Causal Logistic Bandits with Counterfactual Fairness Constraints

Jiajun Chen [1] **Jin Tian** [2] **Christopher John Quinn** [1]

Abstract

Artificial intelligence will play a significant role in decision making in numerous aspects of society. Numerous fairness criteria have been proposed in the machine learning community, but there remains limited investigation into fairness as defined through specified attributes in a sequential decision-making framework. In this paper, we focus on causal logistic bandit problems where the learner seeks to make fair decisions, under a notion of fairness that accounts for counterfactual reasoning. We propose and analyze an algorithm by leveraging primal-dual optimization for constrained causal logistic bandits where the non-linear constraints are a priori unknown and must be learned in time. We obtain sub-linear regret guarantees with leading term similar to that for unconstrained logistic bandits (Lee et al., 2024) while guaranteeing sub-linear constraint violations. We show how to achieve zero cumulative constraint violations with a small increase in the regret bound.

1. Introduction

Artificial intelligence (AI) models, using techniques from statistics and machine learning, are increasingly being used to make affect people's lives. In light of this, a plethora of formal fairness criteria have been proposed (Darlington, 1971; Dwork et al., 2012; Hardt et al., 2016; Zhang et al., 2016; Kusner et al., 2017; Zafar et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Chouldechova & Roth, 2020; Imai & Jiang, 2023; Plecko & Bareinboim, 2024). There has been growing interest in the sequential decision-making community for accounting for fairness, including in settings such as classic and contextual bandits (Joseph et al., 2018), combinatorial bandits (Xu et al., 2020), bandits with long-term constraints (Liu et al., 2022), and reinforcement learning (Jabbari et al., 2017), just to name a few. Notably, rather than addressing fairness through the lens of specified attributes, these studies typically operationalize fairness in a different manner by defining it with respect to one-step rewards and introducing a notion of meritocratic fairness (Joseph et al., 2018). An algorithm should never assign a higher selection probability to a less qualified decision than to a more qualified one, i.e., arms with higher empirical rewards should be picked more frequently than those with lower empirical rewards, which is distinguishable from the fairness criteria based on specified attribute.

In this paper, we focus on a problem structure wherein arms arrive in a sequential and stochastic manner from an underlying fixed distribution and decisions are made in an online fashion by the agent. The objective of the agent is to optimize cumulative rewards while achieving fairness *counterfactually* with respect to specified attributes, i.e. the outcome would not have been substantially different if the specified attributes had different values. In general, this type of task belongs to the setting of dynamic treatment regimes (Murphy, 2003; Lavori & Dawson, 2008; Zhang, 2020) for finding a sequence of decisions over a finite set of treatments which appears across a broad range of applications.

1.1. Our Contributions

In light of the above, the goal of this paper is to analyze the foundations of online causal fair decision-making. More specially, our contributions are as follows:

- We formulate a constrained causal logistic bandits problem where the online decision-making processes are characterized within a causal structure. We formalize a (non-linear) fairness constraint based on the counterfactual outcome effect that is a priori unknown and must be learned in time. To the best of our knowledge, this is the first work to study constrained logistic bandits without a known safe decision subset (see Footnote 1).
- We provide an unified analysis for the confidence set construction, algorithm design, and performance guarantee, i.e., sublinear reward regret and sublinear cumulative constraint violations by leveraging the regret-to-confidence-set conversion and the primal-dual online

[1]Department of Computer Science, Iowa State University, Ames, IA, USA [2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Correspondence to: Christopher John Quinn <cjquinn@iastate.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Table 1: Comparison of reward model, constraint types, frequentist regret guarantees, and cumulative violations upper bounds for select related works. Notation: horizon T, the dimension for arm feature vector n, bounded bandit parameter S, truncated parameter ρ (see Section 3.2), Slater's constant δ (see Assumption 3), decision-set-dependent term $R_{\mathcal{Z}}(T)$, generalized linear model (GLM). Problem dependent constant $\boldsymbol{\kappa}_*, \boldsymbol{\kappa}_{\mathcal{Z}}, \boldsymbol{\kappa}$, where $\sqrt{1/\boldsymbol{\kappa}_*} < 1 \ll \boldsymbol{\kappa}_{\mathcal{Z}} \ll \boldsymbol{\kappa}$ when compared within the same decision and parameter spaces. † Requires prior knowledge (see Footnote 1) of a safe action/policy (i.e. satisfies the constraint $\Delta_{\pi_0} \leq \tau$) (per-round zero constraint violations with high probability).

Algorithm	Model	Constraint	Regret	Violation
Safe UCB-GLM (Amani et al., 2020)	GLM	GLM	$\widetilde{\mathcal{O}}\big(\boldsymbol{\kappa} n\sqrt{T}\big)$	0 †
OFULog-r (Abeille et al., 2021)	Logistic	No	$\widetilde{\mathcal{O}}\big(nS^{\frac{5}{2}}\sqrt{\frac{T}{\boldsymbol{\kappa}_*(T)}} + \min\{n^2S^4\boldsymbol{\kappa}_{\mathcal{Z}}, R_{\mathcal{Z}}(T)\}\big)$	–
F-UCB (Huang et al., 2022b)	Causal MAB	Linear	$\mathcal{O}\Big(\frac{1}{\tau-\Delta_{\pi_0}}\sqrt{\lvert W\rvert T}\Big)$	0 †
OFULog+ (Lee et al., 2024)	Logistic	No	$\widetilde{\mathcal{O}}\Big(nS\sqrt{\frac{T}{\boldsymbol{\kappa}_*(T)}} + \min\{n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}, R_{\mathcal{Z}}(T)\}\Big)$	–
CCLB Theorem 1	Causal logistic	Mixture logistics	$\widetilde{\mathcal{O}}\Big(\rho nS\sqrt{T} + \rho n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}} + \rho\sqrt{T}\Big)$	$\widetilde{\mathcal{O}}\Big(nS\sqrt{T} + n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}} + \sqrt{T}\Big)$
Zero_CCLB Proposition 4	Causal logistic	Mixture logistics	$\widetilde{\mathcal{O}}\Big(\rho nS\sqrt{T} + \rho n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}} + \rho\sqrt{T}(1+\delta)^2\Big)$	0

optimization. We show that the leading term of our regret, $\widetilde{\mathcal{O}}(\rho nS\sqrt{T})$, is significantly better than regret bounds for related works handling constraints and is similar to the bound for the unconstrained problem (Lee et al., 2024) (see Remark 2). Furthermore, by introducing an user-chosen parameter, one can trade off the regret slightly to achieve zero cumulative constraint violations.

1.2. Related Work

We next briefly discuss two lines of literature closely related to our work. See Appendix B for more discussion on additional works.

Firstly, in terms of formulating causal fairness within a multi-armed bandit setting, the closest related work is (Huang et al., 2022b). Like us, they considered a stochastic, contextual MAB problem with a (known) causal graph governing relationships between the stochastic contexts (seen by the learner before making decisions) and the rewards. They assume all variables are discrete. Like us, they proposed characterizing fairness with counterfactual fairness (Kusner et al., 2017; Wu et al., 2019; Chiappa, 2019) w.r.t. specified attributes in the context (e.g. specified user features in an online recommendation system). They make an assumption[1] about a fair policy; they provide high probability guarantees that all actions are fair. We model fairness as a long-term constraint, for which we seek to bound cumulative violations, as it is unclear whether it is possible to certify policies as fair (feasible) before collecting data to estimate the reward parameter upon which the (non-linear) constraint depends. Unlike our work, they considered that all variables except the reward are discrete-valued with non-parametric (thus more flexible) distributions. They proposed simpler empirical estimation methods for rewards, for which the counterfactual constraints became linear. While the structural causal model was discrete-valued but non-parametric, their regret bound in turn depended on $|W|$, the number of realizations of the set of parent variables of the reward, which is exponential in the size of the parent set (see Table 1). In contrast, we model rewards parametrically (using a logistic model), depending on feature maps of the context and decision variables. This dramatically improves the dimensional dependence, though the fairness constraint becomes a mixture of logistic functions for which estimating confidence bounds (to estimate region of fair actions) is more challenging.

Among works on MAB with parametric rewards and unknown (stochastic) constraints, there are numerous works on logistic rewards without constraints and linear rewards with linear constraints (see Appendix B for discussion on those works). The only prior work that like us considered a non-linear (parametric) reward model with non-linear un-

[1] Pacchiano et al. (2021) (which Huang et al. (2022b)'s analysis is based on) requires explicit a priori knowledge of a feasible action/policy (Assumption 5) and states that it is "absolutely necessary" to do so for the problem they study (Remark 1). Huang et al. (2022b)'s Assumption 3 only requires the existence of a safe policy π_0; π_0 is not explicitly used in estimating rewards or estimating a set of feasible policies in the main paper. However, to the best of our knowledge it is unclear how the conservatively estimated sets of policies $\bar{\Phi}_t$ (shown w.h.p. to be feasible for all rounds) that are used to select actions could be guaranteed to be non-empty in early rounds without a known safe policy π_0 or additional assumptions.

known (stochastic) constraints is (Amani et al., 2020). They considered generalized linear rewards where the generalized function is assumed to be twice-differentiable and Lipschitz constant of which logistic rewards is a sub-class. We note that in terms of regret bound alone, their bound specialized to logistic rewards is linearly dependent on the worst case parameter $\boldsymbol{\kappa}$ (see Table 1), which can be arbitrarily large. They considered generalized linear constraints; like in our work, the constraints depend on the unknown parameter vector θ_* in the reward function. However, they consider a priori knowledge of some feasible actions. At a high level, they explore the environment (improving their estimate of θ_*) using those feasible actions and are able to get high probability guarantees of per-round feasibility. We do not assume such prior knowledge. We instead bound long-term constraint violations.

1.3. Preliminaries

In this section, we introduce the basic notations and definitions used throughout the paper. We use capital letters to denote variables (e.g., Y), lowercase letters (e.g., y) represent scalar values, bold lowercase letters (e.g., $\mathbf{y}$) indicate vectors, and bold uppercase letters (e.g., $\mathbf{Y}$) represent matrices. For a twice-differentiable function g, the notation $\dot{g}$ and $\ddot{g}$ denote the first and second derivative of function g respectively. For a random variable Z, let $\mathcal{Z}$ represent the domain of Z and $|\mathcal{Z}|$ the latter's dimension. For two real-valued symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, the notation $\mathbf{A} \succeq \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is positive semi-definite, and when $\mathbf{A}$ is positive definite, we denote $\mathbf{A}$-norm for a vector $\mathbf{z}$ as $\|\mathbf{z}\|_{\mathbf{A}} = \sqrt{\mathbf{z}^\intercal \mathbf{A} \mathbf{z}}$. Finally, for two univariate real-valued functions f and g, we denote $f = \widetilde{\mathcal{O}}(g)$ to indicate that g dominates f up to logarithmic factors; and for an event $E \in \Omega$, we write $\mathbb{1}\{E\}$ the indicator function of E.

We adopt the language of Structural Causal Models (SCMs) (Pearl, 2009, Ch. 7). An SCM M is a tuple $\langle U, V, \mathcal{F}, \mathbb{P}(u) \rangle$, where U is a set of exogenous (unobserved or latent) variables, V is a set of endogenous (observed) variables, $\mathcal{F}$ is a set of structural functions, and $\mathbb{P}(u)$ is a distribution over the latent variables. For the set of structural functions $\mathcal{F}$, $f_{V_i} \in \mathcal{F}$ decides values of an endogenous variable $V_i \in V$ taking as argument a combination of other variables. That is, $V_i \leftarrow f_{V_i}(Pa_{V_i}, U_{V_i}), Pa_{V_i} \subseteq V, U_{V_i} \subseteq U$, where Pa_{V_i} denotes the parent set (explained below) of V_i. Realizations of the set of latent variables $U \sim \mathbb{P}(u)$ induce an observational distribution $\mathbb{P}(v)$ over V. An intervention on a variable $V_1 \in V$, denoted by $do(V_1 = c)$[2] is an operation where value of V_1 is set to a constant c regardless of the structural function $\{f_{V_1} : V_1 \in V\}$. Each SCM is associated with a directed acyclic graph (DAG) $\mathcal{G}$ (e.g., see Figure 1), called the causal diagram, where nodes correspond to endogenous variables V, solid arrows represent arguments of each function f_V. A bi-directed arrow between nodes V_i and V_j indicates an unobserved confounder affecting both V_i and V_j, i.e., $U_{V_i} \cap U_{V_j} \neq \emptyset$. We will use the graph-theoretic family abbreviations, e.g., $Pa(V)_{\mathcal{G}}$ stand for the set of parents of V in $\mathcal{G}$. Two nodes X and Y are said to be d-separated by a third set Z in a DAG $\mathcal{G}$ denoted by $(X \perp\!\!\!\perp Y | Z)_{\mathcal{G}}$ if and only if Z blocks all paths from every node in X to every node in Y. The criterion of blockage follows (Pearl, 2009, Def. 1.2.3), included in Appendix A with formal definitions for completeness.

[2]When the variable being intervened on is clear from context, we write $do(c)$ for short notation.

2. A Theoretical Framework for Constrained Causal Logistic Bandits

In this section, we formalize the constrained causal logistic bandits theoretical framework in the semantics of SCMs and MABs. We start by considering a recruitment example (see Appendix D.1 for more motivating examples), where the decision-making process is characterized by the extended Standard Fairness Model (SFM) (Zhang & Bareinboim, 2018; Plecko & Bareinboim, 2024). See Figure 1 for a graphical model of the SFM. Variable A represents the specified attribute, W is a set of confounded features, and M is a set of intermediate features, D and Y represent the decision and outcome reward. The contextual information $\{\mathbf{w}_t, \mathbf{m}_t, \mathbf{a}_t\}$ is accessible by the learner before making decisions.

2.1. Logistic bandits with structural causal model

At every round t, the learner observes the contextual features $\{\mathbf{w}_t, \mathbf{m}_t, \mathbf{a}_t\}$, which are drawn from a stochastic distribution and then is presented a set of decisions $\mathcal{D}_t$ that depend on the candidate's context. The learner chooses a decision $\mathbf{d}_t \in \mathcal{D}_t$ and receives an outcome reward $y_{t+1} \in \{0, 1\}$. The learner's decision is based on previous round knowledge $\mathcal{F}_t = (\mathcal{F}_0, \{\mathbf{w}_t, \mathbf{m}_t, \mathbf{a}_t, \mathbf{d}_t, y_{t+1}\}_{t=1}^{t-1})$ and causal information, where $\mathcal{F}_0$ represents any prior knowledge. In our problem, we assume that the outcome variable Y has a generalized linear relationship (Filippi et al., 2010; Li et al., 2017) with the features $\mathbf{Z}$, specifically,

$$\mathbb{E}[Y|\mathbf{Z}] = g(f(\mathbf{Z})^\intercal \theta_*), \tag{1}$$

where the fixed but unknown parameters θ_* belongs to $\mathbb{R}^n$, $g(x) = (1 + e^{-x})^{-1}$ is the standard logistic function, f is the mapping function that is known ahead of time to the learner, and the encoded feature vector $f(\mathbf{Z})$ is in $\mathbb{R}^n$. Then the interventional distribution for the expected reward of $do(\mathbf{d}_t)$ and $do(\mathbf{a}_t)$ given the observed contextual features $\mathbf{m}_t$ and $\mathbf{w}_t$ is represented as (Plecko & Bareinboim, 2024):

$$\mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] = g(f(\mathbf{Z}_t)^\intercal \theta_*), \tag{2}$$

where $\mathbf{Z}_t$ is the feature consisting of the decision $\mathbf{d}_t$ and the contexts $\{\mathbf{w}_t, \mathbf{m}_t, \mathbf{a}_t\}$, In this paper, we consider one specified attribute variable A, which in general is a parent of the decision and outcome variable in the causal graph (see Figure 1). Note that the specified attribute value at round t is $\mathbf{a}_t$ and we denote the counterfactual value as $\mathbf{a}'_t$. Both the decision and (hypothetical) counterfactual intervention on the specified attribute value are atomic interventions (Correa & Bareinboim, 2020). Thus, the expected reward for the counterfactual feature $\mathbf{a}'_t$ for any decision $\mathbf{d} \in \mathcal{D}_t$ is

$$\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}'_t), \mathbf{w}_t, \mathbf{m}_t] = g(f(\mathbf{Z}_{\mathbf{a}'_t})^\intercal \theta_*), \quad (3)$$

where $\mathbf{Z}_{\mathbf{a}'_t}$ is the counterfactual feature for the interventions $do(\mathbf{a}'_t)$ and $do(\mathbf{d})$. Notice that for this problem, we consider that the factual feature $\mathbf{Z}_t$ and counterfactual feature $\mathbf{Z}_{\mathbf{a}'_t}$ are both in the feature space $\mathcal{Z}$. The derivation of Equation (2) and Equation (3) follows by the do-calculus rule (Pearl, 1995); readers can refer Appendix D.2 for a more detailed analysis. Therefore, the *counterfactual fairness effect* for decision $\mathbf{d}_t$ is represented as:

$$\Delta(\mathbf{d}_t) \;=\; g(f(\mathbf{Z}_t)^\intercal \theta_*) - g(f(\mathbf{Z}_{\mathbf{a}'_t})^\intercal \theta_*). \quad (4)$$

2.2. Counterfactual fairness modeling via soft constraint

In this section, we discuss modeling fairness as part of the learner's decision making problem. Consider a stochastic bandit optimization with a soft constraint for our decision-making problem. In particular, at every round $t \in [T]$, the learner selects a decision $\mathbf{d}_t$ to maximize the expected reward $\mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$ subject to a constraint on (violations to) counterfactual fairness (Equation (4)),

$$|\, \Delta(\mathbf{d}_t) \,| \;\leq\; \tau, \quad (5)$$

where τ is a predefined fairness threshold. In this work, the counterfactual fairness constraint by Equation (5) requires that the expected reward is similar regardless if the value of the specified attribute had been different. In addition, the learner only receives bandit feedback (the reward). The learner does not not observe feedback on constraint violations.

Huang et al. (2022b) were the first to propose a counterfactual fairness constraint in a bandit framework. We note that their setting confines the learner to decisions from a safe action set (see Footnote 1). To the best of our knowledge, that setting requires strong assumptions on prior knowledge of a subset of safe actions that can be used even before rewards are estimated (the constraint (4) depends on the unknown reward parameter vector θ_*). Prior knowledge of safe actions can be mild in some settings, though we argue counterfactual fairness (a convex combination of logistic functions that depend on the unknown reward parameter vector θ_*) is more complex, thereby it is less obvious for us to construct a prior safe decision without knowing any information about the reward distribution. Therefore, for the setting we consider, since no safe actions might known a priori, we allow for instantaneous violations but bound the cumulative violations.

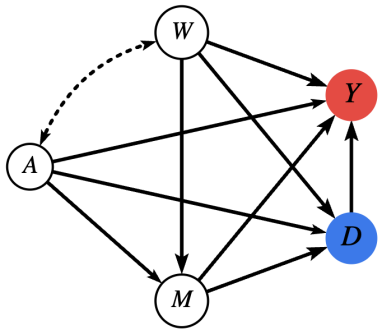


Figure 1: Extended Standard Fairness Model (SFM). A denotes the specified attribute, W is a set of confounded features, M is a set of intermediate features, D is the decision, and Y is the outcome. Let $\{W, M, A\}$ denote contextual information the learner has available to make the decision, and let $Z = \{W, M, A, D\}$ denote variables the reward distribution may depend on.

The goal of the learner is to maximize the cumulative expected outcome reward while minimizing the cumulative expected counterfactual fairness constraint violations throughout the learning process. Define the cumulative expected regret and cumulative expected counterfactual fairness constraint violations as

$$\mathcal{R}(T) \;=\; \sum_{t=1}^{T} \big[\mathbb{E}[Y|do(\mathbf{d}_t^*), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\big], \quad (6)$$

$$\mathcal{V}(T) \;=\; \sum_{t=1}^{T} \big[|\Delta(\mathbf{d}_t)| - \tau\big]_+, \quad (7)$$

where $\mathbf{d}_t^* = \arg\max_{\{\mathbf{d}\in\mathcal{D}_t : |\Delta(\mathbf{d})|\leq\tau\}} \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$ and $[\cdot]_+ = \max\{\cdot, 0\}$. In this paper, we establish a stronger version of regret (Liu et al., 2021; Zhou & Ji, 2022), specifically, let π_t be a probability distribution over the set of actions $\mathcal{D}_t$ at round t, and let $\mathbb{E}_{\pi_t}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] = \sum_{\mathbf{d}\in\mathcal{D}_t} \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\pi_t(\mathbf{d})$ and $\mathbb{E}_{\pi_t}[\,|\Delta(\mathbf{d})| - \tau] = \sum_{\mathbf{d}\in\mathcal{D}_t}[|\Delta(\mathbf{d})| - \tau]\pi_t(\mathbf{d})$. We compare the received outcome reward with the following baseline optimization problem: $\max_{\pi_t}\{\mathbb{E}_{\pi_t}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] : \mathbb{E}_{\pi_t}[|\Delta(\mathbf{d})| - \tau] \leq 0\}$ and π_t^* is the optimal solution at step t. Thus, the stronger regret is defined as:

$$\mathcal{R}_+(T) \;=\; \sum_{t=1}^{T} \big[\mathbb{E}_{\pi_t^*}\, \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$$

$$- \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\big].$$

Note that the probability distribution π_t could include some decisions that violate the constraint but on average the constraint is satisfied, while for a single action it must be a feasible one, therefore, $\mathcal{R}_+(T) \geq \mathcal{R}(T)$.

2.3. Model assumptions and definition

To study our constrained causal logistic bandits problem, we make the following standard assumptions (Yu et al., 2017; Efroni et al., 2020; Zhou & Ji, 2022). Let Θ denote a compact set in $\mathbb{R}^n$. Let $\mathcal{Z}$ denote the feature space domain.

Assumption 1 (Bounded Bandit Parameter)**.** There is a known bound S on the norm of the (unknown) reward parameter vector θ, $\|\theta\|_2 \leq S,\ \forall\, \theta \in \Theta$.

Assumption 2. The feature mapping function $f : \mathcal{Z} \mapsto \mathbb{R}^n$ is in a reproducing kernel Hilbert space (RKHS) with a bounded norm (i.e., a measure of smoothness), such that $\|f(\mathbf{Z})\|_2 \leq 1, \forall\, \mathbf{Z} \in \mathcal{Z}$.

Assumption 3 (Slater's Constraint Qualification)**.** There is a constant $\delta > 0$ such that there exists a feasible probability distribution $\pi_{t,0}$ over decision set $\mathcal{D}_t$ that satisfies $\mathbb{E}_{\pi_{t,0}}[|\Delta(\mathbf{d})| - \tau] \leq -\delta,\ \forall\, t \in [T]$. Without loss of generality, we assume $\delta \leq 1$.

Notice that this is a mild assumption since it only requires that one could find a stochastic policy $\pi_{t,0}$ under which the expected constraint violations will be strictly less than a negative value. Whereas for hard constraints (Amani et al., 2019; Khezeli & Bitar, 2020; Pacchiano et al., 2021), they typically assume that the non-empty initial safe decision set which is stronger than the assumption of existence for a Slater's constant δ about the learner's knowledge.

We next define a problem dependent quantity that impacts learnability.

Definition 1 (Problem Dependent Constant[3])**.**

$$\boldsymbol{\kappa}_{\mathcal{Z}}(\theta_*) = \max_{\mathbf{Z}\in\mathcal{Z}} 1/\dot{g}(f(\mathbf{Z})^\intercal\theta_*). \tag{8}$$

We recall the other problem dependent constants discussed in Table 1: $\boldsymbol{\kappa}_* = 1/\dot{g}(f(\mathbf{Z}_*)^\intercal\theta_*)$, $\boldsymbol{\kappa}_{\mathcal{Z}} = \max_{\mathbf{Z}\in\mathcal{Z}} g(f(\mathbf{Z})^\intercal\theta_*)$, and $\boldsymbol{\kappa} = \max_{\mathbf{Z}\in\mathcal{Z},\theta\in\Theta} 1/\dot{g}(f(\mathbf{Z})^\intercal\theta)$, clearly, $\sqrt{1/\boldsymbol{\kappa}_*} < 1 \ll \boldsymbol{\kappa}_{\mathcal{Z}} \ll \boldsymbol{\kappa}$. Notice that such problem dependent constants are defined through the first order of logistic function, which quantifies the level of non-linearity of plausible expected reward signals with different scales. In particular, $\boldsymbol{\kappa}$ can be significantly large even for reasonable logistic bandits problems. Readers can refer Section 2 of (Faury et al., 2020) for a detailed discussion on the importance of this quantity.

[3]We will drop the dependency on θ_* when there is no ambiguity.

3. Methods for Constrained Causal Logistic Bandits

We next design an online algorithm for the constrained causal logistic bandits problem. We will then develop a unified analysis of regret and constraint violations with rigorous performance guarantees for our decision making strategy. Before proposing the algorithm, we first construct a convex confidence set for the reward parameter θ_* using a regret-to-confidence set conversion (Lee et al., 2024).

3.1. Convex confidence set

For logistic bandit problems, a natural way to estimate the reward parameter θ_* given $\mathcal{F}_t$ is to use maximum-likelihood estimation. We build on the works for the unconstrained problem (Abeille et al., 2021; Lee et al., 2024). At every round t, a reward value y_{t+1} is sampled from a Bernoulli distribution with expected value (or success probability) $g(f(\mathbf{Z}_t)^\intercal\theta_*)$. The unregularized cumulative logistic loss can be written as:

$$\mathcal{L}_t(\theta) = -\sum_{\tau=1}^{t-1} \Big[y_{\tau+1}\log g(f(\mathbf{Z}_\tau)^\intercal\theta) + (1-y_{\tau+1})\log(1-g(f(\mathbf{Z}_\tau)^\intercal\theta))\Big]. \tag{9}$$

The loss $\mathcal{L}_t(\theta)$ is a strongly convex function of θ (Abeille et al., 2021; Lee et al., 2024). The reward parameter is estimated using maximum likelihood estimation (MLE), defined as $\hat{\theta}_t = \arg\min_{||\theta||_2\leq S} \mathcal{L}_t(\theta)$. For $\alpha \in (0, 1]$, we use the confidence set:

$$\mathcal{C}_t(\alpha) = \left\{\theta \in \Theta : \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\alpha)^2\right\}, \tag{10}$$

where $\beta_t(\alpha) = \sqrt{10n\log(\frac{St}{4n}+e) + 2((e-2)+S)\log\frac{1}{\alpha}}$. Then the following proposition ensures that $\mathcal{C}_t(\alpha)$ is a confidence set for θ_* with high probability:

Proposition 1 (Theorem 1 in (Lee et al., 2024))**.**

$$\mathbb{P}\big(\forall t \geq 1,\ \theta_* \in \mathcal{C}_t(\alpha)\big) \ \geq\ 1-\alpha.$$

The proof is provided in Appendix E. The proof uses the approach from the online logistic regression regret guarantee of (Foster et al., 2018) without running the online learning algorithm explicitly. We notice that the radius of the convex confidence set in (Abeille et al., 2021, Lemma 1) is around $\mathcal{O}(\sqrt{nS^3\log(t)})$, while the above tightened loss-based confidence set results in $\mathcal{O}(\sqrt{(n+S)\log(t)})$, leading to an overall improvement in the factor of S, especially when S is large, e.g., $S \geq |\mathcal{D}_t|$.

3.2. Online learning algorithm

We consider a constrained stochastic causal logistic bandit over horizon T as described in Section 2.2. The objec-

tive for the learner is to maximize the cumulative rewards while minimizing cumulative counterfactual fairness violations over time horizon T. To address the challenges on the unknown reward and the unknown counterfactual fairness constraint, we develop a Constrained Causal Logistic Bandits (CCLB) algorithm by leveraging the primal-dual optimization techniques.

The pseudo code for our CCLB algorithm is in Algorithm 1. At every round t, let the Lagrangian of the primal problem $\max_{\pi_t}\{\mathbb{E}_{\pi_t}[\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]] : \mathbb{E}_{\pi_t}\ [|\Delta(\mathbf{d})| - \tau] \leq 0\}$ be $\mathcal{L}_D(\pi_t, \phi) = \mathbb{E}_{\pi_t}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi\mathbb{E}_{\pi_t}[|\Delta(\mathbf{d})| - \tau]$ and then the associated dual function is defined as $q(\phi) = \max_{\pi_t} \mathcal{L}_D(\pi_t, \phi)$. Since both the reward and counterfactual fairness constraint depend on the unknown parameter θ_*, we first estimate it through maximal likelihood estimation and construct a confidence set $\mathcal{C}_t(\alpha)$ using the observed histories, i.e., feature vectors and rewards. The greedy procedure is based on the principle of optimism in the face of uncertainty (OFU) (Auer et al., 2002; 2008; Osband & Van Roy, 2014), where the optimistic estimate $(\tilde{\theta}_t)$ is obtained by maximizing the expected reward across the confidence set $\mathcal{C}_t(\alpha)$, however, we penalize the expected reward for the constraint violations when the greedy action $(\mathbf{d}_t)$ is picked by the learner over the decision domain $\mathcal{D}_t$. The dual update that minimizes $q(\phi)$ with respect to ϕ is by taking a projected gradient descent with $1/\eta$ being the step size. Note that the truncated parameter ρ is chosen to be larger than the optimal dual variable ϕ_*, where can be achieved since the optimal dual variable is bounded under the Slater's constraint qualification, specifically, $\phi_* \leq (\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}_{\pi_0^t}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t])/\delta$ from Theorem 8.42 in (Beck, 2017).

Remark 1. *We remark that the computational complexity of Algorithm 1 is the same as standard algorithms for unconstrained logistic bandits problems (Abeille et al., 2021; Lee et al., 2024), since the dual update is executed via a single-step projection, and the primal optimization retains the character of the unconstrained case without constructing a prior safe subset designed for hard constraints as in (Amani et al., 2020). Additionally, the reward and counterfactual fairness constraint in our algorithm share the same unknown parameter θ_*.*

3.3. Regret and constraint violations bounds

In this section, we provide the theoretic upper bounds for both regret and constraint violations of Algorithm 1 and explain the main idea behind the proof of Theorem 1.

Theorem 1. *Suppose $\rho \geq 2/\delta$, and $\eta = \sqrt{T}/\rho$. For $0 \leq \tau < 1$, under the Slater's constraint qualification in Assumption 3 and regularity assumptions in Assumption 1 and 2, the CCLB algorithm achieves the following bounds simultaneously with probability at least $1-\alpha$ for any $\alpha \in (0, 1]$:*

$$\mathcal{R}_+(T) = \widetilde{\mathcal{O}}\Big(\rho nS\sqrt{T} + \rho n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} + \rho\sqrt{T}\Big),$$
$$\mathcal{V}(T) = \widetilde{\mathcal{O}}\Big(nS\sqrt{T} + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} + \sqrt{T}\Big).$$

Algorithm 1 CCLB Algorithm

1: **Input:** Horizon T, truncated parameter ρ, step size $\eta = \sqrt{T}/\rho$, and the initial dual value $\phi_1 = 0$.
2: **for** $t = 1, 2, 3, \ldots, T$ **do**
3: Use MLE to estimate the reward parameter and build a confidence set $\mathcal{C}_t(\alpha)$ from Equation (10),

$$\mathcal{C}_t(\alpha) = \Big\{\theta \in \Theta : \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\alpha)^2\Big\}.$$

4: Greedy procedure. Choose the optimistic reward parameter $\tilde{\theta}_t$ and select the greedy action $\mathbf{d}_t$:

$$\tilde{\theta}_t = \arg\max_{\theta \in \mathcal{C}_t} \max_{\mathbf{d} \in \mathcal{D}_t} \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t],$$
$$\mathbf{d}_t = \arg\max_{\mathbf{d} \in \mathcal{D}_t} \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t(|\widehat{\Delta}(\mathbf{d})| - \tau).$$

5: Update the estimates of the dual variable:

$$\phi_{t+1} = \text{Proj}_{[0,\rho]}\big[\phi_t + 1/\eta(|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\big].$$

6: Update the estimation and confidence set according to the new received reward y_{t+1}.
7: **end for**

Remark 2. *We remark that: (1) the leading term of our regret $\widetilde{\mathcal{O}}(\rho nS\sqrt{T})$ is similar to the bound $\widetilde{\mathcal{O}}(nS\sqrt{T/\boldsymbol{\kappa}_*})$ established in (Lee et al., 2024) as the logarithmic growth of T, which improves upon (Abeille et al., 2021) (OFULog-r) by a factor of $S^{3/2}$ and improves upon (Zhang & Sugiyama, 2024) by at least a factor of $\sqrt{S}$. Though it acquires a multiplicative factor ρ, one could note that, at an extreme case, when the Slater's constant δ is optimized to $1/\sqrt{\log(T)}$, the leading term scales as $\widetilde{\mathcal{O}}(n\sqrt{T})$. (2) Compared to the unconstrained case (Abeille et al., 2021; Zhang & Sugiyama, 2024; Lee et al., 2024), the regret bound $\mathcal{R}_+(T)$ exhibits an additional term $\rho\sqrt{T}$, which roughly captures the impact of the unknown counterfactual fairness constraint, i.e., a convex combination of logistic functions, which is not logistic function any more. More specifically, the non-convex nature of the logistic mixture introduces a non-linear relationship between the constraint and the reward parameter, thereby resulting in a more complex estimated feasible region of safe decisions at every round. (3) Compared to the constrained generalized linear bandits (Amani et al., 2020), our regret*

bound shows a big improvement on the worst case constant $\boldsymbol{\kappa}$ (see Table 1). (4) If $\tau \geq 1$, the constraint violations bound $\mathcal{V}(T)$ will be zero since the counterfactual fairness constraint is satisfied for all the decisions (see Equation 5) and our problem falls into the setting of logistic bandits without constraint.

See Appendix G for the full proof. We next highlight a few key parts of the proof.

Proof Sketch of Theorem 1. We first derive the following key decomposition of total regret and constraint violations that holds for any $\phi \in [0, \rho] : \mathcal{R}_+(T) + \phi\mathcal{V}(T) \leq \mathcal{R}_1 + \mathcal{R}_2 + \sqrt{T}\rho$, where[4]

$$\begin{aligned}\mathcal{R}_1 \;\leq\; & \sum_{t=1}^{T}(\phi_t - 1)\,\mathbb{E}_{\pi_t^*}\big(\mathbb{E}[\hat{Y}|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \\ & -\; \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\big),\end{aligned}$$

$$\begin{aligned}\mathcal{R}_2 \leq & \sum_{t=1}^{T}\big(\mathbb{E}[\hat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), \\ & do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\big) + \phi\big(\mathbb{E}[\hat{Y}|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] \\ & -\; \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t]\big).\end{aligned}$$

This is attained by employing a dual variable update and necessary algebraic operations. Note that this bound will serve as the cornerstone for the subsequent analysis of both the regret and constraint violations. To further bound $\mathcal{R}_1$ and $\mathcal{R}_2$, we apply the following proposition for logistic bandits regret:

Proposition 2. *With probability at least $1 - \alpha$ for any $\alpha \in (0, 1]$, under the CCLB algorithm, we have:*

$$\sum_{t=1}^{T} g(f(\mathbf{Z}_t)\tilde{\theta}_t) - g(f(\mathbf{Z}_t)\theta_*) = \widetilde{\mathcal{O}}\big(nS\sqrt{T} + n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}\big)$$

The central idea to obtain the above regret (Proposition 2) is by applying Taylor expansion which tightly link estimation errors (e.g. between $\widetilde{\theta}_t$ and θ_*) to prediction errors (e.g. between $g(f(\mathbf{Z}_t)^\intercal\widetilde{\theta}_t)$ and $g(f(\mathbf{Z}_t)^\intercal\theta_*)$), readers can refer Appendix F for more technical details. As for the logistic bandits regret based on the counterfactual feature vectors, i.e., $\sum_{t=1}^{T} g(f(\mathbf{Z}_{\mathbf{a}_t'})\tilde{\theta}_t) - g(f(\mathbf{Z}_{\mathbf{a}_t'})\theta_*)$, we observe that the counterfactual feature vector $\mathbf{Z}_{\mathbf{a}_t'}$ and the factual feature vector $\mathbf{Z}_t$ are both lie in the same feature space $\mathcal{Z}$ for our problem. Thus, $\sum_{t=1}^{T} g(f(\mathbf{Z}_{\mathbf{a}_t'})\tilde{\theta}_t) - g(f(\mathbf{Z}_{\mathbf{a}_t'})\theta_*)$ exhibits the same asymptotic upper bound up to logarithmic factors as $\sum_{t=1}^{T} g(f(\mathbf{Z}_t)\tilde{\theta}_t) - g(f(\mathbf{Z}_t)\theta_*)$, thus $\mathcal{R}_1$ and $\mathcal{R}_2$ are bounded.

Therefore, the regret upper bound $\mathcal{R}_+(T)$ can be obtained by choosing $\phi = 0$. Inspired by (Beck, 2017), we apply tools from constrained convex optimization to obtain the bound on constraint violations $\mathcal{V}(T)$. First, we define the the probability distribution π_t' by $\mathbb{E}_{\pi_t'}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] = \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$ and $\mathbb{E}_{\pi_t'}[|\Delta(\mathbf{d})| - \tau] = [|\Delta(\mathbf{d}_t)| - \tau]$, where the policy π_t' only puts probability mass (equal to 1) on decision $\mathbf{d}_t$ chosen by the learner after the observation of contextual information at every round t. Then, we have,

$$\begin{aligned}\mathcal{R}_+(T) + \phi\mathcal{V}(T) = & \sum_{t=1}^{T}\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \\ & -\; \mathbb{E}_{\pi_t'}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \phi\mathbb{E}_{\pi_t'}[|\Delta(\mathbf{d})| - \tau].\end{aligned}$$

Since $\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$ is convex over π_t^*, both $\mathbb{E}_{\pi_t'}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$ and $\mathbb{E}_{\pi_t'}[|\Delta(\mathbf{d})| - \tau]$ are convex over π_t', by utilizing (Beck, 2017, Theorem 3.60), we obtain the upper bound on $\mathcal{V}(T)$. ■

3.4. Improved regret and constraint violations bounds

In Section 3.3, our analysis demonstrates that the proposed CCLB algorithm (Algorithm 1) achieves both sublinear regret and sublinear constraint violations upper bounds. Another natural question to consider is whether the constraint violations bound can be further improved. It turns out that by introducing a tightness parameter ϵ in the dual update in Algorithm 1, for $\epsilon < \delta$,

$$\phi_{t+1} = \text{Proj}_{[0,\rho]}\big[\phi_t + 1/\eta(|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon)\big], \qquad (11)$$

one can achieve a bounded and in some cases even zero constraint violations by trading the regret slightly while still preserving the same asymptotic order of regret as before. Intuitively, with a tightness parameter $\epsilon > 0$ in the constraint, the learner will be more cautious in selecting actions by effectively working with a stricter constraint (e.g. with fairness threshold $\tau - \epsilon$ instead of τ). Then, under this new hypothetical pessimistic constraint function, the primal problem is modified as: $\max_{\pi_t}\{\mathbb{E}_{\pi_t}[\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]] : \mathbb{E}_{\pi_t}[|\Delta(\mathbf{d})| - \tau + \epsilon] \leq 0\}$. Let $\pi_{t,\epsilon}^*$ be the optimal solution to this new constrained optimization problem, then we have the following relationship between policy $\pi_{t,\epsilon}^*$ and π_t^*:

Proposition 3. *Let policies π_t^* and $\pi_{t,\epsilon}^*$ be the optimal solutions for the constrained problem $\max_{\pi_t}\{\mathbb{E}_{\pi_t}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] : \mathbb{E}_{\pi_t}[|\Delta(\mathbf{d})| - \tau] \leq 0\}$ and $\max_{\pi_{t,\epsilon}}\{\mathbb{E}_{\pi_{t,\epsilon}}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] : \mathbb{E}_{\pi_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau + \epsilon] \leq 0\}$. For $\epsilon < \delta$, we have,*

$$\begin{aligned}& \sum_{t=1}^{T}\mathbb{E}_{\pi_t^*}\,\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \\ & \sum_{t=1}^{T}\mathbb{E}_{\pi_{t,\epsilon}^*}\,\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \;\leq\; \frac{\epsilon T}{\delta}.\end{aligned}$$

[4] $\mathbb{E}[\widehat{Y}|do(\mathbf{d}), \mathbf{X}_t]$ is the estimated expected reward for decision $\mathbf{d}$ and context $\mathbf{X}_t$.

To further investigate how the user-chosen parameter ϵ will impact the regret and constraint violations upper bounds, we define the regret associated with the policy $\pi^*_{t,\epsilon}$ as: $\mathcal{R}^{\epsilon}_{+}(T) = \sum_{t=1}^{T}\left[\mathbb{E}_{\pi^*_{t,\epsilon}}\,\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\right]$, while the constraint violations remain defined by $\mathcal{V}(T) = \sum_{t=1}^{T}\left[|\Delta(\mathbf{d}_t)| - \tau\right]_+$. We then state the following theoretical results for $\mathcal{R}^{\epsilon}_{+}(T)$ and $\mathcal{V}(T)$:

Theorem 2. *Suppose $\rho \geq 2/\delta$, and $\eta = \sqrt{T}/\rho$. For $0 \leq \tau < 1$ and the user-chosen parameter $\epsilon \in [0, \delta)$, under Slater's constraint qualification in Assumption 3 and regularity assumptions in Assumption 1 and 2, the CCLB algorithm with refined constraint condition (see Equation (11)) attains the following theoretical upper bounds with probability at least $1-\alpha$ for any $\alpha \in (0, 1]$:*

$$\mathcal{R}^{\epsilon}_{+}(T) = \widetilde{\mathcal{O}}\Big(\rho n S\sqrt{T} + \rho n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} + \rho\sqrt{T}(1+\epsilon)^2\Big),$$
$$\mathcal{V}(T) = \widetilde{\mathcal{O}}\Big(nS\sqrt{T} + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} + (1+\epsilon)^2\sqrt{T} - \epsilon T\Big).$$

Remark 3. *One could notice that (1) by introducing a tightness parameter ϵ in the dual update, the associated regret $\mathcal{R}^{\epsilon}_{+}(T)$ still achieves a comparable asymptotic upper bound as $\mathcal{R}_{+}(T)$ in Theorem 1; nevertheless, the constraint violations $\mathcal{V}(T)$ upper bound exhibits an ϵT reduction compared to the result in Theorem 1. Consequently, by selecting ϵ appropriately, one can offset the other terms through the subtraction of ϵT, thereby obtaining a constant upper bound (with respect to the horizon T) on the constraint violations. (2) The difference in regret bounds between Theorem 2 with a user selected $\epsilon \in [0, \delta)$ and Theorem 1 is $\rho\sqrt{T}(2\epsilon + \epsilon^2)$. For a problem-dependent (fixed) Slater's constraint qualification constant $\delta > 0$, increasing ϵ only worsens the regret bound, as the learner is increasingly cautious (increasing in ϵ), selecting from a smaller set of actions than the learner would have with $\epsilon = 0$. If δ is large, ρ shrinks towards 2 and so for a fixed tightness ϵ the regret bound reduces. Larger δ also allow for a bigger range of ϵ and thus more room for caution (and regret).*

Proposition 4. *By conditions stated in Theorem 2, for the user-selected parameter $\epsilon' = (\sqrt{T} - \sqrt{T - 4\mathcal{C}_4(\sqrt{T} + \mathcal{C}_1 n\log(T) + (\mathcal{C}_2 + \mathcal{C}_3\boldsymbol{\kappa}_{\mathcal{Z}})n^2((\log(T))^2\sqrt{1/T})})/2\mathcal{C}_4 - 1$, where $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ are the universal constants independent of $n, S, T, \boldsymbol{\kappa}_{\mathcal{Z}}$, if $n \geq 2$ and $\epsilon' < \delta$ for sufficiently large T, then one could achieve a zero upper bound on the constraint violations when selecting $\epsilon \in (\epsilon', \delta)$.*

Note that this user-chosen parameter ϵ trades off between the upper bounds of the regret and constraint violations (Jenatton et al., 2016). Minimizing regret often encourages exploration and adaptability to changing environments, which might lead to occasional violations of constraints. Conversely, strictly adhering to constraints may limit the algorithm's ability to adapt, potentially increasing regret.

4. Numerical Experiments

We next evaluate the empirical performance of our proposed methods on a synthetic data set. See Appendix H for additional experiments for different values of the constraint threshold τ and tightness parameter ϵ.

Data set description:[5] We generated the synthetic dataset from a structural causal model (modified an example from (Plecko & Bareinboim, 2024)), i.e.,

$$\mathscr{F} = \begin{cases} A \leftarrow U_{AW}, \\ W \leftarrow \mathcal{N}(0, 1 - \frac{U_{AW}}{2}), \\ M \leftarrow \begin{cases} \mathcal{N}(0, |W|/2 + |U_M|/3) & \text{if } A = 1, \\ \mathcal{N}(0, |W|/3 + |U_M|/2) & \text{if } A = 0, \end{cases} \\ D_i \leftarrow \mathcal{N}(0, \max\{|W|, |M|\}) \quad i = 1, ..., 20, \\ \mathcal{D} \leftarrow \{D_1, D_2, ..., D_{20}\}, \\ Y \leftarrow \mathbb{1}(U_Y + \frac{1}{3}MD - \frac{1}{5}W > 0), \end{cases}$$
$$\mathbb{P}(U) = \{U_{AW} \sim \text{Bern}(0.5),\ U_M, U_Y \sim \mathcal{N}(0, 1).\}$$

As defined in Figure 1, A denotes the sensitive attribute (binary valued), W is the confounded feature, M represents the intermediate feature, $D \in \mathcal{D}$ is the agent's decision, and Y is the outcome. At every round, we generate a set of 20 feature vectors $\{[A, W, M, D_i]\}_{i=1}^{20}$ along with their corresponding counterfactual feature vectors. We use rejection sampling over the sets to make sure that at least twelve of the feature vectors are feasible.

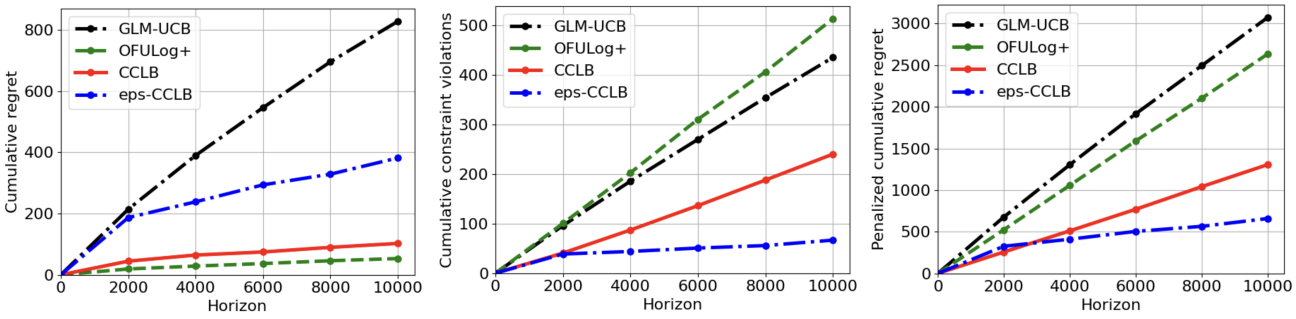
Algorithms:[6] We evaluate four different algorithms: GLM-UCB (Filippi et al., 2010) (unconstrained generalized linear bandits), OFULog+ (Lee et al., 2024) (unconstrained logistic bandits), CCLB (our method, causal logistic bandits with counterfactual fairness constraints, Algorithm 1), and ϵ-CCLB (our method with a user-chosen tightness parameter ϵ, Algorithm 2).

Metrics: We evaluated the algorithms using cumulative regret (6), cumulative constraint violations (7), and a penalized form of cumulative regret for different horizons. For the penalized cumulative regret, when the action picked by the learner violates the counterfactual fairness constraint, the learner still observes the reward value (i.e. the learner can improve the reward parameter estimate $\hat{\theta}$), but we count the reward earned as being 0. In this way, constraint violations are allowed but are not (directly) profitable. This penalized form combines the two primary metrics for simpler analysis.

Results: The results are plotted in Figure 2. Beginning with

[5] The source code is available at https://github.com/jchen-research/CCLB.

[6] Another potential baseline is (Huang et al., 2022b), which also studied counterfactual fairness in the causal bandits framework, though for a different causal graph. Their code was not available at the time of this work.



(a) (b) (c)

Figure 2: Plots for different algorithms GLM-UCB, OFULog+, CCLB ($\tau = 0.16$), and ϵ-CCLB ($\epsilon = 0.1, \tau = 0.16$) on (a) cumulative regret; (b) cumulative constraint violations; (c) penalized cumulative regret.

penalized cumulative regret (Figure 2 (c)), where rewards are only received for fair actions, there is a large gap between our method (CCLB) and the methods of OFULog+ and GLM-UCB, with the gap growing larger for longer horizons. This is expected since GLM-UCB and OFULog+ do not account for constraints. Both baselines have nearly linear penalized cumulative regret across horizons used. OFULog+ does because it frequently violates constraints, and thus large cumulative constraint violations, despite learning good actions (for the unconstrained problem). For cumulative regret (unpenalized), OFULog+ performs better than our method (which seeks to satisfy the constraint).

GLM-UCB performs poorly at identifying good actions within the horizons (Figure 2 (a)). GLM-UCB's regret bound has a linear dependence on the $\boldsymbol{\kappa}$ (see Table 1). GLM-UCB is also designed for a more general class of reward functions. Though ϵ-CCLB has a larger regret than CCLB (but less than GLM-UCB), the cumulative constraint violations of ϵ-CCLB are much smaller than CCLB, especially, its growth rate is nearly 0 from horizon $T = 2,000$ to horizon $T = 10,000$, which rarely violates the constraints.

5. Conclusion

This paper introduced a framework for logistic bandits with counterfactual fairness constraints built within a causal structure. The proposed approach attains satisfactory results, demonstrating sublinear growth in both regret and constraint violations by effectively balancing exploration and exploitation within the environment via primal-dual optimization. By introducing a user-chosen parameter, one can trade the upper bounds between regret and constraint violations to achieve zero cumulative constraint violations.

Several promising directions emerge for future research. (1) One important direction is to extend our method to work with unobserved confounders (i.e. W would be unobserved). (2) Another interesting direction is to extend our model to handle distribution shifts over time. (3) A third interesting direction would be to extend our work to handle budget constraints and consider a fairness notion defined by the resource assignment, potentially building on existing work in bandits with knapsacks (Tran-Thanh et al., 2012; Badanidiyuru et al., 2018; Nie et al., 2024).

Acknowledgments

We thank Wen Huang for discussion. This material is based upon work supported by the National Science Foundation under Award No. 2321786.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. The results in this paper may have implications bringing together strong practical results in causal decision-making. Such integration is anticipated to contribute to the development of more intelligent and reliable AI systems. While a comprehensive exploration of all implications is beyond the scope of this work, an enhanced understanding of causality and decision-making is expected to reduce increase accountability in AI models.

References

Abeille, M., Faury, L., and Calauzènes, C. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021.

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Thompson sampling for the MNL-bandit. In *Conference on Learning Theory*, pp. 76–78. PMLR, 2017.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Generalized linear bandits with safety constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3562–3566. IEEE, 2020.

Amari, S.-i. *Information Geometry and Its Applications*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 4431559779.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. *J. ACM*, 65(3), March 2018. ISSN 0004-5411. doi: 10.1145/3164539. URL https://doi.org/10.1145/3164539.

Beck, A. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017. ISBN 1611974984.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Brekelmans, R., Masrani, V., Wood, F., Steeg, G. V., and Galstyan, A. All in the exponential family: Bregman duality in thermodynamic variational inference. *arXiv preprint arXiv:2007.00642*, 2020.

Chiappa, S. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Chouldechova, A. and Roth, A. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

Correa, J. and Bareinboim, E. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10093–10100, 2020.

Darlington, R. B. Another look at "cultural fairness". *Journal of Educational Measurement*, 8(2):71–82, 1971. ISSN 00220655, 17453984. URL http://www.jstor.org/stable/1433960.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312. PMLR, 2021.

Dong, S., Ma, T., and Van Roy, B. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pp. 1158–1160. PMLR, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Faury, L., Abeille, M., Jun, K.-S., and Calauzènes, C. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Logistic regression: The importance of being improper. In *Conference on Learning Theory*, pp. 167–208. PMLR, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Hu, Y. and Zhang, L. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9549–9557, 2022.

Hu, Y., Wu, Y., and Zhang, L. Long-term fair decision making through deep generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22114–22122, 2024.

Huang, W., Labille, K., Wu, X., Lee, D., and Heffernan, N. Achieving user-side fairness in contextual bandits. *Human-Centric Intelligent Systems*, 2(3):81–94, 2022a.

Huang, W., Zhang, L., and Wu, X. Achieving counterfactual fairness for causal bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6952–6959, 2022b.

Imai, K. and Jiang, Z. Principal Fairness for Human and Algorithmic Decision-Making. *Statistical Science*, 38 (2):317 – 328, 2023. doi: 10.1214/22-STS872. URL https://doi.org/10.1214/22-STS872.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. In *International Conference on Machine Learning*, pp. 1617–1626. PMLR, 2017.

Jenatton, R., Huang, J., and Archambeau, C. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pp. 402–411. PMLR, 2016.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 158–163, 2018.

Khezeli, K. and Bitar, E. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10202–10209, 2020.

Krause, A. and Guestrin, C. Near-optimal observation selection using submodular functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 7, pp. 1650–1654, 2007.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

Lavori, P. W. and Dawson, R. Adaptive treatment strategies in chronic disease. *Annu. Rev. Med.*, 59(1):443–453, 2008.

Lee, J., Yun, S.-Y., and Jun, K.-S. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pp. 4474–4482. PMLR, 2024.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Liu, Q., Xu, W., Wang, S., and Fang, Z. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. *Advances in Neural Information Processing Systems*, 35:2997–3010, 2022.

Liu, X., Li, B., Shi, P., and Ying, L. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.

Murphy, S. A. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Nie, G., Zhu, Y., Nadew, Y. Y., Basu, S., Pavan, A., and Quinn, C. J. Size-constrained k-submodular maximization in near-linear time. In *Uncertainty in Artificial Intelligence*, pp. 1545–1554. PMLR, 2023.

Nie, G., Aggarwal, V., and Quinn, C. J. Gradient methods for online dr-submodular maximization with stochastic long-term constraints. *Advances in Neural Information Processing Systems*, 2024.

Oh, M.-h. and Iyengar, G. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Osband, I. and Van Roy, B. Near-optimal reinforcement learning in factored MDPs. *Advances in Neural Information Processing Systems*, 27, 2014.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835. PMLR, 2021.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2337329.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Plecko, D. and Bareinboim, E. Causal fairness for outcome control. *Advances in Neural Information Processing Systems*, 36:47575–47597, 2023.

Plecko, D. and Bareinboim, E. Causal fairness analysis. *Foundations and Trends® in Machine Learning*, Vol. 17, No. 3, pp 1–238. DOI: 10.1561/2200000106, 2024.

Plecko, D. and Bareinboim, E. Fairness-accuracy trade-offs: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26344–26353, 2025.

Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 1134–1140, 2012.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596.

Wu, H., Srikant, R., Liu, X., and Jiang, C. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Wu, Y., Zhang, L., and Wu, X. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Xu, H., Liu, Y., Lau, W. C., and Li, R. Combinatorial multi-armed bandits with concave rewards and fairness constraints. In *IJCAI*, pp. 2554–2560, 2020.

Yu, H., Neely, M., and Wei, X. Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, 30, 2017.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, 2017.

Zhang, J. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pp. 11012–11022. PMLR, 2020.

Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zhang, L., Wu, Y., and Wu, X. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

Zhang, Y.-J. and Sugiyama, M. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhou, X. and Ji, B. On kernelized multi-armed bandits with constraints. *Advances in Neural Information Processing Systems*, 35:14–26, 2022.

Organization of the Appendix

- In Appendix A, we recall important notations and introduce some useful functions and results.
- In Appendix B, we provide additional related works for our problem.
- In Appendix C, we list some technical lemmas, needed for the analysis.
- In Appendix D, we provide motivation and proofs for the constrained causal logistic bandits framework.
- In Appendix E, we prove the convex confidence set.
- In Appendix F, we prove the logistic bandits regret upper bound.
- In Appendix G, we prove the total regret and constraint violations upper bounds, and the improved results.
- In Appendix H, we provide additional results for the numerical experiments.

A. Preliminaries

We first provide a formal definition for d-separation discussed in Section 1.3,

Definition 2 (d-separation (Pearl, 2009))**.** A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if (1) p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, (2) p contains an intervened fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z.

We then detail below some useful notations that have been used throughout the paper. Below $\theta_* \in \mathbb{R}^n$, $f(\mathbf{Z}_t) \in \mathbb{R}^n$ and $Y \in \{0, 1\}$,

θ_*	true reward parameter vector.
Y	reward variable.
$\mathbf{X}_t$	context vector including the specified attribute, the confounded features, and the intermediate features.
$f(\mathbf{Z}_t)$	mapping feature vector.
λ_t	regularization parameter.
ϕ_t	dual variable.
ρ	truncated parameter.
ϵ	user-chosen tightness parameter.
δ	Slater's constant.
α	failure probability.
$\mathcal{C}_t(\alpha)$	confidence set.
$\mathcal{B}_p^n(1)$	n-dimensional ball of radius 1 under the ℓ^p norm.
$\|\|\cdot\|\|$	ℓ_2 norm.

We further recall and introduce the following functions and use it for the following analysis,

$$\begin{aligned}
\Delta(\mathbf{d}) &= \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}'_t), \mathbf{w}_t, \mathbf{m}_t] \\
&= g(f(\mathbf{Z}_t)^\intercal \theta_*) - g(f(\mathbf{Z}_{\mathbf{a}'_t})^\intercal \theta_*) \qquad (12) \\
\mathbf{V}_t &= \sum_{\tau=1}^{t-1} f(\mathbf{Z}_\tau)\, f(\mathbf{Z}_\tau)^\intercal + \lambda_t \mathbf{I}_n \qquad (13) \\
\mathbf{H}_t(\theta_*) &= \sum_{\tau=1}^{t-1} \dot{g}(f(\mathbf{Z}_\tau)^\intercal \theta_*) f(\mathbf{Z}_\tau) f(\mathbf{Z}_\tau)^\intercal + \lambda_t \mathbf{I}_n \qquad (14) \\
\mathbf{G}_t(\theta, \theta_*) &= \sum_{\tau=1}^{t-1} \int_{v=0}^{1} (1-v) \dot{g}\Big(f(\mathbf{Z}_\tau)^\intercal \theta + v\, f(\mathbf{Z}_\tau)^\intercal (\theta - \theta_*)\Big)\, dv\, f(\mathbf{Z}_\tau)\, f(\mathbf{Z}_\tau)^\intercal \;+\; \lambda_t \mathbf{I}_n \qquad (15) \\
\alpha(f(\mathbf{Z}_\tau), \tilde{\theta}_t, \theta_*) &= \int_{v=0}^{1} \dot{g}\Big(f(\mathbf{Z}_\tau)^\intercal \tilde{\theta}_t + v\, f(\mathbf{Z}_\tau)^\intercal (\theta_* - \tilde{\theta}_t)\Big)\, dv \qquad (16)
\end{aligned}$$

Where the regularized design matrices $\mathbf{V}_t$, $\mathbf{H}_t(\theta_*)$, $\mathbf{G}_t(\theta, \theta_*)$, and $\alpha(f(\mathbf{Z}_\tau), \tilde{\theta}_t, \theta_*)$ are defined for the proof of logistic bandits regret upper bound in Appendix F. In particular, $\mathbf{H}_t(\theta_*)$ measures the local behavior of the logistic function through $\dot{g}\big(f(\mathbf{Z}_\tau)^\intercal \theta_*\big)$.

B. Additional Related Works

Logistic bandits. The logistic bandits model represents a sequential decision-making framework that has attracted substantial attention within the parametric bandits literature (Li et al., 2010; Filippi et al., 2010; Li et al., 2017; Dong et al., 2019). In a recent work, Faury et al. (2020) proposed an optimistic algorithm based on a finer examination of the non-linearities of the reward function to study the prohibitive linear dependencies introduced by $\boldsymbol{\kappa}$ in the regret upper bound. Abeille et al. (2021) proved a minimax-optimal rate by deriving an $\Omega\big(n\sqrt{T/\boldsymbol{\kappa}_*(T)}\big)$ problem-dependent lower-bound, which implies that the non-linearity in logistic bandits can ease the exploration-exploitation trade-off in the long-term regime, i.e. $\kappa_*(T) > 1$. Faury et al. (2022) addressed the issue of computational tractability while preserving statistical efficiency by designing a new convex confidence set. Additionally, another line of research is the multinomial logit contextual bandit problem (Agrawal et al., 2017; Oh & Iyengar, 2019; Zhang & Sugiyama, 2024; Lee et al., 2024), which generalizes the binary logistic bandit by allowing the learner to select a subset of arms. In particular, (Zhang & Sugiyama, 2024; Lee et al., 2024) also improve the logistic bandits on the regret guarantee (with respect to S) and computational complexity, respectively.

Fairness. The body of research in fair machine learning is expanding and encompasses a variety of contexts. Within this field, three distinct tasks can be identified: (1) the detection and quantification of biases in currently deployed policies; (2) the development of fair predictive models for outcomes; and (3) the formulation of fair decision-making policies. Our work falls under the setting of online outcome control (task (3)) that explores fairness through a causal lens (Huang et al., 2022a;b; Plecko & Bareinboim, 2023; 2025). Unlike us, Plecko & Bareinboim (2023; 2025) explored the fairness through the path-specific counterfactual effect in an offline setting along with budget constraint. As for the online setting, Hu & Zhang (2022) studied achieving long-term fairness within a Markov Decision Process (MDP) framework, in which they quantified long-term fairness by evaluating the path-specific effects in a causal graph under interventions on sensitive attributes and predicted decisions. More recently, Hu et al. (2024) studied long-term fair decision-making through deep generative models.

Constrained MABs. There is a large body of work on bandits with different types of constraints, including knapsack bandits (Wu et al., 2015; Agrawal & Devanur, 2016), submodular maximization (Krause & Guestrin, 2007; Nie et al., 2023), bandits with hard safety constraints (Amani et al., 2019; Pacchiano et al., 2021), and bandits with cumulative soft constraints (Liu et al., 2021; Zhou & Ji, 2022). Among them, the bandit setting with cumulative soft constraints is most closely related to ours in that the goal is also to minimize the cumulative constraint violation. In particular, Zhou & Ji (2022) considered a general unknown reward function and a general unknown constraint function in kernelized bandits via primal-dual optimization. More broadly, this type of constrained problem has also been studied in the reinforcement learning (RL) setting (Efroni et al., 2020; Ding et al., 2021) where constraints are managed through convex optimization methods.

C. Technical Lemmas

Lemma 1 ((Abeille et al., 2021) Lemma 11). *Let $\{u_\tau\}_{\tau=1}^{\infty}$ be a sequence in $\mathbb{R}^n$ such that $||u_\tau|| \leq B$ for all $\tau \in \mathbb{N}$, and let λ be a non-negative scalar. For $t \geq 1$ define $\mathbf{V}_t = \sum_{\tau=1}^{t-1} u_\tau u_\tau^\intercal + \lambda \mathbf{I}_n$. The following inequality holds:*

$$\det(\mathbf{V}_t) \;\leq\; \Big(\frac{\mathrm{tr}(\mathbf{V}_t)}{n}\Big)^n \;\leq\; \Big(\lambda + \frac{(t-1)B^2}{n}\Big)^n.$$

Lemma 2 ((Abeille et al., 2021) Lemma 12). *Let $\{u_\tau\}_{\tau=1}^{\infty}$ be a sequence in $\mathbb{R}^n$ such that $||u_\tau|| \leq B$ for all $\tau \in \mathbb{N}$. Further let $\{\lambda_\tau\}_{\tau=1}^{\infty}$ be an non-decreasing sequence in $\mathbb{R}^+$ s.t. $\lambda_1 = 1$. For $t \geq 1$ define $\mathbf{V}_t = \sum_{\tau=1}^{t-1} u_\tau u_\tau^\intercal + \lambda_t \mathbf{I}_n$. Then:*

$$\sum_{t=1}^{T} ||u_t||^2_{\mathbf{V}_t^{-1}} \;\leq\; 2n(1+B^2)\log\Big(\lambda_T + \frac{TB^2}{n}\Big).$$

Proof. By definition of $\mathbf{V}_t$:

$$|\mathbf{V}_{t+1}| \;=\; |\sum_{\tau=1}^{t-1} u_\tau u_\tau^\intercal + u_t u_t^\intercal + \lambda_t \mathbf{I}_n|$$

$$
\begin{aligned}
&\geq |\sum_{\tau=1}^{t-1} u_\tau u_\tau^\intercal + u_t u_t^\intercal + \lambda_{t-1}\mathbf{I}_n| \\
&= |\mathbf{V}_t + u_t u_t^\intercal| \\
&\geq |\mathbf{V}_t|\Big|\mathbf{I}_n + u_t \mathbf{V}_t^{-1} u_t^\intercal\Big| \\
&= |\mathbf{V}_t|\,(1 + ||u_t||^2_{\mathbf{V}_t^{-1}}).
\end{aligned}
$$

Where the second inequality follows by $\lambda_t \geq \lambda_{t-1}$; the forth inequality comes from Matrix Determinant Lemma. Taking the log on both side of the equation and summing from $t = 1$ to T :

$$
\begin{aligned}
\sum_{t=1}^{T} \log(1 + ||u_t||^2_{\mathbf{V}_t^{-1}}) &\leq \sum_{t=1}^{T} \Big[\log|\mathbf{V}_{t+1}| - \log|\mathbf{V}_t|\Big] \\
&= \log|\mathbf{V}_{T+1}| - \log|\lambda_1 \mathbf{I}_n| \\
&= \log\Big(\det(\mathbf{V}_{T+1})\Big) \\
&= n \log\Big(\lambda_T + \frac{TB^2}{n}\Big).
\end{aligned}
$$

Where the second equality is by telescopic sum; and the last equality comes from Lemma 1. Therefore:

$$
\begin{aligned}
n \log\Big(\lambda_T + \frac{TB^2}{n}\Big) &\geq \sum_{t=1}^{T} \log(1 + ||u_t||^2_{\mathbf{V}_t^{-1}}) \\
&\geq \sum_{t=1}^{T} \log(1 + \frac{1}{\max(1, B^2/\lambda_t)} ||u_t||^2_{\mathbf{V}_t^{-1}}) \\
&\geq \frac{1}{2\max(1, B^2/\lambda_1)} \sum_{t=1}^{T} ||u_t||^2_{\mathbf{V}_t^{-1}} \\
&\geq \frac{1}{2(1+B^2)} \sum_{t=1}^{T} ||u_t||^2_{\mathbf{V}_t^{-1}}.
\end{aligned}
$$

Where the second inequality comes from $||u_t||^2_{\mathbf{V}_t^{-1}} \leq B^2/\lambda_t$; and the third inequality follows by $\log(1+x) > x/2,\ \forall x \in (0, 1]$. ■

We then state some useful generalized self-concordance results from (Faury et al., 2020, Lemma 9) and (Abeille et al., 2021, Lemma 7). We provide a proof for the sake of completeness (we also use the properties from (Abeille et al., 2021, Lemma 8)).

Lemma 3 ((Faury et al., 2020) Lemma 9)**.** *Let g be a strictly increasing function such that $|\ddot{g}| \leq |\dot{g}|$, and let $\mathcal{Z}$ be any bounded interval of $\mathbb{R}$. Then, for all $z_1, z_2 \in \mathcal{Z}$:*

$$
\int_{v=0}^{1} \dot{g}(z_1 + v(z_2 - z_1))\, dv \;\geq\; \frac{\dot{g}(z)}{1 + |z_1 - z_2|} \quad \textit{for } z \in \{z_1, z_2\}.
$$

Proof. Since function g is strictly increasing, we have $\dot{g} > 0$ for any $z \in \mathcal{Z}$. Therefore:

$$
\begin{aligned}
\frac{\ddot{g}}{\dot{g}} \geq -1 \;&\Rightarrow\; -|z_1 - z_0| \leq \int_{\min\{z_1, z_0\}}^{\max\{z_1, z_0\}} \frac{\ddot{g}(z)}{\dot{g}(z)}\, dz \\
&\Rightarrow\; -|z_1 - z_0| \leq \log\left(\frac{\dot{g}(\max\{z_1, z_0\})}{\dot{g}(\min\{z_1, z_0\})}\right) \\
&\Rightarrow\; \dot{g}(\min\{z_1, z_0\}) \exp(-|z_1 - z_0|) \leq \dot{g}(\max\{z_1, z_0\}),
\end{aligned}
$$

where the first line follows from $z_0 \in \mathcal{Z}$. Assume that $z_2 \geq z_1$, let $v \geq 0$, and set $z_0 = z_1 + v(z_2 - z_1)$, we then could easily get:

$$\begin{aligned}
&\Rightarrow \dot{g}(z_1)\exp\big(-v|z_2 - z_1|\big) \leq \dot{g}\big(z_1 + v(z_2 - z_1)\big) \\
&\Rightarrow \dot{g}(z_1)\frac{1 - \exp\big(-|z_1 - z_2|\big)}{|z_1 - z_2|} \leq \int_{v=0}^{1} \dot{g}\big(z_1 + v(z_2 - z_1)\big)\, dv \\
&\Rightarrow \dot{g}(z_1)\frac{1}{1 + |z_1 - z_2|} \leq \int_{v=0}^{1} \dot{g}\big(z_1 + v(z_2 - z_1)\big)\, dv.
\end{aligned}$$

Where the second line follows by taking integral of v from 0 to 1 for both sides; and the last line is obtained by using $\exp(-x) \leq (1 + x)^{-1}$ if $x \geq 0$. We note that the same inequality can be proved when $z_2 < z_1$ by following the same steps. ■

Lemma 4 (Polynomial Inequality; (Abeille et al., 2021) Lemma 7)**.** *Let $b, c \in \mathbb{R}^+$, and $u \in \mathbb{R}$. The following implication holds:*

$$u^2 \leq bu + c \implies u \leq b + \sqrt{c}$$

Proof. Let function $f(u) = u^2 - bu - c$. Then f is a strongly-convex function which roots are:

$$u_1 = \frac{1}{2}(b + \sqrt{b^2 + 4c}) \quad u_2 = \frac{1}{2}(b - \sqrt{b^2 + 4c})$$

If $u^2 \leq bu + c$, then $f(u) < 0$ and by convexity of f we obtain:

$$\begin{aligned}
u &\leq \max\{u_1, u_2\} \\
&\leq \frac{1}{2}\left(b + \sqrt{b^2 + 4c}\right) \\
&\leq b + \sqrt{c}.
\end{aligned}$$

Where the last inequality is because $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y},\ \forall x, y \geq 0$. ■

D. Causal Logistic Bandits Framework

D.1. Additional motivation example

Online Recommendation System (Huang et al., 2022b). Customers arrive sequentially according to an underlying stochastic distribution, and an online decision-making model selects and recommends a specific item to each incoming individual based on a predefined strategy. In this context, each arm represents a distinct item or content piece available for recommendation to a user. The reward is determined by the user's interaction with the recommended item, such as whether the user clicks on it or not. The fairness constraint mandates that customers with similar profiles receive similar rewards, irrespective of their specific attributes and the particular items being recommended.

D.2. Derivation for the factual and counterfactual expected reward

Here, we provide proofs for Equation (2) and Equation (3), which follow by the *do*-calculus rule (Pearl, 1995).

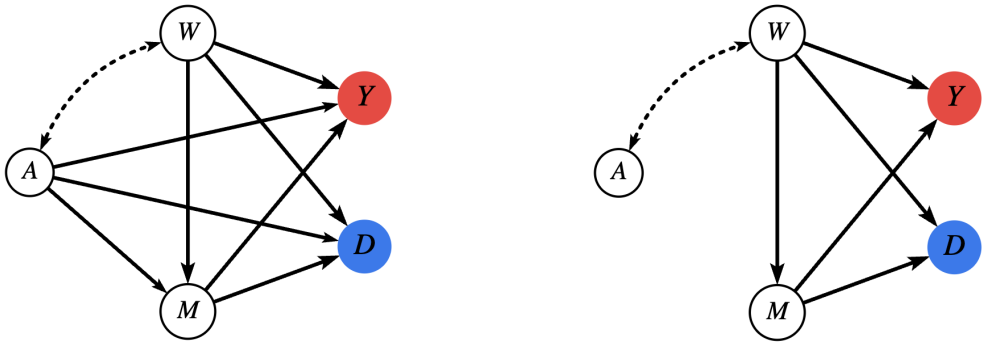
$$\begin{aligned}
\mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] &= \mathbb{E}[Y|\mathbf{d}_t, \mathbf{a}_t, \mathbf{w}_t, \mathbf{m}_t] && (17)\\
&= \mathbb{E}[Y|\mathbf{Z}_t] && (18)\\
&= g(f(\mathbf{Z}_t)^\intercal \theta_*), && (19)
\end{aligned}$$

where (17) follows by $(D, A \perp\!\!\!\perp Y|W, M)_{\mathcal{G}_{\underline{D},A}}$ (see Figure 3b); (18) follows by denoting $\mathbf{Z}_t$ as the features from $\mathbf{d}_t, \mathbf{w}_t, \mathbf{m}_t$ and $\mathbf{a}_t$; and (19) follows by the logistic reward assumption (Equation (1)). As for Equation (3),

$$\begin{aligned}
\mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}'_t), \mathbf{w}_t, \mathbf{m}_t] &= \mathbb{E}[Y|\mathbf{d}_t, \mathbf{a}'_t, \mathbf{w}_t, \mathbf{m}_t] && (20)\\
&= \mathbb{E}[Y|\mathbf{Z}_{\mathbf{a}'_t}] && (21)
\end{aligned}$$

$$= g(f(\mathbf{Z}_{\mathbf{a}'_t})^\intercal \theta_*), \tag{22}$$

where (20) follows by $(D, A \perp\!\!\!\perp Y | W, M)_{\mathcal{G}_{\underline{D,A}}}$ (see Figure 3b); (21) follows by denoting $\mathbf{Z}_{\mathbf{a}'_t}$ as the features from $\mathbf{d}_t, \mathbf{w}_t, \mathbf{m}_t$ and $\mathbf{a}'_t$; and (22) the last equality follows by the logistic reward assumption, similar as above. ■



(a) (b)

Figure 3: (a) A causal diagram representing $\mathcal{G}_{\underline{D}}$; (b) another causal diagram representing $\mathcal{G}_{\underline{D,A}}$.

E. Confidence Sets

In this section, we provide proofs for the construction of the improved convex confidence set for the estimated bandit parameter presented in Section 3.1. We borrow the techniques from (Lee et al., 2024, Section 3) to obtain the results.

Recall the convex confidence set definition:

$$\mathcal{C}_t(\alpha) = \left\{\theta \in \Theta : \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\alpha)^2\right\},$$

where:

$$\beta_t(\alpha) = \sqrt{10n \log(\frac{St}{4n} + e) + 2((e-2) + S) \log \frac{1}{\alpha}}.$$

Proposition 1. *Let* $\alpha \in (0, 1]$*, then*

$$\mathbb{P}\big(\forall t \geq 1,\ \theta_* \in \mathcal{C}_t(\alpha)\big) \ \geq\ 1 - \alpha.$$

Proof. The proof unfolds through three principal technical components similar with (Lee et al., 2024). First, we invoke decomposition identities for the logistic loss, expressing $\mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t)$ as the sum of (i) the regret of the online learning algorithm, (ii) a martingale difference sequence, and (iii) a collection of KL-divergence terms. Second, in controlling the martingale sum, we derive and apply an anytime variant of Freedman's inequality tailored to martingales. Third, to bound the KL-divergence contribution, we fuse the self-concordant analysis of Abeille et al. (2021) with an information-geometric interpretation of the KL divergence.

Firstly, we denote ξ_τ as a real-valued martingale difference noise where $\xi_\tau = g(f(\mathbf{Z}_\tau)^\intercal \theta_*) - y_\tau$, thus for the logistic loss $\ell_\tau(\theta) = -y_\tau \log g(f(\mathbf{Z}_\tau)^\intercal \theta) - (1 - y_\tau) \log\big(1 - g(f(\mathbf{Z}_\tau)^\intercal \theta)\big)$, we have that the following equality holds for any θ:

$$\ell_\tau(\theta_*) = \ell_\tau(\theta) + \xi_\tau \langle f(\mathbf{Z}_\tau), \theta - \theta_* \rangle - KL\big(\mathrm{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta_*)), \mathrm{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta))\big).$$

The equality follows from the first order Taylor expansion with an integral remainder (see (Lee et al., 2024, Appendix C.4.1) for more details). Setting θ to be the optimistic estimate $\tilde{\theta}_\tau$ and taking a sum over time steps τ:

$$0 \ = \ \sum_{\tau=1}^{t} \Big[\ell_\tau(\tilde{\theta}_\tau) - \ell_\tau(\theta_*) - KL\big(\mathrm{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta_*)), \mathrm{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta)) + \xi_\tau \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle\Big]$$

$$= \sum_{\tau=1}^{t} \Big[\ell_\tau(\tilde{\theta}_\tau) - \ell_\tau(\hat{\theta}_t) + \ell_\tau(\hat{\theta}_t) - \ell_\tau(\theta_*) - KL\big(\text{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta_*)), \text{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta)) + \xi_\tau \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle \Big] \tag{23}$$

$$= \sum_{\tau=1}^{t} \Big[\ell_\tau(\hat{\theta}_t) - \ell_\tau(\theta_*) - KL\big(\text{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta_*)), \text{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta)) + \xi_\tau \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle \Big] + \sum_{\tau=1}^{t} \Big[\ell_\tau(\tilde{\theta}_\tau) - \ell_\tau(\hat{\theta}_t) \Big] \tag{24}$$

where in (23) we add and subtract $\ell_\tau(\hat{\theta}_t)$ and in (24) we rearrange terms. We further define $\zeta_1(t) = \sum_{\tau=1}^{t} \xi_\tau \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle$, $\zeta_2(t) = \sum_{\tau=1}^{t} KL\big(\text{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta_*)), \text{Bern}(g(f(\mathbf{Z}_\tau)^\intercal \theta))$, and $\zeta_3(t) = \sum_{\tau=1}^{t} \Big[\ell_\tau(\tilde{\theta}_\tau) - \ell_\tau(\hat{\theta}_t) \Big]$. Using (24), we then have

$$\mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \;=\; \sum_{\tau=1}^{t} \Big[\ell_\tau(\theta_*) - \ell_\tau(\hat{\theta}_t) \Big] \;=\; \zeta_1(t) - \zeta_2(t) + \zeta_3(t) \tag{25}$$

Upper Bounding $\zeta_1(t)$. Recall that $\mathcal{F}_\tau = \sigma\big(\{f(\mathbf{Z}_1), y_1, ..., f(\mathbf{Z}_\tau), y_\tau, f(\mathbf{Z}_{\tau+1})\}\big)$ is the filtration for our bandit model, $f(\mathbf{Z}_\tau)$ and $\tilde{\theta}_\tau$ are $\mathcal{F}_{s-1}$-measurable, and ξ_τ is a martingale difference sequence w.r.t. $\mathcal{F}_{s-1}$. Thus, we have that,

$$\mathbb{E}\big[\xi_\tau^2 \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle^2 | \mathcal{F}_{s-1}\big] = \dot{g}(f(\mathbf{Z}_\tau)^\intercal \theta_*) \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle^2 \quad \text{and} \quad |\xi_\tau \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle| \le 2S.$$

From (Beygelzimer et al., 2011, Theorem 1), we could apply Freedman's inequality to obtain the following result,

Lemma 5. *(Lee et al., 2024, Lemma 3). Let $f(\mathbf{Z}_1), ..., f(\mathbf{Z}_t)$ be martingale difference sequence satisfying $\max_\tau |f(\mathbf{Z}_\tau)| \le R$ a.s., and let $\mathcal{F}_\tau$ be the $\sigma-$field generated by $(f(\mathbf{Z}_1), ..., f(\mathbf{Z}_t)$. Then for any $\alpha \in (0,1)$ and any $\eta \in [0, 1/R]$, the following holds with probability at least $1-\alpha$:*

$$\sum_{\tau=1}^{t} f(\mathbf{Z}_\tau) \;\le\; (e-2)\eta \sum_{\tau=1}^{t} \mathbb{E}[f(\mathbf{Z}_\tau)_\tau^2 | \mathcal{F}_{\tau-1}] + \frac{1}{\eta} \log \frac{1}{\alpha}, \quad \forall t \ge 1.$$

Thus, for $\eta \in [0, \frac{1}{2S}]$ to be chosen later, by invoking Lemma 5 for the martingale difference sequence $f(\mathbf{Z}_1), ..., f(\mathbf{Z}_t)$, the following holds with probability at least $1-\alpha$, $\forall t \ge 1$:

$$\zeta_1(t) \;\le\; (e-2)\eta \sum_{\tau=1}^{t} \dot{g}(f(\mathbf{Z}_\tau)^\intercal \theta_*) \langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_* \rangle^2 + \frac{1}{\eta} \log \frac{1}{\alpha}. \tag{26}$$

Lower Bounding $\zeta_2(t)$. From the standard result in information geometry (Amari, 2016; Brekelmans et al., 2020), we have the following result:

Lemma 6. *(Lee et al., 2024, Lemma 4). Let $m(z) := \log(1+e^z)$ be the log-partition function for the Bernoulli distribution and $g(z) = \frac{1}{1+e^{-z}}$. Then, we have that*

$$KL\big(\text{Bern}(g(z_2)), \text{Bern}(g(z_1))\big) \;=\; D_m(z_1, z_2),$$

where $D_m(z_1, z_2)$ is the Bregman Divergence defined as $D_m(z_1, z_2) = \int_{z_2}^{z_1} \ddot{m}(z)(z_1 - z)\, dz$.

Notice that

$$D_m(z_1, z_2) = \int_{z_2}^{z_1} \ddot{m}(z)(z_1 - z)\, dz = \int_{z_2}^{z_1} \big(\log(1+e^z)\big)''(z_1 - z)\, dz = \int_{z_2}^{z_1} \dot{g}(z)(z_1 - z)\, dz. \tag{27}$$

Thus, we have the following lower bound on $\zeta_2(t)$,

$$\zeta_2(t) \;=\; \sum_{\tau=1}^{t} KL\big(\text{Bern}(g(z_2)), \text{Bern}(g(z_1))\big) \tag{28}$$

$$= \sum_{\tau=1}^{t} D_m\big(f(\mathbf{Z}_\tau)^\intercal \tilde{\theta}_\tau, f(\mathbf{Z}_\tau)^\intercal \theta_*\big) \tag{29}$$

$$= \sum_{\tau=1}^{t} \int_{f(\mathbf{Z}_\tau)^\intercal \theta_*}^{f(\mathbf{Z}_\tau)^\intercal \tilde{\theta}_\tau} \dot{g}(z)(f(\mathbf{Z}_\tau)^\intercal \tilde{\theta}_\tau - z)\, dz \tag{30}$$

$$= \sum_{\tau=1}^{t} \langle f(\mathbf{Z}_\tau), \theta_* - \tilde{\theta}_\tau\rangle^2 \int_0^1 (1-v)\dot{g}\big(f(\mathbf{Z}_\tau)^\intercal(v\tilde{\theta}_\tau + (1-v)\theta_*\big)\, dv \tag{31}$$

$$\geq \sum_{\tau=1}^{t} \langle f(\mathbf{Z}_\tau), \theta_* - \tilde{\theta}_\tau\rangle^2 \frac{\dot{g}(f(\mathbf{Z}_\tau)^\intercal \theta_*)}{2 + |f(\mathbf{Z}_\tau)^\intercal(\theta_* - \tilde{\theta}_\tau)|} \tag{32}$$

$$\geq \sum_{\tau=1}^{t} \langle f(\mathbf{Z}_\tau), \theta_* - \tilde{\theta}_\tau\rangle^2 \frac{\dot{g}(f(\mathbf{Z}_\tau)^\intercal \theta_*)}{2 + 2S}. \tag{33}$$

Where (28) follows by definition of ζ_2; (29) uses Lemma 6; (30) uses (27); (31) follows by change variables; (32) follows by (Abeille et al., 2021, Lemma 8); and (33) follows by Assumption 1 and 2.

Upper Bounding $\zeta_3(t)$ (Lee et al., 2024, Theorem 2). From (Foster et al., 2018, Theorem 3), there exists an (improper learning) algorithm for online logistic regression with the following regret:

$$\zeta_3(t) \;\leq\; 10n\, \log\Big(e + \frac{St}{4n}\Big). \tag{34}$$

Though our selected decisions are more conservative compared with (Lee et al., 2024) (add a penalty term when selecting decisions to account for constraint violations), the estimation method to obtain $\hat{\theta}_t$ (i.e., MLE) and the way to compute the optimistic estimate $\tilde{\theta}_t$ (i.e., $\tilde{\theta}_t = \arg\max_{\theta\in\mathcal{C}_t} \max_{\mathbf{d}\in\mathcal{D}_t} \mathbb{E}[Y|do(\mathbf{d}), \mathbf{a}_t, \mathbf{w}_t, \mathbf{m}_t]$) are the same as (Lee et al., 2024). See the justification for using the improper learning algorithm in (Lee et al., 2024, Appendix B.2).

Combining Equation (25), (26), (33), and (34), with $\eta = \frac{1}{2(e-2)+2S} < \frac{1}{2S}$,

$$\begin{aligned}
\mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \;&=\; \zeta_1(t) - \zeta_2(t) + \zeta_3(t) \\
&\leq\; (e-2)\eta \sum_{\tau=1}^{t} \dot{g}(f(\mathbf{Z}_\tau)^\intercal \theta_*)\langle f(\mathbf{Z}_\tau), \tilde{\theta}_\tau - \theta_*\rangle^2 + \frac{1}{\eta}\log\frac{1}{\alpha} + \sum_{\tau=1}^{t} \langle f(\mathbf{Z}_\tau), \theta_* - \tilde{\theta}_\tau\rangle^2 \frac{\dot{g}(f(\mathbf{Z}_\tau)^\intercal \theta_*)}{2+2S} \\
&\quad + 10n\, \log\Big(e + \frac{St}{4n}\Big) \\
&\leq\; 10n \log(\frac{St}{4n} + e) + 2((e-2) + S)\log\frac{1}{\alpha},
\end{aligned}$$

which finishes the proof. ■

F. Logistic Regret Upper Bounds

In this section, we provide the proofs for logistic bandits regret upper bounds presented in Section 3.3. Some of the details follow from (Faury et al., 2020, Appendix B) and (Abeille et al., 2021, Appendix C). We first define the regret of logistic bandits, and use which to prove the regret and constraint violations upper bound in Appendix G for our problem:

$$\mathcal{R}_{\log} \;=\; \sum_{t=1}^{T} \Big[\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big]$$

$$= \sum_{t=1}^{T} \Big[g(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) - g(f(\mathbf{Z}_t)^\intercal \theta_*)\Big] \tag{35}$$

$$= \underbrace{\sum_{t=1}^{T} \Big[\, \dot{g}(f(\mathbf{Z}_t)^\intercal \theta_*)\; f(\mathbf{Z}_t)^\intercal(\tilde{\theta}_t - \theta_*)\, \Big]}_{\mathcal{R}_{\log_1}} + \underbrace{\sum_{t=1}^{T} \Big[\, \int_{f(\mathbf{Z}_t)^\intercal \theta_*}^{f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t} \ddot{g}(u)(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t - u)\, du\, \Big]}_{\mathcal{R}_{\log_2}}, \tag{36}$$

where the (35) comes from the expected reward in (2); the (36) is by performing the Taylor Series Expansion of $g(f(\mathbf{Z}_t)^\intercal\tilde{\theta}_t)$ on $f(\mathbf{Z}_t)^\intercal\theta_*$ with a first order integral remainder. Then we rewrite the logistic regret $\mathcal{R}_{\log}$ as $\mathcal{R}_{\log_1}$ and $\mathcal{R}_{\log_2}$, where,

$$\mathcal{R}_{\log_1} = \sum_{t=1}^{T} \Big[\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)\, f(\mathbf{Z}_t)^\intercal(\tilde{\theta}_t - \theta_*) \Big]$$
$$\mathcal{R}_{\log_2} = \sum_{t=1}^{T} \Big[\int_{f(\mathbf{Z}_t)^\intercal\theta_*}^{f(\mathbf{Z}_t)^\intercal\tilde{\theta}_t} \ddot{g}(u)(f(\mathbf{Z}_t)^\intercal\tilde{\theta}_t - u)\, du \Big].$$

We separately upper bound both terms. Firstly, we prove the following Lemma used throughout this section.

Lemma 7. *With $\lambda_t = \frac{1}{4S^2(2+2S)}$, for any $\theta \in \mathcal{C}_t(\alpha)$, the following holds with probability at least $1-\alpha$:*

$$\|\theta - \theta_*\|^2_{\mathbf{H}_t(\theta_*)} \le \gamma_t(\alpha)^2 = \mathcal{O}\Big(S^2\Big(n\log\big(e + \frac{St}{4n}\big) + \log\frac{1}{\alpha}\Big)\Big).$$

Proof. By proposition 1, we have that with probability at least $1-\alpha$, $\mathcal{L}_t(\theta_*) - \mathcal{L}_t(\hat{\theta}_t) \le \beta_t(\alpha)^2$, we assume this event is true throughout this proof. Then, by second-order Taylor expansion of $\mathcal{L}_t(\theta)$ around θ_*,

$$\begin{aligned}\mathcal{L}_t(\theta) &= \mathcal{L}_t(\theta_*) + \nabla\mathcal{L}_t(\theta_*)^\intercal(\theta-\theta_*) + \|\theta-\theta_*\|^2_{\mathbf{G}_t(\theta,\theta_*)-\lambda_t\mathbf{I}}\cdot \\ &= \mathcal{L}_t(\theta_*) + \nabla\mathcal{L}_t(\theta_*)^\intercal(\theta-\theta_*) + \|\theta-\theta_*\|^2_{\mathbf{G}_t(\theta,\theta_*)} - \lambda_t\|\theta-\theta_*\|_2^2.\end{aligned}$$

As for the relationship between $\mathbf{G}_t(\theta,\theta_*)$ and $\mathbf{H}_t(\theta_*)$, we have the following result:

$$\begin{aligned}\mathbf{G}_t(\theta,\theta_*) &= \sum_{\tau=1}^{t-1}\int_{v=0}^{1}(1-v)\dot{g}\Big(f(\mathbf{Z}_\tau)^\intercal\theta + v\, f(\mathbf{Z}_\tau)^\intercal(\theta_*-\theta)\Big)\, dv\, f(\mathbf{Z}_\tau)\, f(\mathbf{Z}_\tau)^\intercal + \lambda_t\mathbf{I}_n \\ &\succeq \sum_{\tau=1}^{t-1}\frac{\dot{g}\Big(f(\mathbf{Z}_\tau)^\intercal\theta_*\Big)}{2+|f(\mathbf{Z}_\tau)^\intercal\theta - f(\mathbf{Z}_\tau)^\intercal\theta_*|}\, f(\mathbf{Z}_\tau)\, f(\mathbf{Z}_\tau)^\intercal + \lambda_t\mathbf{I}_n \\ &\succeq \frac{1}{2+2S}\sum_{\tau=1}^{t-1}\dot{g}\Big(f(\mathbf{Z}_\tau)^\intercal\theta_*\Big)\, f(\mathbf{Z}_\tau)\, f(\mathbf{Z}_\tau)^\intercal + \lambda_t\mathbf{I}_n \\ &\succeq \frac{1}{2+2S}\mathbf{H}_t(\theta_*),\end{aligned}$$

Thus, we have that,

$$\begin{aligned}\|\theta-\theta_*\|^2_{\mathbf{H}_t(\theta_*)} &\le (2+2S)\|\theta-\theta_*\|^2_{\mathbf{G}_t(\theta,\theta_*)} \\ &= (2+2S)\big(\mathcal{L}_t(\theta) - \mathcal{L}_t(\theta_*) + \nabla\mathcal{L}_t(\theta_*)^\intercal(\theta_*-\theta) + \lambda_t\|\theta-\theta_*\|_2^2\big) \\ &\le (2+2S)\big(\mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) + \nabla\mathcal{L}_t(\theta_*)^\intercal(\theta_*-\theta) + \lambda_t\|\theta-\theta_*\|_2^2\big) \\ &\le 1 + (2+2S)\beta_t(\alpha)^2 + (2+2S)\nabla\mathcal{L}_t(\theta_*)^\intercal(\theta_*-\theta). \qquad (37)\end{aligned}$$

Where (37) follows by $\lambda_t = \frac{1}{4S^2(2+2S)}$. The next is to bound $\nabla\mathcal{L}_t(\theta_*)^\intercal(\theta_*-\theta)$, which is done via a new concentration-type argument. Let $\mathcal{B}_n(2S)$ be a n ball of radius $2S$ and $v \in \mathcal{B}_n(2S)$. Since

$$\nabla\mathcal{L}_t(\theta_*)^\intercal v = \sum_{\tau=1}^{t}\big(g(f(\mathbf{Z}_\tau)^\intercal\theta_*) - y_\tau\big)f(\mathbf{Z}_\tau)^\intercal v = \sum_{\tau=1}^{t}\zeta_\tau f(\mathbf{Z}_\tau)^\intercal v.$$

As $|\zeta_\tau f(\mathbf{Z}_\tau)^\intercal v| < 2S$ and $\mathbf{E}[(\zeta_\tau f(\mathbf{Z}_\tau)^\intercal v)^2|\mathcal{F}_{\tau-1}] = \dot{g}(f(\mathbf{Z}_\tau)^\intercal\theta_*)(f(\mathbf{Z}_\tau)^\intercal v)^2$, by Freedman's inequality (26), for any $\eta \in [0, \frac{1}{2BS}]$, the following holds:

$$\mathbb{P}\Big[\sum_{\tau=1}^{t}\zeta_\tau f(\mathbf{Z}_\tau)^\intercal v \le (e-2)\eta\sum_{\tau=1}^{t}\dot{g}(f(\mathbf{Z}_\tau)^\intercal\theta_*)(f(\mathbf{Z}_\tau)^\intercal v)^2 + \frac{1}{\eta}\log\frac{1}{\alpha}\Big] \ge 1-\alpha$$

By (Vershynin, Corollary 4.2.13) and (Lee et al., 2024, Appendix C.4.4) the following holds with probability at least $1-\alpha$:

$$\nabla\mathcal{L}_t(\theta_*)^\intercal(\theta_*-\theta) \;\le\; (e-2)\eta||\theta_*-\theta||^2_{\mathbf{H}_t(\theta_*)} + \frac{1}{\eta}\log\frac{1}{\alpha} + \frac{n}{\eta}\log\frac{5S}{\epsilon_t} + \Big(\frac{(e-2)}{4}\big(4S\eta+\eta\epsilon_t\big)+1\Big)\epsilon_t t.$$

Choose $\eta = \frac{1}{2(e-2)(2+2S)} < \frac{1}{2S}$, $\epsilon_t = \frac{n}{t}$, and with Equation (37), we finally have:

$$||\theta-\theta_*||^2_{\mathbf{H}_t(\theta_*)} \;=\; \mathcal{O}\Big(nS^2\log\Big(e+\frac{St}{4n}\Big) + S^2\log\frac{1}{\alpha}\Big).$$

Which finishes the proof. ■

F.1. The regret upper bound of $\mathcal{R}_{\log_1}$.

We start by examining $\mathcal{R}_{\log_1}$ and show the following upper bounds:

$$\begin{aligned}
\mathcal{R}_{\log_1} &= \sum_{t=1}^{T}\Big[\,\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)\; f(\mathbf{Z}_t)^\intercal(\tilde{\theta}_t-\theta_*)\,\Big] && (38)\\
&\le \sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)\;||f(\mathbf{Z}_t)||_{\mathbf{H}_t^{-1}(\theta_*)}\|\tilde{\theta}_t-\theta_*\|_{\mathbf{H}_t(\theta_*)} && (39)\\
&\le \sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)\;||f(\mathbf{Z}_t)||_{\mathbf{H}_t^{-1}(\theta_*)}\;\gamma_t(\alpha) && (40)\\
&\le \gamma_T(\alpha)\sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)\;||f(\mathbf{Z}_t)||_{\mathbf{H}_t^{-1}(\theta_*)} && (41)\\
&\le \gamma_T(\alpha)\sqrt{\sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)}\sqrt{\sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)\;||f(\mathbf{Z}_t)||^2_{\mathbf{H}_t^{-1}(\theta_*)}} && (42)\\
&\le \gamma_T(\alpha)\sqrt{\sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)}\sqrt{\sum_{t=1}^{T}||\mathbf{u}_t||^2_{\tilde{\mathbf{V}}_t^{-1}}} && (43)\\
&\le 2\gamma_T(\alpha)\sqrt{\sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)}\sqrt{n\;\log\Big(\lambda_T+\frac{T}{n}\Big)}. && (44)
\end{aligned}$$

Where (39) and (42) is by Cauchy-Schwarz inequality ($\dot{g}\big(f(\mathbf{Z}_t)^\intercal\theta_*\big)$ is non-negative); (40) comes from Lemma 7 and (41) is because $\gamma_T(\alpha) = \max_{t\in[T]}\gamma_t(\alpha)$; in (43), we define vector $\mathbf{u}_t = \sqrt{\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)}\;f(\mathbf{Z}_t)$ and matrix $\tilde{\mathbf{V}}_t = \sum_{\tau=1}^{t-1}\mathbf{u}_t\,\mathbf{u}_t^\intercal + \lambda_t\mathbf{I}_n$, and obtain:

$$\begin{aligned}
\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)\;||f(\mathbf{Z}_t)||^2_{\mathbf{H}_t^{-1}(\theta_*)} &= ||\sqrt{\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)}\;f(\mathbf{Z}_t)||^2_{\mathbf{H}_t^{-1}(\theta_*)}\\
&= ||\mathbf{u}_t||^2_{\tilde{\mathbf{V}}_t^{-1}};
\end{aligned}$$

and (44) follows by Lemma 2.

We then take a look at the first order of the logistic function $\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*)$ and derive a upper bound for it by a first-order Taylor expansion:

$$\begin{aligned}
\sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\theta_*) &= \sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\tilde{\theta}_t) + \sum_{t=1}^{T}\int_{f(\mathbf{Z}_t)^\intercal\tilde{\theta}_t}^{f(\mathbf{Z}_t)^\intercal\theta_*}\ddot{g}(u)\,du && (45)\\
&= \sum_{t=1}^{T}\dot{g}(f(\mathbf{Z}_t)^\intercal\tilde{\theta}_t) + \sum_{t=1}^{T}\Big[\int_{v=0}^{1}\ddot{g}\Big(f(\mathbf{Z}_t)^\intercal\tilde{\theta}_t + vf(\mathbf{Z}_t)^\intercal(\theta_*-\tilde{\theta}_t)\Big)\,dv\Big]f(\mathbf{Z}_t)^\intercal(\theta_*-\tilde{\theta}_t) && (46)
\end{aligned}$$

$$\leq \sum_{t=1}^{T} \dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) + \sum_{t=1}^{T} \left| \left[\int_{v=0}^{1} \ddot{g}\Big(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t + v f(\mathbf{Z}_t)^\intercal (\theta_* - \tilde{\theta}_t)\Big)\, dv \right] f(\mathbf{Z}_t)^\intercal (\theta_* - \tilde{\theta}_t) \right| \tag{47}$$

$$\leq \sum_{t=1}^{T} \dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) + \sum_{t=1}^{T} \left| \left[\int_{v=0}^{1} \ddot{g}\Big(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t + v f(\mathbf{Z}_t)^\intercal (\theta_* - \tilde{\theta}_t)\Big)\, dv \right] \right| f(\mathbf{Z}_t)^\intercal (\tilde{\theta}_t - \theta_*) \tag{48}$$

$$= \sum_{t=1}^{T} \dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) + \sum_{t=1}^{T} \left[\int_{v=0}^{1} \left| \ddot{g}\Big(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t + v f(\mathbf{Z}_t)^\intercal (\theta_* - \tilde{\theta}_t)\Big) \right| dv \right] f(\mathbf{Z}_t)^\intercal (\tilde{\theta}_t - \theta_*) \tag{49}$$

$$\leq \sum_{t=1}^{T} \dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) + \sum_{t=1}^{T} \left[\int_{v=0}^{1} \dot{g}\Big(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t + v f(\mathbf{Z}_t)^\intercal (\theta_* - \tilde{\theta}_t)\Big)\, dv \right] f(\mathbf{Z}_t)^\intercal (\tilde{\theta}_t - \theta_*) \tag{50}$$

$$= \sum_{t=1}^{T} \dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) + \sum_{t=1}^{T} \alpha(f(\mathbf{Z}_t), \tilde{\theta}_t, \theta_*) f(\mathbf{Z}_t)^\intercal (\tilde{\theta}_t - \theta_*) \tag{51}$$

$$= \sum_{t=1}^{T} \dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) + \sum_{t=1}^{T} \left[g\Big(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t\Big) - g\Big(f(\mathbf{Z}_t)^\intercal \theta_*\Big) \right] \tag{52}$$

$$= \sum_{t=1}^{T} \dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) + \mathcal{R}_{\log} \tag{53}$$

$$\leq T + \mathcal{R}_{\log}. \tag{54}$$

Where (45) comes from the Taylor Expansion; (46) follows by changing variables; (47) is by taking the absolute value; (48) is because the optimistic estimate at step t, hence, $f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t \geq f(\mathbf{Z}_t)^\intercal \theta_*$; (50) follows by the self-concordance property of logistic function $|\dot{g}| \geq |\ddot{g}|$ and $\dot{g} > 0$; (51) follows by (16) and (52) is from the fundamental theorem of calculus; and (54) follows by $\dot{g}(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t) \leq 1$.

Therefore, by (44) and (54), we intermediately obtain the following upper bound on $\mathcal{R}_{\log_1}$:

$$\begin{aligned} \mathcal{R}_{\log_1} &\leq 2\gamma_T(\alpha) \sqrt{n \, \log\left(\lambda_T + \frac{T}{n}\right)} \sqrt{T + \mathcal{R}_{\log}} \\ &\leq 2\gamma_T(\alpha) \sqrt{n \, \log\left(\lambda_T + \frac{T}{n}\right)} \Big(\sqrt{T} + \sqrt{\mathcal{R}_{\log}}\Big), \end{aligned} \tag{55}$$

where (55) is because $\sqrt{T + \mathcal{R}_{\log}} \leq \sqrt{T} + \sqrt{\mathcal{R}_{\log}}$ for $T > 0,\ \mathcal{R}_{\log} > 0$.

F.2. The regret upper bounds of $\mathcal{R}_{\log_2}$.

In order to upper bound the logistic bandits regret $\mathcal{R}_{\log}$ in (36), we still need to upper bound $\mathcal{R}_{\log_2}$ that includes the second order of logistic function:

$$\mathcal{R}_{\log_2} = \sum_{t=1}^{T} \left[\int_{f(\mathbf{Z}_t)^\intercal \theta_*}^{f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t} \ddot{g}(u)(f(\mathbf{Z}_t)^\intercal \tilde{\theta}_t - u)\, du \right] \tag{56}$$

$$= \sum_{t=1}^{T} \left[\int_{v=0}^{1} (1-v) \ddot{g}\Big(f(\mathbf{Z}_t)^\intercal \theta_* + v\, f(\mathbf{Z}_t)^\intercal (\tilde{\theta}_t - \theta_*)\Big)\, dv \right] \Big(f(\mathbf{Z}_t)^\intercal (\tilde{\theta}_t - \theta_*)\Big)^2 \tag{57}$$

$$\leq \sum_{t=1}^{T} \frac{1}{2} \Big(f(\mathbf{Z}_t)^\intercal (\tilde{\theta}_t - \theta_*)\Big)^2 \tag{58}$$

$$\leq \sum_{t=1}^{T} \frac{1}{2} \|f(\mathbf{Z}_t)\|^2_{\mathbf{H}_t^{-1}(\theta_*)} \|\tilde{\theta}_t - \theta_*\|^2_{\mathbf{H}_t(\theta_*)} \tag{59}$$

$$\leq \sum_{t=1}^{T} \frac{1}{2} \|f(\mathbf{Z}_t)\|^2_{\mathbf{H}_t^{-1}(\theta_*)} \gamma_t^2(\alpha) \tag{60}$$

$$\leq \frac{1}{2}\gamma_T^2(\alpha)\sum_{t=1}^{T}\|f(\mathbf{Z}_t)\|^2_{\mathbf{H}_t^{-1}(\theta_*)} \tag{61}$$

$$\leq \frac{1}{2}\gamma_T^2(\alpha)\,\boldsymbol{\kappa}_{\mathcal{Z}}\sum_{t=1}^{T}\|f(\mathbf{Z}_t)\|^2_{\mathbf{V}_t^{-1}(\theta_*)} \tag{62}$$

$$\leq 2n\,\gamma_T^2(\alpha)\,\boldsymbol{\kappa}_{\mathcal{Z}}\log\Big(\lambda_T+\frac{T}{n}\Big). \tag{63}$$

Where (57) follows by changing variables; (58) is because $\ddot{g}\leq 1$; (59) follows by Cauchy-Schwarz inequality; (60) comes from Lemma 7; (61) is by $\mathcal{C}_T(\alpha)=\max_{t\in[T]}\mathcal{C}_t(\alpha)$; as for (62), we note that

$$\begin{aligned}\mathbf{H}_t(\theta_*) &= \sum_{\tau=1}^{t-1}\dot{g}\Big(f(\mathbf{Z}_\tau)^\intercal\theta_*\Big)f(\mathbf{Z}_\tau)\,f(\mathbf{Z}_\tau)^\intercal+\lambda_t\mathbf{I}_n\\ &\succeq \frac{1}{\boldsymbol{\kappa}_{\mathcal{Z}}}\Big[\sum_{\tau=1}^{t-1}f(\mathbf{Z}_\tau)\,f(\mathbf{Z}_\tau)^\intercal+\lambda_t\mathbf{I}_n\Big]\\ &= \frac{1}{\boldsymbol{\kappa}_{\mathcal{Z}}}\mathbf{V}_t(\theta_*),\end{aligned}$$

where the second line comes from Definition 1. Thus,

$$\mathbf{H}_t^{-1}(\theta_*)\preceq\sqrt{\boldsymbol{\kappa}_{\mathcal{Z}}}\mathbf{V}_t^{-1}(\theta_*);$$

and (63)follows by Lemma 2.

Then by the upper bounds on $\mathcal{R}_{\log_1}$ in Equation (55) and $\mathcal{R}_{\log_2}$ in Equation (63), we then finally upper bound the logistic bandits regret $\mathcal{R}_{\log}$ in Equation (36):

$$\mathcal{R}_{\log}=\mathcal{R}_{\log_1}+\mathcal{R}_{\log_2} \tag{64}$$

$$\leq 2\gamma_T(\alpha)\sqrt{n\,\log\Big(\lambda_T+\frac{T}{n}\Big)}\Big(\sqrt{T}+\sqrt{\mathcal{R}_{\log}}\Big)\;+\;2n\,\gamma_T^2(\alpha)\,\boldsymbol{\kappa}_{\mathcal{Z}}\log\Big(\lambda_T+\frac{T}{n}\Big) \tag{65}$$

$$\leq \Bigg[2\gamma_T(\alpha)\sqrt{n\,\log\Big(\lambda_T+\frac{T}{n}\Big)}\;+\;\sqrt{2\gamma_T(\alpha)\sqrt{n\,\log\Big(\lambda_T+\frac{T}{n}\Big)}\,\sqrt{T}+2n\,\gamma_T^2(\alpha)\,\boldsymbol{\kappa}_{\mathcal{Z}}\log\Big(\lambda_T+\frac{T}{n}\Big)}\Bigg]^2 \tag{66}$$

$$\leq 8n\gamma_T^2(\alpha)\log\Big(\lambda_T+\frac{T}{n}\Big)+4\gamma_T(\alpha)\sqrt{n\log\Big(\lambda_T+\frac{T}{n}\Big)}\sqrt{T}+4n\gamma_T^2(\alpha)\,\boldsymbol{\kappa}_{\mathcal{Z}}\log\Big(\lambda_T+\frac{T}{n}\Big), \tag{67}$$

where (66) follows by Lemma 4; and (67) comes from $(x+y)^2\leq 2x^2+2y^2$. To further simplify the logistic bandits regret $\mathcal{R}_{\log}$, we write $\gamma_t(\alpha)$ as:

$$\gamma_t(\alpha)=\mathcal{O}\Big(S\sqrt{\Big(n\log\big(e+\frac{St}{4n}\big)+\log\frac{1}{\alpha}\Big)}\Big).$$

To get an intuitive understanding on how $\gamma_t(\alpha)$ behaves when t grows, we write $\gamma_T(\alpha)$ as an asymptotic notation of T, i.e., $\gamma_T(\alpha)=\mathcal{O}(S\sqrt{n\log(T)})$. Therefore, as for the $\mathcal{R}_{\log}$, we obtain the following bounds:

$$\begin{aligned}\mathcal{R}_{\log} &\leq 8n\gamma_T^2(\alpha)\log\Big(\lambda_T+\frac{T}{n}\Big)+4\gamma_T(\alpha)\sqrt{n\log\Big(\lambda_T+\frac{T}{n}\Big)}\sqrt{T}+4n\gamma_T^2(\alpha)\,\boldsymbol{\kappa}_{\mathcal{Z}}\log\Big(\lambda_T+\frac{T}{n}\Big)\\ &= \mathcal{O}\Big(nS\log(T)\sqrt{T}+n^2S^2(\log(T))^2+n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2\Big),\end{aligned} \tag{68}$$

which finishes the proof. ■

G. Regret guarantee and constraint violations

In this section, we provide proofs for upper bounds of both reward regret and constraint violations. Our proofs build on the greedy procedure in Algorithm 1 and standard convex optimization analysis.

G.1. Proof of Theorem 1

We first prove the regret upper bound. Under Slater's constraint qualification in Assumption 3, we have the boundedness of the optimal dual solution by standard convex optimization analysis from (Beck, 2017, Theorem 8.42), where,

$$0 \;\leq\; \phi_* \;\leq\; \frac{\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}_{\pi_{t,0}}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]}{\delta} \;\leq\; \frac{1}{\delta},$$

the r.h.s. is because the logistic function is less than 1. Now, we turn to establish a bound over $\mathcal{R}_+(T) + \phi\mathcal{V}(T)$. First, note that,

$$\begin{aligned}
&\mathcal{R}_+(T) + \phi\mathcal{V}(T) \\
&= \sum_{t=1}^{T}\Big[\,\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \phi(|\Delta(\mathbf{d}_t)| - \tau)\,\Big] && (69)\\
&\leq \sum_{t=1}^{T}\Big[\,\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \phi(|\Delta(\mathbf{d}_t)| - \tau) - \phi_t\mathbb{E}_{\pi_t^*}(|\Delta(\mathbf{d})| - \tau)\,\Big] && (70)\\
&= \underbrace{\sum_{t=1}^{T}\Big(\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t\mathbb{E}_{\pi_t^*}(|\Delta(\mathbf{d})| - \tau)\Big) - \Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t(|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big)}_{\mathcal{R}_1} + \\
&\quad \underbrace{\sum_{t=1}^{T}\Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi\Big((|\Delta(\mathbf{d}_t)| - \tau) - (|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big)\Big]}_{\mathcal{R}_2} + \\
&\quad \underbrace{\sum_{t=1}^{T}\Big[\phi(|\widehat{\Delta}(\mathbf{d}_t)| - \tau) - \phi_t(|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big]}_{\mathcal{R}_3}. && (71)\\
&\leq \underbrace{\sum_{t=1}^{T}\Big(\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t\mathbb{E}_{\pi_t^*}(|\Delta(\mathbf{d})| - \tau)\Big) - \Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t(|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big)}_{\mathcal{R}_1} + \\
&\quad \underbrace{\sum_{t=1}^{T}\Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi\Big((|\Delta(\mathbf{d}_t)| - \tau) - (|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big)\Big]}_{\mathcal{R}_2} + \sqrt{T}\rho. && (72)
\end{aligned}$$

Where (69) holds since $\phi_t \geq 0$ and $\mathbb{E}_{\pi_t^*}(|\Delta(\mathbf{d}_t)| - \tau) \leq 0$; (71) holds by adding and subtracting $\sum_{t=1}^{T}\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$, $\sum_{t=1}^{T}\phi_t(|\widehat{\Delta}(\mathbf{d}_t)| - \tau)$, $\sum_{t=1}^{T}\phi(|\widehat{\Delta}(\mathbf{d}_t)| - \tau)$; and (72) comes from Lemma 8.

We are then going to bound $\mathcal{R}_1$:

$$\begin{aligned}
\mathcal{R}_1 &= \sum_{t=1}^{T}\Big(\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t\mathbb{E}_{\pi_t^*}(|\Delta(\mathbf{d})| - \tau)\Big) - \Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t(|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big) \\
&= \sum_{t=1}^{T}\mathbb{E}_{\pi_t^*}\Big(\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi_t \cdot \mathbb{E}_{\pi_t^*}\Big((|\widehat{\Delta}(\mathbf{d})| - \tau) - (|\Delta(\mathbf{d})| - \tau)\Big) + \\
&\qquad \mathbb{E}_{\pi_t^*}\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t \cdot (|\widehat{\Delta}(\mathbf{d})| - \tau)\Big) - \Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t \cdot (|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} \mathbb{E}_{\pi_t^*}\Big(\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \sum_{t=1}^{T} \phi_t \mathbb{E}_{\pi_t^*}\Big((|\widehat{\Delta}(\mathbf{d})| - \tau) - (|\Delta(\mathbf{d})| - \tau)\Big) \\
&\leq \sum_{t=1}^{T} \phi_t \cdot \mathbb{E}_{\pi_t^*}\Big((|\widehat{\Delta}(\mathbf{d})| - \tau) - (|\Delta(\mathbf{d})| - \tau)\Big) \\
&\leq \rho \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t^*}\Big((|\widehat{\Delta}(\mathbf{d})| - \tau) - (|\Delta(\mathbf{d})| - \tau)\Big) \\
&= \rho \cdot \mathcal{O}\Big(nS\log(T)\sqrt{T} + n^2S^2(\log(T))^2 + n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2\Big). \qquad (73)
\end{aligned}
$$

Where the second equality follows by adding and subtracting terms; the third inequality comes from the greedy action $\mathbf{d}_t$ chosen at step t in Algorithm 1; the forth inequality comes from the optimistic estimate $\tilde{\theta}_t$ in Algorithm 1; and the fifth inequality is because $\phi_t \leq \rho$; as for the result in the last line, we note that if $\widehat{\Delta}(\mathbf{d}) \geq 0$ then $\Delta(\mathbf{d}) \geq 0$, and by the counterfactual fairness effect (see Equation (4)), we have,

$$
\begin{aligned}
(|\widehat{\Delta}(\mathbf{d})| - \tau) - (|\Delta(\mathbf{d})| - \tau) &= \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] \\
&\quad - \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] \\
&\leq \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \\
&= g(f(\mathbf{Z}_t)\tilde{\theta}_t) - g(f(\mathbf{Z}_t)\theta_*),
\end{aligned}
$$

where the second line follows by optimistic estimation. On the another hand, if $\widehat{\Delta}(\mathbf{d}) < 0$ then $\Delta(\mathbf{d}) < 0$,

$$
\begin{aligned}
(|\widehat{\Delta}(\mathbf{d})| - \tau) - (|\Delta(\mathbf{d})| - \tau) &= \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \\
&\quad + \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] \\
&\leq \mathbb{E}[\widehat{Y}|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] \\
&= g(f(\mathbf{Z}_{\mathbf{a}_t'})\tilde{\theta}_t) - g(f(\mathbf{Z}_{\mathbf{a}_t'})\theta_*),
\end{aligned}
$$

similarly, the second line follows by the optimistic estimate, here, we notice that, the factual feature $\mathbf{Z}_t$ and the counterfactual feature $\mathbf{Z}_{\mathbf{a}_t'}$ reside within the feature space $\mathcal{Z}$, in which the boundness assumption (Assumption 1) and problem dependent constant (Definition 1) are defined by. Therefore, the logistic bandits regret of $g(f(\mathbf{Z}_t)\tilde{\theta}_t) - g(f(\mathbf{Z}_t)\theta_*)$ has the same asymptotic upper bound up to logarithmic factors as $g(f(\mathbf{Z}_{\mathbf{a}_t'})\tilde{\theta}_t) - g(f(\mathbf{Z}_{\mathbf{a}_t'})\theta_*)$ (see Equation (68)).

We further bound $\mathcal{R}_2$. When $\widehat{\Delta}(\mathbf{d}) \geq 0$:

$$
\begin{aligned}
\mathcal{R}_2 &= \sum_{t=1}^{T} \Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi\Big((|\Delta(\mathbf{d}_t)| - \tau) - (|\widehat{\Delta}(\mathbf{d}_t)| - \tau)\Big)\Big] \\
&= \sum_{t=1}^{T} \Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi\Big(\mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \\
&\qquad \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t]\Big) - \phi\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t]\Big)\Big] \\
&= \sum_{t=1}^{T} \Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi\Big(\mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \\
&\qquad \mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t]\Big)\Big] \\
&\leq \sum_{t=1}^{T} \Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) + \phi\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t] - \\
&\qquad \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t'), \mathbf{w}_t, \mathbf{m}_t]\Big)\Big].
\end{aligned}
$$

Where the second equality follows by the counterfactual fairness effect (see Equation (4)); the forth inequality follows by the optimistic estimate.

When $\widehat{\Delta}(\mathbf{d}) < 0$, we have that:

$$\begin{aligned}
\mathcal{R}_2 &= \sum_{t=1}^{T}\Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]-\mathbb{E}[Y|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]\Big)+\phi\Big((|\Delta(\mathbf{d}_t)|-\tau)-(|\widehat{\Delta}(\mathbf{d}_t)|-\tau)\Big)\Big]\\
&= \sum_{t=1}^{T}\Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]-\mathbb{E}[Y|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]\Big)+\phi\Big(-\mathbb{E}[Y|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]+\\
&\qquad \mathbb{E}[Y|do(\mathbf{d}_t),do(\mathbf{a}'_t),\mathbf{w}_t,\mathbf{m}_t]\Big)-\phi\Big(-\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]+\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t),do(\mathbf{a}'_t),\mathbf{w}_t,\mathbf{m}_t]\Big)\Big]\\
&\leq \sum_{t=1}^{T}\Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]-\mathbb{E}[Y|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]\Big)+\phi\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]-\\
&\qquad \mathbb{E}[Y|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]\Big)\Big].
\end{aligned}$$

Since the logistic bandits regret of $g(f(\mathbf{Z}_t)\tilde{\theta}_t)-g(f(\mathbf{Z}_t)\theta_*)$ has the same asymptotic upper bound up to logarithmic factors as $g(f(\mathbf{Z}_{\mathbf{a}'_t})\tilde{\theta}_t)-g(f(\mathbf{Z}_{\mathbf{a}'_t})\theta_*)$ (see discussions on Equation (73)), thus

$$\mathcal{R}_2 \leq (1+\phi)\cdot\mathcal{O}\Big(nS\log(T)\sqrt{T}+n^2S^2(\log(T))^2+n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2\Big). \tag{74}$$

Thus, by Equation (73) and Equation (74), the upper bound on $\mathcal{R}_+(T)+\phi\mathcal{V}(T)$ for any $\phi\in[0,\rho]$ is the following:

$$\begin{aligned}
\mathcal{R}_+(T)+\phi\mathcal{V}(T) &= (1+\phi)\cdot\mathcal{O}\Big(nS\log(T)\sqrt{T}+n^2S^2(\log(T))^2+n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2\Big)+\\
&\qquad \rho\cdot\mathcal{O}\Big(nS\log(T)\sqrt{T}+n^2S^2(\log(T))^2+n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2\Big)+\mathcal{O}(\rho\sqrt{T}).
\end{aligned}$$

Regret $\mathcal{R}_+(T)$. By setting $\phi=0$, then we obtain the upper bounds on the total regret guarantee with high probability:

$$\begin{aligned}
\mathcal{R}_+(T) &= (\rho+1)\cdot\mathcal{O}\Big(nS\log(T)\sqrt{T}+n^2S^2(\log(T))^2+n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2\Big)+\mathcal{O}(\rho\sqrt{T})\\
&= \mathcal{O}\Big(\rho\Big(nS\log(T)\sqrt{T}+n^2S^2(\log(T))^2+n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2\Big)+\rho\sqrt{T}\Big).\\
&= \widetilde{\mathcal{O}}\Big(\rho(nS\sqrt{T}+n^2S^2)\boldsymbol{\kappa}_{\mathcal{Z}}+\rho\sqrt{T}\Big),
\end{aligned} \tag{75}$$

where the second line is because the truncated parameter $\rho\geq 2/\delta$; and the last line is to write the regret upper bound in a logarithmic asymptotic notation.

Constraint violations. Next, to obtain a bound on $\mathcal{V}(T)$, we employ tools from constrained convex optimization. First, we define probability distribution π'_t by

$$\mathbb{E}_{\pi'_t}\mathbb{E}[Y|do(\mathbf{d}),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t] = \mathbb{E}[Y|do(\mathbf{d}_t),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t];\quad \mathbb{E}_{\pi'_t}(|\Delta(\mathbf{d})|-\tau) = (|\Delta(\mathbf{d}_t)|-\tau),$$

Thus,

$$\mathcal{R}_+(T)+\phi\mathcal{V}(T) = \sum_{t=1}^{T}\Big[\mathbb{E}_{\pi^*_t}\mathbb{E}[Y|do(\mathbf{d}),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]-\mathbb{E}_{\pi'_t}\mathbb{E}[Y|do(\mathbf{d}),do(\mathbf{a}_t),\mathbf{w}_t,\mathbf{m}_t]+\phi\mathbb{E}_{\pi'_t}(|\Delta(\mathbf{d})|-\tau)\Big]. \tag{76}$$

Then we apply the following theorem from Theorem 3.60 in (Beck, 2017).

Theorem 3. *Consider the following convex constrained problem $f(\pi^*) = \max_{\pi\in\Pi}\{f(\pi):\ g(\pi)\le 0\}$, where both f and g are convex over the convex set Π in a vector space. Suppose $f(\pi^*)$ is finite and there exists a slater point π_0 such that $g(\pi_0)\le -\delta$, and a constant $\rho \ge 2\phi_*$ where ϕ_* is the optimal dual variable, i.e., $\phi_* = \text{argmin}_{\phi\ge 0}(\max_\pi f(\pi) - \phi g(\pi))$. Assume that $\pi' \in \Pi$ satisfies*

$$f(\pi^*) - f(\pi') + \rho[g(\pi')]_+ \ \le\ \epsilon,$$

for some $\epsilon > 0$, then we have $[g(\pi')]_+ \le \frac{2\epsilon}{\rho}$.

Since $\sum_{t=1}^T \mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$ is convex over $\{\pi_t^*\}_{t=1}^T$, $\sum_{t=1}^T \mathbb{E}_{\pi_t'}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$ and $\sum_{t=1}^T \mathbb{E}_{\pi_t'}(|\Delta(\mathbf{d}_t)| - \tau)$ are convex over $\{\pi_t'\}_{t=1}^T$. Then Equation (76) satisfies the conditions in Theorem 3 and we have that:

$$\begin{aligned}
\mathcal{V}(T) \ &= \ \sum_{t=1}^T (|\Delta(\mathbf{d}_t)| - \tau) \\
&= \ \mathcal{O}\Big(nS\log(T)\sqrt{T} + n^2S^2(\log(T))^2 + n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}}(\log(T))^2 + \sqrt{T}\Big). \\
&= \ \widetilde{\mathcal{O}}\Big(nS\sqrt{T} + n^2S^2\boldsymbol{\kappa}_{\mathcal{Z}} + \sqrt{T}\Big).
\end{aligned}$$

Where the second line follows by $\phi \in [0, \rho]$ and $1/\rho < 1$. ■

Lemma 8. *Under the dual update of ϕ_t in Algorithm 1, we have the following for any $\phi \in [0, \rho]$:*

$$\sum_{t=1}^T (\phi - \phi_t)(|\widehat{\Delta}(\mathbf{d}_t)| - \tau) \ \le\ \frac{1}{2\eta}(\phi_1 - \phi)^2 + \sum_{t=1}^T \frac{\eta}{2}((|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2$$

Proof. By the dual update of ϕ_t in Algorithm 1:

$$\begin{aligned}
(\phi_{t+1} - \phi)^2 \ &= \ (\phi_t \ + \ \frac{1}{\eta}(|\widehat{\Delta}(\mathbf{d}_t)| - \tau) - \phi)^2 \\
&= \ (\phi_t - \phi)^2 + (\frac{1}{\eta}(|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2 + 2(\phi_t - \phi)(\frac{1}{\eta}[|\widehat{\Delta}(\mathbf{d}_t)| - \tau]) \\
&= \ (\phi_t - \phi)^2 + \frac{2}{\eta}\ (\phi_t - \phi)(|\widehat{\Delta}(\mathbf{d}_t)| - \tau) + (\frac{1}{\eta}(|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2.
\end{aligned}$$

Summing over T steps and multiplying both sides by $\frac{\eta}{2}$:

$$\sum_{t=1}^T \frac{\eta}{2}(\phi_{t+1} - \phi)^2 \ = \ \sum_{t=1}^T \frac{\eta}{2}(\phi_t - \phi)^2 + \sum_{t=1}^T (\phi_t - \phi)(|\widehat{\Delta}(\mathbf{d}_t)| - \tau) + \sum_{t=1}^T \frac{1}{2\eta}((|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2.$$

Therefore:

$$\begin{aligned}
\sum_{t=1}^T (\phi - \phi_t)(|\widehat{\Delta}(\mathbf{d}_t)| - \tau) \ &= \ \sum_{t=1}^T \frac{\eta}{2}(\phi_t - \phi)^2 - \sum_{t=1}^T \frac{\eta}{2}(\phi_{t+1} - \phi)^2 + \sum_{t=1}^T \frac{1}{2\eta}((|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2 \\
&= \ \frac{\eta}{2}(\phi_1 - \phi)^2 - \frac{\eta}{2}(\phi_{T+1} - \phi)^2 + \sum_{t=1}^T \frac{1}{2\eta}((|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2 \\
&\le \ \frac{\eta}{2}(\phi_1 - \phi)^2 + \sum_{t=1}^T \frac{1}{2\eta}((|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2 \\
&= \ \frac{\sqrt{T}}{2\rho}\phi^2 + \frac{\rho\sqrt{T}}{2}((|\widehat{\Delta}(\mathbf{d}_t)| - \tau))^2
\end{aligned}$$

$$\leq \frac{\sqrt{T}\rho}{2} + \frac{\rho\sqrt{T}}{2}$$
$$= \mathcal{O}(\sqrt{T}\rho),$$

where the second equality comes from telescopic sum; and the forth equality follows by $\phi_1 = 0$ and $\eta = \sqrt{T}/\rho$ initialized in Algorithm 1; and the fifth inequality is because $|\widehat{\Delta}(\mathbf{d}_t)| \in [0,1]$, $\phi \in [0,\rho]$, and $0 \leq \tau \leq 1$. ■

G.2. Proof of Proposition 3

Proposition 3 states the relationship between policy π_t^* and $\pi_{t,\epsilon}^*$ for the regret upper bounds, we provide a proof in the following.

Proposition 3. *Let policies π_t^* and $\pi_{t,\epsilon}^*$ be the optimal solution for constrained problem $\max_{\pi_t}\{\mathbb{E}_{\pi_t}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] : \mathbb{E}_{\pi_t}[\Delta(\mathbf{d}) - \tau] \leq 0\}$ and $\max_{\pi_t}\{\mathbb{E}_{\pi_t}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] : \mathbb{E}_{\pi_t}[\Delta(\mathbf{d}) - \tau + \epsilon] \leq 0\}$, we have,*

$$\sum_{t=1}^{T} \mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \sum_{t=1}^{T} \mathbb{E}_{\pi_{t,\epsilon}^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \leq \frac{\epsilon T}{\delta}$$

Proof. The policies π_t^* and $\pi_{t,\epsilon}^*$ are defined as:

$$\pi_t^* = \max_{\pi_t}\{\mathbb{E}_{\pi_t}[\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]] : \mathbb{E}_{\pi_t}[|\Delta(\mathbf{d})| - \tau] \leq 0\}$$
$$\pi_{t,\epsilon}^* = \max_{\pi_{t,\epsilon}}\{\mathbb{E}_{\pi_{t,\epsilon}}[\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]] : \mathbb{E}_{\pi_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau] + \epsilon \leq 0\}.$$

Let one policy $\pi_{t,\epsilon} = (1 - \frac{\epsilon}{\delta})\pi_t^* + \frac{\epsilon}{\delta}\pi_{t,0}$, where $\pi_{t,0}$ is the policy satisfies the Slater's constrained qualification, i.e., $\mathbb{E}_{\pi_{t,0}}[|\Delta(\mathbf{d})| - \tau] \leq -\delta,\ \forall\, t \in [T]$. Note that

$$\mathbb{E}_{\pi_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau] = (1 - \frac{\epsilon}{\delta})\mathbb{E}_{\pi_t^*}[|\Delta(\mathbf{d})| - \tau] + \frac{\epsilon}{\delta}\mathbb{E}_{\pi_{t,0}}[|\Delta(\mathbf{d})| - \tau]$$
$$\leq 0 + \frac{\epsilon}{\delta}(-\delta) \leq -\epsilon.$$

Therefore, $\pi_{t,\epsilon}$ is a feasible solution of the baseline problem $\mathbb{E}_{\pi_{t,\epsilon}}[\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]] : \mathbb{E}_{\pi_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau] + \epsilon \leq 0$. Thus,

$$\sum_{t=1}^{T} \mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \sum_{t=1}^{T} \mathbb{E}_{\pi_{t,\epsilon}^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$$
$$\leq \sum_{t=1}^{T} \mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \sum_{t=1}^{T} \mathbb{E}_{\pi_{t,\epsilon}}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]$$
$$\leq \sum_{t=1}^{T} \Big(\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - (1 - \frac{\epsilon}{\delta})\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \frac{\epsilon}{\delta}\mathbb{E}_{\pi_{t,0}}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big)$$
$$= \sum_{t=1}^{T} \frac{\epsilon}{\delta}\Big(\mathbb{E}_{\pi_t^*}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}_{\pi_{t,0}}\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]\Big) \leq \frac{\epsilon T}{\delta}.$$

Where the first inequality follows by that $\pi_{t,\epsilon}^*$ is the optimal solution while $\pi_{t,\epsilon}$ is a feasible solution to $\mathbb{E}_{\pi_{t,\epsilon}}[\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t]] : \mathbb{E}_{\pi_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau] + \epsilon \leq 0$; and the last inequality comes from $\mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \in [0,1]$. ■

G.3. Proof of Theorem 2

In this section, we establish upper bounds on both regret and constraint violation for the revised constraint condition (see Algorithm 2). This is achieved by introducing a slackness variable ϵ, which serves to tighten the constraint. First, we decompose $\mathcal{R}_+^\epsilon(T) + \phi\mathcal{V}^\epsilon(T)$ as follows, where $\mathcal{V}^\epsilon(T) = \sum_{t=1}^{T} |\Delta(\mathbf{d}_t)| - \tau + \epsilon$.

$\mathcal{R}_+^\epsilon(T) + \phi\mathcal{V}^\epsilon(T)$

$$
\begin{aligned}
&= \sum_{t=1}^{T} \Big[\mathbb{E}_{\pi^*_{t,\epsilon}} \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \phi[|\Delta(\mathbf{d}_t)| - \tau + \epsilon] \Big] \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{\pi^*_{t,\epsilon}} \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \phi[|\Delta(\mathbf{d}_t)| - \tau + \epsilon] - \phi_t \mathbb{E}_{\pi^*_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau + \epsilon] \\
&= \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\pi^*_{t,\epsilon}} \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t \mathbb{E}_{\pi^*_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau + \epsilon] - \mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \phi_t[|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon]}_{\mathcal{R}^{\epsilon}_1} \\
&\quad + \underbrace{\sum_{t=1}^{T} \Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \Big) + \phi \Big([|\Delta(\mathbf{d}_t)| - \tau + \epsilon] - [|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon] \Big) \Big]}_{\mathcal{R}^{\epsilon}_2} \\
&\quad + \underbrace{\sum_{t=1}^{T} \Big[\phi[|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon] - \phi_t[|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon] \Big]}_{\mathcal{R}^{\epsilon}_3}.
\end{aligned}
$$

Similar to the techniques in Section G.1, we can upper bound $\mathcal{R}^{\epsilon}_1, \mathcal{R}^{\epsilon}_2, \mathcal{R}^{\epsilon}_3$ as follows:

$$
\begin{aligned}
\mathcal{R}^{\epsilon}_1 &= \sum_{t=1}^{T} \mathbb{E}_{\pi^*_{t,\epsilon}} \mathbb{E}[Y|do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t \mathbb{E}_{\pi^*_{t,\epsilon}}[|\Delta(\mathbf{d})| - \tau + \epsilon] - \mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] + \phi_t[|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon] \\
&= \rho \cdot \mathcal{O}\Big(nS \log(T)\sqrt{T} + n^2 S^2 (\log(T))^2 + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} (\log(T))^2 \Big), \\
\mathcal{R}^{\epsilon}_2 &= \sum_{t=1}^{T} \Big[\Big(\mathbb{E}[\widehat{Y}|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \mathbb{E}[Y|do(\mathbf{d}_t), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] \Big) + \phi \Big([|\Delta(\mathbf{d}_t)| - \tau + \epsilon] - [|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon] \Big) \Big] \\
&= (1+\phi) \cdot \mathcal{O}\Big(nS \log(T)\sqrt{T} + n^2 S^2 (\log(T))^2 + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} (\log(T))^2 \Big), \\
\mathcal{R}^{\epsilon}_3 &= \sum_{t=1}^{T} \Big[\phi[|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon] - \phi_t[|\widehat{\Delta}(\mathbf{d}_t)| - \tau + \epsilon] \Big] \\
&= \mathcal{O}(\rho(1+\epsilon)^2 \sqrt{T}).
\end{aligned}
$$

Thus,

$$
\mathcal{R}_+(T) + \phi \mathcal{V}(T) = (\rho + \phi) \cdot \mathcal{O}\Big(nS \log(T)\sqrt{T} + n^2 S^2 (\log(T))^2 + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} (\log(T))^2 \Big) + \mathcal{O}(\rho(1+\epsilon)^2 \sqrt{T}).
$$

Regret $\mathcal{R}^{\epsilon}_+(T)$. By setting $\phi = 0$, we have:

$$
\begin{aligned}
\mathcal{R}_+(T) &= \mathcal{O}\Big(\rho \Big(nS \log(T)\sqrt{T} + n^2 S^2 (\log(T))^2 + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} (\log(T))^2 \Big) \Big) \\
&= \widetilde{\mathcal{O}}\Big(\rho nS\sqrt{T} + \rho n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} + \rho\sqrt{T}(1+\epsilon)^2 \Big),
\end{aligned}
$$

Constraint violations. By applying (Beck, 2017, Theorem 3.60), we have:

$$
\mathcal{V}^{\epsilon}(T) = \widetilde{\mathcal{O}}\Big(nS\sqrt{T} + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} + (1+\epsilon)^2 \sqrt{T} \Big).
$$

To obtain a bound $\mathcal{V}(T)$, we notice that:

$$
\begin{aligned}
\mathcal{V}(T) &= \mathcal{V}^{\epsilon}(T) - \sum_{t=1}^{T} \epsilon \\
&= \widetilde{\mathcal{O}}\Big(nS\sqrt{T} + n^2 S^2 \boldsymbol{\kappa}_{\mathcal{Z}} + (1+\epsilon)^2 \sqrt{T} - \epsilon T \Big).
\end{aligned}
$$

Which finishes the proof. ■

Algorithm 2 $\epsilon-$CCLB Algorithm

1: **Input:** Horizon T, truncated interval ρ, step size $\eta = \sqrt{T}/\rho$, and the initial dual value $\phi_1 = 0$, user-select parameter $\epsilon \in [0, \delta)$.
2: **for** $t = 1, 2, 3, \ldots, T$ **do**
3: Use MLE to estimate the reward parameter and build a confidence set $\mathcal{C}_t(\alpha)$ from Equation (10),

$$\mathcal{C}_t(\alpha) = \left\{\theta \in \Theta : \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\alpha)^2\right\}.$$

4: Greedy procedure. Choose the optimistic reward parameter and select the greedy action:

$$\tilde{\theta}_t = \text{argmax}_{\theta \in \mathcal{C}_t} \max_{\mathbf{d} \in \mathcal{D}_t} \mathbb{E}[Y | do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t],$$
$$d_t = \text{argmax}_{\mathbf{d} \in \mathcal{D}_t} \mathbb{E}[\widehat{Y} | do(\mathbf{d}), do(\mathbf{a}_t), \mathbf{w}_t, \mathbf{m}_t] - \phi_t(|\widehat{\Delta}(\mathbf{d})| - \tau + \epsilon).$$

5: Update the dual variable:

$$\phi_{t+1} = \text{Proj}_{[0,\rho]}\left[\phi_t + 1/\eta(|\widehat{\Delta}(\mathbf{d})| - \tau + \epsilon)\right].$$

6: Update the estimation and confidence set according to the new received reward y_{t+1}.
7: **end for**

G.4. Proof of Proposition 4

Proposition 4. *By conditions stated in Theorem 2, for the user-selected parameter* $\epsilon' = \Big(\sqrt{T} - \sqrt{T - 4\mathcal{C}_4\big(\sqrt{T} + \mathcal{C}_1 n \log(T) + (\mathcal{C}_2 + \mathcal{C}_3\boldsymbol{\kappa}_{\mathcal{Z}})n^2((\log(T))^2\sqrt{1/T}\big)}\Big)/2\mathcal{C}_4 - 1$, *where* $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ *are the universal constants independent of* $n, S, T, \boldsymbol{\kappa}_{\mathcal{Z}}$, *if* $n \geq 2$ *and* $\epsilon' < \delta$ *for sufficiently large* T, *then one could achieve a zero upper bound on the constraint violations when select* $\epsilon \in [\epsilon', \delta)$.

Proof. To show the result in cumulative zero constraint violations, we write it as the following where $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ are the universal constants which is independent of $n, S, T, \boldsymbol{\kappa}_{\mathcal{Z}}$, and $\epsilon \in [0, \delta)$:

$$\mathcal{V}(T) \leq \mathcal{C}_1 nS \log(T)\sqrt{T} + \mathcal{C}_2 n^2 S^2 (\log(T))^2 + \mathcal{C}_3 \boldsymbol{\kappa}_{\mathcal{Z}}\, n^2 S^2 (\log(T))^2 + \mathcal{C}_4(1+\epsilon)^2\sqrt{T} - \epsilon T,$$

we solve it when the right-hand-side is less than 0:

$$\frac{T - 2\mathcal{C}_4\sqrt{T} - \sqrt{(T - 2\mathcal{C}_4\sqrt{T})^2 - 4\mathcal{C}_4\sqrt{T}\Gamma}}{2\mathcal{C}_4\sqrt{T}} \leq \epsilon \leq \frac{T - 2\mathcal{C}_4\sqrt{T} + \sqrt{(T - 2\mathcal{C}_4\sqrt{T})^2 - 4\mathcal{C}_4\sqrt{T}\Gamma}}{2\mathcal{C}_4\sqrt{T}},$$

where $\Gamma = \mathcal{C}_1 nS \log(T)\sqrt{T} + \mathcal{C}_2 n^2 S^2((\log(T)))^2 + \mathcal{C}_3 \boldsymbol{\kappa}_{\mathcal{Z}}\, n^2 S^2 (\log(T))^2 + \mathcal{C}_4\sqrt{T}$. First, when T is large, the upper bound of ϵ have the following inequality:

$$\delta \leq 1 \leq \frac{\sqrt{T}}{2\mathcal{C}_4} - 1 \leq \frac{T - 2\mathcal{C}_4\sqrt{T} + \sqrt{(T - 2\mathcal{C}_4\sqrt{T})^2 - 4\mathcal{C}_4\sqrt{T}\Gamma}}{2\mathcal{C}_4\sqrt{T}}.$$

Since it is larger than the Slater's constant, therefore $\epsilon < \delta$. Now we look at the lower bound of ϵ,

$$\begin{aligned}
\epsilon' &= \frac{T - 2\mathcal{C}_4\sqrt{T} - \sqrt{(T - 2\mathcal{C}_4\sqrt{T})^2 - 4\mathcal{C}_4\sqrt{T}\Big(\mathcal{C}_1 nS \log(T)\sqrt{T} + \mathcal{C}_2 n^2 S^2((\log(T))^2 + \mathcal{C}_3\boldsymbol{\kappa}_{\mathcal{Z}}\, n^2 S^2 (\log(T))^2 + \mathcal{C}_4\sqrt{T}\Big)}}{2\mathcal{C}_4\sqrt{T}} \\
&= \frac{T - 2\mathcal{C}_4\sqrt{T} - \sqrt{T^2 - 4\mathcal{C}_4\Big(T\sqrt{T} + \mathcal{C}_1 nS \log(T)T + \mathcal{C}_2 n^2 S^2((\log(T))^2\sqrt{T} + \mathcal{C}_3\boldsymbol{\kappa}_{\mathcal{Z}}\, n^2 S^2 (\log(T))^2\sqrt{T}\Big)}}{2\mathcal{C}_4\sqrt{T}}. \\
&= \frac{\sqrt{T} - \sqrt{T - 4\mathcal{C}_4\Big(\sqrt{T} + \mathcal{C}_1 nS \log(T) + (\mathcal{C}_2 + \mathcal{C}_3\boldsymbol{\kappa}_{\mathcal{Z}})n^2 S^2((\log(T))^2\sqrt{1/T}\Big)}}{2\mathcal{C}_4} - 1.
\end{aligned}$$

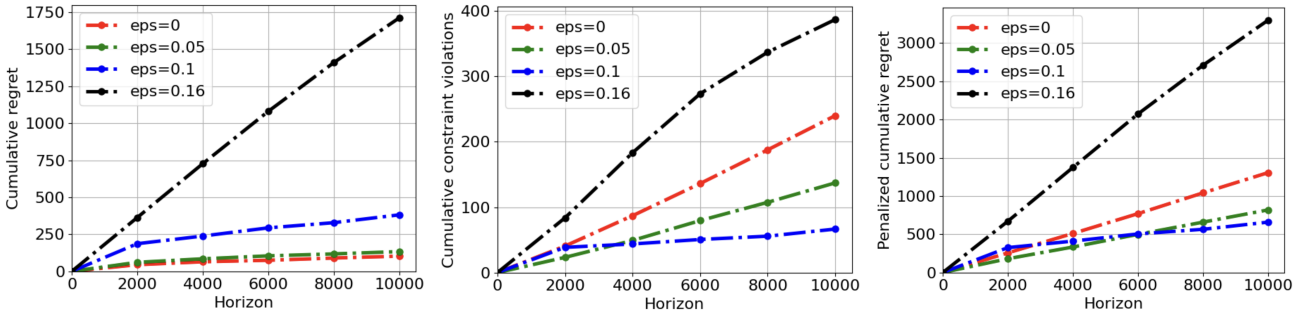
If the lower bound ϵ' is less than the Slater's constant, then when the learner choose $\epsilon \in [\epsilon', \delta)$, we could achieve zero cumulative constraint violations. ■

H. Additional Experiments

In this section, we provide additional evaluations for the ϵ-CCLB and CCLB algorithms when selecting different tightness parameter ϵ (Figure 4) and counterfactual fairness threshold τ (Figure 5), respectively.

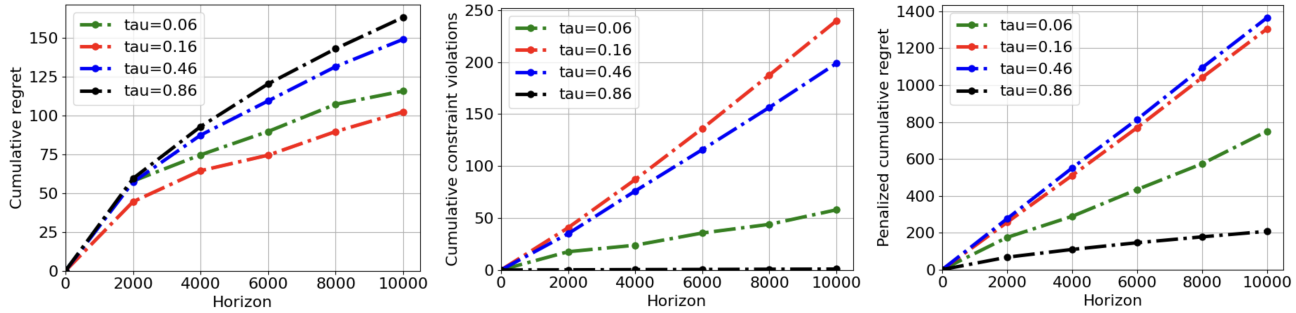
Figure 4. Increasing the user-chosen tightness parameter ϵ, the cumulative regret increases as well, but the cumulative constraint violations decreases (the learner becomes more conservative). When we pick the ϵ equals τ (both are 0.16), we observe that it incurs a high cumulative constraint violations (Figure 4 (b)) since $\mathbb{E}_{\pi_t}[|\Delta(\mathbf{d}, \mathbf{X}_t)| - \tau + \epsilon] > 0$ therefore there does not exist feasible decisions (notice that $\epsilon < \delta \leq \tau$).

Figure 5. As the counterfactual fairness threshold τ increases, the feasible region is larger (more of the actions are feasible), which means the fixed comparator could be better but at the same time easier to avoid violating constraints (since more actions are feasible in the first place when $|\mathcal{D}|$ is fixed), thus reduce the cumulative constraint violations (see Figure 5 (b), the cumulative constraint violations are nearly 0 when $\tau = 0.86$). On the other hand, when τ is small, i.e., $|\Delta(\mathbf{d}, \mathbf{X}_t)| \geq \tau$ almost every round. The dual variable ϕ_t will increase as well, which renders the learner penalizes more on the counterfactual fairness constraint (thus more conservative), therefore decrease the cumulative constraint violations (see Figure 5 (b), the cumulative constraint violations are relative small when $\tau = 0.06$).



(a) (b) (c)

Figure 4: Plots for the ϵ-CCLB (τ=0.16) algorithm when selecting different ϵ on (a) cumulative regret; (b) cumulative constraint violations; (c) penalized cumulative regret.



(a) (b) (c)

Figure 5: Plots for CCLB algorithm when selecting different counterfactual fairness threshold τ on (a) Cumulative regret; (b) Cumulative constraint violations; (c) Penalized cumulative regret.