

Variable Selection for Feature-Based Newsvendor

Zhaoliang Yuan

School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, 225080014@link.cuhk.edu.cn

Jie Wang

School of Artificial Intelligence and School of Data Science, The Chinese University of Hong Kong, Shenzhen, jwang@cuhk.edu.cn

Feature-based newsvendor models use observable covariates to tailor inventory decisions, aiming to balance holding and shortage costs under demand uncertainty. However, high-dimensional feature sets often hinder interpretability and inflate data collection and implementation costs. This paper studies variable selection for the feature-based newsvendor problem under a hard cardinality constraint on the number of selected features. We formulate the resulting ℓ_0 -constrained empirical newsvendor problem with ℓ_2 -regularization, establish its computational hardness, and develop a mixed-integer second-order cone programming reformulation that strengthens the standard Big- M formulation. To enable scalability beyond exact optimization, we develop a randomized-rounding algorithm with a bi-criteria guarantee and a greedy heuristic. Statistically, we provide theoretical analysis of the resulting sparse policy estimator, including finite-sample estimation error, out-of-sample risk bounds, and support recovery guarantees. Extensive experiments on both synthetic and real data illustrate the computational and statistical trade-offs among various baselines. Our results demonstrate that the proposed variable selection framework achieves competitive out-of-sample operational costs while using substantially fewer covariates.

1. Introduction

The newsvendor problem is a classic inventory model that seeks to determine the optimal order quantity by balancing the cost of overstocking and understocking in the face of random demand. It is crucial to obtain an accurate estimate of customer demand, which is influenced by a multitude of features, including weather, time, location, and economic conditions. In the big data era, leveraging available data that pairs these influential features (covariates) with corresponding demand observations enables us to make data-driven predictions of future customer demand and more reliable decisions, while ignoring the presence of covariates can lead to inconsistent decisions [3].

A natural paradigm for incorporating features into decision-making is contextual stochastic optimization (CSO) [31]. Broadly speaking, CSO can be divided into three classes: decision-rule optimization [27, 3], estimate-then-optimize [5], and integrated estimation and optimization [14, 21, 30]. In this paper, we focus on the decision-rule optimization approach, aiming to learn an end-to-end mapping from features to decisions. Linear decision rules are attractive because of their simplicity and tractability [3], but they may not capture complex

feature-to-decision relationships. To improve out-of-sample performance, nonlinear decision rules based on kernel methods [6, 28] and deep learning [29, 40, 20] have also been studied.

Despite this progress, the interpretability of learned decision rules has received comparatively less attention. In many applications, only a small subset of features is truly decision-relevant. Identifying such features can improve interpretability, reduce data collection and monitoring costs, simplify downstream implementation, and improve robustness when irrelevant features are noisy or unstable. This motivates us to consider variable selection for feature-based newsvendor models.

Variable selection has been extensively studied in statistics and machine learning through sparsity-inducing formulations, including Lasso [33], bridge-type penalties [19], the smoothly clipped absolute deviation (SCAD) penalty [37, 16], and the minimax concave penalty (MCP) [39]. These methods can be viewed as tractable surrogates for the ideal but computationally challenging ℓ_0 -constrained formulation. Convex penalties are typically easier to optimize, whereas nonconvex penalties may achieve sharper statistical properties but are harder to solve globally. In this paper, we study the following question:

How can we efficiently solve variable selection for feature-based newsvendor and what statistical guarantees can be established?

Our contributions are summarized as follows.

- (I) **Exact and approximation algorithms.** For the variable selection problem, we add an ℓ_2 -norm regularization and reformulate it as a mixed-integer second-order cone program (MISOCP). Compared with the standard Big- M approach for variable selection, our MISOCP formulation yields a tighter continuous relaxation and enables a more efficient algorithm design. Based on the reformulation, we further develop two efficient approximation algorithms and show that the randomized-rounding method admits a bi-criteria guarantee.
- (II) **Statistical guarantees.** Under the assumption that demand has sparse and linear dependence on features, we establish the concentration error bound between our estimated policy and the ground truth optimal policy, as well as the out-of-sample performance and support recovery guarantees. These performance bounds break the curse of dimensionality because they crucially depend on the sparsity level instead of the data dimension.

- (III) **Numerical validation.** We compare the exact optimization methods, randomized-rounding, greedy, and Lasso-based approaches using synthetic and retail data, illustrating their computational and statistical trade-offs.

1.1. Related Works

Data-Driven Newsvendor. The classical newsvendor model assumes that the demand distribution is known, whereas modern data-driven approaches learn ordering policies directly from historical data. Early work along these lines includes operational-statistics-based approaches [27] and feature-based linear policies for the big-data newsvendor [3]. Subsequent work has developed richer data-driven models based on quantile regression and machine learning [20], reproducing kernel Hilbert space (RKHS)-based policy classes [6, 5], and deep neural networks [29, 40]. Serrano et al. [32] provided a variable selection framework for the feature-based newsvendor problem. They select key features by adding the sparsity budget constraint to the objective and tune penalty parameters using bilevel optimization. Our work contributes to this literature by studying hard-cardinality-constrained policy learning and by providing both computational and statistical guarantees.

Quantile Regression. A fundamental observation in the newsvendor problem is that the optimal order quantity is a conditional quantile of the demand distribution. This connection links contextual newsvendor problems closely to quantile regression, beginning with the seminal work of Koenker and Bassett [23]. Quantile-regression-based methods have been used successfully in feature-based newsvendor models [20]. In high-dimensional settings, sparse quantile regression has been studied under both ℓ_1 -regularized and ℓ_0 -type formulations [4, 13]. The study in this paper can be interpreted as an ℓ_0 -constrained and ℓ_2 -regularized empirical quantile regression problem.

Variable Selection. Variable selection is a central topic in statistics and machine learning. Classical approaches promote sparsity through penalized formulations [33], or through greedy and local-search methods [24, 26, 38, 25, 10] that approximately solve ℓ_0 -constrained optimization problems. Although ℓ_0 -constrained formulations are generally NP-hard and computationally challenging in the worst case, they often perform well in high-dimensional, small-sample, and low-signal-to-noise settings [7, 1, 38]. This motivates the development of scalable algorithms for directly solving the ℓ_0 -constrained variable-selection problem for the

newsvendor model, as well as theoretical analyses that quantify the optimality gap of approximation algorithms. Compared with classical sparse ridge regression, our problem is further complicated by an empirical loss that is piecewise linear and asymmetric. To address these challenges, we combine perspective reformulations and support-selection algorithms with a statistical analysis tailored to the quantile-regression structure of the newsvendor loss.

Notations. For real numbers a, b , we let $a \vee b = \max\{a, b\}$. Given a positive integer n , define $[n] = \{1, \dots, n\}$. For a distribution function $F(\cdot)$, we denote $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}, p \in [0, 1]$. Given an event E , define $\mathbf{1}_E(x) = 1$ if $x \in E$ and otherwise $\mathbf{1}_E = 0$. For a vector $s \in \mathbb{R}^n$, we use $s[i]$ to denote its i -th component, and define its ℓ_0 -norm $\|s\|_0 = \sum_{i \in [n]} \mathbf{1}_{s[i] \neq 0}$. Given an $m \times n$ matrix $\mathbf{X}$ and a set $S \subseteq [n]$, let $\mathbf{X}_S$ denote the submatrix of $\mathbf{X}$ with columns from the set S . Given a set $\mathcal{F}$, let $\text{Conv}(\mathcal{F})$ denote its convex hull, i.e., the smallest convex set containing $\mathcal{F}$.

2. Problem Setup

The newsvendor problem seeks to solve the following stochastic optimization problem:

$$\min_{z \geq 0} \mathbb{E}[b \max\{y - z, 0\} + h \max\{z - y, 0\}], \quad (1)$$

where the expectation is taken with respect to the random demand y , z denotes the order quantity, and b, h represent the unit back-ordering and holding costs, respectively. A central challenge in solving (1) is that the underlying demand distribution is typically unknown.

In the context of big data, auxiliary features can be used for accurate demand estimation. Let $x \in \mathbb{R}^D$ denote a feature vector. For a given x , the optimal order quantity can be expressed as a function $f(x)$ with

$$f(x) = \operatorname{argmin}_{z \geq 0} \left\{ \mathbb{E}_{y|x} [b \max\{y - z, 0\} + h \max\{z - y, 0\}] \right\}.$$

Rather than first estimating the conditional demand distribution and then solving the resulting newsvendor problem, one can directly learn an end-to-end policy function f that maps features to decisions by solving the following contextual stochastic optimization (CSO) problem:

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)} [b \max\{y - f(x), 0\} + h \max\{f(x) - y, 0\}]. \quad (2)$$

The choice of the function class $\mathcal{F}$ critically influences both computational tractability and out-of-sample performance. In this paper, we focus on sparse linear decision rules of the form

$$\mathcal{F} = \left\{ f(x) = \omega^\top x + c : \omega \in \mathbb{R}^D, c \in \mathbb{R}, \|\omega\|_0 \leq d \right\}.$$

The ℓ_0 -constraint ensures that at most d covariates actively affect the decision, thereby facilitating interpretable and explainable policies. Under the linear decision-rule model, the subproblem associated with each fixed support is a convex optimization and therefore enables us to develop mixed-integer convex program reformulations for our variable selection framework. This framework can be extended to nonlinear decision rules through basis expansion.

REMARK 1 (LINEAR DEMAND MODEL). Consider the linear demand model $y = (\omega^*)^\top x + \epsilon$, where ϵ is a random variable independent of x with cumulative distribution function F_ϵ . Let $\tau = b/(b+h)$. The optimal feature-dependent newsvendor decision is the conditional τ -quantile of y given x , namely $f^*(x) = (\omega^*)^\top x + F_\epsilon^{-1}(\tau)$. Under this linear demand model, the slope in the linear decision rule recovers the true coefficient ω^* , while the intercept estimates the τ -quantile of the noise.

REMARK 2 (NONNEGATIVITY OF ORDER QUANTITIES). The optimal newsvendor decision has the nonnegative constraint. For the linear decision rule approach, we clip the estimated optimal order quantity to zero if it is negative. When demand is always non-negative, this procedure does not increase the newsvendor loss. In the main analysis, we focus on the unconstrained formulation in (2) for simplicity.

The objective in (2) involves the population expectation under the joint distribution of (x, y) . In practice, however, only independent and identically distributed (i.i.d.) samples $\{(x_i, y_i)\}_{i=1}^n$ are available. We therefore replace the population expectation with the empirical average and include an ℓ_2 -regularization term to improve both computational tractability and out-of-sample performance, which will be discussed in later sections. In summary, it leads to the optimization problem

$$\min_{\omega, c: \|\omega\|_0 \leq d} F(\omega, c) := \frac{1}{n} \sum_{i=1}^n \left[b \cdot \max\{0, y_i - \omega^\top x_i - c\} + h \cdot \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{1}{2\gamma} \|\omega\|_2^2, \quad (3)$$

Problem (3) is nonconvex because of the ℓ_0 -constraint, and is therefore generally difficult to solve. In fact, we show that it is NP-hard by a reduction from the exact cover by 3-sets problem. Its proof is provided in Appendix EC.1.

PROPOSITION 1 (Computational Hardness). *For fixed $\gamma, b, h > 0$, Problem (3) is NP-hard whenever the sparsity level d is part of the input.*

Problem (3) is also related to sparse quantile linear regression. Indeed, it is equivalent to

$$\min_{\omega, c: \|\omega\|_0 \leq d} \left\{ \frac{(b+h)}{n} \sum_{i=1}^n \rho_\tau(y_i - (\omega^\top x_i + c)) + \frac{1}{2\gamma} \|\omega\|_2^2 \right\},$$

where $\tau = \frac{b}{b+h}$ and $\rho_\tau(u) = u(\tau - \mathbf{1}_{\{u < 0\}})$. In other words, the goal of variable selection for the feature-based newsvendor is equivalent to estimating sparse coefficients that minimize the ℓ_2 -regularized empirical quantile regression loss.

The remainder of the paper is organized as follows. In Section 3, we present a mixed-integer second-order cone reformulation of Problem (3), which can be solved exactly for medium-sized instances. In Section 4, we develop two approximation algorithms with performance guarantees. In Section 5, we establish statistical guarantees for the proposed approach, including consistency, sample complexity, and support recovery.

3. Exact Computational Algorithms

In this section, we develop exact algorithms for solving Problem (3) to global optimality. A standard approach is the *Big-M* reformulation proposed by Dai [13], which converts the sparsity-constrained problem into a mixed-integer convex program.

Big-M Method. Recall that the constraint $\|\omega\|_0 \leq d$ requires that at most d entries of ω be nonzero. To model the support of ω , we introduce a binary vector $s \in \{0, 1\}^D$, where $s[i] = 1$ indicates that $\omega[i]$ is allowed to be nonzero. Equivalently, we impose $\mathbf{1}_{\{\omega[i] \neq 0\}} \leq s[i], i \in [D]$. Assume there exists a vector $M \in \mathbb{R}_+^D$ such that $|\omega[i]| \leq M[i], i \in [D]$, for some optimal solution ω . Then the logical constraint can be equivalently enforced by the linear inequality $|\omega[i]| \leq M[i] s[i], i \in [D]$. This observation leads to the following reformulation of Problem (3).

PROPOSITION 2 (Big-M reformulation [13]). *Assume there exists a vector $M \in \mathbb{R}_+^D$ such that an optimal solution ω^* to Problem (3) satisfies $|\omega^*[i]| \leq M[i], \forall i \in [D]$. Then Problem (3) is equivalent to*

$$\begin{aligned} \min_{\omega, c, s} \quad & \frac{1}{n} \sum_{i=1}^n \left[b \max\{0, y_i - \omega^\top x_i - c\} + h \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{1}{2\gamma} \|\omega\|_2^2 \\ \text{s.t.} \quad & \sum_{i \in [D]} s[i] \leq d, \quad s \in \{0, 1\}^D, \\ & |\omega[i]| \leq M[i] s[i], \quad \forall i \in [D]. \end{aligned} \tag{Big-M}$$

Big- M method is standard but often computationally expensive for large-scale mixed-integer problems. In practice, such formulations are typically solved by branch-and-bound, which repeatedly solves convex relaxations. A major difficulty is that the Big- M formulation usually yields a weak relaxation: the optimality gap between (Big- M) and its continuous relaxation, obtained by replacing $s \in \{0, 1\}^D$ with $s \in [0, 1]^D$, can be substantial.

Some heuristics can help derive tight bounds of M and improve numerical performance. For example, let (ω_0, c_0) be a near-optimal solution to Problem (3), which can be obtained using some tractable algorithms outlined in Section 4. Any optimal solution to Problem (3) satisfies that

$$\frac{\|\omega^*\|_2^2}{2\gamma} \leq F(\omega^*, c^*) \leq F(\omega_0, c_0).$$

Based on this relation, the vector M with $M[i] = \sqrt{2\gamma F(\omega_0, c_0)}$, $i \in [D]$ is a valid choice because $|\omega^*[i]| \leq \|\omega^*\|_2 \leq M[i]$. However, solving the Big- M formulation remains challenging for large-scale datasets.

Motivated by this limitation, we next introduce an alternative reformulation that is more computationally tractable. We show that it has a tighter convex relaxation than the Big- M approach and is therefore better suited for exact computation.

Perspective Reformulation. We introduce a stronger reformulation than the Big- M approach, based on the perspective function. To this end, we first rewrite the quadratic regularization term by introducing auxiliary variables $\mu \in \mathbb{R}_+^D$:

$$\frac{1}{2\gamma} \|\omega\|_2^2 = \min_{\mu \in \mathbb{R}_+^D} \left\{ \frac{\|\mu\|_1}{2\gamma} : \omega[i]^2 \leq \mu[i], \forall i \in [D] \right\}.$$

Therefore, Problem (3) is equivalent to the following extended Big- M formulation:

$$\min_{\omega, c, s, \mu} \frac{1}{n} \sum_{i=1}^n \left[b \cdot \max\{0, y_i - \omega^\top x_i - c\} + h \cdot \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{\|\mu\|_1}{2\gamma}, \quad (4a)$$

$$\text{s.t.} \quad \sum_{j \in [D]} s[j] \leq d, \quad s \in \{0, 1\}^D, \quad (4b)$$

$$(\omega[j], s[j], \mu[j]) \in \mathcal{S}_j, \quad \forall j \in [D], \quad (4c)$$

where for each coordinate $j \in [D]$, define the set

$$\mathcal{S}_j := \left\{ (\omega[j], s[j], \mu[j]) \in \mathbb{R} \times \{0, 1\} \times \mathbb{R}_+ : \omega[j]^2 \leq \mu[j], |\omega[j]| \leq M[j] s[j] \right\}.$$

Günlük and Linderoth [18] have characterized the convex hull of $\mathcal{S}_j$ as

$$\text{Conv}(\mathcal{S}_j) = \left\{ (\omega[j], s[j], \mu[j]) \in \mathbb{R} \times [0, 1] \times \mathbb{R}_+ : \omega[j]^2 \leq \mu[j]s[j], |\omega[j]| \leq M[j]s[j] \right\}.$$

By convexifying the constraint (4c), we can obtain the perspective reformulation of (3).

PROPOSITION 3 (Perspective Reformulation). *Under the same condition as in Proposition 2, Problem (3) is equivalent to*

$$\begin{aligned} \min_{\omega, c, s, \mu} \quad & \frac{1}{n} \sum_{i=1}^n \left[b \cdot \max\{0, y_i - \omega^\top x_i - c\} + h \cdot \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{1}{2\gamma} \sum_{j \in [D]} \mu[j], \\ \text{s.t.} \quad & \sum_{j \in [D]} s[j] \leq d, \quad s \in \{0, 1\}^D, \\ & (\omega[j], s[j], \mu[j]) \in \text{Conv}(\mathcal{S}_j), \quad \forall j \in [D]. \end{aligned} \tag{Perspective}$$

Since the constraint $\omega[j]^2 \leq \mu[j]s[j]$ is representable by a rotated second-order cone when $s[j] \geq 0$ and $\mu[j] \geq 0$, Problem (Perspective) is a mixed-integer second-order cone program (MISOCP). The following theorem compares the Big- M and perspective formulations. Its proof follows techniques in [38] and is provided in Appendix EC.2.

THEOREM 1 (Strength of Perspective Reformulation). (I) *The mixed-integer formulations (4) and (Perspective) are equivalent, and hence have the same optimal value;*
 (II) *Let $V_{\text{rel}}^{\text{Big-M}}$ and $V_{\text{rel}}^{\text{Pers}}$ denote the optimal values of the continuous relaxations of (Big-M) and (Perspective), respectively. It always holds that $V_{\text{rel}}^{\text{Big-M}} \leq V_{\text{rel}}^{\text{Pers}}$.*

REMARK 3 (PERSPECTIVE REFORMULATION IS STRICTLY STRONGER). We provide an example to show that the perspective reformulation is strictly stronger. Take $D = 2, d = 1, b = h = 1, n = 4, M = 2, \gamma = 1$. Assume that one has access to four data samples: $\{(e_1, 1), (-e_1, -1), (e_2, 1), (-e_2, -1)\}$, where e_1, e_2 denote the unit vectors in $\mathbb{R}^2$. In this case, it is easy to verify $V_{\text{rel}}^{\text{Big-M}} = 0.75$ and $V_{\text{rel}}^{\text{Pers}} = 0.875$. Hence, this problem instance ensures that $V_{\text{rel}}^{\text{Big-M}} < V_{\text{rel}}^{\text{Pers}}$.

In summary, perspective reformulation is preferred to the Big- M method in two ways. First, Theorem 1 together with Remark 3 explains that the perspective reformulation is more computationally efficient. Since its continuous relaxation is tighter than that of the Big- M formulation, a branch-and-bound solver typically explores fewer nodes and converges faster

in practice. Second, it is difficult to derive valid and tight Big- M constants for solving both formulations, and using overly conservative Big- M constants makes the formulations numerically ill-conditioned to solve. Even if we drop the constraints $|\omega[j]| \leq M[j]s[j], j \in [D]$ in Problem (Perspective), it is still a valid reformulation of Problem (3) and computationally tractable, leading to the following Big- M -free reformulation:

$$\begin{aligned}
& \min_{\omega, c, s, \mu} \quad \frac{1}{n} \sum_{i=1}^n \left[b \max\{0, y_i - \omega^\top x_i - c\} + h \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{1}{2\gamma} \sum_{j \in [D]} \mu[j] \\
& \text{s.t.} \quad \sum_{j \in [D]} s[j] \leq d, \quad s \in \{0, 1\}^D, \\
& \quad \omega[i]^2 \leq \mu[i]s[i], i \in [D], \quad \mu \in \mathbb{R}_+^D.
\end{aligned}
\tag{Big- M -free-Perspective}$$

4. Scalable Approximation Algorithms

In this section, we study two scalable approximation algorithms to obtain near-optimal solutions to Problem (3) and build their performance guarantees.

4.1. Continuous Relaxation with Randomized Rounding

Our first approach is motivated by the MISOCP (Big- M -free-Perspective). Its continuous relaxation replacing the binary constraint $s \in \{0, 1\}^D$ with $s \in [0, 1]^D$ becomes a convex program, and can be efficiently solved using off-the-shelf solvers. Then, we estimate the support of sparse regression coefficients using a randomized rounding scheme. See Algorithm 1 for detailed implementation.

We provide performance guarantees for this computationally efficient algorithm below. Let $\mathbf{X} = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times D}$ be the data matrix. For $k \in [D]$, define

$$\theta_k := \max_{|S|=k} \sigma_{\max}(\mathbf{X}_S \mathbf{X}_S^\top),$$

where $\sigma_{\max}(\cdot)$ denotes the maximum singular value. Due to the randomized rounding procedure, the support size of the estimated regression coefficient may not satisfy the sparsity budget. The following theorem provides the bi-criteria approximation guarantee of Algorithm 1. It proves that with high probability, the selected support is only moderately larger than d , while the objective value is within an additive tolerance of the optimal value.

Algorithm 1 Continuous Relaxation with Randomized Rounding for Problem (3)

Input: Data $\{(x_i, y_i)\}_{i=1}^n$, sparsity budget d , regularization parameter γ

- 1: Solve the continuous relaxation of (Big- M -free-Perspective) and obtain an optimal solution $(\hat{\omega}, \hat{c}, \hat{s}, \hat{\mu})$
 - 2: Initialize set $S = \emptyset$
 - 3: **for** $i = 1, \dots, D$ **do**
 - 4: Sample a uniform number U on $[0, 1]$.
 - 5: Update $S = S \cup \{i\}$ if $U \leq \hat{s}[i]$.
 - 6: **end for**
 - 7: Obtain $(\tilde{\omega}, \tilde{c})$ as the optimal solution to $\arg \min_{\omega, c} \left\{ F(\omega, c) : \omega[i] = 0 \text{ if } i \notin S \right\}$.
 - 8: **return** $(\tilde{\omega}, \tilde{c})$ as the regression coefficient estimator
-

THEOREM 2 (Approximation Guarantee of Algorithm 1). Given $\delta \in (0, 1)$ and $\epsilon > 0$. Let $(\tilde{\omega}, \tilde{c})$ be the output from Algorithm 1. Suppose the regularization coefficient

$$\frac{1}{2\gamma} \geq \frac{\log(2n/\delta)(h \vee b)^2 \theta_1}{6n\epsilon} + \frac{(h \vee b)^2 \sqrt{2\theta_1 \theta_d \log(2n/\delta)}}{4n\epsilon}.$$

Then, with probability at least $1 - \delta$, it holds that

$$\|\tilde{\omega}\|_0 \leq \left(1 + \sqrt{\frac{2 \log(2/\delta)}{d}} + \frac{\log(2/\delta)}{d} \right) d, \quad F(\tilde{\omega}, \tilde{c}) \leq \text{optval}(3) + \epsilon.$$

The proof of Theorem 2 is provided in Appendix EC.3. The proof step to bound the sparsity level of $\tilde{\omega}$ follows from the usage of concentration inequality discussed in [38, Theorem 5]. The reference therein implicitly assumed that $\log(2/\delta) \leq d/3$. Our concentration result refines the concentration result and does not impose this assumption. The support-size bound in Theorem 2 depends on the target sparsity d , but not on the data dimension D . The objective performance guarantee requires that the regularization parameter $\frac{1}{2\gamma}$ is not too small: its lower bound is related to the dimension D through the maximum singular values θ_1 and θ_d from the design matrix. This assumption is not quite restrictive since adding ℓ_2 -norm regularization ensures the estimated solution is more robust than unregularized models [8].

4.2. Greedy Algorithm

Greedy algorithm is a common heuristic for many variable selection problems. To develop a greedy algorithm for solving (3), we rewrite it as a combinatorial formulation. Indeed, by

introducing a subset $S \subseteq [D]$, Problem (3) is equivalent to

$$\min_{S \subseteq [D], |S| \leq d} \left\{ \min_{\omega \in \mathbb{R}^D, c \in \mathbb{R}} F(\omega, c) : \omega[i] = 0 \text{ if } i \notin S \right\}. \quad (5)$$

For a fixed support S , the inner subproblem of (5) is a convex program. Using strong duality theory, this subproblem has an equivalent dual representation that depends on S only through submatrix $\mathbf{X}_S$.

THEOREM 3 (Minimax Reformulation of (5)). *Problem (5) has the same optimal value as the following minimax optimization problem:*

$$\min_{S \subseteq [D], |S| \leq d} \left\{ G(S) := \max_{z[i] \in [-b, h], i \in [n], \sum_{i=1}^n z[i] = 0} -z^\top \left(\frac{\gamma}{2n^2} \mathbf{X}_S \mathbf{X}_S^\top \right) z - \frac{1}{n} z^\top y \right\}. \quad (6)$$

Based on Theorem 3, Greedy algorithm proceeds by iteratively finding a candidate index such that the objective G is minimized for current subset including the new index. See the detailed procedure in Algorithm 2.

Algorithm 2 Greedy Algorithm for Problem (6)

Input: Data $\{(x_i, y_i)\}_{i=1}^n$, sparsity budget d , regularization parameter γ

- 1: Initialize $S_0 = \emptyset$
 - 2: **for** $k = 0, 1, \dots, d-1$ **do**
 - 3: Choose $j_k \in \arg \min_{j \in [D] \setminus S_k} G(S_k \cup \{j\})$
 - 4: Update $S_{k+1} = S_k \cup \{j_k\}$
 - 5: **end for**
 - 6: Obtain $(\tilde{\omega}^{\text{gr}}, \tilde{c}^{\text{gr}})$ as the optimal solution to $\arg \min_{\omega, c} \left\{ F(\omega, c) : \omega[i] = 0 \text{ if } i \notin S_d \right\}$.
 - 7: **return** $(\tilde{\omega}^{\text{gr}}, \tilde{c}^{\text{gr}})$
-

By Theorem 3, it can be shown that the set function $G(S)$ is non-increasing in S , and the objective of iterates in Algorithm 2 is monotonically decreasing. It is an open question on the approximation guarantees of the final iterate from the greedy algorithm, as the set function G does not satisfy the submodular assumption commonly used in literature. Inspired by [38], the greedy approach can also be integrated with Algorithm 1. Specifically, given threshold $\delta > 0$ and $\hat{\omega}$ obtained from continuous relaxation, define $\mathcal{C} = \{i : |\hat{\omega}[i]| > \delta\}$ as the effective support obtained from continuous relaxation. We can apply the greedy algorithm based on $\mathcal{C}$ instead of $[D]$ and can reduce the number of candidate variables. This algorithm works well especially when $\hat{\omega}$ is sparse.

5. Statistical Performance Guarantees

In this section, we investigate statistical performance guarantees for our proposed variable selection framework. Assume the observed data samples $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. copies of (X, Y) , where

$$Y = (\omega^*)^\top X + c^* + \varepsilon. \quad (7)$$

Here, $\omega^* \in \mathbb{R}^D$ is a fixed d -sparse vector with $S^* = \text{supp}(\omega^*)$, and $|S^*| = d$, and $c^* \in \mathbb{R}$ is a fixed intercept parameter. The distribution of the noise ε is independent of X . We analyze the estimator from the optimization problem

$$(\hat{\omega}_n, \hat{c}_n) = \arg \min_{\omega, c: \|\omega\|_0 \leq d} \left\{ \frac{1}{n} \sum_{i=1}^n \left[b \max\{0, y_i - \omega^\top x_i - c\} + h \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{\|\omega\|_2^2}{2\gamma_n} \right\},$$

where the regularization parameter γ_n is allowed to vary with the sample size n . We impose the following technical assumptions throughout our analysis.

ASSUMPTION 1 (Structure of Data). (I) *The distribution of ε satisfies $F_\varepsilon(0) := \mathbb{P}(\varepsilon \leq 0) = \tau$, where $\tau = b/(b+h)$.*

(II) *The random variable ε satisfies $\mathbb{E}|\varepsilon| < \infty$.*

(III) *The density of the noise exists and is denoted as $f_\varepsilon(\cdot)$. There exists r_0 such that for all $|u| \leq r_0$, it holds that $f_\varepsilon(u) \geq \underline{f} > 0$.*

(IV) *The augmented covariate $\tilde{X} = (X^\top, 1)^\top$ is bounded almost surely: $\|\tilde{X}\|_2 \leq K$.*

(V) *There exists $\kappa > 0$ such that for each $u \in \mathbb{R}^D, t \in \mathbb{R}, \|u\|_0 \leq 2d$, it holds that*

$$\mathbb{E}[(u^\top X + t)^2] \geq \kappa(\|u\|_2^2 + t^2).$$

It is noteworthy that Assumption 1(I) is not restrictive. Indeed, (7) can be reformulated as $Y = (\omega^*)^\top X + \tilde{c}^* + \tilde{\varepsilon}^*$ for constant $\tilde{c}^* = c^* + F_\varepsilon^{-1}(\tau)$ and random noise $\tilde{\varepsilon}^* = \varepsilon - F_\varepsilon^{-1}(\tau)$. Such a scalar transformation always ensures Assumption 1(I) holds. Assumption 1(II) is a sufficient condition that ensures the optimal newsvendor risk

$$\mathbb{E}_{(X,Y)}[b \max(Y - (\omega^*)^\top X - c^*, 0) + h \max((\omega^*)^\top X + c^* - Y, 0)] = (b+h)\mathbb{E}_\varepsilon[\rho_\tau(\varepsilon)]$$

is finite. Assumption 1(III) imposes local regularity of the noise density around the target quantile. Assumption 1(IV) imposes the boundedness of the feature vector. Assumption 1(V)

is a restricted eigenvalue condition, which ensures identifiability over sparse directions. These assumptions are commonly used in sparse statistical analysis [35].

We first establish a finite-sample estimation-error bound regarding the estimator $(\hat{\omega}_n, \hat{c}_n)$ in Theorem 4. It reveals that under proper scaling of the ridge parameter, the estimator converges to the ground truth with a rate crucially dependent on the sparsity level instead of the data dimension. The estimation error vanishes quickly as the sample size n increases.

THEOREM 4 (Estimation Error Bound). *Let $\delta \in (0, 1)$ and suppose Assumption 1 holds. Set the ridge parameter $\frac{1}{2\gamma_n} \lesssim \sqrt{\frac{d \log(eD/d)}{n}}$, then with probability at least $1 - \delta$, it holds that*

$$\left\| \begin{pmatrix} \hat{\omega}_n - \omega^* \\ \hat{c}_n - c^* \end{pmatrix} \right\|_2 = \mathcal{O} \left(\sqrt{\frac{d \log(eD/d) + \log(1/\delta)}{n}} \right),$$

where $\mathcal{O}(\cdot)$ hides the constant depending only on $b, h, r_0, \underline{f}, \kappa, K, \|\omega^*\|_2$.

The proof of Theorem 4 is provided in Appendix EC.4. Building on Theorem 4, we further show the out-of-sample performance and support recovery guarantees of our obtained estimator. Given regression coefficients (ω, c) , define its population newsvendor risk

$$\mathcal{R}(\omega, c) = \mathbb{E}_{(X,Y)} [b \max(Y - \omega^\top X - c, 0) + h \max(\omega^\top X + c - Y, 0)]. \quad (8)$$

Our out-of-sample performance guarantee in Corollary 1 quantifies the gap between the population newsvendor risk of $(\hat{\omega}_n, \hat{c}_n)$ and the optimal risk, which vanishes in the order of $d \log(eD/d)/n$.

COROLLARY 1 (Out-of-Sample Performance Guarantee). *Under the same setup as in Theorem 4, with probability at least $1 - \delta$, it holds that*

$$0 \leq \mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \min_{\|\omega\|_0 \leq d, c} \mathcal{R}(\omega, c) \leq \mathcal{O} \left(\frac{d \log(eD/d) + \log(1/\delta)}{n} \right),$$

where $\mathcal{O}(\cdot)$ hides the constant depending only on $b, h, r_0, \underline{f}, \kappa, K, \|\omega^*\|_2$.

The support recovery guarantee is provided in Corollary 2, whose proof is provided in Appendix EC.6 and uses the fact that $\hat{\omega}_n$ converges to the true sparse vector ω^* , and so is its support.

COROLLARY 2 (Support Recovery Guarantee). *Under the same setup as in Theorem 4, and assume the sample size n is sufficiently large so that with probability at least $1 - \delta$, it holds that*

$$\left\| \begin{pmatrix} \hat{\omega}_n - \omega^* \\ \hat{c}_n - c^* \end{pmatrix} \right\|_2 < \min_{j \in S^*} |\omega^*[j]|.$$

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

$$\hat{\omega}_n[i] \neq 0, \quad \hat{\omega}_n[j] = 0, \quad \forall i \in S^*, \quad j \notin S^*.$$

6. Numerical Study

In this section, we evaluate the performance of the proposed methods. We generate synthetic data from the linear model $Y = (\omega^*)^\top X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(5, \sigma^2)$. The feature vector X follows a multivariate Gaussian distribution with mean zero and covariance matrix Σ , where $\Sigma[i, j] = 0.5^{|i-j|}$, $i, j \in [D]$. Although our statistical analysis assumed that the feature vector has bounded support, we use Gaussian-distributed feature vector with unbounded support for numerical evaluation. We further perform scalar transformation to ensure Assumption 1(I) holds. The first d entries of ω^* are nonzero and are specified as $\omega^*[i] = (-1)^{i+1} \left(1 - \frac{i-1}{d}\right)$, $i \in [d]$. We set the unit shortage and holding costs to $b = 2$ and $h = 1$, respectively.

For baseline comparison, our experiment includes the following methods: (I) the exact algorithms based on the Big- M and perspective formulations in Section 3; (II) the continuous-relaxation algorithm described in Algorithm 1; (III) the greedy algorithm described in Algorithm 2; and (IV) an ℓ_1 -based baseline obtained by replacing the ℓ_0 -constraint in Problem (3) with an ℓ_1 -regularized formulation. For each algorithm, we impose a time limit of 1800 seconds. The mixed-integer optimization algorithms terminate once the relative optimality gap falls below $1e-4$. Unless otherwise specified, the ridge and Lasso tuning parameters are selected by cross-validation. We measure the performance support recovery using false discovery proportion (FDP) and non-discovery proportion (NDP), defined in [2]:

$$\text{FDP}(S) = \frac{|S \setminus S^*|}{|S|}, \quad \text{NDP}(S) = \frac{|S^* \setminus S|}{|S^*|}, \quad (9)$$

where S^* denotes the true support and S denotes the support estimated by a given algorithm. Smaller values of FDP and NDP indicate better support recovery performance.

6.1. Comparison of exact algorithms

We first compare the two exact algorithms in Section 3 under different configurations of (D, n, d) . Throughout this subsection, we fix $\gamma = 50$, $M = 100$, and $\sigma^2 = 1$. Performance is evaluated in terms of the objective value, computational time, and FDP. Table 1 reports the computational results for solving (Big- M) and (Perspective) via branch-and-bound. Overall, the perspective formulation outperforms the Big- M formulation on most instances, and the advantage becomes more pronounced as the ambient dimension D increases. This observation is consistent with Theorem 1, which shows that (Perspective) provides a stronger continuous relaxation. When $D \in \{1000, 2000\}$, both approaches often reach the prescribed time limit, indicating that solving the sparse newsvendor problem to global optimality can be computationally demanding in large-scale settings. We also observe that the FDP generally decreases as D increases in these experiments. A possible explanation is that the regularization parameter γ is not optimally tuned such that the FDP values for different instances are not comparable.

Table 1 Computational results of exact algorithms. Here we fix cardinality $d = 50$. NA indicates the algorithm fails to find any feasible solution within time limit.

D	n	Big- M				Perspective			
		Objective	Gap (%)	Time (s)	FDP (%)	Objective	Gap (%)	Time (s)	FDP (%)
100	500	10.71	3.77	1800.00	52.00	10.65	9.69e-3	11.33	42.00
	1000	10.52	1.02	1800.00	48.00	10.45	9.94e-3	50.79	40.00
	5000	11.06	0.38	1800.00	26.00	10.91	0.14	1800.00	24.00
500	500	12.38	100	1800.00	12.00	10.01	0.56	1800.00	9.70
	1000	10.48	36.71	1800.00	11.00	10.11	1.37	1800.00	9.70
	5000	10.77	6.85	1800.00	7.50	10.71	1.39	1800.00	5.20
1000	500	9.55	100.00	1800.00	5.15	9.17	1.44	1800.00	4.85
	1000	11.54	100.00	1800.00	7.63	10.34	1.44	1800.00	4.15
	5000	10.81	15.52	1800.00	4.26	10.74	3.17	1800.00	3.68
2000	500	12.01	100	1800.00	6.41	9.90	1.99	1800.00	2.46
	1000	12.98	100	1800.00	5.51	9.56	4.34	1800.00	2.41
	5000	NA	NA	1800.00	NA	11.37	41.04	1800.00	2.46

6.2. Results on Support Recovery

Next, we compare the variable-selection performance of all methods. The noise variance σ^2 is chosen so that $\frac{\text{Var}((\omega^*)^\top X)}{\text{Var}(\varepsilon)} = 10$, that is, the signal-to-noise ratio (SNR) is 10.

We first consider the case in which the sparsity budget matches the true support size and is fixed at 10. In this setting, FDP and NDP coincide, so we report only FDP. Figures 1 and 2

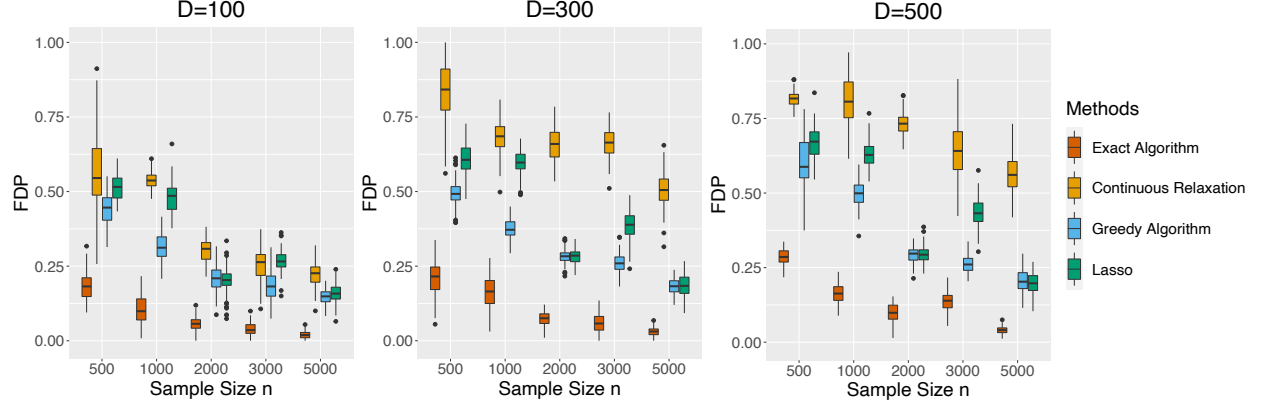


Figure 1 FDP with different choices of sample size n and each subplot represents choices of dimension D .

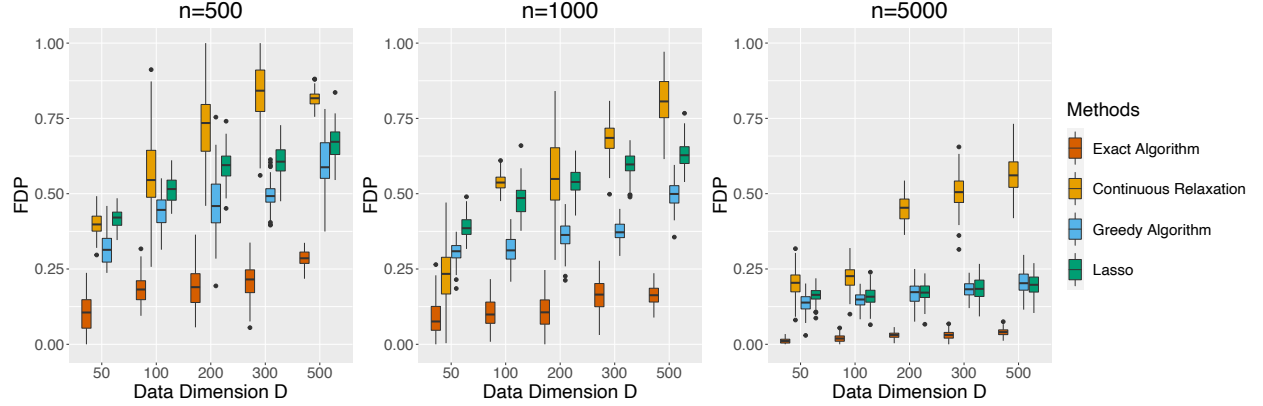


Figure 2 FDP with different choices of dimension D and each subplot represents choices of sample size n .

show the FDP values under different combinations of (n, D) . The exact algorithms achieve the best support recovery performance, especially in the high-dimensional and small-sample regime. Among the scalable baselines, the greedy algorithm generally outperforms both the continuous-relaxation and Lasso baselines.

We next fix the true support size at 10 and vary the sparsity budget to compare the methods in terms of both FDP and NDP. Figure 3 shows the results for $(D, n) \in \{(100, 500), (100, 2000)\}$. The black dashed line represents the ground-truth optimum of these metrics. As shown in the figure, exact algorithm remains close to the oracle benchmark across a wide range of sparsity budgets.

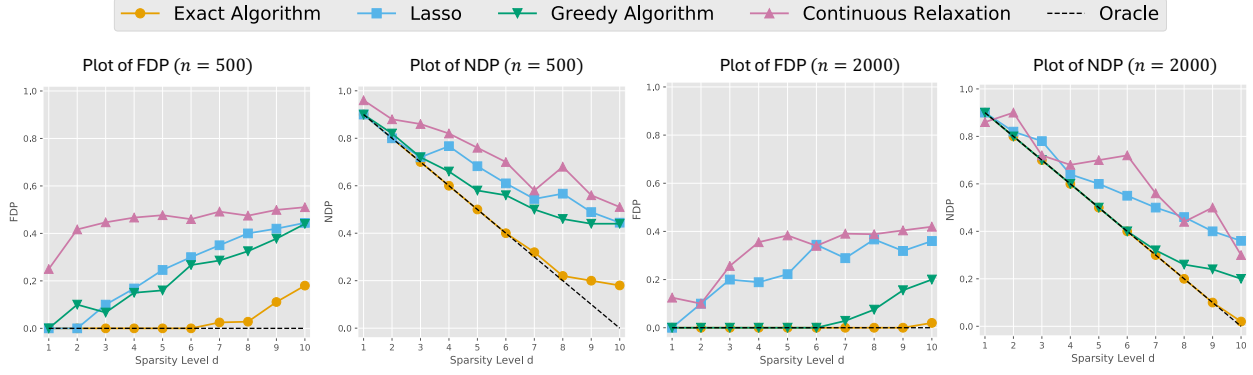


Figure 3 Comparison in terms of FDP and NDP for $n \in \{500, 2000\}$ and $D = 100$.

6.3. Results on Parameter Estimation

We examine the performance of parameter estimation for various methods on problem instances where $D = 100, n \in \{k \cdot 100, k \in [10]\}$, and $\text{SNR} \in \{0.1, 1, 10\}$. The ℓ_1 or ℓ_2 regularization parameters are tuned using cross-validation. Figure 4 summarizes the numerical results, with the sample size n on the x -axis and the logarithm of relative error on the y -axis. The relative error is defined as $|\hat{c} - c^*|/|c^*|$ or $\|\hat{\omega} - \omega^*\|_2/\|\omega^*\|_2$, where $(\hat{\omega}, \hat{c})$ denotes the estimated parameters and (ω^*, c^*) the ground truth. All box plots are generated using 5 independent trials. The results indicate that, for all methods, estimation accuracy improves as the sample size n increases. In particular, the exact and greedy algorithms consistently outperform the continuous relaxation and Lasso-based approaches in estimating the weight vector ω^* , especially in the small-sample and low-SNR regimes.

6.4. Results on Out-of-Sample Performance

We examine the out-of-sample performance of various methods using the population newsvendor risk defined in (8), which is accurately estimated via sample average approximation with 10^5 testing samples. The experimental setup follows that of the previous subsection, except that we also examine different choices of the sparsity level d . Recall the ground truth weight parameter ω^* is fixed with support size $|S^*| = 10$. For the Lasso method, the ℓ_1 regularization parameter is fine-tuned to achieve the desired sparsity level in the estimated parameter. Figure 5 summarizes the numerical results, with all error bars generated from 5 independent trials. The left-hand-side subplots display the sample size n on the x -axis. These results show that all methods improve with increasing n , while the exact and



Figure 4 Comparison in terms of logarithm of relative errors for different (n, D, SNR) .

greedy algorithms consistently yield lower out-of-sample operational costs compared to the continuous relaxation and Lasso-based approaches. The right-hand-side subplots present the sparsity level d on the x -axis. When the true sparsity level $|S^*|$ is underestimated ($d < |S^*|$), the exact and greedy algorithms dominate the others in terms of performance. When it is overestimated, the out-of-sample performance tends to stabilize. Together with the results in Sections 6.2 and 6.3, our numerical findings demonstrate that the variable selection framework delivers satisfactory parameter estimation and out-of-sample performance while using substantially fewer covariates and yielding interpretable decisions.

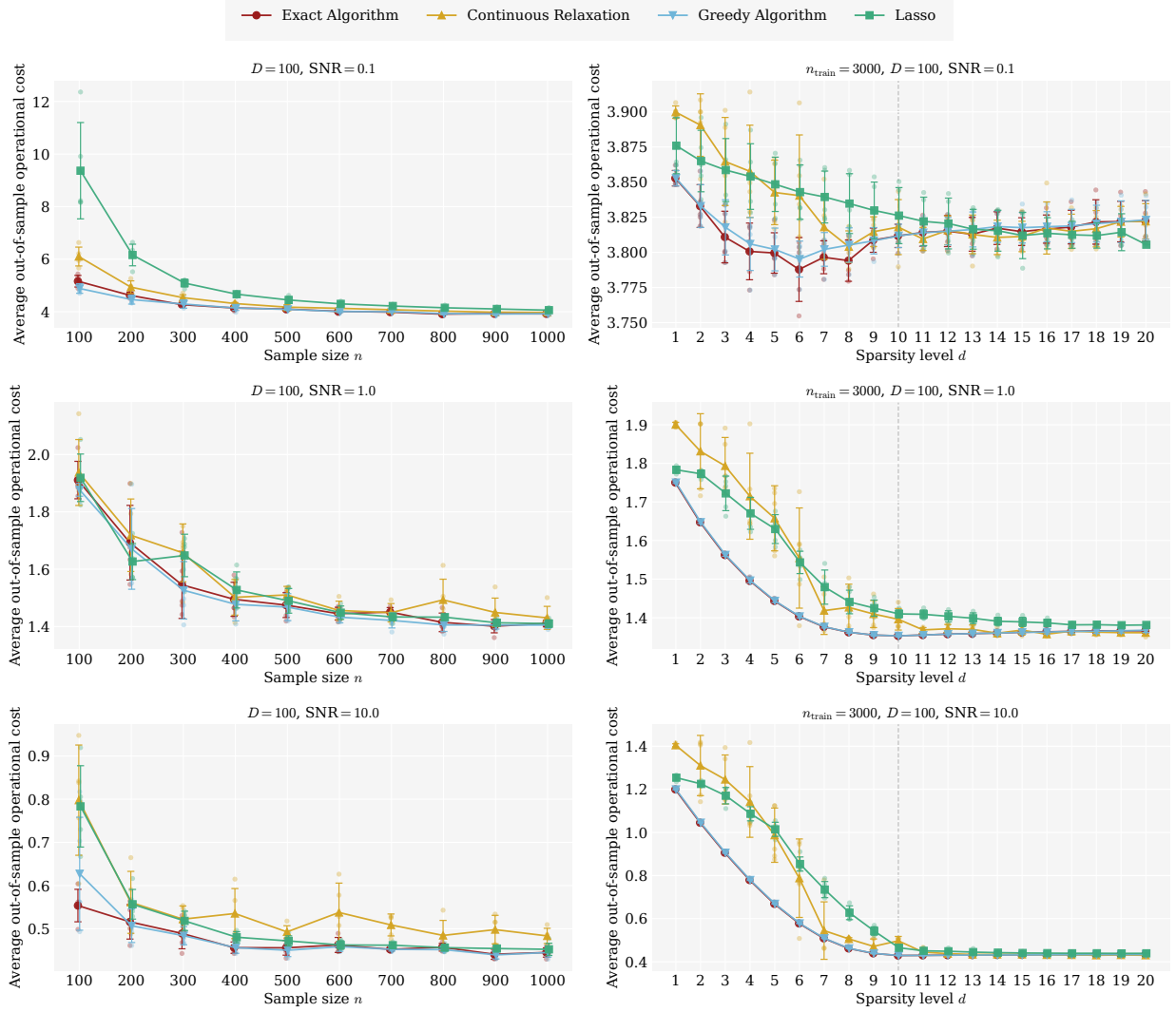


Figure 5 Comparison in terms of out-of-sample operational cost for different (n, d, SNR) .

7. Real-world case study

In this section, we evaluate the performance of various approaches using retail sales data from Kaggle[17]. The real-world dataset is provided by Corporación Favorita, an Ecuadorian-based grocery retailer that sells over 200,000 products. During preprocessing, we focus on hardware products and define the response variable Y as the sum of weekly unit sales across all four hardware product types. The covariate vector X is constructed using standard feature transformations, including one-hot encoding and normalization. The data are restricted to the time period from 2012-12-31 to 2013-12-30. The raw features in the dataset include: store number, city index, week index, month index, weekly averaged oil price, weekly pro-

motion intensity, weekly transaction volume, holiday indicator, weekly number of hardware SKUs sold, and weekly number of active sales days. In addition, we incorporate the following two-way interactions as suggested in Chang et al. [9]: (i) holiday indicator with weekly promotion intensity; (ii) store number with weekly averaged oil price; (iii) month index with city index; (iv) month index with store number; and (v) holiday indicator with store number. After preprocessing, the final dataset consists of $n = 2239$ samples and $D = 1157$ dimensions. We split the data into 70% training and 30% testing sets across 20 independent trials. The training data are used to estimate the weight vector and intercept, while the testing data are used for evaluating the (unit) out-of-sample performance.

Table 2 Out-of-sample operational cost and its standard deviation with different choices of penalty coefficient γ and sparsity level d (bold-faced values represent the lowest cost under the same γ)

method	$\gamma = 0.5$		$\gamma = 1$		$\gamma = 5$		$\gamma = 10$	
	mean	std	mean	std	mean	std	mean	std
Full model	4.5359	0.1709	4.0300	0.1468	3.7305	0.1436	3.7678	0.1769
Exact method ($d = 50$)	4.9280	0.1842	4.2086	0.1672	3.2440	0.1154	3.2073	0.1246
Exact method ($d = 100$)	4.7706	0.1840	4.0875	0.1639	3.3469	0.1289	3.3499	0.1540
Exact method ($d = 200$)	4.6522	0.1748	4.0460	0.1507	3.4718	0.1363	3.5306	0.1586
Exact method ($d = 500$)	4.5707	0.1723	4.0348	0.1479	3.6941	0.1462	3.7306	0.1793
Exact method ($d = 800$)	4.5374	0.1696	4.0281	0.1475	3.7296	0.1435	3.7670	0.1757

We report the out-of-sample performance of the different methods in Table 2. Baseline approaches include the full model without variable selection, as well as our hard-cardinality-constrained model (3) under various sparsity levels d . An ℓ_2 regularization term with different coefficients γ is added into all models to mitigate overfitting. The numerical results indicate that as the sparsity budget d increases, the out-of-sample performance of our variable selection framework approaches that of the full model. Moreover, with an appropriate choice of the ℓ_2 regularization parameter and sparsity budget, our framework not only yields interpretable decisions but also attains the lowest out-of-sample operational costs among all methods considered.

We report the top 20 selected features under different configurations of (γ, d) in Tables 3 and 4. The features are ranked according to their averaged absolute weight values in 20 independent trials. From these selected features, we draw the following operational insights: (I) The number of days per week on which customers purchase hardware products (n_days) and the weekly number of active SKUs (n_items) emerge as the most persistent features. A possible explanation is that customer demand is largely driven by weekend shopping habits and

Table 3 Top 20 features selected using the exact algorithm with $d = 100$ on the real-world data, where the entries with x represent the interaction between two features

Rank	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 5$	$\gamma = 10$
1	n_days	n_days	n_days	n_days
2	n_items	n_items	n_items	n_items
3	transactions	cluster_5_x_oil	cluster_5_x_oil	cluster_5
4	store_44	store_44_x_oil	store_44_x_oil	store_44
5	cluster_5	cluster_5	cluster_5	store_44_x_oil
6	store_44_x_oil	store_44	store_44	cluster_5_x_oil
7	cluster_5_x_oil	transactions	transactions	transactions
8	cluster_8	cluster_8	month_9_x_store_9	city_Cuenca
9	cluster_8_x_oil	cluster_8_x_oil	store_8_x_oil	store_3
10	type_C_x_oil	store_3	store_8	store_8
11	type_C	store_8	store_3	store_9_x_oil
12	store_3	store_8_x_oil	city_Cuenca	month_8_x_store_9
13	store_3_x_oil	store_3_x_oil	store_9	month_9_x_store_9
14	store_9	store_9	type_C	month_2_x_store_44
15	store_8	type_C_x_oil	store_9_x_oil	type_C
16	store_9_x_oil	store_9_x_oil	type_C_x_oil	store_8_x_oil
17	store_8_x_oil	type_C	store_3_x_oil	city_Cuenca_x_oil
18	city_Cuenca	holiday_x_store_44	month_8_x_store_9	month_8_x_store_44
19	holiday_x_store_44	store_41_x_oil	city_Cuenca_x_oil	store_48_x_oil
20	city_Quito	store_41	cluster_10_x_oil	store_9

Table 4 Top 20 features selected using the exact algorithm with $d = 50$ on the real-world data

Rank	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 5$	$\gamma = 10$
1	n_days	n_days	n_days	n_days
2	n_items	n_items	n_items	n_items
3	transactions	cluster_5_x_oil	cluster_5_x_oil	store_44_x_oil
4	store_44	store_44_x_oil	store_44_x_oil	cluster_5_x_oil
5	cluster_5	store_44	cluster_5	cluster_5
6	store_44_x_oil	transactions	store_44	store_44
7	cluster_5_x_oil	cluster_5	transactions	transactions
8	type_C_x_oil	type_C_x_oil	store_8	city_Cuenca
9	type_C	type_C	store_8_x_oil	store_3
10	cluster_8	cluster_8	city_Cuenca	store_8
11	cluster_8_x_oil	cluster_8_x_oil	store_3	store_8_x_oil
12	store_3	store_9	store_9_x_oil	month_9_x_store_9
13	store_3_x_oil	store_9_x_oil	city_Cuenca_x_oil	month_8_x_store_9
14	store_9	store_8	store_9	store_9_x_oil
15	store_9_x_oil	store_8_x_oil	type_C	month_2_x_store_44
16	store_8	store_3	store_3_x_oil	store_3_x_oil
17	store_8_x_oil	store_3_x_oil	month_9_x_store_9	store_9
18	city_Cuenca	holiday_x_store_44	type_C_x_oil	month_8_x_store_44
19	city_Cuenca_x_oil	city_Cuenca	cluster_8	type_C_x_oil
20	city_Quito	city_Cuenca_x_oil	store_48_x_oil	store_48_x_oil

the breadth of product assortment available in stores. (II) The historical weekly transaction volume (transactions) consistently appears as a strong predictor, as it is highly correlated with

future customer demand. (III) The location and type of the store (such as those represented by the store_3, 8, 9, 44, C, city_Cuenca) play a critical role in predicting hardware sales. (IV) The interaction terms between store 9 and the months of August and September, as well as between store 44 and February and August, suggest that the effects of store location and time on operational outcomes are heterogeneous, with certain stores exhibiting pronounced seasonal demand spikes.

8. Conclusion

In this paper, we studied variable selection for the feature-based newsvendor problem under a hard cardinality constraint on the number of selected covariates. We establish both computational algorithms and statistical performance guarantees of this framework. The numerical results illustrate the trade-off among computational effort, support-recovery accuracy, and out-of-sample operating cost.

Several directions remain open for future research. First, it is promising to extend this variable selection framework for nonlinear policy classes, such as the kernel-based decision rules [36]. Second, it is an interesting direction to design more efficient algorithms and hyperparameter selection procedures [12, 11]. Finally, extending the statistical analysis to misspecified, heavy-tailed, dependent, or distributionally robust settings would broaden the applicability of our framework.

References

- [1] Atamturk A, Gomez A (2025) Rank-one convexification for sparse regression. *JMLR* 1–50.
- [2] Bajwa WU, Mixon DG (2012) Group model selection using marginal correlations: The good, the bad and the ugly. *2012 50th Annual Allerton Conference on Communication, Control, and Computing*, 494–501.
- [3] Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.
- [4] Belloni A, Chernozhukov V (2011) ℓ_1 -penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* 39(1):82–130.
- [5] Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* 66(3):1025–1044.
- [6] Bertsimas D, Koduri N (2022) Data-driven optimization: A reproducing kernel hilbert space approach. *Operations Research* 70(1):454–471.
- [7] Bertsimas D, Pauphilet J, Van Parys B (2021) Sparse classification: a scalable discrete optimization perspective. *Machine Learning* 110(11-12):3177–3209.

-
- [8] Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* 56(3):830–857.
 - [9] Chang J, Yang L, Zhang Y, Zhou W (2025) Feature-rich, data-private: A sparse learning framework for the high-dimensional newsvendor. *SSRN* 1–98.
 - [10] Civril A, Magdon-Ismail M (2009) On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science* 410(47-49):4801–4811.
 - [11] Cory-Wright R, Gómez A (2023) Optimal cross-validation for sparse linear regression. *arXiv preprint arXiv:2306.14851*.
 - [12] Cory-Wright R, Gómez A (2025) Stability regularized cross-validation. *arXiv preprint arXiv:2505.06927*.
 - [13] Dai S (2023) Variable selection in convex quantile regression: ℓ_1 -norm or ℓ_0 -norm regularization? *European Journal of Operational Research* 305(1):338–355.
 - [14] Donti P, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems* 30.
 - [15] Dudley RM (1999) *Uniform central limit theorems*, volume 23 (Cambridge university press Cambridge).
 - [16] Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360.
 - [17] Favorita C (2017) Corporación favorita grocery sales forecasting. Kaggle, <https://www.kaggle.com/competitions/favorita-grocery-sales-forecasting/data>.
 - [18] Günlük O, Linderoth J (2011) Perspective reformulation and applications. *Mixed integer nonlinear programming*, 61–89 (Springer).
 - [19] Huang J, Horowitz JL, Ma S (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics* 36(2):587–613.
 - [20] Huber J, Müller S, Fleischmann M, Stuckenschmidt H (2019) A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research* 278(3):904–915.
 - [21] Kallus N, Mao X (2023) Stochastic optimization forests. *Management Science* 69(4):1975–1994.
 - [22] Knight K (1998) Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics* 26(2):755–770, URL <http://dx.doi.org/10.1214/aos/1028144858>.
 - [23] Koenker R, Bassett G (1978) Regression quantiles. *Econometrica* 46(1):33–50.
 - [24] Li Y, Dey SS, Xie W (2024) On sparse canonical correlation analysis. *Advances in Neural Information Processing Systems* 37:10707–10734.
 - [25] Li Y, Xie W (2024) Beyond symmetry: best submatrix selection for the sparse truncated svd: Y. li, w. xie. *Mathematical Programming* 208(1):1–50.
 - [26] Li Y, Xie W (2025) Exact and approximation algorithms for sparse principal component analysis. *INFORMS Journal on Computing* 37(3):582–602.
 - [27] Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Operations research letters* 33(4):341–348.

- [28] Notz PM, Pibernik R (2022) Prescriptive analytics for flexible capacity management. *Management Science* 68(3):1756–1775.
- [29] Oroojlooyjadid A, Snyder LV, Takáč M (2020) Applying deep learning to the newsvendor problem. *IIE Transactions* 52(4):444–463.
- [30] Qi M, Grigas P, Shen ZJ (2025) Integrated conditional estimation-optimization. *Operations Research* .
- [31] Sadana U, Chenreddy A, Delage E, Forel A, Frejinger E, Vidal T (2025) A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research* 320(2):271–289.
- [32] Serrano B, Minner S, Schiffer M, Vidal T (2024) Bilevel optimization for feature selection in the data-driven newsvendor problem. *European Journal of Operational Research* 315(2):703–714.
- [33] Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1):267–288.
- [34] Tropp JA (2010) User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* 12:389–434.
- [35] Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press).
- [36] Wang J, Dey SS, Xie Y (2023) Variable selection for kernel two-sample tests. *arXiv preprint arXiv:2302.07415* .
- [37] Xie H, Huang J (2009) SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics* 37(2):673–696.
- [38] Xie W, Deng X (2020) Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization* 30(4):3359–3386.
- [39] Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2):894–942.
- [40] Zhang Y, Gao J (2017) Assessing the performance of deep learning algorithms for newsvendor problem. *International conference on neural information processing*, 912–921.

Appendix for “Variable Selection for Feature-Based Newsvendor”

EC.1. Proof of Proposition 1

We prove the NP-hardness result by making the reduction from Exact Cover by 3-Sets (X3C) to Problem (3). The formal description of (X3C) is as follows. Let Q be a ground set consisting of $m := 3d$ elements, and $\mathcal{C} = \{C_1, \dots, C_D\}$ be the collection of subsets of Q such that $|C_i| = 3$ for $i \in [D]$. Problem (X3C) aims to determine whether there exists a subset of $\mathcal{C}$, denoted as $\mathcal{C}'$ such that every element in Q occurs exactly once in $\mathcal{C}'$.

Denote by $A \in \mathbb{R}^{m \times D}$ the incidence matrix of Problem (X3C) such that $A[i, j] = 1$ if the i -th element of Q belongs to the subset C_j and otherwise $A[i, j] = 0, \forall i, j$. We define a scalar M_0 such that $M_0 > \frac{m}{3\gamma(b \wedge h)}$. We construct a special instance of Problem (3) by taking $n = 2m$ and the i -th observation as

$$(x_i, y_i) = \begin{cases} (M_0 A[i, :]^\top, M_0) & \text{if } 1 \leq i \leq m, \\ (-M_0 A[i - m, :]^\top, -M_0) & \text{if } m + 1 \leq i \leq n. \end{cases}$$

For weight vector $\omega \in \mathbb{R}^D$, define the error vector $r = \mathbf{1} - A\omega \in \mathbb{R}^m$. For any scalar x , define $x_+ = \max\{x, 0\}$ and $x_- = \max\{-x, 0\}$. Define $b \wedge h = \min(b, h)$. Thus, the fitting error part can be bounded as

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left[b \cdot \max\{0, y_i - \omega^\top x_i - c\} + h \cdot \max\{0, \omega^\top x_i + c - y_i\} \right] \\ &= \frac{1}{2m} \sum_{i=1}^m \left[b \cdot (M_0 r[i] - c)_+ + h \cdot (M_0 r[i] - c)_- + b \cdot (-M_0 r[i] - c)_+ + h \cdot (-M_0 r[i] - c)_- \right] \\ &\geq \frac{b \wedge h}{2m} \sum_{i=1}^m \left[|M_0 r[i] - c| + |M_0 r[i] + c| \right] \\ &\geq \frac{(b \wedge h) M_0}{m} \|r\|_1, \end{aligned}$$

where the last inequality is based on the relation $|u + v| + |u - v| \geq 2|u|$.

Let (ω, c) be a feasible solution to Problem (3) with $\|\omega\|_0 \leq d$. We first show that $F(\omega, c) \geq d/(2\gamma)$.

- Assume that $\|\omega\|_2^2 \leq d$, then

$$|\mathbf{1}^\top \omega| = \left| \sum_{i: \omega[i] \neq 0} \omega[i] \right| \leq \sqrt{\sum_{i: \omega[i] \neq 0} 1} \sqrt{\sum_{i: \omega[i] \neq 0} \omega^2[i]} \leq \sqrt{d} \|\omega\|_2, \quad (\text{EC.1})$$

where the first inequality uses Cauchy-Schwarz inequality. It follows that

$$\begin{aligned} d - \|\omega\|_2^2 &\leq d - \frac{|\mathbf{1}^\top \omega|^2}{d} = (d - \mathbf{1}^\top \omega) \cdot (1 + \mathbf{1}^\top \omega / d) \\ &\leq 2(d - \mathbf{1}^\top \omega) \\ &= \frac{2}{3} \mathbf{1}^\top r \leq \frac{2}{3} \|r\|_1, \end{aligned}$$

where the first inequality is by (EC.1), the second inequality is because $|\mathbf{1}^\top \omega| \leq \sqrt{d} \|\omega\|_2 \leq d$, and the last equality is because each column of A has exactly three ones and

$$\mathbf{1}^\top r = m - \mathbf{1}^\top A\omega = m - (A^\top \mathbf{1})^\top \omega = 3d - 3(\mathbf{1}^\top \omega). \quad (\text{EC.2})$$

then, the objective

$$F(\omega, c) \geq \frac{(b \wedge h)M_0}{m} \|r\|_1 + \frac{1}{2\gamma} \left(d - \frac{2}{3} \|r\|_1 \right) \geq \frac{d}{2\gamma}, \quad (\text{EC.3})$$

where the last inequality is by the choice of M_0 .

- Otherwise, for $\|\omega\|_2^2 > d$, the objective value

$$F(\omega, c) \geq \frac{1}{2\gamma} \|\omega\|_2^2 > \frac{d}{2\gamma}.$$

We now show that Problem (X3C) can be reduced to Problem (3). Assume we solve Problem (3) with optimal value upper bounded by $\frac{d}{2\gamma}$ and denote its optimal solution as (ω, c) . This happens only if $\|\omega\|_2^2 \leq d$, and the inequalities in (EC.3) all hold with equalities. It implies that $r = 0 = \mathbf{1} - A\omega$. By (EC.2), it holds that $\mathbf{1}^\top \omega = d$. By (EC.1), it holds that $\|\omega\|_2^2 = d$. Then, all inequalities in (EC.1) always hold with equalities, and therefore ω has exactly d nonzero entries. The exact cover of (X3C) is given by $\{C_j : \omega[j] \neq 0\}$.

Conversely, suppose Problem (X3C) has an exact cover. Let $\omega \in \mathbb{R}^D$ be the associated weight vector such that $\omega[i] = 1$ if the i -th element of $\mathcal{C}$ is selected to form the cover and otherwise $\omega[i] = 0$. Then the incidence matrix A satisfies that $r := \mathbf{1} - A\omega = 0$ and ω satisfies $\|\omega\|_0 = d$ and $\|\omega\|_2^2 = d$. The objective value of the feasible solution $(\omega, 0)$ to Problem (3) equals

$$\begin{aligned} F(\omega, 0) &= \frac{1}{2m} \sum_{i=1}^m \left[b \cdot (M_0 r[i])_+ + h \cdot (M_0 r[i])_- + b \cdot (-M_0 r[i])_+ + h \cdot (-M_0 r[i])_- \right] + \frac{d}{2\gamma} \\ &= \frac{d}{2\gamma}. \end{aligned}$$

It justifies that the optimal value of Problem (3) is bounded by $d/(2\gamma)$. The proof is completed.

EC.2. Proof of Theorem 1

Both Big- M and perspective reformulation can be written as the following form of optimization:

$$\begin{aligned} \min_{\omega, c, s, \mu} \quad & \frac{1}{n} \sum_{i=1}^n \left[b \max\{0, y_i - \omega^\top x_i - c\} + h \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{1}{2\gamma} \|\mu\|_1 \\ \text{s.t.} \quad & (\omega, c, s, \mu) \in \mathcal{V}^{\text{Big-M}} \text{ or } \mathcal{V}^{\text{Pers}}, \\ & \sum_{j \in [D]} s[j] \leq d, \end{aligned} \tag{EC.4}$$

where

$$\begin{aligned} \mathcal{V}^{\text{Big-M}} &= \left\{ (\omega, s, \mu, c) \in \mathbb{R}^D \times \{0, 1\}^D \times \mathbb{R}_+^D \times \mathbb{R} : \begin{array}{l} \omega[i]^2 \leq \mu[i], i \in [D] \\ |\omega[i]| \leq M[i] s[i], i \in [D] \end{array} \right\}, \\ \mathcal{V}^{\text{Pers}} &= \left\{ (\omega, s, \mu, c) \in \mathbb{R}^D \times \{0, 1\}^D \times \mathbb{R}_+^D \times \mathbb{R} : \begin{array}{l} \omega[i]^2 \leq \mu[i] s[i], i \in [D] \\ |\omega[i]| \leq M[i] s[i], i \in [D] \end{array} \right\}. \end{aligned}$$

These two problems are equivalent because $\mathcal{V}^{\text{Big-M}} = \mathcal{V}^{\text{Pers}}$. Indeed, when the binary variable $s[i] = 0$, it holds that $\omega[i] = 0$ and the first constraints appeared in $\mathcal{V}^{\text{Big-M}}$ and $\mathcal{V}^{\text{Pers}}$ are redundant. Otherwise, the first constraints in $\mathcal{V}^{\text{Big-M}}$ and $\mathcal{V}^{\text{Pers}}$ are equivalent.

For Part (II), if considering the continuous relaxation of these two problems, the binary constraints $s[i] \in \{0, 1\}$ in $\mathcal{V}^{\text{Big-M}}$ and $\mathcal{V}^{\text{Pers}}$ are replaced by $s[i] \in [0, 1]$ for $i \in [D]$. Denoted the continuous relaxation of these two sets as $\bar{\mathcal{V}}^{\text{Big-M}}$ and $\bar{\mathcal{V}}^{\text{Pers}}$. It holds that $\bar{\mathcal{V}}^{\text{Pers}} \subseteq \bar{\mathcal{V}}^{\text{Big-M}}$ as $\omega[i]^2 \leq \mu[i] s[i]$ implies $\omega[i]^2 \leq \mu[i]$. Therefore, the optimal value of continuous relaxation of perspective reformulation is larger than that of Big- M reformulation.

EC.3. Proof of Theorem 2

For Part (I), we bound the support size of solution in Algorithm 1. Let $(\hat{\omega}, \hat{c}, \hat{s}, \hat{\mu})$ be the optimal solution to the continuous relaxation of (Perspective). Let $(\tilde{\omega}, \tilde{c}, \tilde{s})$ be the output of Algorithm 1. The support size of $\tilde{\omega}$ is bounded by the sum of random variables $N := \sum_{j \in [D]} \tilde{s}[j]$, where $\tilde{s}[j] \sim \text{Bernoulli}(\hat{s}[j])$. Its expected value

$$\mathbb{E}[N] = \sum_{j \in [D]} \hat{s}[j] \leq d.$$

For $t > 0$, we compute the moment-generating function of N :

$$\begin{aligned}
M_N(t) &= \mathbb{E}[e^{tN}] = \prod_{j \in [D]} \mathbb{E}[e^{t\hat{s}[j]}] \\
&= \prod_{j \in [D]} (1 + (\hat{s}[j]e^t - \hat{s}[j])) \leq \prod_{j \in [D]} \exp(\hat{s}[j]e^t - \hat{s}[j]) \\
&= \exp\left((e^t - 1) \sum_{j \in [D]} \hat{s}[j]\right) \leq \exp(d(e^t - 1)),
\end{aligned}$$

where the first inequality is based on the relation that $1 + x \leq e^x, \forall x$, and the last inequality is because $\sum_{j \in [D]} \hat{s}[j] \leq d$. For $\Delta > 0$, applying the Markov inequality gives

$$\begin{aligned}
\mathbb{P}(N \geq (1 + \Delta)d) &= \mathbb{P}(e^{tN} \geq e^{t(1+\Delta)d}) \leq \frac{M_N(t)}{e^{t(1+\Delta)d}} \\
&\leq \exp(d(e^t - 1) - t(1 + \Delta)d).
\end{aligned}$$

We choose t such that the upper bound above is as sharp as possible, yielding the optimal choice $t = \log(1 + \Delta)$. Then, it holds that

$$\mathbb{P}(N \geq (1 + \Delta)d) \leq \exp(-d((1 + \Delta)\log(1 + \Delta) - \Delta)).$$

For $\Delta \geq 0$, it is easy to verify that $\log(1 + \Delta) \geq \frac{2\Delta}{1+\Delta}$, and then

$$\mathbb{P}(N \geq (1 + \Delta)d) \leq \exp\left(-\frac{d\Delta^2}{2 + \Delta}\right).$$

Taking $\Delta = \frac{\xi}{2} + \sqrt{2\xi + \frac{\xi^2}{4}}$ with $\xi := \log(2/\delta)/d$ yields

$$\mathbb{P}(\|\tilde{\omega}\|_0 \geq (1 + \Delta)d) \leq \mathbb{P}(N \geq (1 + \Delta)d) \leq \frac{\delta}{2}. \quad (\text{EC.5})$$

We replace this bound with a conservative one ($\Delta = \log(2/\delta)/d + \sqrt{2\log(2/\delta)/d}$) to simplify the notation in the theorem statement.

For Part (II), we first reformulate the continuous relaxation of Problem [\(Perspective\)](#):

$$\begin{aligned}
\min_{\omega, c, s, \mu} \quad & \frac{1}{n} \sum_{i=1}^n \left[b \max\{0, y_i - \omega^\top x_i - c\} + h \max\{0, \omega^\top x_i + c - y_i\} \right] + \frac{1}{2\gamma} \|\mu\|_1 \\
\text{s.t.} \quad & (\omega, c, s, \mu) \in \mathbb{R}^D \times \mathbb{R} \times [0, 1]^D \times \mathbb{R}_+^D, \\
& \sum_{j \in [D]} s[j] \leq d, \\
& \omega[i]^2 \leq \mu[i] s[i], i \in [D].
\end{aligned} \quad (\text{EC.6})$$

The variable μ above can be eliminated by taking $\mu[i] = \frac{\omega[i]^2}{s[i]}$ (by default, we take $\mu[i] = 0$ if $\omega[i] = s[i] = 0$), yielding the reformulation

$$\begin{aligned} \min_{\omega, c, s} \quad & \frac{1}{n} \sum_{i=1}^n \left[b \max\{0, y_i - \omega^\top x_i - c\} + h \max\{0, \omega^\top x_i + c - y_i\} \right] + \sum_{i=1}^D \frac{\omega[i]^2}{2\gamma s[i]} \\ \text{s.t.} \quad & (\omega, c, s) \in \mathbb{R}^D \times \mathbb{R} \times [0, 1]^D, \\ & \sum_{j \in [D]} s[j] \leq d. \end{aligned} \quad (\text{EC.7})$$

By fixing s and taking the duality of the remaining optimization problem, we obtain the min-max reformulation of the continuous relaxation:

$$\min_{\substack{s \in [0, 1]^D, \\ \sum_{j \in [D]} s[j] \leq d}} \left\{ g(s) = \max_{z[i] \in [-b, h], i \in [n], \sum_{i=1}^n z[i] = 0} -z^\top \left(\frac{\gamma}{2n^2} X \text{Diag}(s) X^\top \right) z - \frac{1}{n} z^\top y \right\}. \quad (\text{EC.8})$$

It is also clear that $\text{optval}(\text{3}) = \min \{g(s) : s \in \{0, 1\}^D, \sum_{j \in [D]} s[j] \leq d\}$. By definition,

$$F(\tilde{\omega}, \tilde{c}) - \text{optval}(\text{3}) \leq F(\tilde{\omega}, \tilde{c}) - g(\hat{s}) = g(\tilde{s}) - g(\hat{s}), \quad (\text{EC.9})$$

where the first inequality is because $\hat{s}$ is the optimal solution obtained from continuous relaxation, and the second inequality is by strong duality.

Next, we provide the upper bound on $g(\tilde{s}) - g(\hat{s})$. Define the inner objective of $g(s)$ as

$$\mathfrak{g}(s, z) := -z^\top \left(\frac{\gamma}{2n^2} X \text{Diag}(s) X^\top \right) z - \frac{1}{n} z^\top y.$$

Let $\tilde{z}$ be the maximizer of the inner problem of $g(\tilde{s})$, then

$$\begin{aligned} g(\tilde{s}) - g(\hat{s}) &\leq \mathfrak{g}(\tilde{s}, \tilde{z}) - \mathfrak{g}(\hat{s}, \tilde{z}) \\ &= -\frac{\gamma}{2n^2} \tilde{z}^\top (X \text{Diag}(\tilde{s} - \hat{s}) X^\top) \tilde{z} \\ &\leq \frac{\gamma}{2n^2} \|\tilde{z}\|_2^2 \lambda_{\max} \left(X \text{Diag}(\hat{s} - \tilde{s}) X^\top \right) \\ &\leq \frac{\gamma(b \vee h)^2}{2n} \lambda_{\max} \left(X \text{Diag}(\hat{s} - \tilde{s}) X^\top \right), \end{aligned} \quad (\text{EC.10})$$

where the first inequality is because $\tilde{z}$ is a feasible solution (but perhaps suboptimal) to the inner problem of $g(\hat{s})$, the second inequality is based on the Cauchy-Schwarz inequality, and the last one is because $\|\tilde{z}\|_2^2 \leq n(b \vee h)^2$. It suffices to build the concentration bound on $\lambda_{\max} \left(X \text{Diag}(\hat{s} - \tilde{s}) X^\top \right) = \lambda_{\max} \left(\sum_{i \in [D]} R_i \right)$ with $R_i = (\hat{s}[i] - \tilde{s}[i]) X_{:,i} X_{:,i}^\top$. Following the similar argument as in [38, Proof of Lemma 3], it can be shown that $\mathbb{E}[R_i] = 0$, $\lambda_{\max}(R_i) \leq \theta_1$

and $\|\sum_{i \in [D]} \mathbb{E}[R_i^2]\|_2 \leq \theta_1 \theta_d$. Applying the matrix concentration bound in [34, Theorem 1.4] yields

$$\mathbb{P}\left\{\lambda_{\max}\left(X\text{Diag}(\hat{s} - \tilde{s})X^\top\right) \geq t\right\} \leq n \exp\left(-\frac{t^2}{2\theta_1\theta_d + 2/3\theta_1 t}\right). \quad (\text{EC.11})$$

Taking $t = \frac{2}{3}\theta_1 \log(2n/\delta) + \sqrt{2\theta_1\theta_d \log(2n/\delta)}$ ensures the right-hand-side of (EC.11) being bounded by $\delta/2$. Combining the inequalities in (EC.9), (EC.10), and (EC.11) yields

$$\begin{aligned} & \mathbb{P}\left\{F(\tilde{\omega}, \tilde{c}) - \text{optval}(\mathbf{3}) \leq \frac{\gamma(b \vee h)^2}{2n} \cdot \left(\frac{2}{3}\theta_1 \log(2n/\delta) + \sqrt{2\theta_1\theta_d \log(2n/\delta)}\right)\right\} \\ & \geq \mathbb{P}\left\{\lambda_{\max}\left(X\text{Diag}(\hat{s} - \tilde{s})X^\top\right) \leq \frac{2}{3}\theta_1 \log(2n/\delta) + \sqrt{2\theta_1\theta_d \log(2n/\delta)}\right\} \\ & \geq 1 - \frac{\delta}{2}. \end{aligned}$$

The choice of γ in the main statement of Theorem 2 further ensures that

$$\mathbb{P}\{F(\tilde{\omega}, \tilde{c}) - \text{optval}(\mathbf{3}) \leq \epsilon\} \geq 1 - \delta/2. \quad (\text{EC.12})$$

Applying the union bound of the concentration results in (EC.5) and (EC.12) gives the desired result.

EC.4. Proof of Theorem 4

Let us define the empirical cost

$$\mathcal{R}_n(\omega, c) = \frac{1}{n} \sum_{i=1}^n \left[b \cdot \max\{0, y_i - \omega^\top x_i - c\} + h \cdot \max\{0, \omega^\top x_i + c - y_i\} \right].$$

and the population cost

$$\mathcal{R}(\omega, c) = \mathbb{E}_{(X,Y)} \left[b \cdot \max\{0, Y - \omega^\top X - c\} + h \cdot \max\{0, \omega^\top X + c - Y\} \right].$$

We rely on the following two technical lemmas that are helpful to show the proof of Theorem 4.

LEMMA EC.1 (Global Risk Lower Bound). *Under Assumption 1, for any $\Delta := (u^\top, t)^\top \in \mathbb{R}^{D+1}$ with $\|u\|_0 \leq 2d$, it holds that*

$$\mathcal{R}(\omega^* + u, c^* + t) - \mathcal{R}(\omega^*, c^*) \geq \frac{(b+h)f\kappa r_0}{2(r_0 + K\|\Delta\|_2)} \cdot \|\Delta\|_2^2.$$

Proof. By definition,

$$\begin{aligned}
& \mathcal{R}(\omega^* + u, c^* + t) - \mathcal{R}(\omega^*, c^*) \\
&= (b + h) \mathbb{E}_{(X, Y)} [\rho_\tau(Y - (\omega^* + u)^\top X - (c^* + t)) - \rho_\tau(Y - (\omega^*)^\top X - c^*)] \\
&= (b + h) \mathbb{E}_{(\varepsilon, X)} [\rho_\tau(\varepsilon - \Delta^\top \tilde{X}) - \rho_\tau(\varepsilon)],
\end{aligned}$$

where the first equality is by the definition of quantile function $\rho_\tau(u) = u(\tau - \mathbf{1}_{\{u < 0\}})$, and the second one is based on the model assumption (7) and the expressions that $\tilde{X} = (X^\top, 1)^\top$, $\Delta = (u^\top, t)^\top$. Based on the Knight's identity [22], it holds that

$$\begin{aligned}
& \mathbb{E}_{(\varepsilon, X)} [\rho_\tau(\varepsilon - \Delta^\top \tilde{X}) - \rho_\tau(\varepsilon)] \\
&= \mathbb{E}_{(\varepsilon, X)} \left[-\Delta^\top \tilde{X}(\tau - \mathbf{1}_{\{\varepsilon < 0\}}) + \int_0^{\Delta^\top \tilde{X}} (\mathbf{1}_{\{\varepsilon \leq s\}} - \mathbf{1}_{\{\varepsilon < 0\}}) ds \right] \tag{EC.13} \\
&= \mathbb{E}_{(\varepsilon, X)} \left[\int_0^{\Delta^\top \tilde{X}} (\mathbf{1}_{\{\varepsilon \leq s\}} - \mathbf{1}_{\{\varepsilon < 0\}}) ds \right] = \mathbb{E}_X \left[\int_0^{\Delta^\top \tilde{X}} (F_\varepsilon(s) - F_\varepsilon(0)) ds \right],
\end{aligned}$$

where the second and last equalities are due to the assumption that ε and X are independent together with the definition of the cumulative density function (cdf) $F_\varepsilon(s) := \mathbb{P}(\varepsilon \leq s)$.

We provide lower bound on $\int_0^{\Delta^\top \tilde{X}} (F_\varepsilon(s) - F_\varepsilon(0)) ds$ as follows. When $|\Delta^\top \tilde{X}| \leq r_0$, by Assumption 1(III),

$$\int_0^{\Delta^\top \tilde{X}} (F_\varepsilon(s) - F_\varepsilon(0)) ds \geq \int_0^{|\Delta^\top \tilde{X}|} \underline{f}s ds = \frac{\underline{f}(\Delta^\top \tilde{X})^2}{2},$$

where the inequality is because $F_\varepsilon(s) - F_\varepsilon(0) = \int_0^s f_\varepsilon(t) dt \geq \int_0^s \underline{f} dt = \underline{f}s$ for $s \geq 0$. When $\Delta^\top \tilde{X} > r_0 > 0$, it holds that

$$\begin{aligned}
& \int_0^{\Delta^\top \tilde{X}} (F_\varepsilon(s) - F_\varepsilon(0)) ds \geq \int_0^{r_0} (F_\varepsilon(s) - F_\varepsilon(0)) ds + \int_{r_0}^{\Delta^\top \tilde{X}} (F_\varepsilon(r_0) - F_\varepsilon(0)) ds \\
& \geq \frac{\underline{f}r_0^2}{2} + (\Delta^\top \tilde{X} - r_0)\underline{f}r_0 \geq \frac{\underline{f}r_0\Delta^\top \tilde{X}}{2},
\end{aligned}$$

where the first inequality is by separating the integral on $[0, \Delta^\top \tilde{X}]$ into two parts and by the monotonicity of the cdf, the second inequality is by Assumption 1(III), and the last one is by the relation $\Delta^\top \tilde{X} > r_0$. It can be shown using the similar argument that $\int_0^{\Delta^\top \tilde{X}} (F_\varepsilon(s) - F_\varepsilon(0)) ds \geq -\underline{f}r_0\Delta^\top \tilde{X}/2$ for $\Delta^\top \tilde{X} < -r_0 < 0$. Combining these cases yields the global lower bound

$$\int_0^{\Delta^\top \tilde{X}} (F_\varepsilon(s) - F_\varepsilon(0)) ds \geq \underline{f}|\Delta^\top \tilde{X}|/2 \cdot \min\{r_0, |\Delta^\top \tilde{X}|\} \geq \frac{\underline{f}r_0(\Delta^\top \tilde{X})^2}{2(r_0 + |\Delta^\top \tilde{X}|)}.$$

Substituting this lower bound to (EC.13) yields

$$\begin{aligned} & \mathcal{R}(\omega^* + u, c^* + t) - \mathcal{R}(\omega^*, c^*) \\ & \geq \mathbb{E}_X \left[\frac{(b+h)\underline{f}r_0(\Delta^\top \tilde{X})^2}{2(r_0 + |\Delta^\top \tilde{X}|)} \right] \geq \mathbb{E}_X \left[\frac{(b+h)\underline{f}r_0(\Delta^\top \tilde{X})^2}{2(r_0 + \|\Delta\|_2 K)} \right] \geq \frac{(b+h)\underline{f}r_0 \kappa \|\Delta\|_2^2}{2(r_0 + \|\Delta\|_2 K)}, \end{aligned}$$

where the second and last inequalities leverage Assumptions 1(IV) and 1(V), respectively. $\square$

LEMMA EC.2 (Uniform Concentration Bound on Risk). *Under Assumption 1, there exists a constant C_0 depending on b, h, K such that, with probability at least $1 - \delta$, simultaneously for every $\Delta := (u^\top, t)^\top \in \mathbb{R}^{D+1}$ with $\|u\|_0 \leq 2d$, it holds that*

$$\begin{aligned} & \left| \left(\mathcal{R}_n(\omega^* + u, c^* + t) - \mathcal{R}_n(\omega^*, c^*) \right) - \left(\mathcal{R}(\omega^* + u, c^* + t) - \mathcal{R}(\omega^*, c^*) \right) \right| \\ & \leq C_0 \|\Delta\|_2 \cdot \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}}. \end{aligned}$$

Proof. The inequality for $\Delta = 0$ holds trivially. We focus on the case where $\Delta \neq 0$ instead. Define the function $\ell(y, z) = b \max\{0, y - z\} + h \max\{0, z - y\}$, and the function class

$$\mathcal{H} = \left\{ (x, y) \mapsto \frac{\ell(y, x^\top(\omega^* + u) + c^* + t) - \ell(y, x^\top \omega^* + c^*)}{\|\Delta\|_2} : \Delta \neq 0, \Delta := (u^\top, t)^\top, \|u\|_0 \leq 2d \right\}.$$

It suffices to provide concentration bound on

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{X,Y}[h(X, Y)] - \frac{1}{n} \sum_{i=1}^n h(x_i, y_i) \right|,$$

which can be shown based on the covering number argument.

Each element in the function class $\mathcal{H}$ can be indexed with two terms: $r := \|\Delta\|_2$ and $v := \frac{\Delta}{\|\Delta\|_2}$. Hence, we reformulate

$$\mathcal{H} = \left\{ (x, y) \mapsto h_{r,v}(x, y) := \frac{\ell(y, x^\top(\omega^* + rv_u) + c^* + rv_t) - \ell(y, x^\top \omega^* + c^*)}{r}, \right. \\ \left. r > 0, v = (v_u^\top, v_t)^\top \in \mathbb{R}^{D+1}, \|v\|_2 = 1, \|v_u\|_0 \leq 2d \right\}.$$

Define the empirical Rademacher complexity

$$\hat{\mathfrak{R}}_n(\mathcal{H}) := \mathbb{E}_{\sigma_i, i \in [n]} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i, y_i) \right| \middle| (x_i, y_i), i \in [n] \right],$$

where $\sigma_i, i \in [n]$ are i.i.d. Rademacher random variables. We provide upper bound on $\hat{\mathfrak{R}}_n(\mathcal{H})$ using the following procedure.

- For fixed $v \in \mathbb{R}^{D+1}$, denote the subset of $\mathcal{H}$ as

$$\mathcal{H}_v = \{h_{r,v} : r > 0\}.$$

Define the empirical L_2 pseudo-metric as

$$\mathbf{d}_n(f, g) = \left\{ \frac{1}{n} \sum_{i=1}^n \left(f(x_i, y_i) - g(x_i, y_i) \right)^2 \right\}^{1/2}.$$

Let $r_1, \dots, r_M$ with $0 < r_1 < \dots < r_M$ index an δ -packing of $\mathcal{H}_v$ under the metric $\mathbf{d}_n(\cdot)$.

Therefore,

$$(M-1)\delta^2 < \sum_{j=1}^{M-1} \mathbf{d}_n(h_{r_j,v}, h_{r_{j+1},v})^2 = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{M-1} \left(h_{r_j,v}(x_i, y_i) - h_{r_{j+1},v}(x_i, y_i) \right)^2.$$

According to the formulation of $h_{r,v}$ and by the convexity of $\ell(y, \cdot)$, it is easy to show that $h_{r,v}(x, y)$ is non-decreasing in r , and therefore

$$\begin{aligned} (M-1)\delta^2 &< \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{M-1} \left(h_{r_j,v}(x_i, y_i) - h_{r_{j+1},v}(x_i, y_i) \right)^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \left[\sum_{j=1}^{M-1} \left(h_{r_{j+1},v}(x_i, y_i) - h_{r_j,v}(x_i, y_i) \right) \right]^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left(h_{r_M,v}(x_i, y_i) - h_{r_1,v}(x_i, y_i) \right)^2 \leq 4(b \vee h)^2 K^2, \end{aligned}$$

where the last inequality uses the fact that for any (x, y) and $r > 0$,

$$|h_{r,v}(x, y)| \leq (b \vee h) \cdot \frac{|x^\top (rv_u) + rv_t|}{r} \leq (b \vee h) \|(x^\top, 1)^\top\|_2 \|v\|_2 \leq (b \vee h)K.$$

The packing number $M \leq 1 + 4(b \vee h)^2 K^2 / \delta^2$. Therefore, the covering number of $\mathcal{H}_v$ under $\mathbf{d}_n(\cdot)$ is bounded:

$$\mathcal{N}(\delta, \mathcal{H}_v, \mathbf{d}_n) \leq 1 + 4(b \vee h)^2 K^2 / \delta^2.$$

- Next, we study the covering number of the set

$$\mathcal{V} = \left\{ v = (v_u^\top, v_t)^\top \in \mathbb{R}^{D+1} : \|v\|_2 = 1, \|v_u\|_0 \leq 2d \right\}.$$

For fixed support $S \subseteq [D]$ with $|S| \leq 2d$, the δ -covering number of $\mathcal{V}_J = \left\{ v = (v_u^\top, v_t)^\top \in \mathbb{R}^{D+1} : \|v\|_2 = 1, \text{supp}(v_u) = J \right\}$ under $\|\cdot\|_2$ is at most $(1 + 2/\delta)^{2d+1}$. By taking the union of all sparse supports, the covering number

$$\mathcal{N}(\delta, \mathcal{V}, \|\cdot\|_2) \leq \left(\sum_{k=0}^{\min\{2d, D\}} \binom{D}{k} \right) (1 + 2/\delta)^{2d+1}.$$

- For fixed $r > 0$ and any two unit vectors $v = (v_u, v_t), v' = (v'_u, v'_t)$, it holds that

$$\left| h_{r,v}(x, y) - h_{r,v'}(x, y) \right| \leq (b \vee h) \cdot \frac{|x^\top (rv_u - rv'_u) + rv_t - rv'_t|}{r} \leq (b \vee h)K \|v - v'\|_2.$$

For any $h_{r,v} \in \mathcal{H}$, we select v' in the δ -cover of $\mathcal{V}$ and $h_{r',v'}$ in the δ -cover of $\mathcal{H}_{v'}$. It follows that

$$\begin{aligned} \mathbf{d}_n(h_{r,v}, h_{r',v'}) &= \left[\frac{1}{n} \sum_{i=1}^n \left(h_{r,v}(x_i, y_i) - h_{r',v'}(x_i, y_i) \right)^2 \right]^{1/2} \\ &\leq \left[\frac{2}{n} \sum_{i=1}^n \left(h_{r,v}(x_i, y_i) - h_{r,v'}(x_i, y_i) \right)^2 + \frac{2}{n} \sum_{i=1}^n \left(h_{r',v'}(x_i, y_i) - h_{r,v'}(x_i, y_i) \right)^2 \right]^{1/2} \\ &\leq [2(b \vee h)^2 K^2 \delta^2 + 2\delta^2]^{1/2} = (2 + 2(b \vee h)^2 K^2)^{1/2} \delta. \end{aligned}$$

Therefore,

$$\mathcal{N}\left((2 + 2(b \vee h)^2 K^2)^{1/2} \delta, \mathcal{H}, \mathbf{d}_n\right) \leq \mathcal{N}(\delta, \mathcal{H}_v, \mathbf{d}_n) \cdot \mathcal{N}(\delta, \mathcal{V}, \|\cdot\|_2).$$

It further implies that

$$\begin{aligned} \log \mathcal{N}(\delta, \mathcal{H}, \mathbf{d}_n) &\leq \log \left[1 + \frac{8(b \vee h)^2 K^2 (1 + (b \vee h)^2 K^2)}{\delta^2} \right] + \log \left(\sum_{k=0}^{\min\{2d, D\}} \binom{D}{k} \right) \\ &\quad + (2d+1) \log \left[1 + \frac{2\sqrt{2}(1 + (b \vee h)^2 K^2)^{1/2}}{\delta} \right] \\ &\leq C_1 \cdot \left[d \log \left(\frac{eD}{d} \right) + 1 + d \log \left(\frac{1}{\delta} \right) \right], \end{aligned}$$

where C_1 is a constant depending on b, h, K . As each element in $\mathcal{H}$ is uniformly bounded by $(b \vee h)K$, by Dudley's entropy integral bound [15], the Rademacher complexity

$$\begin{aligned} \mathfrak{R}_n(\mathcal{H}) &\leq \frac{C_2}{\sqrt{n}} \int_0^{(b \vee h)K} \sqrt{\log \mathcal{N}(\delta, \mathcal{H}, \mathbf{d}_n)} d\delta \\ &\leq C_3 \sqrt{\frac{d \log(eD/d) + 1}{n}}, \end{aligned}$$

where C_2, C_3 are some constants depending on b, h, K .

Based on the symmetrization and McDiarmid's inequalities, it holds with probability at least $1 - \delta$ that

$$\begin{aligned} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(x_i, y_i) - \mathbb{E}[h(X, Y)] \right| &\leq C_4 \sqrt{\frac{d \log(eD/d) + 1}{n}} + (b \vee h)K \sqrt{\frac{2 \log(2/\delta)}{n}} \\ &\leq C_0 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}}, \end{aligned}$$

where C_0, C_4 are some constants depending on b, h, K . Re-arranging the inequality above yields the desired result. $\square$

Finally, we provide the bound on the residual error

$$a_n := \left\| \begin{pmatrix} \hat{\omega}_n - \omega^* \\ \hat{c}_n - c^* \end{pmatrix} \right\|_2.$$

Proof of Theorem 4. As $\|\hat{\omega}_n - \omega^*\|_0 \leq 2d$, by Lemma EC.1, we obtain the lower bound

$$\mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}(\omega^*, c^*) \geq \frac{(b+h)f\kappa r_0}{2(r_0 + Ka_n)} \cdot a_n^2. \quad (\text{EC.14})$$

Next, we derive its upper bound. By the definition of $(\hat{\omega}_n, \hat{c}_n)$, it holds that

$$\mathcal{R}_n(\hat{\omega}_n, \hat{c}_n) + \frac{1}{2\gamma_n} \|\hat{\omega}_n\|_2^2 \leq \mathcal{R}_n(\omega^*, c^*) + \frac{1}{2\gamma_n} \|\omega^*\|_2^2.$$

Leveraging this relation and re-arranging the terms yields

$$\begin{aligned} & \mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}(\omega^*, c^*) \\ &= \left[\mathcal{R}_n(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}_n(\omega^*, c^*) \right] - \left[\left(\mathcal{R}_n(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}_n(\omega^*, c^*) \right) - \left(\mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}(\omega^*, c^*) \right) \right] \\ &\leq \frac{1}{2\gamma_n} \left[\|\omega^*\|_2^2 - \|\hat{\omega}_n\|_2^2 \right] - \left[\left(\mathcal{R}_n(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}_n(\omega^*, c^*) \right) - \left(\mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}(\omega^*, c^*) \right) \right]. \end{aligned}$$

The first term above can be further bounded as follows:

$$\frac{1}{2\gamma_n} \left[\|\omega^*\|_2^2 - \|\hat{\omega}_n\|_2^2 \right] \leq \frac{\langle \omega^*, \omega^* - \hat{\omega}_n \rangle}{\gamma_n} \leq \frac{\|\omega^*\|_2 a_n}{\gamma_n}.$$

By Lemma EC.2, with probability at least $1 - \delta$, the second term above can be bounded as

$$C_0 a_n \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}}.$$

Combining these two relations implies the following relation holds with probability at least $1 - \delta$:

$$\mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}(\omega^*, c^*) \leq a_n \cdot \left(\frac{\|\omega^*\|_2}{\gamma_n} + C_0 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}} \right). \quad (\text{EC.15})$$

Comparing this with (EC.14) yields

$$\frac{(b+h)f\kappa r_0}{2(r_0 + Ka_n)} \cdot a_n^2 \leq a_n \cdot \left(\frac{\|\omega^*\|_2}{\gamma_n} + C_0 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}} \right).$$

Specify $\frac{1}{2\gamma_n} \lesssim \sqrt{\frac{d \log(eD/d)}{n}}$, then there exists constant C_1 depending on $C_0, b, h, K, \|\omega^*\|_2$ such that

$$\frac{(b+h)\underline{f}\kappa r_0}{2(r_0 + \underline{K}a_n)} \cdot a_n \leq C_1 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}}. \quad (\text{EC.16})$$

Assume n is sufficiently large such that

$$C_1 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}} \leq \frac{(b+h)\underline{f}\kappa r_0}{4\underline{K}},$$

then re-arranging (EC.16) yields

$$\begin{aligned} a_n &\leq \frac{\left[C_1 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}} r_0 \right]}{\frac{(b+h)\underline{f}\kappa r_0}{2} - \left[C_1 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}} \right] \underline{K}} \\ &\leq \frac{\left[C_1 \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}} r_0 \right]}{\frac{(b+h)\underline{f}\kappa r_0}{4}} = \frac{4C_1}{(b+h)\underline{f}\kappa} \cdot \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}} \end{aligned}$$

holds with probability at least $1 - \delta$. □

EC.5. Proof of Corollary 1

By definition,

$$\min_{\|\omega\|_0 \leq d, c} \mathcal{R}(\omega, c) = \min_{\|\omega\|_0 \leq d, c} \left\{ (b+h) \mathbb{E}_{(X,Y)} [\rho_\tau(Y - (\omega^\top X + c))] \right\}.$$

By the likelihood model, for each x , $(\omega^*)^\top x + c^*$ is the τ -quantile of $Y \mid X = x$, and therefore the optimization problem above takes the optimal solution (ω^*, c^*) . Recall from the proof of Theorem 4 that for sufficiently large n , with probability at least $1 - \delta$, the following two relations hold:

$$\begin{aligned} \mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}(\omega^*, c^*) &\leq a_n C_1 \cdot \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}}, \\ a_n &\leq \frac{4C_1}{(b+h)\underline{f}\kappa} \cdot \sqrt{\frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}}. \end{aligned}$$

Combining these two relations yields

$$\mathcal{R}(\hat{\omega}_n, \hat{c}_n) - \mathcal{R}(\omega^*, c^*) \leq \frac{4C_1^2}{(b+h)\underline{f}\kappa} \cdot \frac{d \log(eD/d) + 1 + \log(2/\delta)}{n}.$$

EC.6. Proof of Corollary 2

Under the event described in Corollary 2, for each $j \in S^*$, it holds that

$$|\hat{\omega}_n[j]| \geq |\omega^*[j]| - |\hat{\omega}_n[j] - \omega^*[j]| \geq |\omega_j^*| - \left\| \begin{pmatrix} \hat{\omega}_n - \omega^* \\ \hat{c}_n - c^* \end{pmatrix} \right\|_2 > 0.$$

Therefore, $S^* \subseteq \text{supp}(\hat{\omega}_n)$. On the other hand, by the feasibility of $\hat{\omega}_n$, it holds that $|\text{supp}(\hat{\omega}_n)| \leq d = |S^*|$. This fact, together with $S^* \subseteq \text{supp}(\hat{\omega}_n)$ implies that $S^* = \text{supp}(\hat{\omega}_n)$.