Functional Linear Regression of Cumulative Distribution Functions

Anonymous authors
Paper under double-blind review

Abstract

The estimation of cumulative distribution functions (CDF) is an important learning task with a great variety of downstream applications, such as risk assessments in predictions and decision making. In this paper, we study functional regression of contextual CDFs where each data point is sampled from a linear combination of context dependent CDF basis functions. We propose functional ridge-regression-based estimation methods that estimate CDFs accurately everywhere. In particular, given n samples with d basis functions, we show estimation error upper bounds of $\widetilde{O}(\sqrt{d/n})$ for fixed design, random design, and adversarial context cases. We also derive matching information theoretic lower bounds, establishing minimax optimality for CDF functional regression. Furthermore, we remove the burn-in time in the random design setting using an alternative penalized estimator. Then, we consider agnostic settings where there is a mismatch in the data generation process. We characterize the error of the proposed estimators in terms of the mismatched error, and show that the estimators are well-behaved under model mismatch. Moreover, to complete our study, we formalize infinite dimensional models where the parameter space is an infinite dimensional Hilbert space, and establish a self-normalized estimation error upper bound for this setting. Notably, the upper bound reduces to the $\widetilde{O}(\sqrt{d/n})$ bound when the parameter space is constrained to be d-dimensional. Our comprehensive numerical experiments validate the efficacy of our estimation methods in both synthetic and practical settings.

1 Introduction

Estimating cumulative distribution functions (CDF) of random variables is a salient theoretical problem that underlies the study of many real-world phenomena. For example, Huang et al. (2021) and Leqi et al. (2022) recently showed that estimating CDFs is sufficient for risk assessment, thereby making CDF estimation a key building block for such decision-making problems. In a similar vein, it is known that CDFs can also be used to directly compute distorted risk functions (Wirch & Hardy, 2001), coherent risks (Artzner et al., 1999), conditional value-at-risk and mean-variance (Cassel et al., 2023), and cumulative prospect theory risks (Prashanth et al., 2016). Furthermore, CDFs are also useful in calculating various risk functionals appearing in insurance premium design, portfolio design, behavioral economics, behavioral finance, and healthcare applications (Rockafellar et al., 2000; Shapiro et al., 2014; Prashanth et al., 2016; Wong et al., 2022). Given the broad utility of estimating CDFs, there is a vast (and fairly classical) literature that tries to understand this problem.

In particular, the renowned *Glivenko-Cantelli theorem* (Cantelli, 1933; Glivenko, 1933) states that given independent samples of a random variable, one can construct a consistent estimator for its CDF. Tight non-asymptotic sample complexity rates for such estimation using the Kolmogorov-Smirnov (KS) distance as the loss have also been established in the literature (Cantelli, 1933; Glivenko, 1933; Dvoretzky et al., 1956; Massart, 1990). However, these results are all limited to the setting of a single random variable. In contrast, many modern learning problems, such as doubly-robust estimators in contextual bandits, treatment effects, and Markov decision processes (Huang et al., 2021; Kallus et al., 2019; Huang et al., 2022), require us to simultaneously learn the CDFs of potentially infinitely many random variables from limited data. Hence, the classical results on CDF estimation do not address the needs of such emerging learning applications.

Contributions. In this work, as a first step towards developing general CDF estimation methods that fulfill the needs of the aforementioned learning problems, we study *functional linear regression of CDFs*, where samples are generated from CDFs that are convex combinations of context-dependent CDF bases. Our model resembles the well-studied linear regression and stochastic linear bandits problem. In linear regression, researchers analyzed finite-dimensional parametric models with pre-selected feature functions. These pre-designed features result from extensive feature engineering processes carried out for the underlying task. Similarly, within the domain of contextual bandits, researchers studied the stochastic linear bandit problem using a linear model (Lattimore & Szepesvári, 2020, Equation (19.1)) with finite dimension and known feature map. Thus, it is natural to commence the analysis assuming the access to known "feature" CDFs, which ultimately bestows the advantages intrinsic to linear regression. As our main contribution, we define both least-squares regression and ridge regression estimators for the unknown linear weight parameter, and establish corresponding estimation error bounds for the fixed design, random design, adversarial, and self-normalized settings. In particular, given n samples with d CDF bases, we prove estimation error upper bounds that scale like $\widetilde{O}(\sqrt{d/n})$ (neglecting sub-dominant factors). Our results achieve the same problem-dependent scaling as in canonical finite dimensional linear regression (Abbasi-Yadkori et al., 2011b;a; Hsu et al., 2012b). Moreover, we derive $\Omega(\sqrt{d/n})$ information theoretic lower bounds for functional linear regression of CDFs. This establishes minimax estimation rates of $\widetilde{\Theta}(\sqrt{d/n})$ for the CDF functional regression problem. We later show that this result directly implies the concentration of CDFs in KS distance. We also propose a new penalized estimator that theoretically eliminates the requirement on the burn-in time of sample size in the random design setting. Then, we consider agnostic settings where there is a mismatch between our linear model and the actual data generation process. We characterize the estimation error of the proposed estimator in terms of the mismatch error, and demonstrate that the estimator is well-behaved under model mismatch. To complete our study, we generalize the parameter space in the linear model from finite-dimensional Euclidean spaces to general infinite-dimensional Hilbert spaces, extend the ridge regression estimator to the infinite-dimensional model with proper regularization, and establish a corresponding self-normalized estimation error upper bound which immediately recovers our previous $\widetilde{O}(\sqrt{d/n})$ upper bound when the parameter space is restricted to be d-dimensional. Finally, we present numerical results for synthetic and real data experiments to illustrate the performance of our estimation methods.

Related works. A complementary approach to the proposed CDF regression framework is quantile regression (Koenker & Bassett Jr, 1978). Although quantile regression may appear to be closely related to CDF regression at first glance, the two problems have very different flavors. Indeed, unlike CDFs, quantiles are not sufficient for law invariant risk assessment. Besides, due to their infinite range, quantile estimation is quite challenging, resulting in analyses that only consider pointwise estimation (Takeuchi et al., 2006). However, as it is necessary to estimate multiple quantiles for CDF estimation, a simultaneous analysis of multiple quantile estimates is needed theoretically, which typically requires a union bound on the failure probability that increases linearly with the number of estimates. Furthermore, the estimated multiple quantiles may not be monotonically increasing with respect to the probability values, requiring extra effort to construct a valid CDF from a finite series of quantile estimates. Moreover, any such construction will incur a non-convergent KS distance between the estimated CDF and the true CDF for some distribution, as a general CDF can exhibit jumps or flat regions at any position. Additionally, the quality of the estimated CDF from multiple estimated quantiles relies heavily on the selection of grid points of probability values, which is instance-dependent and may require knowledge of the distribution the learner seeks to estimate. Thus, establishing a universal rule for choosing grid points that yield reasonable CDF estimates via quantile regression across diverse distributions proves challenging. In practice, the introduction of grid points introduces numerous hyperparameters to tune, adding artificial complexity to the methodology. Perhaps more importantly, quantile regression can be ill-posed in many machine learning settings. For example, quantiles are not estimatable in decision-making problems and games with mixed random variables (which take both discrete and continuous values). For these reasons, our focus in this paper will be on CDF regression.

Several works have delved into the realm of conditional CDF estimation. Hall et al. (1999) estimated conditional CDFs for fixed cutoff y and context x using local logistic methods and adjusted Nadaraya-Watson estimators. However, their analysis necessitates the assumption of strong regularity conditions on the conditional CDF (including at least continuous second-order derivatives), the marginal CDF of the context,

and the data generating process. They established asymptotic convergence only for fixed cutoff and context. Ferraty et al. (2006) introduced a kernel-type nonparametric estimator for conditional CDFs at a fixed context x. Their analysis mandates that the samples are independent and identically distributed (iid), in addition to some regularity assumptions concerning the marginal distribution of x and the smoothness of the conditional CDF. Their theoretical findings, too, revolve around asymptotic scenarios and apply solely to fixed contexts. Chung & Dunson (2009) proposed a special class of conditional CDFs based on probit stick-breaking process mixture models. They developed an MCMC algorithm for posterior sampling of parameters but did not furnish theoretical assurances regarding consistency. Distinguishing itself from existing endeavors, this paper introduces a novel linear model (1) or (18) where we presume knowledge of an arbitrary family of contextual CDFs and aim to estimate the weight parameter θ_*. Consequently, our model possesses the capability to encompass any conditional CDF, enabling the estimation of the conditional CDF across all values of the context x and cutoff y by estimating one parameter. Furthermore, we embrace an adversarial data generation process (see Scheme I in Section 2), which surpasses the limitations of the iid setting in terms of generality. We provide tight non-asymptotic analysis of the estimation error by showing matching upper bounds and lower bounds of the error. Additionally, our model (1) or (18) readily accommodates the integration of estimated CDFs from previous works on conditional CDF estimation into the family of feature contextual CDFs, thereby enhancing the overall quality of the conditional CDF estimates.

Chernozhukov et al. (2013) and Koenker et al. (2013) study "distribution regression" where for a fixed cutoff y, they estimate parameters in conditional CDF models by maximizing log likelihood of $\mathbb{1}\{y \geqslant Y_i\}$ for outcome samples $Y_1, \ldots, Y_n$. Thus, both works require specific models for conditional CDFs. Chernozhukov et al. (2013) introduced a "distribution regression" model where the conditional CDF takes the form of a link function evaluated at the inner product of vector transformations of the context X and outcome Y. However, due to the dependence of the log likelihood on the cutoff y within this model, their estimator is inherently pointwise. They established asymptotic convergence of the estimated conditional CDF. Nonetheless, this hinges on certain assumptions concerning the true parameter functions, which is challenging to validate. Koenker et al. (2013) considered the "linear local-scale model" where the outcome is the summation of a linear local function of the context and the product of a linear scale function of the context and an independent random error boasting a smooth density. Their convergence results are of an asymptotic nature, assuming iid samples, alongside other conditions on the expected log likelihood and the asymptotic covariance function which also pose substantial verification challenges. Furthermore, the maximum likelihood estimation (MLE) used in both papers only accesses the indicators denoting whether the samples $Y_1, \ldots, Y_n$ surpass a fixed cutoff y, which underutilizes the wealth of information inherent in the samples. In stark contrast, our estimator (2) or (21) uses the one-sample empirical CDFs ($\mathbb{1}\{Y_i \leqslant \cdot\}$) which fully exploit the sample information. Moreover, as previously mentioned, the estimated CDFs derived in the above distribution regression problems can be seamlessly integrated into our proposed model.

In some literature, "distribution regression" takes on a distinctive meaning, referring to the model where the context is a sequence of samples from some distribution which, together with the outcome, is sampled from some meta joint distribution (Póczos et al., 2013; Szabó et al., 2016). The task is to learn a mapping from the distribution of the context to the outcome. Contrastingly, our model (1) or (18) operates in a different realm: the outcome is a sample from a mixture of contextual CDFs and the task is to learn the weight parameter θ_*. Thus, our model diverges from the above notion of distribution regression. Our focus is not on estimating a mapping from distributions to outcomes but on estimating a parameter that governs the condition distribution. Moreover, there is no meta distribution that the samples follow in our adversarial data generating process.

Outline. We briefly outline the rest of the paper. Notation and formal setup for our problem are given in Section 2. We propose our estimation paradigm and analyze its theoretical performance in Section 3. We derive corresponding lower bounds on the estimation error in Section 4. We establish upper bounds on the estimation error under the existence of a mismatch in our proposed model in Section 5. We generalize the problem from estimating finite dimensional parameters to estimating infinite dimensional parameters, extend our estimation paradigm to this infinite dimensional setting, and prove an upper bound on estimation error in Section 6. Numerical results are displayed in Section 7. Conclusions are drawn and future research directions are suggested in Section 8. All the proofs and additional results are presented in the appendices.

2 Preliminaries

In this section, we introduce the notation used in the paper and set up the learning problem of contextual CDF regression.

Notation. Let $\mathbb{N}$ denote the set of positive integers. For any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, \dots, n\}$. For any measure space $(\Omega, \mathcal{F}, \mathfrak{m})$, define the Hilbert space $\mathcal{L}^2(\Omega, \mathfrak{m}) := \{f : \Omega \to \mathbb{R} \mid \int_\Omega |f|^2 d\mathfrak{m} < \infty\}$ with $\mathcal{L}^2$-norm $\|f\|_{\mathcal{L}^2(\Omega,\mathfrak{m})} := \sqrt{\int_\Omega |f|^2 d\mathfrak{m}}$ for $f \in \mathcal{L}^2(\Omega, \mathfrak{m})$. For any positive definite matrix $A \in \mathbb{R}^{d\times d}$, define $\|\cdot\|_A$ to be the weighted ℓ^2-norm in $\mathbb{R}^d$ induced by A, i.e., $\|x\|_A = \sqrt{x^\top A x}$ for $x \in \mathbb{R}^d$. For the standard Euclidean (or ℓ^2-) norm $\|\cdot\|_{I_d}$, where I_d denotes the $d \times d$ identity matrix, we omit the subscript I_d and simply write $\|\cdot\|$. For any square matrix A, let $\mu_{\min}(A)$ denote the smallest eigenvalue of A, $\mu_{\max}(A)$ denote the largest eigenvalue of A and $\|A\|_2$ denote the spectral norm of the matrix A, i.e., $\|A\|_2 := \sqrt{\mu_{\max}(A^\top A)}$. Let $\mathrm{KS}(F_1, F_2) := \sup_{x\in\mathbb{R}} |F_1(x) - F_2(x)|$ denote the KS distance between two CDFs F_1 and F_2. Finally, let $\mathbb{1}\{\cdot\}$ denote the indicator function. More technical notation dealing with measurability issues is provided at the beginning of Appendix B.

Problem setup. In this paper, we consider the problem of functional linear regression of CDFs. To define this problem, let $\mathcal{X}$ denote the context space, and let $F(x, \cdot) : \mathbb{R} \to [0, 1]$ be the CDF of some $\mathbb{R}$-valued random variable for any $x \in \mathcal{X}$. We assume that $\mathcal{X}$ is a Polish space throughout the paper. For a context $x \in \mathcal{X}$, we observe a sample y from its corresponding CDF $F(x, \cdot)$. We next summarize two schemes to generate (x, y) samples:

- **Scheme I (Adversarial).** For each $j \in \mathbb{N}$, an adversary picks $x^{(j)} \in \mathcal{X}$ (either deterministically or randomly) in an adaptive way given knowledge of the previous $y^{(i)}$'s for $i < j$, and then $y^{(j)} \in \mathbb{R}$ is sampled from $F(x^{(j)}, \cdot)$. This includes the canonical **fixed design** setting as a special case, where all $x^{(j)}$'s are fixed a priori without knowledge of $y^{(j)}$'s.
- **Scheme II (Random).** For each $j \in \mathbb{N}$, $x^{(j)} \in \mathcal{X}$ is sampled from some probability distribution $P_X^{(j)}$ on $\mathcal{X}$ independently, and then $y^{(j)} \in \mathbb{R}$ is sampled from $F(x^{(j)}, \cdot)$ independently. This is known as the **random design** setting in the regression context.

Scheme I and Scheme II generalize the assumptions of the data generation process in canonical ridge regression in Abbasi-Yadkori et al. (2011a) and Hsu et al. (2012b) to the problem of CDF estimation, respectively. Note that although the random design setting in Scheme II is a special case of Scheme I, we emphasize it because it has specific properties that deserve a separate treatment. The adversarial setting in Scheme I is more general than what is typically considered for regression, and our corresponding self-normalized analysis has several potential future applications in risk assessment for reinforcement learning, e.g., in contextual bandits (Abbasi-Yadkori et al., 2011a).

The task of contextual CDF regression is to recover F from a sample $\{(x^{(j)}, y^{(j)})\}_{j\in[n]}$ of size n. As an initial step towards this problem, inspired by the well-studied linear regression and linear contextual bandits problems (Lattimore & Szepesvári, 2020, Equation (19.1)), where finite-dimensional parametric models with pre-selected feature functions are assumed, we consider a *linear model* for F. Let d be a fixed positive integer. For each $i \in [d]$ and $x \in \mathcal{X}$, let $\phi_i(x, \cdot) : \mathbb{R} \to [0, 1]$ be a feature function that is a CDF of a $\mathbb{R}$-valued random variable with range contained in some Borel set $S \subseteq \mathbb{R}$, and assume that ϕ_i is measurable. Then, we define the vector-valued function $\Phi : \mathcal{X} \times \mathbb{R} \to [0, 1]^d$, $\Phi(x, t) = [\phi_1(x, t), \dots, \phi_d(x, t)]^\top$. We assume that there exists some *unknown* $\theta_* \in \Delta^{d-1}$, where $\Delta^{d-1} := \{(\theta_1, \dots, \theta_d) \in \mathbb{R}^d : \sum_{i=1}^d \theta_i = 1, \theta_i \geqslant 0 \text{ for } 1 \leqslant i \leqslant d\}$ denotes the probability simplex in $\mathbb{R}^d$, such that,

$$F(x, t) = \theta_*^\top \Phi(x, t), \quad \forall\ x \in \mathcal{X},\ t \in \mathbb{R}. \tag{1}$$

Thus, we can view Φ as a "basis" for contextual CDF learning.

We visualize the sample generation process in Figure 1 where the contextual CDFs are shown in the left column and the one-sample empirical CDFs ($\mathbb{1}\{y \leqslant \cdot\}$ for sample y) are shown in the right column. It is

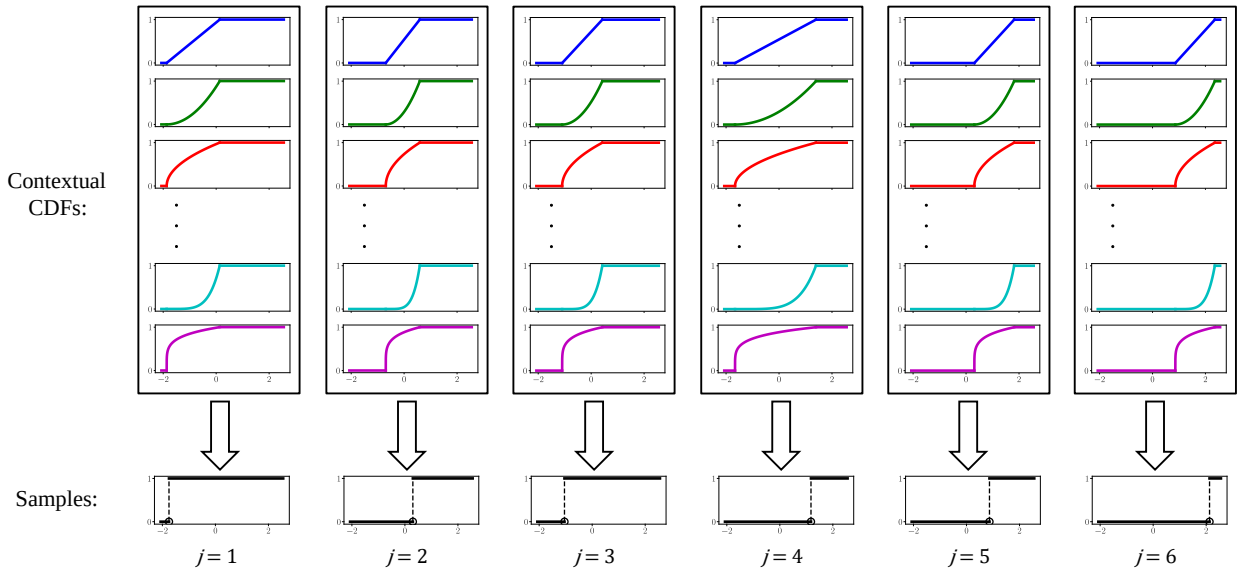


Figure 1: A visualization of the data generating process. For each $j \in [6]$ with context $x^{(j)} \in \mathcal{X}$, the upper row shows the d contextual CDFs ($\phi_i(x^{(j)}, \cdot)$, $i \in [d]$) under the context $x^{(j)}$. For $y^{(j)}$ drawn from the CDF $F(x^{(j)}, \cdot) = \theta_*^\top \Phi(x^{(j)}, \cdot)$ where $\Phi(x^{(j)}, \cdot) := [\phi_i(x^{(j)}, \cdot), \ldots, \phi_d(x^{(j)}, \cdot)]^\top$, the bottom row shows the sample empirical CDF $\mathrm{I}_{y^{(j)}}(\cdot) := \mathbb{1}\{y^{(j)} \leqslant \cdot\}$.

worth mentioning the differences between our model and the mixture model with known basis distributions in the statistics literature. First, the basis distributions in our model depend on the context of the sample and are not fixed. Second, in mixture models, the samples are assumed to be independent while in our Scheme I, the samples can be dependent since $x^{(j)}$ is picked adversarially given knowledge of the previous $y^{(i)}$'s. Thus, the mixture model with known basis distributions only corresponds to the fixed design setting with the same context $x^{(j)} = x$ for all samples.

As explained in the sampling schemes above, given $x^{(j)}$ at the jth sample, the observation $y^{(j)}$ is generated according to the CDF $F(x^{(j)}, \cdot) = \theta_*^\top \Phi(x^{(j)}, \cdot)$. For notational convenience, we will often refer to the vector-valued function $\Phi(x^{(j)}, \cdot)$ as $\Phi_j(\cdot)$ for all $j \in [n]$, so that $F(x^{(j)}, \cdot) = \theta_*^\top \Phi_j(\cdot)$. Under the linear model in (1), our *goal is to estimate the unknown parameter* θ_* from the sample $\{(x^{(j)}, y^{(j)})\}_{j \in [n]}$ in a (regularized) least-squares error sense. This in turn recovers the contextual CDF function F.

3 Upper bounds on estimation error

In this section, we propose an estimation paradigm for the unknown parameter θ_* in Section 3.1, derive the upper bounds on the associated estimation error in Section 3.2, and propose a new penalized estimator that theoretically eliminates the burn-in time of the sample size in the random setting in Section 3.3.

3.1 Ridge regression estimator

We begin by formally stating our least-squares functional regression optimization problem to learn θ_*. Given a probability measure $\mathfrak{m}$ on S, the sample $\{(x^{(j)}, y^{(j)})\}_{j \in [n]}$, and the set of basis functions $\{\Phi_j\}_{j \in [n]}$, we propose to estimate θ_* by minimizing the (ridge or) ℓ^2-regularized squared $\mathcal{L}^2(S, \mathfrak{m})$-distance between the estimated and empirical CDFs:

$$\widehat{\theta}_\lambda := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{j=1}^{n} \|\mathrm{I}_{y^{(j)}} - \theta^\top \Phi_j\|^2_{\mathcal{L}^2(S,\mathfrak{m})} + \lambda \|\theta\|^2, \tag{2}$$

where $\lambda \geqslant 0$ is the hyper-parameter that determines the level of regularization, and the function observation $\mathrm{I}_{y^{(j)}}(t) := \mathbb{1}\{y^{(j)} \leqslant t\}$ is an empirical CDF of $y^{(j)}$ that forms an unbiased estimator for $F(x^{(j)}, \cdot)$ conditioned

on past contexts and observations. Hence, in Scheme I, we only require that $\mathrm{I}_{y^{(j)}} - \theta^\top \Phi_j$ is a zero-mean function given past contexts and observations, making our analysis suitable for online learning problems where the later contexts can depend on the past contexts and observations. We remark that the adoption of $\mathcal{L}^2$-distance in (2) is natural. Indeed, researchers have considered the $\mathcal{L}^2$-distance between a one-sample empirical CDF and a CDF estimate in the definition of Continuous Ranked Probability Score (CRPS) (Hersbach, 2000) to assess the performance of the CDF estimate in approximating data distributions. In fact, viewing the one-sample empirical CDF as the response and the basis contextual CDFs as the feature in linear regression, it is natural to consider the least squares method, precisely corresponding to minimizing the $\mathcal{L}^2$-distance in our functional setting. Notice further that $\widehat{\theta}_\lambda$ in (2) is an *improper estimator* since it may not lie in Δ^{d-1}. However, since Δ^{d-1} is compact in $\mathbb{R}^d$, $\widetilde{\theta}_\lambda := \arg\min_{\vartheta \in \Delta^{d-1}} \|\vartheta - \widehat{\theta}_\lambda\|_A$ exists for any positive definite $A \in \mathbb{R}^{d\times d}$. Moreover, since Δ^{d-1} is also convex, we have $\|\widetilde{\theta}_\lambda - \theta\|_A \leqslant \|\widehat{\theta}_\lambda - \theta\|_A$ (Beck, 2014, Theorem 9.9) for any $\theta \in \Delta^{d-1}$ including θ_*. This means that an upper bound on $\|\widehat{\theta}_\lambda - \theta_*\|_A$ is also an upper bound on $\|\widetilde{\theta}_\lambda - \theta_*\|_A$. Additionally, as we will see later, the $\widehat{\theta}_\lambda$ has a closed-form analytic solution which benefits the analysis of the estimation error. Therefore, we focus our analysis on the improper estimator $\widehat{\theta}_\lambda$, noting that its projection onto Δ^{d-1} yields an estimator $\widetilde{\theta}_\lambda$ for which the same upper bounds hold.

When $\lambda > 0$, the objective function in (2) is a (2λ)-strongly convex function of $\theta \in \mathbb{R}^d$ (see, e.g., Bertsekas et al., 2003, for the definition), and is uniquely minimized at

$$\widehat{\theta}_\lambda = \left(\sum_{j=1}^n \int_S \Phi_j \Phi_j^\top d\mathfrak{m} + \lambda I_d\right)^{-1} \left(\sum_{j=1}^n \int_S \mathrm{I}_{y^{(j)}} \Phi_j d\mathfrak{m}\right). \tag{3}$$

For the unregularized case where $\lambda = 0$, we omit the subscript λ and write $\widehat{\theta}$ to denote a corresponding estimator in (2). Note that when $\lambda = 0$, if $\mu_{\min}(\sum_{j=1}^n \int_S \Phi_j \Phi_j^\top d\mathfrak{m}) > 0$, the objective function in (2) is still strongly convex, and is uniquely minimized at $\widehat{\theta}$ given in (3) with $\lambda = 0$. In practice, one can deploy standard numerical methods to compute the integral in (3), and the computational complexity of the matrix inversion is cubic in the dimension d. However, iterative methods can be used to obtain better dimension dependence in the running time. As a remark, since the probability density functions (PDFs) of the basis distributions may not exist, the samples in Scheme I can be dependent, and the distributions of the contexts in Scheme II are unknown, the likelihood function of the samples generally does not exist in our problem setting, which rules out the usage of MLE. But our estimator (2) always exists. Moreover, we focus on non-asymptotic analysis of our estimator and prove self-normalized upper bounds for the estimation error, which is rarely analyzed for MLEs.

Lastly, it is worth remarking upon the choice of measure $\mathfrak{m}$ used above. In order for the estimator in (2) to be well-defined, since $\mathrm{I}_y(t), \theta^\top \Phi(x,t) \in [0,1]$ for any $t, y \in \mathbb{R}$ and $x \in \mathcal{X}$, it suffices to ensure that $\mathfrak{m}(S) < \infty$ (i.e., $\mathfrak{m}$ is a finite measure). This is the reason why we restrict $\mathfrak{m}$ to be a probability measure on S. Furthermore, the probability measure $\mathfrak{m}$ can in general be chosen to adapt to specific problem settings. For example, the uniform measure $\mathfrak{m}_{\mathsf{U}}$ on S is often easy to compute for some choices of S. Specifically, if $0 < \mathsf{Leb}(S) < \infty$, where Leb denotes the Lebesgue measure, $\mathfrak{m}_{\mathsf{U}}$ is defined by $\frac{d\mathfrak{m}_{\mathsf{U}}}{d\mathsf{Leb}} = \frac{1}{\mathsf{Leb}(S)}$, where $\frac{d\mathfrak{m}_{\mathsf{U}}}{d\mathsf{Leb}}$ is the Radon-Nikodym derivative. If S is a finite set with cardinality $\#S$, $\mathfrak{m}_{\mathsf{U}} = \frac{1}{\#S}\sum_{s\in S} \delta_s$, where δ_s denotes the Dirac measure at s. On the other hand, when $S = \mathbb{R}$, $\mathfrak{m}$ can be set to the Gaussian measure γ_{c,σ^2} defined by $\gamma_{c,\sigma^2}(dx) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-c)^2/(2\sigma^2)} dx$ with $c \in \mathbb{R}$ and $\sigma^2 > 0$.

3.2 Self-normalized bounds in various settings

For samples generated according to Scheme I, we prove self-normalized upper bounds on the error $\widehat{\theta}_\lambda - \theta_*$. For any probability measure $\mathfrak{m}$ on S, define $U_n := \sum_{j=1}^n \int_S \Phi_j \Phi_j^\top d\mathfrak{m}$ and $U_n(\lambda) = U_n + \lambda I_d$ for $n \in \mathbb{N}$ and $\lambda \geqslant 0$. For $n, d \in \mathbb{N}$, $\lambda, \tau \in (0,\infty)$, and $\delta \in (0,1)$, define

$$\varepsilon_\lambda(n,d,\delta) := \sqrt{d\log\left(1 + n/\lambda\right) + 2\log(1/\delta)} + \sqrt{\lambda}\|\theta_*\| \quad \text{and} \tag{4}$$

$$\varepsilon(n,d,\delta,\tau) := \left(\sqrt{d} + \sqrt{8d\log(1/\delta)} + \frac{4}{3}\sqrt{d/n}\log(1/\delta)\right)/\sqrt{\tau}. \tag{5}$$

The next theorem states our self-normalized upper bound on the estimation error.

Theorem 1 (Self-normalized bound in adversarial setting)**.** *Assume $\mathfrak{m}$ is a probability measure on S and $\{(x^{(j)}, y^{(j)})\}_{j\in\mathbb{N}}$ is sampled according to Scheme I with F defined in (1). For any $\lambda > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, for all $n \in \mathbb{N}$, the estimator defined in (2) satisfies*

$$\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)} \leqslant \varepsilon_\lambda(n, d, \delta). \tag{6}$$

Moreover, for the unregularized case, we have the following result.

Proposition 2 (Self-normalized bound in adversarial setting for unregularized estimator)**.** *Under the same assumptions as Theorem 1, if U_N is positive definite for a fixed $N \in \mathbb{N}$, then for any $\delta \in (0,1)$ and $n \geqslant N$, with probability at least $1-\delta$, the estimator defined in (2) with $\lambda = 0$ satisfies*

$$\|\widehat{\theta} - \theta_*\|_{U_n} \leqslant \varepsilon\left(n, d, \delta, \mu_{\min}(U_n)/n\right). \tag{7}$$

The proofs of Theorem 1 and Proposition 2 are provided in Appendix B.1. Informally, Theorem 1 and Proposition 2 convey that with high probability, the self-normalized errors $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ and $\|\widehat{\theta} - \theta_*\|_{U_n}$ scale as $\widetilde{O}(\sqrt{d})$ in the ℓ^2-regularized and unregularized cases, where $\widetilde{O}(\cdot)$ ignores logarithmic and other sub-dominant factors. We note that Theorem 1 and Proposition 2 also imply upper bounds on the (un-normalized) error $\|\widehat{\theta}_\lambda - \theta_*\|$. Indeed, for any positive definite matrix $A \in \mathbb{R}^{d\times d}$ and vector $a \in \mathbb{R}^d$, we have $\|a\| \leqslant \mu_{\min}(A)^{-1/2}\|a\|_A$. Thus, for example, (6) in Theorem 1 implies that $\|\widehat{\theta}_\lambda - \theta_*\| \leqslant \mu_{\min}(U_n(\lambda))^{-1/2}\varepsilon_\lambda(n,d,\delta) = \widetilde{O}\big(\sqrt{d/(1+\mu_{\min}(U_n))}\big)$ with high probability. Then, for the projected estimator $\widetilde{\theta}_\lambda \in \Delta^{d-1}$, we have $\|\widetilde{\theta}_\lambda - \theta_*\| \leqslant \widetilde{O}\big(\min\{1, \sqrt{d/(1+\mu_{\min}(U_n))}\}\big)$ by the property of Δ^{d-1}. When $\mu_{\min}(U_n) = \Theta(n)$, we have $\|\widehat{\theta}_\lambda - \theta_*\| = \widetilde{O}\big(\sqrt{d/n}\big)$.

The key idea in the proof of Theorem 1 is to first notice that $\widehat{\theta}_\lambda - \theta_* = U_n(\lambda)^{-1}W_n - U_n(\lambda)^{-1}(\lambda\theta_*)$, where $W_n := \sum_{j=1}^n \int_S (\mathrm{I}_{y^{(j)}}\Phi_j - \theta_*^\top\Phi_j\Phi_j)d\mathfrak{m}$. We next show that $\{\overline{M}_n\}_{n\geqslant 0}$ where $\overline{M}_n := \frac{\lambda^{d/2}}{\det(U_n(\lambda))^{1/2}}\exp\left(\frac{1}{2}\|W_n\|^2_{U_n(\lambda)^{-1}}\right)$ is a super-martingale. *Doob's maximal inequality* for super-martingales is then used in conjunction with some careful algebra to establish (6). To prove Proposition 2, we use a vector *Bernstein inequality* for bounded martingale difference sequences (Hsu et al., 2012a, Proposition 1.2) to show a high probability upper bound for $\|W_n\|$. Note that U_N being positive definite implies that U_n is positive definite for $n \geqslant N$. Since $\|\widehat{\theta} - \theta_*\|_{U_n} = \|W_n\|_{U_n^{-1}} \leqslant \|W_n\|/\sqrt{\mu_{\min}(U_n)}$, we establish (7).

Since the fixed design is a special case of the adversarial setting, Theorem 1 and Proposition 2 imply the same $\widetilde{O}(\sqrt{d})$-style upper bounds as a corollary in the fixed design setting.

Corollary 3 (Self-normalized bound in fixed design setting)**.** *For an arbitrary probability measure $\mathfrak{m}$ on S and an arbitrary sequence $\{x^{(j)}\}_{j\in\mathbb{N}} \in \mathcal{X}^{\mathbb{N}}$, assume that $y^{(j)}$ is sampled from $F(x^{(j)}, \cdot)$ independently for each $j \in \mathbb{N}$ with F defined in (1). For any $\lambda > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, the estimator defined in (2) satisfies (6) for all $n \in \mathbb{N}$.*

If U_N is positive definite for some fixed $N \in \mathbb{N}$, then for any $\delta \in (0,1)$ and $n \geqslant N$, with probability at least $1-\delta$, the estimator defined in (2) with $\lambda = 0$ satisfies (7).

The proof of Corollary 3 is inline with those of Theorem 1 and Proposition 2.

Furthermore, based on Theorem 1 and Proposition 2, we prove self-normalized upper bounds on the estimation error under Scheme II, which corresponds to the random design setting in linear regression. For any probability measure $\mathfrak{m}$ on $S \subseteq \mathbb{R}$, define $\Sigma^{(j)} := \mathbb{E}_{x^{(j)}\sim P_X^{(j)}}\left[\int_S \Phi_j\Phi_j^\top d\mathfrak{m}\right]$ and $\Sigma_n := \sum_{j=1}^n \Sigma^{(j)}$ for $j, n \in \mathbb{N}$.

Theorem 4 (Self-normalized bound in random design setting)**.** *Assume $\mathfrak{m}$ is a probability measure on S, $\{(x^{(j)}, y^{(j)})\}_{j\in\mathbb{N}}$ is sampled according to Scheme II with F defined in (1), and $\mu_{\min}\left(\Sigma^{(j)}\right) \geqslant \sigma_{\min}$ for some constant $\sigma_{\min} > 0$ and all $j \in \mathbb{N}$. For any $\delta \in (0, 1/2)$ and $n \geqslant \frac{32d^2}{\sigma_{\min}^2}\log(\frac{d}{\delta})$, with probability at least $1-2\delta$, the estimator in (2) with $\lambda = 0$ satisfies*

$$\|\widehat{\theta} - \theta_*\|_{\Sigma_n} \leqslant 2\varepsilon\left(n, d, \delta, \sigma_{\min}\right). \tag{8}$$

Moreover, for regularized estimators, we have the following result.

Proposition 5 (Self-normalized bound in random design setting for regularized estimator)**.** *Under the same assumptions as Theorem 4, for any* $\lambda > 0$, $\delta \in (0, 1/2)$, *and* $n \geqslant \frac{32d^2}{\sigma_{\min}^2} \log\left(\frac{d}{\delta}\right)$, *with probability at least* $1 - 2\delta$, *the estimator defined in* (2) *satisfies*

$$\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n} \leqslant \sqrt{2}\varepsilon_\lambda(n, d, \delta). \tag{9}$$

The proofs of Theorem 4 and Proposition 5 are given in Appendix B.2. As before, they convey that in the random design setting, the self-normalized errors $\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n}$ and $\|\widehat{\theta} - \theta_*\|_{\Sigma_n}$ scale as $\widetilde{O}\left(\sqrt{d}\right)$ with high probability in the ℓ^2-regularized and unregularized cases. Moreover, we once again note that Theorem 4 and Proposition 5 imply upper bounds on the (un-normalized) error $\|\widehat{\theta}_\lambda - \theta_*\|$. For example, since $\sigma_{\min}$ is a positive constant, (8) implies that $\|\widehat{\theta} - \theta_*\| \leqslant 2\mu_{\min}(\Sigma_n)^{-1/2}\varepsilon\left(n, d, \delta, \sigma_{\min}\right) = \widetilde{O}\left(\sqrt{d/n}\right)$ with high probability since $\mu_{\min}(\Sigma_n) \geqslant n\sigma_{\min}$ by Weyl's inequality (Weyl, 1912). Moreover, it is not hard to show that for general Σ_n and $\lambda > 0$, (9) can be generalized to $\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n(\lambda)} = \widetilde{O}(\sqrt{d})$ which again implies that $\|\widehat{\theta}_\lambda - \theta_*\| = \widetilde{O}(\min\{1, \sqrt{d/(\mu_{\min}(\Sigma_n) + 1)}\})$.

The main idea in the proofs of Theorem 4 and Proposition 5 is to establish a high probability lower bound on $\mu_{\min}(\Delta_n)$, where $\Delta_n := \Sigma_n^{-\frac{1}{2}}\left(U_n - \Sigma_n\right)\Sigma_n^{-\frac{1}{2}}$. This can be achieved using the *matrix Hoeffding's inequality* (Tropp, 2012, Theorem 1.3). Then, we show that for any $\lambda \geqslant 0$, $\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n} \leqslant (1+\mu_{\min}\left(\Delta_n\right))^{-1/2}\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$. For Theorem 4, we prove that $\mu_{\min}(U_n) \geqslant \mu_{\min}(\Sigma_n)(\mu_{\min}(\Delta_n) + 1)$. Then, we can lower bound $\mu_{\min}(U_n)$ in (7) by a multiple of $\mu_{\min}(\Sigma_n)$ with high probability. Thus, (8) follows from (7) and the high probability lower bound on $\mu_{\min}(\Delta_n)$. For Proposition 5, (9) follows from (6) and the high probability lower bound on $\mu_{\min}(\Delta_n)$.

We briefly compare our results in this section with related results in the literature. In the (canonical, finite dimensional) adversarial linear regression setting, Abbasi-Yadkori et al. (2011a) and Zhou et al. (2021) show an $\widetilde{O}\left(\sqrt{d}\right)$ upper bound for the self-normalized error of the ridge least-squares estimator. Specifically, the upper bound in Hsu et al. (2012b) is $\widetilde{O}(R\sqrt{d})$ for the case where the noise term is R-sub-Gaussian and the upper bound in Zhou et al. (2021) is $\widetilde{O}(\sigma\sqrt{d} + R)$ for the case where the noise term is bounded by R with variance bounded by σ^2. The functional regression upper bound in (6) aligns precisely with this scaling (neglecting sub-dominant factors) with respect to (wrt) d and n. Moreover, the upper bounds of Abbasi-Yadkori et al. (2011a) and Zhou et al. (2021) are susceptible to the magnitudes of the responses, as evidenced by their multiplicative constants of d. In contrast, the multiplicative constant in our upper bound is 1, ensuring that our upper bound remains independent of response scales. This independence constitutes a notable advantage, distinguishing our linear model from those explored in previous works. In the (canonical, finite dimensional) random design linear regression setting, Hsu et al. (2012b) show $\widetilde{O}\left(\sqrt{d}\right)$ upper bounds for the self-normalized error of the unregularized least-squares estimator under some conditions on the distribution of covariates. The upper bound in (8) for the unregularized case also matches this scaling (neglecting sub-dominant factors). Nevertheless, it's crucial to acknowledge that our linear model (1) is characterized by a unique complexity. Unlike the canonical linear regression framework, where the features are finite-dimensional vectors, and the response is a scalar, both features and response are functions in our model. Consequently, the theoretical results of Abbasi-Yadkori et al. (2011a), Zhou et al. (2021), and Hsu et al. (2012b) are not applicable to our estimators. This intricacy introduces numerous analytical challenges, setting it apart from the conventional linear regression paradigm. Furthermore, in our infinite dimensional model (18) studied later, we elevate the parameter from a finite-dimensional vector to a function, ushering in even more formidable complexities and challenges during the analysis.

Finally, we note that an upper bound on $\|\widehat{\theta}_\lambda - \theta_*\|$ immediately implies an upper bound on the KS distance between our estimated CDF and the true one. Let $\widehat{F}_\lambda(x, \cdot) := \widetilde{\theta}_\lambda^\top \Phi(x, \cdot)$ denote the estimated CDF for any $x \in \mathcal{X}$. Then, under the linear model (1), we have

$$\begin{aligned}\sup_{x\in\mathcal{X}} \mathrm{KS}(\widehat{F}_\lambda(x, \cdot), F(x, \cdot)) = \sup_{x\in\mathcal{X}, t\in S} |(\widetilde{\theta}_\lambda - \theta_*)^\top \Phi(x, t)| &\leqslant \|\widetilde{\theta}_\lambda - \theta_*\| \sup_{x\in\mathcal{X}, t\in S} \|\Phi(x, t)\| \\ &\leqslant \sqrt{d}\|\widehat{\theta}_\lambda - \theta_*\|,\end{aligned}$$

where we use the Cauchy-Schwarz inequality and the fact that $\sup_{x\in\mathcal{X},t\in S}\|\Phi(x,t)\|\leqslant\sqrt{d}$. Since $\|\widehat{\theta}_\lambda-\theta_*\|=\widetilde{O}\big(\min\{1,\sqrt{d/(1+\mu_{\min}(U_n))}\}\big)$ (see discussion below Proposition 2 and 5) and $\widehat{F}_\lambda,F\in[0,1]$, we have $\sup_{x\in\mathcal{X}}\mathrm{KS}(\widehat{F}_\lambda(x,\cdot),F(x,\cdot))=\widetilde{O}\big(\min\{1,d/\sqrt{(1+\mu_{\min}(U_n))}\}\big)$. It is worth mentioning that the above upper bound on the estimation error in KS distance may not be sharp because we focus on a tight analysis of the estimation of θ_* instead of $F(x,\cdot)$ for some $x\in\mathcal{X}$. Nevertheless, in Appendix A, we show that when $\mu_{\min}(U_n)=0$ ($\mu_{\min}(\Sigma_n)=0$), the minimax risk in terms of the uniform KS distance for the estimation of F is lower bounded by $\Omega(1)$ for the adversarial (random) setting.

3.3 Burn-in-time-free upper bound

Note that the theoretical guarantees in Theorem 4 and Proposition 5 require a burn-in time of the sample size n: $n\geqslant\frac{32d^2}{\sigma_{\min}^2}\log(\frac{d}{\delta})$. Motivated by Pires & Szepesvári (2012), we propose a new estimator $\breve{\theta}_\lambda$ in (10) to eliminate the burn-in time of n:

$$\breve{\theta}_\lambda\in\arg\min_{\theta\in\mathbb{R}^d}\left(\|U_n(\lambda)\theta-u_n\|+\Delta_n^U(\delta)\|\theta\|\right), \tag{10}$$

where $\lambda\geqslant0$, $\delta\in(0,1)$, $u_n:=\sum_{j=1}^n\int_S \mathrm{I}_{y^{(j)}}\Phi_j d\mathfrak{m}$, and $\Delta_n^U(\delta)$ is a positive number such that $\Delta_n^U(\delta)\geqslant\|U_n-\Sigma_n\|$ with probability at least $1-\delta$. For notatoinal convenience, we use $\breve{\theta}$ to denote $\breve{\theta}_0$. To calculate $\breve{\theta}_\lambda$ in (10), it is necessary to first choose $\Delta_n^U(\delta)$ for which we prove a lower bound in the following lemma.

Lemma 6. *Assume $\mathfrak{m}$ is a probability measure on S and $\{(x^{(j)},y^{(j)})\}_{j\in\mathbb{N}}$ is sampled according to Scheme II with F defined in (1). For any $\delta\in(0,1)$ and $n\in\mathbb{N}$, any $\Delta_n^U(\delta)\geqslant d\sqrt{8n\log(d/\delta)}$ satisfies $\Delta_n^U(\delta)\geqslant\Delta_n^U$ with probability at least $1-\delta$.*

The proof of Lemma 6 follows from the *matrix Hoeffding's inequality* (Tropp, 2012, Theorem 1.3) and the boundedness of CDFs, and is provided in Appendix E. Then, we show the following upper bound on the estimation error of $\breve{\theta}_\lambda$.

Theorem 7 (Self-normalized bound in random setting without burn-in time)**.** *Under the same assumptions as Lemma 6, for any $\delta\in(0,1/2)$ and $n\in\mathbb{N}$, if $\mu_{\min}(\Sigma_n)>0$, then, with probability at least $1-2\delta$, the estimator defined in (10) with $\lambda=0$ satisfies*

$$\|\breve{\theta}-\theta_*\|\leqslant\frac{1}{\mu_{\min}(\Sigma_n)}\left[2d\sqrt{8n\log(d/\delta)}\|\theta_*\|+2\left(\sqrt{nd}+\sqrt{8nd\log(1/\delta)}+\frac{4}{3}\sqrt{d}\log(1/\delta)\right)\right]. \tag{11}$$

The proof of Theorem 7 is provided in Appendix B.3. It conveys that for any $n\in\mathbb{N}$, as long as $\mu_(\Sigma_n)>0$, $\|\breve{\theta}-\theta_*\|\leqslant\widetilde{O}(\frac{d\sqrt{n}}{\mu_{\min}(\Sigma_n)})$ holds with high probability. Under the assumption that $\mu_{\min}(\Sigma^{(j)})\geqslant\sigma_{\min}$ for any $j\in\mathbb{N}$ as in Theorem 4 and Proposition 5, we have that $\|\breve{\theta}-\theta_*\|\leqslant\widetilde{O}\big(d/\sqrt{n}\big)$ with high probability for any $n\in\mathbb{N}$. Compared with the $\widetilde{O}\big(\sqrt{d/n}\big)$ upper bound of the estimation error of $\widehat{\theta}$ in Theorem 4, $\breve{\theta}$ suffers a larger error rate wrt the dimension d in order to eliminate the burn-in time of the sample size n. Thus, $\breve{\theta}$ is more applicable to the estimation of θ_* for small sample size and small dimension.

The proof of Theorem 7 builds on the upper bound shown in Pires & Szepesvári (2012) for the estimator that minimizes the unsquared penalized loss as in (10). By Pires & Szepesvári (2012, Theorem 3.4), we have that with probability at least $1-\delta$,

$$\|\Sigma_n(\lambda)\breve{\theta}_\lambda-\Sigma_n\theta_*\|\leqslant(\lambda+2\Delta_n^U(\delta))\|\theta_*\|+2\|u_n-\mathbb{E}[u_n]\|.$$

Then, we can bound $\|u_n-\mathbb{E}[u_n]\|$ with high probability by the vector Bernstein inequality (Hsu et al., 2012a, Proposition 1.2). By setting $\lambda=0$ and $\Delta_n^U(\delta)=d\sqrt{8n\log(d/\delta)}$ as is guaranteed by Lemma 6, we obtain (11) after some derivation.

4 Minimax lower bounds

To show that our estimator (2) is minimax optimal, we prove information theoretic lower bounds on the ℓ^2-norm of the estimation error for any estimator. Recall that for a distribution family $\mathcal{Q}$ and (parameter)

function $\xi : \mathcal{Q} \to \mathbb{R}^d$, the minimax ℓ^2-risk is defined as,

$$\mathfrak{R}(\xi(\mathcal{Q})) := \inf_{\hat{\xi}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{z \sim Q}[\|\hat{\xi}(z) - \xi(Q)\|], \tag{12}$$

where the infimum is over all (possibly randomized) estimators $\hat{\xi}$ of ξ based on a sample z, and the supremum is over all distributions in the family $\mathcal{Q}$. To specialize this definition for our problem, for any $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^d$, let $P^{\Phi}_{Y|x;\theta}$ denote the probability measure defined by the CDF $\theta^\top \Phi(x, \cdot)$. Moreover, for any sequence $x^{1:n} := (x^{(1)}, \ldots, x^{(n)}) \in \mathcal{X}^n$, define the collection of product measures, $\mathcal{P}^d_{x^{1:n}} := \left\{ \otimes_{j=1}^n P^{\Phi}_{Y|x^{(j)};\theta} : \theta \in \Delta^{d-1}, \Phi \in \mathfrak{B}_d \right\}$, where

$$\mathcal{B}_d := \{[\phi_1, \ldots, \phi_d]^\top : \phi_i : \mathcal{X} \times \mathbb{R} \to [0,1] \text{ is measurable and } \phi_i(x, \cdot) \text{ is a CDF on } \mathbb{R}, \forall i \in [d]\}.$$

For any distribution $P \in \mathcal{P}^d_{x^{1:n}}$, let $\theta(P)$ be a parameter in Δ^{d-1} such that $P = \otimes_{j=1}^n P^{\Phi}_{Y|x^{(j)};\theta}$. Then, we have the following theorem in the adversarial setting.

Theorem 8 (Information theoretic lower bound in adversarial setting)**.** *For any $d \geqslant 2$ and any sequence $x^{1:n} = (x^{(1)}, \ldots, x^{(n)}) \in \mathcal{X}^n$, we have*

$$\mathfrak{R}(\theta(\mathcal{P}^d_{x^{1:n}})) = \Omega\left(\min\{1, \sqrt{d/(1 + \mu_{\min}(U_n))}\}\right). \tag{13}$$

The proof uses *Fano's method* (Fano, 1961) and is given in Appendix C.1. Note that strictly speaking, the above theorem is written for the fixed design setting. However, a lower bound in the fixed design setting also implies the same lower bound in adversarial setting. Furthermore, by our discussion below Theorem 1, (6) implies that in the adversarial setting,

$$\mathbb{P}\left[\|\widehat{\theta}_\lambda - \theta_*\|^2 \geqslant \frac{C_1 d \log(n) + C_2 + C_3 r}{1 + \mu_{\min}(U_n)}\right] \leqslant e^{-r}$$

for $r > 0$ and some constants C_1, C_2, and C_3, which immediately implies that $\mathbb{E}[\|\widehat{\theta}_\lambda - \theta_*\|] = \widetilde{O}(\sqrt{d/(1 + \mu_{\min}(U_n))})$ and $\mathbb{E}[\|\widetilde{\theta}_\lambda - \theta_*\|] = \widetilde{O}(\min\{1, \sqrt{d/(1 + \mu_{\min}(U_n))}\})$. Thus, our estimator $\widetilde{\theta}_\lambda$ is minimax optimal. When $\mu_{\min}(U_n) = \Theta(n)$, the optimal rate is $\widetilde{\Theta}(\sqrt{d/n})$ in the adversarial setting.

In the proof of Theorem 8, we construct a family of $\Omega(a/\sqrt{d})$-packing subsets of Δ^{d-1} for $a \in (0,1)$ under ℓ^2-distance. We then show that when $\phi_1, \ldots, \phi_d$ are the CDFs of d Bernoulli distributions, for any $\theta^{(1)} \neq \theta^{(2)}$ in such a packing subset, the Kullback-Leibler (KL) divergence (see definition in Appendix C.1) satisfies

$$D(P_{Y|x^{(j)};\theta^{(1)}} \| P_{Y|x^{(j)};\theta^{(2)}}) = O(a^2(1 + \mu_{\min}(U_n))/d)$$

for any $j \in [n]$. Since the above family of Bernoulli distributions is a subset of $\mathcal{P}^d_{x^{1:n}}$, we are able to show that $\mathfrak{R}(\theta(\mathcal{P}_{x^{1:n}})) = \Omega\big(\sqrt{d/(1 + \mu_{\min}(U_n))}\big)$ using Fano's method and the aforementioned bound on KL divergence.

Next, to analyze minimax ℓ^2-risk under the random setting, let $\mathcal{D}_{\mathcal{X}}$ denote the set of all probability distributions on $\mathcal{X}$. For any $P_X \in \mathcal{D}_{\mathcal{X}}$, let $P_X P^{\Phi}_{Y|X;\theta}$ denote the joint distribution of (X, Y) such that the marginal distribution of X is P_X and the conditional distribution of Y given $X = x$ is $P^{\Phi}_{Y|x;\theta}$. Define the distribution family

$$\mathcal{P}^d_n := \left\{ \otimes_{j=1}^n P_X^{(j)} P^{\Phi}_{Y|X;\theta} : \theta \in \mathbb{R}^d,\ \Phi \in \mathcal{B}_d,\ P_X^{(j)} \in \mathcal{D}_{\mathcal{X}} \right\},$$

and for any $P \in \mathcal{P}^d_n$, let $\theta(P)$ denote the parameter in Δ^{d-1} such that $P = \otimes_{j=1}^n P_X^{(j)} P^{\Phi}_{Y|X;\theta}$. Clearly, for any $x^{1:n} \in \mathcal{X}^n$, we have $\left\{ \otimes_{j=1}^n \delta_{x^{(j)}} P_{Y|X;\theta} : \theta \in \Delta^{d-1} \right\} \subseteq \mathcal{P}^d_n$. Thus, each $\mathcal{P}^d_{x^{1:n}}$ is a collection of marginal distributions of elements belonging to such subsets of $\mathcal{P}^d_n$. Then, by the definition of minimax ℓ^2-risk, Theorem 8 immediately implies the following corollary.

Corollary 9 (Information theoretic lower bound in random setting)**.** *For any $d \geqslant 2$,*

$$\mathfrak{R}(\theta(\mathcal{P}^d_n)) = \Omega\left(\min\{1, \sqrt{d/(1 + \mu_{\min}(\Sigma_n))}\}\right) \tag{14}$$

The proof is given in Appendix C.2. By the discussion below Proposition 5, our estimator $\widetilde{\theta}_\lambda$ ($\lambda > 0$) is minimax optimal. When $\mu_{\min}(\Sigma_n) = \Theta(n)$ as in Theorem 4 and Corollary 5, the lower bound on the Euclidean norm of the estimation error is also $\Omega\big(\sqrt{d/n}\big)$ in random setting. Following the discussion below Theorem 4, (8) implies that in random setting, $\mathbb{P}[\|\widehat{\theta} - \theta_*\| \geqslant C_1\sqrt{d/n} + C_2\sqrt{rd/n} + C_3 r\sqrt{d}/n] \leqslant e^{-r}$ for $r > 0$ and constants C_1, C_2, and C_3, which immediately implies that $\mathbb{E}[\|\widehat{\theta} - \theta_*\|] = \widetilde{O}(\sqrt{d/n})$. Thus, the estimator (2) is minimax optimal with rate $\widetilde{\Theta}(\sqrt{d/n})$ in random setting when $\mu_{\min}(\Sigma_n) = \Theta(n)$.

5 Mismatched model

In general, a mismatch may exist between the true target function and our linear model (1) with basis Φ. So, in analogy with canonical linear regression where additive Gaussian random variables are used to model the error term (Montgomery et al., 2021), we consider the following *mismatched model*:

$$F(x,t) = \theta_*^\top \Phi(x,t) + e(x,t), \quad \forall\ x \in \mathcal{X}, t \in \mathbb{R}, \tag{15}$$

where an additive error function depending on the context is included to model the mismatch in (1). Note that in (15), each $F(x,\cdot)$ is a CDF and $e : \mathcal{X} \times S \to [-1,1]$ is a measurable function. One equivalent interpretation of (15) is as follows. Suppose that their exists another contextual CDF function ϕ_e such that $F(x,\cdot)$ is a mixture of the linear model $\theta_*^\top \Phi(x,\cdot)$ and the new feature function $\phi_e(x,\cdot)$, i.e., for some $q \in [0,1]$,

$$F(x,t) = (1-q)\theta_*^\top \Phi(x,t) + q\phi_e(x,t) = \theta_*^\top \Phi(x,t) + q\left(\phi_e(x,t) - \theta_*^\top \Phi(x,t)\right),\ \forall\ x \in \mathcal{X}, t \in \mathbb{R}.$$

Then, we naturally obtain an additive error function $e(x,t) = q\left(\phi_e(x,t) - \theta_*^\top \Phi(x,t)\right)$.

Given a sample $\{(x^{(j)}, y^{(j)})\}_{j\in[n]}$ generated using the mismatched model (15), let $e_j(t)$ denote $e(x^{(j)},t)$ for $j \in [n]$. Moreover, define $E_n := \sum_{j=1}^n \int_S e_j \Phi_j d\mathfrak{m}$ and $B_n := \mathbb{E}[E_n] = \sum_{j=1}^n \mathbb{E}\left[\int_S e_j \Phi_j d\mathfrak{m}\right]$. Then, we have the following theoretical guarantees for the task of estimating θ_* using the estimator in (2) in the adversarial and random settings.

Theorem 10 (Self-normalized bound in mismatched adversarial setting)**.** *Assume $\mathfrak{m}$ is a probability measure on S and $\{(x^{(j)}, y^{(j)})\}_{j\in\mathbb{N}}$ is sampled according to Scheme I with F defined in (15). For any $\lambda > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, the estimator defined in (2) satisfies that for all $n \in \mathbb{N}$,*

$$\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)} \leqslant \varepsilon_\lambda(n,d,\delta) + \|E_n\|/\sqrt{\lambda}. \tag{16}$$

The proof of Theorem 10 follows the same approach as the proof of Theorem 1, and it is provided in Appendix I.1. Furthermore, Theorem 10 implies Corollary 11 for the mismatched random setting.

Corollary 11 (Self-normalized bound in mismatched random setting)**.** *Assume $\mathfrak{m}$ is a probability measure on S, $\{(x^{(j)}, y^{(j)})\}_{j\in\mathbb{N}}$ is sampled according to Scheme II with F defined in (15), and $\mu_{\min}(\Sigma^{(j)}) \geqslant \sigma_{\min}$ for some $\sigma_{\min} > 0$ and all $j \in \mathbb{N}$. For any $\lambda > 0$, $\delta \in (0, 1/2)$, and $n \geqslant \frac{32d^2}{\sigma_{\min}^2}\log\left(\frac{d}{\delta}\right)$, with probability at least $1-2\delta$, the estimator defined in (2) satisfies*

$$\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n} \leqslant \sqrt{2}\varepsilon_\lambda(n,d,\delta) + \sqrt{2/\lambda}\|B_n\|. \tag{17}$$

The proof of Corollary 11 is given in Appendix I.2. It follows from the proofs of Theorem 10 and Proposition 5.

In the adversarial setting, comparing (16) in Theorem 10 with (6) in Theorem 1, we see that the effect of the additive error in the mismatched model is captured by the additional $\|E_n\|/\sqrt{\lambda}$ term in our self-normalized error upper bound. Similarly, in the random setting, comparing (17) in Corollary 11 with (9), we again see that the effect of the additive error is captured by the additional $\sqrt{2/\lambda}\|B_n\|$ term in the self-normalized upper bound.

6 Infinite dimensional model

So far, we have been assuming finite-dimensional models where the number of base contextual CDFs ϕ_i's per sample is finite. It is natural to consider generalizing the linear model to be infinite-dimensional and estimating an infinite dimensional parameter θ_* which shall be considered as a function on the "index" space of the base functions. In Section 6.1, we formally introduce the infinite-dimensional linear model. We present necessary definitions and technical facts for the statement of the estimator and theorem in Section 6.2. We extend the estimator $\widehat{\theta}_\lambda$ in (2) with properly chosen regularization and provide a high probability upper bound on the estimation error of the generalized estimator in Section 6.3.

6.1 Formal model

First, we introduce the infinite dimensional index space Ω and the generalized basis function Φ. Assume that $(\Omega, \mathcal{F}_\Omega, \mathfrak{n})$ is a measure space with $\mathfrak{n}(\Omega) < \infty$ and $\Phi: \mathcal{X} \times \Omega \times \mathbb{R} \to [0,1]$, $(x, \omega, t) \mapsto \Phi(x, \omega, t)$ is a $(\mathcal{B}(\mathcal{X}) \otimes \mathcal{F}_\Omega \otimes \mathcal{B}(\mathbb{R}))/\mathcal{B}([0,1])$-measurable function (see Appendix B for the explanations of notation) such that for any $x \in \mathcal{X}$ and $\mathfrak{n}$-a.e. $\omega \in \Omega$, $\Phi(x, \omega, \cdot)$ is the CDF of some $\mathbb{R}$-valued random variable with its range contained in some Borel set $S \subseteq \mathbb{R}$. Define the following mapping,

$$\langle \cdot, \cdot \rangle : \ \mathcal{L}^2(\Omega, \mathfrak{n}) \times \mathcal{L}^2(\Omega, \mathfrak{n}) \to \mathbb{R}, \ (f, g) \mapsto \int_\Omega f g d\mathfrak{n}.$$

Then, $\langle \cdot, \cdot \rangle$ is an inner product on $\mathcal{L}^2(\Omega, \mathfrak{n})$ and $(\mathcal{L}^2(\Omega, \mathfrak{n}), \langle \cdot, \cdot \rangle)$ is a Hilbert space. Let $\| \cdot \|$ denote the norm by induced $\langle \cdot, \cdot \rangle$ on $\mathcal{L}^2(\Omega, \mathfrak{n})$. Assume that $(\mathcal{L}^2(\Omega, \mathfrak{n}), \langle \cdot, \cdot \rangle)$ is separable. Then, there exists a countable orthonormal basis on $(\mathcal{L}^2(\Omega, \mathfrak{n}), \langle \cdot, \cdot \rangle)$. For notational convenience, we write $\mathcal{L}^2(\Omega, \mathfrak{n})$ to represent the Hilbert space $(\mathcal{L}^2(\Omega, \mathfrak{n}), \langle \cdot, \cdot \rangle)$. Let $\boldsymbol{e} = \{e_i\}_{i=1}^\infty$ be an arbitrary countable orthonormal basis of $\mathcal{L}^2(\Omega, \mathfrak{n})$ and $\boldsymbol{\sigma} = \{\sigma_i\}_{i \in \mathbb{N}}$ be an arbitrary real sequence such that $\sum_{i=1}^\infty |\sigma_i| < \infty$. Assume that there exists some *unknown* $\theta_* \in \mathcal{H}_{\boldsymbol{\sigma}, \boldsymbol{e}}$ such that $\theta_* \geqslant 0$ $\mathfrak{n}$-a.e., $\int_\Omega \theta_* \mathfrak{n} = 1$, and the target function F satisfies the following model

$$F(x, t) = \langle \theta_*(\cdot), \Phi(x, \cdot, t) \rangle, \quad \forall \ x \in \mathcal{X}, \ t \in \mathbb{R}. \tag{18}$$

6.2 Technical preliminaries

The proofs of the theoretical results in this section are provided in Appendix F. Given the sample $\{(x^{(j)}, y^{(j)})\}_{j \in \mathbb{N}} \subseteq \mathcal{X} \times \mathbb{R}$, define the function $\Phi_j : \ \Omega \times \mathbb{R} \to [0,1]$, $(\omega, t) \mapsto \Phi(x_j, \omega, t)$ for any $j \in \mathbb{N}$. Since $\mathfrak{n}(\Omega) < \infty$ and $|\Phi(x, \omega, t)| \leqslant 1$ for any $x \in \mathcal{X}$, $\mathfrak{n}$-a.e. $\omega \in \Omega$, and any $t \in \mathbb{R}$, we have $\Phi_j \in \mathcal{L}^2(\Omega, \mathfrak{n})$ for any $j \in \mathbb{N}$. Then, for any $j \in \mathbb{N}$, we define,

$$\Psi_j : \mathcal{L}^2(\Omega, \mathfrak{n}) \times S \to \mathbb{R}, \ (\theta, t) \mapsto \langle \theta(\cdot), \Phi_j(\cdot, t) \rangle.$$

Then, by Holder's inequality, for any $j \in \mathbb{N}$ and $\theta \in \mathcal{L}^2(\Omega, \mathfrak{n})$, we have $\sup_{t \in \mathbb{R}} |\Psi_j(\theta, t)| \leqslant \mathfrak{n}(\Omega) \int_\Omega |\theta|^2 d\mathfrak{n} < \infty$. It follows that $\Psi_j(\theta, \cdot) \in \mathcal{L}^2(S, \mathfrak{m})$. Moreover, we have that for any $n \in \mathbb{N}$, any $\theta \in \mathcal{L}^2(\Omega, \mathfrak{n})$, and $\mathfrak{n}$-a.e. $\omega \in \Omega$,

$$\left| \sum_{j=1}^n \int_S \Psi_j(\theta, t) \Phi_j(\omega, t) \mathfrak{m}(dt) \right| \leqslant \sum_{j=1}^n \int_S |\Psi_j(\theta, t) \Phi_j(\omega, t)| \, \mathfrak{m}(dt) \leqslant n \mathfrak{n}(\Omega) \int_\Omega |\theta|^2 d\mathfrak{n}.$$

Since $\mathfrak{n}(\Omega) < \infty$, it follows that the function $\omega \mapsto \sum_{j=1}^n \int_S \Psi_j(\theta, t) \Phi_j(\omega, t) \mathfrak{m}(dt)$ is in $\mathcal{L}^2(\Omega, \mathfrak{n})$. Thus, for any $n \in \mathbb{N}$, we can define an operator $U_n : \mathcal{L}^2(\Omega, \mathfrak{n}) \to \mathcal{L}^2(\Omega, \mathfrak{n})$ by

$$(U_n \theta)(\omega) := \sum_{j=1}^n \int_S \Psi_j(\theta, t) \Phi_j(\omega, t) \mathfrak{m}(dt) = \sum_{j=1}^n \int_S \langle \theta(\cdot), \Phi_j(\cdot, t) \rangle \Phi_j(\omega, t) \mathfrak{m}(dt) \tag{19}$$

for any $\theta \in \mathcal{L}^2(\Omega, \mathfrak{n})$. We show the following properties of U_n.

Lemma 12. *For any $n \in \mathbb{N}$, U_n is a self-adjoint positive Hilbert-Schmidt integral operator with $\|U_n\| \leqslant n \mathfrak{n}(\Omega)$. Thus, it is also a compact operator.*

Now, we assume that U_n satisfies Assumption 13 for some $n \in \mathbb{N}$.

Assumption 13. *Assume that e_i is an eigenfunction of U_n with the corresponding eigenvalue denoted with λ_i for any $i \in \mathbb{N}$.*

Under the Assumption 13 on U_n, we can conclude from Lemma 12 that:

Corollary 14. *Assume that U_n satisfies Assumption 13 for some $n \in \mathbb{N}$. Then, we have $0 \leqslant \lambda_i \leqslant n\mathfrak{n}(\Omega)$ for any $i \in \mathbb{N}$ and $\lambda_i \to 0$.*

Define the set $\mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n}) := \left\{\theta \in \mathcal{L}^2(\Omega, \mathfrak{n}) : \sum_{i=1}^{\infty} \frac{|\langle e_i, \theta\rangle|^2}{\sigma_i^4} < \infty\right\}$. Then, we have that

Lemma 15. *For any $\boldsymbol{\sigma} = \{\sigma_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R}$ satisfying $\sum_{i=1}^{\infty} |\sigma_i| < \infty$, $\mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n})$ is a linear subspace of $\mathcal{L}^2(\Omega, \mathfrak{n})$.*

For any $\theta \in \mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n})$, we have

$$\sum_{i=1}^{\infty} \left|\lambda_i + \frac{1}{\sigma_i^2}\right|^2 |\langle e_i, \theta\rangle|^2 \leqslant \sum_{i=1}^{\infty} 2\lambda_i^2 |\langle e_i, \theta\rangle|^2 + \sum_{i=1}^{\infty} \frac{2}{\sigma_i^4} |\langle e_i, \theta\rangle|^2 \leqslant 2\|U_n\theta\|^2 + 2\sum_{i=1}^{\infty} \frac{|\langle e_i, \theta\rangle|^2}{\sigma_i^4} < \infty,$$

which implies that $\{\sum_{i=1}^{m}(\lambda_i + \frac{1}{\sigma_i^2})\langle e_i, \theta\rangle e_i\}_{m\in\mathbb{N}}$ is a Cauchy sequence and thus converges in $\mathcal{L}^2(\Omega, \mathfrak{n})$ to $\sum_{i=1}^{\infty}(\lambda_i + \frac{1}{\sigma_i^2})\langle e_i, \theta\rangle e_i \in \mathcal{L}^2(\Omega, \mathfrak{n})$. Therefore, we can define the operator $U_{n,\boldsymbol{\sigma}} : \mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n}) \to \mathcal{L}^2(\Omega, \mathfrak{n})$, $\theta \mapsto \sum_{i=1}^{\infty}\left(\lambda_i + \frac{1}{\sigma_i^2}\right)\langle e_i, \theta\rangle e_i$ for which we show the following lemma.

Lemma 16. *$U_{n,\boldsymbol{\sigma}}$ is bijective linear operator from $\mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n})$ onto $\mathcal{L}^2(\Omega, \mathfrak{n})$. $U^{-1}_{n,\boldsymbol{\sigma}}$ is a bounded linear operator on $\mathcal{L}^2(\Omega, \mathfrak{n})$ with $\|U^{-1}_{n,\boldsymbol{\sigma}}\| \leqslant \sup_{i\in\mathbb{N}} \sigma_i^2$ and $U^{-1}_{n,\boldsymbol{\sigma}}\theta = \sum_{i=1}^{\infty} \frac{\sigma_i^2\langle e_i, \theta\rangle}{1+\lambda_i\sigma_i^2} e_i$ for any $\theta \in \mathcal{L}^2(\Omega, \mathfrak{n})$. Moreover, $U^{-1}_{n,\boldsymbol{\sigma}}$ is positive and self-adjoint.*

Consequently, we can define the following mapping

$$\|\cdot\|_{U_{n,\boldsymbol{\sigma}}} : \mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n}) \to [0, \infty),\ \theta \mapsto \sqrt{\langle \theta, U_{n,\boldsymbol{\sigma}}\theta\rangle} = \sqrt{\|\theta\|^2_{U_n} + \sum_{i=1}^{\infty} \frac{|\langle e_i, \theta\rangle|^2}{\sigma_i^2}}, \tag{20}$$

where $\|\theta\|_{U_n} := \sqrt{\langle \theta, U_n\theta\rangle}$ for any $\theta \in \mathcal{L}^2(\Omega, \mathfrak{n})$. Define the set

$$\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}} := \left\{\theta \in \mathcal{L}^2(\Omega) : \sum_{i=1}^{\infty} |\langle e_i, \theta\rangle|^2/\sigma_i^2 < \infty\right\}$$

and the mapping $\langle\cdot,\cdot\rangle_{\boldsymbol{\sigma},\boldsymbol{e}} :\ \mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}} \times \mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}} \to \mathbb{R},\ (f, g) \mapsto \sum_{i=1}^{\infty} \frac{\langle e_i, f\rangle\langle e_i, g\rangle}{\sigma_i^2}$. Similar to the proofs of Lemma 15, we can show that $\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$ is a linear subspace of $\mathcal{L}^2(\Omega, \mathfrak{n})$. Moreover, $(\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}, \langle\cdot,\cdot\rangle_{\boldsymbol{\sigma},\boldsymbol{e}})$ is also a separable Hilbert space with $\{\sigma_i e_i\}_{i\in\mathbb{N}}$ being an orthonormal basis. For notational convenience, we write $\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$ to represent the Hilbert space $(\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}, \cdot, \langle\cdot\rangle_{\boldsymbol{\sigma},\boldsymbol{e}})$ and use $\|\cdot\|_{\boldsymbol{\sigma},\boldsymbol{e}}$ to denote the induced norm on $\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$. Moreover, we show the following lemma.

Lemma 17. *For any real sequence $\boldsymbol{\sigma} = \{\sigma_i\}_{i\in\mathbb{N}}$ with $\lim_{i\to\infty} \sigma_i = 0$, $\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}} \subseteq \mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n})$.*

6.3 Self-normalized upper bound

Since we have proved that $\Psi_j(\theta, \cdot) \in \mathcal{L}^2(S, \mathfrak{m})$ for any $j \in \mathbb{N}$ and $\theta \in \mathcal{L}^2(\Omega, \mathfrak{n})$, the following loss function is well-defined on $\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$,

$$L(\theta; \boldsymbol{\sigma}) := \sum_{j=1}^{n} \|\mathrm{I}_{y^{(j)}}(\cdot) - \Psi_j(\theta, t)\|^2_{\mathcal{L}^2(S,\mathfrak{m})} + \sum_{i=1}^{\infty} \frac{|\langle e_i, \theta\rangle|^2}{\sigma_i^2}.$$

In fact, assuming the convention that $0/0 = 0$ and $1/0 = \infty$, we can extend the domain of $L(\cdot; \boldsymbol{\sigma})$ to $\mathcal{L}^2(\Omega, \mathfrak{n})$ by extending its codomain from $[0, \infty)$ to $[0, \infty]$.

We propose to estimate θ_* by minimizing the above loss function over $\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$:

$$\widehat{\theta}_{\boldsymbol{\sigma}} := \operatorname*{arg\,min}_{\theta \in \mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}} L(\theta; \boldsymbol{\sigma}). \tag{21}$$

Since $\sum_{i=1}^{\infty} \frac{|\langle e_i, \theta \rangle|^2}{\sigma_i^2} = \infty$ for any $\theta \in \mathcal{L}^2(\Omega, \mathfrak{n})$, we also have $\widehat{\theta}_{\boldsymbol{\sigma}} = \arg\min_{\theta \in \mathcal{L}^2(\Omega,\mathfrak{n})} L(\theta; \lambda)$. We have the following formula for $\widehat{\theta}_{\boldsymbol{\sigma}}$ in (21).

Proposition 18. *The solution to the optimization problem (21) is given as the following,*

$$\widehat{\theta}_{\boldsymbol{\sigma}} = U_{n,\boldsymbol{\sigma}}^{-1} \left(\sum_{j=1}^{n} \int_S \mathrm{I}_{y^{(j)}}(t) \Phi_j(\cdot, t) \mathfrak{m}(dt) \right). \tag{22}$$

The proof of Proposition 18 is provided in Appendix G. Under the adversarial setting, we show the following upper bound for the self-normalized estimation error of $\widehat{\theta}_{\boldsymbol{\sigma}}$ in (21).

Theorem 19 (Self-normalized bound in adversarial setting for infinite dimensional model)**.** *Assume $\mathfrak{m}$ is a probability measure on $(S, \mathcal{B}(S))$, $\mathfrak{n}$ is a finite measure on $(\Omega, \mathcal{F}_\Omega)$, $\boldsymbol{e} = \{e_i\}_{i=1}^{\infty}$ is an orthonormal basis of $\mathcal{L}^2(\Omega, \mathfrak{n})$, $\boldsymbol{\sigma} = \{\sigma_i\}_{i \in \mathbb{N}}$ is a real sequence satisfying $\sum_{i=1}^{\infty} |\sigma_i| < \infty$, $\theta_* \in \mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$ satisfies $\theta_* \geqslant 0$ $\mathfrak{n}$-a.e. and $\int_\Omega \theta_* \mathfrak{n} = 1$, and $\{(x^{(j)}, y^{(j)})\}_{j \in \mathbb{N}}$ is sampled according to Scheme I with F defined in (18).*

For any given $n \in \mathbb{N}$ and any $\delta \in (0,1)$, if U_n defined in (19) satisfies Assumption 13 and $\boldsymbol{\sigma}$ satisfies that $|\sigma_i| < \frac{1}{\sqrt{\lambda_i}}$ for any $i \in \mathbb{N}$, then, with probability at least $1-\delta$, the estimator $\widehat{\theta}_{\boldsymbol{\sigma}}$ defined in (21) satisfies

$$\|\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_*\|_{U_{n,\boldsymbol{\sigma}}} \leqslant \sqrt{\left(\sum_{i=1}^{\infty} \log\left(1 + \lambda_i \sigma_i^2\right) \right) + 2 \log \frac{1}{\delta}} + \|\theta_*\|_{\boldsymbol{\sigma},\boldsymbol{e}}. \tag{23}$$

In particular, for any given $n \in \mathbb{N}$ and any $\delta \in (0,1)$, if U_n defined in (19) satisfies Assumption 13 and $\boldsymbol{\sigma}$ satisfies that $|\sigma_i| < \frac{1}{\sqrt{n\mathfrak{n}(\Omega)}}$ for any $i \in \mathbb{N}$, then, with probability at least $1-\delta$, the estimator $\widehat{\theta}_{\boldsymbol{\sigma}}$ defined in (21) satisfies

$$\|\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_*\|_{U_{n,\boldsymbol{\sigma}}} \leqslant \sqrt{\left(\sum_{i=1}^{\infty} \log\left(1 + n\mathfrak{n}(\Omega)\sigma_i^2\right) \right) + 2 \log \frac{1}{\delta}} + \|\theta_*\|_{\boldsymbol{\sigma},\boldsymbol{e}}. \tag{24}$$

The detailed proof of Theorem 19 is provided in Appendix D. Since $\sum_{i=1}^{\infty} |\sigma_i| < \infty$ and $\theta_* \in \mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$, we have that $\|\theta_*\|_{\boldsymbol{\sigma},\boldsymbol{e}} < \infty$ and $\sum_{i=1}^{\infty} |\sigma_i|^2 < \infty$ which implies that

$$\sum_{i=1}^{\infty} \log\left(1 + \lambda_i \sigma_i^2\right) \leqslant \sum_{i=1}^{\infty} \log\left(1 + n\mathfrak{n}(\Omega)\sigma_i^2\right) < \infty.$$

Thus, the RHS terms in (23) and (24) are finite and $\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_* \in \mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega, \mathfrak{n})$. (24) conveys that with high probability,

$$\|\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_*\|_{U_{n,\boldsymbol{\sigma}}} \leqslant \widetilde{O}\left(1 + \sqrt{\sum_{i=1}^{\infty} \log(1 + n\mathfrak{n}(\Omega)\sigma_i^2)}\right).$$

When $\Omega = [d]$ for some $d \in \mathbb{N}$ and $\mathfrak{n}$ is the counting measure on Ω, (21) reduces to (2) after setting $e_i = \mathbb{1}_{\{i\}}$ and $\sigma_i = \frac{1}{\sqrt{\lambda}}$ for any $i \in [d]$ and some $\lambda > 0$. Then, by (20) and (24), we have $\|\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_*\|_{U_n} \leqslant \widetilde{O}\big(\sqrt{d}\big)$ and

$$\|\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_*\|_{U_n} \leqslant \widetilde{O}\big(\sqrt{d/(1 + \mu_{\min}(U_n))}\big),$$

which also recovers the result in Theorem 1. Thus, Theorem 19 is a generalization of Theorem 1 for the possibly infinite dimensional model (18).

The proof of Theorem 19 generalizes the approach used in the proof of Theorem 1 to the setting of the infinite dimensional model (18). However, there are plenty of technical challenges in dealing with the infinite dimensional $\mathcal{L}^2$ space. First of all, since the vectors in the proof of Theorem 1 are generalized to functions and the matrices are generalized to operators, we need to ensure that these functions are well-defined in some proper spaces and figure out the domain/codomain and properties (e.g., linearity, boundedness, self-adjointness, positivity, compactness, invertibility, etc) of those operators. As in the proof of Theorem 1, we would like to write $\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_* = U_{n,\boldsymbol{\sigma}}^{-1} W_n - U_{n,\boldsymbol{\sigma}}^{-1}(\varsigma\theta_*)$ where,

$$W_n := \sum_{j=1}^{n} \int_S \mathrm{I}_{y^{(j)}}(t)\Phi_j(\omega,t)\mathfrak{m}(dt) - \int_S \Psi_j(\theta_*,t)\Phi_j(\omega,t)\mathfrak{m}(dt),$$

and $\varsigma\theta_* := \sum_{i=1}^{\infty} \frac{\langle e_i,\theta_*\rangle}{\sigma_i^2} e_i$. However, this sequence $\{\sum_{i=1}^{m} \frac{\langle e_i,\theta_*\rangle}{\sigma_i^2} e_i\}_{m\in\mathbb{N}}$ only converges for $\theta_* \in \mathcal{L}^2_{\boldsymbol{\sigma}}(\Omega,\mathfrak{n})$ but not $\mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$. Thus, for general $\theta_* \in \mathcal{H}_{\boldsymbol{\sigma},\boldsymbol{e}}$, $\varsigma\theta_*$ does not exist and we instead consider the finite-rank operator $\varsigma_m : \theta \mapsto \sum_{i=1}^{m} \frac{\langle e_i,\theta\rangle}{\sigma_i^2} e_i$ on $\mathcal{L}^2(\Omega,\mathfrak{n})$ and the sequence $\{\theta_{*,m} := U_{n,\boldsymbol{\sigma}}^{-1}(U_n\theta_* + \varsigma_m\theta_*)\}_{m\in\mathbb{N}}$ which we show satisfies $\|\theta_{*,m} - \theta_*\|_{U_{n,\boldsymbol{\sigma}}} \to 0$ as $m \to \infty$. Then, since it suffices to bound

$$\|\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_{*,m}\|_{U_{n,\boldsymbol{\sigma}}} \leqslant \|U_{n,\boldsymbol{\sigma}}^{-1} W_n\|_{U_{n,\boldsymbol{\sigma}}} + \|U_{n,\boldsymbol{\sigma}}^{-1}\varsigma_m\theta_*\|_{U_{n,\boldsymbol{\sigma}}} \leqslant \|W_n\|_{U_{n,\boldsymbol{\sigma}}^{-1}} + \|\theta_*\|_{\boldsymbol{\sigma},\boldsymbol{e}}.$$

To bound $\|W_n\|_{U_{n,\boldsymbol{\sigma}}^{-1}}$, we use the martingale approach as in the proof of Theorem 1. However, after proving that $\left\{M_n(\alpha) := \exp\left(\langle\alpha, W_n\rangle - \frac{1}{2}\|\alpha\|^2_{U_n}\right)\right\}_{n\geqslant 0}$ is a super-martingale for any $\alpha \in \mathcal{L}^2(\Omega,\mathfrak{n})$ wrt the natural filtration $\{\mathcal{F}_n := \sigma(x_1, y_1, \ldots, x_n, y_n, x_{n+1})\}_{n\geqslant 0}$, it is difficult to pick a properly defined "Gaussian" random variable in $\mathcal{L}^2(\Omega,\mathfrak{n})$. Inspired by Lifshits (2012, Example 2.2), we define $\beta = \sum_{i=1}^{\infty} \sigma_i\zeta_i e_i$ with $\{\zeta_i\}_{i\in\mathbb{N}}$ being a sequence of independent $N(0,1)$-random variables. Note that $\beta \in \mathcal{L}^2(\Omega,\mathfrak{n})$ a.s. if $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$. Thus, we can define $\overline{M}_n := \mathbb{E}[M_n(\beta)|\mathcal{F}_\infty]$ with $\mathcal{F}_\infty := \sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$. Then, we prove that $\{M_n\}_{n\geqslant 0}$ is also a super-martingale wrt $\{\mathcal{F}_n\}_{n\geqslant 0}$ and the question remained is to calculate M_n. However, directly generalizing (41), we would get

$$\text{“}\|W_n\|^2_{U_{n,\boldsymbol{\sigma}}^{-1}} - \|\beta - U_{n,\boldsymbol{\sigma}}^{-1} W_n\|^2_{U_{n,\boldsymbol{\sigma}}} = 2\langle\beta, W_n\rangle - \|\beta\|^2_{U_{n,\boldsymbol{\sigma}}}\text{”}$$

which does not make sense because $\|\beta\|_{U_{n,\boldsymbol{\sigma}}}$ could be ∞ with positive probability. Since it is hard to deal with this in the integration over the the law of β, we instead adopt the similar approach as we do for θ_*. Define $\beta_m := \sum_{i=1}^{m} \sigma_i\zeta_i e_i$ and $W_{n,m} := \sum_{i=1}^{m}\langle e_i, W_n\rangle e_i$. Then, after some calculation, we get

$$\|W_{n,m}\|^2_{U_{n,\boldsymbol{\sigma}}^{-1}} - \|\beta_m - U_{n,\boldsymbol{\sigma}}^{-1} W_{n,m}\|^2_{U_{n,\boldsymbol{\sigma}}} = 2\langle\beta_m, W_{n,m}\rangle - \|\beta_m\|^2_{U_{n,\boldsymbol{\sigma}}} \text{ and}$$

$$\mathbb{E}[\exp(H_m)|\mathcal{F}_\infty] = \frac{1}{\sqrt{\prod_{i=1}^{m}(1+\lambda_i\sigma_i^2)}} \exp\left(\frac{1}{2}\|W_{n,m}\|^2_{U_{n,\boldsymbol{\sigma}}^{-1}}\right),$$

where $\exp(H_m) := \exp\left\{\langle\beta_m, W_{n,m}\rangle - \frac{1}{2}\|\beta_m\|^2_{U_n}\right\}$. Afterwards, we use dominated convergence theorem to conclude that,

$$\lim_{m\to\infty} \mathbb{E}[\exp(H_m)|\mathcal{F}_\infty] = \mathbb{E}[M_n|\mathcal{F}_\infty] = \overline{M}_n, \ a.s..$$

The verification the integrability of the dominating function $\exp\left(n\sum_{i=1}^{\infty}|\sigma_i\zeta_i| + \frac{1}{2}\sum_{i=1}^{\infty}\lambda_i\sigma_i^2\zeta_i^2\right)$ is also quite technical, during which the condition that $\sum_{i=1}^{\infty}|\sigma_i| < \infty$ is used. Finally, we obtain that $\overline{M}_n = \frac{1}{\sqrt{\prod_{i=1}^{\infty}(1+\lambda_i\sigma_i^2)}} \exp\left(\frac{1}{2}\|W_n\|^2_{U_{n,\boldsymbol{\sigma}}^{-1}}\right)$. Then, by applying *Doob's maximal inequality* for super-martingales, we can bound $\|W_n\|^2_{U_{n,\boldsymbol{\sigma}}^{-1}}$ which yields the final bound on $\|\widehat{\theta}_{\boldsymbol{\sigma}} - \theta_*\|_{U_{n,\boldsymbol{\sigma}}}$ in (23). (24) immediately follows from (23) and Corollary 14.

7 Numerical studies

In this section, we demonstrate the scaling of estimation errors of the proposed estimator empirically in our synthetic data experiments in Section 7.1 and illustrate the practical utility of the proposed estimator in our real data experiments in Section 7.2.

7.1 Synthetic data experiments

This section contains the experimental results on discrete and continuous synthetic data.

Bernoulli data experiments. To illustrate that our estimator (2) achieves the ℓ^2-error rate of $\widetilde{\Theta}(\sqrt{d/(1+\mu_{\min}(U_n))})$ in the estimation of θ under model (1), we consider the Bernoulli data generated according to the hard instance used to show the lower bound in the proof of Theorem 8 in Appendix C.1. Specifically, after choosing a true parameter $\theta_* \in \Delta^{d-1}$ of dimension $d \in \mathbb{N}$, for any $j \in \mathbb{N}$, we set $\phi_i(x_j, \cdot)$ as the CDF of Bernoulli(p_{ji}) for $i \in [d]$, where $p_j := [p_{j1}, \ldots, p_{jd}]^\top \in [0,1]^d$ is defined as follows. When $j \in [d]$, we set $p_{ji} = 1 - \frac{c_j}{2d^3} - \frac{c_j \mathbb{1}\{i=j\}}{2d^3}$; when $j > d$, we set $p_{ji} = 1 - \frac{c_j \mu_{\min}(R_{j-1})}{2d^2} - \frac{c_j \mu_{\min}(R_{j-1}) \mathbb{1}\{i=(j \mod d)\}}{2d^2}$, where mod denotes the modulo operation, c_j's are constants independent of d, and $R_j := q_j q_j^\top + \frac{1}{n}\sum_{k=1}^{j-1} q_k q_k^\top$ for any $j \geqslant d$ with $q_j := [1-p_{j1}, \ldots, 1-p_{jd}]^\top$. Then, we sample y_j independently from Bernoulli$(\theta_*^\top p_j)$ whose CDF is $\theta_*^\top \Phi_j$. Given n samples, we calculate $\widehat{\theta}_\lambda$ using different values of λ according to (21) with $S = [0,1]$ and $\mathfrak{m} = \mathsf{Leb}([0,1])$. We evaluate the performance using the un-normalized ℓ^2-error $\|\widehat{\theta}_\lambda - \theta_*\|$, the self-normalized error $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$, and the KS distance $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$ (for KS distance, we consider the family of Bernoulli distributions with parameters in $[1-\frac{1}{d^2}, 1-\frac{1}{2d^2}]$ to align with the setting of p_j in data generation). We repeat the experiments 100 times to calculate means and 90% confidence intervals of the errors.

We first study the dependence of estimation errors of our estimator (2) on sample size n with the dimension $d = 5$. Specifically, for $\lambda = 0.001$, 0.1, and 10, we run the experiments with n ranging from 10^4 to 10^6 and plot the means and 90% confidence intervals of the errors against n (both in logarithmic scale) in Figure 2. According to Figure 2, for different values of λ, the slopes of the curves of $\log\|\widehat{\theta}_\lambda - \theta_*\|$, $\log \mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$, and $\log\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ against $\log n$ are around -0.5, -0.5, and 0.025, which obeys the $\widetilde{\Theta}(\sqrt{d/n})$, $\widetilde{O}(d/\sqrt{n})$ (assuming $\mu_{\min}(U_n)$ grows linearly with n), and $O(\sqrt{d\log(1+n/\lambda)})$ upper bounds on the errors $\|\widehat{\theta}_\lambda - \theta_*\|$, $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$, and $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ respectively according to Theorem 1 and 8.

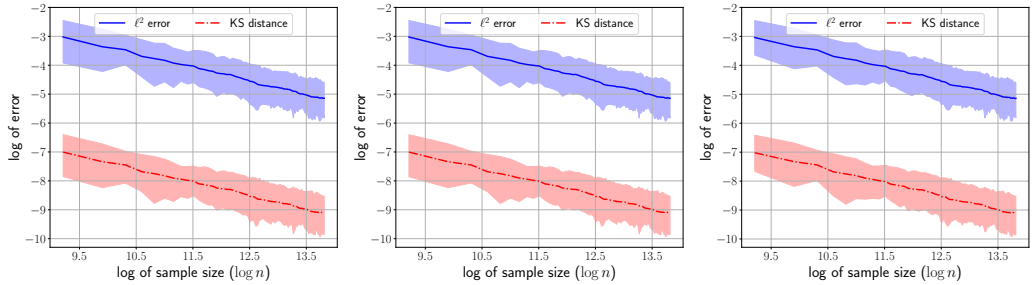
Then, we study the dependence of estimation errors of estimator (2) on dimension d with the sample size $n = 10^6$. For $\lambda = 0.001$, 0.1, and 10, we run the experiments with d ranging from 10 to 100. Then, we plot the means and 90% confidence intervals of $\log\|\widehat{\theta}_\lambda - \theta_*\|$ and $\log \mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$ against $\log d - \log \mu_{\min}(U_n(\lambda))$ as well as $\log\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ against $\log d$ in Figure 3. According to Figure 3, for different values of λ, the slopes of the curves of $\log\|\widehat{\theta}_\lambda - \theta_*\|$ and $\log \mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$ against $\log d - \log \mu_{\min}(U_n(\lambda))$ are around 0.5 and -0.5 respectively, and the slopes of the curves of $\log\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ against $\log d$ are around 0.5. These results also obey the $\widetilde{\Theta}(\sqrt{d/(1+\mu_{\min}(U_n))})$, $\widetilde{O}(d/\sqrt{(1+\mu_{\min}(U_n))})$, and $O(\sqrt{d\log(1+n/\lambda)})$ upper bounds on the errors $\|\widehat{\theta}_\lambda - \theta_*\|$, $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$, and $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ respectively according to Theorem 1 and 8.

Polynomial CDF data experiments. For $d \in \mathbb{N}$, $r(i) := i$ if $1 \leqslant i \leqslant \frac{d+1}{2}$, and $r(i) := \frac{2}{2i-d+1}$ if $\frac{d+1}{2} < i \leqslant d$, we consider the following basis CDFs:

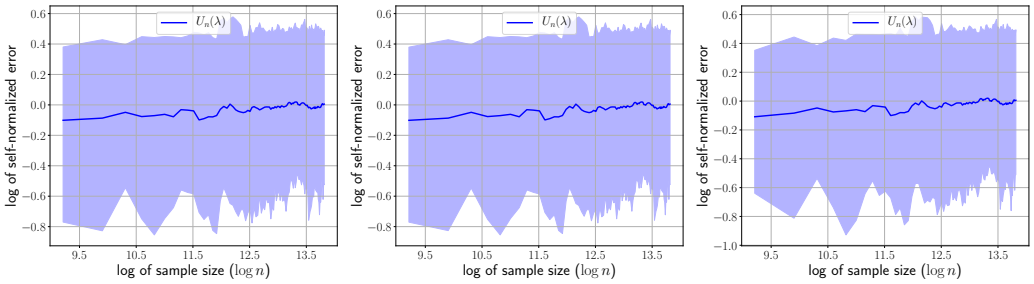
$$\phi_i(x,t) = \mathbb{1}\{t \in [0, 1/x]\}(xt)^{r(i)} + \mathbb{1}\{t > 1/x\}, \ i \in [d]. \tag{25}$$

To simulate n samples, we first choose a true parameter θ_*. For each $j \in [n]$, x_j is sampled independently from the uniform distribution on $[0.5, 2]$. Then, we sample y_j independently from the CDF $\theta_*^\top \Phi(x_j, \cdot)$ using the inverse CDF method for $j \in [n]$. Given the simulated sample, we calculate $\widehat{\theta}_\lambda$ using (3) with $S = [0,2]$, $\mathfrak{m}$ chosen as the uniformly distribution $\mathfrak{m}_{\mathsf{U}}$ on S, and different values of λ. We evaluate the performance by calculating ℓ^2-error $\|\widehat{\theta}_\lambda - \theta_*\|$, the self-normalized errors $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ and $\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n}$, and the KS distance $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$. To obtain stable results, we repeat the simulation independently 100 times in each setting to calculate 90% confidence intervals and means of the errors.

Fixing $d = 5$, we study the dependence of estimation errors of our estimator (2) on sample size n using $\lambda = 0.001$, 0.1, and 10. We run the experiments with n ranging from 10^4 to 10^6 and plot the means and 90% confidence intervals of the errors against n (both in logarithmic scale) in Figure 4. According to Figure



(a) $\lambda = 0.001$ (b) $\lambda = 0.1$ (c) $\lambda = 10$



(d) $\lambda = 0.001$ (e) $\lambda = 0.1$ (f) $\lambda = 10$

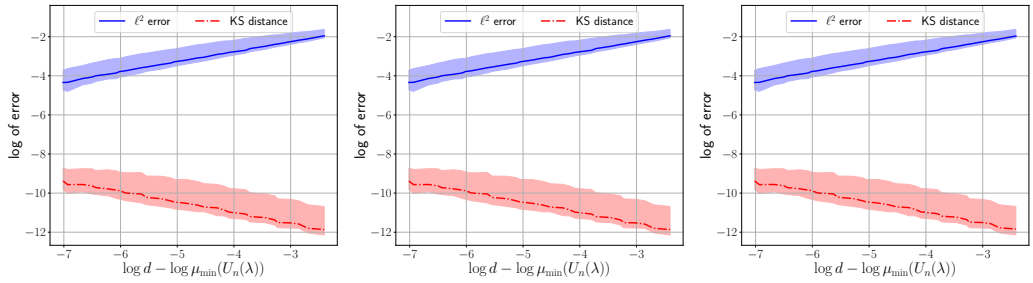
Figure 2: Means and 90% confidence intervals of un-normalized ℓ^2-errors $\|\widehat{\theta}_\lambda - \theta_*\|$, KS distances $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$, and self-normalized errors $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ against sample size n in logarithmic scale in Bernoulli synthetic data experiments.

4, for different values of λ, the slopes of the curves of $\log \|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ and $\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n}$ against $\log n$ are around 0, which obeys the $O(\sqrt{d \log(1 + n/\lambda)})$ upper bounds proved in Theorem 1 and Proposition 5. When λ is negligible compared to $\mu_{\min}(U_n))$, the $\widetilde{O}(\sqrt{d/(\lambda + \mu_{\min}(U_n))})$ bound on ℓ^2-error followed from Theorem 1 implies the $\widetilde{O}(\sqrt{d/n})$ ℓ^2-error bound if $\mu_{\min}(U_n))$ grows linearly with n. Indeed, for small $\lambda = 0.001$, the slope of the curve of $\log \|\widehat{\theta}_\lambda - \theta_*\|$ against $\log n$ in Figure 4a is around -0.5. When λ is comparable with $\mu_{\min}(U_n))$, as is observed in Figure 4b and 4c, the slopes of the curves of ℓ^2-errors are larger than -0.5, which is expected from the $\widetilde{O}(\sqrt{d/(\lambda + \mu_{\min}(U_n))})$ bound. The slopes of the curves of the KS distances against $\log n$ are smaller than 0.5, also obeying the $\widetilde{O}(d/\sqrt{(\lambda + \mu_{\min}(U_n))})$ bound implied by Theorem 1.

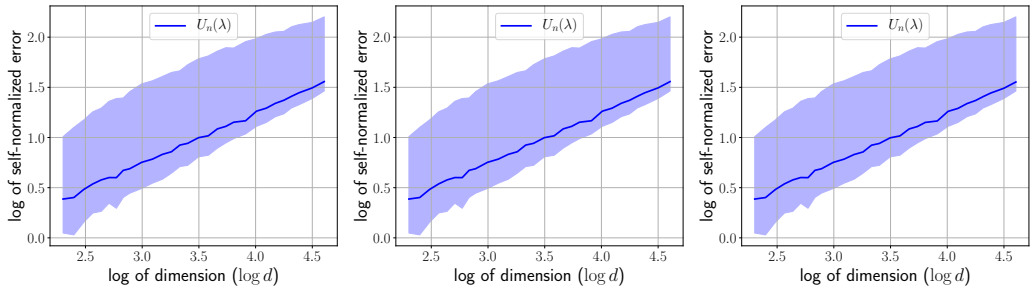
Next, fixing $n = 10^5$, we run the experiments with d ranging from 10 to 100 using $\lambda = 0.001$, 0.1, and 10. We plot the means and 90% confidence intervals of the errors against d (both in logarithmic scale) in Figure 5. According to Figure 5, for different values of λ, the slopes of the curves of $\log \|\widehat{\theta}_\lambda - \theta_*\|$, $\log \|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$, and $\log \|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n}$ against $\log d$ are around 0, obeying the respective $\widetilde{O}(\sqrt{d/(\lambda + \mu_{\min}(U_n))})$, $O(\sqrt{d \log(1 + n/\lambda)})$, and $O(\sqrt{d \log(1 + n/\lambda)})$ bounds proved in Theorem 1 and Proposition 5. The slopes of the curves of the KS distances are smaller than 1, which also obeys the $\widetilde{O}(d/\sqrt{(\lambda + \mu_{\min}(U_n))})$ bound implied by Theorem 1. Since the lower bounds are proved for the worst case of any estimator, the results above do not violate our theoretical results on lower bound.

7.2 Real data experiments

We compare the empirical performance of our estimator (2) and other methods on two real-world datasets: the California house price dataset and the adult income dataset.



(a) $\lambda = 0.001$ (b) $\lambda = 0.1$ (c) $\lambda = 10$



(d) $\lambda = 0.001$ (e) $\lambda = 0.1$ (f) $\lambda = 10$

Figure 3: Means and 90% confidence intervals of un-normalized ℓ^2-errors $\|\widehat{\theta}_\lambda - \theta_*\|$, KS distances $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$, and self-normalized errors $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ against $d/\mu_{\min}(U_n(\lambda))$ and dimension d in logarithmic scale in Bernoulli synthetic data experiments.

California house price dataset. We evaluate the performance of estimator (2) on the California house price dataset (Mohapatra, 2022) of size $n = 20,640$ from Kaggle. There are 10 attributes among which we use median house value as the samples y from target CDFs and all other attributes as the contexts x ($d = 9$). We standardize all the ordinal variables.

We apply the proposed estimator (2) and three other methods, MLE, empirical CDF (ECDF), and kernel density estimation (KDE) to estimate contextual CDFs for this dataset. Specifically, for ECDF, given samples $y^{(1)}, \ldots, y^{(n)}$, the empirical CDF is as follows,

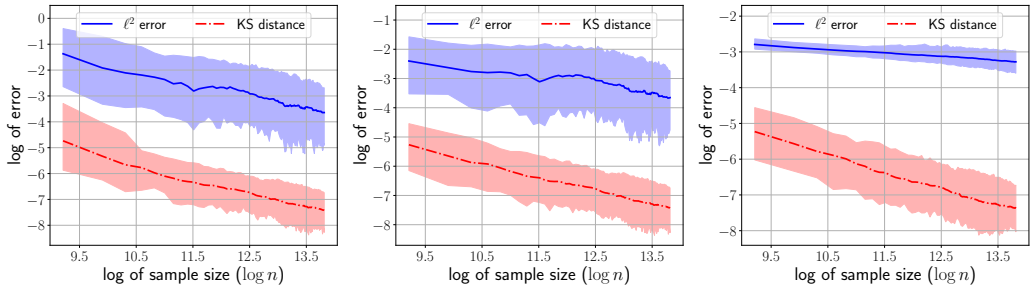
$$\widehat{F}_E(t) := \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}_{y^{(j)}}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{y^{(j)} \leqslant t\}. \tag{26}$$

For KDE, we apply the function "density" in the R package "stats" with Gaussian, rectangular, and triangular kernels. Note that only the samples y are used to estimate one CDF without considering the contexts x in ECDF and KDE. For the proposed estimator and MLE, we assume the linear model (1). We consider the following family of basis CDFs:

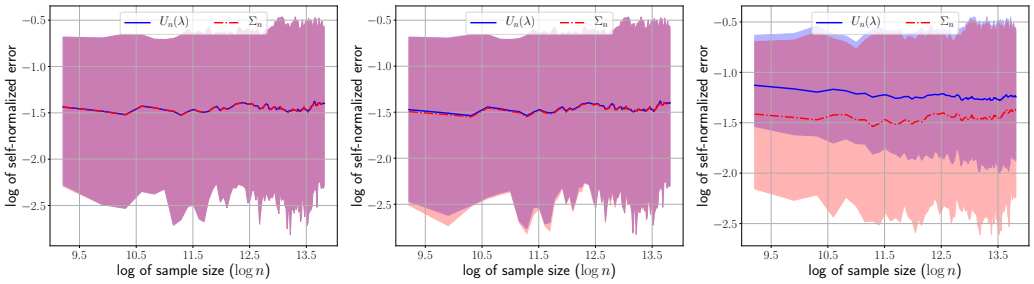
$$\phi_i(x,t) = (1-w)F_N(t; \beta_{N,i}^{(1)} x_i + \beta_{N,i}^{(0)}, \sigma_i^2) + wF_L(t; \beta_{L,i}^{(1)} x_i + \beta_{L,i}^{(0)}, b_i),\ t \in \mathbb{R},\ i \in [d], \tag{27}$$

where $x = (x_1, \ldots, x_d)$ is the context, $F_N(\cdot; \mu, \sigma^2)$ is the CDF of the Gaussian distribution $N(\mu, \sigma^2)$, $F_L(\cdot; \mu, b)$ is the CDF of the Laplace distribution $\mathrm{Laplace}(\mu, b)$, w is the weight of Laplace distributions, $\beta_{N,i}^{(1)}$ ($\beta_{N,i}^{(0)}$) is the coefficient (intercept) in the Gaussian linear model of x_i, and $\beta_{L,i}^{(1)}$ ($\beta_{L,i}^{(0)}$) is the coefficient (intercept) in the Laplace linear model of x_i.

We split the whole dataset into subsets of fractions 1/3, 1/2, and 1/6. 1/3 data points are used to estimate the coefficients and intercepts under Gaussian or Laplace linear models separately by maximizing log likelihood. For Laplace linear model, it corresponds to the least absolute residual regression which we solve using the



(a) $\lambda = 0.001$ (b) $\lambda = 0.1$ (c) $\lambda = 10$



(d) $\lambda = 0.001$ (e) $\lambda = 0.1$ (f) $\lambda = 10$

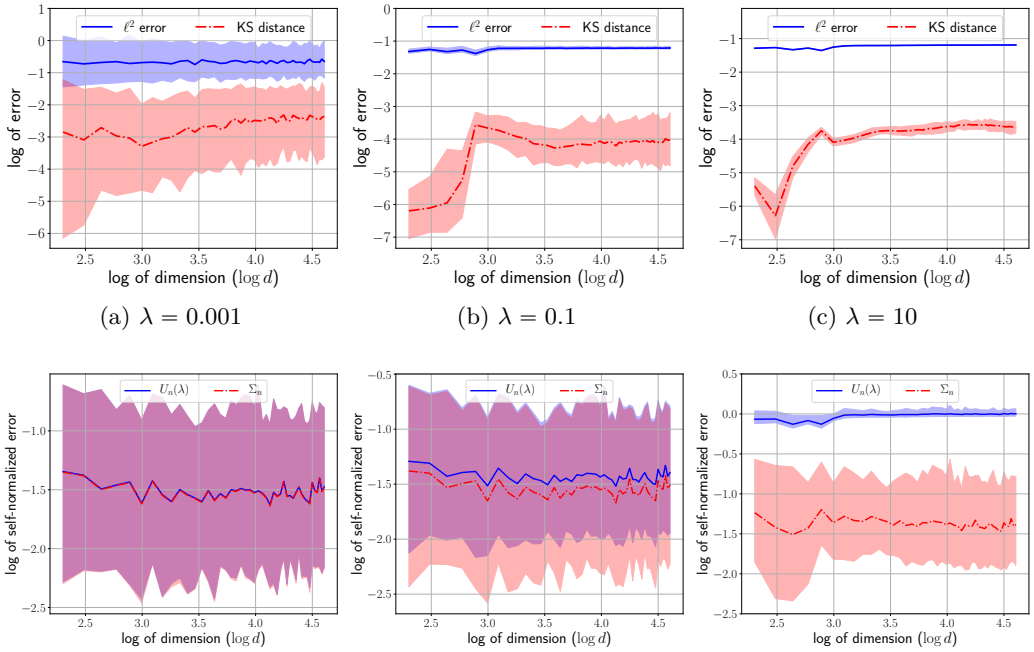
Figure 4: Means and 90% confidence intervals of un-normalized ℓ^2-errors $\|\widehat{\theta}_\lambda - \theta_*\|$, KS distances $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$, and self-normalized errors $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ and $\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n}$ against sample size n in logarithmic scale in polynomial CDF synthetic data experiments.

function "lad" in the R package "L1pack" (Osorio & Wolodzko, 2023). Afterwards, we estimate σ_i^2's and b_i's using the sample variance and the mean absolute deviation of the their corresponding residuals respectively. Then, we apply different methods on the second subset (training dataset) of 1/2 data points. For the proposed estimator, we calculate $\widehat{\theta}_\lambda$ using (3) with $S = \mathbb{R}$, $\mathfrak{m} = \gamma_{0,100}$, and $\lambda = 0.1, 1, 5$. For MLE, we can formulate the likelihood function of the parameter θ in (1) with Φ specified in (27). Let $\widehat{\theta}_{MLE}$ denote a maximizer of likelihood function. Since under (1), MLE corresponds to solving a convex minimization problem in a convex set (the probability simplex Δ^{d-1}), we use the solver "SCS" in the R package "CVXR" (Fu et al., 2020) to calculate $\widehat{\theta}_{MLE}$. Let $\widehat{F}_E$ denote the ECDF calculated by (26) using the training dataset. Let $\widehat{F}_{KG}$, $\widehat{F}_{KR}$, and $\widehat{F}_{KT}$ denote the CDF calculated by KDE with Gaussian, rectangular, and triangular kernels respectively using the training dataset. Given samples $y^{(1)}, \ldots, y^{(n)}$, we define the $\mathcal{L}^2$-error of an estimated CDF $\widehat{F}$ as

$$\frac{1}{n}\sum_{j=1}^{n} \|\mathrm{I}_{y^{(j)}} - \widehat{F}\|^2_{\mathcal{L}^2(S,\mathfrak{m})}, \tag{28}$$

where we also set $S = \mathbb{R}$ and $\mathfrak{m} = \gamma_{0,100}$. Note that when $S = \mathbb{R}$ and $\mathfrak{m} = \mathsf{Leb}(\mathbb{R})$, the $\mathcal{L}^2$-error in (28) corresponds to the renowned Continuous Ranked Probability Score (CRPS) (Hersbach, 2000) used to assess the performance of a CDF in approximating data distribution. We calculate $\mathcal{L}^2$-errors on the third subset (test dataset) of 1/6 data points for the four methods described previously. For ECDF and KDE, we plug $\widehat{F}_E$, $\widehat{F}_{KG}$, $\widehat{F}_{KR}$, and $\widehat{F}_{KT}$ in (28). For MLE and the proposed estimator, we calculate $\mathcal{L}^2$-errors using $\frac{1}{n}\sum_{j=1}^n \|\mathrm{I}_{y^{(j)}} - \widehat{\theta}_{MLE}^\top \Phi_j\|^2_{\mathcal{L}^2(S,\mathfrak{m})}$ and $\frac{1}{n}\sum_{j=1}^n \|\mathrm{I}_{y^{(j)}} - \widehat{\theta}_\lambda^\top \Phi_j\|^2_{\mathcal{L}^2(S,\mathfrak{m})}$ with different values of λ.

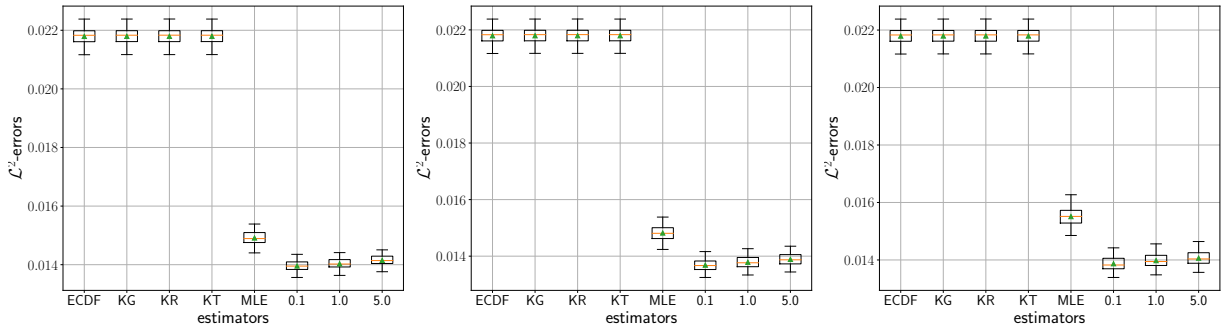
We run the experiments with $w = 0$, 0.5, and 1 in (27). To get stable results, we permute the dataset uniformly at random independently and repeat the experiments 100 times to calculate $\mathcal{L}^2$-errors. We draw the box plots of the $\mathcal{L}^2$-errors of different methods with different values of w in Figure 6. As is shown in the figure, ECDF and KDE have comparable $\mathcal{L}^2$-errors which are much larger than the other two methods. For all choices of w and λ, our estimator (2) achieves the smallest $\mathcal{L}^2$-error than any other method, indicating



(a) $\lambda = 0.001$ (b) $\lambda = 0.1$ (c) $\lambda = 10$

(d) $\lambda = 0.001$ (e) $\lambda = 0.1$ (f) $\lambda = 10$

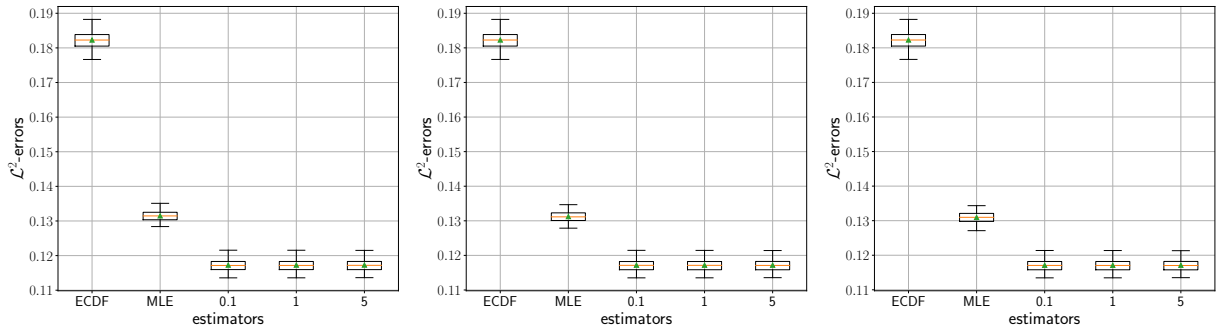
Figure 5: Means and 90% confidence intervals of un-normalized ℓ^2-errors $\|\widehat{\theta}_\lambda - \theta_*\|$, KS distances $\mathrm{KS}(\widehat{F}_\lambda(x,\cdot), F(x,\cdot))$, and self-normalized errors $\|\widehat{\theta}_\lambda - \theta_*\|_{U_n(\lambda)}$ and $\|\widehat{\theta}_\lambda - \theta_*\|_{\Sigma_n}$ against dimension d in logarithmic scale in polynomial CDF synthetic data experiments.



(a) $w = 0$ (b) $w = 0.5$ (c) $w = 1$

Figure 6: Box plots of $\mathcal{L}^2$-errors in the California house price data experiment. "ECDF" refers to the empirical CDF defined in (26). "KG", "KR", and "KT" refer to the kernel density estimation method using Gaussian, rectangular, and triangular kernels respectively. "0.1", "1.0", and "5.0" refer to our estimator $\widehat{\theta}_\lambda$ in (2) with $\lambda = 0.1$, 1.0, and 5.0 respectively.

that its performance is very robust in the choices of basis CDFs and regularization level. Also, $\mathcal{L}^2$-error of our estimator decreases with the value of λ as expected. Thus, with different basis contextual CDFs, our estimator (2) has better performance in approximating target data distributions and the performance is stable wrt the value of λ in (2).



(a) $w = 0$ (b) $w = 0.5$ (c) $w = 1$

Figure 7: Box plots of $\mathcal{L}^2$-errors in adult income data experiments. "ECDF" refers to the empirical CDF defined in (26). "0.1", "1.0", and "5.0" refer to the estimator $\widehat{\theta}_\lambda$ (2) with $\lambda = 0.1$, 1.0, and 5.0 respectively.

Adult income dataset. The adult income dataset (Becker & Kohavi, 1996) was extracted from the 1994 census bureau database. The typical learning task is to predict whether income exceeds \$50K/yr based on other attributes in the census data. Thus, we use the attributes age, workclass, education, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country as the contexts x ($d = 12$), and use income (i.e., whether income exceeds \$50K/yr) as the samples y from target CDFs. We standardize all of the ordinal attributes. The total number of samples is $n = 48,842$.

Since the samples follow Bernoulli distributions, KDE is not considered. We apply our estimator (2), MLE, and ECDF on this dataset. For our estimator and MLE, we assume model (1) and use the following mixtures of logistic and probit models as basis CDFs:

$$\phi_i(x,t) = wF_B(t; f_{\text{logi}}(\beta_{L,i}^{(1)}x_i + \beta_{L,i}^{(0)})) + (1-w)F_B(t; F_N(\beta_{P,i}^{(1)}x_i + \beta_{P,i}^{(0)}; 0, 1)),\ i \in [d], \tag{29}$$

where $t \in \mathbb{R}$, $x = (x_1, \ldots, x_d)$ denotes the context, $F_B(\cdot; p)$ denotes the CDF of the Bernoulli distribution with parameter p, $f_{\text{logi}}(a) := 1/(1+e^{-a})$ for any $a \in \mathbb{R}$, w is the weight of the logistic model, $\beta_{L,i}^{(1)}$ ($\beta_{L,i}^{(0)}$) denotes the coefficient (intercept) in the logistic model of x_i, and $\beta_{P,i}^{(1)}$ ($\beta_{P,i}^{(0)}$) denotes the coefficient (intercept) in the probit model of x_i. We split the whole dataset into subsets of fractions 1/3, 1/2, and 1/6. 1/3 data points are used to estimate the coefficients and intercepts in (29) with the function "glm" in the R package "stats".

We apply all methods on the second subset (training dataset) of 1/2 data points. For our estimator, we calculate $\widehat{\theta}_\lambda$ using (3) with $S = [0,1]$, $\mathfrak{m} = \mathsf{Leb}([0,1])$, and $\lambda = 0.1, 1, 5$. For MLE, we also use the solver "SCS" in the R package "CVXR" (Fu et al., 2020) to calculate $\widehat{\theta}_{MLE}$ as in the previous example. We use $\widehat{F}_E$ to denote the ECDF calculated by (26) using the training dataset. Then, we calculate $\mathcal{L}^2$-errors (28) with $S = [0,1]$ and $\mathfrak{m} = \mathsf{Leb}([0,1])$ for the three methods on the third subset (test dataset) of 1/6 data points.

We run the experiments described above using $w = 0$, 0.5, and 1 in (29) 100 times with the dataset permuted randomly in each run to get stable results. In Figure 7, we report the calculated $\mathcal{L}^2$-errors in box plots. According to the figure, our estimator (2) achieves the smallest $\mathcal{L}^2$-errors for all choices of λ and weight w and ECDF has the largest $\mathcal{L}^2$-error for all choices of w. Moreover, the performance of our estimator is very robust wrt λ and w. Thus, with a wide range of the basis contextual CDFs, our estimator (2) achieves good and robust performance in approximating target data distributions.

8 Conclusion

In this paper, we propose a linear model for contextual CDFs and estimators for the coefficient parameter in this model. We prove $\widetilde{O}(\sqrt{d/n})$ upper bounds on the estimation error of our estimator under the adversarial and random settings, and show that the upper bounds are tight up to logarithmic factors by proving $\Omega(\sqrt{d/n})$

information theoretic lower bounds. Additionally, when a mismatch exists in the linear model, we prove that the estimation error of our estimator only increases by an amount commensurate with the mismatch error. Furthermore, we increase the generality of our linear model by expanding the parameter space into an infinite dimensional Hilbert space. Within this framework, we generalize our estimator and subsequently establish self-normalized upper bounds for this general estimator. Moreover, we elucidate the scaling of the estimation error of our estimator empirically and showcase its practical utility on real-world datasets. Our current work assumes that the bases are known a priori. So, a fruitful future research direction would be to focus on the basis selection problem for CDF regression with possibly infinitely many base functions.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011b.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Kamyar Azizzadenesheli. Importance weight estimation and generalization in domain adaptation under label shift, 2020. URL https://arxiv.org/abs/2011.14251.

Necdet Batir. Inequalities for the gamma function. *Archiv der Mathematik*, 91(6):554–563, Dec 2008. ISSN 1420-8938. doi: 10.1007/s00013-008-2856-9. URL https://doi.org/10.1007/s00013-008-2856-9.

Amir Beck. *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. SIAM, 2014.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

Vladimir Bogachev. *Measure Theory*, volume 2. 01 2007. ISBN 978-3-540-34513-8. doi: 10.1007/978-3-540-34514-5.

Francesco Paolo Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4 (421-424), 1933.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general framework for bandit problems beyond cumulative objectives. *Mathematics of Operations Research*, 2023.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Yeonseung Chung and David B Dunson. Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660, 2009.

DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.5 of 2022-03-15. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

R.M. Fano. *Transmission of Information: A Statistical Theory of Communication*. MIT Press Classics. MIT Press, 1961. ISBN 9780262561693.

Frédéric Ferraty, Ali Laksaci, and Philippe Vieu. Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes*, 9:47–76, 2006.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020. doi: 10.18637/jss.v094.i14.

Valery Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4:92–99, 1933.

Peter Hall, Rodney CL Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, 94:154–163, 1999.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559 – 570, 2000. doi: https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0559_dotcrp_2_0_co_2.xml.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(none):1 – 6, 2012a. doi: 10.1214/ECP.v17-2079. URL https://doi.org/10.1214/ECP.v17-2079.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012b.

Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

Audrey Huang, Liu Leqi, Zachary C Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment for markov decision processes. In *Artificial Intelligence and Statistics*, 2022.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *arXiv preprint arXiv:1912.12945*, 2019.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Roger Koenker, Samantha Leorato, and Franco Peracchi. Distributional vs. quantile regression. 2013.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Liu Leqi, Audrey Huang, Zachary Lipton, and Kamyar Azizzadenesheli. Supervised learning with general risk functionals. In *International Conference on Machine Learning*, pp. 12570–12592. PMLR, 2022.

Mikhail Lifshits. Lectures on gaussian processes. In *Lectures on Gaussian Processes*, pp. 1–117. Springer, 2012.

Anuran Makur. *Information Contraction and Decomposition*. Sc.D. thesis in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, May 2019.

Anuran Makur and Lizhong Zheng. Comparison of contraction coefficients for f-divergences. *Problems of Information Transmission*, 56(2):103–156, April 2020.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Shibu Mohapatra. California House Price, 2022. Retrieved August 2023 from https://www.kaggle.com/datasets/shibumohapatra/house-price.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

F. Osorio and T. Wolodzko. *Routines for L1 estimation*, 2023. URL http://l1pack.mat.utfsm.cl. R package version 0.41-24.

Bernardo Ávila Pires and Csaba Szepesvári. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In *artificial intelligence and statistics*. PMLR, 2013.

LA Prashanth, Cheng Jie, Michael Fu, Steve Marcus, and Csaba Szepesvári. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, pp. 1406–1415. PMLR, 2016.

Michael Reed and Barry Simon. Vi - bounded operators. In *I: Functional Analysis*, pp. 182–220. Elsevier Inc, 1972. ISBN 9780125850018.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42, 2000.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

P. Stein. A note on the volume of a simplex. *The American Mathematical Monthly*, 73(3):299–301, 1966. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2315353.

Francis Edward Su. *Methods for quantifying rates of convergence for random walks on groups*. Harvard University, 1995.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(45):1231–1264, 2006.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Julia L Wirch and Mary R Hardy. Distortion risk measures: Coherence and stochastic dominance. In *International congress on insurance: Mathematics and economics*, 2001.

William Wong, Audrey Huang, Liu Leqi, Kamyar Azizzadenesheli, and Zachary C Lipton. Riskyzoo: A library for risk-sensitive supervised learning. 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.