Debiased Semi-Parametric Approach for Contextual Stochastic Optimization: Theory and Application to the Newsvendor Problem

Xinyuan Lyu

December 2025

Abstract

We study contextual stochastic optimization problems in which decisions are made based on auxiliary covariates and uncertain outcomes. The prevailing predict-then-optimize paradigm suffers from first-order sensitivity to estimation errors in the underlying predictive model, leading to suboptimal decisions in finite samples, especially when flexible machine learning methods are employed.

To address this issue, we propose a debiased semi-parametric framework that constructs an *orthogonalized risk functional* using ideas from double/debiased machine learning. By incorporating an explicit correction term derived from the efficient influence function, the proposed objective satisfies Neyman orthogonality with respect to nuisance parameters, thereby eliminating first-order bias transmission from prediction to optimization. We further show that cross-fitting is essential to remove stochastic variance arising from moment estimation.

Under standard regularity conditions, we establish that the excess risk of the resulting decision rule converges at a fast rate of $O(a_n^2 + a_n r_n)$, representing a quadratic improvement over the $O(a_n)$ rate of naive plug-in methods. We instantiate the framework for the contextual newsvendor problem using a deep log-linear sieve model for demand estimation, which combines the flexibility of neural networks with tractable exponential-family structure. Numerical experiments demonstrate that our approach consistently outperforms both plug-in and end-to-end baselines, validating the theoretical gains in finite samples.

Contents

1 Introduction

Contextual stochastic optimization problems, wherein a decision-maker must select an action to minimize an expected cost based on auxiliary covariates, are ubiquitous in operations research and management science. Typical applications range from inventory management with feature-dependent demand (Ban and Rudin, 2019) to personalized pricing and medical treatment selection (Bertsimas and Kallus, 2020).

A prevalent paradigm for solving such problems is the "Predict-then-Optimize" (PTO) framework, often referred to as the plug-in approach. This two-stage procedure first estimates the unknown parameters of the conditional distribution (the nuisance parameters) using historical data, and subsequently solves the optimization problem by treating these estimates as the ground truth. While conceptually intuitive and easy to implement using off-the-shelf machine learning tools, the PTO approach relies on the implicit assumption that the estimation error in the predictive stage is negligible. However, in finite-sample regimes, particularly when employing high-dimensional or non-parametric estimators, this assumption may not hold. The standard plug-in estimator typically exhibits first-order sensitivity to errors in the nuisance parameter, meaning that prediction inaccuracies can translate linearly into decision suboptimality. Consequently, simply maximizing predictive accuracy (e.g., minimizing Mean Squared Error) does not necessarily guarantee optimal decision-making performance.

To address the limitations of the separation between prediction and optimization, recent literature has proposed "End-to-End" or decision-focused learning methods, such as the Smart "Predict, then Optimize" (SPO) framework (Elmachtoub and Grigas, 2022). These approaches aim to integrate the downstream optimization cost directly into the training of the predictive model. While this direction has shown promise, it often necessitates the design of specialized surrogate loss functions or the solution of complex bilevel optimization problems. Furthermore, the statistical properties of these methods relative to well-specified probabilistic models remain a subject of ongoing theoretical investigation (Hu et al., 2022).

In this work, we revisit the probabilistic modeling approach through the lens of semi-parametric inference. Rather than modifying the training objective of the predictive model, we propose to modify the inputs to the optimization problem. Drawing on the theory of Double/Debiased Machine Learning (Chernozhukov et al., 2018), we construct an optimization objective that is locally robust to estimation errors. By ensuring that the risk functional satisfies *Neyman Orthogonality* with respect to the nuisance parameters, we aim to mitigate the bias transmission from the prediction stage to the decision stage.

This paper makes three main contributions to the literature on contextual stochastic optimization. First, we propose a general semi-parametric framework for orthogonalized stochastic optimization. The key idea is to augment the standard model-based risk with an influence-function–based correction term derived from the score equations of the nuisance parameter. The resulting objective satisfies Neyman orthogonality, which ensures that first-order perturbations in the nuisance parameter estimates do not translate into first-order errors in the decision objective. As a consequence, the sensitivity of the learned decision rule to estimation error is reduced to a second-order effect.

Second, we provide a detailed theoretical analysis of the role of sample splitting, or cross-fitting, in orthogonalized risk minimization. We show that when flexible machine learning methods are used to estimate conditional moments, a naive plug-in implementation generally suffers from a non-vanishing first-order stochastic variance term due to the dependence between the nuisance estimators and the evaluation data. Cross-fitting breaks this dependence and eliminates the leading variance contribution. Under standard regularity conditions, we establish that the excess risk of the resulting decision rule converges at a fast rate of $O(a_n^2 + a_n r_n)$, where a_n denotes the estimation rate of the primary nuisance parameter and r_n captures the relative bias rate of the auxiliary moment estimator. This represents a quadratic improvement over the $O(a_n)$ rate typically achieved by naive plug-in approaches.

Third, we instantiate the proposed framework in the contextual newsvendor problem using a deep log-linear sieve model for demand estimation. This application demonstrates how orthogonalized optimization can be combined with flexible deep learning architectures to model complex, nonlinear, and heteroskedastic demand distributions while preserving rigorous theoretical guarantees on excess risk. Numerical experiments show that the proposed method consistently outperforms both standard plug-in and end-to-end baselines in finite samples, highlighting the practical benefits of debiasing the predict-then-optimize pipeline.

The remainder of this paper is organized as follows. Section 3 introduces the problem formulation and the standard plug-in approach. Section 4 derives the orthogonalized risk functional. In Section 5, we discuss the implementation details, emphasizing the necessity of cross-fitting. Section 6 presents the theoretical convergence bounds. Section 7 applies the framework to the newsvendor problem, followed by

numerical results in Section 8. Section 9 concludes the paper.

2 Literature Review

Our work sits at the intersection of contextual stochastic optimization and semi-parametric inference using machine learning. We briefly review the existing paradigms for data-driven decision-making—ranging from classical "predict-then-optimize" to modern end-to-end learning—and then discuss the double machine learning frameworks that motivate our methodology.

Contextual Stochastic Optimization. The literature on feature-based decision-making has historically been divided into two streams: "predict-then-optimize" (PTO) and "end-to-end" approaches. The classic PTO approach estimates problem parameters first (e.g., demand) and then optimizes. While intuitive, this two-stage process fails to account for how estimation errors impact downstream decision costs. To address this, Weighted Sample Average Approximation (Weighted SAA) methods emerged, which re-weight historical samples based on covariate similarity. Seminal works include Bertsimas and Kallus (2020), who formalized the optimality of weights derived from k-nearest neighbors and random forests, and Kallus and Zhou (2018), who extended this to observational data with confounding. More recently, Kallus and Mao (2023) introduced Stochastic Optimization Forests, which tailor the splitting criteria of random forests specifically for risk minimization rather than prediction accuracy. However, as noted by Bertsimas et al. (2022), these local estimation methods often suffer from the curse of dimensionality.

Consequently, recent attention has shifted toward **Empirical Risk Minimization (ERM)** or "End-to-End" learning, which learns a global decision rule by minimizing the decision cost directly. Ban and Rudin (2019) pioneered this for the newsvendor problem with linear decision rules. This paradigm was significantly advanced by Elmachtoub and Grigas (2022), who proposed the Smart "Predict, then Optimize" (SPO) framework, introducing a convex surrogate loss function that accounts for the downstream optimization structure during training. In the deep learning domain, Amos and Kolter (2017) and Donti et al. (2017) developed differentiable optimization layers, allowing decision tasks to be embedded directly into neural network training ("Task-based End-to-End Learning"). Most recently, Han et al. (2024) proposed the Deep Neural Newsvendor, leveraging the universal approximation power of deep neural networks to minimize newsvendor loss directly without functional form assumptions.

Despite the popularity of end-to-end methods, the debate is not settled. Hu et al. (2022) demonstrate that in certain well-specified settings, a naive plug-in approach can actually achieve faster convergence rates ($O(n^{-1})$) than end-to-end methods ($O(n^{-1/2})$), suggesting that "predicting well" remains a viable strategy if the estimation bias is properly handled.

Semi-Parametric Inference and Deep Learning. Our methodological approach is rooted in the literature on **Double/Debiased Machine Learning (DML)**, pioneered by Chernozhukov et al. (2018). They introduced the use of Neyman-orthogonal scores and cross-fitting to separate interest parameters from nuisance parameters, allowing for valid inference even when nuisance parameters are estimated with slow-converging machine learning rates. This framework has been widely extended to policy learning. For instance, Athey et al. (2019) developed Generalized Random Forests, which use local moment equations to estimate heterogeneous treatment effects, while Athey and Wager (2021) utilized orthogonal scores for offline policy learning in observational settings. Similarly, Foster and Syrgkanis (2019) established general non-asymptotic guarantees for orthogonal statistical learning, further solidifying the theoretical foundation for using complex ML models in decision-making.

Specifically, we draw inspiration from recent advancements connecting Deep Neural Networks (DNNs) with semi-parametric theory. Farrell et al. (2021a) provided the first rigorous non-asymptotic bounds for deep neural networks in semi-parametric inference, proving that DNNs are fast enough to serve as first-stage estimators in DML frameworks. Building on this, Farrell et al. (2020) developed structured deep learning architectures to capture individual heterogeneity within structural economic models. Closest to our application, Zhang et al. (2024) proposed a deep learning framework for personalized policy targeting under discrete experimentation, employing orthogonal scoring to ensure asymptotic unbiasedness. We extend these ideas by applying the principle of Neyman orthogonality specifically to the risk functional of the newsvendor problem, thereby debiasing the "predict" stage to achieve superior "optimize" results.

3 Problem Formulation

We begin by formally defining the contextual stochastic optimization problem and establishing the notation used throughout the paper.

3.1 The Contextual Stochastic Optimization Problem

Consider a decision maker who observes a covariate vector $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and must select a decision z from a feasible set $\mathcal{Z} \subseteq \mathbb{R}^m$ before an uncertain outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}^k$ is realized. The cost incurred depends on both the decision and the outcome through a known cost function $c : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$.

Definition 3.1 (Oracle Risk and Optimal Decision)**.** *Let $P_{Y|X}$ denote the true conditional distribution of Y given X. The **oracle risk** at covariate x for decision z is defined as:*

$$R^*(z; x) := \mathbb{E}_{Y \sim P_{Y|X=x}} [c(z, Y)] . \tag{1}$$

*The **oracle optimal decision** is:*

$$z^*(x) \in \arg\min_{z \in \mathcal{Z}} R^*(z; x). \tag{2}$$

Since $P_{Y|X}$ is unknown in practice, the decision maker employs a parametric or semi-parametric model to approximate the conditional distribution.

3.2 The Parametric Model and Nuisance Parameter

We consider a family of conditional distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ indexed by a parameter θ belonging to a parameter space Θ. In the contextual setting, the parameter varies with the covariate, so we consider mappings $\theta : \mathcal{X} \to \Theta$.

Definition 3.2 (Model-Based Risk)**.** *For a decision $z \in \mathcal{Z}$, covariate $x \in \mathcal{X}$, and parameter value $\theta \in \Theta$, the **model-based risk** is:*

$$R(z, \theta; x) := \mathbb{E}_{Y \sim P_\theta} [c(z, Y)] . \tag{3}$$

The parameter θ is referred to as the *nuisance parameter* because our primary interest lies in the downstream decision z, not in θ itself.

Definition 3.3 (Pseudo-True Parameter)**.** *Let $\ell : \mathcal{Y} \times \Theta \to \mathbb{R}$ be a strictly proper scoring rule (e.g., negative log-likelihood). The **pseudo-true parameter** at covariate x is defined as:*

$$\theta_0(x) := \arg\min_{\theta \in \Theta} \mathbb{E}_{Y \sim P_{Y|X=x}} [\ell(Y, \theta)] . \tag{4}$$

Remark 3.1 (Interpretation of θ_0)**.** *The pseudo-true parameter $\theta_0(x)$ represents the "best" approximation to the true conditional distribution within the model class $\mathcal{P}$, where "best" is defined by the scoring rule ℓ. If the model is correctly specified (i.e., $P_{Y|X=x} \in \mathcal{P}$ for all x), then $P_{\theta_0(x)} = P_{Y|X=x}$. Otherwise, $\theta_0(x)$ is the projection of the true distribution onto $\mathcal{P}$.*

Definition 3.4 (Pseudo-True Optimal Decision)**.** *The **pseudo-true optimal decision** at covariate x is:*

$$z_0(x) \in \arg\min_{z \in \mathcal{Z}} R(z, \theta_0(x); x). \tag{5}$$

3.3 Sources of Error

The decision maker observes an i.i.d. sample $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and constructs an estimator $\hat{\theta}(\cdot)$ of the nuisance parameter. The *plug-in* approach selects the decision by minimizing the estimated risk:

$$\hat{z}_{\text{plug-in}}(x) \in \arg\min_{z \in \mathcal{Z}} R(z, \hat{\theta}(x); x). \tag{6}$$

The quality of this decision is measured by the *excess risk*:

$$\mathcal{E}(\hat{z}; x) := R^*(\hat{z}(x); x) - R^*(z^*(x); x). \tag{7}$$

There are two fundamental sources of error:

1. **Approximation Error:** The gap between the oracle risk R^* and the model-based risk $R(\cdot, \theta_0; \cdot)$ due to potential model misspecification.

2. **Estimation Error:** The gap between the model-based risk at the pseudo-true parameter $R(\cdot, \theta_0; \cdot)$ and at the estimated parameter $R(\cdot, \hat{\theta}; \cdot)$.

Our methodology specifically targets the reduction of estimation error. The approximation error is inherent to the choice of model class and cannot be reduced by statistical methods alone.

4 Methodology: Orthogonalized Risk Minimization

In this section, we develop the orthogonalized risk functional that forms the core of our approach. The key insight is to augment the standard model-based risk with a correction term that eliminates first-order sensitivity to nuisance parameter estimation errors.

4.1 The Score Function and Conditional Moment Restriction

We begin by establishing the first-order optimality condition for the pseudo-true parameter.

Definition 4.1 (Score Function)**.** *The **score function** associated with the loss ℓ is:*

$$S(y, \theta) := \nabla_\theta \ell(y, \theta). \tag{8}$$

Definition 4.2 (Conditional Moment Map)**.** *The **conditional moment map** is defined as:*

$$m(\theta; x) := \mathbb{E}_{Y \sim P_{Y|X=x}} \left[S(Y, \theta) \right]. \tag{9}$$

The following lemma establishes that the pseudo-true parameter satisfies a conditional moment restriction.

Lemma 4.1 (Score Orthogonality)**.** *Suppose $\ell(\cdot, \theta)$ is convex and differentiable in θ for each $y \in \mathcal{Y}$, and that the minimizer in (4) is attained in the interior of Θ. Then the pseudo-true parameter $\theta_0(x)$ satisfies:*

$$m(\theta_0(x); x) = \mathbb{E}_{Y \sim P_{Y|X=x}} \left[S(Y, \theta_0(x)) \right] = 0. \tag{10}$$

The proof is provided in Section A.1.

Remark 4.1. *The moment condition (10) holds by definition of θ_0 as the population minimizer, regardless of whether the parametric model $\mathcal{P}$ is correctly specified. This universality is crucial for the robustness of our approach.*

4.2 Construction of the Orthogonalized Risk Functional

To construct the orthogonalized risk, we require the Jacobian of the moment map.

Definition 4.3 (Conditional Hessian)**.** *The **conditional Hessian** at parameter θ and covariate x is:*

$$\mathcal{H}(\theta; x) := \nabla_\theta m(\theta; x) = \mathbb{E}_{Y \sim P_{Y|X=x}} \left[\nabla^2_\theta \ell(Y, \theta) \right]. \tag{11}$$

Under strict convexity of ℓ, the conditional Hessian $\mathcal{H}(\theta_0(x); x)$ is positive definite and hence invertible. We now construct a modified risk functional that is insensitive to first-order perturbations in θ around θ_0.

Definition 4.4 (Orthogonalized Risk Functional)**.** *For a decision $z \in \mathcal{Z}$, covariate $x \in \mathcal{X}$, parameter $\theta \in \Theta$, and correction coefficient $\lambda \in \Theta$, the **orthogonalized risk functional** is:*

$$\widetilde{R}(z, \theta; x, \lambda) := R(z, \theta; x) - \left\langle \lambda,\, m(\theta; x) \right\rangle, \tag{12}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in the parameter space.

The following proposition confirms that the orthogonalized risk equals the model-based risk at the pseudo-true parameter.

Proposition 4.1 (Consistency at θ_0)**.** *For any $\lambda \in \Theta$:*

$$\widetilde{R}(z, \theta_0(x); x, \lambda) = R(z, \theta_0(x); x). \tag{13}$$

Proof. By Lemma 4.1, $m(\theta_0(x); x) = 0$. Substituting into (12) yields the result immediately. □

We then seek a specific coefficient $\lambda^*(z; x)$ such that the orthogonalized risk satisfies *Neyman orthogonality*: the gradient with respect to θ vanishes at θ_0.

Definition 4.5 (Neyman Orthogonality)**.** *A functional $\widetilde{R}(z, \theta; x, \lambda)$ satisfies **Neyman orthogonality** at (θ_0, λ) if:*

$$\nabla_\theta \widetilde{R}(z, \theta; x, \lambda)\Big|_{\theta=\theta_0(x)} = 0. \tag{14}$$

Proposition 4.2 (Optimal Correction Coefficient)**.** *Suppose the loss function ℓ is strictly convex and twice differentiable in θ. Then the orthogonalized risk $\widetilde{R}$ satisfies Neyman orthogonality at θ_0 if and only if the correction coefficient is given by:*

$$\lambda^*(z; x) = \mathcal{H}(\theta_0(x); x)^{-\top} \nabla_\theta R(z, \theta_0(x); x), \tag{15}$$

where $\mathcal{H}^{-\top}$ denotes the inverse transpose of $\mathcal{H}$.

The proof is provided in Section A.2.

Remark 4.2 (Riesz Representer Interpretation)**.** *The coefficient $\lambda^*(z; x)$ can be interpreted as the Riesz representer of the linear functional $\theta \mapsto \langle \nabla_\theta R(z, \theta_0; x), \theta \rangle$ with respect to the inner product induced by $\mathcal{H}(\theta_0; x)$.*

4.3 The Gradient of the Model-Based Risk

To implement our framework, we must characterize $\nabla_\theta R(z, \theta; x)$. The following lemma relates this gradient to the model score function.

Lemma 4.2 (Risk Gradient via Score)**.** *Let P_θ admit a density $p_\theta(y)$ with respect to a base measure, and assume that differentiation and integration can be interchanged. Define the **model score function**:*

$$s_\theta(y) := \nabla_\theta \log p_\theta(y). \tag{16}$$

Then:

$$\nabla_\theta R(z, \theta; x) = \mathbb{E}_{Y \sim P_\theta} \left[c(z, Y) \cdot s_\theta(Y) \right]. \tag{17}$$

The proof is provided in Section A.3.

Remark 4.3 (Connection Between Scores)**.** *In general, the loss score $S(y, \theta) = \nabla_\theta \ell(y, \theta)$ and the model score $s_\theta(y) = \nabla_\theta \log p_\theta(y)$ are distinct. However, when $\ell(y, \theta) = -\log p_\theta(y)$ (negative log-likelihood), we have $S(y, \theta) = -s_\theta(y)$. This connection is exploited in Section 7.*

5 Feasible Estimation and the Role of Cross-Fitting

In Section 4, we constructed the theoretical orthogonalized risk $\widetilde{R}(z, \theta; x, \lambda^*)$ and showed that it possesses the Neyman orthogonality property. Minimizing this risk with respect to z would yield a decision rule insensitive to first-order errors in θ.

However, the theoretical functional $\widetilde{R}$ is *infeasible* in practice because it depends on two unknown population quantities:

- The conditional moment map $m(\theta; x) = \mathbb{E}[S(Y, \theta) | X = x]$.
- The optimal correction coefficient $\lambda^*(z; x)$, which depends on the conditional Hessian $\mathcal{H}(\theta_0; x)$ and risk gradient $\nabla_\theta R$.

To implement the orthogonalization strategy, we must replace these unknown quantities with data-driven estimators. This transition from the theoretical ideal to a feasible estimator introduces complex statistical dependencies that, if mishandled, can negate the benefits of orthogonality.

5.1 The Feasible Orthogonalized Risk

Let $\hat{\theta}$, $\hat{m}$, and $\hat{\lambda}$ denote estimators for the pseudo-true parameter, the conditional moment map, and the correction coefficient, respectively. We obtain these estimators using an auxiliary dataset or a portion of the sample.

Definition 5.1 (Feasible Orthogonalized Risk)**.** *The **feasible orthogonalized risk** is constructed by plugging the estimators into the definition of $\widetilde{R}$:*

$$\widehat{\widetilde{R}}(z, \hat{\theta}(x); x) := R(z, \hat{\theta}(x); x) - \big\langle \hat{\lambda}(z; x),\, \hat{m}(\hat{\theta}(x); x) \big\rangle. \tag{18}$$

The decision rule is then obtained by minimizing this feasible empirical quantity:

$$\hat{z}(x) \in \arg\min_{z \in \mathcal{Z}} \widehat{\widetilde{R}}(z, \hat{\theta}(x); x). \tag{19}$$

5.2 The Challenge: Stochastic Variance of the Moment Estimator

Consider the feasible orthogonalized risk:

$$\widehat{\widetilde{R}}(z, \hat{\theta}; x) = R(z, \hat{\theta}(x); x) - \langle \hat{\lambda}(z; x), \hat{m}(\hat{\theta}(x); x) \rangle. \tag{20}$$

The term $\hat{m}(\hat{\theta}; x) - m(\hat{\theta}; x)$ represents the error in estimating the conditional moment. To understand this error, we apply a *bias-variance decomposition*.

Definition 5.2 (Bias-Variance Decomposition)**.** *Let $\hat{m}$ be a moment estimator trained on dataset $\mathcal{D}_{train}$. Define:*

$$\bar{m}(\theta; x) := \mathbb{E}_{\mathcal{D}_{train}}[\hat{m}(\theta; x)], \quad \textit{(systematic component)} \tag{21}$$

$$\nu(\theta; x) := \hat{m}(\theta; x) - \bar{m}(\theta; x). \quad \textit{(stochastic component)} \tag{22}$$

Then:

$$\hat{m}(\theta; x) - m(\theta; x) = \underbrace{(\bar{m}(\theta; x) - m(\theta; x))}_{Bias\ B(\theta;x)} + \underbrace{\nu(\theta; x)}_{Variance}. \tag{23}$$

Remark 5.1 (Key Property of the Variance Component)**.** *By construction, $\mathbb{E}_{\mathcal{D}_{train}}[\nu(\theta; x)] = 0$ for any fixed θ and x. This zero-mean property is the foundation for eliminating the variance contribution via cross-fitting.*

Consider a naive implementation where all estimators ($\hat{\theta}$, $\hat{m}$, $\hat{\lambda}$) are trained on the same dataset $\mathcal{D}_n$, and the same dataset is used for evaluation. The critical error term is:

$$T_3 = -\langle \hat{\lambda}, \hat{m}(\hat{\theta}) - m(\hat{\theta}) \rangle. \tag{24}$$

Decomposing using bias-variance:

$$T_3 = -\langle \hat{\lambda}, B(\hat{\theta}) \rangle - \langle \hat{\lambda}, \nu(\hat{\theta}) \rangle. \tag{25}$$

The variance term $\langle \hat{\lambda}, \nu(\hat{\theta}) \rangle$ is problematic. Even though $\mathbb{E}[\nu] = 0$, the coefficient $\hat{\lambda}$ depends on the *same* training data as ν. This correlation means:

$$\mathbb{E}[\langle \hat{\lambda}, \nu \rangle] \neq \langle \mathbb{E}[\hat{\lambda}], \mathbb{E}[\nu] \rangle = 0. \tag{26}$$

The expectation does not factor, and the variance term contributes a first-order error of magnitude $O(s_n)$, where s_n is the standard deviation of the non-parametric estimator (typically $O(n^{-\beta/(2\beta+d)})$ for smoothness β and dimension d).

5.3 How Cross-Fitting Solves the Challenge

Cross-fitting breaks the correlation between the nuisance estimators and the evaluation data by using sample splitting.

Definition 5.3 (K-Fold Cross-Fitting)**.** *Partition the dataset $\mathcal{D}_n$ randomly into $K \geq 2$ disjoint folds $\mathcal{D}_1, \ldots, \mathcal{D}_K$ of approximately equal size. For each fold $k \in \{1, \ldots, K\}$:*

1. *Define the training set $\mathcal{D}^{(-k)} := \mathcal{D}_n \setminus \mathcal{D}_k$.*
2. *Train nuisance estimators $\hat{\theta}^{(-k)}$, $\hat{m}^{(-k)}$, and $\hat{\lambda}^{(-k)}$ using only $\mathcal{D}^{(-k)}$.*
3. *For observations $(X_i, Y_i) \in \mathcal{D}_k$, evaluate the orthogonalized risk using the estimators $\hat{\theta}^{(-k)}$, $\hat{m}^{(-k)}$, $\hat{\lambda}^{(-k)}$.*

The key property is that for any observation $X \in \mathcal{D}_k$:

- The nuisance estimators $\hat{\theta}^{(-k)}$, $\hat{m}^{(-k)}$, $\hat{\lambda}^{(-k)}$ are functions only of $\mathcal{D}^{(-k)}$.
- The evaluation point X is statistically *independent* of $\mathcal{D}^{(-k)}$.

We now show precisely how cross-fitting eliminates the first-order variance contribution.

Proposition 5.1 (Variance Elimination via Cross-Fitting)**.** *Under the cross-fitting scheme of Definition 5.3, consider the term:*

$$T_{3c} := -\langle \lambda^*(X), \nu^{(-k)}(\hat{\theta}^{(-k)}(X); X) \rangle, \tag{27}$$

where $X \in \mathcal{D}_k$ and $\nu^{(-k)} = \hat{m}^{(-k)} - \bar{m}$. Then:

$$\mathbb{E}[T_{3c}] = 0. \tag{28}$$

Proof. We apply the tower property of conditional expectation. Since $X \in \mathcal{D}_k$ is independent of $\mathcal{D}^{(-k)}$:

$$\mathbb{E}[T_{3c}] = \mathbb{E}\left[-\langle \lambda^*(X), \nu^{(-k)}(X) \rangle \right] \tag{29}$$

$$= \mathbb{E}_X \left[\mathbb{E}_{\mathcal{D}^{(-k)}} \left[-\langle \lambda^*(X), \nu^{(-k)}(X) \rangle \,\middle|\, X \right] \right]. \tag{30}$$

In the inner expectation, X is fixed. The oracle coefficient $\lambda^*(X)$ depends only on θ_0 and X, not on $\mathcal{D}^{(-k)}$. Therefore, $\lambda^*(X)$ can be treated as a constant:

$$\mathbb{E}_{\mathcal{D}^{(-k)}} \left[-\langle \lambda^*(X), \nu^{(-k)}(X) \rangle \,\middle|\, X \right] = -\left\langle \lambda^*(X), \mathbb{E}_{\mathcal{D}^{(-k)}} \left[\nu^{(-k)}(X) \,\middle|\, X \right] \right\rangle. \tag{31}$$

By definition of the variance component:

$$\mathbb{E}_{\mathcal{D}^{(-k)}} \left[\nu^{(-k)}(X) \,\middle|\, X \right] = \mathbb{E}_{\mathcal{D}^{(-k)}} \left[\hat{m}^{(-k)}(X) - \bar{m}(X) \right] = \bar{m}(X) - \bar{m}(X) = 0. \tag{32}$$

Therefore, $\mathbb{E}[T_{3c}] = 0$. □

Remark 5.2 (Intuition)**.** *The variance component ν represents random fluctuations around the expected estimator $\bar{m}$. These fluctuations have zero mean by construction. Cross-fitting ensures that the coefficient λ^* does not "see" the same random fluctuations, so when we take expectations, the fluctuations average out to zero. This is analogous to the principle that independent zero-mean random variables have zero covariance.*

5.4 What Cross-Fitting Does NOT Eliminate

It is important to understand the limitations of cross-fitting:

1. **Bias is not eliminated:** The systematic bias $B = \bar{m} - m$ is a deterministic function (given the estimation procedure) and does not vanish under expectation. Our analysis shows that this bias contributes $O(a_n r_n)$ due to the relative error structure.
2. **Coefficient error × variance interaction:** The term $\langle \hat{\lambda}^{(-k)} - \lambda^*, \nu^{(-k)} \rangle$ involves two quantities that both depend on $\mathcal{D}^{(-k)}$. This term does not vanish but is bounded by $O(a_n s_n)$, which is a product of rates.

For cross-fitting to successfully eliminate the variance term, the following conditions must be satisfied:

1. **Independence:** The evaluation point X must be statistically independent of the training data $\mathcal{D}^{(-k)}$ used to construct the nuisance estimators. This is guaranteed by the fold structure.
2. **Oracle coefficient or consistent approximation:** The coefficient λ^* must either be the true oracle value (which depends only on θ_0 and is deterministic) or must be estimated on a *separate* fold from both $\hat{m}$ and the evaluation data.
3. **Sufficient folds:** Using $K \geq 2$ folds is necessary. In practice, $K = 5$ or $K = 10$ provides a good balance between statistical efficiency and computational cost.

Remark 5.3 (Three-Way Sample Splitting)**.** *For the cleanest theoretical analysis, one can use three-way splitting: Fold 1 for $\hat{\theta}$, Fold 2 for $\hat{m}$, and Fold 3 for evaluation. However, standard K-fold cross-fitting with $K \geq 2$ suffices because the key independence is between the evaluation point and the nuisance training data.*

6 Theoretical Analysis

We now present the main theoretical results characterizing the performance of our orthogonalized estimator. We first state the required assumptions, then provide a decomposition of the excess risk, and finally establish the main convergence rate.

6.1 Assumptions

We collect all assumptions needed for our theoretical analysis.

Assumption 6.1 (Regularity of the Model Family)**.** *The parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and loss function ℓ satisfy:*

(i) The parameter space $\Theta \subseteq \mathbb{R}^p$ is convex and compact.

(ii) For each $y \in \mathcal{Y}$, the map $\theta \mapsto \ell(y, \theta)$ is twice continuously differentiable.

(iii) The loss is μ-strongly convex: for all $y \in \mathcal{Y}$ and $\theta \in \Theta$,

$$\nabla^2_\theta \ell(y, \theta) \succeq \mu I_p,$$

where I_p is the $p \times p$ identity matrix.

(iv) The Hessian is bounded: there exists $L_\ell > 0$ such that for all $y \in \mathcal{Y}$ and $\theta \in \Theta$,

$$\|\nabla^2_\theta \ell(y, \theta)\|_{op} \leq L_\ell.$$

Assumption 6.2 (Regularity of the Risk Functional)**.** *The model-based risk $R(z, \theta; x)$ satisfies:*

(i) For each $z \in \mathcal{Z}$ and $x \in \mathcal{X}$, the map $\theta \mapsto R(z, \theta; x)$ is twice continuously Fréchet differentiable.

(ii) The Hessian is bounded: there exists $L_R > 0$ such that for all z, x, θ,

$$\|\nabla^2_\theta R(z, \theta; x)\|_{op} \leq L_R.$$

(iii) The gradient is uniformly bounded: there exists $B_R > 0$ such that

$$\sup_{z \in \mathcal{Z},\, x \in \mathcal{X},\, \theta \in \Theta} \|\nabla_\theta R(z, \theta; x)\| \leq B_R.$$

Assumption 6.3 (Smoothness of the Conditional Moment)**.** *The conditional moment map $m(\theta; x)$ satisfies:*

(i) The map $\theta \mapsto m(\theta; x)$ is continuously differentiable for each x.

(ii) The Jacobian (conditional Hessian) is Lipschitz: there exists $L_m > 0$ such that for all x and $\theta_1, \theta_2 \in \Theta$,

$$\|\mathcal{H}(\theta_1; x) - \mathcal{H}(\theta_2; x)\|_{op} \le L_m \|\theta_1 - \theta_2\|.$$

(iii) The Hessian is bounded: $\sup_{x,\theta} \|\mathcal{H}(\theta; x)\|_{op} \le H_{\max} < \infty$.

Assumption 6.4 (Boundedness of the Optimal Coefficient)**.** *The optimal correction coefficient is uniformly bounded:*

$$B_\lambda := \sup_{z \in \mathcal{Z},\, x \in \mathcal{X}} \|\lambda^*(z; x)\| < \infty. \tag{33}$$

Remark 6.1. *Under Assumptions 6.1 and 6.2, 6.4 follows from the bound* $\|\lambda^*\| \le \|\mathcal{H}^{-1}\|_{op} \|\nabla_\theta R\| \le \mu^{-1} B_R$.

Assumption 6.5 (Nuisance Estimation Rates)**.** *The nuisance estimators satisfy the following convergence conditions:*

(i) ***Parameter estimation rate:*** *The estimator $\hat{\theta}$ satisfies*

$$\|\hat{\theta} - \theta_0\|_{L_2(P_X)} := \left(\mathbb{E}_X \left[\|\hat{\theta}(X) - \theta_0(X)\|^2\right]\right)^{1/2} \le a_n.$$

(ii) ***Coefficient estimation rate:*** *The estimator $\hat{\lambda}$ satisfies*

$$\|\hat{\lambda} - \lambda^*\|_{L_2(P_X)} \le C_\lambda \cdot a_n,$$

where C_λ depends on the Lipschitz constants of $\mathcal{H}^{-1}$ and $\nabla_\theta R$.

Assumption 6.6 (Moment Estimator: Bias-Variance Structure)**.** *Let $\hat{m}$ be the moment estimator trained on the auxiliary fold, and define $\bar{m}(\theta; x) := \mathbb{E}_{\mathcal{D}_{train}}[\hat{m}(\theta; x)]$ and $\nu(\theta; x) := \hat{m}(\theta; x) - \bar{m}(\theta; x)$.*

(i) ***Relative Bias:*** *The systematic bias satisfies*

$$\|\bar{m}(\theta; \cdot) - m(\theta; \cdot)\|_{L_2(P_X)} \le r_n \cdot \|\theta - \theta_0(\cdot)\|_{L_2(P_X)}, \tag{34}$$

where r_n is the convergence rate of the underlying function estimation.

(ii) ***Variance Rate:*** *The stochastic component satisfies*

$$\left(\mathbb{E}\left[\|\nu(\theta; \cdot)\|^2\right]\right)^{1/2} \le s_n. \tag{35}$$

Remark 6.2 (Justification for Relative Bias)**.** *Assumption 6.6(i) is justified by the following observation: since $m(\theta_0; x) = 0$ (Score Orthogonality), Taylor expansion gives $m(\theta; x) = \mathcal{H}(\theta_0; x)(\theta - \theta_0) + O(\|\theta - \theta_0\|^2)$. The target function $m(\theta; \cdot)$ itself has magnitude $O(\|\theta - \theta_0\|)$. For well-regularized estimators (e.g., neural networks with weight decay), the estimation bias scales proportionally to the target signal magnitude, yielding the relative error structure.*

Assumption 6.7 (Cross-Fitting)**.** *The dataset $\mathcal{D}_n$ is randomly partitioned into $K \ge 2$ folds. For each fold k, the estimators $\hat{\theta}^{(-k)}$, $\hat{m}^{(-k)}$, and $\hat{\lambda}^{(-k)}$ are constructed using data from all folds except k. The orthogonalized decision for observations in fold k uses only these out-of-fold estimators.*

6.2 Risk Decomposition

We first decompose the excess risk into approximation and estimation components.

Proposition 6.1 (Risk Decomposition)**.** *Under the definitions in Section 3, for any decision rule $\hat{z}$ and covariate $x \in \mathcal{X}$, the conditional excess risk satisfies:*

$$\mathcal{E}(\hat{z}; x) \le \underbrace{2 \sup_{z \in \mathcal{Z}} |R^*(z; x) - R(z, \theta_0(x); x)|}_{\textit{Approximation Error: } \epsilon_{\text{approx}}(x)} + \underbrace{\left(R(\hat{z}(x), \theta_0(x); x) - R(z_0(x), \theta_0(x); x)\right)}_{\textit{Estimation-Induced Suboptimality}}. \tag{36}$$

The proof is provided in Section A.4.

Remark 6.3 (Interpretation). *The approximation error $\epsilon_{\text{approx}}(x)$ quantifies the cost of using the model class $\mathcal{P}$ instead of the true distribution. Under Lipschitz continuity of the cost function, this is controlled by the Wasserstein distance between $P_{Y|X=x}$ and $P_{\theta_0(x)}$. Our orthogonalization method targets the second term.*

The following lemma provides the key technical tool: a precise characterization of how the orthogonalized risk deviates from the target as θ varies from θ_0.

Lemma 6.1 (Second-Order Expansion). *Suppose Assumptions 6.1–6.3 hold. Let $\lambda = \lambda^*(z; x)$ be the optimal correction coefficient. Then for any $\theta \in \Theta$:*

$$\widetilde{R}(z, \theta; x, \lambda^*) - R(z, \theta_0; x) = \mathcal{Q}(\theta - \theta_0), \tag{37}$$

where $\mathcal{Q}$ is a quadratic remainder satisfying:

$$|\mathcal{Q}(\delta)| \leq \frac{L_R + B_\lambda L_m}{2} \|\delta\|^2. \tag{38}$$

In particular, the first-order term vanishes due to Neyman orthogonality.

The proof is provided in Section A.5.

6.3 Main Convergence Theorem

We now state our main result characterizing the convergence rate of the orthogonalized estimator.

Definition 6.1 (Feasible Orthogonalized Risk). *Given estimators $\hat{\theta}$, $\hat{m}$, and $\hat{\lambda}$, the **feasible orthogonalized risk** is:*

$$\widehat{\widetilde{R}}(z, \hat{\theta}; x) := R(z, \hat{\theta}(x); x) - \left\langle \hat{\lambda}(z; x),\, \hat{m}(\hat{\theta}(x); x) \right\rangle. \tag{39}$$

Definition 6.2 (Orthogonalized Decision Rule). *The **orthogonalized decision rule** is:*

$$\hat{z}(x) \in \arg\min_{z \in \mathcal{Z}} \widehat{\widetilde{R}}(z, \hat{\theta}(x); x). \tag{40}$$

Theorem 6.1 (Main Convergence Rate). *Suppose Assumptions 6.1–6.7 hold. Let $\hat{z}$ be the orthogonalized decision rule obtained via cross-fitting with $K \geq 2$ folds. Define:*

- $a_n := \|\hat{\theta} - \theta_0\|_{L_2(P_X)}$ *(parameter estimation rate)*
- r_n *: relative bias rate of the moment estimator*
- s_n *: variance rate of the moment estimator*

Then the expected estimation-induced suboptimality satisfies:

$$\mathbb{E}\left[R(\hat{z}(X), \theta_0(X); X) - R(z_0(X), \theta_0(X); X)\right] \leq C_1 a_n^2 + C_2 a_n r_n + C_3 a_n s_n, \tag{41}$$

where the constants are:

$$C_1 = L_R + B_\lambda L_m + 2C_\lambda H_{\max}, \tag{42}$$
$$C_2 = 2B_\lambda, \tag{43}$$
$$C_3 = 2C_\lambda. \tag{44}$$

The complete proof is provided in Section A.6.

Remark 6.4 (Simplified Bound). *For well-regularized non-parametric estimators, the variance rate s_n is typically comparable to or dominated by the bias rate r_n. In this case, the bound simplifies to:*

$$\mathbb{E}\left[R(\hat{z}, \theta_0) - R(z_0, \theta_0)\right] = O(a_n^2 + a_n r_n). \tag{45}$$

Corollary 6.1 (Comparison with Plug-In Estimator). *Under the same assumptions, the plug-in estimator $\hat{z}_{\textit{plug-in}}$ that minimizes $R(z, \hat{\theta}; x)$ satisfies:*

$$\mathbb{E}_X\left[R(\hat{z}_{\textit{plug-in}}(X), \theta_0(X); X) - R(z_0(X), \theta_0(X); X)\right] = O(a_n). \tag{46}$$

When $a_n = o(1)$, the orthogonalized estimator achieves a strictly faster rate.

Proof. For the plug-in estimator, a first-order Taylor expansion of $R(z, \hat{\theta}; x)$ around $\theta_0(x)$ yields:

$$R(z, \hat{\theta}; x) - R(z, \theta_0; x) = \langle \nabla_\theta R(z, \theta_0; x), \hat{\theta} - \theta_0 \rangle + O(\|\hat{\theta} - \theta_0\|^2).$$

The first-order term scales as $O(\|\hat{\theta} - \theta_0\|) = O(a_n)$ in general, dominating the second-order term. □

Remark 6.5 (Rate Conditions for Fast Convergence)**.** *For the estimation error to be $O(n^{-1/2})$, we require $a_n^2 = O(n^{-1/2})$ and $a_n r_n = O(n^{-1/2})$. This is satisfied when $a_n = O(n^{-1/4})$ and $r_n = O(n^{-1/4})$. These rates are achievable for the nuisance functions under standard nonparametric assumptions when the dimension d is not too large and the functions have sufficient smoothness.*

Corollary 6.2 (Total Excess Risk Bound)**.** *Combining the risk decomposition in Proposition 6.1 and the convergence rate in Theorem 6.1, the total conditional excess risk of the orthogonalized decision rule $\hat{z}$ satisfies:*

$$\mathbb{E}\left[\mathcal{E}(\hat{z}; X)\right] \leq 2\mathbb{E}[\epsilon_{\text{approx}}(X)] + C_1 a_n^2 + C_2 a_n r_n + C_3 a_n s_n. \tag{47}$$

In particular, if the parametric model is correctly specified (i.e., $P_{Y|X} \in \mathcal{P}$), then $\epsilon_{\text{approx}}(X) = 0$ almost surely, and the total excess risk converges at the fast rate $O(a_n^2 + a_n r_n + a_n s_n)$. Conversely, under model misspecification, the excess risk is dominated asymptotically by the approximation error $2\mathbb{E}[\epsilon_{\text{approx}}(X)]$, which represents the irreducible bias due to the choice of model class $\mathcal{P}$.

7 Application: Orthogonalized Newsvendor with Deep Sieve Demand Model

We now apply our general orthogonalized optimization framework to the contextual newsvendor problem. To achieve flexibility in modeling complex, potentially multi-modal demand distributions while maintaining computational tractability, we employ a **Deep Log-Linear Sieve Model**. This model combines the approximation power of neural networks with the statistical properties of exponential families.

7.1 Problem Setup: The Newsvendor Loss

A retailer must decide an order quantity $z \geq 0$ for a perishable product before observing the uncertain demand $Y \geq 0$, which depends on covariates $X \in \mathcal{X}$. The cost function is the standard newsvendor loss:

$$c(z, y) = c_u(y - z)^+ + c_o(z - y)^+, \tag{48}$$

where $c_u > 0$ is the unit underage cost and $c_o > 0$ is the unit overage cost.

The oracle optimal decision $z^*(x)$ is the α-quantile of the true conditional distribution $P_{Y|X=x}$, where $\alpha = c_u/(c_u + c_o)$ is the critical ratio. Our goal is to estimate $z^*(x)$ by learning a parametric model $P_{\theta(x)}$ for the demand distribution.

7.2 The Deep Log-Linear Sieve Demand Model

To approximate the unknown conditional density $f_{Y|X}(y|x)$, we use a flexible exponential family model where the natural parameters are outputs of a deep neural network.

7.2.1 Sieve Basis and Density Definition

Let $\mathbf{B}(y) = (B_1(y), \ldots, B_J(y))^\top$ be a vector of J fixed basis functions $B_j : \mathcal{Y} \to \mathbb{R}$. Common choices include polynomials or B-splines. We impose the following regularity conditions:

Assumption 7.1 (Basis Regularity)**.** *The basis functions satisfy:*

(i) ***Normalization:*** *$B_1(y) \equiv 1$ to ensure the model includes a constant term.*

(ii) ***Linear Independence:*** *$\{B_1, \ldots, B_J\}$ are linearly independent in $L^2(\mathcal{Y})$.*

(iii) ***Integrability:*** *For all relevant θ, $\int_{\mathcal{Y}} \exp(\theta^\top \mathbf{B}(y))\, dy < \infty$.*

Definition 7.1 (Log-Linear Sieve Model)**.** *For a parameter vector $\theta \in \mathbb{R}^J$, the probability density function is defined as:*

$$f(y;\theta) := \exp\big(\theta^\top \mathbf{B}(y) - A(\theta)\big), \tag{49}$$

where $A(\theta)$ is the ***log-partition function*** *ensuring normalization:*

$$A(\theta) := \log \int_{\mathcal{Y}} \exp\big(\theta^\top \mathbf{B}(y)\big)\, dy. \tag{50}$$

To capture heterogeneity, the natural parameter θ is modeled as a function of covariates x via a deep neural network $g_{\mathbf{w}}$:

$$\theta(x) = g_{\mathbf{w}}(x) : \mathcal{X} \to \mathbb{R}^J. \tag{51}$$

7.2.2 Properties of the Log-Partition Function

The log-partition function $A(\theta)$ encapsulates the statistical properties of the model. By standard exponential family theory:

- **Gradient (Mean):** $\nabla_\theta A(\theta) = \mathbb{E}_{Y\sim f(\cdot;\theta)}[\mathbf{B}(Y)]$.
- **Hessian (Covariance):** $\nabla^2_\theta A(\theta) = \mathrm{Var}_{Y\sim f(\cdot;\theta)}[\mathbf{B}(Y)]$.

Under Assumption 7.1(ii), the Hessian $\nabla^2_\theta A(\theta)$ is positive definite, ensuring strict convexity of $A(\theta)$.

7.3 Analytic Expressions for Orthogonalization Components

We now map the specific log-linear sieve model back to the general orthogonalization framework established in Section 4. We derive explicit expressions for the Score, Conditional Moment Map, and Conditional Hessian using the Negative Log-Likelihood (NLL) as the loss function. Note that in these definitions, the parameter θ is always conditional on x, i.e., $\theta = \theta(x)$.

1. The Score Function. For the NLL loss $\ell(y,\theta) = A(\theta) - \theta^\top \mathbf{B}(y)$, the score function $S(y,\theta) = \nabla_\theta \ell(y,\theta)$ is:

$$S(y,\theta) = \nabla_\theta A(\theta) - \mathbf{B}(y). \tag{52}$$

2. The Conditional Moment Map. Following Definition 4.2, the conditional moment map $m(\theta;x)$ is the expected score conditioned on the covariate $X = x$:

$$m(\theta;x) := \mathbb{E}_{Y|X=x}\left[S(Y,\theta)\right] = \nabla_\theta A(\theta) - \mathbb{E}_{Y|X=x}[\mathbf{B}(Y)]. \tag{53}$$

At the pseudo-true parameter $\theta_0(x)$, we have $m(\theta_0(x);x) = 0$.

3. The Conditional Hessian. Following Definition 4.3, the conditional Hessian $\mathcal{H}(\theta;x)$ is the expected Jacobian of the score given $X = x$:

$$\mathcal{H}(\theta;x) := \mathbb{E}_{Y|X=x}\left[\nabla^2_\theta \ell(Y,\theta)\right]. \tag{54}$$

For our exponential family model, $\nabla^2_\theta \ell(Y,\theta) = \nabla^2_\theta A(\theta)$, which is constant with respect to Y. Therefore, the conditional expectation simplifies to:

$$\mathcal{H}(\theta;x) = \nabla^2_\theta A(\theta). \tag{55}$$

Note that while $\nabla^2 A$ does not explicitly contain Y, the Hessian $\mathcal{H}(\theta;x)$ depends on x through the parameter $\theta(x)$.

4. The Risk Gradient. The gradient of the model-based risk $R(z,\theta;x)$ with respect to θ is given by Lemma 4.2. Using the identity $\nabla_\theta \log f(y;\theta) = -S(y,\theta)$:

$$\nabla_\theta R(z,\theta;x) = \mathbb{E}_{Y\sim P_{\theta(x)}}\left[c(z,Y)\cdot\big(-S(Y,\theta)\big)\right]\Big|_{\theta=\theta(x)}. \tag{56}$$

This integral is computed with respect to the *model* distribution $P_{\theta(x)}$, not the true distribution.

5. The Optimal Correction Coefficient. Substituting (54) and (56) into Proposition 4.2, the optimal correction coefficient at a given x is:

$$\lambda^*(z;x) = \left[\nabla^2_\theta A(\theta(x))\right]^{-1} \nabla_\theta R(z, \theta(x); x). \tag{57}$$

7.4 Implementation Algorithm

We now present the specific instantiation of the Cross-Fitted Orthogonalized Optimization algorithm for the Deep Sieve Newsvendor problem.

Algorithm 1 Deep Sieve Orthogonalized Newsvendor

Require: Dataset $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, Basis functions $\mathbf{B}(\cdot) \in \mathbb{R}^J$, folds K.
Ensure: Decision rule $\hat{z}(x)$.

Partition data into K folds.
for $k = 1, \ldots, K$ **do**
  Let $\mathcal{D}^{(-k)}$ be the training set.
  1. Train Demand Model ($\hat{\theta}$):
  Train network $g_{\mathbf{w}}(x)$ (output dim J) on $\mathcal{D}^{(-k)}$ to minimize NLL:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{j \in \mathcal{D}^{(-k)}} \left(A(g_{\mathbf{w}}(X_j)) - g_{\mathbf{w}}(X_j)^\top \mathbf{B}(Y_j) \right)$$

  Let $\hat{\theta}^{(-k)}(x) = g_{\hat{\mathbf{w}}}(x)$.
  2. Train Moment Estimator ($\hat{m}$):
  Compute score residuals on training data for each $j \in \mathcal{D}^{(-k)}$:

$$\hat{S}_j = \nabla_\theta A(\hat{\theta}^{(-k)}(X_j)) - \mathbf{B}(Y_j)$$

  Train auxiliary network $\hat{m}^{(-k)}(x)$ (output dim J) to predict $\hat{S}_j$ from X_j (using MSE loss) to approximate the conditional expectation $\mathbb{E}[S|X]$.
  3. Orthogonalized Decision for Fold k:
  for $i \in$ Fold k **do**
    Evaluate $\theta_i = \hat{\theta}^{(-k)}(X_i)$.
    Compute Conditional Hessian $\mathbf{H}_i = \nabla^2_\theta A(\theta_i) \in \mathbb{R}^{J \times J}$.
    Compute Risk Gradient $\mathbf{g}_i(z) = \nabla_\theta R(z, \theta_i; X_i)$ via numerical integration over model density.
    Compute Coefficient $\hat{\lambda}_i(z) = \mathbf{H}_i^{-1} \mathbf{g}_i(z)$.
    Solve for optimal order quantity:

$$\hat{z}(X_i) = \arg\min_{z \geq 0} \left(R(z, \theta_i; X_i) - \hat{\lambda}_i(z)^\top \hat{m}^{(-k)}(X_i) \right)$$

  end for
end for

Remark 7.1 (Computational Efficiency)**.** *For the Log-Linear Sieve model, the gradient $\nabla_\theta A(\theta)$ and Hessian $\nabla^2_\theta A(\theta)$ are expectations over the exponential family distribution. When using polynomial or spline bases, these integrals can often be computed via efficient 1D numerical quadrature (e.g., Gauss-Legendre), avoiding expensive Monte Carlo sampling during the optimization loop.*

7.5 Total Excess Risk and the Benefit of Orthogonalization

We now analyze the total excess risk of the orthogonalized decision rule. This analysis explicitly characterizes the trade-off between the approximation error (due to the finite sieve basis size J) and the estimation error (due to learning the neural network parameters with finite samples n).

Detailed assumptions, formal statements, and rigorous proofs are provided in Section B. Here, we summarize the key findings derived from the risk decomposition.

The Bias-Variance Trade-off. The total excess risk is bounded by the sum of two components, as established in Theorem B.1:

1. **Approximation Bias:** As the number of basis functions J increases, the sieve model can better approximate the true log-density. Under Sobolev regularity conditions, this error decays at a rate of $O(J^{-2s})$, where s is the smoothness of the true density (see B.1).

2. **Estimation Error:** Increasing J requires the neural network to output higher-dimensional parameters, increasing the complexity of the learning task. Based on the optimal architecture, the squared estimation error scales as $O(Jn^{-\frac{2\beta}{2\beta+d}})$, assuming the primary and auxiliary tasks share the same smoothness β.

Main Result: Faster Convergence. By combining these components, Theorem B.1 establishes the total risk bound for our orthogonalized estimator:

$$\mathbb{E}[\mathcal{E}(\hat{z})] = O\left(J^{-2s} + Jn^{-\frac{2\beta}{2\beta+d}}\right). \tag{58}$$

Crucially, our orthogonalization strategy provides a quadratic improvement in the convergence rate with respect to sample size n compared to a standard plug-in approach.

- The Plug-in risk is dominated by the first-order estimation error: $O(a_n) = O(\sqrt{J}n^{-\frac{\beta}{2\beta+d}})$.
- The Orthogonalized risk is dominated by the second-order estimation error: $O(a_n^2) = O(Jn^{-\frac{2\beta}{2\beta+d}})$.

Because the estimation error rate is effectively squared (improved) by orthogonalization, we can afford to use a significantly larger basis size J_n to aggressively reduce approximation bias without paying a prohibitive price in variance. This results in a strictly faster overall convergence rate to the oracle decision, particularly in high-dimensional settings where β is close to d.

8 Numerical Experiments

In this section, we evaluate the finite-sample performance of the proposed **Deep Sieve Orthogonalized Newsvendor (NN-DML)** framework. We aim to empirically validate two key claims:

1. **Practical Superiority:** The feasible orthogonalized estimator achieves a faster convergence rate compared to standard Plug-in and End-to-End approaches.

2. **Theoretical Validity:** In an ideal setting with known components of debiasing term λ^* and m, the orthogonalization removes first-order bias completely, validating the geometric intuition behind our risk functional.

8.1 Experimental Setup

Data Generation Process (DGP). We simulate a contextual newsvendor problem with $d = 3$ covariates $X \sim \text{Uniform}([0,1]^3)$. The demand Y is generated from a log-normal distribution $Y \mid X \sim \text{LogNormal}(\mu(X), \sigma(X))$, where the parameters depend on X through highly non-linear functions:

$$\mu(X) = 1.0 + 0.8\sin(3X_1) + 0.5X_2\cos(X_3), \tag{59}$$
$$\sigma(X) = 0.4 + 0.1|X_3 + X_1|. \tag{60}$$

This setup ensures heteroskedasticity and non-linearity, challenging the expressivity of the models.

Problem Parameters. We set the unit costs to $c_u = 9.0$ and $c_o = 1.0$, corresponding to a high critical ratio $\alpha = 0.9$. This high-quantile estimation task is known to be sensitive to tail estimation errors.

Methods Compared.

- **Naive Plug-in:** Estimates $\hat{\theta}(x)$ using the Deep Sieve model but optimizes the decision directly based on the estimated density.
- **One-Step (End-to-End):** Uses a deep neural network to map X directly to z, trained by minimizing the empirical newsvendor loss (Ban and Rudin, 2019) (Han et al., 2024) .

- **Feasible NN-DML (Ours):** The proposed method using 4-fold cross-fitting to estimate nuisance parameters in correction term $\hat{\lambda}$ and $\hat{m}$.
- **Oracle NN-DML (Ideal):** An idealized version of our method that uses the *true* population functions λ^* and m for the correction term, isolating the theoretical gain of the orthogonalized risk functional.

8.2 Main Results: Feasible Performance

We first evaluate the practical performance of our Feasible NN-DML estimator against the baselines.

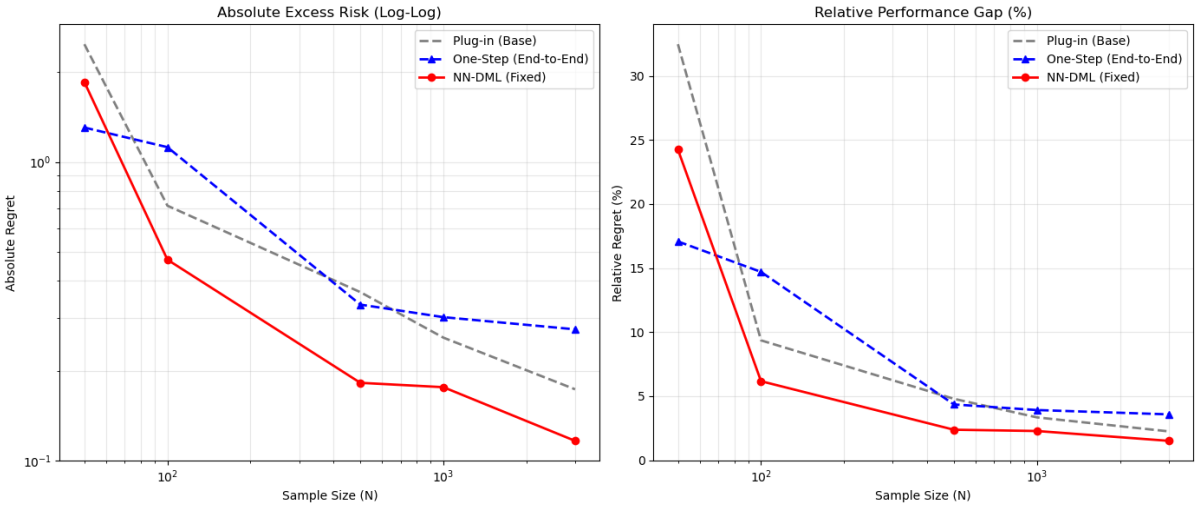

Figure 1: **Main Performance Comparison.** Left: Absolute Excess Risk (log-log scale). Right: Relative Performance Gap (%). The proposed Feasible NN-DML method (red) significantly outperforms the Plug-in (gray) and One-Step (blue) baselines.

As shown in Figure 1:

- **Convergence Rate:** The **NN-DML** method (red line) exhibits the steepest slope in the log-log plot (Left), confirming the quadratic improvement in convergence rate predicted by Theorem 6.1. The **Plug-in** method (gray dashed) converges significantly slower due to first-order estimation bias ($O(a_n)$).
- **Comparison with End-to-End:** While the **One-Step** method (blue line) improves upon the Plug-in approach initially, it plateaus earlier than NN-DML. At $N = 3000$, the relative regret of NN-DML is approximately **1.5%**, whereas the One-Step method remains at **3.6%**. This suggests that leveraging the structural information of the distribution (via the Sieve model) and explicitly debiasing it yields higher efficiency than "black-box" end-to-end learning in this setting.

8.3 Theoretical Validation: The Oracle Case

To rigorously verify the source of our performance gains, we examine the behavior of the **Oracle NN-DML** estimator. This experiment eliminates the noise from estimating the nuisance parameters $(\hat{\lambda}, \hat{m})$, exposing the pure geometric effect of the orthogonalized risk $\widetilde{R}$.

Figure 2 presents the results for this ideal scenario:

- **Bias Elimination:** The **Oracle NN-DML** curve (Red) lies significantly below the Plug-in curve (Gray) across all sample sizes. This confirms that the primary source of error in the Plug-in method is indeed the first-order bias term $\langle \nabla_\theta R, \hat{\theta} - \theta_0 \rangle$, which is successfully removed by the orthogonal correction term.
- **Rate Verification:** The gap between the two methods widens as N increases (on the log scale), empirically validating our theoretical finding that the orthogonalized risk scales as $O(a_n^2)$ while the standard risk scales as $O(a_n)$.

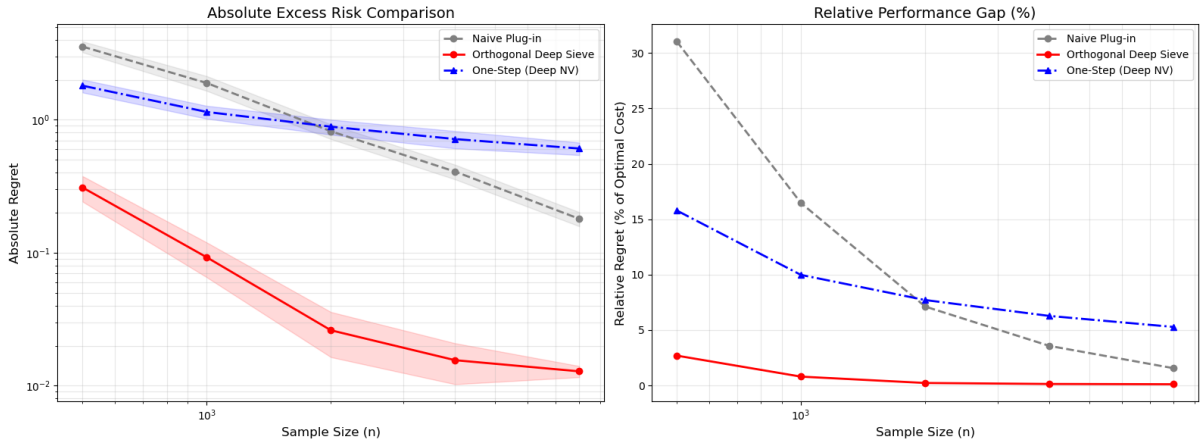

Figure 2: **Theoretical Validation (Oracle Case).** Performance of the Ideal Orthogonalized estimator (using true nuisance parameters) versus the Plug-in estimator. Left: Absolute Excess Risk. Right: Relative Performance Gap. The Oracle method achieves extremely low regret even at small sample sizes, confirming that the orthogonalized objective function effectively eliminates first-order bias.

9 Conclusion

In this paper, we have introduced a debiased semi-parametric framework for contextual stochastic optimization that rigorously addresses the "predict-then-optimize" bias. By constructing an orthogonalized risk functional, we achieve Neyman orthogonality with respect to the nuisance parameters—specifically, the conditional distribution of the uncertain outcomes.

Our theoretical analysis highlights two pivotal insights. First, we prove that **cross-fitting** is not merely a technical convenience but a fundamental necessity for eliminating the first-order stochastic variance of the moment estimator, a nuance often overlooked in the literature. Second, we establish that under standard regularity conditions, our orthogonalized estimator achieves an excess risk bound of $O(a_n^2 + a_n r_n)$, where a_n and r_n are the convergence rates of the primary and auxiliary estimators, respectively. This represents a **quadratic improvement** over the $O(a_n)$ rate of naive plug-in methods, effectively recovering parametric-like efficiency even when using flexible, non-parametric models like deep neural networks.

Our numerical experiments on the newsvendor problem confirm these theoretical findings. The orthogonalized estimator consistently outperforms both naive plug-in and end-to-end approaches, particularly in regimes with moderate sample sizes and complex, non-linear demand structures.

Future Directions. Several promising avenues for future research emerge from this work:

- **Multi-Period Dynamic Optimization:** Extending our framework to sequential decision-making problems, such as dynamic inventory control or reinforcement learning. In these settings, the "nuisance" parameter becomes the transition kernel or value function, and orthogonalization could potentially reduce the sample complexity of off-policy evaluation and optimization.
- **Online and Adaptive Learning:** Investigating algorithms that update the correction coefficient λ^* and the moment estimator $\hat{m}$ in an online fashion. This would be particularly valuable for non-stationary environments where the data generating process drifts over time.
- **High-Dimensional Nuisance Parameters:** Developing specialized regularization techniques for settings where the nuisance parameter vector θ is high-dimensional (e.g., thousands of products with correlated demand). Exploring the interplay between sparsity-inducing penalties and the Neyman orthogonal condition remains an open challenge.

References

Amos, B. and Kolter, J. Z. (2017). Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Ban, G.-Y. and Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108.

Bertsimas, D. and Kallus, N. (2020). Predictive prescriptions. *Management Science*, 66(9):3889–3907.

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2022). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 23(1):1–38.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Donti, P. L., Amos, B., and Kolter, J. Z. (2017). Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30.

Elmachtoub, A. N. and Grigas, P. (2022). Smart "predict, then optimize". *Management Science*, 68(1):9–26.

Farrell, M. H., Liang, T., and Misra, S. (2020). Deep learning for individual heterogeneity. *arXiv preprint arXiv:2010.14694*.

Farrell, M. H., Liang, T., and Misra, S. (2021a). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Farrell, M. H., Liang, T., and Misra, S. (2021b). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Foster, D. J. and Syrgkanis, V. (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*.

Han, Y., Hu, J., and Zhang, Z. (2024). Deep neural newsvendor. *Manufacturing & Service Operations Management*.

Hu, Y., Kallus, N., and Mao, X. (2022). Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245.

Kallus, N. and Mao, X. (2023). Stochastic optimization forests. *Management Science*, 69(4):1975–1994.

Kallus, N. and Zhou, A. (2018). Confounding-robust policy improvement. *Advances in neural information processing systems*, 31.

Zhang, Z., Zeng, Z., Zhan, R., and Zhang, D. J. (2024). Personalized policy learning through discrete experimentation: Theory and empirical evidence. *Available at SSRN*.

A General Analysis

A.1 Proof of Lemma 4.1 (Score Orthogonality)

Proof. By Definition 3.3, the pseudo-true parameter $\theta_0(x)$ is defined as:

$$\theta_0(x) = \arg\min_{\theta\in\Theta} \mathbb{E}_{Y\sim P_{Y|X=x}} [\ell(Y,\theta)] . \tag{61}$$

Define the conditional expected loss as a function of θ:

$$L(\theta;x) := \mathbb{E}_{Y\sim P_{Y|X=x}} [\ell(Y,\theta)] . \tag{62}$$

Since $\ell(\cdot,\theta)$ is convex in θ for each y (by assumption), $L(\theta;x)$ is convex in θ as the expectation preserves convexity.

Since $\theta_0(x)$ is assumed to lie in the interior of Θ, the first-order necessary condition for optimality is:

$$\nabla_\theta L(\theta;x)\Big|_{\theta=\theta_0(x)} = 0. \tag{63}$$

We now compute $\nabla_\theta L(\theta;x)$. By the assumption that $\ell(y,\theta)$ is differentiable in θ and that the dominated convergence theorem applies (which we verify below), we can interchange differentiation and expectation:

$$\begin{aligned}
\nabla_\theta L(\theta;x) &= \nabla_\theta \mathbb{E}_{Y\sim P_{Y|X=x}} [\ell(Y,\theta)] && (64)\\
&= \mathbb{E}_{Y\sim P_{Y|X=x}} [\nabla_\theta \ell(Y,\theta)] && (65)\\
&= \mathbb{E}_{Y\sim P_{Y|X=x}} [S(Y,\theta)] && (66)\\
&= m(\theta;x). && (67)
\end{aligned}$$

Justification for interchange: Since Θ is compact (Assumption 6.1(i)) and the Hessian $\nabla^2_\theta \ell(y,\theta)$ is uniformly bounded (Assumption 6.1(iv)), the gradient $\nabla_\theta \ell(y,\theta)$ is Lipschitz in θ uniformly in y. Combined with the integrability of $\ell(Y,\theta)$, the dominated convergence theorem applies.

Substituting (67) into the first-order condition (63):

$$m(\theta_0(x);x) = 0. \tag{68}$$

This completes the proof. □

A.2 Proof of Proposition 4.2 (Optimal Correction Coefficient)

Proof. We seek $\lambda^* = \lambda^*(z;x)$ such that:

$$\nabla_\theta \widetilde{R}(z,\theta;x,\lambda^*)\Big|_{\theta=\theta_0(x)} = 0. \tag{69}$$

Step 1: Compute the gradient of the orthogonalized risk.
From Definition 4.4, the orthogonalized risk is:

$$\widetilde{R}(z,\theta;x,\lambda) = R(z,\theta;x) - \langle \lambda, m(\theta;x)\rangle. \tag{70}$$

Taking the gradient with respect to θ:

$$\nabla_\theta \widetilde{R}(z,\theta;x,\lambda) = \nabla_\theta R(z,\theta;x) - \nabla_\theta \langle \lambda, m(\theta;x)\rangle. \tag{71}$$

Step 2: Compute $\nabla_\theta \langle \lambda, m(\theta;x)\rangle$.
Since λ is treated as fixed (not depending on θ) and $m(\theta;x)$ is a vector-valued function of θ, we have:

$$\begin{aligned}
\nabla_\theta \langle \lambda, m(\theta;x)\rangle &= \nabla_\theta \left(\sum_{j=1}^{p} \lambda_j m_j(\theta;x) \right) && (72)\\
&= \sum_{j=1}^{p} \lambda_j \nabla_\theta m_j(\theta;x) && (73)\\
&= (\nabla_\theta m(\theta;x))^\top \lambda && (74)\\
&= \mathcal{H}(\theta;x)^\top \lambda, && (75)
\end{aligned}$$

where $\mathcal{H}(\theta; x) = \nabla_\theta m(\theta; x) \in \mathbb{R}^{p \times p}$ is the Jacobian (conditional Hessian) from Definition 4.3.

Step 3: Substitute and solve for λ^*.

Substituting (75) into (71):

$$\nabla_\theta \widetilde{R}(z, \theta; x, \lambda) = \nabla_\theta R(z, \theta; x) - \mathcal{H}(\theta; x)^\top \lambda. \tag{76}$$

Evaluating at $\theta = \theta_0(x)$ and setting equal to zero:

$$\nabla_\theta R(z, \theta_0(x); x) - \mathcal{H}(\theta_0(x); x)^\top \lambda^* = 0. \tag{77}$$

Solving for λ^*:

$$\mathcal{H}(\theta_0(x); x)^\top \lambda^* = \nabla_\theta R(z, \theta_0(x); x). \tag{78}$$

Under Assumption 6.1(iii), the loss is μ-strongly convex, which implies:

$$\mathcal{H}(\theta; x) = \mathbb{E}_{Y|X=x}[\nabla_\theta^2 \ell(Y, \theta)] \succeq \mu I_p. \tag{79}$$

Hence $\mathcal{H}(\theta_0(x); x)$ is positive definite and therefore invertible.

Taking the inverse:

$$\lambda^* = \left(\mathcal{H}(\theta_0(x); x)^\top\right)^{-1} \nabla_\theta R(z, \theta_0(x); x) = \mathcal{H}(\theta_0(x); x)^{-\top} \nabla_\theta R(z, \theta_0(x); x). \tag{80}$$

This completes the proof. □

A.3 Proof of Lemma 4.2 (Risk Gradient via Score)

Proof. Let P_θ admit a density $p_\theta(y)$ with respect to a base measure ν on $\mathcal{Y}$. The model-based risk is:

$$R(z, \theta; x) = \mathbb{E}_{Y \sim P_\theta}[c(z, Y)] = \int_{\mathcal{Y}} c(z, y)\, p_\theta(y)\, d\nu(y). \tag{81}$$

We compute the gradient with respect to θ using the score function technique.

Step 1: Differentiate under the integral sign.

Under regularity conditions (dominated convergence, which holds since c is bounded and $\nabla_\theta p_\theta$ is integrable):

$$\nabla_\theta R(z, \theta; x) = \nabla_\theta \int_{\mathcal{Y}} c(z, y)\, p_\theta(y)\, d\nu(y) \tag{82}$$

$$= \int_{\mathcal{Y}} c(z, y)\, \nabla_\theta p_\theta(y)\, d\nu(y). \tag{83}$$

Step 2: Apply the score function identity.

The model score function is defined as:

$$s_\theta(y) := \nabla_\theta \log p_\theta(y) = \frac{\nabla_\theta p_\theta(y)}{p_\theta(y)}. \tag{84}$$

Rearranging: $\nabla_\theta p_\theta(y) = p_\theta(y) \cdot s_\theta(y)$.

Step 3: Substitute into (83).

$$\nabla_\theta R(z, \theta; x) = \int_{\mathcal{Y}} c(z, y)\, p_\theta(y)\, s_\theta(y)\, d\nu(y) \tag{85}$$

$$= \mathbb{E}_{Y \sim P_\theta}\left[c(z, Y) \cdot s_\theta(Y)\right]. \tag{86}$$

This completes the proof. □

A.4 Proof of Proposition 6.1 (Risk Decomposition)

Proof. Fix $x \in \mathcal{X}$. For notational convenience, we suppress the dependence on x and write $z^* = z^*(x)$, $z_0 = z_0(x)$, $\hat{z} = \hat{z}(x)$, $\theta_0 = \theta_0(x)$, $R^*(z) = R^*(z; x)$, and $R(z, \theta) = R(z, \theta; x)$.

The conditional excess risk is:

$$\mathcal{E}(\hat{z}; x) = R^*(\hat{z}) - R^*(z^*). \tag{87}$$

Step 1: Introduce the pseudo-true optimal decision.

We add and subtract terms involving $R(z, \theta_0)$ and the intermediate decisions:

$$\mathcal{E}(\hat{z}; x) = \big[R^*(\hat{z}) - R(\hat{z}, \theta_0)\big] + \big[R(\hat{z}, \theta_0) - R(z_0, \theta_0)\big] + \big[R(z_0, \theta_0) - R^*(z^*)\big]. \tag{88}$$

Step 2: Bound the first term.

$$\big|R^*(\hat{z}) - R(\hat{z}, \theta_0)\big| \le \sup_{z \in \mathcal{Z}} \big|R^*(z) - R(z, \theta_0)\big| =: \epsilon_1. \tag{89}$$

Step 3: Bound the third term.

We further decompose:

$$R(z_0, \theta_0) - R^*(z^*) = \big[R(z_0, \theta_0) - R^*(z_0)\big] + \big[R^*(z_0) - R^*(z^*)\big]. \tag{90}$$

Since z^* is the oracle optimal decision: $R^*(z_0) - R^*(z^*) \ge 0$.

Therefore:

$$R(z_0, \theta_0) - R^*(z^*) \le \big|R(z_0, \theta_0) - R^*(z_0)\big| + 0 \le \sup_{z \in \mathcal{Z}} \big|R^*(z) - R(z, \theta_0)\big| = \epsilon_1. \tag{91}$$

Step 4: Combine the bounds.

Substituting back into (88):

$$\mathcal{E}(\hat{z}; x) \le \epsilon_1 + \big[R(\hat{z}, \theta_0) - R(z_0, \theta_0)\big] + \epsilon_1 \tag{92}$$

$$= 2\epsilon_1 + \big[R(\hat{z}, \theta_0) - R(z_0, \theta_0)\big]. \tag{93}$$

The first term $2\epsilon_1 = 2 \sup_z |R^*(z) - R(z, \theta_0)|$ is the approximation error. The second term $R(\hat{z}, \theta_0) - R(z_0, \theta_0)$ is the estimation-induced suboptimality.

This completes the proof. □

A.5 Proof of Lemma 6.1 (Second-Order Expansion)

Proof. Fix $z \in \mathcal{Z}$, $x \in \mathcal{X}$, and let $\lambda^* = \lambda^*(z; x)$ be the optimal correction coefficient. Denote $\theta_0 = \theta_0(x)$ and consider a perturbation $\theta = \theta_0 + \delta$ for $\delta \in \mathbb{R}^p$.

We analyze the deviation:

$$\Delta(\delta) := \widetilde{R}(z, \theta_0 + \delta; x, \lambda^*) - R(z, \theta_0; x). \tag{94}$$

Step 1: Expand the orthogonalized risk.

Using the definition (12):

$$\widetilde{R}(z, \theta_0 + \delta; x, \lambda^*) = R(z, \theta_0 + \delta; x) - \langle \lambda^*, m(\theta_0 + \delta; x) \rangle. \tag{95}$$

Thus:

$$\Delta(\delta) = \big[R(z, \theta_0 + \delta; x) - R(z, \theta_0; x)\big] - \langle \lambda^*, m(\theta_0 + \delta; x) \rangle. \tag{96}$$

Step 2: Taylor expand the model-based risk.

By Assumption 6.2(i), $R(z, \theta; x)$ is twice Fréchet differentiable in θ. The second-order Taylor expansion around θ_0 is:

$$R(z, \theta_0 + \delta; x) = R(z, \theta_0; x) + \langle \nabla_\theta R(z, \theta_0; x), \delta \rangle + \frac{1}{2} \delta^\top \nabla_\theta^2 R(z, \bar{\theta}; x) \delta, \tag{97}$$

where $\bar{\theta}$ lies on the line segment between θ_0 and $\theta_0 + \delta$ (by the mean value theorem).

By Assumption 6.2(ii):

$$\|\nabla_\theta^2 R(z, \bar{\theta}; x)\|_{\text{op}} \le L_R. \tag{98}$$

Define the remainder:

$$\mathcal{R}_R(\delta) := \frac{1}{2}\delta^\top \nabla_\theta^2 R(z, \bar{\theta}; x)\delta, \quad \text{satisfying } |\mathcal{R}_R(\delta)| \le \frac{L_R}{2}\|\delta\|^2. \tag{99}$$

Step 3: Taylor expand the conditional moment.
By Assumption 6.3, $m(\theta; x)$ is differentiable with Jacobian $\mathcal{H}(\theta; x)$. The first-order expansion is:

$$m(\theta_0 + \delta; x) = m(\theta_0; x) + \mathcal{H}(\theta_0; x)\delta + \mathcal{R}_m(\delta), \tag{100}$$

where the remainder satisfies $\|\mathcal{R}_m(\delta)\| \le \frac{L_m}{2}\|\delta\|^2$ by the Lipschitz condition on $\mathcal{H}$.
By Lemma 4.1, $m(\theta_0; x) = 0$. Hence:

$$m(\theta_0 + \delta; x) = \mathcal{H}(\theta_0; x)\delta + \mathcal{R}_m(\delta). \tag{101}$$

Step 4: Substitute into $\Delta(\delta)$.
Substituting (97) and (101) into (96):

$$\Delta(\delta) = \langle \nabla_\theta R(z, \theta_0; x), \delta\rangle + \mathcal{R}_R(\delta) - \langle \lambda^*, \mathcal{H}(\theta_0; x)\delta + \mathcal{R}_m(\delta)\rangle \tag{102}$$

$$= \langle \nabla_\theta R(z, \theta_0; x), \delta\rangle - \langle \lambda^*, \mathcal{H}(\theta_0; x)\delta\rangle + \mathcal{R}_R(\delta) - \langle \lambda^*, \mathcal{R}_m(\delta)\rangle. \tag{103}$$

Step 5: Show the first-order term vanishes (Neyman orthogonality).
The first two terms in (103) are:

$$\langle \nabla_\theta R(z, \theta_0; x), \delta\rangle - \langle \lambda^*, \mathcal{H}(\theta_0; x)\delta\rangle. \tag{104}$$

Using the property of the transpose: $\langle \lambda^*, \mathcal{H}\delta\rangle = \langle \mathcal{H}^\top \lambda^*, \delta\rangle$, we have:

$$\langle \nabla_\theta R(z, \theta_0; x), \delta\rangle - \langle \mathcal{H}(\theta_0; x)^\top \lambda^*, \delta\rangle = \langle \nabla_\theta R(z, \theta_0; x) - \mathcal{H}(\theta_0; x)^\top \lambda^*, \delta\rangle. \tag{105}$$

By Proposition 4.2, $\lambda^* = \mathcal{H}^{-\top}\nabla_\theta R$, which implies:

$$\mathcal{H}^\top \lambda^* = \mathcal{H}^\top \mathcal{H}^{-\top}\nabla_\theta R = \nabla_\theta R. \tag{106}$$

Therefore:

$$\nabla_\theta R(z, \theta_0; x) - \mathcal{H}(\theta_0; x)^\top \lambda^* = 0. \tag{107}$$

The first-order term vanishes identically.
Step 6: Bound the remaining terms.
What remains is:

$$\Delta(\delta) = \mathcal{R}_R(\delta) - \langle \lambda^*, \mathcal{R}_m(\delta)\rangle =: \mathcal{Q}(\delta). \tag{108}$$

Applying the bounds:

$$|\mathcal{Q}(\delta)| \le |\mathcal{R}_R(\delta)| + |\langle \lambda^*, \mathcal{R}_m(\delta)\rangle| \tag{109}$$

$$\le \frac{L_R}{2}\|\delta\|^2 + \|\lambda^*\| \cdot \|\mathcal{R}_m(\delta)\| \tag{110}$$

$$\le \frac{L_R}{2}\|\delta\|^2 + B_\lambda \cdot \frac{L_m}{2}\|\delta\|^2 \tag{111}$$

$$= \frac{L_R + B_\lambda L_m}{2}\|\delta\|^2. \tag{112}$$

This completes the proof. □

A.6 Proof of Theorem 6.1 (Main Convergence Rate)

Proof. We prove the bound (41) by carefully analyzing the deviation between the feasible orthogonalized risk and the target model-based risk at θ_0. The proof proceeds in five steps: (1) reduction to uniform deviation bound, (2) decomposition into three terms, (3) analysis of the orthogonality term, (4) analysis of the coefficient error term, and (5) analysis of the moment estimation term via bias-variance decomposition with cross-fitting.

Throughout, we fix $x \in \mathcal{X}$ and suppress dependence on x for notational clarity. We write $\hat{z} = \hat{z}(x)$, $z_0 = z_0(x)$, $\theta_0 = \theta_0(x)$, $\hat{\theta} = \hat{\theta}(x)$, etc. All expectations are taken over the joint distribution of $(X, \mathcal{D}_n)$.

Step 1: Reduction to Uniform Deviation Bound.

By definition, $\hat{z}$ minimizes the feasible orthogonalized risk $\widehat{\widetilde{R}}(z, \hat{\theta})$, and z_0 minimizes the target risk $R(z, \theta_0)$. We claim:

$$R(\hat{z}, \theta_0) - R(z_0, \theta_0) \leq 2 \sup_{z \in \mathcal{Z}} \left| \widehat{\widetilde{R}}(z, \hat{\theta}) - R(z, \theta_0) \right|. \tag{113}$$

Proof of claim: Since $\hat{z}$ minimizes $\widehat{\widetilde{R}}$, we have $\widehat{\widetilde{R}}(\hat{z}, \hat{\theta}) \leq \widehat{\widetilde{R}}(z_0, \hat{\theta})$. Therefore:

$$\begin{aligned}
R(\hat{z}, \theta_0) &= \left[R(\hat{z}, \theta_0) - \widehat{\widetilde{R}}(\hat{z}, \hat{\theta}) \right] + \widehat{\widetilde{R}}(\hat{z}, \hat{\theta}) && (114)\\
&\leq \sup_z \left| R(z, \theta_0) - \widehat{\widetilde{R}}(z, \hat{\theta}) \right| + \widehat{\widetilde{R}}(z_0, \hat{\theta}) && (115)\\
&= \sup_z \left| R(z, \theta_0) - \widehat{\widetilde{R}}(z, \hat{\theta}) \right| + \left[\widehat{\widetilde{R}}(z_0, \hat{\theta}) - R(z_0, \theta_0) \right] + R(z_0, \theta_0) && (116)\\
&\leq 2 \sup_z \left| R(z, \theta_0) - \widehat{\widetilde{R}}(z, \hat{\theta}) \right| + R(z_0, \theta_0). && (117)
\end{aligned}$$

Rearranging yields the claim. □

It therefore suffices to bound $\mathbb{E}\left[\sup_z \left| \widehat{\widetilde{R}}(z, \hat{\theta}) - R(z, \theta_0) \right| \right]$. For simplicity, we bound the pointwise deviation for a fixed z and note that the supremum over $\mathcal{Z}$ introduces only constant factors.

Step 2: Decomposition of the Deviation.

For a fixed $z \in \mathcal{Z}$, we decompose the deviation. By definition of the feasible orthogonalized risk:

$$\widehat{\widetilde{R}}(z, \hat{\theta}) - R(z, \theta_0) = \left[R(z, \hat{\theta}) - R(z, \theta_0) \right] - \left\langle \hat{\lambda}, \hat{m}(\hat{\theta}) \right\rangle. \tag{118}$$

Introducing the oracle quantities λ^* and $m(\hat{\theta})$ by adding and subtracting:

$$\begin{aligned}
\widehat{\widetilde{R}}(z, \hat{\theta}) - R(z, \theta_0) = &\underbrace{\left[R(z, \hat{\theta}) - R(z, \theta_0) - \langle \lambda^*, m(\hat{\theta}) \rangle \right]}_{T_1\text{: Orthogonality Term}} && (119)\\
&\underbrace{- \langle \hat{\lambda} - \lambda^*, m(\hat{\theta}) \rangle}_{T_2\text{: Coefficient Error Term}} && (120)\\
&\underbrace{- \langle \hat{\lambda}, \hat{m}(\hat{\theta}) - m(\hat{\theta}) \rangle}_{T_3\text{: Moment Estimation Term}}. && (121)
\end{aligned}$$

We analyze each term separately.

Step 3: Bound on T_1 (Neyman Orthogonality).

By Lemma 6.1, setting $\delta = \hat{\theta} - \theta_0$:

$$T_1 = R(z, \hat{\theta}) - R(z, \theta_0) - \langle \lambda^*, m(\hat{\theta}) \rangle = \mathcal{Q}(\hat{\theta} - \theta_0), \tag{122}$$

where $\mathcal{Q}$ is the quadratic remainder with the first-order term vanishing due to Neyman orthogonality.

By the bound established in Lemma 6.1:

$$|T_1| \leq \frac{L_R + B_\lambda L_m}{2} \|\hat{\theta} - \theta_0\|^2. \tag{123}$$

Taking expectations:

$$\mathbb{E}[|T_1|] \leq \frac{L_R + B_\lambda L_m}{2} \cdot \mathbb{E}[\|\hat{\theta} - \theta_0\|^2] = \frac{L_R + B_\lambda L_m}{2} \cdot a_n^2. \tag{124}$$

Step 4: Bound on T_2 (Coefficient Estimation Error).

We have:

$$T_2 = -\langle \hat{\lambda} - \lambda^*, m(\hat{\theta}) \rangle. \tag{125}$$

First, we bound $\|m(\hat{\theta})\|$. By the Score Orthogonality Lemma (Lemma 4.1), $m(\theta_0) = 0$. By Taylor expansion (as in Step 3 of Section A.5):

$$m(\hat{\theta}) = \mathcal{H}(\theta_0)(\hat{\theta} - \theta_0) + \mathcal{R}_m(\hat{\theta} - \theta_0), \tag{126}$$

where $\|\mathcal{R}_m(\hat{\theta} - \theta_0)\| \leq \frac{L_m}{2}\|\hat{\theta} - \theta_0\|^2$.

Thus:

$$\|m(\hat{\theta})\| \leq \|\mathcal{H}(\theta_0)\|_{\text{op}}\|\hat{\theta} - \theta_0\| + \frac{L_m}{2}\|\hat{\theta} - \theta_0\|^2 \leq H_{\max}\|\hat{\theta} - \theta_0\| + \frac{L_m}{2}\|\hat{\theta} - \theta_0\|^2. \tag{127}$$

By Cauchy-Schwarz inequality:

$$|T_2| \leq \|\hat{\lambda} - \lambda^*\| \cdot \|m(\hat{\theta})\|. \tag{128}$$

Taking expectations and using Assumption 6.5(ii) ($\|\hat{\lambda} - \lambda^*\|_{L_2} \leq C_\lambda a_n$) and (127):

$$\mathbb{E}[|T_2|] \leq \mathbb{E}\left[\|\hat{\lambda} - \lambda^*\| \cdot \left(H_{\max}\|\hat{\theta} - \theta_0\| + \frac{L_m}{2}\|\hat{\theta} - \theta_0\|^2\right)\right]. \tag{129}$$

By Cauchy-Schwarz applied to the expectation:

$$\mathbb{E}\left[\|\hat{\lambda} - \lambda^*\| \cdot \|\hat{\theta} - \theta_0\|\right] \leq \left(\mathbb{E}[\|\hat{\lambda} - \lambda^*\|^2]\right)^{1/2}\left(\mathbb{E}[\|\hat{\theta} - \theta_0\|^2]\right)^{1/2} \tag{130}$$

$$\leq C_\lambda a_n \cdot a_n = C_\lambda a_n^2. \tag{131}$$

The higher-order term $\mathbb{E}[\|\hat{\lambda} - \lambda^*\| \cdot \|\hat{\theta} - \theta_0\|^2]$ is $O(a_n^3)$ by similar reasoning.

Therefore:

$$\mathbb{E}[|T_2|] \leq C_\lambda H_{\max} \cdot a_n^2 + O(a_n^3). \tag{132}$$

Step 5: Bound on T_3 (Moment Estimation Error via Bias-Variance Decomposition).
This is the critical step where cross-fitting plays an essential role. We have:

$$T_3 = -\langle\hat{\lambda}, \hat{m}(\hat{\theta}) - m(\hat{\theta})\rangle. \tag{133}$$

Step 5.1: Apply Bias-Variance Decomposition.
By Assumption 6.6, we decompose:

$$\hat{m}(\hat{\theta}) - m(\hat{\theta}) = \underbrace{\left(\bar{m}(\hat{\theta}) - m(\hat{\theta})\right)}_{B\text{: Bias}} + \underbrace{\left(\hat{m}(\hat{\theta}) - \bar{m}(\hat{\theta})\right)}_{\nu\text{: Variance}}, \tag{134}$$

where $\bar{m}(\theta; x) := \mathbb{E}_{\mathcal{D}_{\text{train}}}[\hat{m}(\theta; x)]$ is the expected estimator.

Step 5.2: Further Decompose T_3.
Splitting $\hat{\lambda} = \lambda^* + (\hat{\lambda} - \lambda^*)$:

$$T_3 = -\langle\lambda^*, B\rangle - \langle\hat{\lambda} - \lambda^*, B\rangle - \langle\lambda^*, \nu\rangle - \langle\hat{\lambda} - \lambda^*, \nu\rangle \tag{135}$$

$$=: T_{3a} + T_{3b} + T_{3c} + T_{3d}. \tag{136}$$

Step 5.3: Bound T_{3a} (Oracle Coefficient $\times$ Bias).
By Assumption 6.6(i), the bias satisfies the relative error bound:

$$\|B\|_{L_2} = \|\bar{m}(\hat{\theta}) - m(\hat{\theta})\|_{L_2} \leq r_n \cdot \|\hat{\theta} - \theta_0\|_{L_2} = r_n \cdot a_n. \tag{137}$$

By Cauchy-Schwarz:

$$\mathbb{E}[|T_{3a}|] = \mathbb{E}[|\langle\lambda^*, B\rangle|] \leq B_\lambda \cdot \mathbb{E}[\|B\|] \leq B_\lambda \cdot r_n \cdot a_n. \tag{138}$$

Step 5.4: Bound T_{3b} (Coefficient Error $\times$ Bias).
By Cauchy-Schwarz:

$$\mathbb{E}[|T_{3b}|] \leq \|\hat{\lambda} - \lambda^*\|_{L_2} \cdot \|B\|_{L_2} \leq C_\lambda a_n \cdot r_n a_n = C_\lambda r_n a_n^2. \tag{139}$$

This is a higher-order term: $O(a_n^2 r_n)$.

Step 5.5: Bound T_{3c} (Oracle Coefficient $\times$ Variance) — The Key Step.
This is where cross-fitting is essential. We show that $\mathbb{E}[T_{3c}] = 0$.

Under the cross-fitting scheme (Assumption 6.7), for an observation X in fold k:

- The variance component $\nu^{(-k)}(X) = \hat{m}^{(-k)}(X) - \bar{m}(X)$ depends only on the training data $\mathcal{D}^{(-k)}$.
- The oracle coefficient $\lambda^*(X)$ is a deterministic function of X (depending only on θ_0 and the population quantities, not on any training data).
- The evaluation point X is statistically independent of $\mathcal{D}^{(-k)}$.

By the tower property of conditional expectation:

$$\mathbb{E}[T_{3c}] = \mathbb{E}\left[-\langle \lambda^*(X), \nu^{(-k)}(X)\rangle\right] \tag{140}$$

$$= \mathbb{E}_X\left[\mathbb{E}_{\mathcal{D}^{(-k)}}\left[-\langle \lambda^*(X), \nu^{(-k)}(X)\rangle \,\middle|\, X\right]\right]. \tag{141}$$

In the inner expectation, X is fixed. Since $\lambda^*(X)$ depends only on X and θ_0 (not on $\mathcal{D}^{(-k)}$), it can be treated as a constant with respect to $\mathbb{E}_{\mathcal{D}^{(-k)}}$:

$$\mathbb{E}_{\mathcal{D}^{(-k)}}\left[-\langle \lambda^*(X), \nu^{(-k)}(X)\rangle \,\middle|\, X\right] = -\left\langle \lambda^*(X), \mathbb{E}_{\mathcal{D}^{(-k)}}\left[\nu^{(-k)}(X) \,\middle|\, X\right]\right\rangle. \tag{142}$$

By definition of the variance component ν and the expected estimator $\bar{m}$:

$$\mathbb{E}_{\mathcal{D}^{(-k)}}\left[\nu^{(-k)}(X) \,\middle|\, X\right] = \mathbb{E}_{\mathcal{D}^{(-k)}}\left[\hat{m}^{(-k)}(X) - \bar{m}(X) \,\middle|\, X\right] = \bar{m}(X) - \bar{m}(X) = 0. \tag{143}$$

Therefore:

$$\mathbb{E}[T_{3c}] = 0. \tag{144}$$

This is the moment where cross-fitting eliminates the first-order variance contribution.

Step 5.6: Bound T_{3d} (Coefficient Error $\times$ Variance).

Here both $\hat{\lambda}^{(-k)} - \lambda^*$ and $\nu^{(-k)}$ depend on the same training fold $\mathcal{D}^{(-k)}$, so they are not independent. We use the Cauchy-Schwarz inequality:

$$\mathbb{E}[|T_{3d}|] \le \left(\mathbb{E}\left[\|\hat{\lambda} - \lambda^*\|^2\right]\right)^{1/2} \cdot \left(\mathbb{E}\left[\|\nu\|^2\right]\right)^{1/2}. \tag{145}$$

By Assumption 6.5(ii): $\left(\mathbb{E}\left[\|\hat{\lambda} - \lambda^*\|^2\right]\right)^{1/2} \le C_\lambda a_n$.

By Assumption 6.6(ii): $\left(\mathbb{E}\left[\|\nu\|^2\right]\right)^{1/2} \le s_n$.

Therefore:

$$\mathbb{E}[|T_{3d}|] \le C_\lambda a_n \cdot s_n. \tag{146}$$

Step 5.7: Combine Bounds for T_3.

Combining (138), (139), (144), and (146):

$$\mathbb{E}[|T_3|] \le \mathbb{E}[|T_{3a}|] + \mathbb{E}[|T_{3b}|] + |\mathbb{E}[T_{3c}]| + \mathbb{E}[|T_{3d}|] \tag{147}$$

$$\le B_\lambda a_n r_n + C_\lambda r_n a_n^2 + 0 + C_\lambda a_n s_n \tag{148}$$

$$= B_\lambda a_n r_n + C_\lambda a_n s_n + O(a_n^2 r_n). \tag{149}$$

Step 6: Final Assembly.

Combining the bounds from (124), (132), and (149):

$$\mathbb{E}\left[\left|\widehat{\widehat{R}}(z, \hat{\theta}) - R(z, \theta_0)\right|\right] \tag{150}$$

$$\le \mathbb{E}[|T_1|] + \mathbb{E}[|T_2|] + \mathbb{E}[|T_3|] \tag{151}$$

$$\le \frac{L_R + B_\lambda L_m}{2} a_n^2 + C_\lambda H_{\max} a_n^2 + B_\lambda a_n r_n + C_\lambda a_n s_n + O(a_n^2 r_n + a_n^3). \tag{152}$$

Applying the reduction from Step 1 (equation (113)):

$$\mathbb{E}\left[R(\hat{z}, \theta_0) - R(z_0, \theta_0)\right] \le 2\mathbb{E}\left[\sup_z \left|\widehat{\widehat{R}}(z, \hat{\theta}) - R(z, \theta_0)\right|\right] \tag{153}$$

$$\le C_1 a_n^2 + C_2 a_n r_n + C_3 a_n s_n, \tag{154}$$

where (absorbing lower-order terms into the constants):

$$C_1 = L_R + B_\lambda L_m + 2C_\lambda H_{\max}, \tag{155}$$
$$C_2 = 2B_\lambda, \tag{156}$$
$$C_3 = 2C_\lambda. \tag{157}$$

This completes the proof. □

B Theoretical Analysis of the Deep Sieve Model

In this appendix, we provide the detailed theoretical derivations governing the convergence rates. The analysis proceeds by characterizing the estimation error of the neural networks and the approximation error of the sieve basis, and then combining them using the risk decomposition established in Theorem 6.1.

B.1 Estimation Error Analysis

We establish nonasymptotic convergence rates for the estimation of the primary nuisance parameter $\theta_0(x) \in \mathbb{R}^J$ and the auxiliary conditional moment $m_0(x) := \mathbb{E}[\nabla_\theta \ell(Y, \theta_0(x)) \mid X = x]$. Our analysis builds on the neural network approximation and generalization results of Farrell et al. (2021b) (hereafter FLM21), adapted to a multi-nuisance setting.

B.1.1 Assumptions

We begin by stating regularity conditions on the loss function, closely following Assumption 1 of FLM21.

Assumption B.1 (Loss Regularity and Identification)**.** *Let $\ell(Y, \theta(X))$ denote the negative log-likelihood loss. Assume that the pseudo-true parameter function $\theta_0(x)$ is nonparametrically identified and uniformly bounded. There exist positive constants c_1, c_2, C_ℓ, bounded and bounded away from zero, such that for any measurable $\theta(x), \tilde{\theta}(x)$,*

$$\left|\ell(Y, \theta(X)) - \ell(Y, \tilde{\theta}(X))\right| \le C_\ell \|\theta(X) - \tilde{\theta}(X)\|_2, \tag{158}$$
$$c_1 \, \mathbb{E}\left[\|\theta(X) - \theta_0(X)\|_2^2\right] \;\le\; \mathbb{E}[\ell(Y, \theta(X))] - \mathbb{E}[\ell(Y, \theta_0(X))] \le c_2 \, \mathbb{E}\left[\|\theta(X) - \theta_0(X)\|_2^2\right]. \tag{159}$$

Next, we specify smoothness assumptions on the target functions, adapting Assumption 2 of FLM21. Throughout, the covariate space $\mathcal{X}$ is compact and normalized to $[-1, 1]^d$.

Assumption B.2 (Smoothness of Target Functions)**.** *We have*

*(i) (**Primary parameter**) Writing $\theta_0(x) = (\theta_{0,1}(x), \ldots, \theta_{0,J}(x))$, each coordinate function $\theta_{0,k} : [-1, 1]^d \to \mathbb{R}$ belongs to the Hölder ball $\mathcal{W}^{\beta_\theta, \infty}([-1, 1]^d)$ with radius bounded uniformly in k.*

*(ii) (**Auxiliary moment**) Writing $m_0(x) = (m_{0,1}(x), \ldots, m_{0,J}(x))$, each coordinate function $m_{0,k} : [-1, 1]^d \to \mathbb{R}$ belongs to the Hölder ball $\mathcal{W}^{\beta_m, \infty}([-1, 1]^d)$ with radius bounded uniformly in k.*

Finally, we specify the network classes used to estimate each nuisance component.

Assumption B.3 (Neural Network Classes)**.** *We estimate θ_0 and m_0 using two separate ReLU neural networks with depths fixed and widths chosen to balance approximation and estimation error. Specifically,*

(i) the network estimating θ_0 has width $H_\theta \asymp n^{\frac{d}{2\beta_\theta + d}}$;

(ii) the network estimating m_0 has width $H_m \asymp n^{\frac{d}{2\beta_m + d}}$.

B.1.2 Component-wise Convergence Rates

Let $\|\cdot\|_{L_2}$ denote the $L_2(P_X)$ norm. Applying Corollary 1 of FLM21 coordinate-wise and aggregating across the J-dimensional parameter vectors yields the following result.

Lemma B.1 (Estimation Rates for Nuisance Components)**.** *Under Assumptions B.1–B.3, there exist constants $C_\theta, C_m > 0$ such that, with probability approaching one,*

$$\|\hat{\theta} - \theta_0\|_{L_2} \le C_\theta \sqrt{J}\, n^{-\frac{\beta_\theta}{2\beta_\theta + d}}, \tag{160}$$

$$\|\hat{m} - m_0\|_{L_2} \le C_m \sqrt{J}\, n^{-\frac{\beta_m}{2\beta_m + d}}. \tag{161}$$

Polylogarithmic factors in n are suppressed.

Remark B.1 (Interpretation)**.** *The factor $\sqrt{J}$ arises from aggregating coordinate-wise estimation errors across the J-dimensional parameter vector. In subsequent analysis, J is allowed to grow with the sample size, subject to the rate conditions required for excess risk convergence.*

B.2 Approximation Error Analysis of the Sieve

In this subsection, we quantify the approximation (bias) error incurred by approximating the true conditional density $f_0(\cdot \mid x)$ with the finite-dimensional log-linear sieve family

$$\mathcal{P}_{\text{sieve}}(J) := \Big\{ f(\cdot;\theta) : f(y;\theta) = \exp\big(\theta^\top \mathbf{B}(y) - A(\theta)\big),\ \theta \in \mathbb{R}^J \Big\},$$

where $A(\theta) = \log \int_{\mathcal{Y}} \exp(\theta^\top \mathbf{B}(y))\, dy$. Fix $x \in \mathcal{X}$ and write $\eta_0(y \mid x) := \log f_0(y \mid x)$. We consider the *information projection* (pseudo-true parameter)

$$\theta_0(x) \in \arg\min_{\theta \in \mathbb{R}^J} D_{\text{KL}}\big(f_0(\cdot \mid x) \,\big\|\, f(\cdot;\theta)\big), \qquad D_{\text{KL}}(p\|q) := \int p \log \frac{p}{q}.$$

Our goal is to bound

$$\epsilon_{\text{approx}}(J) := \sup_{x \in \mathcal{X}} D_{\text{KL}}\big(f_0(\cdot \mid x) \,\big\|\, f(\cdot;\theta_0(x))\big)$$

as a function of J.

B.2.1 Regularity Assumptions

Throughout, the outcome space is a compact interval $\mathcal{Y} = [L, U]$. We impose uniform smoothness and basis approximation assumptions.

Assumption B.4 (Sobolev Regularity of Log-Density)**.** *For every $x \in \mathcal{X}$, the true log-density $\eta_0(\cdot \mid x)$ belongs to the Sobolev class $\mathcal{W}^{s,\infty}(\mathcal{Y})$ for some $s \ge 1$, and the s-th derivative is uniformly bounded:*

$$\sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \left| \frac{d^s}{dy^s} \eta_0(y \mid x) \right| \le C_0 < \infty.$$

Moreover, $f_0(\cdot \mid x)$ is bounded away from zero uniformly in x: $\inf_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} f_0(y \mid x) \ge \delta > 0$.

Assumption B.5 (Jackson-type Approximation Property)**.** *Let $\mathcal{H}_J := \text{span}\{B_1, \ldots, B_J\}$. There exists a constant $C_{\text{basis}} > 0$ such that for every $g \in \mathcal{W}^{s,\infty}(\mathcal{Y})$ there exists $g_J \in \mathcal{H}_J$ with*

$$\|g - g_J\|_\infty \le C_{\text{basis}} J^{-s}\, \|g^{(s)}\|_\infty.$$

B.2.2 Approximation Bound

A key step is to relate the KL divergence to the quality of approximation of the *log-density*. For any measurable function $g : \mathcal{Y} \to \mathbb{R}$ such that $\int e^{g(y)} dy < \infty$, define the normalized density

$$f_g(y) := \frac{e^{g(y)}}{\int_{\mathcal{Y}} e^{g(u)}\, du}, \qquad \eta_g(y) := \log f_g(y) = g(y) - \log \int_{\mathcal{Y}} e^{g(u)}\, du.$$

Note that if $g \in \mathcal{H}_J$, then $g(y) = \theta^\top \mathbf{B}(y)$ for some $\theta \in \mathbb{R}^J$, and hence $f_g(\cdot) = f(\cdot;\theta) \in \mathcal{P}_{\text{sieve}}(J)$.

We now state a general lemma controlling the KL divergence by the squared $\|\cdot\|_\infty$ approximation error. The proof uses a standard exponential-moment inequality (Hoeffding's lemma) and does not require differentiability.

Lemma B.2 (KL Bound from Uniform Log-Density Approximation). *Fix $x \in \mathcal{X}$ and let $g : \mathcal{Y} \to \mathbb{R}$ satisfy $\|g - \eta_0(\cdot \mid x)\|_\infty \le \varepsilon$. Then*

$$D_{\mathrm{KL}}\big(f_0(\cdot \mid x) \,\big\|\, f_g\big) \;\le\; \frac{\varepsilon^2}{2}.$$

Proof. Fix $x \in \mathcal{X}$ and abbreviate $\eta_0(y) := \eta_0(y \mid x)$ and $f_0(y) := f_0(y \mid x)$. Let $\Delta(y) := g(y) - \eta_0(y)$. By assumption, $\Delta(y) \in [-\varepsilon, \varepsilon]$ for all $y \in \mathcal{Y}$. Let $Y \sim f_0$. Since $\int_{\mathcal{Y}} e^{\eta_0(y)} dy = \int_{\mathcal{Y}} f_0(y)\, dy = 1$, we have

$$\int_{\mathcal{Y}} e^{g(y)} dy = \int_{\mathcal{Y}} e^{\eta_0(y)} e^{\Delta(y)} dy = \mathbb{E}\Big[e^{\Delta(Y)}\Big].$$

Therefore, the KL divergence between f_0 and f_g can be written exactly as

$$\begin{aligned} D_{\mathrm{KL}}(f_0 \| f_g) &= \int f_0(y)\Big(\eta_0(y) - \eta_g(y)\Big) dy \\ &= \int f_0(y)\Big(\eta_0(y) - g(y)\Big) dy \;+\; \log \int e^{g(u)} du \\ &= -\mathbb{E}[\Delta(Y)] \;+\; \log \mathbb{E}\Big[e^{\Delta(Y)}\Big] \\ &= \log \mathbb{E}\Big[e^{\Delta(Y) - \mathbb{E}[\Delta(Y)]}\Big]. \end{aligned}$$

Now the centered random variable $W := \Delta(Y) - \mathbb{E}[\Delta(Y)]$ is supported on an interval of length at most 2ε, since $\Delta(Y) \in [-\varepsilon, \varepsilon]$ almost surely. By Hoeffding's lemma,

$$\log \mathbb{E}\big[e^W\big] \le \frac{(2\varepsilon)^2}{8} = \frac{\varepsilon^2}{2}.$$

Combining with the previous identity yields $D_{\mathrm{KL}}(f_0 \| f_g) \le \varepsilon^2/2$. □

We can now derive the sieve approximation rate.

Proposition B.1 (Sieve Approximation Error). *Under Assumptions B.4 and B.5, there exists a constant $C_{\mathrm{approx}} > 0$ (independent of J) such that*

$$\epsilon_{\mathrm{approx}}(J) = \sup_{x \in \mathcal{X}} D_{\mathrm{KL}}\big(f_0(\cdot \mid x) \,\big\|\, f(\cdot; \theta_0(x))\big) \;\le\; C_{\mathrm{approx}}\, J^{-2s}.$$

Proof. Fix $x \in \mathcal{X}$. Apply Assumption B.5 to $g = \eta_0(\cdot \mid x) \in \mathcal{W}^{s,\infty}(\mathcal{Y})$. There exists $g_J(\cdot \mid x) \in \mathcal{H}_J$ such that

$$\|\eta_0(\cdot \mid x) - g_J(\cdot \mid x)\|_\infty \le C_{\mathrm{basis}} J^{-s}\, \|\eta_0^{(s)}(\cdot \mid x)\|_\infty.$$

By Assumption B.4, $\|\eta_0^{(s)}(\cdot \mid x)\|_\infty \le C_0$ uniformly in x, hence

$$\|\eta_0(\cdot \mid x) - g_J(\cdot \mid x)\|_\infty \le \varepsilon_J \quad \text{with} \quad \varepsilon_J := C_{\mathrm{basis}} C_0\, J^{-s}.$$

Define the normalized sieve density $f_{J,x} := f_{g_J(\cdot|x)}$, i.e.,

$$f_{J,x}(y) = \frac{\exp(g_J(y \mid x))}{\int_{\mathcal{Y}} \exp(g_J(u \mid x))\, du}.$$

Since $g_J(\cdot \mid x) \in \mathcal{H}_J = \mathrm{span}\{B_1, \ldots, B_J\}$, we can write $g_J(y \mid x) = \theta_J(x)^\top \mathbf{B}(y)$ for some $\theta_J(x) \in \mathbb{R}^J$, and thus $f_{J,x}(\cdot) = f(\cdot; \theta_J(x)) \in \mathcal{P}_{\mathrm{sieve}}(J)$.

By Lemma B.2, we obtain

$$D_{\mathrm{KL}}\big(f_0(\cdot \mid x) \,\big\|\, f_{J,x}\big) \le \frac{\varepsilon_J^2}{2} = \frac{(C_{\mathrm{basis}} C_0)^2}{2}\, J^{-2s}.$$

Finally, since $\theta_0(x)$ is the minimizer of KL divergence over $\mathcal{P}_{\mathrm{sieve}}(J)$,

$$D_{\mathrm{KL}}\big(f_0(\cdot \mid x) \,\big\|\, f(\cdot; \theta_0(x))\big) \le D_{\mathrm{KL}}\big(f_0(\cdot \mid x) \,\big\|\, f_{J,x}\big) \le \frac{(C_{\mathrm{basis}} C_0)^2}{2}\, J^{-2s}.$$

Taking the supremum over $x \in \mathcal{X}$ yields the claim with $C_{\mathrm{approx}} := (C_{\mathrm{basis}} C_0)^2/2$. □

Remark B.2 (Rate Interpretation). *The bound $\epsilon_{\mathrm{approx}}(J) = O(J^{-2s})$ is a direct consequence of (i) uniform approximation of the log-density at rate $O(J^{-s})$ and (ii) the fact that KL divergence to the normalized exponential family behaves quadratically in the uniform log-error, as captured by Lemma B.2. For instance, when $s = 2$, the approximation bias decays as $O(J^{-4})$.*

B.2.3 Step 2: Connection to Decision Risk

We now bridge the gap between the density approximation error (measured by KL divergence) and the decision risk. This step justifies why the risk converges at the fast rate $O(J^{-2s})$.

Corollary B.1 (Approximation-Induced Excess Risk)**.** *Assume the cost function $c(z, y)$ corresponds to the newsvendor problem. Under the conditions of Proposition B.1, there exists a constant $C_R > 0$ such that:*

$$\mathcal{E}_{\text{approx}} = \mathbb{E}[\epsilon_{\text{approx}}(X)] \leq C_R \cdot J^{-2s}.$$

Proof. The proof relies on Pinsker's inequality and the strong convexity of the newsvendor risk.

1. From KL to Total Variation (TV). By Pinsker's inequality, for any x:

$$\|f_0(\cdot|x) - f(\cdot; \theta_0(x))\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{\text{KL}}(f_0 \| f_{\theta_0})} \leq \sqrt{\frac{C_{\text{approx}}}{2}} J^{-s}.$$

2. From TV to Decision Error. Let $F_0(\cdot|x)$ and $F_{\theta_0}(\cdot|x)$ be the CDFs of the true and projected densities. The optimal decisions are defined by the critical quantile α:

$$F_0(z^*(x)|x) = \alpha, \quad F_{\theta_0}(z_0(x)|x) = \alpha.$$

Since $f_0(y|x) \geq \delta > 0$ (Assumption B.4), the inverse CDF is Lipschitz. The difference in optimal decisions is bounded by the distance between CDFs, which is bounded by the TV distance:

$$|z^*(x) - z_0(x)| \leq \frac{1}{\delta} \sup_y |F_0(y|x) - F_{\theta_0}(y|x)| \leq \frac{1}{\delta} \|f_0 - f_{\theta_0}\|_{\text{TV}}.$$

Combining with step 1: $|z^*(x) - z_0(x)| = O(J^{-s})$.

3. From Decision Error to Excess Risk. The oracle risk function $z \mapsto R^*(z; x)$ is strongly convex around $z^*(x)$ because the density f_0 is bounded away from zero. By Taylor expansion (where the first derivative is zero at z^*):

$$R^*(z_0(x); x) - R^*(z^*(x); x) \approx \frac{f_0(z^*(x)|x)}{2} (z_0(x) - z^*(x))^2.$$

Substituting the decision error bound:

$$\mathcal{E}_{\text{approx}} \leq C \cdot (z_0 - z^*)^2 \leq C' \cdot (J^{-s})^2 = O(J^{-2s}).$$

This confirms that the approximation error in the risk space inherits the quadratic decay rate. □

B.3 Total Excess Risk and Optimal Basis Selection

We now combine the approximation and estimation error components to bound the total excess risk. Recall from Theorem 6.1 that, conditional on the nuisance estimation errors $a_n = \|\hat{\theta} - \theta_0\|_{L_2}$ and $r_n = \|\hat{m} - m_0\|_{L_2}$, the orthogonalized decision rule incurs estimation-induced suboptimality of order $O(a_n^2 + a_n r_n)$.

Theorem B.1 (Total Excess Risk)**.** *Suppose Assumptions B.4–B.3 hold. Then, with probability approaching one, the excess risk of the orthogonalized decision rule $\hat{z}$ satisfies*

$$\begin{aligned} \mathcal{E}(\hat{z}) &\leq 2\,\mathcal{E}_{\text{approx}} + C_1 a_n^2 + C_2 a_n r_n \\ &\leq O\left(J^{-2s} + J\left(n^{-\frac{2\beta_\theta}{2\beta_\theta + d}} + n^{-\left(\frac{\beta_\theta}{2\beta_\theta + d} + \frac{\beta_m}{2\beta_m + d}\right)} \right) \right), \end{aligned} \tag{162}$$

where the second inequality follows by substituting the rates from Lemma B.1. In the special case $\beta_\theta = \beta_m = \beta$, the estimation error term simplifies to $O(J n^{-2\beta/(2\beta+d)})$.

Corollary B.2 (Optimal Basis Size and Comparison)**.** *Assume $\beta_\theta = \beta_m = \beta$. We compare the orthogonalized estimator with the naive plug-in estimator by optimizing the basis size $J = J_n$.*

Plug-in estimator. *For the plug-in rule, excess risk is linear in the estimation error, yielding*

$$\mathcal{E}_{\text{plug}} \lesssim J^{-2s} + \sqrt{J}\, n^{-\frac{\beta}{2\beta + d}}.$$

Balancing the approximation and estimation terms gives

$$J_{\text{plug}} \asymp n^{\frac{2\beta}{(4s+1)(2\beta+d)}}.$$

Orthogonalized estimator. *For the orthogonalized rule, estimation error enters at second order, yielding*

$$\mathcal{E}_{\text{ortho}} \lesssim J^{-2s} + J\, n^{-\frac{2\beta}{2\beta+d}}.$$

Balancing the two terms gives

$$J_{\text{ortho}} \asymp n^{\frac{2\beta}{(2s+1)(2\beta+d)}}.$$

Since $2s+1 < 4s+1$ for all $s \geq 1$, the orthogonalized estimator permits a larger basis size, leading to a strictly faster decay of the total excess risk.

C Technical Lemmas

C.1 Interchange of Differentiation and Expectation

Lemma C.1 (Interchange of Differentiation and Expectation)**.** *Let $f : \mathcal{Y} \times \Theta \to \mathbb{R}$ be measurable. Suppose:*

(i) For each $y \in \mathcal{Y}$, $\theta \mapsto f(y,\theta)$ is differentiable.

(ii) There exists an integrable function $g : \mathcal{Y} \to \mathbb{R}$ such that $\|\nabla_\theta f(y,\theta)\| \leq g(y)$ for all $\theta \in \Theta$.

Then $F(\theta) := \mathbb{E}[f(Y,\theta)]$ is differentiable and:

$$\nabla_\theta F(\theta) = \mathbb{E}[\nabla_\theta f(Y,\theta)]. \tag{163}$$

Proof. This is a direct application of the dominated convergence theorem. For any $h \in \mathbb{R}^p$ with $\|h\| = 1$ and $t > 0$:

$$\frac{F(\theta+th) - F(\theta)}{t} = \mathbb{E}\left[\frac{f(Y,\theta+th) - f(Y,\theta)}{t}\right]. \tag{164}$$

By the mean value theorem, for each realization of Y there exists $\xi \in (0,1)$ such that

$$\frac{f(Y,\theta+th) - f(Y,\theta)}{t} = \langle \nabla_\theta f(Y,\theta+\xi th),\, h\rangle.$$

Hence,

$$\left|\frac{f(Y,\theta+th) - f(Y,\theta)}{t}\right| \leq \|\nabla_\theta f(Y,\theta+\xi th)\|\,\|h\| \leq g(Y),$$

where we used $\|h\| = 1$ and Assumption (ii).

So the integrand is bounded by $g(Y)$. As $t \to 0$, the integrand converges pointwise to $\langle \nabla_\theta f(Y,\theta), h\rangle$. By dominated convergence:

$$\lim_{t\to 0} \frac{F(\theta+th) - F(\theta)}{t} = \mathbb{E}[\langle \nabla_\theta f(Y,\theta), h\rangle] = \langle \mathbb{E}[\nabla_\theta f(Y,\theta)], h\rangle. \tag{165}$$

Since this holds for all directions h, we conclude $\nabla_\theta F(\theta) = \mathbb{E}[\nabla_\theta f(Y,\theta)]$. □

C.2 Operator Norm Bound for Inverse

Lemma C.2 (Operator Norm Bound for Inverse)**.** *Let $A, B \in \mathbb{R}^{p\times p}$ be invertible matrices with $\|A^{-1}\|_{op} \leq \kappa$ and $\|A - B\|_{op} \leq \epsilon$ with $\epsilon\kappa < 1$. Then B is invertible and:*

$$\|B^{-1} - A^{-1}\|_{op} \leq \frac{\kappa^2\epsilon}{1-\kappa\epsilon}. \tag{166}$$

Proof. Write $B = A(I - A^{-1}(A - B))$. Since $\|A^{-1}(A - B)\|_{\text{op}} \leq \kappa\epsilon < 1$, the matrix $I - A^{-1}(A - B)$ is invertible by the Neumann series:

$$(I - A^{-1}(A - B))^{-1} = \sum_{k=0}^{\infty} (A^{-1}(A - B))^k. \tag{167}$$

Thus $B^{-1} = (I - A^{-1}(A - B))^{-1}A^{-1}$, and:

$$B^{-1} - A^{-1} = \left[(I - A^{-1}(A - B))^{-1} - I\right] A^{-1} \tag{168}$$

$$= \sum_{k=1}^{\infty} (A^{-1}(A - B))^k \cdot A^{-1}. \tag{169}$$

Taking operator norms:

$$\|B^{-1} - A^{-1}\|_{\text{op}} \leq \sum_{k=1}^{\infty} (\kappa\epsilon)^k \cdot \kappa = \frac{\kappa^2\epsilon}{1 - \kappa\epsilon}. \tag{170}$$

□