{x})\boldsymbol{y} &= \sum_{i=1}^n y_i^2 \sum_{j\neq i} x_j + \sum_{i=1}^n x_i \left(\sum_{j\neq i} y_j\right)^2 + a\left(\sum_{i=1}^n y_i^2 + \left(\sum_{i=1}^n y_i\right)^2\right) \\
&\leq \left[(n+1)\sum_{i=1}^n x_i\right]\|\boldsymbol{y}\|_2^2 + a(n+1)\|\boldsymbol{y}\|_2^2 \\
&= \left[(n+1)\left(a+\sum_{i=1}^n x_i\right)\right]\|\boldsymbol{y}\|_2^2,
\end{aligned}
$$

therefore, by (C.7) we have

$$
\begin{aligned}
\|\boldsymbol{y}^\top H_{\mathcal{G}}\boldsymbol{y}\| &\leq \frac{n+1}{(a+\sum_{i=1}^n x_i)^2}\|\boldsymbol{y}\|_2^2 + \max_{i\in[n]}\{\beta_i\}\|\boldsymbol{y}\|_2^2 \\
&\leq \left(\frac{n+1}{a^2} + \max_{i\in[n]}\{\beta_i\}\right)\|y\|_2^2.
\end{aligned}
$$

Hence, an upperbound of $\rho(H_{\mathcal{G}})$ is $\frac{n+1}{a^2} + \max_{i\in[n]}\{\beta_i\}$.

□

C.2 Proofs in Section 4.3

In this section, we present the missing proofs from Section 4.3, including the proofs of Lemma 4.3.1 and Theorem 4.3.1 which fully characterize the theoretical guarantees of SUSA, and the proof of Proposition 4 along with an extended argument regarding an additional attack success guarantee of SUSA.

Notably, we establish these results under a broader setting where the adversary can attack multiple agents, rather than being restricted to a single agent. Specifically, we consider an adversary targeting a subset of agents $\mathcal{V} \subseteq [n]$, and for each victim agent-$k \in \mathcal{V}$, the adversary performs SUSA$(k, \boldsymbol{\delta}_k, \Delta_k)$ on agent-k. We specify such an utility poisoning attack as the tuple $(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$, and we denote the corresponding adversary as

SUSA$(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$.

The proofs provided in this section apply to any SUSA$(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$, and thus also apply to the special case when $\mathcal{V} = \{k\}$, which is the setting considered in the main paper. For simplicity of notations, we denote the vector $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_n) \in \mathbb{R}^{nd}$, where $\boldsymbol{\delta}_i = \boldsymbol{\delta}_i$ if $i \in \mathcal{V}$, and $\boldsymbol{\delta}_i = \mathbf{0}$ if $i \notin \mathcal{V}$. In the following, we restate the generalized versions of our results in Section 4.3 under SUSA$(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$ and then present their proofs.

C.2.1 Proof of Lemma 4.3.1

We prove the following generalized version of Lemma 4.3.1 as follows:

Lemma C.2.1. *For any β-strongly monotone and second-order L-smooth game $\mathcal{G}(n, \{\mathcal{X}_i\}_{i=1}^n, \{u_i\}_{i=1}^n)$ under SUSA$(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$, the corrupted game $\tilde{G}(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$ with $\|\boldsymbol{\delta}\|_2 < \beta/L$ remains strongly monotone.*

Proof. By the definition of strongly monotonicity, for any $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n), \boldsymbol{x}' = (\boldsymbol{x}'_1, \cdots, \boldsymbol{x}'_n) \in \mathcal{X}$, it holds that

$$\sum_{i=1}^{n} (\boldsymbol{x}'_i - \boldsymbol{x}_i)^\top (v_i(\boldsymbol{x}'_i, \boldsymbol{x}'_{-i}) - v_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})) \leq -\beta \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2, \tag{C.12}$$

where $v_i(\boldsymbol{x}) = \nabla_{\boldsymbol{x}_i} u_i(\boldsymbol{x})$. Without loss of generality let's assume the SUSA targets at all agents (for a agent $i \notin \mathcal{V}$, simply let $\boldsymbol{\delta}_i = \mathbf{0}$ and $\Delta_i \equiv 0$). Then for agent-1's corrupted utility function $\tilde{u}_1$, it holds that $\tilde{v}_1(\boldsymbol{x}) = \nabla_{\boldsymbol{x}_1} \tilde{u}_1(\boldsymbol{x}) = v_1(\boldsymbol{x}_1 + \boldsymbol{\delta}_1, \boldsymbol{x}_{-1})$. By the definition of strictly monotonicity, what we need to show is that there exists a $\beta_0 > 0$ such that

$$\sum_{i=1}^{n} (\boldsymbol{x}'_i - \boldsymbol{x}_i)^\top (v_i(\boldsymbol{x}'_i + \boldsymbol{\delta}_i, \boldsymbol{x}'_{-i}) - v_i(\boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{x}_{-i})) < -\beta_0 \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2. \tag{C.13}$$

According to the second-order L-smooth condition, for any $i \in [n]$ and $\boldsymbol{\delta}_i \in \mathbb{R}^d$, the

function $h_i(\boldsymbol{x}) : \mathbb{R}^{nd} \to \mathbb{R}^d$ defined as

$$h_i(\boldsymbol{x}) = \frac{v_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}) - v_i(\boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{x}_{-i})}{\|\boldsymbol{\delta}_i\|} \tag{C.14}$$

is L-Lipschitz in $\boldsymbol{x}$. As a result, we have

$$\begin{aligned}
&|\text{LHS of (C.12)} - \text{LHS of (C.13)}| \\
\leq &\sum_{i=1}^n \left| (\boldsymbol{x}_i' - \boldsymbol{x}_i)^\top (v_i(\boldsymbol{x}_i', \boldsymbol{x}_{-i}') - v_i(\boldsymbol{x}_i' + \boldsymbol{\delta}_i, \boldsymbol{x}_{-i}') + v_i(\boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{x}_{-i}) - v_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})) \right| \\
= &\sum_{i=1}^n \|\boldsymbol{\delta}_i\| \cdot \left| (\boldsymbol{x}_i' - \boldsymbol{x}_i)^\top (h_i(\boldsymbol{x}') - h_i(\boldsymbol{x})) \right| \\
\leq &\sum_{i=1}^n \|\boldsymbol{\delta}_i\| \cdot \|\boldsymbol{x}_i' - \boldsymbol{x}_i\| \cdot \|h_i(\boldsymbol{x}') - h_i(\boldsymbol{x})\| \\
\leq &\sum_{i=1}^n \|\boldsymbol{\delta}_i\| \cdot \|\boldsymbol{x}_i' - \boldsymbol{x}_i\| \cdot L\|\boldsymbol{x}' - \boldsymbol{x}\| \\
= &L\|\boldsymbol{x}' - \boldsymbol{x}\| \cdot \sum_{i=1}^n \|\boldsymbol{\delta}_i\| \cdot \|\boldsymbol{x}_i' - \boldsymbol{x}_i\| \\
\leq &L\|\boldsymbol{x}' - \boldsymbol{x}\| \cdot \left(\sum_{i=1}^n \|\boldsymbol{\delta}_i\|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^n \|\boldsymbol{x}_i' - \boldsymbol{x}_i\|^2 \right)^{\frac{1}{2}} \\
\leq &L\|\boldsymbol{x}' - \boldsymbol{x}\|^2 \cdot \left(\sum_{i=1}^n \|\boldsymbol{\delta}_i\|^2 \right)^{\frac{1}{2}}.
\end{aligned}$$

Since $\left(\sum_{i=1}^n \|\boldsymbol{\delta}_i\|^2\right)^{\frac{1}{2}} < \frac{\beta}{L}$, there must exist a $\beta_0 \in (0, \beta]$ such that $\left(\sum_{i=1}^n \|\boldsymbol{\delta}_i\|^2\right)^{\frac{1}{2}} \leq \frac{\beta - \beta_0}{L}$. As a result, the difference between the LHS of (C.12) and the LHS of (C.13) does not exceed $(\beta - \beta_0)\|\boldsymbol{x}' - \boldsymbol{x}\|^2$. Hence, (C.13) holds and the corrupted game $\tilde{\mathcal{G}}$ remains β_0-strongly monotone.

□

C.2.2 Proof of Theorem 4.3.1

We prove the following generalized version of Theorem 4.3.1:

Theorem C.2.1 (Vulnerability of MAL algorithms for strongly monotone games under SUSA$(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$)**.** *Consider a SUSA$(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in\mathcal{V}}, \{\Delta_k\}_{k\in\mathcal{V}})$ to a β-strongly monotone game $\mathcal{G}(n, \{\mathcal{X}_i\}_{i=1}^n, \{u_i\}_{i=1}^n)$ (with unique NE $\boldsymbol{x}^*$). Suppose the victim agent-k's utility function u_k is second-order L_0-smooth for all $k \in \mathcal{V}$. All agents run an (α, p)-MAL algorithm $\mathcal{A}$ for T rounds. If the following conditions are satisfied:*

1. $\|\boldsymbol{\delta}\|_2 < \beta/L_0$,
2. $\Delta(\boldsymbol{x}_{-k}, \boldsymbol{\delta}) = -u_k(BR_k(\boldsymbol{x}_{-k}), \boldsymbol{x}_{-k}) + u_k(BR_k(\boldsymbol{x}_{-k}) - \boldsymbol{\delta}, \boldsymbol{x}_{-k}), \forall k \in \mathcal{V}$, *where* $BR_k(\boldsymbol{x}_{-k}) \triangleq \arg\max_{\boldsymbol{x}_k\in\mathcal{X}_k} u_k(\boldsymbol{x}_k, \boldsymbol{x}_{-k})$ *is the best response mapping of agent-k,*
3. *u_i is L_1-Lipschitz in $\boldsymbol{x}_i, \forall i \in [n]$ and BR_k is L_2-Lipschitz in $\boldsymbol{x}_{-k}, \forall k \in \mathcal{V}$.*

Then, the resulting dynamics induced by $\mathcal{A}$ converges to some $\tilde{\boldsymbol{x}}^*$, such that

1. *The deviation to the original NE satisfies*

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*\|_2 \geq \beta\|\boldsymbol{\delta}\|_2 / \sup_{\boldsymbol{x}\in[\boldsymbol{x}^*, \tilde{\boldsymbol{x}}^*]} \{\rho(H_{\mathcal{G}}(\boldsymbol{x}))\}, \tag{C.15}$$

   where $\rho(H_{\mathcal{G}}(\boldsymbol{x}))$ represents the spectral norm of the game's Hessian $H_{\mathcal{G}}$ at $\boldsymbol{x}$.

2. *The expected total budget satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} |c_t|\right] \leq C_0 \cdot T^{1-\frac{p\alpha}{p+1}}, \tag{C.16}$$

   where the constant $C_0 = [CL_1(4L_2 + 5) + 2] \cdot |\mathcal{V}|$.

Proof. From Lemma C.2.1 we know that the corrupted game $\tilde{G}$ is strongly monotone and thus also has a unique NE $\tilde{\boldsymbol{x}}^*$. We first give an estimation of the distance between $\boldsymbol{x}^*$ and $\tilde{\boldsymbol{x}}^*$.

Define the best response mapping $f_i(\boldsymbol{x}_{-i}) = \arg\max_t u_i(t, \boldsymbol{x}_{-i})$ for any $i \in [n]$. Then by the definition of NE, $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_n^*)$ is the unique stationary point of the system

$$\begin{cases} \boldsymbol{x}_1 = f_1(\boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n), \\ \boldsymbol{x}_2 = f_2(\boldsymbol{x}_1, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n), \\ \vdots \\ \boldsymbol{x}_n = f_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}), \end{cases} \tag{C.17}$$

while $\tilde{\boldsymbol{x}}^* = (\tilde{\boldsymbol{x}}_1^*, \cdots, \tilde{\boldsymbol{x}}_n^*)$ is the unique stationary point of the system

$$\begin{cases} \boldsymbol{x}_1 = f_1(\boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n) - \boldsymbol{\delta}_1, \\ \boldsymbol{x}_2 = f_2(\boldsymbol{x}_1, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n) - \boldsymbol{\delta}_2, \\ \vdots \\ \boldsymbol{x}_n = f_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}) - \boldsymbol{\delta}_n, \end{cases} \tag{C.18}$$

where $\boldsymbol{\delta}_i = \mathbf{0}$ if $i \notin \mathcal{V}$. If we let $F(\boldsymbol{x}) = (f_1(\boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n), f_2(\boldsymbol{x}_1, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n),$ $\cdots, f_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}))$ as an n-valued function and $G(\boldsymbol{x}) = F(\boldsymbol{x}) - \boldsymbol{x}$, (C.17) and (C.18) can be further expressed as

$$\begin{aligned} G(\boldsymbol{x}^*) &= \mathbf{0}, \\ G(\tilde{\boldsymbol{x}}^*) &= (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_n). \end{aligned}$$

When $\mathcal{G}$ is twice differentiable, from the mean value theorem for vector valued functions

[Hall and Newell, 1979], there exists $\lambda \in [0,1]$ such that

$$\|G(\boldsymbol{x}^*) - G(\tilde{\boldsymbol{x}}^*)\| \leq \max_{\lambda\in[0,1]} \rho(J(\lambda\tilde{\boldsymbol{x}}^* + (1-\lambda)\boldsymbol{x}^*)) \cdot \|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}^*\|, \tag{C.19}$$

where $J(\boldsymbol{x})$ is the Jacobian of $\mathcal{G}$ at $\boldsymbol{x}$, and $\rho(J) \triangleq \sup_{\|\boldsymbol{u}\|_2=1} \|J\boldsymbol{u}\|_2$ is the spectral norm of J. Let $\rho_{\max}(J) \triangleq \max_{\lambda\in[0,1]}(\rho(J(\lambda\tilde{\boldsymbol{x}}^* + (1-\lambda)\boldsymbol{x}^*)))$, we immediately obtain

$$\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}^*\| \geq \|\boldsymbol{\delta}\|/\rho_{\max}(J). \tag{C.20}$$

Next we estimate the upper bound of $\rho(J)$. In fact, from implicit function theorem, we can derive the (i,j)-th element of J as

$$\begin{aligned}
&J_{i,i} = -1, 1 \leq i \leq n, \\
&J_{i,j} = \frac{\partial f_i(\boldsymbol{x}_{-i})}{\partial \boldsymbol{x}_j} = \frac{\partial \arg\max_t u_i(t, \boldsymbol{x}_{-i})}{\partial \boldsymbol{x}_j} = -\left(\frac{\partial^2 u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})}{\partial \boldsymbol{x}_i^2}\right)^{-1} \cdot \frac{\partial^2 u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_j}.
\end{aligned}$$

Note that the n-by-n matrix $H = \left[\frac{\partial^2 u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_j}\right]$ is exactly the Hessian of $\mathcal{G}$ (i.e., $H_{\mathcal{G}}$), and the diagonal matrix $D = \mathrm{diag}(H_{\mathcal{G}})$. Thus, the Jacobian J can be represented as

$$J = D^{-1} H_{\mathcal{G}}. \tag{C.21}$$

Since $\mathcal{G}$ is β-strongly monotone, each diagonal element of D is lower bounded by $\beta > 0$. Hence, the spectral norm of J can be upper bounded by

$$\rho_{\max}(J) \leq \frac{\rho_{\max}(H_{\mathcal{G}})}{\beta}, \tag{C.22}$$

and

$$\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}^*\| \geq \beta\|\boldsymbol{\delta}\|/\rho_{\max}(H_{\mathcal{G}}),$$

where $\rho_{\max}(H_{\mathcal{G}}) \triangleq \max_{\lambda\in[0,1]}(\rho(H_{\mathcal{G}}(\lambda\tilde{\boldsymbol{x}}^* + (1-\lambda)\boldsymbol{x}^*)))$.

Next, we derive a upper bound of the total corruption budget needed by an SUSA adversary. Since $\mathcal{A}$ is an (α, p)-MAL algorithm and $\tilde{\mathcal{G}}$ is strictly monotone, the resulting playing sequence $\{\boldsymbol{x}_t\}$ converges to the unique NE of $\tilde{\mathcal{G}}$ with a polynomial rate specified by

$$\mathbb{E}[\|\boldsymbol{x}_t - \tilde{\boldsymbol{x}}^*\|_2^p] \leq C^p \cdot t^{-p\alpha}.$$

Let γ be any number such that $0 < \gamma < p\alpha$. From Chebyshev's inequality, with probability at least $1 - t^{-p\alpha+\gamma}$, $\|\boldsymbol{x}_t - \tilde{\boldsymbol{x}}^*\|_2^p \leq C^p \cdot t^{-\gamma}$, and the following bounds also hold:

$$\left\|\boldsymbol{x}_{t,i} - \tilde{\boldsymbol{x}}_i^*\right\|_2^p \leq C^p \cdot t^{-\gamma}, \forall i \in [n], t \in [T]. \tag{C.23}$$

Denote the attacking cost at round t for a agent $i \in \mathcal{V}$ as $c_{t,i}$, which is given by $c_{t,i} = |\tilde{u}_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - u_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i})|$. By the choice of Δ_i, it holds that

$$\max_{\boldsymbol{x}_i\in\mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}) = u_i(\tilde{BR}_i(\boldsymbol{x}_{-i}), \boldsymbol{x}_{-i}), \forall \boldsymbol{x}_{-i} \in \mathcal{X}_{-i}. \tag{C.24}$$

Hence, we can estimate an upper bounded of $c_{t,i}$ as the following:

$$\begin{aligned}
c_{t,i} &= |\tilde{u}_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - u_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i})| \\
&\leq |u_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - u_i(\tilde{BR}_i(\boldsymbol{x}_{t,-i}), \boldsymbol{x}_{t,-i})| + |\tilde{u}_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - u_i(\tilde{BR}_i(\boldsymbol{x}_{t,-i}), \boldsymbol{x}_{t,-i})| \\
&= |u_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - u_i(\tilde{BR}_i(\boldsymbol{x}_{t,-i}), \boldsymbol{x}_{t,-i})| + |\tilde{u}_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - \max_{\boldsymbol{x}_i\in\mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{t,-i})|.
\end{aligned} \tag{C.25}$$

Since u_i is L_1-Lipschitz in $\boldsymbol{x}_i$, and $\tilde{BR}_i(\cdot) = BR(\cdot) - \boldsymbol{\delta}_i$ is L_2-Lipschitz, the first part of

of RHS of (C.25) can be upper bounded by

$$\begin{aligned}
|u_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - u_i(\tilde{BR}_i(\boldsymbol{x}_{t,-i}), \boldsymbol{x}_{t,-i})| &\leq L_1 \|\boldsymbol{x}_{t,i} - \tilde{BR}_i(\boldsymbol{x}_{t,-i})\| \\
&= L_1 \|\boldsymbol{x}_{t,i} - \tilde{\boldsymbol{x}}_i^* + \tilde{BR}_i(\tilde{\boldsymbol{x}}_{-i}^*) + \tilde{BR}_i(\boldsymbol{x}_{t,-i})\| \\
&\leq L_1 \|\boldsymbol{x}_{t,i} - \tilde{\boldsymbol{x}}_i^*\| + L_1 \|\tilde{BR}_i(\tilde{\boldsymbol{x}}_{-i}^*) - \tilde{BR}_i(\boldsymbol{x}_{t,-i})\| \\
&\leq L_1 \|\boldsymbol{x}_{t,i} - \tilde{\boldsymbol{x}}_i^*\| + L_1 L_2 \|\tilde{\boldsymbol{x}}_{-i}^* - \boldsymbol{x}_{t,-i}\| \\
&\leq C L_1 (1 + L_2) t^{-\gamma/p}. \qquad \text{(C.26)}
\end{aligned}$$

For the second part, we have

$$\begin{aligned}
&|\tilde{u}_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - \max_{\boldsymbol{x}_i \in \mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{t,-i})| \\
\leq &|\tilde{u}_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - \tilde{u}_i(\tilde{\boldsymbol{x}}_i^*, \tilde{\boldsymbol{x}}_{-i}^*)| + |\tilde{u}_i(\tilde{\boldsymbol{x}}_i^*, \tilde{\boldsymbol{x}}_{-i}^*) - \max_{\boldsymbol{x}_i \in \mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{t,-i})|. \qquad \text{(C.27)}
\end{aligned}$$

Because $\tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}) = u_i(\boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{x}_{-i}) - u_i(BR_i(\boldsymbol{x}_{-i}), \boldsymbol{x}_{-i}) + u_i(BR_i(\boldsymbol{x}_{-i}) - \boldsymbol{\delta}_i, \boldsymbol{x}_{-i})$ and u_i, BR_i are L_1, L_2-Lipschitz continuous, $\tilde{u}_i$ is $(L_1 + 2L_1(1 + L2)) = L_1(2L_2 + 3)$-Lipschitz continuous. Therefore, the first part of the RHS of (C.27) can be upper bounded by

$$|\tilde{u}_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - \tilde{u}_i(\tilde{\boldsymbol{x}}_i^*, \tilde{\boldsymbol{x}}_{-i}^*)| \leq C L_1 (2L_2 + 3) t^{-\gamma/p}. \qquad \text{(C.28)}$$

To upper bound the second term of the RHS of (C.27), observe that

$$\begin{aligned}
\max_{\boldsymbol{x}_i \in \mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}) &= \max_{\boldsymbol{x}_i \in \mathcal{X}_i} u_i(\boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{x}_{-i}) - u_i(BR_i(\boldsymbol{x}_{-i}), \boldsymbol{x}_{-i}) + u_i(BR_i(\boldsymbol{x}_{-i}) - \boldsymbol{\delta}_i, \boldsymbol{x}_{-i}) \\
&= u_i(BR_i(\boldsymbol{x}_{-i}), \boldsymbol{x}_{-i}) - u_i(BR_i(\boldsymbol{x}_{-i}), \boldsymbol{x}_{-i}) + u_i(BR_i(\boldsymbol{x}_{-i}) - \boldsymbol{\delta}_i, \boldsymbol{x}_{-i}) \\
&= u_i(BR_i(\boldsymbol{x}_{-i}) - \boldsymbol{\delta}_i, \boldsymbol{x}_{-i}). \qquad \text{(C.29)}
\end{aligned}$$

Using (C.29), we conclude that,

$$\begin{aligned}
&\left|\tilde{u}_i(\tilde{\boldsymbol{x}}_i^*, \tilde{\boldsymbol{x}}_{-i}^*) - \max_{\boldsymbol{x}_i \in \mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{t,-i})\right| \\
=&\left|\max_{\boldsymbol{x}_i \in \mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_{-i}^*) - \max_{\boldsymbol{x}_i \in \mathcal{X}_i} \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{x}_{t,-i})\right| \\
=&\left|u_i(BR_i(\tilde{\boldsymbol{x}}_{-i}^*) - \boldsymbol{\delta}_i, \tilde{\boldsymbol{x}}_{-i}^*) - u_i(BR_i(\boldsymbol{x}_{t,-i}) - \boldsymbol{\delta}_i, \boldsymbol{x}_{t,-i})\right| \\
\leq& L_1(\|BR_i(\tilde{\boldsymbol{x}}_{-i}^*) - BR_i(\boldsymbol{x}_{t,-i})\| + \|\tilde{\boldsymbol{x}}_{-i}^* - \boldsymbol{x}_{t,-i}\|) \\
\leq& CL_1(L_2+1)t^{-\gamma/p}. \qquad \text{(C.30)}
\end{aligned}$$

Assembling (C.26), (C.28), (C.27) and (C.30) together, from (C.25) we conclude that with probability at least $1 - t^{-p\alpha+\gamma}$, it holds that

$$\begin{aligned}
c_{t,i} &\leq CL_1(1+L_2)t^{-\gamma/p} + CL_1(2L_2+3)t^{-\gamma/p} + CL_1(L_2+1)t^{-\gamma/p} \\
&= CL_1(4L_2+5)t^{-\gamma/p}. \qquad \text{(C.31)}
\end{aligned}$$

On the other hand, in the event (with a probability at most $t^{-p\alpha+\gamma}$) that Eq. (C.23) do not hold, we can use the fact that $u_i \in [0, M]$ to give a trivial upper bound of $c_{t,i}$ as follows:

$$\begin{aligned}
c_{t,i} &= |\tilde{u}_1(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i}) - u_1(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i})| \\
&= |u_i(\boldsymbol{x}_{t,i} + \boldsymbol{\delta}_i, \boldsymbol{x}_{t,-i}) - u_i(BR_i(\boldsymbol{x}_{t,-i}), \boldsymbol{x}_{t,-i}) + u_i(BR_i(\boldsymbol{x}_{t,-i}) - \boldsymbol{\delta}_i, \boldsymbol{x}_{t,-i}) - u_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i})| \\
&\leq |u_i(\boldsymbol{x}_{t,i} + \boldsymbol{\delta}_i, \boldsymbol{x}_{t,-i}) - u_i(BR_i(\boldsymbol{x}_{t,-i}), \boldsymbol{x}_{t,-i})| + |u_i(BR_i(\boldsymbol{x}_{t,-i}) - \boldsymbol{\delta}_i, \boldsymbol{x}_{t,-i}) - u_i(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,-i})| \\
&\leq 2M. \qquad \text{(C.32)}
\end{aligned}$$

Putting (C.31) and (C.32) together, the expected corruption on agent-i the adversary

needs at round t can be thus upper bounded by

$$\begin{aligned}\mathbb{E}[c_{t,i}] &\leq (1 - t^{-p\alpha+\gamma}) \cdot CL_1(4L_2 + 5)t^{-\gamma/p} + t^{-p\alpha+\gamma} \cdot 2M \\ &\leq [CL_1(4L_2 + 5) + 2M]t^{-\frac{p\alpha}{p+1}},\end{aligned}$$

and the optimal order is achieved when $\gamma = \frac{p^2}{p+1} \cdot \alpha$. As a result, the total expected corruption is upper bounded by

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} c_{t,i}\right] \leq [CL_1(4L_2 + 5) + 2M]|\mathcal{V}| \cdot T^{1-\frac{p\alpha}{p+1}}. \tag{C.33}$$

□

C.2.3 A General Argument of Attacking Ability and Proof of Proposition 4

We first provide an argument regarding the relationship between $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*$ in general games, and then present the proof of Proposition 4, which rigorously establishes the corresponding results for the case of Cournot competition.

As pointed out in the proof of Theorem 4.3.1, the converged point $\tilde{\boldsymbol{x}}^* = (\tilde{\boldsymbol{x}}_1^*, \cdots, \tilde{\boldsymbol{x}}_n^*)$ under SUSA$(\mathcal{V}, \{\boldsymbol{\delta}_k\}_{k\in[n]}, \{\Delta_k\}_{k\in[n]})$ can be described by the solution to the following system:

$$\begin{cases} \boldsymbol{x}_1 = f_1(\boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n) - \boldsymbol{\delta}_1, \\ \boldsymbol{x}_2 = f_2(\boldsymbol{x}_1, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n) - \boldsymbol{\delta}_2, \\ \vdots \\ \boldsymbol{x}_n = f_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}) - \boldsymbol{\delta}_n, \end{cases} \tag{C.34}$$

where the function $f_i(\boldsymbol{x}_{-i}) = \arg\max_t u_i(t, \boldsymbol{x}_{-i})$ is the best response mapping for agent i.

And $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_n^*)$ is the solution to

$$\begin{cases} \boldsymbol{x}_1 = f_1(\boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n), \\ \boldsymbol{x}_2 = f_2(\boldsymbol{x}_1, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n), \\ \vdots \\ \boldsymbol{x}_n = f_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}). \end{cases} \tag{C.35}$$

Let $F(\boldsymbol{x}) = (f_1(\boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n), f_2(\boldsymbol{x}_1, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n), \cdots, f_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}))$ as an n-valued function and $G(\boldsymbol{x}) = F(\boldsymbol{x}) - \boldsymbol{x}$, we have

$$\begin{aligned} G(\boldsymbol{x}^*) &= \mathbf{0}, \\ G(\tilde{\boldsymbol{x}}^*) &= (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_n). \end{aligned}$$

For a small perturbation $\Delta_{\boldsymbol{x}}$, we can approximate the system G around $\boldsymbol{x}^*$ with a linearized version:

$$G(\boldsymbol{x}^* + \Delta_{\boldsymbol{x}}) = G(\boldsymbol{x}^*) + J(G(\boldsymbol{x}^*))\Delta_{\boldsymbol{x}} + \mathcal{O}(\|\Delta_{\boldsymbol{x}}\|^2). \tag{C.36}$$

If we view the target point to steer $\tilde{\boldsymbol{x}}^*$ as $\boldsymbol{x}^* + \Delta_{\boldsymbol{x}}$, it holds that

$$\boldsymbol{\delta} = G(\tilde{\boldsymbol{x}}^*) = G(\boldsymbol{x}^*) + J(G(\boldsymbol{x}^*))(\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*) + \mathcal{O}(\|\Delta_{\boldsymbol{x}}\|^2) = J(G(\boldsymbol{x}^*))(\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*) + \mathcal{O}(\|(\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*)\|^2). \tag{C.37}$$

From (C.21) we know that $J(G(\boldsymbol{x}^*)) = D^{-1}(\boldsymbol{x}^*)H_{\mathcal{G}}(\boldsymbol{x}^*)$, where $H_{\mathcal{G}}(\boldsymbol{x}^*)$ is the game Hessian at $\boldsymbol{x} = \boldsymbol{x}^*$, and $D^{-1}(\boldsymbol{x}^*)$ is the block diagonal matrix with each block element being the inverse Hessian of u_i w.r.t. $\boldsymbol{x}_i^*$. Therefore, we argue the following in the sense of approximation:

1. if the adversary wants to induce a particular deviation direction $\boldsymbol{v}$, it can simply pick

$\boldsymbol{\delta}$ to be $D^{-1}(\boldsymbol{x}^*)H_{\mathcal{G}}(\boldsymbol{x}^*)\boldsymbol{v}$;

2. if the goal is to induce largest deviation as possible, it can pick δ to align with the direction corresponding to the smallest eigenvalues of $D^{-1}(\boldsymbol{x}^*)H_{\mathcal{G}}(\boldsymbol{x}^*)$.

And for the special case when SUSA only has one victim agent k, it need to pick

$$\boldsymbol{\delta}_k = \mathrm{diag}([\nabla_{ii}^{-1}u_i]_{i=1}^n)H_{\mathcal{G}}[:,k]\boldsymbol{v} \tag{C.38}$$

in order to induce an NE shift $\boldsymbol{v} = \tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*$, and for any $\boldsymbol{\delta}$, it will cause an NE shifting characterized by

$$\boldsymbol{v} = H_{\mathcal{G}}^{-1}[:,k][\nabla_{kk}u_k]\boldsymbol{\delta}, \tag{C.39}$$

which echoes our argument after Theorem 4.3.1. In addition, since β-strongly monotonicity implies that each diagonal element of D is lower bounded by $\beta > 0$, we know the smallest eigenvalue of J can be upper bounded by

$$\rho_{\min}(J) \leq \frac{\rho_{\min}(H_{\mathcal{G}})}{\min_{k\in[n]}\beta_k} = \frac{\beta}{\min_{k\in[n]}\beta_k} \leq 1, \tag{C.40}$$

where $\rho_{\min}(H_{\mathcal{G}}) = \beta$ because $\mathcal{G}$ is β-strongly monotone, and β_k is a constant such that u_k is β_k-strongly concave, meaning $\beta_k \geq \beta$. As a result, we have

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*\| = \rho_{\min}(J(G(\boldsymbol{x}^*)))^{-1}\|\boldsymbol{\delta}\| \geq \|\boldsymbol{\delta}\|. \tag{C.41}$$

We note that such an argument is not rigorous and only holds in the sense of approximation, as we omit the high-order term $\mathcal{O}(\|\Delta_{\boldsymbol{x}}\|^2)$ in (C.36). However, this argument becomes a rigorous proof for Cournot competition, as the Jacobian of system G becomes a constant matrix and the high-order term $\mathcal{O}(\|\Delta_{\boldsymbol{x}}\|^2)$ in (C.36) disappears.

Proof. For Cournot competition, since the Jacobian of G is a constant matrix and we do not

have the high-order term $\mathcal{O}(\|\Delta_{\boldsymbol{x}}\|^2)$ in (C.36). Hence, from (C.36) we have

$$G(\boldsymbol{x}^* + \Delta_{\boldsymbol{x}}) = G(\boldsymbol{x}^*) + D^{-1}H_{\mathcal{G}}\Delta_{\boldsymbol{x}}, \tag{C.42}$$

which means that for any particular $\boldsymbol{\delta}$, it will induce

$$\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^* = H_{\mathcal{G}}^{-1}D\boldsymbol{\delta}. \tag{C.43}$$

As a result, to induce a particular NE deviation direction $\boldsymbol{v}$, we only need to pick $\boldsymbol{\delta} = D^{-1}H_{\mathcal{G}}\boldsymbol{v}$. To induce the largest possible deviation distance $\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*\|$, it need to pick $\boldsymbol{\delta}$ such that $\boldsymbol{\delta}$ aligns with the direction corresponding to the largest eigenvalues of $H_{\mathcal{G}}^{-1}D$, which is also the direction corresponding to the smallest eigenvalues of $D^{-1}H_{\mathcal{G}}$. □

C.3 Technical Details for Section 4.4

In this section, we provide the technical details related to Section 4.4 that we do not have space to include in the main paper. These include essential details for Algorithm 7 and 9, as well as the proofs of Theorem 4.4.1, Proposition 7, and Corollary 2.

C.3.1 Essential Details for Algorithm 7

In this subsection, we outline all the necessary details for Algorithm 7. The algorithm requires selecting a self-concordant barrier $\mathcal{R}_i$ (see Definition C.3.1) and using a specific prox-mapping, $\mathcal{P}_{\mathcal{R}_i}(\boldsymbol{x}_i^{(t)}, \tilde{\boldsymbol{v}}_i^{(t)}, \eta_t, \lambda_i, \beta)$ (see Definition C.3.2), to update the action for the subsequent round.

Algorithm 7: MD-SCB under Utility Poisoning Attack [Ba et al., 2024b]

Input: step-size $\eta_t = \frac{t^{-\phi}}{2d}$, weight $\lambda_i > 0$, game's strongly concave parameter $\beta > 0$, and self-concordant barrier $\mathcal{R}_i : \text{int}(\mathcal{X}_i) \to \mathbb{R}$

Initialize: Let $\boldsymbol{x}_i^{(t)} = \arg\min_{\boldsymbol{x}_i \in \mathcal{X}_i} \mathcal{R}_i(\boldsymbol{x}_i)$ for all $i \in [n]$

for $t \in [T]$ **do**

  for $i \in [n]$ **do**

    Set $A_i^{(t)} \leftarrow (\nabla^2 \mathcal{R}_i(\boldsymbol{x}_i^{(t)}) + \frac{\eta_t \beta (t+1)}{\lambda_i} I_{d_i})^{-\frac{1}{2}}$

    Draw direction $\boldsymbol{z}_i^{(t)}$ uniformly from unit sphere $\mathbb{S}^{d_i}$

    Play action $\hat{\boldsymbol{x}}_i^{(t)} \leftarrow \boldsymbol{x}_i^{(t)} + A_i^{(t)} \boldsymbol{z}_i^{(t)}$

  Receive $\tilde{\mu}_{\tilde{i}}^{(t)} \leftarrow \mu_{\tilde{i}}(\hat{\boldsymbol{x}}_t) + c_{t,\tilde{i}}(\hat{\boldsymbol{x}}_t)$ for $\tilde{i} \in \tilde{\mathcal{I}}_t$, $\tilde{\mu}_i^{(t)} \leftarrow \mu_i(\hat{\boldsymbol{x}}_t)$ for $i \in [n] \setminus \tilde{\mathcal{I}}_t$

  for $i \in [n]$ **do**

    Compute gradient $\tilde{\boldsymbol{v}}_i^{(t)} \leftarrow d_i \tilde{\mu}_i^{(t)} (A_i^{(t)})^{-1} \boldsymbol{z}_i^{(t)}$

    Update action $\boldsymbol{x}_i^{(t+1)} \leftarrow \mathcal{P}_{\mathcal{R}_i}(\boldsymbol{x}_i^{(t)}, \tilde{\boldsymbol{v}}_i^{(t)}, \eta_t, \lambda_i, \beta)$

Definition C.3.1 (Self-concordance)**.** *A function $\mathcal{R} : int(\mathcal{A}) \to \mathbb{R}$ is self-concordant if it satisfies*

1. *$\mathcal{R}$ is three times continuously differentiable, convex, and approaches infinity along any sequence of points approaching the boundary of $int(\mathcal{A})$.*
2. *For every $h \in \mathbb{R}^d$ and $x \in int(\mathcal{A})$, $|\nabla^3 \mathcal{R}(x)[h,h,h]| \leq 2 \left(h^\top \nabla^2 \mathcal{R}(x) h\right)^{\frac{3}{2}}$ holds, where $|\nabla^3 \mathcal{R}(x)[h,h,h] := \frac{\partial^3 \mathcal{R}}{\partial t_1 \partial t_2 \partial t_3}(x + t_1 h + t_2 h + t_3 h)|_{t_1 = t_2 = t_3 = 0}$.*

In addition to these two conditions, if for every $h \in \mathbb{R}^d$ and $x \in int(\mathcal{A})$, $|\nabla \mathcal{R}(x)^\top h \leq \nu^{\frac{1}{2}} (h^\top \nabla^2 \mathcal{R}(x) h)^{\frac{1}{2}}|$ for a positive real number ν, $\mathcal{R}$ is ν-self-concordant.

Definition C.3.2 (Specific Prox-mapping)**.** *The prox-mapping exploited in Algorithm 7 is defined as*

$$\mathcal{P}_{\mathcal{R}}(x, \hat{v}, \lambda, \eta) = \arg\min_{x' \in x} \langle \hat{v}, x - x' \rangle + \frac{\eta \beta (t+1)}{2\lambda} \|x - x'\|_2^2 + D_{\mathcal{R}}(x', x),$$

where $D_{\mathcal{R}}(x', x)$ is the Bregman divergence that measures the difference between the values of a convex $\mathcal{R}$ at two points, x and x', combined with the local linear approximation of $\mathcal{R}$ around x, described by the following equation

$$D_{\mathcal{R}}(x', x) = \mathcal{R}(x') - \mathcal{R}(x) - \langle \nabla \mathcal{R}(x), x' - x \rangle.$$

C.3.2 Proof of Theorem 4.4.1

Theorem (Theorem 4.4.1 restated)**.** *Consider an adversary that attacks a set of agents $\tilde{\mathcal{I}}_t$ at round t with a total budget satisfying $\sum_{j=1}^{t} \sum_{\tilde{i} \in \tilde{\mathcal{I}}_j} |c_{j,\tilde{i}}| \leq \mathcal{O}(t^{\rho})$ for all t and some $\rho \in [0, 1]$. By choosing a learning rate sequence $\eta_t = \frac{1}{2d} t^{-\phi}$, for sufficiently large T the last-iterate convergence rate of Algorithm 7 satisfies $\mathbb{E}\left[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^*\|_2^2\right] \leq \max\left\{\mathcal{O}(dT^{\phi-1}), \mathcal{O}(dT^{-\phi}), \mathcal{O}(dT^{\rho - \frac{1}{2}(\phi+1)})\right\}$. Specifically, $\mathbb{E}\left[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^*\|_2^2\right] \leq \mathcal{O}\left(dT^{\frac{2(\rho-1)}{3}}\right)$ can be achieved by choosing $\phi = \frac{2}{3}(\rho + \frac{1}{2})$.*

Proof. The proof outlined below is motivated by the observation that we can quantify the bias caused by adversarial attack in utility observation in the a Simultaneous Perturbation Stochastic Approximation (SPSA) estimator, which shares a similar spirit to the gradient estimation under corruption lemma in Cheng et al. [2024]. In particular, we can extend Lemma 3.5 in Ba et al. [2024b] to incorporate adversarial attacks as follows.

Lemma C.3.1 (Extended Lemma 3.5 in [Ba et al., 2024b])**.** *Suppose that μ_i is a concave function and $A_i \in \mathbb{R}^{d_i \times d_i}$ is an convertible matrix for each $i \in [n]$, we define the smoothed version of μ_i with respect to A_i by $\hat{\mu}_i(\boldsymbol{x}) = \mathbb{E}_{w_i \sim \mathbb{B}^{d_i}} \mathbb{E}_{\boldsymbol{z}_{-i} \sim \prod_{j \neq i} \mathbb{S}^{d_j}} [\mu_i(\boldsymbol{x}_i + A_i w_i; \hat{\boldsymbol{x}}_{-i})]$ where $\mathbb{S}^{d_i}$ is a d_i dimensional unit sphere, $\mathbb{B}^{d_i}$ is a d_i-dimensional unit ball and $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + A_i \boldsymbol{z}_i$ for all $i \in [n]$. Then, the following statement hold true*

$$\nabla_i \hat{\mu}_i(\boldsymbol{x}) + \boldsymbol{b}_i = \mathbb{E}[d_i \mu_i(\hat{\boldsymbol{x}}_i; \hat{\boldsymbol{x}}_{-i})(A_i)^{-1} z_i | \boldsymbol{x}],$$

where $\boldsymbol{b}_i = d_i \mathbb{E}[c_i(\hat{\boldsymbol{x}}_i; \hat{\boldsymbol{x}}_{-i})(A_i)^{-1} \boldsymbol{z}_i | \boldsymbol{x}]$.

Remark C.3.1. *The key difference between utility shifting attack and random noise is* $\boldsymbol{b}_i \neq 0$. *This is because the corruption* $c_i(\hat{\boldsymbol{x}}_i; \hat{\boldsymbol{x}}_{-i})$ *depends on the randomly sampled vector* $\boldsymbol{z}_i$ *through* $\hat{\boldsymbol{x}}_i$, *which makes*

$$\mathbb{E}[c_i(\hat{\boldsymbol{x}}_i; \hat{\boldsymbol{x}}_{-i})(A_i)^{-1}\boldsymbol{z}_i|\boldsymbol{x}] \neq \mathbb{E}[c_i(\hat{\boldsymbol{x}}_i; \hat{\boldsymbol{x}}_{-i})(A_i)^{-1}|\boldsymbol{x}]\mathbb{E}[\boldsymbol{z}_i|\boldsymbol{x}].$$

Thus, we need to control the cumulative impact of the bias in gradient caused by adversarial attack.

We then carry this bias $\boldsymbol{b}_i$ throughout the analysis, closely following the original proof developed by Ba et al. [2024b]. Finally, we derive a result demonstrating how the decreasing speed of the learning rate affects convergence and its associated robustness to adversarial attacks.

To prepare for the proof, we first extend the Lemma 2.12 and Lemma 3.10 in Ba et al. [2024b] to incorporate adversarial attacks.

Lemma C.3.2 (Extended Lemma 2.12 in [Ba et al., 2024b])**.** *Suppose that the iterate* $\{\boldsymbol{x}^{(t)}\}_{t\geq 1}$ *is generated by Algorithm 8 and let each corrupted utility value* $\tilde{\mu}^t$ *satisfy that* $|\tilde{\mu}^t(\boldsymbol{x})| \leq 1$ *for all* $\boldsymbol{x} \in \mathcal{X}$ *and* $0 < \eta_t \leq \frac{1}{2d}$, *we have*

$$D_{\mathcal{R}}(p, \boldsymbol{x}_{t+1}) + \frac{\eta_t \beta(t+1)}{2}\|x^{(t+1)} - p\|^2 \leq D_{\mathcal{R}}(p, \boldsymbol{x}_t) + \frac{\eta_t \beta(t+1)}{2}\|\boldsymbol{x}_t - p\|^2 + 2\eta_t^2\|A^{(t)}\tilde{\boldsymbol{v}}^t\|^2 + \eta_t\langle \tilde{\boldsymbol{v}}^t, x^{(t)} - p\rangle,$$

where $p \in \mathcal{X}$ *and the sequence* $\{\eta_t\}_{t\geq 1}$ *is assumed to be non-increasing.*

Algorithm 8: Single Agent Mirror Descent Self-Concordant Barrier with Bandit Feedback under Utility Shifting Attack

Input: step size $\eta_t > 0$, module $\beta > 0$, and barrier $\mathcal{R} : \text{int}(\mathcal{X}) \to \mathbb{R}$

Initialize $\boldsymbol{x}^{(1)} = \arg\min_{a \in \mathcal{X}} \mathcal{R}(\boldsymbol{x})$

for $t \in [T]$ **do**

- set $A^{(t)} \leftarrow (\nabla^2 R(\boldsymbol{x}^{(t)}) + \eta_t \beta (t+1) I_d)^{-1/2}$
- draw $\boldsymbol{z}^{(t)} \sim \mathbb{S}^d$
- play $\hat{\boldsymbol{x}}^{(t)} \leftarrow A^{(t)} \boldsymbol{z}^{(t)}$
- receive $\tilde{\mu}^{(t)} \leftarrow \mu^{(t)}(\hat{\boldsymbol{x}}^{(t)}) + c_t(\hat{\boldsymbol{x}}^{(t)})$
- set $\tilde{\boldsymbol{v}}^{(t)} \leftarrow \hat{\boldsymbol{v}}^{(t)} + d c_t(\hat{\boldsymbol{x}}^{(t)})(A^{(t)})^{-1}\boldsymbol{z}^{(t)}$, $\hat{v}^t = n\mu^{(t)}(\hat{\boldsymbol{x}}^{(t)})(A^{(t)})^{-1}\boldsymbol{z}^{(t)}$
- update $\boldsymbol{x}^{(t+1)} \leftarrow \mathcal{P}_{\mathcal{R}}(\boldsymbol{x}^{(t)}, \tilde{\boldsymbol{v}}^{(t)}, \eta_t)$

Remark C.3.2. *Algorithm 8 is the single agent version of Algorithm 7. The proof Lemma C.3.2 is similar to the proof in Appendix A of the paperBa et al. [2024b], with a simple modification for* $\lambda(\boldsymbol{x}^{(t)}, g) = \eta_t \|\tilde{\boldsymbol{v}}^{(t)}\|_{\boldsymbol{x}^{(t)},*} = d\eta_t |\tilde{\mu}^{(t)}(\hat{\boldsymbol{x}}^{(t)})| \|\boldsymbol{z}^{(t)}\|_2 \leq d\eta_t \leq \frac{1}{2}$, *which still holds.*

Lemma C.3.3 (Extended Lemma 3.10 in Ba et al. [2024b])**.** *Suppose that the iterate* $\{\boldsymbol{x}^{(t)}\}_{t \geq 1}$ *is generated by Algorithm 7 and let each corrupted utility value* $\tilde{\mu}_i^t$ *satisfy that* $|\tilde{\mu}_i^{(t)}(\boldsymbol{x})| \leq 1$ *for all* $\boldsymbol{x} \in \mathcal{X}$ *and* $0 < \eta_t \leq \frac{1}{2d}$, *we have*

$$\sum_{i \in [n]} \lambda_i D_{\mathcal{R}}(p, \boldsymbol{x}_i^{(t+1)}) + \frac{\eta_t \beta (t+1)}{2} \left(\sum_{i \in \mathcal{N}} \|\boldsymbol{x}_i^{(t+1)} - p_i\|^2 \right) \leq \sum_{i \in [n]} \lambda_i D_{\mathcal{R}}(p_i, \boldsymbol{x}_i^{(t)})$$
$$+ \frac{\eta_t \beta (t+1)}{2} \left(\sum_{i \in [n]} \|\boldsymbol{x}_i^{(t)} - p_i\|^2 \right) + 2\eta_t^2 \left(\sum_{i \in [n]} \lambda_i \|A_i^{(t)} \tilde{\boldsymbol{v}}_i^{(t)}\|^2 \right) + \eta_t \left(\sum_{i \in [n]} \lambda_i \langle \tilde{\boldsymbol{v}}_i^{(t)}, \boldsymbol{x}_i^{(t)} - p_i \rangle \right),$$

where $p_i \in \mathcal{X}_i$ *and the sequence* $\{\eta_t\}_{t \geq 1}$ *is assumed to be non-increasing.*

Now, we will prove Theorem 4.4.1. Taking the expectation of both sides for equation

proved in Lemma C.3.3 conditioned on $\boldsymbol{x}^{(t)}$, we have

$$\mathbb{E}\left[\sum_{i\in[n]}\lambda_i D_{\mathcal{R}}(p,\boldsymbol{x}_i^{(t+1)})|\boldsymbol{x}_t\right]+\frac{\eta_t\beta(t+1)}{2}\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t+1)}-p_i\|^2|\boldsymbol{x}_t\right]\leq$$
$$\sum_{i\in[n]}\lambda_i D_{\mathcal{R}}(p_i,\boldsymbol{x}_i^{(t)})+\frac{\eta_t\beta(t+1)}{2}\left(\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t)}-p_i\|^2\right)$$
$$+2\eta_t^2\mathbb{E}\left(\sum_{i\in[n]}\lambda_i\|A_i^{(t)}\tilde{\boldsymbol{v}}_i^{(t)}\|^2|\boldsymbol{x}_t\right)+\eta_t\mathbb{E}\left(\sum_{i\in[n]}\lambda_i\langle\tilde{\boldsymbol{v}}_i^{(t)},\boldsymbol{x}_i^{(t)}-p_i\rangle|\boldsymbol{x}_t\right).$$

By the definition of $\tilde{\boldsymbol{v}}_i^{(t)}$ and using Lemma 3.5 in Ba et al. [2024b], we have

$$\mathbb{E}(\|A_i^{(t)}\hat{\boldsymbol{v}}_i^{(t)}\|^2|\boldsymbol{x}_t)\leq d_i^2.$$

In addition, using Lemma 3.5 in Ba et al. [2024b] again and the Young's inequality, we have

$$\mathbb{E}\left(\langle\hat{\boldsymbol{v}}_i^{(t)},\boldsymbol{x}_i^{(t)}-p_i\rangle|\boldsymbol{x}_t\right)\leq\langle\nabla_i\mu_i(\boldsymbol{x}_t),\boldsymbol{x}_i^{(t)}-p_i\rangle+\frac{\lambda_i l_i^2}{2\beta}\left(\sum_{i\in[n]}(\sigma_{\max}(A_i^{(t)}))^2\right)$$
$$+\frac{\beta}{2\lambda_i}\|\boldsymbol{x}_i^{(t)}-p_i\|^2+\boldsymbol{b}_i^{(t)\top}(\boldsymbol{x}_i^{(t)}-p_i),$$

where we have $\boldsymbol{b}_i^{(t)}$ equals the following $\boldsymbol{b}_i^{(t)}=n_i\mathbb{E}\left[c_{t,i}(\hat{\boldsymbol{x}}^{(t)})(A_i^{(t)})^{-1}\boldsymbol{z}_i^{(t)}|\boldsymbol{x}_t\right]$. $\boldsymbol{b}_i^{(t)}\neq 0$, this

is because $c_{t,i}(\hat{\boldsymbol{x}})$ is a function of $\boldsymbol{z}_i^{(t)}$. Since $\nabla^2\mathcal{R}_i(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}\in\mathcal{X}$ and

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[n]}\lambda_i\langle\tilde{\boldsymbol{v}}_i^{(t)},\boldsymbol{x}_i^{(t)}-p_i\rangle|\boldsymbol{x}_t\right]\leq\sum_{i\in[n]}\lambda_i\langle\nabla_i\mu_i(\boldsymbol{x}^{(t)}),\boldsymbol{x}_i^{(t)}-p_i\rangle+\frac{1}{2\eta_t\beta^2(t+1)}\left(\sum_{i\in[n]}\lambda_i\right)\left(\sum_{i\in[n]}\lambda_i^2l_i^2\right)\\
&+\frac{\beta}{2}\left(\|\boldsymbol{x}_i^{(t)}-p_i\|^2\right)+\sum_{i\in[n]}\lambda_i\boldsymbol{b}_i^{(t)\top}(\boldsymbol{x}_i^{(t)}-p_i)\\
&\leq\frac{1}{2\eta_t\beta^2(t+1)}\left(\sum_{i\in[n]}\lambda_i\right)\left(\sum_{i\in[n]}\lambda_i^2l_i^2\right)-\frac{\beta}{2}\left(\|\boldsymbol{x}_i^{(t)}-p_i\|^2\right)+\sum_{i\in[n]}\lambda_i\boldsymbol{b}_i^{(t)\top}(\boldsymbol{x}_i^{(t)}-p_i)\\
&+\sum_{i\in[n]}\lambda_i\langle\nabla_i\mu_i(p_i),\boldsymbol{x}_i^{(t)}-p_i\rangle.
\end{aligned}$$

In this way, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[n]}\lambda_iD_{\mathcal{R}_i}(p_i,\boldsymbol{x}_i^{(t+1)})\right]+\frac{\eta_{t+1}\beta(t+1)}{2}\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t+1)}-p_i\|^2\right]\leq\eta_t\mathbb{E}\left[\sum_{i\in[n]}\lambda_i\langle\nabla_iu_i(p),\boldsymbol{x}_i^{(t)}-p_i\rangle\right]\\
&+\mathbb{E}\left[\sum_{i\in[n]}\lambda_iD_{\mathcal{R}_i}(p_i,\boldsymbol{x}_i^{(t)})\right]+\frac{\eta_t\beta t}{2}\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t)}-p_i\|^2\right]+2\eta_t^2\left(\sum_{i\in[n]}d_i^2\lambda_i\right)\\
&+\frac{1}{2\beta^2(t+1)}\left(\sum_{i\in[n]}\lambda_i\right)\left(\sum_{i\in[n]}\lambda_i^2\ell_i^2\right)+\eta_t\mathbb{E}\left[\sum_{i\in[n]}\lambda_i\boldsymbol{b}_i^{(t)\top}(\boldsymbol{x}_i^{(t)}-p_i)\right].
\end{aligned}$$

Summing up the above inequality over $t=1,2,\ldots,T-1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[n]}\lambda_iD_{\mathcal{R}_i}(p_i,\boldsymbol{x}_i^{(t)})\right]+\frac{\eta_T\beta T}{2}\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t)}-p_i\|^2\right]\leq\sum_{t=1}^{T-1}\eta_t\mathbb{E}\left[\sum_{i\in[n]}\lambda_i\langle\nabla_iu_i(p),\boldsymbol{x}_i^{(t)}-p_i\rangle\right]\\
&+\sum_{i\in[n]}\lambda_iD_{\mathcal{R}_i}(p_i,\boldsymbol{x}_i^{(1)})+\frac{\eta_1\beta}{2}\left(\sum_{i\in N}\|\boldsymbol{x}_i^{(1)}-p_i\|^2\right)+2\left(\sum_{i\in[n]}d_i^2\lambda_i\right)\left(\sum_{t=1}^{T-1}\eta_t^2\right)\\
&+\frac{1}{2\beta^2}\left(\sum_{i\in[n]}\lambda_i\right)\left(\sum_{i\in[n]}\lambda_i^2\ell_i^2\right)\left(\sum_{t=1}^{T-1}\frac{1}{t+1}\right)+\sum_{t=1}^{T-1}\eta_t\mathbb{E}\left[\sum_{i\in[n]}\lambda_i\boldsymbol{b}_i^{(t)\top}(\boldsymbol{x}_i^{(t)}-p_i)\right]
\end{aligned}$$

We should notice that for each round, there is freedom for the adversary to choose who

to attack. To be consistent with the lower bound, in this upper bound, we assume that the adversary consistently choose $i := \tilde{i}$ to attack. Before proceeding afterwards, now let's analyze the cumulative impact of bias on the convergence of action profile. Notice that for $i \neq \tilde{i}, \boldsymbol{b}_i^{(t)} = 0$. Therefore we can use Cauchy Schwartz to bound the following inequality

$$\sum_{t=1}^{T-1} \eta_t \mathbb{E}\left[\sum_{i\in[n]} \lambda_i \boldsymbol{b}_i^{(t)\top}(\boldsymbol{x}_i^{(t)} - p_i)\right] \leq \tilde{C}'\sqrt{\beta}\sum_{t=1}^{T-1}\eta_t^{3/2}\sqrt{t+1}|c_{t,\tilde{i}}(\hat{\boldsymbol{x}}_t)|, \tilde{C}' = \tilde{C}dB\max_{i\in[n]}\sqrt{\lambda_i},$$

where $B = \max_{i\in[n]} B_i, B_i$ is the radius of $\mathcal{X}_i$. In particular, $\tilde{C}$ is the constant such that for all t, $\sigma_{\max}\left(\nabla^2\mathcal{R}_{\tilde{i}}(\boldsymbol{x}_{\tilde{i}}^{(t)}) + \frac{\eta_t\beta(t+1)}{\lambda_{\tilde{i}}}I_{d_{\tilde{i}}}\right) \leq \tilde{C}\frac{\eta_t\beta(t+1)}{\lambda_{\tilde{i}}}$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[n]} \lambda_i D_{\mathcal{R}_i}(p_i, \boldsymbol{x}_i^{(t)})\right] + \frac{\eta_T\beta T}{2}\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t)} - p_i\|^2\right] \leq \sum_{t=1}^{T-1}\eta_t\mathbb{E}\left[\sum_{i\in[n]}\lambda_i\langle\nabla_i u_i(p), \boldsymbol{x}_i^{(t)} - p_i\rangle\right] \\
&+ \sum_{i\in[n]}\lambda_i D_{\mathcal{R}_i}(p_i, \boldsymbol{x}_i^{(1)}) + \frac{\eta_1\beta}{2}\left(\sum_{i\in[n]}\|\boldsymbol{x}_i^{(1)} - p_i\|^2\right) + 2\left(\sum_{i\in[n]} d_i^2\lambda_i\right)\left(\sum_{t=1}^{T-1}\eta_t^2\right) \\
&+ \frac{1}{2\beta^2}\left(\sum_{i\in[n]}\lambda_i\right)\left(\sum_{i\in[n]}\lambda_i^2\ell_i^2\right)\left(\sum_{t=1}^{T-1}\frac{1}{t+1}\right) + \tilde{C}'\sqrt{\beta}\sum_{t=1}^{T-1}\eta_t^{3/2}\sqrt{t+1}|c_{t,\tilde{i}}(\hat{\boldsymbol{x}}_t)|.
\end{aligned}$$

Since $\mathbb{E}\left[\sum_{i\in[n]}\lambda_i D_{\mathcal{R}_i}(p_i, \boldsymbol{x}_i^{(t)})\right] > 0$, we have the following equation

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t)} - p_i\|^2\right] \leq \frac{2}{\eta_T\beta T}\sum_{t=1}^{T-1}\eta_t\mathbb{E}\left[\sum_{i\in[n]}\lambda_i\langle\nabla_i u_i(p), \boldsymbol{x}_i^{(t)} - p_i\rangle\right] \\
&+ \frac{2}{\eta_T\beta T}\sum_{i\in[n]}\lambda_i D_{\mathcal{R}_i}(p_i, \boldsymbol{x}_i^{(1)}) + \frac{\eta_1}{\eta_T T}\left(\sum_{i\in[n]}\|\boldsymbol{x}_i^{(1)} - p_i\|^2\right) + \frac{4}{\eta_T\beta T}\left(\sum_{i\in[n]} d_i^2\lambda_i\right)\left(\sum_{t=1}^{T-1}\eta_t^2\right) \\
&+ \frac{2}{\eta_T\beta^3 T}\left(\sum_{i\in[n]}\lambda_i\right)\left(\sum_{i\in[n]}\lambda_i^2\ell_i^2\right)\left(\sum_{t=1}^{T-1}\frac{1}{t+1}\right) + \frac{2}{\eta_T\sqrt{\beta}T}\tilde{C}'\sum_{t=1}^{T-1}\eta_t^{3/2}\sqrt{t+1}|c_{t,\tilde{i}}(\hat{\boldsymbol{x}}_t)|.
\end{aligned} \tag{C.44}$$

By the initialization $\boldsymbol{x}_i^{(1)} = \arg\min_{x\in\mathcal{X}} \mathcal{R}_i(x)$, we have $\nabla\mathcal{R}_i(\boldsymbol{x}_i^{(1)}) = 0$ which implies that $D_{\mathcal{R}_i}(p_i, \boldsymbol{x}_i^{(1)}) = \mathcal{R}_i(p_i) - \mathcal{R}(\boldsymbol{x}_i^{(1)})$. Then, let us inspect each coordinate of a unique Nash equilibrium x^* and set p coordinate by coordinate in C.44; indeed, we consider the following two cases:

- A point $\boldsymbol{x}_i^* \in \mathcal{X}_i$ satisfies that $\pi_{x^1}(\boldsymbol{x}_i^*) \leq 1 - \frac{1}{\sqrt{T}}$. By Lemma 2.4 [Ba et al., 2024b], we have $D_{\mathcal{R}_i}(\boldsymbol{x}_i^*, \boldsymbol{x}_i^{(1)}) = \mathcal{R}_i(\boldsymbol{x}_i^*) - \mathcal{R}_i(\boldsymbol{x}_i^{(1)}) \leq \nu_i \log(T)$. In this case, we set $p_i = \boldsymbol{x}_i^*$.
- A point $\boldsymbol{x}_i^* \in \mathcal{X}_i$ satisfies that $\pi_{x^1}(\boldsymbol{x}_i^*) > 1 - \frac{1}{\sqrt{T}}$. Thus, we can find $\bar{\boldsymbol{x}}_i \in \mathcal{X}_i$ such that $\|\bar{\boldsymbol{x}}_i - \boldsymbol{x}_i^*\| = O(\frac{1}{\sqrt{T}})$ and $\pi_{x^1}(\bar{\boldsymbol{x}}_i) \leq 1 - \frac{1}{\sqrt{T}}$. By Lemma 2.4 [Ba et al., 2024b], we have $D_{\mathcal{R}_i}(\bar{\boldsymbol{x}}_i, \boldsymbol{x}_i^{(1)}) = \mathcal{R}_i(\bar{\boldsymbol{x}}_i) - \mathcal{R}_i(\boldsymbol{x}_i^{(1)}) \leq \nu_i \log(T)$. In this case, we set $p_i = \bar{\boldsymbol{x}}_i$.

Therefore, we have the following equation. By choosing $\eta_t = \frac{1}{2d}t^{-\phi}$, we have

$$\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}_i^*\|^2\right] \le \mathbb{E}\left[\sum_{i\in[n]\setminus\mathcal{I}}\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}_i^*\|\|\boldsymbol{x}_i^*-\bar{\boldsymbol{x}}_i\|\right] \tag{C.45}$$

$$+\frac{4d}{\beta T^{1-\phi}}\sum_{t=1}^{T-1}\eta_t(\mathbb{E}[\left(\sum_{i\in[n]}\lambda_i\ell_i\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}_i^*\|\right)\left(\sum_{i\in[n]\setminus\mathcal{I}}\|\boldsymbol{x}_i^*-\bar{\boldsymbol{x}}_i\|\right) \tag{C.46}$$

$$+\sum_{i\in[n]}\lambda_i\|\nabla_i u_i(\boldsymbol{x}_{\mathcal{I}}^{(t)},\bar{\boldsymbol{x}}_{\mathcal{N}\setminus\mathcal{I}})\|\|\boldsymbol{x}_i^*-\bar{\boldsymbol{x}}_i\|]) \tag{C.47}$$

$$+\frac{1}{T^{1-\phi}}\left(\sum_{i\in[n]}B_i^2\right) \tag{C.48}$$

$$+\frac{2}{d\beta T^{1-\phi}}\left(\sum_{i\in[n]}d_i^2\lambda_i\right)\sum_{t=1}^{T-1}t^{-2\phi} \tag{C.49}$$

$$+\frac{4d\log T}{\beta T^{1-\phi}}\left(\sum_{i\in[n]}\lambda_i\right)\left(\sum_{i\in[n]}\lambda_i^2\ell_i^2\right) \tag{C.50}$$

$$+\frac{2\tilde{C}'}{d\sqrt{\beta}T^{1-\phi}}\sum_{t=1}^{T-1}t^{\frac{1}{2}-\frac{3}{2}\phi}|c_{t,\tilde{i}}(\hat{\boldsymbol{x}}_t)| \tag{C.51}$$

$$+\frac{2\tilde{C}'}{d\sqrt{\beta}T^{1-\phi}}\sum_{t=1}^{T-1}t^{-\frac{3}{2}\phi}|c_{t,\tilde{i}}(\hat{\boldsymbol{x}}_t)| \tag{C.52}$$

$$+\frac{4d}{\beta T^{1-\phi}}\sum_{i\in[n]}\nu_i\lambda_i\log T \tag{C.53}$$

We will analyze the convergence of $\mathbb{E}\left[\sum_{i\in[n]}\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}_i^*\|^2\right]$ by bounding the convergence of the right-hand size one by one. By Lemma 2.4 [Ba et al., 2024b], we have (C.45) and (C.47) are upper bounded by $O\left(\frac{d}{\sqrt{T}}\right)$. (C.48) is upper bounded by $O(\frac{d}{T^{1-\phi}})$. (C.49) is upper bounded by $O(\frac{d}{T^{\phi}})$. (C.50) and (C.53) are upper bounded by $\tilde{O}(\frac{d}{T^{1-\phi}})$. Now it remains to analyze the convergence rate for (C.51), which we regard as the impact of observation error. Note that (C.53) can never be the dominant term, therefore, we drop it from the analysis. Since $t^{-\phi}-(t+1)^{-\phi} \le t^{-(\phi+1)}$ for integer t and $\phi>0$ and if $\sum_{j=1}^{t}|c_{j,\tilde{i}}(\hat{\boldsymbol{x}}_j)| \le t^{\rho}$ holds

for all t, then we can bound

$$\frac{2\tilde{C}'}{d\sqrt{\beta}T^{1-\phi}}\sum_{t=1}^{T-1} t^{\frac{1}{2}-\frac{3}{2}\phi}|c_i^t(\hat{\boldsymbol{x}}^{(t)})| \leq O(T^{\rho-\frac{1}{2}(\phi+1)}).$$

Noticing that

$$\mathbb{E}\left[\sum_{i\in[n]}\|\hat{\boldsymbol{x}}_i^{(t)}-\boldsymbol{x}_i^{(t)}\|^2\right] \leq \mathbb{E}\left[\sum_{i\in[n]}(\sigma_{\max}(A_i^{(t)}))^2\right] \leq \frac{2d\sum_{i\in[n]}\lambda_i}{\beta T^{1-\phi}}$$

We have the following

$$\mathbb{E}\left[\sum_{i\in[n]}\|\hat{\boldsymbol{x}}_i^{(t)}-\boldsymbol{x}_i^{*}\|^2\right] \leq \max\left\{O(dT^{\phi-1}), O(dT^{-\phi}), O\left(T^{\rho-\frac{1}{2}(\phi+1)}\right)\right\},$$

which completes the proof. □

C.3.3 Essential Details for MAMD [Bravo et al., 2018]

At the beginning of this section, we provide necessary details for Muti-Agent Mirror Descent (MAMD), initially developed by Bravo et al. [2018] for multi-agent learning (see Algorithm 9). We attempt to evaluate the robustness of Algorithm 9 under utility shifting attack, highlighted by the red line (Line 7 of Algorithm 9).

Algorithm 9: MAMD [Bravo et al., 2018] under Utility Shifting Attack

Input: step-size $\eta_{t,1} = \gamma t^{-3\phi}$, query radius $\eta_{t,2} = \delta t^{-\phi}$, safety ball $\mathbb{B}_{r_i}(p_i) \subset \mathcal{X}_i$ for all agents

Initialize action $\boldsymbol{x}_i^{(0)} \in \mathcal{X}_i$ for each agent $i \in [n]$

for $t \in [T]$ **do**

  for $i \in [n]$ **do**

    Draw direction $\boldsymbol{z}_i^{(t)}$ uniformly from $\mathbb{S}^{d_i}$

    Set $\boldsymbol{w}_i^{(t)} \leftarrow \boldsymbol{z}_i^{(t)} - r_i^{-1}(\boldsymbol{x}_i^{(t)} - p_i)$

    play $\hat{\boldsymbol{x}}_i^{(t)} \leftarrow \boldsymbol{x}_i^{(t)} + \eta_{t,2}\boldsymbol{w}_i^{(t)}$

  Receive $\tilde{\mu}_{\tilde{i}}^{(t)} \leftarrow \mu_{\tilde{i}}(\hat{\boldsymbol{x}}_t) + c_{t,\tilde{i}}(\hat{\boldsymbol{x}}_t)$ for $i \in \tilde{\mathcal{I}}_t$, $\tilde{\mu}_i^{(t)} \leftarrow \mu_i(\hat{\boldsymbol{x}}_t)$ for $i \in [n] \setminus \tilde{\mathcal{I}}_t$

  for $i \in [n]$ **do**

    Set $\hat{\boldsymbol{v}}_i^{(t)} \leftarrow (d_i/\eta_{t,2})\tilde{\mu}_i^{(t)}\boldsymbol{z}_i^{(t)}$

    Update $\boldsymbol{x}_i^{(t+1)} \leftarrow \mathcal{P}_{\boldsymbol{x}_i^{(t)}}\left(\eta_{t,1}\hat{\boldsymbol{v}}_i^{(t)}\right)$ [a]

a. The proximal mapping $\mathcal{P}_X(y) := \arg_{x' \in \mathcal{X}} \min\{y^\top(x - x') + D(x', x)\}$, where $D_\mathcal{R}(x', x)$ is the Bregman divergence, where $D_\mathcal{R}(x', x) = \mathcal{R}(x') - \mathcal{R}(x) - \nabla\mathcal{R}(x)^\top(x' - x)$, for some strongly convex function $\mathcal{R}$. Additionally, Algorithm 9 creates a safety ball $\mathbb{B}_{r_i}(p_i)$, with $\delta/r_i < 1$, and generate $\hat{\boldsymbol{x}}_i^{(t)}$ bases on the adjusted randomly sampled direction $\boldsymbol{z}_i^{(t)}$ to ensure the feasibility.

C.3.4 Proof of Proposition 7

Proposition 7. *Consider an adversary that attacks a set of agents $\tilde{\mathcal{I}}_t$ at round t with a total budget satisfying $\sum_{j=1}^t \sum_{\tilde{i} \in \tilde{\mathcal{I}}_j} |c_{j,\tilde{i}}| \leq \mathcal{O}(t^\rho)$ for all t and some $\rho \in [0, 1]$. By choosing parameters $\eta_{t,1} := \gamma t^{-3\phi}$ and $\eta_{t,2} := \delta t^{-\phi}$, for constants $\gamma > \frac{1}{3\beta}$ and $\delta > 0$, $\frac{1}{4} < \phi \leq \frac{1}{3}$, running Algorithm 9 a sufficiently large T, the last-iterate convergence rate satisfies:* $\mathbb{E}\left(\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2\right) \leq \mathcal{O}(d^2 \max\{T^{-\phi}, T^{\rho-2\phi}\})$.

Proof. Similar to the proof structure in Appendix C.3.2, at core of the proof of Proposition 7 is to quantify the bias, $\boldsymbol{b}_i$, caused by utility shifting attack in gradient estimation for a SPSA estimator developed by Bravo et al. [2018], resulting the following lemma.

Lemma C.3.4 (Extended Lemma C.1 in [Bravo et al., 2018]). *Consider* $\hat{\boldsymbol{v}}_i^{(t)}$ *(defined in Line 9 of Algorithm 9) and Let* $\hat{\mu}_i(\boldsymbol{x}) := \mathbb{E}_{w_i \sim \mathbb{B}^{d_i}} \mathbb{E}_{\boldsymbol{z}_{-i} \sim \prod_{j \neq i} \mathbb{S}^{d_j}} [\mu_i(\boldsymbol{x}_i + \eta_{t,2} w_i; \hat{\boldsymbol{x}}_{-i})]$*, then we have*

$$\mathbb{E}(\hat{\boldsymbol{v}}_i | \boldsymbol{x}) := \nabla_i \hat{\mu}_i(\boldsymbol{x}) + \boldsymbol{b}_i,$$

where $\boldsymbol{b}_i := \frac{d_i}{\eta_{t,2}} \mathbb{E}\left(c_i(\hat{\boldsymbol{x}}) \boldsymbol{z}_i | \boldsymbol{x}\right)$.

Using Lemma C.3.4, by choosing $\eta_{t,1} = \gamma t^{-p}, \eta_{t,2} = \delta t^{-q}$, we can extend Equation D.13 (see (C.54)) in Bravo et al. [2018] to incorporate bias (see term C in (C.55)), where we use $\bar{D}_t$ to denote the average Bergman divergence $\mathbb{E}(D_t)$ between $\boldsymbol{x}^*$ and $\boldsymbol{x}_t$ induced by some K-strongly convex function $\mathcal{R}$, and $V^2 = \sum_i d_i^2$, we have

$$\bar{D}_{t+1} \leq (1 - \frac{\beta\gamma}{t^p}) \bar{D}_t + \frac{\gamma}{t^p} \left(\frac{\delta}{t^q}\right) + \frac{V^2}{2K} \frac{\gamma^2 \delta^2}{t^{2(p-q)}}, \tag{C.54}$$

$$\bar{D}_{t+1} \leq \underbrace{(1 - \frac{\beta\gamma}{t^p}) \bar{D}_t}_{A} + \underbrace{\frac{\gamma}{t^p} \left(\frac{\delta}{t^q}\right)}_{B} + \underbrace{\eta_{t,1} \left(\sum_{i \in [n]} \boldsymbol{b}_i^{(t)}\right)}_{C} + \underbrace{\frac{V^2}{2K} \frac{\gamma^2 \delta^2}{t^{2(p-q)}}}_{D}. \tag{C.55}$$

To analyze the convergence rate of $\bar{D}_{t+1}$, we need further extend the convergence lemma (Lemma D.2 in Bravo et al. [2018]) from smooth decaying $b_t \sim \mathcal{O}(\frac{1}{t^q})$ to only a sublinear summation bound (see (C.56)) to incorporate utility shifting attack.

Lemma C.3.5 (Extended Lemma D.2 in [Bravo et al., 2018]). *Let* a_n *be a non-negative sequence satisfying that*

$$a_{t+1} \leq a_t \left(1 - \frac{P}{t^p}\right) + \frac{Q}{t^p} \cdot b_t,$$

$$\sum_{t=1}^{T} b_t < T^{\alpha}, \forall T > 0. \tag{C.56}$$

where $P, Q > 0, p \in (0, 1], \alpha \in [\frac{1}{2}, 1)$, $b_t \in [0, 1]$. *Then, there exists a constant* $C > 0$ *such that for sufficiently large* t, *we have*

1. *there exists positive constants* C, T_0 *such that for any* $t > T_0$,

$$a_t \leq \frac{C}{t^{p-\alpha}},$$

2. *if the sequence* $\{b_t\}_{t=0}^{\infty}$ *is non-increasing starting from* $t = t_0$, *there exists positive constants* C, T_0 *such that for any* $t > T_0$,

$$a_t \leq \frac{C}{t^{1-\alpha}}.$$

Proof. For the first situation, it holds that

$$\begin{aligned}
a_{t+1} &\leq a_t \left(1 - \frac{P}{t^p}\right) + \frac{Q}{t^p} \cdot b_t \\
&\leq \left(a_{t-1}\left(1 - \frac{P}{(t-1)^p}\right) + \frac{Q}{(t-1)^p} \cdot b_{t-1}\right)\left(1 - \frac{P}{t^p}\right) + \frac{Q}{t^p} \cdot b_t \\
&\leq \cdots \\
&= Q \cdot \sum_{k=0}^{t-1} \left[\frac{b_{t-k}}{(t-k)^p} \prod_{i=1}^{k} \left(1 - \frac{P}{(t-i+1)^p}\right)\right] \qquad \text{(C.57)} \\
&\leq Q \cdot \sum_{k=0}^{t^{\alpha}} \left[\frac{1}{(t-k)^p} \prod_{i=1}^{k} \left(1 - \frac{P}{(t-i+1)^p}\right)\right] \qquad \text{(C.58)} \\
&\leq Q \cdot \sum_{k=0}^{t^{\alpha}} \left[\frac{1}{(t-k)^p}\right] \\
&\leq Q t^{\alpha - p},
\end{aligned}$$

where (C.58) holds because the sequence $\left\{\frac{1}{(t-k)^p}\prod_{i=1}^{k}\left(1-\frac{P}{(t-i+1)^p}\right)\right\}_{k=0}^{\infty}$ is non-increasing when $p \in (0,1]$. Therefore, the maximum possible value of the RHS of (C.57) is achieved when $b_t = \cdots = b_{t-t^\alpha+1} = 1$. When $\{b_t\}$ is non-increasing, the RHS of (C.57) achieves its maximum possible value when $b_1 = \cdots = b_t = t^{\alpha-1}$. Therefore, it holds that

$$\begin{aligned} a_{t+1} &\leq Q \cdot \sum_{k=0}^{t-1}\left[\frac{b_{t-k}}{(t-k)^p}\prod_{i=1}^{k}\left(1-\frac{P}{(t-i+1)^p}\right)\right] \\ &\leq t^{\alpha-1}Q \cdot \sum_{k=0}^{t-1}\left[\frac{1}{(t-k)^p}\prod_{i=1}^{k}\left(1-\frac{P}{(t-i+1)^p}\right)\right] \\ &\leq O(t^{\alpha-1}). \end{aligned}$$

□

Armed with Lemma C.3.5, we can analyze the convergence of (C.55). Firstly, we note that $\eta_{t,1}\left(\sum_{i\in[n]}\boldsymbol{b}_i^{(t)}\right) \leq \mathcal{O}(dT^{\rho-q})$. Second, we need to choose $p, q > 0$ such that the cumulative sum with respect to the term A, B, C, D converge to a constant. To satisfy this constraint, we need $\frac{1}{4} < q \leq \frac{1}{3}$ and we choose q to make the order of term B and D equal, we get $p = 3q$. In this way, the convergence rate is determined by $\mathcal{O}(d^2\max\{t^{-q}, t^{\rho-2q}\})$, completing the proof. □

C.4 Additional Discussion about Absolute Robustness

In Section 4.4, we explored the efficiency-robustness trade-off based on a specific concept of robustness—whether a utility poisoning adversary can induce NSA. However, this is not the only measure of robustness. In this section, we introduce a new concept called absolute robustness and establish a formal efficiency-robustness trade-off, analogous to our results in Section 4.4, under this broader notion of robustness.

To introduce the notion of absolute robustness, we need to first define a generalized concept of utility poisoning attack outcome, named γ-Successful Attack (γ-SA), as shown in Definition C.4.1. Intuitively, γ-SA depicts a broader and milder goal (i.e., slowing down the convergence rate or causing non-convergence) compared to NSA.

Definition C.4.1 (γ-Successful Attack and Absolute Robustness)**.** *We say an adversary successfully implements a γ-Successful Attack (γ-SA) for some $\gamma \in [0,1]$ to an (α, p)-MAL dynamics running a game if it induces a joint strategy sequence $\{\boldsymbol{x}^{(t)}\}_t^T$ such that for any sufficiently large T, the cumulative error*

$$\sum_{t=1}^{T} \mathbb{E}[\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2^p] = \Omega\left(T^{1-p\alpha(1-\gamma)}\right). \tag{C.59}$$

In addition, we say an MAL dynamics $\mathcal{A}$ is absolutely robust to a utility poisoning adversary $\mathcal{C}$ with budget $\mathcal{O}(T^\rho)$ if $\mathcal{C}$ with budget $\mathcal{O}(T^\rho)$ cannot induce γ-SA for any $\gamma > 0$.

Naturally, NSA is a form of 1-SA, as shifting the NE inevitably results in a linearly growing cumulative error. Therefore, γ-SA is a weaker concept in terms of the attack's impact, while absolute robustness represents a stronger defensive notion, since it implies that not only can a new convergence point not be imposed, but the original convergence rate remains unaffected. Our next result shown in Corollary 3 indicates that if an adversary aims to achieve γ-SA for some $\gamma \in (0,1)$, it requires a smaller total budget, as expected. To accomplish this, the adversary can divide the time horizon into different phases, alternating between performing SUSA and doing nothing. By adjusting the lengths of these phases, the adversary can balance the attack budget and the desired level of regret.

Corollary 3. *Under the same conditions as stated in Theorem 4.3.1, an adversary can implement a γ-SA within a total budget*

$$\mathbb{E}\left[\sum_{t=1}^{T} |c_t|\right] = \mathcal{O}\left(T^{\left(1-\frac{p\alpha}{p+1}\right)(1-p\alpha(1-\gamma))}\right). \tag{C.60}$$

Proof. Define the p-th moment regret $Reg_p(T) = \sum_{t=1}^{T} \mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^p]$. Then from Theorem 4.3.1, we know that for any sufficiently large T_0, the adversary can use an expected total budget of $\mathcal{O}\left(T_0^{1-\frac{p\alpha}{p+1}}\right)$ to incur a regret $\Omega(T_0)$. Consequently, for any given T, setting $T_0 = T^{1-p\alpha(1-\gamma)}$, the adversary can use $\mathcal{O}\left(T^{(1-\frac{p\alpha}{p+1})(1-p\alpha(1-\gamma))}\right)$ total budget in expectation to incur a regret of the order $\Omega(T^{1-p\alpha(1-\gamma)})$ by performing SUSA for the initial T_0 rounds. For the remaining rounds, the adversary can refrain from injecting corruptions and allow the agents to run their MAL algorithm without interference. In this way, the adversary can incur at least $\Omega(T^{1-p\alpha(1-\gamma)})$ regret while using at most $\mathcal{O}\left(T^{(1-\frac{p\alpha}{p+1})(1-p\alpha(1-\gamma))}\right)$ total expected budget. □

Theorem 4.3.1 and Corollary 3 together highlight an intriguing trade-off between the efficiency and robustness of MAL learning dynamics. That is, the faster an MAL algorithm converges (i.e., a larger α), the smaller the total corruption budget required to implement a successful attack (either NSA or γ-SA). Next, we establish an analogous result to Corollary 2 under a new (and stronger) robustness concept, that is, whether an MAL dynamics is absolutely robust to a utility poisoning adversary with certain budget, as defined below:

$$\begin{aligned}\bar{\rho}(\mathcal{C}, \mathcal{A}) &= \inf\{\rho \in [0,1] | \mathcal{C} \text{ with budget } \mathcal{O}(T^\rho) \textit{ can} \text{ achieve } \gamma\text{-SA against } \mathcal{A}.\},\\ &= \sup\{\rho \in [0,1] | \mathcal{A} \text{ is } \textit{absolutely robust} \text{ to } \mathcal{C} \text{ with budget } \mathcal{O}(T^\rho).\},\end{aligned}$$

where $\mathcal{C} \in ADV$ represents a generic utility poisoning adversary. Similarly, we fix $p = 2$ and define the following quantities:

$$\bar{\rho}_-(\alpha) = \inf_{\mathcal{C}\in ADV} \sup_{\mathcal{A}\in MAL(\alpha,2)} \rho(\mathcal{C}, \mathcal{A}), \quad \bar{\rho}_+(\alpha) = \sup_{\mathcal{A}\in MAL(\alpha,2)} \inf_{\mathcal{C}\in ADV} \rho(\mathcal{C}, \mathcal{A}). \tag{C.61}$$

Here, the function $\bar{\rho}_-(\alpha)$ defined in (C.61) represents the optimal strategy of a utility poi-

soning adversary. Specifically, for any MAL algorithm known to enjoy convergence rate α, $\bar{\rho}_-(\alpha)$ denotes the minimum budget level ρ required for a successful γ-SA by the adversary. In contrast, the function $\bar{\rho}_+(\alpha)$ characterizes the optimal defense from the perspective of the algorithm designer. Given a required convergence rate α, $\bar{\rho}_+(\alpha)$ identifies the most robust algorithm that can withstand γ-SA of level ρ from any utility poisoning adversary. From minimax theorem [v. Neumann, 1928, Sion, 1958] we still have $\bar{\rho}_-(\alpha) \geq \bar{\rho}_+(\alpha), \forall \alpha \in [0, \frac{1}{4}]$. Furthermore, Corollary 3 and Theorem 4.4.1 lead to the following result

Corollary 4. *Given the definitions of $\bar{\rho}_-(\alpha)$ and $\bar{\rho}_+(\alpha)$ in (C.61), for some particular MAL algorithm $\mathcal{A}_0$ and adversary $\mathcal{C}_0 \in ADV$, we further define*

$$\bar{\rho}(\mathcal{A}_0(\alpha)) = \inf_{\mathcal{C} \in ADV} \rho(\mathcal{C}, \mathcal{A} = \mathcal{A}_0(\alpha)), \quad \bar{\rho}(\alpha, \mathcal{C}_0) = \sup_{\mathcal{A} \in MAL(\alpha,2)} \rho(\mathcal{C} = \mathcal{C}_0, \mathcal{A}),$$

where $\mathcal{A}_0(\alpha)$ denotes algorithm $\mathcal{A}_0$ with a set of hyper-parameters that guarantees $MAL(\alpha, 2)$ convergence. Then for any $\alpha \in [0, \frac{1}{4}]$, it holds that

$$1 - 3\alpha \leq \bar{\rho}(\textit{MD-SCB}(\alpha)) \leq \bar{\rho}_+(\alpha) \leq \bar{\rho}_-(\alpha) \leq \bar{\rho}(\alpha; \textit{SUSA}) \leq \left(1 - \frac{2\alpha}{3}\right)(1 - 2\alpha). \quad \text{(C.62)}$$

The proof of Corollary 4 follows exactly from the idea of Corollary 2, and thus we put them together in Appendix C.4.1.

Figure C.1 illustrates this trade-off. The yellow dashed line represents $\bar{\rho}_-(\alpha)$; above it, a budget of $\mathcal{O}(T^\rho)$ suffices for a powerful adversary to induce γ-SA in any $MAL(\alpha, 2)$ dynamics. The red region, a subset of this area, corresponds to the actual budget required for γ-SA for the specific attacking strategy outlined in Corollary 3. Conversely, the blue dashed

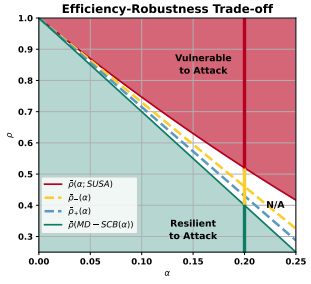

Figure C.1: Illustration of the four quantities in Corollary 4, revealing the intrinsic efficiency-robustness trade-off under γ-SA.

line represents $\bar{\rho}_+(\alpha)$; below it, the most robust $MAL(\alpha, 2)$ dynamics is absolutely robust to any adversary with a total corruption of $\mathcal{O}(T^\rho)$. For a specific MAL dynamics like MD-SCB, the green region shows the budget within which it remains absolutely robust. We only present $\rho \in [0.25, 1]$ because the MAL dynamics with the best achievable convergence rate in such a setting ($alpha = 0.25$)[Ba et al., 2024b] can withstand a corruption level of $\rho = 0.25$ (see Theorem 4.4.1). The three gaps in Figure C.1 correspond to three open problems similar to those we proposed in Figure 4.1.

C.4.1 Proof of Corollary 2 and Corollary 4

Proof. By the definitions of $\rho_-(\alpha), \rho_+(\alpha), \bar{\rho}_-(\alpha), \bar{\rho}_+(\alpha)$ and minimax theorem, it holds that

$$\rho(\text{MD-SCB}(\alpha)) \leq \rho_+(\alpha) \leq \rho_-(\alpha) \leq \rho(\alpha; \text{SUSA}),$$
$$\bar{\rho}(\text{MD-SCB}(\alpha)) \leq \bar{\rho}_+(\alpha) \leq \bar{\rho}_-(\alpha) \leq \bar{\rho}(\alpha; \text{SUSA}).$$

Taking $p = 2$ in Theorem 4.3.1, we have $\rho(\alpha; \text{SUSA}) \leq 1 - \frac{2\alpha}{3}$, and taking $p = 2$ in Corollary 3, we have $\bar{\rho}(\alpha; \text{SUSA}) \leq \left(1 - \frac{2\alpha}{3}\right)(1 - 2\alpha)$.

Next, we show that $\rho(\text{MD-SCB}(\alpha)) \geq 1 - \alpha$ and $\bar{\rho}(\text{MD-SCB}(\alpha)) \geq 1 - 3\alpha$. By Theorem 4.4.1, we know that for MD-SCB, in order to incorporate a corruption $\mathcal{O}(T^\rho)$ with $\rho > 0$, we have to slow down the learning rate by increase ϕ. Therefore, for any particular choice of ϕ, the resultant convergence rate $\alpha = \frac{1-\phi}{2}$. Hence, only when $-\rho + \frac{\phi+1}{2} \geq 1 - \phi$, this choice does not affect the convergence rate, and we obtain $\rho \leq \frac{3}{2}\phi - \frac{1}{2} = 1 - 3\alpha$, meaning $\bar{\rho}(\text{MD-SCB}(\alpha)) \geq 1 - 3\alpha$. On the other hand, if $\rho \geq 1 - 3\alpha$, as long as $\rho - \frac{\phi+1}{2} < 0$ (i.e., $\rho < 1 - \alpha$), NSA is impossible to induce as MD-SCB will still converge although at a slower speed. This implies that $\rho(\text{MD-SCB}(\alpha)) \geq 1 - \alpha$. □

Notably, the efficiency-robustness trade off also exists in Algorithm 9 for γ-SA. (C.55) suggest that as long as $1 - \rho - q > q, \frac{1}{t^{p+q}}$, will be the dominant order in deciding the

convergence $\mathbb{E}(\|\boldsymbol{x}_T - \boldsymbol{x}^*\|_2^2)$. This implies that $\bar{\rho}(\text{MAMD}(\alpha)) \geq 1 - 4\alpha$, which also shows as long as $\rho < 1 - 4\alpha$, the convergence rate of Algorithm 9 will not be affected – robust to γ-SA. However, since q is restricted to $\frac{1}{4} < q \leq \frac{1}{3}$ to ensure the convergence of $\bar{D}_t$, when $\rho > \frac{1}{2}$, the dominant order will be $t^{\rho-2q}$ and we can see optimal choice of $q = \frac{1}{3}$, implying $\rho(\text{MAMD}(\alpha)) \geq \frac{2}{3}$, which we do not observe such a efficiency-robustness trade-off of Algorithm 9 for SUSA, theoretically.

C.5 Additional Experiments

C.5.1 Experiment Result for MAMD

In this section, consider the same experimental setup as of Section 3.5 and evaluate the performance of Algorithm 9. We let each agent run MAMD with the game size $n \in \{10, 50, 100\}$. The learning rate schedule of MAMD is set to $\eta_{t,1} \propto t^{-3\phi}$, and $\eta_{t,2} \propto t^{-\phi}, \phi = 0.33, 0.30, 0.25$, corresponding to convergence rates $\alpha = 0.165, 0.15, 0.125, p = 2$ when no attack is present. We implement SUSA against agent 1, with a fixed $\delta = 10.0$. For each game instance specified by n and the attacked algorithm specified by the convergence rate α, we compare the dynamics' behavior both without attack and under SUSA, and report the actual total budget used.

Result: In the upper row of Figure C.2, the left panel plots the convergence curve of the L_2 square error for $n = 10$, serving as a sanity check to confirm that different learning rate schedules induce different convergence rates for MAMD in absence of an adversary. The middle panel shows the outcome of SUSA against MAMD with $\alpha = 0.165$, with the y-axis representing L_2 distance between $\boldsymbol{x}_t$ and NE. The dashed lines display the convergence curve without the adversary, while the solid lines represent the dynamics under SUSA, with different colors indicating results for various game sizes n. As shown, for different values of n, the attacked dynamics exhibit divergence, as the solid lines saturate and stop decreasing,

indicating that the dynamics are being steered to converge to a new point. However, the trend between induced NE deviation magnitude with respect to n is unclear for MAMD, possibly because the lower bound proved in (4.10) of Theorem 4.3.1 is not tight. The right panel shows the cumulative attack budget $C_t = \sum_{\tau=1}^{t} |c_\tau|$ at each time step t against MAMD with different values of α. As the results indicate, although all exhibit a sublinear trend, a faster-converging dynamics ($\alpha = 0.165$) requires a smaller attack budget, corroborating our findings in Theorem 4.3.1 and 4.4.1. The bottom row of Figure C.2 evaluates the performance of MAMD under a different sets of parameters nevertheless conveys a similar message.

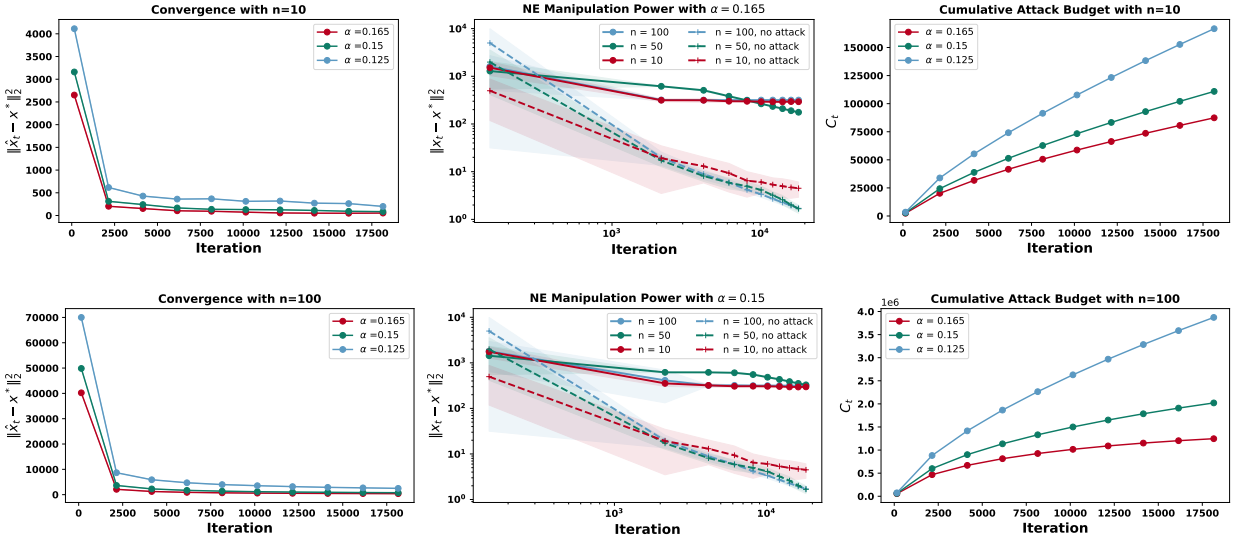

Figure C.2: Left: square error of MAMD with varying convergence rates. Middle: square error of MAMD on different sizes of game instances against SUSA. Right: cumulative attack budget used by SUSA against MAMD with varying convergence rates. Error bars represent the 1-σ region from 20 independent simulations.

C.5.2 Additional Results for MD-SCB

Figure C.3 evaluates the performance of MD-SCB under a different sets of parameters nevertheless conveys a similar message as of Figure 4.2.

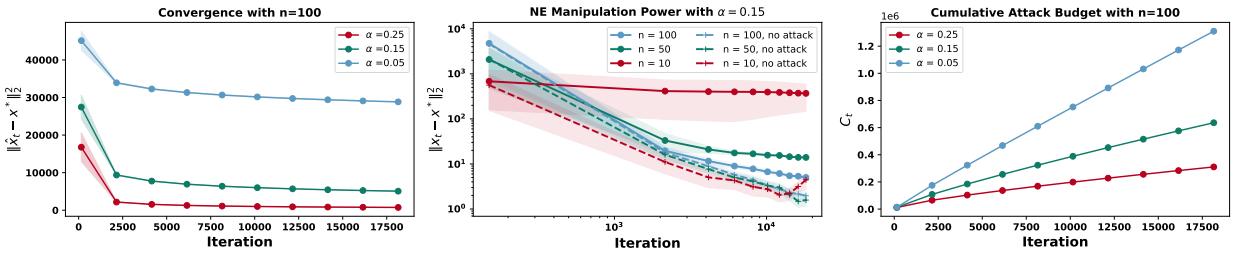

Figure C.3: Left: square error of MD-SCB with varying convergence rates when $n = 100$. Middle: square error of MD-SCB on different sizes of game instances against SUSA with $\alpha = 0.15$. Right: cumulative attack budget used by SUSA against MD-SCB with varying convergence rates when $n = 100$. Error bars represent the 1-σ region from 20 independent simulations.

C.6 Experiments

In this section, we validate our theoretical results from Theorem 4.3.1 and Theorem 4.4.1 by implementing SUSA against MD-SCB on n-person Cournot games, in which each player-i has a utility function $u_i(x_i, x_{-i}) = x_i \left(a - b\sum_{j=1}^{n} x_j\right) - c_i x_i$. The formal definition is given in Appendix C.1. We also direct readers to Appendix C.5 for additional results for attacks against another algorithm MAMD (see Algorithm 9). We run simulations on two types of Cournot game instances. In the first type (homogeneous), all players share the same marginal cost c_i, resulting in a symmetric game. In the second type (heterogeneous), players have varying costs. The results for the homogeneous case are presented in the main paper, while those for the heterogeneous case are included in Appendix C.5.2, as they basically convey the same message.

***Homogeneous* n-person Cournot Game Experimental Setup:** We set $a = 10, b = 0.05$, and we consider the cost $c_i = 1$ for all agents $i \in [n]$. The action space is set to $\mathcal{X}_i = [0, 50]$ and the unique NE of such games can be verified as $\boldsymbol{x}_i^* = \frac{180}{n+1}, i \in [n]$. We let each agent run MD-SCB with the game size $n \in \{10, 50, 100\}$. The learning rate schedule of MD-SCB is set to $(\eta_t \propto t^{-\phi}, \phi = 0.5, 0.7, 0.9)$, corresponding to convergence rates $\alpha = 0.25, 0.15, 0.05, p = 2$ when no attack is present. We implement SUSA against agent 1, with a

fixed $\delta = -10.0$. For each game instance specified by n and the attacked algorithm specified by the convergence rate α, we compare the dynamics' behavior both without attack and under SUSA, and report the actual total budget used.

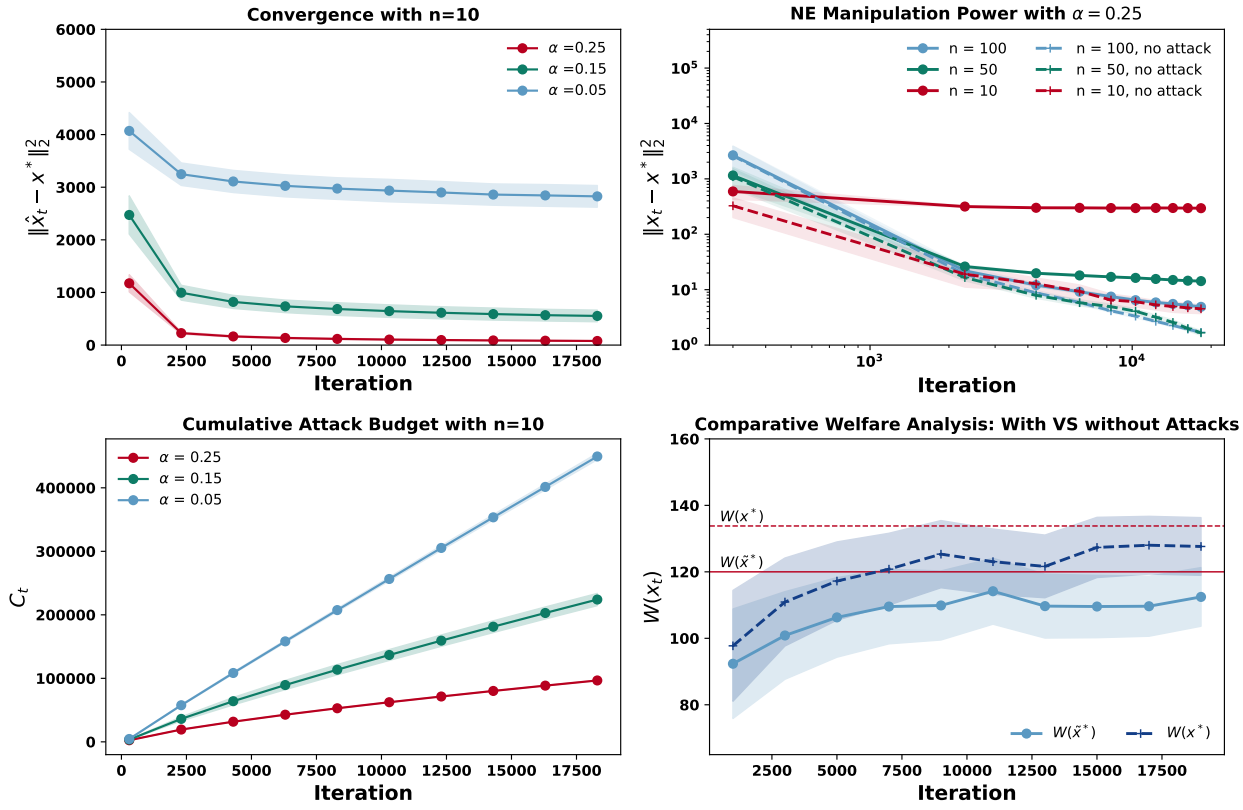

Figure C.4: This figure displays results for *Homogeneous* n-person Cournot Game. Top left: square error of MD-SCB with varying convergence rates. Top right: square error of MD-SCB on different sizes of game instances against SUSA. Bottom left: cumulative attack budget used by SUSA against MD-SCB with varying convergence rates. Bottom right: social welfare $W(\boldsymbol{x}_t)$ under SUSA and without attack for $n = 10$. The red dashed line denote the social welfare under NE without attack, $W(x^*)$, while the red solid line denote the social welfare under manipulated NE under SUSA, $W(\tilde{x}^*)$. We plot 0.1-σ region from 20 independent simulations for social welfare (the bottom right figure) and 0.5-σ region for the rest figures.

Result: The top left panel of Figure C.4 plots the convergence curve of the L_2 square error for $n = 10$, serving as a sanity check to confirm that different learning rate schedules induce different convergence rates for MD-SCB in absence of an adversary. The top right panel shows the outcome of SUSA against MD-SCB with $\alpha = 0.25$, with the y-axis representing

L_2 distance between $\boldsymbol{x}_t$ and NE. The dashed lines display the convergence curve without the adversary, while the solid lines represent the dynamics under SUSA, with different colors indicating results for various game sizes n. As shown, for different values of n, the attacked dynamics exhibit divergence, as the solid lines saturate and stop decreasing, indicating that the dynamics are being steered to converge to a new point. Additionally, we observe that the induced NE deviation decreases with respect to n, which aligns with our theoretical predictions in Theorem 4.3.1 and Remark 4.3.1. The bottom left panel shows the cumulative attack budget $C_t = \sum_{\tau=1}^{t} |c_\tau|$ at each time step t against MD-SCB with different values of α. As the results indicate, although all exhibit a sublinear trend, a faster-converging dynamics ($\alpha = 0.25$) requires a smaller attack budget, corroborating our findings in Theorem 4.3.1 and 4.4.1. The bottom right panel plots the social welfare $W(\boldsymbol{x}_t)$ without attack (indicated by the dashed line) and under SUSA (indicated by the solid line). The panel shows that SUSA with $\delta = -10$ can lead the learning dynamic to a new NE with a lower social welfare, aligning with our theoretical results. These observations also hold for Cournot games with *heterogeneous* costs, and the corresponding results are presented in Appendix C.5.2.

***Heterogeneous* n-person Cournot Game Experimental Setup:** Similar to the *homogeneous* n-person Cournot game, we set $a = 10, b = 0.05, n \in [10, 50, 100], \alpha = [0.25, 0.15, 0.05]$, and $\mathcal{X}_i \in [0, 50], \forall i \in [n]$. However, in the heterogeneous setting, we consider different configurations of c_i based on the value of n. For $n = 10$, the costs are assigned as follows: $c_1 = 0.5$, $c_i = 0.6$ for $i \in [2, 7]$, and $c_i = 0.7$ for $i \in [8, 10]$. For larger values of n, specifically $n = 50$ and $n = 100$, we replicate the structure of the $n = 10$ case, dividing the sequence into 5 groups when $n = 50$ and into 10 groups when $n = 100$. Each group follows the same pattern of costs assignments as in the $n = 10$ case. We implement SUSA against agent 1 with $\delta = -15$.

Result: Figure C.5 suggests that all results hold in the heterogeneous setting just as they do in the homogeneous setting.

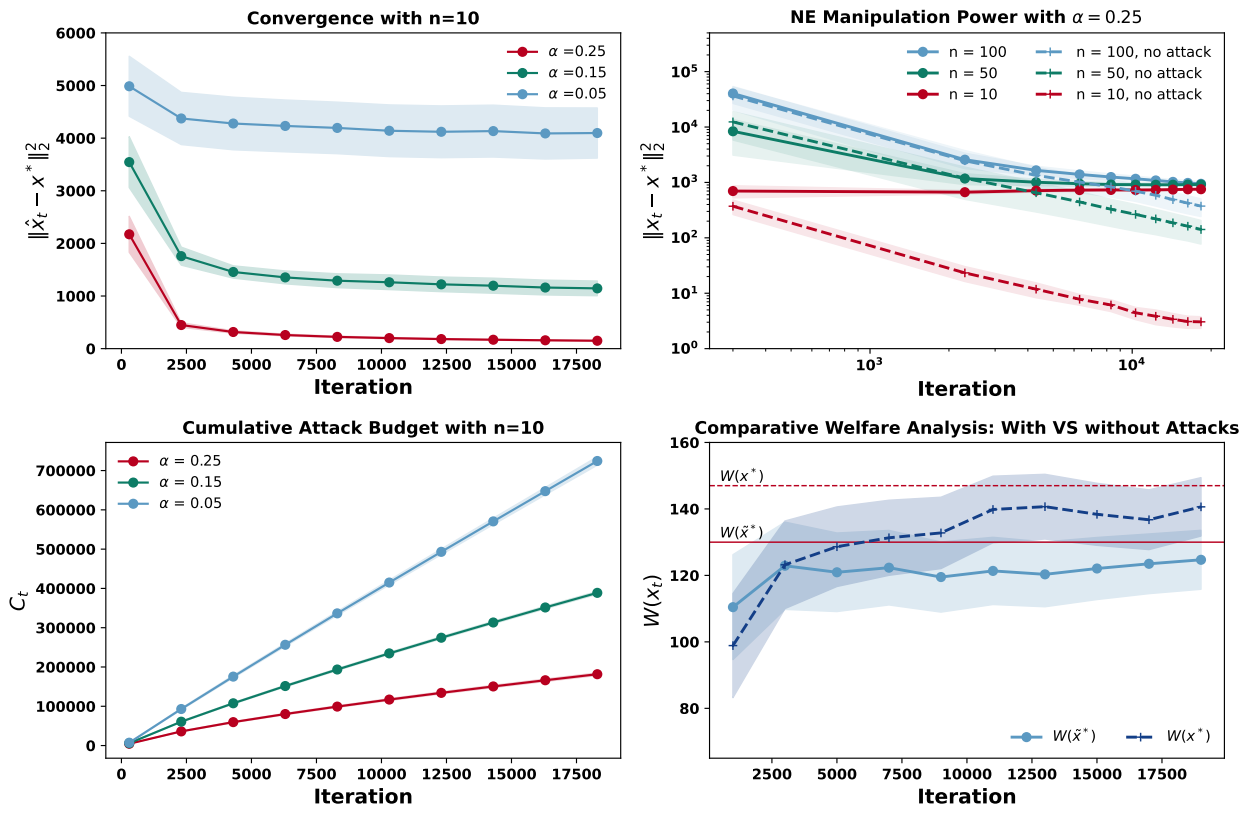

Figure C.5: This figure displays results for *Heterogeneous* n-person Cournot Game.

APPENDIX D

APPENDIX TO INFORMATION CONVEYANCE IN LANGUAGE MODELS

D.1 Missing proofs and algorithms

D.1.1 Equivalence to a two-stage stochastic decision process

Theorem D.1.1 (Equivalence between autoregressive generation and a two-stage decision process)**.** *Fix a prompt $X = x$ and a generation history $G_t = g_t = (y_1, \ldots, y_{t-1})$. Suppose the autoregressive model defines a probability distribution over the extended vocabulary $Z := \mathcal{V} \cup \{\texttt{EOS}\}$, denoted by $q(z \mid g_t, x), z \in Z$. Then this one-step generation rule is equivalent to the following two-stage stochastic decision process: $A_t \in \{\mathsf{STOP}, \mathsf{CONT}\}$, where $\mathbb{P}(A_t = \mathsf{STOP} \mid g_t, x) = q(\texttt{EOS} \mid g_t, x)$, and $\mathbb{P}(A_t = \mathsf{CONT} \mid g_t, x) = 1 - q(\texttt{EOS} \mid g_t, x)$. Conditional on $A_t = \mathsf{CONT}$, the next token $Y_t \in \mathcal{V}$ is sampled from $\mathbb{P}(Y_t = v \mid g_t, x, A_t = \mathsf{CONT}) = \frac{q(v|g_t,x)}{1-q(\texttt{EOS}|g_t,x)}, v \in \mathcal{V}$, whenever $1-q(\texttt{EOS} \mid g_t, x) > 0$. This two-stage process induces the same distribution over generated sequences as the original autoregressive model.*

Proof. Fix a prompt x and history g_t. Under the original autoregressive model, the next outcome is sampled directly from $q(z \mid g_t, x), z \in \mathcal{V} \cup \{\texttt{EOS}\}$. In particular, generation terminates at step t with probability $q(\texttt{EOS} \mid g_t, x)$. Now consider the two-stage process. The probability of termination is $\mathbb{P}(A_t = \mathsf{STOP} \mid g_t, x) = q(\texttt{EOS} \mid g_t, x)$, which matches the probability assigned to EOS by the original model. Next, for any token $v \in \mathcal{V}$, the probability that the two-stage process generates v is

$$\begin{aligned}
\mathbb{P}(Y_t = v \mid g_t, x) &= \mathbb{P}(A_t = \mathsf{CONT} \mid g_t, x)\, \mathbb{P}(Y_t = v \mid g_t, x, A_t = \mathsf{CONT}) \\
&= (1 - q(\texttt{EOS} \mid g_t, x))\, \frac{q(v \mid g_t, x)}{1 - q(\texttt{EOS} \mid g_t, x)} \\
&= q(v \mid g_t, x).
\end{aligned}$$

Thus, the two-stage process assigns exactly the same probability to every possible next outcome:

$$\mathbb{P}(\mathsf{STOP} \mid g_t, x) = q(\texttt{EOS} \mid g_t, x),\ \mathbb{P}(Y_t = v \mid g_t, x) = q(v \mid g_t, x), \quad v \in \mathcal{V}.$$

Therefore, the one-step transition distribution is identical under the two formulations. Since this equality holds for every prompt x, every generation history g_t, and every generation step t, the induced probability of any finite sequence $y_{1:N}$ terminating at length N is the same under both processes. Thus, the two-stage stochastic decision process is equivalent to the original autoregressive generation model. □

D.1.2 Proof of Theorem 5.3.1

Proof. Let $X \sim p_X$ denote the prompt distribution. For each realization $X = x$, consider the sequence of random variables $(Z_t)_{t \geq 1}$ with $Z_t \in \mathcal{Z} := \mathcal{V} \cup \{\texttt{EOS}\}$, where selecting EOS indicates termination. Define the stopping time $N := \min\{t : Z_t = \texttt{EOS}\}$, and let $Y_t = Z_t$ for $t < N$. Adopt the convention that $Z_t = \texttt{EOS}$ for all $t > N$, which induces a one-to-one correspondence between the infinite sequence $(Z_1, Z_2, \ldots)$ and the pair $(N, Y_{1:N})$, conditional on X. Therefore,

$$H(N, Y_{1:N} \mid X) = H(Z_1, Z_2, \ldots \mid X).$$

By the chain rule of entropy, $H(Z_1, Z_2, \ldots \mid X) = \sum_{t=1}^{\infty} H(Z_t \mid Z_{1:t-1}, X)$. Since the history G_t is completely determined by $(Z_{1:t-1}, X)$, we have

$$H(Z_t \mid Z_{1:t-1}, X) = H(Z_t \mid G_t, X) = \mathbb{E}_{g_t, X}\big[\widetilde{H}(g_t \mid X = x)\big].$$

Therefore, $H(Z_1, Z_2, \ldots \mid X) = \sum_{t=1}^{\infty} \mathbb{E}_{g_t, X}\big[\widetilde{H}(g_t \mid X = x)\big]$.

Since $\widetilde{H}(g_t \mid X = x) \geq 0$, we can apply Tonelli's theorem [Folland, 1999] to exchange the

sum and expectation.

$$\sum_{t=1}^{\infty} \mathbb{E}_{g_t,X}\big[\widetilde{H}(g_t \mid X = x)\big] = \mathbb{E}_{g_t,X}\left[\sum_{t=1}^{\infty} \widetilde{H}(g_t \mid X = x)\right].$$

Moreover, for all $t > N$, $Z_t = \texttt{EOS}$ deterministically, so $\widetilde{H}(g_t \mid X = x) = 0$. Therefore, $\sum_{t=1}^{\infty} \widetilde{H}(g_t \mid X = x) = \sum_{t=1}^{N(X)} \widetilde{H}(g_t \mid X = x)$, almost surely. Combining the above, we have $H(Z_1, Z_2, \ldots \mid X) = \mathbb{E}_{g_t,X}\left[\sum_{t=1}^{N(X)} \widetilde{H}(g_t \mid X = x)\right]$. Finally, since $(Z_1, Z_2, \ldots)$ is in one-to-one correspondence with $(N, Y_{1:N})$ given X, we obtain $H(N, Y_{1:N} \mid X) = \mathbb{E}_{g_t,X}\left[\sum_{t=1}^{N(X)} \widetilde{H}(g_t \mid X = x)\right] = \mathbb{E}\left[\sum_{t=1}^{N} \widetilde{H}(G_t \mid X)\right] = \mathrm{CE}^{\star}$. □

D.1.3 Algorithms

Algorithm 10: Estimating $\mathrm{CE}^{\star}_{\max}$ over a prompt distribution

Input: Prompt distribution p_X; number of prompts P; rollout count per prompt M; maximum length $T_{\max}$; sampling policy π (e.g. temperature/top-p/top-k).

for $p = 1$ **to** P **do**
  Sample prompt $X_p \sim p_X$;
  for $i = 1$ **to** M **do**
    Initialize history $G_1^{(p,i)} \leftarrow X_p$; set $S_{\max}^{(p,i)} \leftarrow 0$;
    for $t = 1$ **to** $T_{\max}$ **do**
      Compute model distribution $q_t^{(p,i)}(\cdot \mid G_t^{(p,i)}, X_p)$ under sampling policy π, including `EOS` as an outcome;
      Compute one-step local entropy $\tilde{g}_t^{(p,i)} \leftarrow -\sum_{z \in \mathcal{Z}} q_t^{(p,i)}(z \mid G_t^{(p,i)}, X_p) \log q_t^{(p,i)}(z \mid G_t^{(p,i)}, X_p)$;
      Update $S_{\max}^{(p,i)} \leftarrow S_{\max}^{(p,i)} + \tilde{g}_t^{(p,i)}$;
      Sample next outcome $z_t^{(p,i)} \sim q_t^{(p,i)}(\cdot \mid G_t^{(p,i)}, X_p)$;
      If $z_t^{(p,i)} = \texttt{EOS}$ then set $N_{\max}^{(p,i)} \leftarrow t$ and break;
      Else update history $G_{t+1}^{(p,i)} \leftarrow G_t^{(p,i)} \oplus z_t^{(p,i)}$.
    If no `EOS` is sampled by $T_{\max}$: set $N_{\max}^{(p,i)} \leftarrow T_{\max}$.

return $\widehat{\mathrm{CE}}^{\star}_{\max,P,M} = \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M} S_{\max}^{(p,i)}$, $\widehat{\mathbb{E}}[N_{\max}] = \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M} N_{\max}^{(p,i)}$.

D.1.4 Examples of prompt-averaged bias decay rates

Corollary 5. *By Remark 5.4.1, $g(T_{\max})$ admits the following rates.*

Uniform exponential tail. *If there exist constants $c, \lambda > 0$ such that, for all prompts x and all $t \geq 1$, $\mathbb{P}(N(x) \geq t \mid X = x) \leq ce^{-\lambda t}$, then $g(T_{\max}) \leq c \log|\mathcal{Z}| \sum_{t=T_{\max}+1}^{\infty} e^{-\lambda t} = c \log|\mathcal{Z}| \frac{e^{-\lambda(T_{\max}+1)}}{1-e^{-\lambda}} = O(e^{-\lambda T_{\max}})$.*

Uniform polynomial tail. *If there exist constants $c > 0$ and $\alpha > 1$ such that, for all*

prompts x and all $t \geq 1$, $\mathbb{P}(N(x) \geq t \mid X = x) \leq ct^{-\alpha}$, then $g(T_{\max}) \leq c \log |\mathcal{Z}| \sum_{t=T_{\max}+1}^{\infty} t^{-\alpha} \leq \frac{c \log |\mathcal{Z}|}{\alpha - 1} T_{\max}^{1-\alpha} = O(T_{\max}^{1-\alpha})$.

D.1.5 Proof of Theorem 5.4.1 and MSE analysis

Proof. Denote $\mu_{\max}(x) := \mathbb{E}[S_{\max} \mid X = x]$, and $S_{\max} = \sum_{t=1}^{N_{\max}(x)} \widetilde{H}(g_t \mid X = x)$. Then $\mathrm{CE}^{\star}_{\max} = \mathbb{E}_{X \sim p_X}[\mu_{\max}(X)]$. Since $|\mathcal{Z}| < \infty$, we have $0 \leq \widetilde{H}(g_t \mid X = x) \leq H_{\max} := \log |\mathcal{Z}|$. Therefore $0 \leq S_{\max} \leq H_{\max} N_{\max}(x)$. Therefore, $\mathbb{E}[S_{\max}^2] \leq H_{\max}^2 \mathbb{E}[N_{\max}(X)^2] = H_{\max}^2 \nu_{\max}^2$, where $\nu_{\max}^2 := E[N_{\max}(X)^2]$.

Now decompose the variance of $\widehat{\mathrm{CE}}^{\star}_{\max,P,M}$. Conditional on $X_1, \ldots, X_P$, the M rollouts for each prompt are independent, by law of total variance, we have

$$\mathrm{Var}\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M}\right) = \mathrm{Var}\left(\frac{1}{P}\sum_{p=1}^{P} \mu_{\max}(X_p)\right) + \mathbb{E}\left[\mathrm{Var}\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M} \mid X_1, \ldots, X_P\right)\right].$$

By Jensen's inequality, we have $\mu_{\max}^2(X) = H_{\max}^2 \mathbb{E}(N_{\max}(X))^2 \leq H_{\max}^2 \nu_{\max}^2$, therefore

$$\mathrm{Var}\left(\frac{1}{P}\sum_{p=1}^{P} \mu_{\max}(X_p)\right) = \frac{1}{P}\mathrm{Var}(\mu_{\max}(X)) \leq \frac{1}{P}\mathbb{E}[\mu_{\max}(X)^2] \leq \frac{H_{\max}^2 \nu_{\max}^2}{P}.$$

Because, $\mathbb{E}\left[\mathrm{Var}\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M} \mid X_1, \ldots, X_P\right)\right] \leq \frac{H_{\max}^2 \nu_{\max}^2}{PM}$. Thus we have,

$$\mathrm{Var}\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M}\right) \leq H_{\max}^2 \nu_{\max}^2 \left(\frac{1}{P} + \frac{1}{PM}\right).$$

Since $\mathbb{E}\left[\widehat{\mathrm{CE}}^{\star}_{\max,P,M}\right] = \mathrm{CE}^{\star}_{\max}$, Chebyshev's inequality gives that, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, $\left|\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}_{\max}\right| \leq H_{\max}\nu_{\max}\sqrt{\frac{1}{\delta}\left(\frac{1}{P} + \frac{1}{PM}\right)}$. By the truncation bias assumption, $0 \leq \mathrm{CE}^{\star} - \mathrm{CE}^{\star}_{\max} \leq g(T_{\max})$. Thus,

$$\left|\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}\right| \leq H_{\max}\nu_{\max}\sqrt{\frac{1}{\delta}\left(\frac{1}{P} + \frac{1}{PM}\right)} + g(T_{\max}).$$

Equivalently, $\left|\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}\right| = O_{\mathbb{P}}\left(\nu_{\max}\sqrt{\frac{1}{P}+\frac{1}{PM}} + g(T_{\max})\right)$. Thus, if $T_{\max} = T_{\max}(P,M) \to \infty$, $g(T_{\max}) \to 0$, and if $\nu_{\max}(P,M)\sqrt{\frac{1}{P}+\frac{1}{PM}} \to 0$, then we have $\widehat{\mathrm{CE}}^{\star}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathrm{CE}^{\star}$. □

Theorem D.1.2 (MSE of the prompt-averaged truncated CE estimator)**.**

$$\mathbb{E}\left[\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}\right)^2\right] \le \frac{H^2_{\max}}{P}\operatorname{Var}_X(\mathbb{E}[N_{\max} \mid X]) + \frac{H^2_{\max}}{PM}\mathbb{E}[N^2_{\max}] + B^2_{T_{\max}}.$$

In Theorem D.1.2, we further characterize the mean squared error (MSE) of $\widehat{\mathrm{CE}}^{\star}_{\max,P,M}$ as the sum of the squared truncation bias and the variance of the estimator. The variance decreases with both the number of prompts P and the total number of rollouts PM, highlighting two distinct sources of statistical efficiency: diversity across prompts and repeated sampling within each prompt.

The variance admits a hierarchical decomposition with two distinct sources of randomness: a $1/\sqrt{P}$ term due to variability across prompts, and a $1/\sqrt{PM}$ term arising from rollout-level stochasticity within each prompt. Importantly, the prompt-level variance cannot be reduced by increasing the number of rollouts per prompt, reflecting the intrinsic hierarchical structure of the estimator. Moreover, the Monte Carlo error is scaled by the effective trajectory length (captured by $T_{\max}$ or, more precisely, $N_{\max}$), since longer generations accumulate more uncertainty along the trajectory.

These results reveal a fundamental trade-off. Increasing $T_{\max}$ reduces the truncation bias by capturing more of the generation process, but also increases computational cost and variance through longer trajectories.

Proof. Define the per-prompt rollout average $\bar{S}^{(p)}_{\max} := \frac{1}{M}\sum_{i=1}^{M} S^{(p,i)}_{\max}$. Then $\widehat{\mathrm{CE}}^{\star}_{\max,P,M} = \frac{1}{P}\sum_{p=1}^{P}\bar{S}^{(p)}_{\max}$. Conditional on X_p, $\mathbb{E}[\bar{S}^{(p)}_{\max} \mid X_p] = \mu_{\max}(X_p)$, and, since the M rollouts are conditionally independent, $\operatorname{Var}(\bar{S}^{(p)}_{\max} \mid X_p) = \frac{1}{M}\sigma^2_{\max}(X_p)$. By the law of total variance, we

have

$$\mathrm{Var}(\bar{S}^{(p)}_{\max}) = \mathrm{Var}_X\left(\mathbb{E}[\bar{S}^{(p)}_{\max} \mid X_p]\right) + \mathbb{E}_X\left[\mathrm{Var}(\bar{S}^{(p)}_{\max} \mid X_p)\right].$$

Substituting the conditional mean and variance gives $\mathrm{Var}(\bar{S}^{(p)}_{\max}) = \mathrm{Var}_X(\mu_{\max}(X)) + \frac{1}{M}\mathbb{E}_X[\sigma^2_{\max}(X)]$. Since the prompts $X_1, \ldots, X_P$ are i.i.d., the variables $\bar{S}^{(1)}_{\max}, \ldots, \bar{S}^{(P)}_{\max}$ are i.i.d. Therefore, $\mathrm{Var}\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M}\right) = \frac{1}{P}\,\mathrm{Var}(\bar{S}^{(p)}_{\max}) = \frac{1}{P}\,\mathrm{Var}_X(\mu_{\max}(X)) + \frac{1}{PM}\mathbb{E}_X[\sigma^2_{\max}(X)]$.

For the MSE, write $\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star} = \left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}_{\max}\right) - B_{T_{\max}}$. Taking squares and expectations yields $\mathbb{E}\left[\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}\right)^2\right] = \mathbb{E}\left[\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}_{\max}\right)^2\right] + B^2_{T_{\max}}$, because $\mathbb{E}[\widehat{\mathrm{CE}}^{\star}_{\max,P,M} - \mathrm{CE}^{\star}_{\max}] = 0$.

Since pathwise, $0 \le S_{\max}(x) = \sum_{t=1}^{N_{\max}(x)} \widetilde{H}(g_t \mid x) \le H_{\max} N_{\max}(x)$. Thus, conditional on $X = x$, $\mathrm{Var}(S_{\max}(x) \mid X = x) \le \mathbb{E}[S^2_{\max}(x) \mid X = x] \le H^2_{\max}\mathbb{E}[N^2_{\max}(x) \mid X = x]$. Averaging over prompts gives $\mathbb{E}_X[\sigma^2_{\max}(X)] \le H^2_{\max}\mathbb{E}_X\mathbb{E}[N^2_{\max}(X) \mid X] = H^2_{\max}\mathbb{E}[N^2_{\max}]$.

Similarly, $\mu_{\max}(X) = \mathbb{E}[S_{\max}(X) \mid X] \le H_{\max}\mathbb{E}[N_{\max} \mid X]$. Since both quantities are nonnegative, this gives the prompt-level bound $0 \le \mu_{\max}(X) \le H_{\max}\mathbb{E}[N_{\max} \mid X]$. Therefore, $\mathrm{Var}_X(\mu_{\max}(X)) \le H^2_{\max}\,\mathrm{Var}_X(\mathbb{E}[N_{\max} \mid X])$. Combining these two refined bounds with the exact variance decomposition gives

$$\mathrm{Var}\left(\widehat{\mathrm{CE}}^{\star}_{\max,P,M}\right) \le \frac{H^2_{\max}}{P}\,\mathrm{Var}_X(\mathbb{E}[N_{\max} \mid X]) + \frac{H^2_{\max}}{PM}\mathbb{E}[N^2_{\max}].$$

Adding the squared truncation bias $B^2_{T_{\max}}$ gives the stated refined MSE upper bound. □

D.1.6 Estimation and consistency of GenPPL, BF, and length–uncertainty correlation

Lemma D.1.1 (Consistency of the truncated length estimator)**.** *Let* $\widehat{N}_{\max,P,M} := \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M} N^{(p,i)}_{\max}$, *where* $N^{(p,i)}_{\max} := \min\{N^{(p,i)}, T_{\max}\}$. *Assume* $\mathbb{E}[N] < \infty$. *If* $T_{\max} = T_{\max}(P, M) \to \infty$ *and* $\frac{\sqrt{\mathbb{E}[N^2_{\max}]}}{\sqrt{P}} \to 0$, *then* $\widehat{N}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathbb{E}[N]$.

Proof. Let $\eta_{\max}(x) := \mathbb{E}[N_{\max} \mid X = x]$. By the law of total variance, $\mathrm{Var}(\widehat{N}_{\max,P,M}) = \frac{1}{P}\mathrm{Var}(\eta_{\max}(X)) + \frac{1}{PM}\mathbb{E}\left[\mathrm{Var}(N_{\max} \mid X)\right]$. Using $\mathrm{Var}(\eta_{\max}(X)) \le \mathbb{E}[\eta_{\max}(X)^2] \le \mathbb{E}[N_{\max}^2]$, and $\mathbb{E}[\mathrm{Var}(N_{\max} \mid X)] \le \mathbb{E}[N_{\max}^2]$, we obtain $\mathrm{Var}(\widehat{N}_{\max,P,M}) \le \mathbb{E}[N_{\max}^2]\left(\frac{1}{P} + \frac{1}{PM}\right)$. Therefore, under the stated condition, $\mathrm{Var}(\widehat{N}_{\max,P,M}) \to 0$.

In addition, $\mathbb{E}[\widehat{N}_{\max,P,M}] = \mathbb{E}[N_{\max}]$. Since $N_{\max} = \min\{N, T_{\max}\} \uparrow N$ as $T_{\max} \to \infty$, the monotone convergence theorem [Rudin, 2021] gives $\mathbb{E}[N_{\max}] \to \mathbb{E}[N]$. Hence,

$$\widehat{N}_{\max,P,M} - \mathbb{E}[N] = \left(\widehat{N}_{\max,P,M} - \mathbb{E}[N_{\max}]\right) + \left(\mathbb{E}[N_{\max}] - \mathbb{E}[N]\right).$$

The first term converges to 0 in probability by Chebyshev's inequality, and the second term converges to 0 deterministically. Therefore, $\widehat{N}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathbb{E}[N]$. □

Corollary 6 (Consistency of GenPPL)**.** *Let* $\widehat{\mathrm{CE}}^{\star}_{\max,P,M} = \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M} S_{\max}^{(p,i)}$ *be the Monte Carlo estimator of* $\mathrm{CE}^{\star}$*, and let* $\widehat{N}_{\max,P,M} := \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M} N_{\max}^{(p,i)}$ *be the corresponding estimator of* $\mathbb{E}[N]$*. Since* $\widehat{\mathrm{CE}}^{\star}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathrm{CE}^{\star}$*, (by Theorem 5.4.1)* $\widehat{N}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathbb{E}[N]$*, (by Lemma D.1.1). Define* $\widehat{\mathrm{GenPPL}}_{\max,P,M} := \exp\left(\frac{\widehat{\mathrm{CE}}^{\star}_{\max,P,M}}{\widehat{N}_{\max,P,M}}\right)$, $\mathrm{GenPPL} := \exp\left(\frac{\mathrm{CE}^{\star}}{\mathbb{E}[N]}\right)$*. Then* $\widehat{\mathrm{GenPPL}}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathrm{GenPPL}$.

Proof. We first prove consistency of GenPPL. Since we have $\widehat{\mathrm{CE}}^{\star}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathrm{CE}^{\star}$ and $\widehat{N}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathbb{E}[N]$. and $\mathbb{E}[N] > 0$, the function $f(a,b) = \exp(a/b)$ is continuous at $(\mathrm{CE}^{\star}, \mathbb{E}[N])$. Therefore, by the continuous mapping theorem [Mann and Wald, 1943], $\exp\left(\frac{\widehat{\mathrm{CE}}^{\star}_{\max,P,M}}{\widehat{N}_{\max,P,M}}\right) \xrightarrow{\mathbb{P}} \exp\left(\frac{\mathrm{CE}^{\star}}{\mathbb{E}[N]}\right)$. Hence, $\widehat{\mathrm{GenPPL}}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathrm{GenPPL}$, which completes the proof. □

Corollary 7 (Consistency of the BF estimator)**.** *Let* $\widehat{L}_{\max,P,M} := \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M}\frac{S_{\max}^{(p,i)}}{N_{\max}^{(p,i)}}$, $\widehat{\mathrm{BF}}_{\max,P,M} := \exp\left(\widehat{L}_{\max,P,M}\right)$*, where* $S_{\max}^{(p,i)} = \sum_{t=1}^{N_{\max}^{(p,i)}} \widetilde{H}(g_t^{(p,i)} \mid X_p)$, $N_{\max}^{(p,i)} = \min\{N^{(p,i)}, T_{\max}\}$. *Assume that* $N \ge 1$ *almost surely and that the normalized truncated entropy variables* $r_{\max} := \frac{S_{\max}}{N_{\max}}$ *are integrable. Then, for fixed* $T_{\max}$*, as* $P, M \to \infty$, $\widehat{L}_{\max,P,M} \xrightarrow{\mathbb{P}} L_{\max} := \mathbb{E}\left[\frac{S_{\max}}{N_{\max}}\right]$*, and therefore, by the continuous mapping theorem,* $\widehat{\mathrm{BF}}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathrm{BF}_{\max} := \exp(L_{\max})$.

Moreover, if $\frac{S_{\max}}{N_{\max}} \xrightarrow[T_{\max}\to\infty]{a.s.} \frac{S}{N}$ *and the sequence* $\{S_{\max}/N_{\max}\}_{T_{\max}\geq 1}$ *is uniformly integrable, then* $L_{\max} \to L := \mathbb{E}\left[\frac{1}{N}\sum_{t=1}^{N}\widetilde{H}(g_t \mid X)\right]$, *and consequently* $\mathrm{BF}_{\max} \to \mathrm{BF} := \exp(L)$. *Thus, under these conditions,* $\widehat{\mathrm{BF}}_{\max,P,M} \xrightarrow{\mathbb{P}} \mathrm{BF}$ *as* $P, M \to \infty$ *and* $T_{\max} \to \infty$.

Proof. For fixed $T_{\max}$, the variables $\hat{r}_{\max}^{(p,i)} = \frac{S_{\max}^{(p,i)}}{N_{\max}^{(p,i)}}$ are integrable rollout-level observations. Since $N \geq 1$ almost surely, the denominator is bounded away from zero. Applying the law of large numbers to the empirical average over prompts and rollouts gives $\widehat{L}_{\max,P,M} = \frac{1}{PM}\sum_{p=1}^{P}\sum_{i=1}^{M}\hat{r}_{\max}^{(p,i)} \xrightarrow{\mathbb{P}} \mathbb{E}[r_{\max}] = L_{\max}$. Since the exponential map is continuous, $\widehat{\mathrm{BF}}_{\max,P,M} = \exp(\widehat{L}_{\max,P,M}) \xrightarrow{\mathbb{P}} \exp(L_{\max}) = \mathrm{BF}_{\max}$. Finally, if $S_{\max}/N_{\max} \to S/N$ almost surely and the normalized entropy ratios are uniformly integrable, then convergence of expectations follows, yielding $L_{\max} \to L$. Another application of continuity of the exponential function gives $\mathrm{BF}_{\max} \to \mathrm{BF}$. □

Empirical support for the truncation condition. The consistency result above requires the truncation bias induced by $T_{\max}$ to vanish, or at least to be negligible in finite-sample estimation. While this is a population-level condition and cannot be proven from data alone, our diagnostics provide empirical support for it in our experimental setting. With $T_{\max} = 4000$, the observed truncation rates are uniformly small across tasks, model families, and variants: most are below 1%, and the largest observed rate is 4.32% (see subsection D.3.1). In addition, the empirical length distributions concentrate well below the truncation threshold (see Figure 5.3), and the number of active rollouts decays rapidly with token position (see Figure 5.2). Together with the bounded observed entropy-rate trajectories (see Figure 5.2), these results suggest that the normalized entropy ratios are well behaved and that the finite-$T_{\max}$ truncation bias is small in practice.

Estimating entropy rate. For each rollout (p, i), Algorithm 10 returns the truncated stopping time $N_{\max}^{(p,i)} = \min\{N^{(p,i)}, T_{\max}\}$ and the accumulated truncated canopy entropy $S_{\max}^{(p,i)} = \sum_{t=1}^{N_{\max}^{(p,i)}}\widetilde{H}(g_t^{(p,i)} \mid X_p)$. We estimate the rollout-level entropy rate by $\hat{r}_{\max}^{(p,i)} := \frac{S_{\max}^{(p,i)}}{N_{\max}^{(p,i)}}$.

Thus, for each prompt X_p, the prompt-specific length–uncertainty relationship is estimated from the rollout-level pairs $\left\{\left(N_{\max}^{(p,i)}, \widehat{r}_{\max}^{(p,i)}\right)\right\}_{i=1}^{M}$.

Rank-based length–uncertainty correlations. In addition to Pearson correlation, which measures linear association, we also report two rank-based measures. Spearman correlation [Spearman, 1961] is Pearson correlation applied to the ranked variables and therefore captures monotone dependence. Kendall's τ_b [Kendall, 1938] measures ordinal association by comparing concordant and discordant rollout pairs, with tie adjustment. Since Kendall's τ_b is defined through pairwise comparisons, we aggregate it using pair-count weights rather than Fisher-z averaging.

Proposition 8 (Consistency of prompt-controlled length–uncertainty correlation estimators)**.** *For each prompt $X = x$, define $r_{\max} := \frac{S_{\max}}{N_{\max}}$, $S_{\max} = \sum_{t=1}^{N_{\max}} \widetilde{H}(g_t \mid X)$, $N_{\max} = \min\{N, T_{\max}\}$. Let $\rho_{\max}^{\mathrm{P}}(x) := \mathrm{Corr}(N_{\max}, r_{\max} \mid X = x)$ denote the Pearson correlation, let $\rho_{\max}^{\mathrm{S}}(x) := \mathrm{Corr}\left(F_{N_{\max}|x}(N_{\max}), F_{r_{\max}|x}(r_{\max}) \mid X = x\right)$ denote the population Spearman rank correlation, and let $\tau_{\max}^{\mathrm{K}}(x)$ denote Kendall's τ_b between $N_{\max}$ and $r_{\max}$ conditional on $X = x$. For sampled prompts $X_1, \ldots, X_P$, suppose that for each prompt X_p we observe n_p conditionally i.i.d. rollouts and compute the prompt-specific estimates $\widehat{\rho}_p^{\mathrm{P}}$, $\widehat{\rho}_p^{\mathrm{S}}$, $\widehat{\tau}_p^{\mathrm{K}}$, $\widehat{\mathrm{Cov}}_p$ as in Algorithm 11. Assume that $N_{\max} \geq 1$ almost surely. For Pearson and covariance, assume that the conditional fourth moments of $N_{\max}$ and $r_{\max}$ are uniformly bounded and that $0 < c \leq \mathrm{Var}(N_{\max} \mid X = x)$, $0 < c \leq \mathrm{Var}(r_{\max} \mid X = x)$ uniformly over x. For Spearman and Kendall, assume the corresponding rank-correlation functionals are continuous at the conditional law of $(N_{\max}, r_{\max}) \mid X = x$, uniformly over x; in particular, ties are either absent or handled by the mid-rank and τ_b conventions used in Algorithm 11. Finally, assume $|\rho_{\max}^{\mathrm{P}}(x)| \leq 1 - \epsilon_0$, $|\rho_{\max}^{\mathrm{S}}(x)| \leq 1 - \epsilon_0$ for some $\epsilon_0 > 0$, and that $\min_p n_p \to \infty$ as $P \to \infty$. Then, as $P \to \infty$ and $\min_p n_p \to \infty$, $\widehat{\rho}_{\mathrm{pc,max}}^{\mathrm{P}} \xrightarrow{\mathbb{P}} \rho_{\mathrm{pc,max}}^{\mathrm{P}}$, $\widehat{\rho}_{\mathrm{pc,max}}^{\mathrm{S}} \xrightarrow{\mathbb{P}} \rho_{\mathrm{pc,max}}^{\mathrm{S}}$, where $\rho_{\mathrm{pc,max}}^{\mathrm{P}} = \tanh\left(\frac{\mathbb{E}_X[w^{\mathrm{F}}(X)\,\mathrm{atanh}\{\rho_{\max}^{\mathrm{P}}(X)\}]}{\mathbb{E}_X[w^{\mathrm{F}}(X)]}\right)$, $\rho_{\mathrm{pc,max}}^{\mathrm{S}} = \tanh\left(\frac{\mathbb{E}_X[w^{\mathrm{F}}(X)\,\mathrm{atanh}\{\rho_{\max}^{\mathrm{S}}(X)\}]}{\mathbb{E}_X[w^{\mathrm{F}}(X)]}\right)$.*

Also, $\widehat{\tau}^{\mathrm{K}}_{\mathrm{pc,max}} \xrightarrow{\mathbb{P}} \tau^{\mathrm{K}}_{\mathrm{pc,max}} := \frac{\mathbb{E}_X[w^{\mathrm{K}}(X)\tau^{\mathrm{K}}_{\max}(X)]}{\mathbb{E}_X[w^{\mathrm{K}}(X)]}$, *and* $\widehat{\mathrm{Cov}}_{\mathrm{pc,max}} \xrightarrow{\mathbb{P}} \mathrm{Cov}_{\mathrm{pc,max}} := \frac{\mathbb{E}_X[w^{\mathrm{Cov}}(X)\,\mathrm{Cov}(N_{\max}, r_{\max}|X)]}{\mathbb{E}_X[w^{\mathrm{Cov}}(X)]}$. *If additionally* $T_{\max} \to \infty$ *and the corresponding truncated targets converge to their untruncated counterparts, then the same estimators are consistent for the corresponding prompt-controlled untruncated Pearson, Spearman, Kendall, and covariance targets.*

Proof. We prove the result in two steps: first within each prompt, and then across prompts. For a fixed prompt $X_p = x$, the rollout-level pairs $\left(N^{(p,i)}_{\max}, r^{(p,i)}_{\max}\right)$, $i = 1, \ldots, n_p$, are conditionally i.i.d. By Algorithm 11, both coordinates are observed from the same rollout, with $r^{(p,i)}_{\max} = \frac{S^{(p,i)}_{\max}}{N^{(p,i)}_{\max}}$. The moment and nondegeneracy assumptions imply consistency of the sample covariance and variances. Hence, $\widehat{\rho}^{\mathrm{P}}_p \xrightarrow{\mathbb{P}} \rho^{\mathrm{P}}_{\max}(X_p)$, $\widehat{\mathrm{Cov}}_p \xrightarrow{\mathbb{P}} \mathrm{Cov}(N_{\max}, r_{\max} \mid X_p)$. Similarly, by consistency of empirical ranks and the assumed continuity of the rank-correlation functionals, $\widehat{\rho}^{\mathrm{S}}_p \xrightarrow{\mathbb{P}} \rho^{\mathrm{S}}_{\max}(X_p)$, $\widehat{\tau}^{\mathrm{K}}_p \xrightarrow{\mathbb{P}} \tau^{\mathrm{K}}_{\max}(X_p)$. Since $|\rho^{\mathrm{P}}_{\max}(x)|$ and $|\rho^{\mathrm{S}}_{\max}(x)|$ are bounded away from one, and the clipping constant is chosen smaller than this margin, the Fisher transform is continuous in a neighborhood of the limiting values. Therefore, $\mathrm{atanh}\left(\mathrm{clip}(\widehat{\rho}^{\mathrm{P}}_p, -1+\epsilon, 1-\epsilon)\right) \xrightarrow{\mathbb{P}} \mathrm{atanh}\{\rho^{\mathrm{P}}_{\max}(X_p)\}$, and $\mathrm{atanh}\left(\mathrm{clip}(\widehat{\rho}^{\mathrm{S}}_p, -1+\epsilon, 1-\epsilon)\right) \xrightarrow{\mathbb{P}} \mathrm{atanh}\{\rho^{\mathrm{S}}_{\max}(X_p)\}$. Now consider the aggregation across prompts. For Pearson, define $Z^{\mathrm{P}}_p = \mathrm{atanh}\{\rho^{\mathrm{P}}_{\max}(X_p)\}$. By the previous step and the uniform assumptions, $\frac{\sum_{p=1}^{P} w^{\mathrm{F}}_p [\mathrm{atanh}(\mathrm{clip}(\widehat{\rho}^{\mathrm{P}}_p, -1+\epsilon, 1-\epsilon)) - Z^{\mathrm{P}}_p]}{\sum_{p=1}^{P} w^{\mathrm{F}}_p} \xrightarrow{\mathbb{P}} 0$. Since prompts are sampled i.i.d., the weighted law of large numbers gives $\frac{\sum_{p=1}^{P} w^{\mathrm{F}}_p Z^{\mathrm{P}}_p}{\sum_{p=1}^{P} w^{\mathrm{F}}_p} \xrightarrow{\mathbb{P}} \frac{\mathbb{E}_X[w^{\mathrm{F}}(X)\,\mathrm{atanh}\{\rho^{\mathrm{P}}_{\max}(X)\}]}{\mathbb{E}_X[w^{\mathrm{F}}(X)]}$. Applying the continuous mapping theorem with $\tanh(\cdot)$ yields $\widehat{\rho}^{\mathrm{P}}_{\mathrm{pc,max}} \xrightarrow{\mathbb{P}} \rho^{\mathrm{P}}_{\mathrm{pc,max}}$. The Spearman result is identical after replacing $\rho^{\mathrm{P}}_{\max}$ by $\rho^{\mathrm{S}}_{\max}$. For Kendall's τ_b, Algorithm 11 aggregates directly on the τ scale using pair-count weights $w^{\mathrm{K}}_p = \frac{n_p(n_p-1)}{2}$. Thus, by the prompt-level consistency of $\widehat{\tau}^{\mathrm{K}}_p$ and the weighted law of large numbers, $\widehat{\tau}^{\mathrm{K}}_{\mathrm{pc,max}} = \frac{\sum_{p=1}^{P} w^{\mathrm{K}}_p \widehat{\tau}^{\mathrm{K}}_p}{\sum_{p=1}^{P} w^{\mathrm{K}}_p} \xrightarrow{\mathbb{P}} \frac{\mathbb{E}_X[w^{\mathrm{K}}(X)\tau^{\mathrm{K}}_{\max}(X)]}{\mathbb{E}_X[w^{\mathrm{K}}(X)]}$. The covariance result follows in the same way, using weights $w^{\mathrm{Cov}}_p = (n_p - 1)_+$ and the prompt-level consistency of $\widehat{\mathrm{Cov}}_p$. Finally, if $T_{\max} \to \infty$ and the truncated conditional covariance,

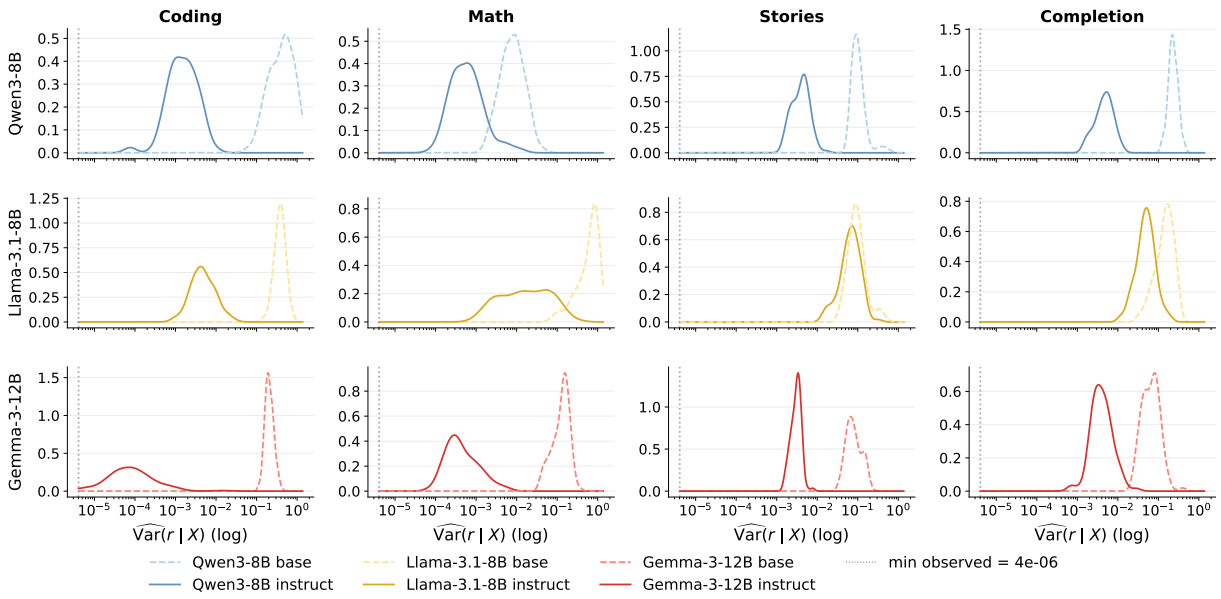

Figure D.1: Gaussian KDEs of $\log \widehat{\mathrm{Var}}(r \mid X_p)$ over P=100 prompts. The dotted vertical marks $\min_p \widehat{\mathrm{Var}}(r \mid X_p) \approx 4 \times 10^{-6}$. All densities sit well to the right of zero, providing empirical support for the bounded-away-from-zero assumption.

variance, and rank-association functionals converge to their untruncated counterparts, then the truncated prompt-controlled targets converge to the corresponding untruncated targets. Combining this approximation step with the fixed-$T_{\max}$ consistency above completes the proof. □

Empirical support for the consistency assumptions. The assumptions underlying Proposition 8 are supported by several empirical diagnostics observed in our experiments. First, the truncation rates under $T_{\max} = 4000$ are uniformly small across model families, tasks, and variants, with most rates below 1% and the maximum observed rate equal to 4.32% (see Table D.3). This suggests that the truncated quantities $(N_{\max}, r_{\max})$ closely approximate their untruncated counterparts (N, r_N) in practice. Second, the empirical sequence-length distributions exhibit rapidly decaying tails, and the number of active roll-outs decreases quickly as generation proceeds, providing evidence that the stopping-time distribution is sufficiently light-tailed and that truncation bias is negligible for the chosen

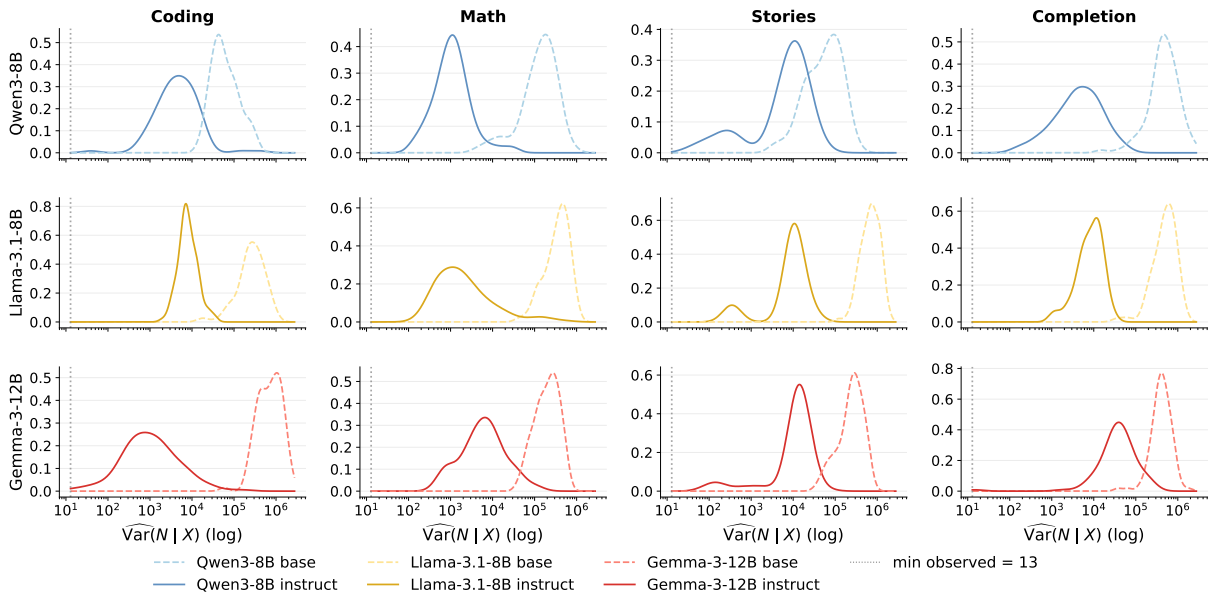

Figure D.2: Gaussian KDEs of $\log \widehat{\mathrm{Var}}(N \mid X_p)$ over P=100 prompts. The dotted vertical line marks $\min_p \widehat{\mathrm{Var}}(N \mid X_p) \approx 13$, the smallest per-prompt variance observed across all model-dataset combinations. All densities concentrate at $\widehat{\mathrm{Var}}(N \mid X_p) \gg 0$, empirically supporting the bounded-away-from-zero assumption. Instruct variants consistently shift to smaller values, indicating that fine-tuning produces tighter length distributions.

$T_{\max}$ (see Figure 5.3).

Third, the rollout-level entropy rates remain bounded and stable throughout generation trajectories, supporting the bounded-moment and continuity assumptions required for consistency of the covariance and correlation estimators (see Figure 5.2). Finally, substantial variability is consistently observed in both generation length and entropy rate across prompts and rollouts, indicating that the conditional variances of $N_{\max}$ and $r_{\max}$ are bounded away from zero with high probability (see Figure D.1 and Figure D.2). Together, these empirical observations suggest that the regularity conditions required for the consistency of the prompt-controlled Pearson, Spearman, Kendall, and covariance estimators are well approximated in our experimental regime.

D.2 Discussions

D.2.1 When to use BF vs. GenPPL across different tasks.

We refer readers to Table D.2 for a concise practical comparison of when to use BF versus GenPPL across different tasks.

D.2.2 Comparison with perplexity.

Perplexity [Jelinek et al., 1977] is a widely used evaluation metric in the LLM literature. To avoid confusion, we briefly contrast it with GenPPL and BF. Perplexity is defined on a fixed dataset $(y_1, \ldots, y_T)$, where $y_t \in \mathcal{V}$ is the observed token and $y_{<t}$ denotes its context. Given the model distribution $q(\cdot \mid y_{<t})$, it is defined as $\mathrm{PPL} = \exp\left(\frac{1}{T}\sum_{t=1}^{T} -\log q(y_t \mid y_{<t})\right)$, where $-\log q(y_t \mid y_{<t})$ measures how surprising the true token is under the model. Thus, perplexity evaluates predictive accuracy, with lower values indicating better fit to the data. In contrast, GenPPL and BF are defined with respect to the model's generative distribution and measure generation capacity. Namely, the size and diversity of the space of possible outputs rather than accuracy on the observed data.

D.2.3 An essay analogy: how GenPPL and BF capture different notions of diversity

We illustrate the difference between GenPPL and BF through an essay–word analogy, where trajectories correspond to essays and tokens correspond to individual words. For a fixed prompt x, suppose a model generates two types of essays. With probability $1/2$, it produces a short essay of length $N = 2$, where each word is selected uniformly from 100 plausible words. With probability $1/2$, it produces a long essay of length $N = 10$, where each word is selected uniformly from only 2 plausible words. For the short essays, $H(Y_{1:2} \mid N = 2, x) = 2\log 100, r_2 = \log 100$.For the long essays, $H(Y_{1:10} \mid N = 10, x) = 10\log 2, r_{10} = \log 2$.

BF treats each *trajectory* equally: $\log \mathrm{BF} = \frac{1}{2}\log 100 + \frac{1}{2}\log 2$, which yields $\mathrm{BF} \approx 14$. This means that a typical generated trajectory behaves as if each token has roughly 14 plausible continuations. Because BF weights trajectories equally, short but highly diverse generations contribute substantially even if they contain relatively few tokens overall. In contrast, GenPPL treats each *token* equally. Since most generated tokens come from the long low-diversity essays, we obtain GenPPL ≈ 3.Thus, a typical token effectively has only about 3 plausible continuations.

The distinction arises from the averaging mechanism. BF averages uncertainty at the trajectory level, while GenPPL averages at the token level and therefore assigns greater weight to longer generations. Consequently, BF reflects the diversity of a *typical trajectory*, whereas GenPPL reflects the uncertainty of a *typical token*. These two measures therefore provide complementary views of generation space that cannot be captured by a single metric alone.

D.3 Experiment details

In this section, we provide additional experimental and implementation details underlying our empirical analysis. In particular, we discuss the stopping behavior of autoregressive generation and the resulting truncation bias, as well as the statistical modeling procedures used throughout the paper. We further present detailed outputs from the Beta mixed-effects regression analysis, including coefficient estimates, interaction effects, and dispersion modeling results, together with additional discussions on uncertainty allocation, semantic diversity, and length–entropy dynamics across model families and tasks.

D.3.1 Stopping criteria and empirical truncation bias

To ensure the generative process remains consistent with the theoretical framework, we include the `EOS` token in the sampling vocabulary $\mathcal{V} \cup \{\texttt{EOS}\}$. Crucially, during top-$k$ sampling,

a rollout only terminates if `EOS` is present within the top k most probable tokens and is subsequently sampled.

While our theoretical framework allows for infinite generation length, practical estimation requires a finite truncation point $T_{\max}$. To ensure that our estimator $\widehat{\mathrm{CE}}^{*}_{\max,P,M}$ remains a consistent proxy for the true Canopy Entropy CE^{*}, we set a generous token budget of $T_{\max} = 4000$. As shown in Table D.3, the empirical truncation rate remains low across all model families and tasks. Notably, fine-tuned models exhibit a 0.00% truncation rate in almost all settings, confirming that aligned models naturally terminate well within the provided budget. While truncation rates are higher for base models, they remain sufficiently low (≤ 0.0432). Overall, these results demonstrate that the truncation bias term $g(T_{\max})$ defined in Theorem 5.4.1 is sufficiently small, ensuring that the observed differences in generation space are not artifacts of the maximum length constraint.

D.3.2 Completion traps

Instead of answering the instruction, base models sometimes treat the prompt as an unfinished text prefix and continue it with a short completion fragment. Here we present some empirical examples where base models fall into such "completion traps". Some examples are listed below.

Prompt:

> Come up with an original and creative solution for the following real-world problem: Clara, a junior pre-med student, is working part-time and taking a 15 hour credit load at school /.../ Clara is not sure how to solve her problem..

Qwen3-8B Base:

```
"In order to entice donors, you would like to take the following approach:"
```

Llama-3.1-8B Base:

```
"How can she decide what steps she should take to solve her problem?"
```

Gemma-3-12B Base:

```
"There will be partial credit for good ideas that might just work (even
if they're not perfect).  Your solution will be judged on its creativity
and how well it works."
```

Coding prompts were especially vulnerable to premature termination caused by structural delimiters. Prompts ending with a closing code block delimiter (‘‘‘) frequently produced empty responses from base models, as they likely interpret the closing backticks as an end-of-document signal from the training distribution. To ensure valid evaluations, we manually stripped these trailing delimiters to force the model to generate the intended logic rather than treating the task as already finalized.

D.3.3 CE^* *decomposition analysis*

By defining the rollout-level entropy rate $r_N = \frac{1}{N}\sum_{t=1}^{N} \tilde{H}(G_t|X)$, we can express $\mathrm{CE}^\star = \mathbb{E}\left[\sum_{t=1}^{N} \tilde{H}(G_t|X)\right]$ as $\mathrm{CE}^\star = \mathbb{E}[N \cdot r_N]$. We thus obtain the $\mathrm{CE}^\star$ decomposition as

$$\mathrm{CE}^\star = \underbrace{\mathbb{E}[N]\mathbb{E}[r_N]}_{\text{Length-driven component}} + \underbrace{\mathrm{Cov}(N, r_N)}_{\text{Structural coupling}}$$

This formulation reveals that total uncertainty is not just a product of average length and average entropy rate, but also depends on how the information density (r_N) changes as generations grow longer (N).

Empirically, we observe that fine-tuning consistently shifts this covariance term $\mathrm{Cov}(N, r_N)$ from negative toward positive values, indicating that the interaction between generation length and entropy rate contributes positively to total uncertainty, rather than longer generations being dominated by low-information continuation (see Figure D.3).

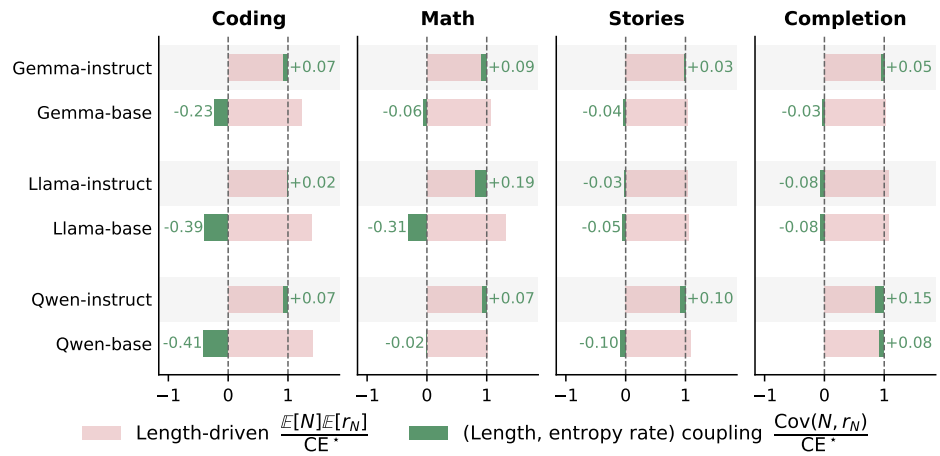

Figure D.3: Decomposition of total generation uncertainty across Qwen3-8B, LLaMA-3.1-8B, and Gemma-3-12B families on different tasks. Pink bars represent the normalized length-driven component $\mathbb{E}[N]\mathbb{E}[r_N]/\mathrm{CE}^\star$, while green bars represent the normalized length–entropy-rate coupling term $\mathrm{Cov}(N, r_N)/\mathrm{CE}^\star$.

D.3.4 Bootstrap procedure

All standard errors and confidence intervals reported in Table 5.1, Table 5.2, and Table D.1 are estimated using a prompt-level cluster bootstrap with $B = 2000$ replicates. In each replicate, we sample $P = 100$ prompt indices uniformly with replacement from the original set of prompts. All $M = 100$ rollouts associated with a sampled prompt are resampled jointly as a single cluster, preserving the within-prompt rollout structure.

Point estimates are always computed on the original, non-resampled data; bootstrap replicates are used only for uncertainty quantification. Reported standard errors correspond to the sample standard deviation of the bootstrap replicates with degree of freedom $= n - 1$.

D.3.5 Beta mixed-effects model for semantic diversity

The analysis is based on 2400 observations, corresponding to 4 tasks, 3 model families, and 2 model variants (base and instruct), with 100 prompts evaluated for each configuration.

Conditional mean modeling. Formally, we model $D_{tpm} \sim \mathrm{Beta}\big(\mu_{tpm}\phi_{tpm}, (1-\mu_{tpm})\phi_{tpm}\big)$, where $\mu_{tpm} \in (0,1)$ denotes the conditional mean and $\phi_{tpm} > 0$ is the precision parameter.

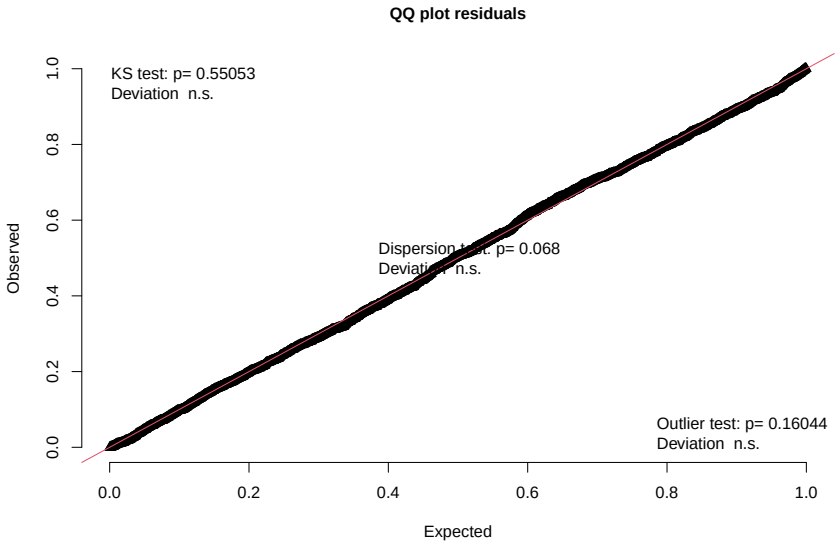

Figure D.4: DHARMa residual diagnostics for the fitted Beta mixed-effects regression model. The QQ plot shows no significant deviation from the expected uniform residual distribution.

The conditional mean model is specified as $\text{logit}(\mu_{tpm}) = \alpha + \beta_1 R_{tpm} + \beta_2 \mathbf{1}\{m = \text{FT}\} + f\left(\text{invN}_{tpm}\right) + \sum_k \tau_k \mathbf{1}\{t = k\} + \beta_3 (R_{tpm} \cdot \mathbf{1}\{m = \text{FT}\}) + g\left(\text{invN}_{tpm}\right) \mathbf{1}\{m = \text{FT}\} + u_p + v_m$ where R_{tpm} is the entropy rate, invN_{tpm} is the inverse sequence length, and $f(\cdot)$, $g(\cdot)$ are natural spline functions corresponding to nonlinear length effects and their interaction with instruction tuning. The coefficients τ_k represent fixed task effects. To account for hierarchical dependence, we include random intercepts $u_p \sim \mathcal{N}(0, \sigma_p^2), v_m \sim \mathcal{N}(0, \sigma_m^2)$, for prompt identity and model family, respectively.

In addition to the finding presented in section 5.5, we also observe strong nonlinear dependence on inverse sequence length. Several spline terms and their interactions with fine-tuning are significant, confirming that the relationship between diversity and length is highly nonlinear and differs between base and fine-tuned models. This reinforces that simple linear controls for length are insufficient.

Task-level effects are also pronounced. Relative to completion, mathematical reasoning exhibits significantly higher semantic diversity ($+0.169$, $p < 10^{-6}$), while story generation shows substantially lower embedding-based diversity (-0.276, $p < 2 \times 10^{-16}$), despite ap-

pearing lexically diverse. This suggests that open-ended generation can remain semantically concentrated even when surface variation is high (see Table D.4).

Finally, random effects indicate that prompt-level variability remains larger than model-level variability, highlighting the strong influence of prompt semantics on generation diversity (see Table D.5).

Dispersion modeling. The dispersion model specifies how the precision parameter ϕ_{tpm} varies across model variants, tasks, sequence length, entropy rate, and model family. Specifically, we use the log-link specification $\log(\phi_{tpm}) = \gamma_0 + \gamma_1 \mathbf{1}\{m = \text{FT}\} + \sum_k \delta_k \mathbf{1}\{t = k\} + h\big(\text{invN}_{tpm}\big) + \gamma_2 R_{tpm} + \sum_j \eta_j \mathbf{1}\{\text{model} = j\} + \sum_k \zeta_k \mathbf{1}\{m = \text{FT}\}\mathbf{1}\{t = k\} + \gamma_3 \left(R_{tpm} \cdot \mathbf{1}\{m = \text{FT}\}\right)$, where $h(\cdot)$ is a natural spline function capturing nonlinear length-dependent dispersion effects. The task indicators δ_k allow the residual precision to vary across tasks, while the model-family indicators η_j account for systematic differences in dispersion across model families. The interaction terms between model variant and task allow fine-tuning to affect dispersion differently across tasks, and the interaction between entropy rate and model variant tests whether the relationship between entropy rate and precision changes after instruction tuning.

Empirically, the dispersion model reveals substantial heterogeneity in residual precision. Instruction-tuned models have significantly higher baseline precision than base models ($\hat{\gamma}_1 = 1.415$, $p < 0.001$), indicating lower conditional variability after accounting for the mean structure. Precision also differs strongly by task, with coding, math, and story generation all showing significantly higher precision relative to the reference task. The spline terms for inverse length indicate that dispersion varies nonlinearly with output length, while the positive main effect of entropy rate ($\hat{\gamma}_2 = 0.975$, $p < 0.001$) suggests that higher entropy rate is associated with greater precision among base models. However, the negative interaction between instruction tuning and entropy rate ($\hat{\gamma}_3 = -1.609$, $p < 0.001$) shows that this relationship is substantially weakened or reversed for fine-tuned models. The sig-

nificant model-variant-by-task interactions further indicate that fine-tuning changes residual variability in a task-dependent manner (see Table D.6).

Residual diagnosis. The figure (see Figure D.4) presents a DHARMa residual diagnostic QQ plot [Hartig, 2016] for the fitted Beta mixed-effects regression model, comparing the empirical residual distribution against the expected uniform distribution. The residuals closely follow the diagonal reference line, indicating good overall model calibration and no substantial systematic deviation from the assumed distribution. The associated diagnostic tests further support model adequacy: the Kolmogorov–Smirnov test ($p = 0.551$) suggests no significant deviation from uniformity, the dispersion test ($p = 0.068$) indicates no significant over- or under-dispersion, and the outlier test ($p = 0.160$) shows no evidence of excessive outliers. Overall, these diagnostics suggest that the fitted model provides an adequate characterization of the observed data and that the regression results are statistically reliable.

D.3.6 Compute resources and runtime

All experiments were run on a single internal SLURM cluster. Each generation job uses one NVIDIA GPU (A100, H100, or H200), 8 CPU cores, and 100 GB host RAM, with a 6-hour wall-clock timeout per job. Overall, we estimate total project compute at approximately 30-40 GPU-hours, overwhelmingly dominated by rollout generation; metric computation and statistical analysis were negligible by comparison and can be completed on CPU.

Algorithm 11: Estimating prompt-controlled length–uncertainty correlations

Input: Prompt distribution p_X; number of prompts P; rollout counts $\{n_p\}_{p=1}^{P}$; maximum length $T_{\max}$; sampling policy π; clipping constant $\epsilon > 0$.

for $p = 1$ **to** P **do**
  Sample prompt $X_p \sim p_X$;
  for $i = 1$ **to** n_p **do**
    Generate one rollout using Algorithm 10;
    Record $N_{\max}^{(p,i)} = \min\{N^{(p,i)}, T_{\max}\}$, $S_{\max}^{(p,i)} = \sum_{t=1}^{N_{\max}^{(p,i)}} \widetilde{H}(g_t^{(p,i)} \mid X_p)$. Compute the rollout-level entropy rate $r_{\max}^{(p,i)} = \frac{S_{\max}^{(p,i)}}{N_{\max}^{(p,i)}}$.
  Compute the prompt-specific Pearson correlation $\widehat{\rho}_p^{\mathrm{P}} = \mathrm{Corr}\left(\{N_{\max}^{(p,i)}\}_{i=1}^{n_p}, \{r_{\max}^{(p,i)}\}_{i=1}^{n_p}\right)$.
  Compute the prompt-specific Spearman correlation $\widehat{\rho}_p^{\mathrm{S}} = \mathrm{Corr}\left(\{\mathrm{rank}(N_{\max}^{(p,i)})\}_{i=1}^{n_p}, \{\mathrm{rank}(r_{\max}^{(p,i)})\}_{i=1}^{n_p}\right)$, using mid-ranks for ties.
  Compute the prompt-specific Kendall's τ_b $\widehat{\tau}_p^{\mathrm{K}} = \mathrm{KendallTauB}\left(\{N_{\max}^{(p,i)}\}_{i=1}^{n_p}, \{r_{\max}^{(p,i)}\}_{i=1}^{n_p}\right)$.
  Compute the prompt-specific covariance $\widehat{\mathrm{Cov}}_p = \widehat{\mathrm{Cov}}\left(\{N_{\max}^{(p,i)}\}_{i=1}^{n_p}, \{r_{\max}^{(p,i)}\}_{i=1}^{n_p}\right)$.
Aggregate Pearson and Spearman using Fisher-z weighted averages: $w_p^{\mathrm{F}} = \max\{n_p - 3, 0\}$, $\widehat{\rho}_{\mathrm{pc}}^{\mathrm{P}} = \tanh\left(\frac{\sum_{p=1}^{P} w_p^{\mathrm{F}} \operatorname{atanh}(\mathrm{clip}(\widehat{\rho}_p^{\mathrm{P}}, -1+\epsilon, 1-\epsilon))}{\sum_{p=1}^{P} w_p^{\mathrm{F}}}\right)$, and $\widehat{\rho}_{\mathrm{pc}}^{\mathrm{S}} = \tanh\left(\frac{\sum_{p=1}^{P} w_p^{\mathrm{F}} \operatorname{atanh}(\mathrm{clip}(\widehat{\rho}_p^{\mathrm{S}}, -1+\epsilon, 1-\epsilon))}{\sum_{p=1}^{P} w_p^{\mathrm{F}}}\right)$.
Aggregate Kendall's τ_b using pair-count weights: $w_p^{\mathrm{K}} = \max\left\{\frac{n_p(n_p-1)}{2}, 0\right\}$, $\widehat{\tau}_{\mathrm{pc}}^{\mathrm{K}} = \frac{\sum_{p=1}^{P} w_p^{\mathrm{K}} \widehat{\tau}_p^{\mathrm{K}}}{\sum_{p=1}^{P} w_p^{\mathrm{K}}}$.
Aggregate covariance using degrees-of-freedom weights: $w_p^{\mathrm{Cov}} = \max\{n_p - 1, 0\}$, and $\widehat{\mathrm{Cov}}_{\mathrm{pc}} = \frac{\sum_{p=1}^{P} w_p^{\mathrm{Cov}} \widehat{\mathrm{Cov}}_p}{\sum_{p=1}^{P} w_p^{\mathrm{Cov}}}$.
return $\widehat{\rho}_{\mathrm{pc}}^{\mathrm{P}}$, $\widehat{\rho}_{\mathrm{pc}}^{\mathrm{S}}$, $\widehat{\tau}_{\mathrm{pc}}^{\mathrm{K}}$, $\widehat{\mathrm{Cov}}_{\mathrm{pc}}$.

Table D.1: Base–instruct comparison of length–entropy-rate coupling measured by Spearman's ρ and Kendall's τ. Each entry reports estimate $\pm$ standard error, and absolute change is computed as instruct minus base. All numbers are rounded to two significant digits. Warm colors indicate increases, while cold colors indicate decreases; darker colors correspond to larger magnitude changes.

Task	Type	ER vs len (ρ)			ER vs len (τ)		
		Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B
MATH	Base	-0.19 ± 0.023	-0.56 ± 0.031	-0.27 ± 0.023	-0.12 ± 0.015	-0.36 ± 0.024	-0.17 ± 0.015
	Instruct	0.039 ± 0.032	0.13 ± 0.029	0.096 ± 0.034	0.026 ± 0.020	0.082 ± 0.019	0.056 ± 0.021
	Abs. Change	0.23 ± 0.037	0.68 ± 0.046	0.36 ± 0.045	0.15 ± 0.024	0.44 ± 0.033	0.22 ± 0.028
CODING	Base	-0.28 ± 0.018	-0.66 ± 0.014	-0.40 ± 0.011	-0.18 ± 0.013	-0.47 ± 0.012	-0.26 ± 0.0079
	Instruct	0.49 ± 0.030	0.30 ± 0.023	0.38 ± 0.032	0.32 ± 0.021	0.20 ± 0.015	0.25 ± 0.022
	Abs. Change	0.76 ± 0.035	0.97 ± 0.022	0.78 ± 0.033	0.50 ± 0.024	0.67 ± 0.016	0.51 ± 0.023
SENTENCE COMPLETION	Base	-0.0049 ± 0.025	-0.39 ± 0.019	-0.32 ± 0.016	0.0026 ± 0.017	-0.26 ± 0.013	-0.22 ± 0.011
	Instruct	0.52 ± 0.024	-0.22 ± 0.025	0.13 ± 0.035	0.36 ± 0.018	-0.14 ± 0.017	0.086 ± 0.021
	Abs. Change	0.53 ± 0.033	0.17 ± 0.029	0.46 ± 0.037	0.35 ± 0.023	0.12 ± 0.019	0.30 ± 0.023
STORY	Base	-0.36 ± 0.015	-0.32 ± 0.015	-0.30 ± 0.014	-0.25 ± 0.011	-0.21 ± 0.011	-0.20 ± 0.0099
	Instruct	0.16 ± 0.021	-0.12 ± 0.019	-0.049 ± 0.027	0.10 ± 0.014	-0.081 ± 0.012	-0.032 ± 0.017
	Abs. Change	0.52 ± 0.025	0.20 ± 0.024	0.25 ± 0.032	0.35 ± 0.017	0.13 ± 0.016	0.16 ± 0.021

Table D.2: When to use BF vs. GenPPL across different tasks.

Task	**Preferred Metric**	**Level**	**Reason**
MMLU (QA)	GenPPL	Token	Requires selecting a single correct answer. Performance depends on calibrated token probabilities. Diversity across full trajectories is unnecessary.
Story Generation	BF	Trajectory	Outputs are evaluated as complete units. Diversity across stories reflects creativity and coverage of different narrative possibilities.
Math	GenPPL BF	Token Trajectory	GenPPL is primary because correctness depends on precise step-by-step reasoning. BF is secondary and useful for analyzing diversity across alternative solution strategies.
Coding	GenPPL BF	Token Trajectory	GenPPL is primary because code correctness requires token-level precision. BF is secondary and useful for comparing multiple valid implementations.
Sentence Completion	GenPPL	Token	This is a next-token prediction task, so GenPPL aligns directly with likelihood and perplexity-based evaluation.

Table D.3: Empirical Truncation Rates. We report the proportion of Monte Carlo rollouts that reached the maximum generation budget ($T_{\max} = 4000$ tokens) without sampling an `EOS` token. Rates are calculated over $P = 100$ prompts with $M = 100$ rollouts per prompt.

Model	MATH		STORY		CODING		COMPLETION	
	Base	Instruct	Base	Instruct	Base	Instruct	Base	Instruct
Qwen3-8B	0.0073	0.0000	0.0005	0.0000	0.0008	0.0004	0.0313	0.0000
Llama-3.1-8B	0.0105	0.0006	0.0325	0.0000	0.0095	0.0000	0.0174	0.0000
Gemma-3-12B	0.0036	0.0000	0.0066	0.0000	0.0432	0.0000	0.0119	0.0000

Table D.4: Full table of conditional mean model estimates for the Beta mixed-effects regression.

Term	Estimate	Std. Error	p-value
Intercept	-0.87953	0.07007	$< 2e{-}16$
R	0.70803	0.03340	$< 2e{-}16$
Instruct	-0.37650	0.08609	1.23e−05
$\text{ns}(\text{invN}, 4)_1$	0.15197	0.04085	0.000199
$\text{ns}(\text{invN}, 4)_2$	0.39690	0.23890	0.096641
$\text{ns}(\text{invN}, 4)_3$	0.62065	1.47024	0.672924
$\text{ns}(\text{invN}, 4)_4$	0.61022	3.10032	0.843966
Coding	-0.03721	0.03640	0.306726
Math	0.16948	0.03448	8.83e−07
Stories	-0.27563	0.02864	$< 2e{-}16$
$R \times$ Instruct	1.18254	0.04642	$< 2e{-}16$
Instruct $\times\text{ns}(\text{invN}, 4)_1$	-0.13029	0.07868	0.097734
Instruct $\times\text{ns}(\text{invN}, 4)_2$	1.35321	0.31701	1.97e−05
Instruct $\times\text{ns}(\text{invN}, 4)_3$	1.00650	1.48303	0.497340
Instruct $\times\text{ns}(\text{invN}, 4)_4$	0.81421	3.10203	0.792954

Table D.5: Estimated random intercept variances for prompt identity and model family in the Beta mixed-effects regression model.

Random effect	Group	Variance	Std. Dev.
Intercept	prompt_uid	0.013937	0.11806
Intercept	model_name	0.006817	0.08256

Table D.6: Dispersion model estimates for the Beta mixed-effects regression. The dispersion submodel uses a log link for the precision parameter ϕ_{tpm}.

Term	Estimate	Std. Error	p-value
Intercept	2.7544	0.2328	$< 2e{-}16$
Instruct	1.4153	0.2452	7.83e−09
Coding	2.5511	0.2186	$< 2e{-}16$
Math	2.1418	0.2149	$< 2e{-}16$
Stories	1.6534	0.1479	$< 2e{-}16$
$\text{ns}(\text{invN}, 3)_1$	-3.3780	0.5323	2.22e−10
$\text{ns}(\text{invN}, 3)_2$	-2.0039	0.8870	0.0239
$\text{ns}(\text{invN}, 3)_3$	2.6200	1.8507	0.1569
R	0.9751	0.1290	4.12e−14
LLaMA-3.1-8B	1.0091	0.1162	$< 2e{-}16$
Qwen3-8B	0.7270	0.1151	2.71e−10
Instruct × Coding	-2.1969	0.2852	1.32e−14
Instruct × Math	-1.5440	0.2861	6.76e−08
Instruct × Stories	-2.9273	0.2085	$< 2e{-}16$
Instruct $\times R$	-1.6085	0.2047	3.96e−15

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Arpit Agarwal, Shivani Agarwal, and Prathamesh Patil. Stochastic dueling bandits with adversarial corruption. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 217–248. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/agarwal21a.html.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

Gagan Aggarwal, Ashwinkumar Badanidiyuru, Santiago R Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Gagan Goel, Christopher Liaw, Haihao Lu, Mohammad Mahdian, et al. Auto-bidding and auctions in online advertising: A survey. *ACM SIGecom Exchanges*, 22(1):159–183, 2024.

Nir Ailon, Thorsten Joachims, and Zohar Karnin. Reducing dueling bandits to cardinal bandits, 2014.

Amazon Ads. Dynamic Bidding for Sponsored Products Guide, 2025. URL https://advertising.amazon.com/library/guides/dynamic-bidding-sponsored-products. Accessed: 2025-01-25.

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 736–749, 2022.

Foozhan Ataiefard and Hadi Hemmati. Gray-box adversarial attack of deep reinforcement learning-based trading agents. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 675–682. IEEE, 2023.

Donald R Atkinson, Bruce E Wampold, Susana M Lowe, Linda Matthews, and Hyun-Nie Ahn. Asian american preferences for counselor characteristics: Application of the bradley-terry-luce model to paired comparison data. *The Counseling Psychologist*, 26(1):101–123, 1998.

Wenjia Ba, Tianyi Lin, Jiawei Zhang, and Zhengyuan Zhou. Doubly optimal no-regret online learning in strongly monotone games with bandit feedback, 2024a.

Wenjia Ba, Tianyi Lin, Jiawei Zhang, and Zhengyuan Zhou. Doubly optimal no-regret online learning in strongly monotone games with bandit feedback, 2024b. URL https://arxiv.org/abs/2112.02856.

Ashwinkumar Badanidiyuru, Zhe Feng, and Guru Guruganesh. Learning to bid in contextual first price auctions, 2021.

Ashwinkumar Badanidiyuru, Zhe Feng, Tianxi Li, and Haifeng Xu. Incrementality bidding via reinforcement learning under mixed and delayed rewards, 2023. URL https://arxiv.org/abs/2206.01293.

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.

Pablo Ballesteros-Pérez and Martin Skitmore. On the distribution of bids for construction contract auctions. *Construction Management and Economics*, 35(3):106–121, 2017.

Gordon M Becker, Morris H DeGroot, and Jacob Marschak. Stochastic models of choice behavior. *Behavioral science*, 8(1):41–55, 1963.

Raoul Bell, Laura Mieth, and Axel Buchner. Coping with high advertising exposure: a source-monitoring perspective. *Cognitive Research: Principles and Implications*, 7(1):82, 2022.

E Veronica Belmega, Panayotis Mertikopoulos, Romain Negrel, and Luca Sanguinetti. Online convex optimization and no-regret learning: Algorithms, guarantees and applications. *arXiv preprint arXiv:1804.04529*, 2018.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021a.

Viktor Bengs, Robert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey, 2021b.

Carole Bernard, Xuedong He, Jia-An Yan, and Xun Yu Zhou. Optimal insurance design under rank-dependent expected utility. *Mathematical Finance*, 25(1):154–186, 2015.

doi:https://doi.org/10.1111/mafi.12027. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mafi.12027.

Avrim Blum, Meghal Gupta, Gene Li, Naren Sarayu Manoj, Aadirupa Saha, and Yuanyuan Yang. Dueling optimization with a monotone adversary, 2023.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks, 2020.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021a.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 991–999. PMLR, 13–15 Apr 2021b. URL https://proceedings.mlr.press/v130/bogunovic21a.html.

Ilija Bogunovic, Zihan Li, Andreas Krause, and Jonathan Scarlett. A robust phased elimination algorithm for corruption-tolerant gaussian process bandits, 2022.

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.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Mei Cai, Li Yan, Zaiwu Gong, and Guo Wei. A voting mechanism designed for talent shows in mass media: weighted preference of group decision makers in social networks using fuzzy measures and choquet integral. *Group Decision and Negotiation*, 30:1261–1284, 2021.

Yang Cai and Weiqiang Zheng. Doubly optimal no-regret learning in monotone games. *arXiv preprint arXiv:2301.13120*, 2023.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Ilayda Canyakmaz, Iosif Sakos, Wayne Lin, Antonios Varvitsiotis, and Georgios Piliouras. Steering game dynamics towards desired outcomes. *arXiv preprint arXiv:2404.01066*, 2024.

Christopher D. Carroll and Miles S. Kimball. On the concavity of the consumption function. *Econometrica*, 64(4):981–992, 1996. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2171853.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 61(1):549–564, 2014.

T Chatfield. The trouble with big data? it's called the 'recency bias'. *BBC. Retrieved April*, 28:2023, 2016.

Bangrui Chen and Peter I Frazier. Dueling bandits with weak regret. In *International Conference on Machine Learning*, pages 731–739. PMLR, 2017.

Haokun Chen, Xinyi Dai, Han Cai, Weinan Zhang, Xuejian Wang, Ruiming Tang, Yuzhou Zhang, and Yong Yu. Large-scale interactive recommendation with tree-structured policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3312–3320, 2019.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021a.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021b.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models, 2024.

Yuwei Cheng, Fan Yao, Xuefeng Liu, and Haifeng Xu. Learning from imperfect human feedback: a tale from corruption-robust dueling, 2024. URL https://arxiv.org/abs/2405.11204.

Pierre-André Chiappori and Jean-Charles Rochet. Revealed preferences and differentiable demand. *Econometrica*, 55(3):687–691, 1987. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913607.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

Subhasish M Chowdhury and Roman M Sheremeta. A generalized tullock contest. *Public Choice*, 147:413–420, 2011.

Wenchang Chu. Abel's method on summation by parts and balanced q-series identities. *Applicable Analysis and Discrete Mathematics*, pages 54–65, 2010.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.

David J Cooper and Hanming Fang. Understanding overbidding in second price auctions: An experimental study. *The Economic Journal*, 118(532):1572–1595, 2008.

Luis C Corchón, Marco Serena, et al. Contest theory. *Handbook of game theory and industrial organization*, 2:125–146, 2018.

Antoine Augustin Cournot. *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. L. Hachette, Paris, 1838. URL https://gallica.bnf.fr/ark:/12148/bpt6k8804262.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Annual Conference Computational Learning Theory*, 2008. URL https://api.semanticscholar.org/CorpusID:9134969.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably sample efficient rlhf via active preference optimization. *arXiv preprint arXiv:2402.10500*, 2024.

Evert De Haan, Thorsten Wiesel, and Koen Pauwels. The effectiveness of different forms of online advertising for purchase conversion in a multiple-channel attribution framework. *International journal of research in marketing*, 33(3):491–507, 2016.

André De Palma, Gordon M Myers, and Yorgos Y Papageorgiou. Rational choice under an imperfect ability to choose. *The American Economic Review*, pages 419–440, 1994.

Gerard Debreu. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893, 1952.

Yuan Deng, Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. Fairness in the autobidding world with machine-learned advice. *arXiv preprint arXiv:2209.04748*, 4, 2022.

Yash Deshpande and Andrea Montanari. Linear bandits in high dimension and recommendation systems, 2013.

Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits, 2023.

Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. *arXiv preprint arXiv:2404.10776*, 2024.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7111–7123. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/ding22c.html.

Jing Dong, Baoxiang Wang, and Yaoliang Yu. Uncoupled and convergent learning in monotone games under bandit feedback. *arXiv preprint arXiv:2408.08395*, 2024.

Juncheng Dong, Suya Wu, Mohammadreza Soltani, and Vahid Tarokh. Multi-agent adversarial attacks for multi-channel communications. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1580–1582, 2022.

David M Drukker, Ingmar R Prucha, and Rafal Raciborski. Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances. *The Stata Journal*, 13(2):221–241, 2013.

Yihan Du, Siwei Wang, and Longbo Huang. Dueling bandits: From two-dueling to multi-dueling, 2022.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2892–2902. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/duan21a.html.

Agupta et al. Knowledge distillation in large language models. 2023.

Eyal Even-dar, Yishay Mansour, and Uri Nadav. On the convergence of regret minimization dynamics in concave games. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 523–532, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585062. doi:10.1145/1536414.1536486. URL https://doi.org/10.1145/1536414.1536486.

Eyal Even-Dar, Yishay Mansour, and Uri Nadav. On the convergence of regret minimization dynamics in concave games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 523–532, 2009.

Christian Ewerhart. The lottery contest is a best-response potential game. *Economics Letters*, 155:168–171, 2017. ISSN 0165-1765. doi:https://doi.org/10.1016/j.econlet.2017.03.030. URL https://www.sciencedirect.com/science/article/pii/S0165176517301325.

Facebook Ads. About Automated Bidding on Facebook Ads, 2025. URL https://www.facebook.com/business/help/1619591734742116?id=2196356200683573. Accessed: 2025-01-25.

Ky Fan. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proceedings of the National Academy of Sciences*, 38(2):121–126, 1952.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. *Advances in Neural Information Processing Systems*, 35:39076–39089, 2022.

Yiding Feng, Ronen Gradwohl, Jason Hartline, Aleck Johnsen, and Denis Nekipelov. Bias-variance games, 2022.

Zhe Feng, Chara Podimata, and Vasilis Syrgkanis. Learning to bid without knowing your value. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 505–522, 2018.

Zhe Feng, Christopher Liaw, and Zixin Zhou. Improved online learning algorithms for ctr prediction in ad auctions. In *International Conference on Machine Learning*, pages 9921–9937. PMLR, 2023.

Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

Jorge I Figueroa-Zúñiga, Reinaldo B Arellano-Valle, and Silvia LP Ferrari. Mixed beta regression: A bayesian perspective. *Computational Statistics & Data Analysis*, 61:137–147, 2013.

Martin Figura, Krishna Chaitanya Kosaraju, and Vijay Gupta. Adversarial attacks in consensus-based multi-agent reinforcement learning. In *2021 American Control Conference (ACC)*, pages 3050–3055, 2021. doi:10.23919/ACC50511.2021.9483080.

Peter C Fishburn, Peter C Fishburn, et al. *Utility theory for decision making*. Krieger NY, 1979.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857.

Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making, 2023.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Yuwei Fu, Wu Di, and Benoit Boulet. Batch reinforcement learning in the real world: A survey. In *Offline RL Workshop, NeuroIPS*, 2020.

Jiarui Gan, Adish Singla, and Goran Radanovic. Envy-free policy teaching to multiple agents. *Advances in Neural Information Processing Systems*, 35:13290–13302, 2022.

Haoxiang Gao, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. *Advances in Neural Information Processing Systems*, 33:14362–14373, 2020.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint, 2022.

Abheek Ghosh and Paul W Goldberg. Best-response dynamics in lottery contests. *arXiv preprint arXiv:2305.10881*, 2023.

Irving L Glicksberg. A further generalization of the kakutani fixed theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

Google Ads Attribution. Guide To Google Ads Attribution Models in 2025 – Is Data-Driven Attribution The Future?, 2025. URL https://www.datafeedwatch.com/blog/google-ads-attribution-models. Accessed: 2025-05-11.

Google Ads Support. About Automated Bidding in Google Ads, 2025. URL https://support.google.com/google-ads/answer/2979071?hl=en. Accessed: 2025-01-25.

Google Developers. Authorized buyers real-time bidding (rtb) protocol guide, 2024. URL https://developers.google.com/authorized-buyers/rtb/openrtb-guide. Accessed: 2024-01-30.

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225, 2019.

Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. *arXiv preprint arXiv:2402.08567*, 2024.

Rui Guo and Zhengrui Jiang. Optimal dynamic advertising policy considering consumer ad fatigue. *Decision Support Systems*, 187:114323, 2024.

Wenbo Guo, Xian Wu, Sui Huang, and Xinyu Xing. Adversarial policy learning in two-player competitive games. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3910–3919. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/guo21b.html.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1562–1578. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/gupta19a.html.

William S Hall and Martin L Newell. The mean value theorem for vector valued functions: a simple proof. *Mathematics Magazine*, 52(3):157–158, 1979.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Yanjun Han, Tsachy Weissman, and Zhengyuan Zhou. Optimal no-regret learning in repeated first-price auctions. *Operations Research*, 2024a.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024b.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

Florian Hartig. Dharma: residual diagnostics for hierarchical (multi-level/mixed) regression models. *CRAN: contributed packages*, 2016.

Elad Hazan and Kfir Levy. Bandit convex optimization: Towards tight bounds. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions, 2022.

Yu He, Fan Yao, Yang Yu, Xiaoyun Qiu, Minming Li, and Haifeng Xu. The complexity of tullock contests. *arXiv preprint arXiv:2412.06444*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ICLR*, 2020.

Edward J Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. *arXiv preprint arXiv:2310.04363*, 2023.

Yizheng Hu and Zhihua Zhang. Sparse adversarial attack in multi-agent reinforcement learning. *arXiv e-prints*, pages arXiv–2205, 2022.

Jiawei Huang, Vinzenz Thoma, Zebang Shen, Heinrich H Nax, and Niao He. Learning to steer markovian agents under model uncertainty. *arXiv preprint arXiv:2407.10207*, 2024.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. *arXiv preprint arXiv:1811.06026*, 2018.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 647–648, 2020.

Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke van der Plas, and Roger E. Beaty. Creative preference optimization, 2025. URL https://arxiv.org/abs/2505.14442.

ITA. E-commerce Sales Size and Forecast, 2025. URL https://www.trade.gov/ecommerce-sales-size-forecast. Accessed: 2025-01-25.

Meena Jagadeesan, Michael Jordan, Jacob Steinhardt, and Nika Haghtalab. Improved bayes risk can yield reduced social welfare under competition. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 66940–66952. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d3602fc92fb8b9e0d55356c9e8815e2b-Paper-Conference.pdf.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Popart: Efficient sparse regression and experimental design for optimal sparse linear bandits, 2023.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing with reusable resources: Linear bandits with sub-exponential rewards. 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 10135–10160. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/jia22c.html.

Yiling Jia, Weitong Zhang, Dongruo Zhou, Quanquan Gu, and Hongning Wang. Learning neural contextual bandits through perturbed rewards. *arXiv preprint arXiv:2201.09910*, 2022b.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 2193–2201, 2018.

Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *Advances in Neural Information Processing Systems*, 35:33469–33481, 2022.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. *Advances in neural information processing systems*, 31, 2018.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914185.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Robust lipschitz bandits to adversarial corruptions. *Advances in Neural Information Processing Systems*, 36, 2024.

Kawaljeet Kaur Kapoor, Yogesh K Dwivedi, and Niall C Piercy. Pay-per-click advertising: A literature review. *The Marketing Review*, 16(2):183–202, 2016.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Frank P Kelly, Aman K Maulloo, and David Kim Hong Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49:237–252, 1998.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Johannes Kirschner and Andreas Krause. Bias-robust bayesian optimization via dueling bandits. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5595–5605. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kirschner21a.html.

Vinnie Ko and Nils Lid Hjort. Model robust inference with two-stage maximum likelihood estimation for copulas. *Journal of Multivariate Analysis*, 171:362–381, 2019.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pages 1141–1154. PMLR, 2015.

Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

Nikki Lijing Kuang, Ming Yin, Mengdi Wang, Yu-Xiang Wang, and Yian Ma. Posterior sampling with delayed feedback for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 36:6782–6824, 2023.

Wataru Kumagai. Regret analysis for continuous dueling bandit. *Advances in Neural Information Processing Systems*, 30, 2017.

Jean-Jacques Laffont, Herve Ossard, and Quang Vuong. Econometrics of first-price auctions. *Econometrica: Journal of the Econometric Society*, pages 953–980, 1995.

Thom Lake, Eunsol Choi, and Greg Durrett. From distributional to overton pluralism: Investigating large language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6794–6814, 2025.

Vicki R Lane. The impact of ad repetition and ad content on consumer perceptions of incongruent extensions. *Journal of Marketing*, 64(2):80–91, 2000.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi:10.1017/9781108571401.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3703–3712. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/le19a.html.

R Leo, RS Milton, and A Kaviya. Multi agent reinforcement learning based distributed optimization of solar microgrid. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–7. IEEE, 2014.

Jonathan Levin. Supermodular games. *Lectures Notes, Department of Economics, Stanford University*, 2003.

Orin Levy and Yishay Mansour. Learning efficiently function approximation for contextual mdp, 2022. URL https://arxiv.org/abs/2203.00995.

Orin Levy and Yishay Mansour. Optimism in face of a context:regret guarantees for stochastic contextual mdp. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8510–8517, Jun. 2023. doi:10.1609/aaai.v37i7.26025. URL https://ojs.aaai.org/index.php/AAAI/article/view/26025.

Orin Levy, Alon Cohen, Asaf Cassel, and Yishay Mansour. Efficient rate optimal regret for adversarial contextual MDPs using online function approximation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19287–19314. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/levy23a.html.

Randall Lewis and Jeffrey Wong. Incrementality bidding and attribution, 2022. URL https://arxiv.org/abs/2208.12809.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119, 2016.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Chuang-Chieh Lin and Chi-Jen Lu. Efficient mechanisms for peer grading and dueling bandits. In *Asian Conference on Machine Learning*, pages 740–755. PMLR, 2018.

Jieyu Lin, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. On the robustness of cooperative multi-agent reinforcement learning. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 62–68. IEEE, 2020.

Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050. PMLR, 2019.

Guanlin Liu and Lifeng Lai. Efficient adversarial attacks on online multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:24401–24433, 2023.

Hao Liu, Yunze Li, Qinyu Cao, Guang Qiu, and Jiming Chen. Estimating individual advertising effect in e-commerce. *arXiv preprint arXiv:1903.04149*, 2019.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

Wenjie Liu, Lidong Li, Jian Sun, Fang Deng, Gang Wang, and Jie Chen. Online data-driven control against false data injection attacks, 2023.

Wei Lu, Rachel K Luu, and Markus J Buehler. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(1):84, 2025.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions, 2018.

Yuzhe Ma, Young Wu, and Xiaojin Zhu. Game redesign in no-regret game playing, 2021. URL https://arxiv.org/abs/2110.11763.

Omer Madmon, Idan Pipano, Itamar Reinman, and Moshe Tennenholtz. On the convergence of no-regret dynamics in information retrieval games with proportional ranking functions. *arXiv preprint arXiv:2405.11517*, 2024.

Binita Manandhar. Effect of advertisement in consumer behavior. *Management Dynamics*, 21(1):46–54, 2018.

Debmalya Mandal, Andi Nika, Parameswaran Kamalaruban, Adish Singla, and Goran Radanović. Corruption robust offline reinforcement learning with human feedback, 2024.

Henry B Mann and Abraham Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Lucas Maystre and Matthias Grossglauser. Just sort it! a simple and effective approach to active preference learning. In *International Conference on Machine Learning*, pages 2344–2353. PMLR, 2017.

Jeremy McMahan, Young Wu, Xiaojin Zhu, and Qiaomin Xie. Optimal attack and defense for reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14332–14340, Mar. 2024. doi:10.1609/aaai.v38i13.29346. URL https://ojs.aaai.org/index.php/AAAI/article/view/29346.

Clara Meister and Ryan Cotterell. Typical decoding for natural language generation. *TACL*, 2023.

Microsoft Ads. Automated Bidding Tools for Performance Optimization, 2025. URL https://about.ads.microsoft.com/en/tools/performance/automated-bidding. Accessed: 2025-01-25.

Aditya Modi and Ambuj Tewari. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 829–838. PMLR, 2020.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

Mohammad Mohammadi, Jonathan Nöther, Debmalya Mandal, Adish Singla, and Goran Radanovic. Implicit poisoning attacks in two-agent reinforcement learning: Adversarial policies for training-time attacks. In *22nd International Conference on Autonomous Agents and Multiagent Systems*, pages 1835–1844. IFAAMAS, 2023.

Luigi Montrucchio. Lipschitz continuous policy functions for strongly concave optimization problems. *Journal of mathematical economics*, 16(3):259–273, 1987.

Jamie Murphy, Charles Hofacker, and Richard Mizerski. Primacy and recency effects on clicking behavior. *Journal of computer-mediated communication*, 11(2):522–535, 2006.

Abraham Neyman. Correlated equilibrium and potential games. *International Journal of Game Theory*, 26(2):223–227, 1997.

Andi Nika, Debmalya Mandal, Adish Singla, and Goran Radanovic. Corruption-robust offline two-player zero-sum markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR, 2024a.

Andi Nika, Debmalya Mandal, Adish Singla, and Goran Radanović. Corruption-robust offline two-player zero-sum markov games, 2024b.

Ann L Oberg and Douglas W Mahoney. Linear mixed effects models. *Topics in biostatistics*, pages 213–234, 2007.

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.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishers, San Mateo, CA, 1988.

Cornelia Pechmann and David W Stewart. Advertising repetition: A critical review of wearin and wearout. *Current issues and research in advertising*, 11(1-2):285–329, 1988.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gloria Phillips-Wren, Daniel J Power, and Manuel Mora. Cognitive bias, decision styles, and risk attitudes in decision making and dss, 2019.

David Pollard. *Empirical processes: Theory and applications*. Inst. of Math. Statistics u.a., 1990.

Richard A Posner. Rational choice, behavioral economics, and the law. *StAn. l. reV.*, 50: 1551, 1997.

Pearl Pu, Li Chen, and Rong Hu. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22: 317–355, 2012.

Dawei Qiu, Jianhong Wang, Junkai Wang, and Goran Strbac. Multi-agent reinforcement learning for automated peer-to-peer energy trading in double-side auction market. In *IJCAI*, pages 2913–2920, 2021.

Quartile. Quartile Product Overview, 2025. URL https://www.quartile.com/product. Accessed: 2025-01-25.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 235–247. Springer, 1992.

Danda B. Rawat, Laurent Njilla, Kevin Kwiat, and Charles Kamhoua. ishare: Blockchain-based privacy-aware multi-agent information sharing games for cybersecurity. In *2018 International Conference on Computing, Networking and Communications (ICNC)*, pages 425–431, 2018. doi:10.1109/ICCNC.2018.8390264.

Kan Ren, Weinan Zhang, Ke Chang, Yifei Rong, Yong Yu, and Jun Wang. Bidding machine: Learning to bid for directly optimizing profits in display advertising. *IEEE Transactions on Knowledge and Data Engineering*, 30(4):645–659, 2017.

Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits, 2022.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

Tim Roughgarden and Florian Schoppmann. Local smoothness and the price of anarchy in splittable congestion games. *Journal of Economic Theory*, 156:317–342, 2015.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Walter Rudin. Principles of mathematical analysis. 2021.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30050–30062. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fc3cf452d3da8402bebb765225ce8c0e-Paper.pdf.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021b.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. 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 19011–19026. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saha22a.html.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Dueling convex optimization with general preferences, 2022. URL https://arxiv.org/abs/2210.02562.

Virgilijus Sakalauskas and Dalia Kriksciuniene. Personalized advertising in e-commerce: Using clickstream data to target high-value customers. *Algorithms*, 17(1):27, 2024.

Pier Giuseppe Sessa, Ilija Bogunovic, Andreas Krause, and Maryam Kamgarpour. Contextual games: Multi-agent learning with side information, 2021.

Mian Ibad Ali Shah, Abdul Wahid, Enda Barrett, and Karl Mason. Multi-agent systems in peer-to-peer energy trading: A comprehensive survey. *Engineering Applications of Artificial Intelligence*, 132:107847, 2024.

Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*, 2024.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization, 2013.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948a. doi:10.1002/j.1538-7305.1948.tb01338.x.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948b.

Han Shao, Lee Cohen, Avrim Blum, Yishay Mansour, Aadirupa Saha, and Matthew Walter. Eliciting user preferences for personalized multi-objective decision making through comparative feedback. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 12192–12221. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/286e7ab0ce6a68282394c92361c27b57-Paper-Conference.pdf.

Han Shao, Lee Cohen, Avrim Blum, Yishay Mansour, Aadirupa Saha, and Matthew Walter. Eliciting user preferences for personalized multi-objective decision making through comparative feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Aizaz Sharif and Dusica Marijan. Adversarial deep reinforcement learning for improving the robustness of multi-agent autonomous driving policies. In *2022 29th Asia-Pacific Software Engineering Conference (APSEC)*, pages 61–70. IEEE, 2022.

Lei Shi, Jingshen Wang, and Tianhao Wu. Statistical inference on multi-armed bandits with delayed feedback. In *International Conference on Machine Learning*, pages 31328–31352. PMLR, 2023.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Does instruction tuning reduce diversity? a case study using code generation.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Maurice Sion. On general minimax theorems. 1958.

Martin Skitmore. First and second price independent values sealed bid procurement auctions: some scalar equilibrium results. *Construction Management and Economics*, 26(8):787–803, 2008.

Eric Smith, J Doyne Farmer, László Gillemot, and Supriya Krishnamurthy. Statistical theory of the continuous double auction. *Quantitative finance*, 3(6):481, 2003.

Michael Smithson and Jay Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54, 2006.

Gerhard Sorger. On the sensitivity of optimal growth paths. *Journal of Mathematical Economics*, 24(4):353–369, 1995.

Charles Spearman. The proof and measurement of association between two things. 1961.

Spotify. Music recommendation system using spotify dataset. https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset, 2020.

Richard Stengel. *Information wars: How we lost the global battle against disinformation & what we can do about it*. Atlantic Monthly Press, 2019.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Yanan Sui, Masrour Zoghi, Katja Hofmann, and Yisong Yue. Advancements in dueling bandits. In *IJCAI*, pages 5502–5510, 2018.

Sabine Swoboda. Confidentiality for the protection of national security interests. *Revue internationale de droit pénal*, 81(1):209–229, 2010.

Jing Tan, Ramin Khalili, and Holger Karl. Learning to bid long-term: Multi-agent reinforcement learning with long-term and sparse reward in repeated auction games. *arXiv preprint arXiv:2204.02268*, 2022.

Tatiana Tatarenko and Maryam Kamgarpour. Bandit learning in convex non-strictly monotone games. *arXiv preprint arXiv:2009.04258*, 2020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, 2021.

Topsort. Auto-bidding Technology, 2025. URL https://www.topsort.com/product/tech/autobidding. Accessed: 2025-01-25.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.

James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7768–7777, 2021.

Gordon Tullock. Efficient rent seeking. In *40 Years of Research on Rent Seeking 1*, pages 105–120. Springer, 2008.

J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Demetrios Vakratsas and Tim Ambler. How advertising works: what do we really know? *Journal of marketing*, 63(1):26–43, 1999.

Alain Venditti. Strong concavity properties of indirect utility functions in multisector optimal growth models. *Journal of Economic Theory*, 74(2):349–367, 1997a. ISSN 0022-0531. doi:https://doi.org/10.1006/jeth.1996.2267. URL https://www.sciencedirect.com/science/article/pii/S002205319692267X.

Alain Venditti. Strong concavity properties of indirect utility functions in multisector optimal growth models. *Journal of Economic Theory*, 74(2):349–367, 1997b.

Roman Vershynin. *High-dimensional probability: An introduction with applications in Data Science*. Cambridge University Press, 2018.

Martin J. Wainwright. *High-dimensional statistics a non-asymptotic viewpoint*. Cambridge University Press, 2019a.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019b.

Yuanyu Wan, Wei-Wei Tu, and Lijun Zhang. Online strongly convex optimization with unknown delays. *Machine Learning*, 111(3):871–893, 2022.

Chaoqi Wang, Ziyu Ye, Zhe Feng, Ashwinkumar Badanidiyuru, and Haifeng Xu. Follow-ups also matter: Improving contextual bandits via post-serving contexts, 2023.

Jun Wang, Li Zhang, Yanjun Huang, and Jian Zhao. Safety of autonomous vehicles. *Journal of advanced transportation*, 2020(1):8867757, 2020.

Shumin Wang, Yuexiang Xie, Wenhao Zhang, Yuchang Sun, Yanxi Chen, Yaliang Li, and Yanyong Zhang. On the entropy dynamics in reinforcement fine-tuning of large language models. *arXiv preprint arXiv:2602.03392*, 2026.

Tengyun Wang, Haizhi Yang, Han Yu, Wenjun Zhou, Yang Liu, and Hengjie Song. A revenue-maximizing bidding strategy for demand-side platforms. *IEEE Access*, 7:68692–68706, 2019.

Yichen Wang, Chenghao Yang, Tenghao Huang, Muhao Chen, Jonathan May, and Mina Lee. Optimizing diversity and quality through base-aligned model collaboration. *arXiv preprint arXiv:2511.05650*, 2025.

Yuanhao Wang, Dingwen Kong, Yu Bai, and Chi Jin. Learning rationalizable equilibria in multiplayer games. *arXiv preprint arXiv:2210.11402*, 2022.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1562–1583, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/weed16.html.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*, 2025.

Guillem Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generation. *TACL*, 2022.

WE Williams. A class of integral equations. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 59, pages 589–597. Cambridge University Press, 1963.

Robert Wilson. Sequential equilibria of asymmetric ascending auctions: The case of log-normal distributions. *Economic Theory*, 12:433–440, 1998.

Robert C Wilson, Andra Geana, John M White, Elliot A Ludvig, and Jonathan D Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of experimental psychology: General*, 143(6):2074, 2014.

Jibang Wu, Haifeng Xu, and Fan Yao. Uncoupled bandit learning towards rationalizability: Benchmarks, barriers, and algorithms. *arXiv preprint arXiv:2111.05486*, 2021.

Jibang Wu, Haifeng Xu, and Fan Yao. Multi-agent learning for iterative dominance elimination: Formal barriers and new algorithms. In *COLT*, page 543, 2022.

Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Reward poisoning attacks on offline multi-agent reinforcement learning, 2023. URL https://arxiv.org/abs/2206.01888.

Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Data poisoning to fake a nash equilibria for markov games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15979–15987, Mar. 2024. doi:10.1609/aaai.v38i14.29529. URL https://ojs.aaai.org/index.php/AAAI/article/view/29529.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference*

on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 11404–11413. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xie21d.html.

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, 2024.

Wenjie Xu, Wenbin Wang, Yuning Jiang, Bratislav Svetozarevic, and Colin N Jones. Principled preferential bayesian optimization. *arXiv preprint arXiv:2402.05367*, 2024.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

Wanqi Xue, Wei Qiu, Bo An, Zinovi Rabinovich, Svetlana Obraztsova, and Chai Kiat Yeo. Mis-spoke or mis-lead: Achieving robustness in multi-agent communicative reinforcement learning. *arXiv preprint arXiv:2108.03803*, 2021.

Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. Prefrec: Recommender systems with human preferences for reinforcing long-term user engagement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2874–2884, 2023.

Xinyi Yan, Chengxi Luo, Charles LA Clarke, Nick Craswell, Ellen M Voorhees, and Pablo Castells. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 567–577, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Chenghao Yang, Sida Li, and Ari Holtzman. Llm probability concentration: How alignment shrinks the generative horizon. *arXiv preprint arXiv:2506.17871*, 2025b.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Learning from a learning user for optimal recommendations. In *International Conference on Machine Learning*, pages 25382–25406. PMLR, 2022a.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Learning the optimal recommendation from explorative users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9457–9465, 2022b.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Human vs. generative ai in content creation competition: Symbiosis or conflict? In *International Conference on Machine Learning*. PMLR, 2024a.

Fan Yao, Yiming Liao, Jingzhou Liu, Shaoliang Nie, Qifan Wang, Haifeng Xu, and Hongning Wang. Unveiling user satisfaction and creator productivity trade-offs in recommendation platforms. *Advances in Neural Information Processing Systems*, 2024b.

Fan Yao, Yiming Liao, Mingzhe Wu, Chuanhao Li, Yan Zhu, James Yang, Qifan Wang, Haifeng Xu, and Hongning Wang. User welfare optimization in recommender systems with competing content creators. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024c.

Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pages 39834–39863. PMLR, 2023.

Chenlu Ye, Jiafan He, Quanquan Gu, and Tong Zhang. Towards robust model-based reinforcement learning against adversarial corruption, 2024a.

Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes, 2024b.

Soeun You, Taeha Kim, and Hoon S Cha. The smartphone user's dilemma among personalization, privacy, and advertisement fatigue: An empirical examination of personalized smartphone advertisement. *Information Systems Review*, 17(2):77–100, 2015.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1201–1208, New York, NY, USA, 2009a. Association for Computing Machinery. ISBN 9781605585161. doi:10.1145/1553374.1553527. URL https://doi.org/10.1145/1553374.1553527.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *International Conference on Machine Learning*, 2009b. URL https://api.semanticscholar.org/CorpusID:13141692.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl, 2021.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Qingli Zeng and Farid Nait-Abdesselam. Multi-agent reinforcement learning-based extended boid modeling for drone swarms. In *ICC 2024 - IEEE International Conference on Communications*, pages 1551–1556, 2024. doi:10.1109/ICC51166.2024.10622479.

Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. When demonstrations meet generative world models: A maximum likelihood framework for offline inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 36:65531–65565, 2023.

Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Alexander Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Steering no-regret learners to a desired equilibrium. *arXiv preprint arXiv:2306.05221*, 2023.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

Mengxiao Zhang and Haipeng Luo. Online learning in contextual second-price pay-per-click auctions, 2023.

Mengxiao Zhang and Haipeng Luo. Online learning in contextual second-price pay-per-click auctions. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2395–2403. PMLR, 02–04 May 2024a. URL https://proceedings.mlr.press/v238/zhang24f.html.

Mengxiao Zhang and Haipeng Luo. Online learning in contextual second-price pay-per-click auctions. In *International Conference on Artificial Intelligence and Statistics*, pages 2395–2403. PMLR, 2024b.

Wei Zhang, Yanjun Han, Zhengyuan Zhou, Aaron Flores, and Tsachy Weissman. Leveraging the hints: Adaptive bidding in repeated first-price auctions. *Advances in Neural Information Processing Systems*, 35:21329–21341, 2022.

Haoyu Zhao and Wei Chen. Online second price auction with semi-bandit feedback under the non-stationary setting. *arXiv preprint arXiv:1911.05949*, 2019.

Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. Smarts: An open-source scalable multi-agent rl training school for autonomous driving. In *Conference on robot learning*, pages 264–285. PMLR, 2021.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2023a.

Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline policy evaluation and optimization with heavy-tailed rewards. *arXiv preprint arXiv:2310.18715*, 2023b.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*, 2024.

Julian Zimmert and Yevgeny Seldin. Factored bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. URL http://jmlr.org/papers/v22/19-753.html.

Jinhang Zuo, Zhiyao Zhang, Xuchuang Wang, Cheng Chen, Shuai Li, John C. S. Lui, Mohammad Hajiesmaili, and Adam Wierman. Adversarial attacks on cooperative multi-agent bandits, 2023. URL https://arxiv.org/abs/2311.01698.

Shiliang Zuo. Near optimal adversarial attacks on stochastic bandits and defenses with smoothed responses. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2098–2106. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/zuo24a.html.