Sparse Contextual CDF Regression

Anonymous authors
Paper under double-blind review

Abstract

Estimating cumulative distribution functions (CDFs) of context-dependent random variables is a central statistical task underpinning numerous applications in machine learning and economics. In this work, we extend a recent line of theoretical inquiry into this domain by analyzing the problem of *sparse contextual CDF regression*, wherein data points are sampled from a convex combination of s context-dependent CDFs chosen from a set of d basis functions. We show that adaptations of several canonical regression methods serve as tractable estimators in this functional sparse regression setting under standard assumptions on the conditioning of the basis functions. In particular, given n data samples, we prove estimation error upper bounds of $\tilde{O}(\sqrt{s/n})$ for functional versions of the lasso and Dantzig selector estimators, and $\tilde{O}(\sqrt{s}/\sqrt[4]{n})$ for a functional version of the elastic net estimator. Our results match the corresponding error bounds for finite dimensional regression and improve upon CDF ridge regression which has $\tilde{O}(\sqrt{d/n})$ estimation error. Finally, we obtain a matching information-theoretic lower bound which establishes the minimax optimality of the lasso and Dantzig selector estimators up to logarithmic factors.

1 Introduction

The estimation of cumulative distribution functions (CDFs) is a classical problem in mathematical statistics stemming back to the Glivenko-Cantelli theorem (Cantelli, 1933; Glivenko, 1933; Devroye et al., 2013), which states that empirical CDFs constructed from independent samples of a single random variable converge uniformly to the random variable's true CDF. Subsequent classical research in this area has focused on deriving tight non-asymptotic sample complexity results in terms of the Kolmogorov-Smirnov distance among others, such as the Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky et al., 1956) and improved bounds by Massart (1990).

Motivated by applications to modern learning tasks such as contextual bandits and Markov decision processes (Huang et al., 2021; 2022), a recent line of research (Zhang et al., 2024) introduced the problem of *context-dependent* CDF estimation, which requires a learner to simultaneously estimate a (possibly infinite) family of CDFs parameterized by some context variable. As an initial simplification, the authors considered the restricted setting of *contextual CDF regression*, wherein the true contextual CDF is a convex combination of d context-dependent basis functions. The authors generalized the classical ridge regression method (Hoerl & Kennard, 1970; Abbasi-Yadkori et al., 2011) to this functional regression problem and derived a tight $\tilde{O}(\sqrt{d/n})$ estimation error bound given n samples in a variety of data generation settings. However, when d is large, specifically in unstructured settings where a massive set of potential CDFs are considered, ridge regression utilizes all CDF basis functions for the purpose of estimation without regard to their relevance.

In this paper, as a further step towards developing general algorithms for contextual CDF estimation, we propose *sparse regression* and *basis selection* techniques for the aforementioned CDF regression problem based on functional versions of lasso (Tibshirani, 1996), elastic net (Zou & Hastie, 2005), and Dantzig selector (Candes & Tao, 2007) methods. Crucially, all of our techniques achieve estimation bounds with no polynomial dependence on d, allowing accurate recovery of the true contextual CDF from bases containing exponentially many irrelevant functions. We also establish minimax optimality results for sparse CDF regression.

1.1 Outline

We briefly delineate the structure of our paper. We introduce the sparse contextual CDF regression problem under present investigation in Section 1.2 and define relevant notation in Appendix A. In Section 2, we outline some applications of CDFs in machine learning and economics, and provide an overview of related functional regression schemes in the previous scientific literature. In Section 3, we state our main contributions and discuss the overarching proof techniques used throughout the paper. Section 4 provides the formal derivation of our upper bound on lasso estimation error in the fixed design setting. We include numerical simulations of our data generation and parameter estimation processes in Section 5. We defer remaining proofs and technical details to Appendices B to E, provide examples of CDF bases which satisfy the preconditions for our main results in Appendix F, and briefly discuss the computational complexity of our estimators in Appendix G.

1.2 Model and Setup

In this subsection, we formally define the sparse contextual CDF regression problem under investigation in this paper. (Recall that we utilize the notation defined in Appendix A.) Let $\mathcal{X}$ denote a general context space. We refer to a function $f(x,t) : \mathcal{X} \times \mathbb{R} \to [0,1]$ as a *contextual CDF* if for any $x \in \mathcal{X}$, $f(x,\cdot)$ is a valid CDF for a real-valued random variable with range contained in some set $S \subseteq \mathbb{R}$. Let $\mathfrak{m}$ be any probability measure on S. Let $\{\phi_1, \dots, \phi_d\}$ denote a fixed basis of $d \in \mathbb{N}$ contextual CDFs indexed by $i \in [d]$. For convenience, we often conceptualize this basis as a single vector-valued function $\Phi : \mathcal{X} \times \mathbb{R} \to [0,1]^d$ defined by $[\Phi(x,t)]_i = \phi_i(x,t)$ for all $i \in [d]$.

Let $F : \mathcal{X} \times \mathbb{R} \to [0,1]$ denote the true contextual CDF we aim to recover. Following the precedent established in Zhang et al. (2024), we assume that F is a convex combination of the basis functions $\{\phi_1, \dots, \phi_d\}$. Hence, the problem reduces to recovering the true parameter vector $\theta_* \in \Delta^{d-1}$ such that

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

Let $S_* = \operatorname{supp}(\theta_*)$ denote the support of the true parameter. For notational simplicity, let $s = |S_*| = \|\theta_*\|_0$. Let $\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$ be a set of $n \in \mathbb{N}$ observed samples indexed by $j \in [n]$, generated according to the *fixed design* or *random design* settings described below:

- **Fixed design**. For each $j \in [n]$, the context variable $x^{(j)} \in \mathcal{X}$ is fixed a priori, and the response variable $y^{(j)} \in \mathbb{R}$ is independently sampled from the CDF $F(x^{(j)}, \cdot)$.
- **Random design**. For each $j \in [n]$, the context variable $x^{(j)} \in \mathcal{X}$ is independently sampled from an unknown probability distribution $P_X^{(j)}$ over $\mathcal{X}$. Then, conditional on $x^{(j)}$, the response variable $y^{(j)} \in \mathbb{R}$ is independently sampled from the CDF $F(x^{(j)}, \cdot)$.

These design settings extend the data generation processes described in Abbasi-Yadkori et al. (2011) and Hsu et al. (2012). Let $\Phi_j(t) = \Phi(x^{(j)}, t)$ denote the CDF basis at the jth context variable, and let $x^{1:n} = (x^{(1)}, \dots, x^{(n)})$ denote a collection of sampled variables, with analogous definitions for $y^{1:n}$ and $(x,y)^{1:n}$. Let $U_n \in \mathbb{R}^{d \times d}$ denote the *empirical n-sample Gramian matrix* given by

$$\forall i, i' \in [d],\;\; [U_n]_{i,i'} = \frac{1}{n}\sum_{j=1}^{n} \left\langle [\Phi_j]_i \,, [\Phi_j]_{i'} \right\rangle \qquad\qquad \therefore U_n = \frac{1}{n}\sum_{j=1}^{n} \int_S \Phi_j \Phi_j^\top \, d\mathfrak{m}\,.$$

In the random design setting, let $\Sigma_n \in \mathbb{R}^{d \times d}$ denote the *expected n-sample Gramian matrix* given by

$$\forall i, i' \in [d],\;\; [\Sigma_n]_{i,i'} = \frac{1}{n}\sum_{j=1}^{n} \mathop{\mathbb{E}}_{X^{(j)} \sim P_X^{(j)}} \left[\left\langle [\Phi_j]_i \,, [\Phi_j]_{i'} \right\rangle \right] \qquad\qquad \therefore \Sigma_n = \mathop{\mathbb{E}}_{X^{1:n}} \left[\frac{1}{n}\sum_{j=1}^{n} \int_S \Phi_j \Phi_j^\top \, d\mathfrak{m} \right].$$

As mentioned previously, the objective of contextual CDF regression is to recover θ_* given the samples $(x,y)^{1:n}$. Each $Y^{(j)}$ defines a one-sample empirical CDF $\mathrm{I}_{Y^{(j)}}(t) = \mathbb{1}\{t \geq Y^{(j)}\}$ which approximates the true CDF $F(x^{(j)}, \cdot) = \mathbb{E}[\mathrm{I}_{Y^{(j)}}(\cdot) \mid X^{(j)} = x^{(j)}]$ in expectation conditioned on $X^{(j)} = x^{(j)}$. We investigate three estimators for θ_* based on the paradigm of empirical risk minimization. Firstly, the *lasso estimator* $\hat{\theta}_\lambda$ imposes an ℓ^1-penalty on the parameter vector, weighted by $\lambda > 0$:

$$\hat{\theta}_\lambda = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{j=1}^{n} \left\| \mathrm{I}_{y^{(j)}} - \theta^\top \Phi_j \right\|^2_{\mathcal{L}^2(S,\mathfrak{m})} + \lambda \left\| \theta \right\|_1 \right\} . \tag{1}$$

Secondly, the *elastic net estimator* $\hat{\theta}_{\lambda_1,\lambda_2}$ imposes both ℓ^1- and ℓ^2-penalties on the parameter vector, weighted by $\lambda_1 > 0$ and $\lambda_2 > 0$ respectively:

$$\hat{\theta}_{\lambda_1,\lambda_2} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{j=1}^{n} \left\| \mathrm{I}_{y^{(j)}} - \theta^\top \Phi_j \right\|^2_{\mathcal{L}^2(S,\mathfrak{m})} + \lambda_1 \left\| \theta \right\|_1 + \lambda_2 \left\| \theta \right\|^2 \right\} . \tag{2}$$

Thirdly, the *Dantzig selector* $\bar{\theta}_\lambda$ is given by the following variation on the constrained optimization formulation of lasso estimation:

$$\bar{\theta}_\lambda = \underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \left\| \theta \right\|_1 \tag{3}$$

$$\text{subject to} \quad \left\| \frac{1}{n} \sum_{j=1}^{n} \left\langle \mathrm{I}_{y^{(j)}} - \theta^\top \Phi_j, \Phi_j \right\rangle \right\|_\infty \leq \lambda \, . \tag{4}$$

For all three estimators, we analyze the *estimation error* given by the Euclidean distance $\|\hat{\theta} - \theta_*\|_A$ between the estimated and true parameter vectors, possibly weighted by some matrix A. We remark that Equations (1) to (3) define *improper estimators* which do not necessarily lie in Δ^{d-1} and are thus not guaranteed to define valid CDFs. However, since $\theta_* \in \Delta^{d-1}$ and Δ^{d-1} is closed and convex, any upper bound on the estimation error of an improper estimator also holds for its projection onto Δ^{d-1}, by Beck (2014, Equation 9.10).

Lastly, we remark that our results hold for any $s \leq d$, although we are principally interested in the advantages of our proposed methods over the ridge estimator baseline (Zhang et al., 2024, Section 3.1) for small values of s. We emphasize that our results have no dependence on the cardinality of the context space $\mathcal{X}$.

2 Related Work

Broadly speaking, CDFs underpin the computation of risk functionals which inform decision-making in mathematical finance and actuarial science. For example, generic law invariant risk functions such as conditional value-at-risk (Rockafellar et al., 2000; Artzner et al., 1999) are parameterized by CDFs, and a notion of distortion risk measure (Wirch & Hardy, 2001) arises from the composition of a CDF with a distortion function. Coherent risk measures are instrumental in portfolio management and optimization (Krokhmal, 2007) and can be formulated in terms of CDFs (Shapiro et al., 2014). Lastly, spectral risk measures are computed as weighted averages of outcomes (Acerbi, 2002) and hence depend on the entire trajectory of the underlying random variable's CDF.

Similar considerations exist in further scopes of learning theory, where the aforementioned risk functionals are incorporated into supervised learning tasks (Liu et al., 2022) and multi-armed bandit problems (Cassel et al., 2023) to model fairness, risk aversion, and distribution shift (Wong et al., 2022). The recently proposed off-policy risk assessment framework (Huang et al., 2021) includes CDF estimation as a key building block and has been applied to contextual bandit problems and Markov decision processes (Huang et al., 2022). Other work in this vein has employed the mean-variance (Sani et al., 2013; Zimin et al., 2014) and value-at-risk (Vakili & Zhao, 2015) paradigms in the multi-armed bandit setting to analyze risk-reward trade-offs. Furthermore, cumulative prospect theory is intimately tied to risk distortion and the ensuing dependence on CDF estimation (Prashanth et al., 2016), and has found relevance in reinforcement learning and stochastic optimization (Jie et al., 2018).

We touch on the well-established precedent in the scientific community for framing function estimation through the lens of linear regression and basis selection. Romo et al. (2013) generalized lasso variable selection to linear models with scalar regressors and functional responses to develop interpretable analyses of car accident data. A related method exploiting feature sparsity and output function smoothness was proposed in Barber et al. (2017) to perform genome-wide association studies. Linear models with functional regressors and scalar responses have been used in genomics, MRI data analysis, and chemometrics, spurring the development of functional group-lasso (Pannu & Billor, 2017) and wavelet-based lasso methods (Zhao et al., 2012; 2015) for such models. Prior studies on genetic regulatory networks (Hong & Lian, 2011) have incorporated function-on-function regression models with ℓ^1-regularization to encode sparsity in pairwise interactions between genes. Lastly, recent developments on lasso estimation for function-on-function regression (Centofanti et al., 2022; Maranzano et al., 2023) find downstream application in geostatistical models and mortality data analysis.

Compared to the prior literature, the main novelty of our contributions lies in the simultaneous estimation of an entire family of CDFs. In this sense, our work may be interpreted as a generalization of the canonical mixture model with known basis distributions (Murphy, 2012, Section 11.2) to the context-dependent setting. For maximum generality, we focus on CDF estimation instead of PDFs or quantiles which require restrictions for existence and well-definedness. We formulate estimation through the lens of functional linear regression and show that our task reduces to integral computation and finite dimensional optimization, making our approach computationally efficient using existing numerical methods. Lastly, our derivations utilize intrinsic properties of CDFs instead of performing discretization to avoid introducing approximation error, in contrast to the approach taken in Hong & Lian (2011) among others.

3 Main Results

3.1 Lasso

Before presenting our main contributions, we introduce the following condition on symmetric matrices which is used in the statements of our main results:

Definition 1 (Restricted eigenvalue condition). *Fix $d \in \mathbb{N}$, $\kappa \geq 0$, $\gamma \geq 0$, and $S_* \subseteq [d]$. Let*

$$C_\gamma(S_*) = \left\{ v \in \mathbb{R}^d : \left\| [v]_{S_*^c} \right\|_1 \leq \gamma \left\| [v]_{S_*} \right\|_1 \right\} .$$

A symmetric matrix $A \in \mathbb{R}^{d \times d}$ satisfies the (κ, γ)-restricted eigenvalue condition over S_* iff

$$\forall v \in C_\gamma(S_*),\ v^\top A v \geq \kappa \left\| v \right\|^2 .$$

Intuitively, the restricted eigenvalue condition specializes the notions of positive definiteness and strong convexity to the subset of directions whose support is close to S_*. The significance of this condition lies in its application to the Gramian matrices U_n and Σ_n, since lower-bounding the Gramian eigenvalues is sufficient to ensure the well-conditioning of the basis functions comprising the true CDF. Our formulation of the restricted eigenvalue condition extends the standard definition from Wainwright (2019, Definition 7.12) to the general inner product spaces used in the definitions of U_n and Σ_n. We provide examples of non-trivial CDF bases which satisfy the restricted eigenvalue condition in Appendix F.

Under this assumption, we state our first main result, a high-probability upper bound on the error of the lasso estimator in the fixed design setting:

Theorem 1 (Lasso fixed design upper bound). *Fix $\delta \in (0, 1)$ and $\kappa > 0$. Assume the samples $(x, y)^{1:n}$ are generated according to the fixed design setting. Assume U_n satisfies the $(\kappa, 3)$-restricted eigenvalue condition over S_*. Let $\hat{\theta}_\lambda$ be the lasso estimator (1) with regularization hyperparameter $\lambda = 4\sqrt{(2/n)\log(2d/\delta)}$. Then with probability at least $1 - \delta$, the estimation error satisfies the bounds*

$$\left\| \hat{\theta}_\lambda - \theta_* \right\|_{U_n} \leq 6 \sqrt{\frac{2s}{\kappa n} \log\left(\frac{2d}{\delta}\right)} \quad \textit{and} \quad \left\| \hat{\theta}_\lambda - \theta_* \right\| \leq \frac{6}{\kappa} \sqrt{\frac{2s}{n} \log\left(\frac{2d}{\delta}\right)} .$$

We provide the technical details of our proof in Section 4. The crux of our argument lies in the definition of $\hat{\theta}_\lambda$ as the minimizer of the empirical risk objective in (1). Substituting $\hat{\theta}_\lambda$ and θ_* into this objective results in an inequality (11) in terms of $\hat{\theta}_\lambda$ and θ_*. Rearranging yields an upper-bound (13) on the estimation error $\|\Delta\|_{U_n}$ in terms of two quantities:

- The scalar projections of the sampling errors $\mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j$ onto the d basis functions Φ_j, which we upper-bound with high probability (8) using various concentration inequalities.
- The difference between the ℓ^1-regularizers $\|\theta_*\|_1 - \|\hat{\theta}_\lambda\|_1$, which we upper-bound by utilizing the sparsity of θ_* and the definition of ℓ^1-norm, among other techniques.

Combining these results gives rise to a bound on the relative magnitudes of the components of Δ corresponding to the true support S_* and its complement (16). Then, upper-bounding the component $\|[\Delta]_{S_*}\|_1$ is sufficient (17) to characterize the overall estimation error $\|\Delta\|_{U_n}$. Applying various norm equivalences and invoking the restricted eigenvalue condition on U_n completes the proof.

Theorem 1 establishes the asymptotic complexity of lasso CDF regression as $O(\sqrt{s\log(d)/n}) = \tilde{O}(\sqrt{s/n})$, analogously to the $\tilde{O}(\sqrt{d/n})$ result for ridge regression in Zhang et al. (2024). The dependence on $\sqrt{s}$ arises from our sparsity analysis on $\|\theta_*\|_1 - \|\hat{\theta}_\lambda\|_1$ and the upper-bound on $\|[\Delta]_{S_*}\|_1$ as discussed above. The factors of $1/\sqrt{n}$ and $\sqrt{\log(d)}$ arise from the probabilistic arguments in our analysis of the sampling errors. Ultimately, the lasso estimator's characteristic ℓ^1-penalty term plays an instrumental role in deriving a result with sub-polynomial dependence on d.

Our second main result is a high-probability upper bound on the error of the lasso estimator in the random design setting, under similar assumptions on the well-conditioning of the CDF basis:

Theorem 2 (Lasso random design upper bound)**.** *Fix $\delta_1 \in (0,1)$, $\delta_2 \in (0, 1-\delta_1)$, and $\kappa > 0$. Assume the samples $(x,y)^{1:n}$ are generated according to the random design setting. Assume that for any $n \in \mathbb{N}$, the matrix Σ_n satisfies the $(\kappa, 3)$-restricted eigenvalue condition over S_*. Assume the sample size is at least $n \geq (32d^2/\kappa^2)\log(d/\delta_1)$. Let $\hat{\theta}_\lambda$ be the lasso estimator (1) with regularization hyperparameter $\lambda = 4\sqrt{(2/n)\log(2d/\delta_2)}$. Then with probability at least $1-\delta_1-\delta_2$, the estimation error satisfies the bounds*

$$\left\|\hat{\theta}_\lambda - \theta_*\right\|_{U_n} \leq 12\sqrt{\frac{s}{\kappa n}\log\left(\frac{2d}{\delta_2}\right)} \quad \text{and} \quad \left\|\hat{\theta}_\lambda - \theta_*\right\| \leq \frac{12}{\kappa}\sqrt{\frac{2s}{n}\log\left(\frac{2d}{\delta_2}\right)}.$$

We defer the technical details of our proof to Appendix B. In a nutshell, since U_n is a sum of n independent matrices and is an unbiased estimator of Σ_n, we can employ a matrix analogue of Hoeffding's inequality (18) to justify approximating Σ_n with U_n given sufficiently many samples n. Subsequently, we analyze the estimation error weighted by U_n to complete the proof.

3.2 Elastic Net

Our third main result is a high-probability upper bound on the error of the elastic net estimator in the fixed design setting:

Theorem 3 (Elastic net fixed design upper bound)**.** *Fix $\delta \in (0,1)$. Assume the samples $(x,y)^{1:n}$ are generated according to the fixed design setting. Assume U_n satisfies the $(\kappa, 3+4\lambda_2/\lambda_1)$-restricted eigenvalue condition over S_*. Let $\hat{\theta}_{\lambda_1,\lambda_2}$ be the elastic net estimator (2) with ℓ^1-regularization hyperparameter $\lambda_1 = 4\sqrt{(2/n)\log(2d/\delta)}$. Then with probability at least $1-\delta$, the estimation error satisfies the bounds*

$$\left\|\hat{\theta}_{\lambda_1,\lambda_2} - \theta_*\right\|_{U_n+\lambda_2 I_d} \leq \left(6\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta}\right)} + 2\lambda_2\right)\sqrt{\frac{s}{\kappa+\lambda_2}}, \tag{5}$$

$$\left\|\hat{\theta}_{\lambda_1,\lambda_2} - \theta_*\right\| \leq \left(6\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta}\right)} + 2\lambda_2\right)\frac{\sqrt{s}}{\kappa+\lambda_2}.$$

Furthermore, if $\kappa < (3/2)\sqrt{(2/n)\log(2d/\delta)}$ and $\lambda_2 = 3\sqrt{(2/n)\log(2d/\delta)} - 2\kappa$, the bound (5) implies

$$\left\|\hat{\theta}_{\lambda_1,\lambda_2} - \theta_*\right\|_{U_n+\lambda_2 I_d} \le 4\sqrt{s}\left(3\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta}\right)} - \kappa\right)^{\frac{1}{2}} \tag{6}$$
$$\le 4\sqrt[4]{\frac{18s^2}{n}\log\left(\frac{2d}{\delta}\right)}.$$

We emphasize that Theorem 3 produces non-trivial estimation bounds even when no assumptions are placed on U_n (i.e., $\kappa = 0$), in which case the bound in (5) has $\tilde{O}(\sqrt{s}/\sqrt[4]{n})$ complexity. However, when $\kappa > 0$ and the ℓ^2-regularization hyperparameter is sufficiently small with $\lambda_2 = O(1/\sqrt{n})$, the elastic net estimation bound exhibits $\tilde{O}(\sqrt{s/n})$ scaling, comparable with the result for the lasso estimator in Theorem 1.

At a high level, the elastic net estimator with its characteristic ℓ^2-penalty term may be perceived as "regularizing" the empirical Gramian matrix to produce $U_n + \lambda_2 I_d$, whose smallest eigenvalue is strictly positive. Additional conditions on U_n further tighten the estimation bounds by bolstering the minimum eigenvalue of $U_n + \lambda_2 I_d$. Hence, the elastic net estimator simultaneously selects features and regularizes the problem in an integrated fashion by ensuring strong convexity of the objective function (2).

We defer the technical details of our proof to Appendix C. We carry out an analysis of the sampling errors and upper-bound the difference between the ℓ^1-regularizers $\|\theta_*\|_1 - \|\hat{\theta}_{\lambda_1,\lambda_2}\|_1$, with additional accommodations for the elastic net estimator's ℓ^2-penalty term. Notably, substituting $\hat{\theta}_{\lambda_1,\lambda_2}$ and θ_* into the empirical risk objective (2) and rearranging yields an upper-bound on $\|\Delta\|_{U_n+\lambda_2 I_d}$ instead of $\|\Delta\|_{U_n}$. The resulting expression (25) contains an additional quantity $\lambda_2 \Delta^\top \theta_*$, which we upper-bound in terms of the error component $\|[\Delta]_{S_*}\|_1$ in (26). Combining this contribution with the terms resulting from the ℓ^1-regularizers gives rise to the parenthesized sum in (5), and choosing the optimal value of λ_2 to balance the summands achieves the bound in (6).

By the same technique based on matrix Hoeffding introduced in Section 3.1, we also obtain the following high-probability upper bound on the error of the elastic net estimator in the random design setting, which we prove in Appendix C:

Theorem 4 (Elastic net random design upper bound)**.** *Fix $\delta_1 \in (0,1)$, $\delta_2 \in (0, 1-\delta_1)$, and $\kappa > 0$. Assume the samples $(x,y)^{1:n}$ are generated according to the random design setting. Assume that for any $n \in \mathbb{N}$, the matrix Σ_n satisfies the $(\kappa, 3 + 4\lambda_2/\lambda_1)$-restricted eigenvalue condition over S_*. Assume the sample size is at least $n \ge (32d^2/\kappa^2)\log(d/\delta_1)$. Let $\hat{\theta}_{\lambda_1,\lambda_2}$ be the elastic net estimator (2) with ℓ^1-regularization hyperparameter $\lambda_1 = 4\sqrt{(2/n)\log(2d/\delta_2)}$. Then with probability at least $1-\delta_1-\delta_2$, the estimation error satisfies the bounds*

$$\left\|\hat{\theta}_{\lambda_1,\lambda_2} - \theta_*\right\|_{U_n+\lambda_2 I_d} \le \left(6\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta_2}\right)} + 2\lambda_2\right)\sqrt{\frac{2s}{\kappa + 2\lambda_2}}, \tag{7}$$
$$\left\|\hat{\theta}_{\lambda_1,\lambda_2} - \theta_*\right\| \le \left(6\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta_2}\right)} + 2\lambda_2\right)\frac{2\sqrt{s}}{\kappa + 2\lambda_2}.$$

Furthermore, if $\kappa < 3\sqrt{(2/n)\log(2d/\delta_2)}$ and $\lambda_2 = 3\sqrt{(2/n)\log(2d/\delta_2)} - \kappa$, the bound (7) implies

$$\left\|\hat{\theta}_{\lambda_1,\lambda_2} - \theta_*\right\|_{U_n+\lambda_2 I_d} \le 2\sqrt{2s}\left(6\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta_2}\right)} - \kappa\right)^{\frac{1}{2}}$$
$$\le 4\sqrt[4]{\frac{18s^2}{n}\log\left(\frac{2d}{\delta_2}\right)}.$$

3.3 Dantzig Selector

Before presenting our fourth main result, we introduce two more conditions on symmetric matrices which underlie standard assumptions in the statistics and compressed sensing literature (Bandeira et al., 2013):

Definition 2 (Restricted isometry property). *Fix $\epsilon \geq 0$ and $p \in \mathbb{N}$. A symmetric matrix $A \in \mathbb{R}^{d\times d}$ satisfies the (ϵ, p)-restricted isometry property iff*

$$\forall v \in \mathbb{R}^d, \ \ \|v\|_0 \leq p \implies (1-\epsilon)\|v\|^2 \leq v^\top A v \leq (1+\epsilon)\|v\|^2 .$$

The restricted isometry property states that the curvature of the quadratic form represented by A is bounded around 1 along directions residing in sparse axis-aligned subspaces, or alternatively, that the linear operator represented by A is approximately scale-preserving when applied to sparse inputs. A weaker variant of this property bears resemblance to the Cauchy-Schwarz inequality:

Definition 3 (Restricted orthogonality property). *Fix $\zeta \geq 0$, $p \in \mathbb{N}$, and $q \in \mathbb{N}$. A symmetric matrix $A \in \mathbb{R}^{d\times d}$ satisfies the (ζ, p, q)-restricted orthogonality property iff*

$$\forall u, v \in \mathbb{R}^d, \ \ \|u\|_0 \leq p \wedge \|v\|_0 \leq q \wedge \operatorname{supp}(u) \cap \operatorname{supp}(v) = \emptyset \implies \left|u^\top A v\right| \leq \zeta \|u\| \|v\| .$$

Under the restricted orthogonality property, A maps sparse vectors with disjoint support to dissimilar outputs. Our fourth main result is a high-probability upper bound on the error of the Dantzig selector in the fixed design setting, under assumptions on U_n similar to the prior literature (Candes & Tao, 2007):

Theorem 5 (Dantzig selector fixed design upper bound). *Fix $\delta \in (0,1)$, $\epsilon \in [0,1)$, and $\zeta \in [0, 1-\epsilon)$. Assume the samples $(x,y)^{1:n}$ are generated according to the fixed design setting. Assume U_n satisfies the $(\epsilon, 2s)$-restricted isometry property and the $(\zeta, s, 2s)$-restricted orthogonality property. Let $\bar{\theta}_\lambda$ be the Dantzig selector (3) with regularization hyperparameter $\lambda = \sqrt{(2/n)\log(2d/\delta)}$. Then with probability at least $1-\delta$, the estimation error satisfies the bound*

$$\left\|\bar{\theta}_\lambda - \theta_*\right\| \leq \frac{4}{1-\epsilon-\zeta}\sqrt{\frac{2s}{n}\log\left(\frac{2d}{\delta}\right)} .$$

We include the Dantzig selector in our investigation of sparse contextual CDF regression as an example of variable selection formulated as a linear programming problem (4), as opposed to the usual quadratic programming perspective of lasso estimation. We remark that the $(\epsilon, 3p)$-restricted isometry property implies the $(\epsilon, p, 2p)$-restricted orthogonality property, by Candes & Tao (2005, Lemma 1.1). Hence, Theorem 5 also holds when the sole assumption placed on U_n is the $(\epsilon, 3s)$-restricted isometry property.

We defer the technical details of our proof to Appendix D and provide a high-level summary below. At the outset, we upper-bound the estimation error $\|\Delta\|$ in terms of the components $\|[\Delta]_{S_\dagger}\|$ and $\|[\Delta]_{S_*^c}\|_1$, where $S_\dagger$ is a superset of S_* (34). Utilizing the sparsity of θ_*, we bound the latter component in terms of the former (35). Subsequently, we upper-bound $\|[\Delta]_{S_\dagger}\|$ in terms of $\|[U_n]_{\langle S_\dagger\rangle}\Delta\|$ (37), which in turn depends on two quantities (38):

- The scalar projections of the sampling errors $\mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j$ onto the d basis functions Φ_j, which we upper-bound using various probabilistic arguments and concentration inequalities.
- The scalar projections of $\mathrm{I}_{y^{(j)}} - \bar{\theta}_\lambda^\top \Phi_j$ onto Φ_j, which we upper-bound using the constraint in (4).

Combining these bounds and applying various norm equivalences completes the proof. We note that the restricted isometry and restricted orthogonality assumptions on U_n are preconditions to upper-bound $\|\Delta\|$ and $\|[\Delta]_{S_\dagger}\|$ in the proof.

As previously done for the lasso and elastic net estimators, we use a matrix version of Hoeffding's inequality to derive the following high-probability upper bound on the error of the Dantzig selector in the random design setting, which we prove in Appendix D:

Theorem 6 (Dantzig selector random design upper bound). *Fix $\delta_1 \in (0,1)$, $\delta_2 \in (0, 1-\delta_1)$, $\epsilon \in (0, 1/2)$, and $\zeta \in (0, 1/2-\epsilon)$. Assume the samples $(x,y)^{1:n}$ are generated according to the random design setting. Assume that for any $n \in \mathbb{N}$, the matrix Σ_n satisfies the $(\epsilon, 2s)$-restricted isometry property and the $(\zeta, s, 2s)$-restricted* *orthogonality property. Assume the sample size is at least* $n \geq (8d^2/\min\{\epsilon,\zeta\}^2)\log(2d/\delta_1)$*. Let* $\bar{\theta}_\lambda$ *be the Dantzig selector (3) with regularization hyperparameter* $\lambda = \sqrt{(2/n)\log(2d/\delta_2)}$*. Then with probability at least* $1-\delta_1-\delta_2$*, the estimation error satisfies the bound*

$$\left\|\bar{\theta}_\lambda - \theta_*\right\| \leq \frac{4}{1-2\epsilon-2\zeta}\sqrt{\frac{2s}{n}\log\left(\frac{2d}{\delta_2}\right)}.$$

3.4 Lower Bound

Our last main result is a lower bound on the *minimax* ℓ^2*-risk* of sparse contextual CDF regression. To begin this discussion, we introduce some relevant notation. For any $d \in \mathbb{N}$, let $\mathcal{B}_d$ be the universe of all d-dimensional bases of contextual CDFs, i.e.,

$$\mathcal{B}_d = \left\{\Phi : \mathcal{X} \times \mathbb{R} \to [0,1]^d : \forall i \in [d],\, \forall x \in \mathcal{X},\, [\Phi(x,\cdot)]_i \text{ is a CDF}\right\}.$$

For any $x \in \mathcal{X}$, $\theta \in \Delta^{d-1}$, and $\Phi \in \mathcal{B}_d$, let $P^{\Phi}_{Y|x,\theta}$ denote the probability distribution corresponding to the CDF $\theta^\top \Phi(x,\cdot)$. Given context variables $x^{1:n} \in \mathcal{X}^n$, let $\mathcal{P}^d_{x^{1:n}}$ denote the family of product distributions of $Y^{1:n}$ which are convex combinations of d contextual CDFs, i.e.,

$$\mathcal{P}^d_{x^{1:n}} = \left\{\bigotimes_{j=1}^{n} P^{\Phi}_{Y|x^{(j)},\theta} : \theta \in \Delta^{d-1} \wedge \Phi \in \mathcal{B}_d\right\}.$$

For any $s \leq d$, let $\mathcal{P}^{d,s}_{x^{1:n}} \subset \mathcal{P}^d_{x^{1:n}}$ denote the family of product distributions of $Y^{1:n}$ which are convex combinations of s contextual CDFs chosen from a d-dimensional basis, i.e.,

$$\mathcal{P}^{d,s}_{x^{1:n}} = \left\{\bigotimes_{j=1}^{n} P^{\Phi}_{Y|x^{(j)},\theta} : \theta \in \Delta^{d-1} \wedge \|\theta\|_0 = s \wedge \Phi \in \mathcal{B}_d\right\}.$$

Given a distribution $P \in \mathcal{P}^d_{x^{1:n}}$, let $\theta(P) \in \Delta^{d-1}$ denote its parameter and let $S_*(P) = \operatorname{supp}(\theta(P)) \subseteq [d]$. Consequently, $|S_*(P)| = s$ for any $P \in \mathcal{P}^{d,s}_{x^{1:n}}$. For any $n \in \mathbb{N}$, let $\hat{\Theta}_{d,n}$ be the universe of all (possibly randomized) estimators $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^d$.

Our main result establishes a lower bound on the estimation error of any estimator for sparse contextual CDF regression, and is obtained as a consequence of the lower bound for general contextual CDF regression in Zhang et al. (2024, Theorem 8):

Proposition 1 (Lower bound)**.** *Fix* $d \in \mathbb{N}$*,* $s \leq d$*, and any sufficiently large* $n \geq s/2$*. Fix* $x^{1:n} \in \mathcal{X}^n$*. Then, the minimax* ℓ^2*-risk of sparse contextual CDF regression satisfies the bound*

$$\mathfrak{R}\left(\theta\left(\mathcal{P}^{d,s}_{x^{1:n}}\right)\right) = \inf_{\hat{\theta} \in \hat{\Theta}_{d,n}} \sup_{P \in \mathcal{P}^{d,s}_{x^{1:n}}} \mathop{\mathbb{E}}_{Y^{1:n}\sim P}\left[\left\|\hat{\theta}\left(Y^{1:n}\right) - \theta(P)\right\|\right] = \Omega\left(\sqrt{\frac{s}{n}}\right).$$

We defer the technical details of our proof to Appendix E. In a nutshell, it suffices to consider only the estimation error component corresponding to the indices in $S_*(P)$, thereby reducing the problem to general s-dimensional CDF regression. Furthermore, we obtain minimax upper bounds from the high-probability upper bounds in Theorems 1 and 5 by the arguments from Zhang et al. (2024, p. 12). Hence, Proposition 1 establishes the minimax optimality of the lasso and Dantzig selector estimators up to logarithmic factors.

4 Proof of Lasso Fixed Design Upper Bound

In this section, we prove Theorem 1. We begin by deriving a concentration bound on the inner products between the sampling errors $\mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j$ and the basis functions Φ_j:

Lemma 1 (Concentration bound)**.** *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, it holds that*

$$\left\| \frac{1}{n} \sum_{j=1}^{n} \left\langle \mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j, \Phi_j \right\rangle \right\|_\infty \leq \sqrt{\frac{2}{n} \log\left(\frac{2d}{\delta}\right)} . \tag{8}$$

Proof of Lemma 1. For each $j \in [n]$ and $i \in [d]$, define a random variable

$$Z_{i,j} = \left\langle \mathrm{I}_{Y^{(j)}} - \theta_*^\top \Phi_j, [\Phi_j]_i \right\rangle .$$

Let $\mathcal{F}_j = \sigma(Y^{1:j-1})$ denote the σ-algebra generated by the random variables $Y^{1:j-1}$.[1] The mean of $Z_{i,j}$ is

$$\begin{aligned}
\mathbb{E}[Z_{i,j}] &\overset{(a)}{=} \mathbb{E}[\mathbb{E}[Z_{i,j} \mid \mathcal{F}_j]] \\
&\overset{(b)}{=} \mathbb{E}\big[\mathbb{E}\big[\big\langle \mathrm{I}_{Y^{(j)}} - \theta_*^\top \Phi_j, [\Phi_j]_i \big\rangle \bigm| \mathcal{F}_j\big]\big] \\
&\overset{(c)}{=} \mathbb{E}\big[\big\langle \mathbb{E}\big[\mathrm{I}_{Y^{(j)}} - \theta_*^\top \Phi_j \bigm| \mathcal{F}_j\big], [\Phi_j]_i \big\rangle\big] \\
&\overset{(d)}{=} \mathbb{E}\big[\big\langle 0, [\Phi_j]_i \big\rangle\big] = 0 ,
\end{aligned}$$

where (a) holds by the tower rule of expectation, (b) holds by definition of $Z_{i,j}$, (c) holds by Fubini's theorem, and (d) holds because $\mathbb{E}[\mathrm{I}_{Y^{(j)}} \mid \mathcal{F}_j] = \theta_*^\top \Phi_j$. Furthermore, the support of $Z_{i,j}$ is $[0,1]$ because

$$\begin{aligned}
|Z_{i,j}| &\overset{(a)}{=} \left| \int_S \left(\mathrm{I}_{y^{(j)}}(t) - \theta_*^\top \Phi_j(t)\right) \phi_i\left(x^{(j)}, t\right) d\mathfrak{m} \right| \\
&\overset{(b)}{\leq} \int_S \left|\mathrm{I}_{y^{(j)}}(t) - \theta_*^\top \Phi_j(t)\right| \left|\phi_i\left(x^{(j)}, t\right)\right| d\mathfrak{m} \\
&\overset{(c)}{\leq} \mathfrak{m}(S) \overset{(d)}{=} 1 ,
\end{aligned}$$

where (a) holds by definition of inner product, (b) holds by the triangle inequality, (c) holds because CDFs are bounded between 0 and 1, and (d) holds because $\mathfrak{m}$ is a probability measure. Thus, for any $\tau > 0$,

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{j=1}^{n} \left\langle \mathrm{I}_{Y^{(j)}} - \theta_*^\top \Phi_j, \Phi_j \right\rangle \right\|_\infty \geq \tau \right) \overset{(a)}{\leq} \sum_{i=1}^{d} \mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^{n} Z_{i,j} \right| \geq \tau \right) \overset{(b)}{\leq} 2d \exp\left(-\frac{n\tau^2}{2}\right) ,$$

where (a) holds by definition of $Z_{i,j}$ and the union bound, and (b) holds by Hoeffding's inequality. Choosing $\tau = \sqrt{2/n \log(2d/\delta)}$ and rearranging, we get $\delta = 2d \exp(-n\tau^2/2)$, and thus

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{j=1}^{n} \left\langle \mathrm{I}_{Y^{(j)}} - \theta_*^\top \Phi_j, \Phi_j \right\rangle \right\|_\infty \leq \sqrt{\frac{2}{n} \log\left(\frac{2d}{\delta}\right)} \right) \geq 1 - \delta$$

as desired. ■

Now, we are ready to prove Theorem 1.

Proof of Theorem 1. Throughout this proof, we restrict to the subset of the probability space where

$$\left\| \frac{1}{n} \sum_{j=1}^{n} \left\langle \mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j, \Phi_j \right\rangle \right\|_\infty \leq \sqrt{\frac{2}{n} \log\left(\frac{2d}{\delta}\right)} = \frac{\lambda}{4} , \tag{9}$$

[1]This lemma also applies in the random design setting by taking $\mathcal{F}_j = \sigma(X^{1:j}, Y^{1:j-1})$.

which holds with probability at least $1-\delta$ by Lemma 1. For notational simplicity, let $\Delta = \hat{\theta}_\lambda - \theta_*$. We have

$$\begin{aligned}
\|\Delta\|_{U_n}^2 &\overset{(a)}{=} \Delta^\top U_n \Delta \\
&\overset{(b)}{=} \Delta^\top \left(\frac{1}{n} \sum_{j=1}^n \int_S \Phi_j \Phi_j^\top \, d\mathfrak{m} \right) \Delta \\
&\overset{(c)}{=} \frac{1}{n} \sum_{j=1}^n \int_S \Delta^\top \Phi_j \Phi_j^\top \Delta \, d\mathfrak{m} \\
&\overset{(d)}{=} \frac{1}{n} \sum_{j=1}^n \left\| \Delta^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 \\
&\overset{(e)}{=} \frac{1}{n} \sum_{j=1}^n \left(\left\| \hat{\theta}_\lambda^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 + \left\| \theta_*^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 \right) - \frac{2}{n} \sum_{j=1}^n \left\langle \hat{\theta}_\lambda^\top \Phi_j, \theta_*^\top \Phi_j \right\rangle \\
&\overset{(f)}{=} \frac{1}{n} \sum_{j=1}^n \left(\left\| \hat{\theta}_\lambda^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 - \left\| \theta_*^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 \right) - \frac{2}{n} \sum_{j=1}^n \left\langle \Delta^\top \Phi_j, \theta_*^\top \Phi_j \right\rangle \,,
\end{aligned} \tag{10}$$

where (a) holds by definition of weighted ℓ^2-norm induced by U_n, (b) holds by definition of U_n, (c) holds by the linearity of integration, (d) holds by definition of inner product between functions, (e) holds by definition of Δ, and (f) holds by definition of Δ and the linearity of inner product. Next, since $\hat{\theta}_\lambda$ minimizes the objective in (1), it follows that

$$\frac{1}{n} \sum_{j=1}^n \left\| \mathrm{I}_{y^{(j)}} - \hat{\theta}_\lambda^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 + \lambda \left\| \hat{\theta}_\lambda \right\|_1 \leq \frac{1}{n} \sum_{j=1}^n \left\| \mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 + \lambda \left\| \theta_* \right\|_1 \,. \tag{11}$$

Expanding the squared norms and rearranging, it follows that (cf. Wainwright (2019))

$$\frac{1}{n} \sum_{j=1}^n \left(\left\| \hat{\theta}_\lambda^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 - \left\| \theta_*^\top \Phi_j \right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 \right) \leq \frac{2}{n} \sum_{j=1}^n \left\langle \mathrm{I}_{y^{(j)}}, \Delta^\top \Phi_j \right\rangle + \lambda \left(\|\theta_*\|_1 - \left\| \hat{\theta}_\lambda \right\|_1 \right) \,. \tag{12}$$

Combining Equations (10) and (12), we obtain

$$\|\Delta\|_{U_n}^2 \leq \underbrace{\frac{2}{n} \sum_{j=1}^n \left\langle \mathrm{I}_{y^{(j)}}, \Delta^\top \Phi_j \right\rangle - \frac{2}{n} \sum_{j=1}^n \left\langle \Delta^\top \Phi_j, \theta_*^\top \Phi_j \right\rangle}_{①} + \lambda \underbrace{\left(\|\theta_*\|_1 - \left\| \hat{\theta}_\lambda \right\|_1 \right)}_{②} \,. \tag{13}$$

Next, we upper-bound ①. We have

$$① \overset{(a)}{=} \Delta^\top \left(\frac{2}{n} \sum_{j=1}^n \left\langle \mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j, \Phi_j \right\rangle \right) \overset{(b)}{\leq} \|\Delta\|_1 \left\| \frac{2}{n} \sum_{j=1}^n \left\langle \mathrm{I}_{y^{(j)}} - \theta_*^\top \Phi_j, \Phi_j \right\rangle \right\|_\infty \overset{(c)}{\leq} \frac{\lambda}{2} \|\Delta\|_1 \,, \tag{14}$$

where (a) holds by the linearity of inner product, (b) follows from Hölder's inequality, and (c) holds by substituting in (9). Next, we upper-bound ②. We have

$$\begin{aligned}
② &\overset{(a)}{=} \sum_{i=1}^d |[\theta_*]_i| - \sum_{i=1}^d |[\theta_* + \Delta]_i| \\
&\overset{(b)}{=} \sum_{i \in S_*} |[\theta_*]_i| - \sum_{i \in S_*} |[\theta_* + \Delta]_i| - \sum_{i \in S_*^c} |[\Delta]_i| \\
&\overset{(c)}{=} \left\| [\theta_*]_{S_*} \right\|_1 - \left\| [\theta_*]_{S_*} + [\Delta]_{S_*} \right\|_1 - \left\| [\Delta]_{S_*^c} \right\|_1
\end{aligned}$$

$$\stackrel{(d)}{\leq} \left\|[\Delta]_{S_*}\right\|_1 - \left\|[\Delta]_{S_*^c}\right\|_1 , \tag{15}$$

where (a) holds by definition of ℓ^1-norm and Δ, (b) holds because $\text{supp}(\theta_*) = S_*$, (c) holds by definition of ℓ^1-norm, and (d) holds by the triangle inequality. Combining Equations (13) to (15), we obtain

$$\|\Delta\|_{U_n}^2 \leq \lambda \left(\frac{3}{2} \left\|[\Delta]_{S_*}\right\|_1 - \frac{1}{2} \left\|[\Delta]_{S_*^c}\right\|_1 \right) \tag{16}$$

$$\leq \frac{3\lambda}{2} \left\|[\Delta]_{S_*}\right\|_1 \tag{17}$$

$$\stackrel{(a)}{\leq} \frac{3\lambda}{2}\sqrt{s} \left\|[\Delta]_{S_*}\right\| \leq \frac{3\lambda}{2}\sqrt{s}\,\|\Delta\|$$

$$\stackrel{(b)}{\leq} \frac{3\lambda}{2}\sqrt{\frac{s}{\kappa}}\,\|\Delta\|_{U_n} ,$$

where (a) holds by the equivalence between ℓ^1- and ℓ^2-norms, and (b) follows from the restricted eigenvalue condition on U_n since rearranging (16) yields $\|[\Delta]_{S_*^c}\|_1 \leq 3\|[\Delta]_{S_*}\|_1$. Thus, the estimation error satisfies the bounds

$$\|\Delta\|_{U_n} \leq \frac{3\lambda}{2}\sqrt{\frac{s}{\kappa}} \stackrel{(a)}{=} 6\sqrt{\frac{2s}{\kappa n}\log\left(\frac{2d}{\delta}\right)} \quad \text{and} \quad \|\Delta\| \stackrel{(b)}{\leq} \frac{1}{\sqrt{\kappa}}\|\Delta\|_{U_n} = \frac{6}{\kappa}\sqrt{\frac{2s}{n}\log\left(\frac{2d}{\delta}\right)} ,$$

where (a) holds by substituting in λ and (b) follows from the restricted eigenvalue condition on U_n. ■

5 Numerical Simulations

In this section, we numerically simulate the data generation processes described in Section 1.2 and empirically evaluate the accuracy of our proposed lasso (1) and elastic net (2) estimators on synthetic data. For mathematical conciseness, we consider a basis of contextual Bernoulli CDFs, which yields a closed-form expression for the induced norm $\|\cdot\|_{\mathcal{L}^2(S,\mathfrak{m})}$ in the training objective. Formally, let $\mathcal{X} = [0,1]^d$ be the space of all d-tuples $x = (x_1, \ldots, x_d)$ of Bernoulli parameters, and define the basis functions

$$\forall i \in [d], \;\; \phi_i(x,t) = \begin{cases} 1 - x_i , & \text{if } 0 \leq t < 1 , \\ 1 , & \text{if } t = 1 , \end{cases}$$

where $\phi_i(x,\cdot)$ is the CDF of a $\mathsf{Bernoulli}(x_i)$ random variable. It follows that $F(x,\cdot) = \theta_*^\top \Phi(x,\cdot)$ is the CDF of a $\mathsf{Bernoulli}(\theta_*^\top x)$ random variable, because

$$F(x,t) \;=\; \sum_{i=1}^{d} [\theta_*]_i \cdot \begin{cases} 1 - x_i , & \text{if } 0 \leq t < 1 \\ 1 , & \text{if } t = 1 \end{cases} \stackrel{(a)}{=} \begin{cases} 1 - \theta_*^\top x , & \text{if } 0 \leq t < 1 , \\ 1 , & \text{if } t = 1 , \end{cases}$$

where (a) holds because θ_* is a PMF. Let $\mathfrak{m}$ be the uniform measure over the support $S = [0,1]$. Then, the log-likelihood term in Equations (1) and (2) is equivalent to the canonical least squares formulation because

$$\sum_{j=1}^{n} \left\|\mathrm{I}_{y^{(j)}} - \theta^\top \Phi_j\right\|_{\mathcal{L}^2(S,\mathfrak{m})}^2 \stackrel{(a)}{=} \sum_{j=1}^{n} \int_0^1 \left(\mathrm{I}_{y^{(j)}}(t) - \theta^\top \Phi_j(t)\right)^2 dt$$

$$\stackrel{(b)}{=} \sum_{j=1}^{n} \left(\mathbb{1}\left\{y^{(j)} = 0\right\} - \left(1 - \theta^\top x^{(j)}\right)\right)^2$$

$$= \sum_{j=1}^{n} \left(\theta^\top x^{(j)} - y^{(j)}\right)^2$$

$$\stackrel{(c)}{=} \|A\theta - b\|^2 ,$$

where (a) follows from the support $S = [0, 1]$ and the choice of measure, (b) holds by substituting in the definitions of $\mathrm{I}_{y^{(j)}}$ and Φ_j, and the matrix $A \in \mathbb{R}^{n\times d}$ and vector $b \in \mathbb{R}^n$ in (c) are defined as $[A]_{\langle j\rangle} = x^{(j)\top}$ and $[b]_j = y^{(j)}$. For specific details regarding our implementation, we refer interested readers to our Python code at https://anonymous.4open.science/r/SparseContextualCDFRegression-E57B/.

In our experiments, we compare the ℓ^2-norm estimation error of our proposed lasso and elastic net estimators against the ridge regression baseline introduced in Zhang et al. (2024, Section 3.1). For the ℓ^2-regularization hyperparameter of ridge regression and the ℓ^1-regularization hyperparameter of lasso and elastic net regression, we use $\lambda = 4\sqrt{(2/n)\log(2d/\delta)}$ as specified in Theorems 1 and 3, with $\delta = 0.001$. For experimental convenience, we use various fixed values of λ_2 for the elastic net estimator, which we report in Figures 1 and 2. Different values of λ and δ produced qualitatively similar results. We investigate how the estimation errors scale with various problem dimensions, and report means and standard deviations over 30 independent random trials in Figures 1 and 2 for each configuration under consideration. We train all models on the same set of generated samples in each random trial.

5.1 Fixed Design

First, we consider the fixed design setting with context variables $x^{(j)} = (x_1^{(j)}, \dots, x_d^{(j)})$ given by

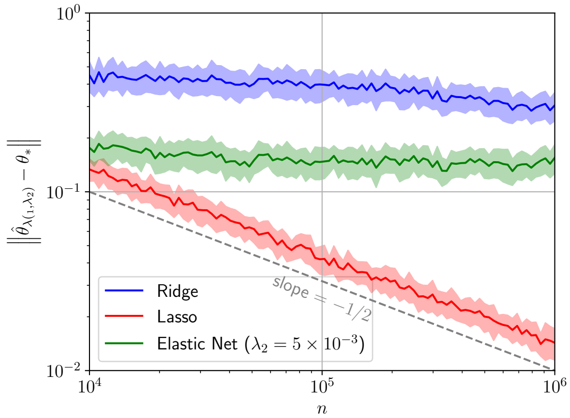
$$x_i^{(j)} = \begin{cases} 1 - 2x_{\mathsf{val}}^{(j)}, & \text{if } i \equiv j \pmod d \\ 1 - x_{\mathsf{val}}^{(j)}, & \text{if } i \not\equiv j \pmod d \end{cases} \qquad \text{with} \quad x_{\mathsf{val}}^{(j)} = \begin{cases} \frac{1}{2}, & \text{if } j \le d \\ \frac{\mu_{\min}(M_{j-1})}{\alpha_j}, & \text{if } j > d \end{cases}$$

$$\text{and} \quad M_j = \left(1 - x^{(j)}\right)\left(1 - x^{(j)}\right)^\top + \frac{1}{n}\sum_{k=1}^{j-1}\left(1 - x^{(k)}\right)\left(1 - x^{(k)}\right)^\top,$$

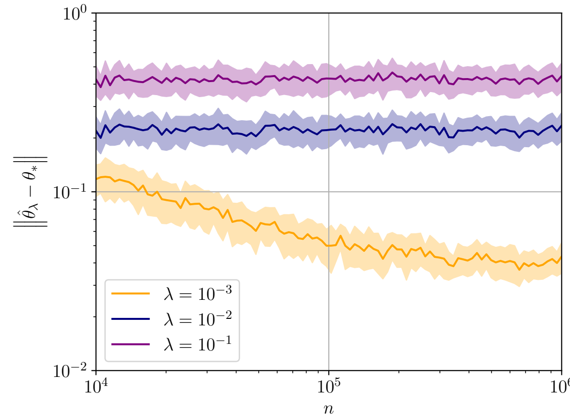
where α_j is initialized as $\alpha_{d+1} = \mu_{\min}(M_d)/2$ and is doubled as necessary on each j iteration to ensure $x^{(j)} \in \mathcal{X}$. To investigate the effect of the sample size n, we choose 100 logarithmically spaced points for n from 10^4 to 10^6, and fix the CDF basis dimension $d = 10$ and parameter sparsity $s = 5$. Figure 1(a) graphs the estimation errors against n on a log-log plot. The lasso trend line has slope $-1/2$, matching the theoretical $O(1/\sqrt{n})$ bound in Theorem 1 and substantially outperforming the ridge baseline. The elastic net estimator achieves results in between the lasso and ridge methods. As a further point of comparison, we repeat this experiment using handpicked regularization hyperparameters for the ridge estimator and plot the results in Figure 1(b). Values of λ less than 10^{-3} or greater than 10^{-1} produced results comparable to the yellow and purple trend lines, respectively. Observe that the accuracy of the ridge estimator with $\lambda = 10^{-3}$ matches the lasso estimator when $n \le 10^5$, but quickly plateaus for larger sample sizes. This control experiment confirms that the superior accuracy of our lasso estimator in the large n setting cannot be emulated by the ridge baseline regardless of hyperparameter tuning.

Secondly, we investigate how the estimation errors scale with the sparsity s of the true parameter θ_*. We fix the sample size $n = 10^5$ and CDF basis dimension $d = 30$, and consider s from 1 to 30. Figure 1(c) graphs the estimation errors against s on a linear plot. The lasso trend line reflects the theoretical $O(\sqrt{s})$ bound in Theorem 1, and our lasso estimator notably outperforms the ridge baseline in the sparse regime $s \ll d$. To interpret the decreasing ridge trend line, note that $\|\theta_*\|_1 = 1$ for any value of s (because θ_* is a PMF), and so denser parameter vectors (with greater s) tend to have smaller $\|\theta_*\|$. Thus, the ridge regularization penalty $\lambda\|\theta_*\|^2$ at the true parameter vector decreases as s increases, leading to improved accuracy. For additional comparison, we repeat this experiment for the elastic net estimator using handpicked values of λ_2 and plot the results in Figure 1(d). As expected, the elastic net estimation error interpolates continuously between the lasso and ridge estimation errors as λ_2 increases. Values of λ_2 outside the range visualized in Figure 1(d) produced similar results.

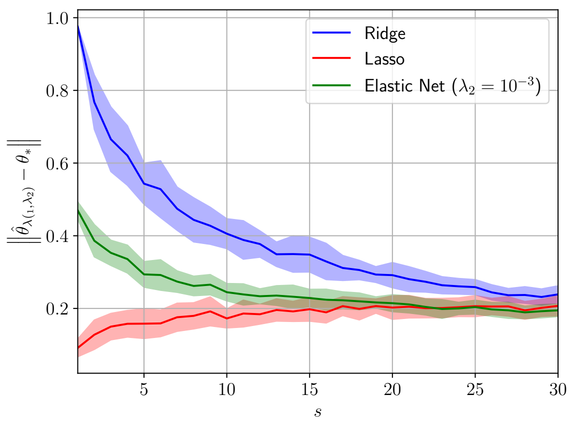
Lastly, we investigate the dependence of the estimation errors on the CDF basis dimension d. We fix the sample size $n = 10^5$ and parameter sparsity $s = 10$, and consider d from 10 to 100. Figure 1(e) graphs the estimation errors against d on a linear plot. The ridge and lasso trend lines indicate the respective theoretical bounds of $O(\sqrt{d})$ (Zhang et al., 2024, Section 3.2) and $O(\sqrt{\log d})$ (Theorem 1).



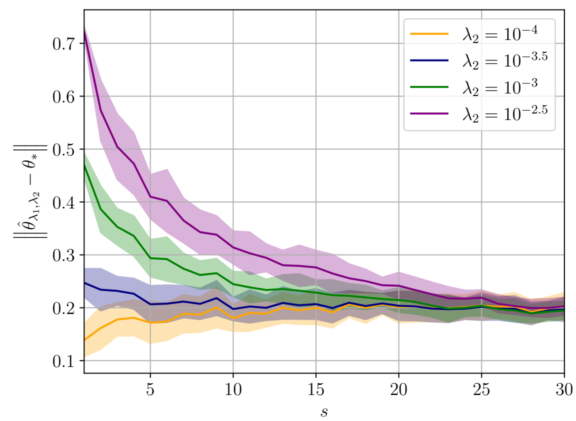
(a) Effect of sample size n



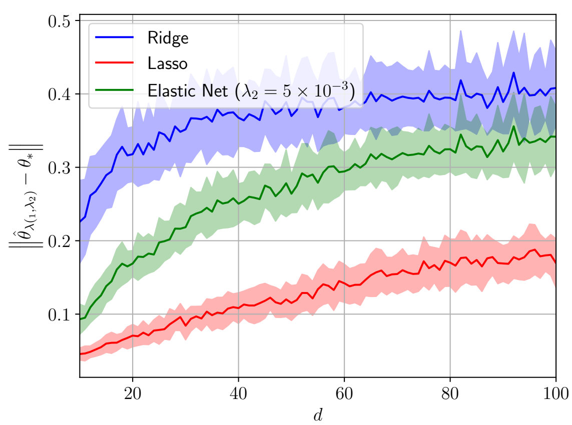
(b) Effect of ridge regularization λ



(c) Effect of sparsity s



(d) Effect of elastic net regularization λ_2

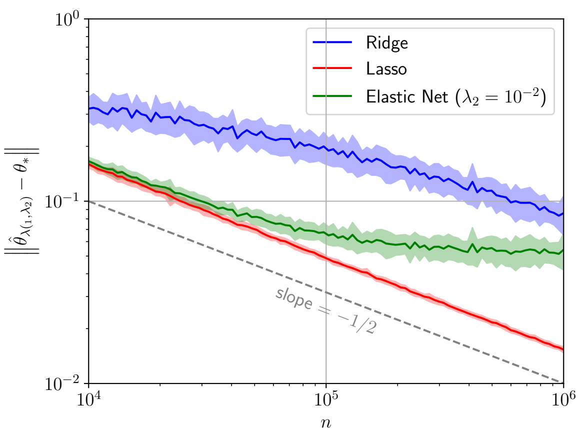


(e) Effect of CDF basis dimension d

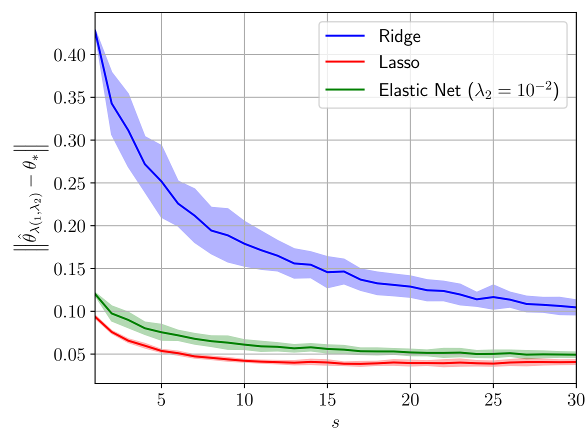
Figure 1: Means and standard deviations of ridge, lasso, and elastic net estimation errors against various hyperparameters in the synthetic Bernoulli experiments for the fixed design setting. Figures 1(a), 1(c) and 1(e) contrast the ridge, lasso, and elastic net estimators with common regularization hyperparameter $\lambda = \lambda_1 = 4\sqrt{(2/n)\log(2d/\delta)}$, as defined in Theorems 1 and 3. For further comparison, Figure 1(b) plots the ridge estimation error against n for various values of λ, and Figure 1(d) plots the elastic net estimation error against s for various values of λ_2.

5.2 Random Design

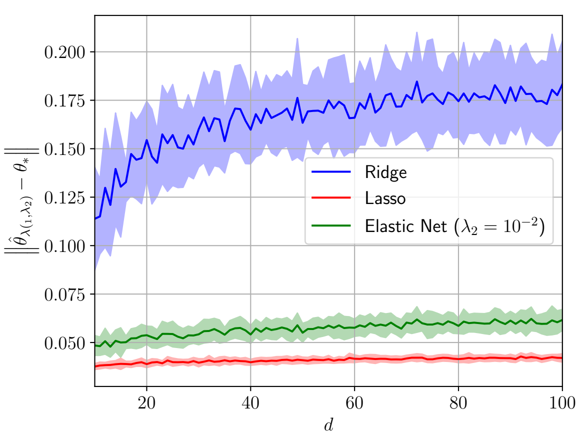
Next, we consider the random design setting with context vectors $x^{(j)} = (x_1^{(j)}, \ldots, x_d^{(j)})$ sampled i.i.d. from a $\mathsf{Uniform}(0,1)^d$ distribution. We investigate the effect of the sample size n, sparsity s, and CDF basis dimension d on the ridge, lasso, and elastic net estimation errors, as was done for the fixed design setting above, and plot the results in Figures 2(a) to 2(c) respectively. The qualitative trends in the random design experiments are comparable to the corresponding fixed design experiments, although the variance among the 30 trials for each configuration is lower in the random design setting. We conjecture that the variance is reduced by virtue of the context vectors in the random design experiments being more typical instances of the problem, as opposed to the fixed design setting where the context vectors were specifically chosen to be hard instances which may amplify the effect of the randomness in θ_*. On a different note, we remark that the elastic net error asymptotes for $n \approx 10^6$ in Figure 2(a) because λ_2 is fixed for experimental convenience and does not decrease with n, although a more refined hyperparameter setting could avoid this effect in principle. Lastly, we compare different values of λ_2 for the elastic net estimator in Figure 2(d) and observe the expected interpolation between the ridge and lasso trendlines as d ranges from 10 to 100.



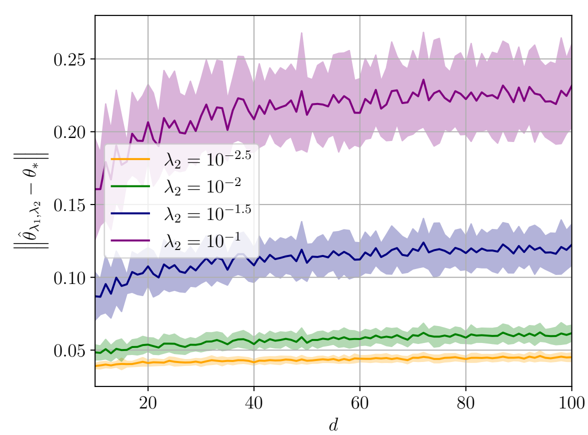
(a) Effect of sample size n



(b) Effect of sparsity s



(c) Effect of CDF basis dimension d



(d) Effect of elastic net regularization λ_2

Figure 2: Means and standard deviations of ridge, lasso, and elastic net estimators against various hyperparameters in the synthetic Bernoulli experiments for the random design setting. Figures 2(a) to 2(c) contrast the ridge, lasso, and elastic net estimators with common regularization hyperparameter $\lambda = \lambda_1 = 4\sqrt{(2/n)\log(2d/\delta)}$, as defined in Theorems 1 and 3. For further comparison, Figure 2(d) plots the elastic net estimation error against d for various values of λ_2.

6 Conclusion

In this paper, we introduced the task of sparse contextual CDF regression and proposed three basis selection techniques for this problem stemming from the canonical lasso, elastic net, and Dantzig selector regression methods. We derived upper bounds of $\tilde{O}(\sqrt{s/n})$ and $\tilde{O}(\sqrt{s}/\sqrt[4]{n})$ on estimation error, and obtained a matching lower bound on minimax risk to establish the optimality of our proposed lasso and Dantzig selector estimators. In particular, our estimation bounds have sub-polynomial dependence on the dimension d of the CDF regression basis, enabling our methods to perform basis selection with exponentially many irrelevant features and furthering progress towards the ultimate goal of general contextual CDF estimation.

We suggest three directions for future work. Firstly, our present analysis holds only when d is finite. A natural continuation of our research may investigate similar basis selection methods for CDF regression with infinite-dimensional feature maps. Another promising follow-up direction is to generalize our results to CDF regression with the least absolute deviation (ℓ^1-) loss, in the vein of previous work which combines robust regression and variable selection (Wang et al., 2007). Lastly, prior literature has established looser estimation error bounds on finite-dimensional lasso regression without the restricted isometry property (Zhao & Yu, 2006; Meinshausen & Yu, 2009). In this spirit, future work may aim to determine unified and weaker necessary conditions for analyzing the functional lasso, elastic net, and Dantzig estimators proposed in this paper.

Overall, our main contributions and proposed future directions indicate that contextual CDF estimation remains a fruitful area of theoretical investigation, accompanied by immediate relevance to a profusion of downstream scientific applications.

References

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

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

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

Afonso S Bandeira, Matthew Fickus, Dustin G Mixon, and Percy Wong. The road to deterministic matrices with the restricted isometry property. *Journal of Fourier Analysis and Applications*, 19(6):1123–1149, 2013.

Rina Foygel Barber, Matthew Reimherr, and Thomas Schill. The function-on-scalar LASSO with applications to longitudinal GWAS. *Electronic Journal of Statistics*, 11(1):1351 – 1389, 2017. doi: 10.1214/17-EJS1260. URL https://doi.org/10.1214/17-EJS1260.

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

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542. URL https://doi.org/10.1137/080716542.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51 (12):4203–4215, 2005. doi: 10.1109/TIT.2005.858979.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313 – 2351, 2007. doi: 10.1214/009053606000001523. URL https://doi.org/10.1214/009053606000001523.

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.

Fabio Centofanti, Matteo Fontana, Antonio Lepore, and Simone Vantini. Smooth lasso estimator for the function-on-function linear regression model. *Computational Statistics & Data Analysis*, 176:107556, 2022.

Michael B. Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pp. 938–942, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. doi: 10.1145/3313276.3316303. URL https://doi.org/10.1145/3313276.3316303.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

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.

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

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Zhaoping Hong and Heng Lian. Inference of genetic networks from time course expression data using functional regression with lasso penalty. *Communications in Statistics-Theory and Methods*, 40(10): 1768–1779, 2011.

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, 2012.

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

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

Cheng Jie, LA Prashanth, Michael Fu, Steve Marcus, and Csaba Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9):2867–2882, 2018.

PAVLO A. Krokhmal. Higher moment coherent risk measures. *Quantitative Finance*, 7(4):373–387, 2007. doi: 10.1080/14697680701458307.

Liang Li, Yongcheng Li, and Qing Ling. Fast algorithms to solve the dantzig selector. In *2013 10th IEEE International Conference on Control and Automation (ICCA)*, pp. 1544–1549. IEEE, 2013.

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

Paolo Maranzano, Philipp Otto, and Alessandro Fassò. Adaptive lasso estimation for functional hidden dynamic geostatistical models. *Stochastic Environmental Research and Risk Assessment*, pp. 1–23, 2023.

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

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1), 2009.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Jasdeep Pannu and Nedret Billor. Robust group-lasso for functional regression model. *Communications in statistics-simulation and computation*, 46(5):3356–3374, 2017.

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.

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

Juan Romo, Rosa E. Lillo, and Nicola Mingotti. Lasso variable selection in functional regression. In *DES - Working Papers. Statistics and Econometrics*, 2013. URL https://api.semanticscholar.org/CorpusID:118891754.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. *arXiv preprint arXiv:1301.1936*, 2013.

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

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

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

Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1330–1335. IEEE, 2015.

Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '20, pp. 259–278, USA, 2020. Society for Industrial and Applied Mathematics.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, and Renfei Zhou. New bounds for matrix multiplication: from alpha to omega. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 3792–3835. SIAM, 2024.

Julia L Wirch and Mary R Hardy. Distortion risk measures: Coherence and stochastic dominance. In *International congress on insurance: Mathematics and economics*, pp. 15–17, 2001.

William Wong, Audrey Huang, Liu Leqi, Kamyar Azizzadenesheli, and Zachary C Lipton. Riskyzoo: A library for risk-sensitive supervised learning. In *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments*, 2022.

Qian Zhang, Anuran Makur, and Kamyar Azizzadenesheli. Functional linear regression of cumulative distribution functions. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ZOqJCP4eMk.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7: 2541–2563, 2006.

Yihong Zhao, R Todd Ogden, and Philip T Reiss. Wavelet-based lasso in functional linear regression. *Journal of computational and graphical statistics*, 21(3):600–617, 2012.

Yihong Zhao, Huaihou Chen, and R Todd Ogden. Wavelet-based weighted lasso and screening approaches in functional linear regression. *Journal of Computational and Graphical Statistics*, 24(3):655–675, 2015.

Yujie Zhao and Xiaoming Huo. A survey of numerical algorithms that can solve the lasso problems. *WIREs Computational Statistics*, 15(4):e1602, 2023. doi: https://doi.org/10.1002/wics.1602. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1602.

Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.