ROBUST PRICING POLICY IMPROVEMENT WITH INVALID INSTRUMENTAL VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study offline policy improvement over a continuous price space when historical prices are endogenous and instrumental variables (IVs) are potentially invalid. Standard IV-based methods rely on instrument exogeneity. When this assumption is violated, the demand function cannot be identified, potentially leading to suboptimal pricing policies. To address this challenge, we relax the exact exogeneity assumption by constructing an ambiguity set of demand functions whose deviations from instrument exogeneity are bounded by a radius. We propose a robust policy that maximizes the worst-case expected revenue improvement over a baseline policy across this set. To overcome the computational challenge of the resulting max-min optimization, we derive a closed-form solution for the inner minimization problem. Furthermore, we develop a safe policy improvement procedure that maximizes the allowable violation radius of the ambiguity set subject to a prespecified worst-case revenue improvement. Theoretically, we establish convergence rates for the max-min policy regret and show that the proposed safe policy achieves the target revenue improvement up to a vanishing error. Simulation studies and synthetic dataset analysis in the context of airline-ticket pricing demonstrate the superior performance of our approach under instrument invalidity compared with existing methods.

1 INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Data-driven pricing has become central to modern retail decision-making. Leveraging massive observational datasets from historical transactions, sellers can employ machine learning to estimate complex demand functions and optimize pricing policies to maximize expected revenue. This decision problem is naturally formulated as an offline policy learning problem with a continuous action space. However, a fundamental challenge is unobserved confounding. In practice, sellers often set prices based on unrecorded market information, such as competitor promotions or sudden influencer endorsements, which affect customer demand. This unobserved confounding makes historical prices endogenous. Consequently, standard supervised learning fails to identify the true structural demand function, and optimizing policies based on biased demand estimates can lead to suboptimal decisions.

In causal inference, IV-based methods provide a principled framework for addressing unobserved confounding. A valid instrumental variable must be correlated with price and exogenous with respect to the unobserved noise. For the non-parametric regression task, Newey & Powell (2003) approximate the true regression function via a linear combination of basis functions. They first regress the basis functions on instrumental variables, and then regress the outcome on the fitted conditional expectations of these basis functions. Recently, Singh et al. (2019) extended this framework to reproducing kernel Hilbert spaces, while Xu et al. (2021) parameterized the regression function using deep neural networks. However, these methods rely heavily on the exact exogeneity assumption, which is frequently violated in retail pricing. For example, while coupon assignment is commonly used as an instrumental variable, a seller may issue coupons specifically to counter a competitor's discount. Because unobserved competitor discounts simultaneously affect customer demand and trigger coupon assignment, the instrument becomes correlated with the unobserved confounder.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.2 CONTRIBUTIONS

To address the challenge of invalid instruments (exogeneity assumption is violated), this paper proposes an offline continuous policy learning framework that explicitly accommodates instrument invalidity. Rather than relying on exact exogeneity, we allow for bounded violations of the conditional moment restrictions governing the IV exogeneity. Specifically, we construct an ambiguity set of demand functions whose conditional moment violations are bounded by a user-specified sensitivity radius. To construct a robust policy, we maximize the worst-case revenue improvement over a baseline policy across all demand functions within this ambiguity set. Furthermore, to trade off the policy value improvement and robustness, we propose a safe policy that maximizes the allowable violation radius of the ambiguity set subject to a targeted policy value improvement.

Our main contributions are summarized as follows:

(i) Robust Policy Improvement under Instrument Exogeneity Violations: We develop an offline pricing framework that accommodates bounded violations of instrument exogeneity. We construct an ambiguity set through conditional moment restrictions and learn a policy that maximizes worst-case revenue improvement over a baseline policy. This formulation supports policy improvement without requiring point identification of the demand function.

(ii) Closed-Form Solution and Targeted Robustness: Using a variational representation and double-sieve approximation, we express the ambiguity set through unconditional moments and derive a closed-form solution to the inner minimization over demand functions. For a prespecified improvement target, we characterize the largest admissible sensitivity radius as the unique solution to a one-dimensional equation, providing a practical procedure for selecting the sensitivity radius and the associated policy.

(iii) Regret Rates and Safe Policy Improvement: We derive convergence rates for the max–min policy regret, allowing the dimensions of sieve spaces to grow and the projected Gram matrix to become ill-conditioned. The variational sieve approximation error contributes only a second-order term. We further establish that the proposed safe policy achieves the prespecified policy value improvement up to a term that vanishes in probability.

1.3 RELATED WORK

Offline Policy Learning. Offline policy learning has been extensively studied in statistics, economics, and machine learning (Manski, 2004; Qian & Murphy, 2011; Zhao et al., 2012; Zhang et al., 2012; Athey & Wager, 2021; Wang et al., 2025). For continuous treatments, Chen et al. (2016) leveraged propensity scores alongside an interval-based indicator function to adjust observed outcomes, proposing an outcome-weighted learning method to maximize policy value and estimate the optimal policy. Kallus & Zhou (2018b) developed kernel-smoothed inverse-propensity-weighted and doubly robust methods for policy evaluation and optimization. Furthermore, Cai et al. (2021) addressed continuous treatments by discretizing the treatment space for off-policy evaluation. More recently, Leung et al. (2025) studied policy learning under distribution shifts and Ai et al. (2026) studied policy learning using kernel-based welfare estimation and sieve approximations to the policy class. However, these methods rely on the unconfoundedness assumption. This assumption may be violated when historical prices are endogenous.

Identification under Unmeasured Confounding. Instrumental variable methods have long been used to identify structural regression functions under unmeasured confounding (Newey & Powell, 2003; Singh et al., 2019; Xu et al., 2021; Fonseca et al., 2024; Kim et al., 2025). Cui & Tchetgen Tchetgen (2021) studied optimal treatment regimes using a valid instrument, while Shao et al. (2024) developed a doubly robust IV method for policy learning. These methods assume instrument validity. Miao et al. (2023) studied personalized pricing with instruments that may directly affect revenue, relaxing the exclusion restriction while retaining conditional independence between instruments and unobserved confounders. However, these studies do not focus on pricing policy learning under violations of the instrument exogeneity assumption.

Robust Policy Learning. Relaxing exact identifying restrictions via sensitivity analysis has received significant attention for treatment effect estimation (Armstrong & Kolesár, 2021; Masten &

Poirier, 2020). In the binary treatment setting, worst-case policy improvement over a baseline has been studied under unobserved confounding (Kallus & Zhou, 2018a). Hess et al. (2026) derived a semiparametrically efficient estimator for a sharp bound on policy value. These methods are formulated for discrete treatments and do not directly apply to continuous pricing. Joshi et al. (2024) combined confounded observational data with randomized trials that observe different subsets of covariates to derive sharp bounds on conditional treatment effects and maximize a resulting lower bound on policy value. These studies do not address pricing policy learning under violations of instrument exogeneity.

2 PRELIMINARIES

2.1 FEATURE-BASED PRICING UNDER ENDOGENEITY

Consider observational data $(\boldsymbol{X}, P, Y)$, where $\boldsymbol{X}$ contains customer and product features, $P \in [\underline{p}, \bar{p}]$ is the historical price, and Y is demand. Let $Y(p)$ denote potential demand under a price intervention p. We assume consistency: $Y(p) = Y$ whenever $P = p$. Define the conditional mean demand function as

$$f_0(\boldsymbol{x}, p) = \mathbb{E}\{Y(p) \mid \boldsymbol{X} = \boldsymbol{x}\}.$$

We assume that $f_0 \in \mathcal{S}$, where $\mathcal{S}$ is a Hölder class. Define the residual by $\epsilon = Y - f_0(\boldsymbol{X}, P)$. Therefore, we have $Y = f_0(\boldsymbol{X}, P) + \epsilon$.

A pricing policy $\pi \in \Pi$ maps $\boldsymbol{X}$ to $[\underline{p}, \bar{p}]$. The policy value is the expected revenue

$$V_{f_0}(\pi) = \mathbb{E}[\pi(\boldsymbol{X}) Y\{\pi(\boldsymbol{X})\}] = \mathbb{E}[\pi(\boldsymbol{X}) f_0\{\boldsymbol{X}, \pi(\boldsymbol{X})\}]. \tag{1}$$

Our objective is to maximize $V_{f_0}(\pi)$ over $\pi \in \Pi$.

In practice, sellers may set prices using market information, such as competitor discounts, that is omitted from the feature vector $\boldsymbol{X}$. Because this information also affects demand, it induces $\mathbb{E}(\epsilon \mid \boldsymbol{X}, P) \neq 0$. This endogeneity poses a significant challenge for identifying the true demand function f_0. Specifically, for any function f, the mean squared error can be decomposed as follows:

$$\mathbb{E}\left[\{Y - f(\boldsymbol{X}, P)\}^2\right] = \mathbb{E}\left[\mathrm{Var}(\epsilon \mid \boldsymbol{X}, P)\right] + \mathbb{E}\left[\{f_0(\boldsymbol{X}, P) + \mathbb{E}(\epsilon \mid \boldsymbol{X}, P) - f(\boldsymbol{X}, P)\}^2\right].$$

Its population minimizer is given by $f_{\text{ols}}(\boldsymbol{x}, p) = f_0(\boldsymbol{x}, p) + \mathbb{E}(\epsilon \mid \boldsymbol{X} = \boldsymbol{x}, P = p)$. Therefore, minimizing the mean squared error fails to identify f_0 due to the price endogeneity.

2.2 INVALID INSTRUMENTAL VARIABLE

To address the challenge of endogeneity, a large body of literature assumes that there exists a valid instrumental variable Z satisfying the following assumption:

Assumption 1. (a) (*Relevance*) $Z \not\perp P \mid \boldsymbol{X}$. (b) (*Instrument Exogeneity or Conditional IV Moment Validity*) $\mathbb{E}[\epsilon \mid \boldsymbol{X}, Z] = 0$.

Assumption 1(b) is not directly testable using observational data because the residual ϵ is unobserved. In practice, this condition is frequently violated.

Example 1 (Motivating example)**.** *Define $Z = 1$ if a customer receives a coupon, and $Z = 0$ otherwise. Consider the following model:*

$$P = \mu(\boldsymbol{X}) - Z - U, \quad Y = f_0(\boldsymbol{X}, P) + \theta U + \xi,$$

where ξ is a mean-zero random variable independent of $(\boldsymbol{X}, U, Z)$. The noise is $\epsilon = \theta U + \xi$, where $\theta \neq 0$ and U is an unobserved confounder, such as the magnitude of a competitor's discount. Since the variable U influences both the price and customer demand simultaneously, the observed price is endogenous. Moreover, in practice, the coupon assignment Z is dependent on the market environment. For instance, the seller may issue coupons to counter a competitor's price cut. Assume that $E(U \mid \boldsymbol{X}) = 0$ and $E(U \mid \boldsymbol{X}, Z) \neq 0$. Thus, in this setting, Z and U become dependent given $\boldsymbol{X}$, which implies

$$\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z) = \theta E(U \mid \boldsymbol{X}, Z) \neq 0.$$

Consequently, the instrumental variable Z violates Assumption 1 (b).

3 METHODOLOGY

3.1 INFORMATIVE AMBIGUITY SET

In practice, the instrument exogeneity condition in Assumption 1(b) may fail. We therefore relax this condition by allowing $\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)$ to deviate from zero.

Accordingly, we construct the following ambiguity set

$$\mathcal{F}(\rho; \mathcal{S}) = \left\{ f : \mathbb{E}\left[\{\mathbb{E}[Y - f(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2 \right] \leq \rho^2, f \in \mathcal{S} \right\}. \tag{2}$$

This formulation relaxes the instrument exogeneity condition in Assumption 1(b). Instead of requiring $\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2 \right] = 0$, we permit a bounded violation up to a threshold ρ^2. This relaxation guarantees that the ambiguity set $\mathcal{F}(\rho; \mathcal{S})$ contains the true demand function f_0 whenever $\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2 \right] \leq \rho^2$. Specifically, in Example 1, this condition translates to $|\theta| \left[\mathbb{E}\left[\{E(U \mid \boldsymbol{X}, Z)\}^2 \right] \right]^{1/2} \leq \rho$. The parameter $\rho \geq 0$ quantifies the allowed magnitude of violation of Assumption 1 (b). A larger value of ρ implies a higher tolerance for the violation of this condition. When $\rho = 0$, the ambiguity set reduces to the class of demand functions that exactly satisfy the instrument exogeneity condition in Assumption 1(b).

To illustrate why the ambiguity set $\mathcal{F}(\rho; \mathcal{S})$ can be more informative than ordinary least squares under endogenous pricing, we consider a finite-dimensional sieve approximation. Define

$$\mathcal{S}_K = \left\{ f_{\boldsymbol{\beta}}(\boldsymbol{x}, p) = \boldsymbol{\phi}_K(\boldsymbol{x}, p)^\top \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^K \right\},$$

where $\boldsymbol{\phi}_K(\boldsymbol{x}, p) = (\phi_1(\boldsymbol{x}, p), \ldots, \phi_K(\boldsymbol{x}, p))^\top$ is a vector of prespecified square-integrable basis functions. Define

$$Q_K = \mathbb{E}\left[\boldsymbol{\phi}_K(\boldsymbol{X}, P) \boldsymbol{\phi}_K(\boldsymbol{X}, P)^\top \right] \text{ and } G_K = \mathbb{E}\left[\mathbb{E}\{\boldsymbol{\phi}_K(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z\} \mathbb{E}\{\boldsymbol{\phi}_K(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z\}^\top \right].$$

The following proposition compares the ambiguity set with the population least-squares regression function. It provides a sufficient condition under which every demand function in the ambiguity set is closer to the true structural demand function than f_{ols}.

Proposition 1. *(i) Suppose that $(\boldsymbol{X}, P)$ is supported on a compact rectangular domain $\mathcal{D}_f \subset \mathbb{R}^{d_f}$ and $f_0 \in \mathcal{S} = \mathcal{H}^{s_f}(\mathcal{D}_f, B_f)$, where $\mathcal{H}^{s_f}(\mathcal{D}_f, B_f)$ denotes the Hölder class with smoothness parameter s_f and radius B_f on the domain $\mathcal{D}_f$ (see Appendix B.5 for details). (ii) Assume that the sieve approximation error satisfies $\inf_{\widetilde{f} \in \mathcal{S}_K} \|\widetilde{f} - f_0\|_\infty \leq \alpha_K$, where $\alpha_K \to 0$ as $K \to \infty$. (iii) Assume that Q_K and G_K are positive definite, and let $\lambda_{\min}$ denote the minimum eigenvalue of $Q_K^{-1/2} G_K Q_K^{-1/2}$. For any ρ satisfying $\sqrt{\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2 \right]} + \alpha_K < \rho < \frac{\sqrt{\lambda_{\min}}}{2} \left\{ \sqrt{\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, P)\}^2 \right]} - \alpha_K \right\}$, we have, for every $f \in \mathcal{F}(\rho; \mathcal{S}_K)$*

$$\mathbb{E}\left[\{f(\boldsymbol{X}, P) - f_0(\boldsymbol{X}, P)\}^2 \right] < \mathbb{E}\left[\{f_{\text{ols}}(\boldsymbol{X}, P) - f_0(\boldsymbol{X}, P)\}^2 \right].$$

The proposition demonstrates that every demand function in the ambiguity set has smaller mean squared error than f_{ols}, provided the instrument exogeneity violation $\sqrt{\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2 \right]}$ and the sieve approximation error α_K are sufficiently small compared with the endogeneity bias $\sqrt{\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, P)\}^2 \right]}$. The sieve approximation error α_K is negligible. For polynomial, spline, orthogonal wavelet and tensor-product bases, the sieve approximation results in Newey (1997); Chen (2007); Belloni et al. (2015) yield $\alpha_K = O(K^{-s_f/d_f})$. Thus, an approximately valid instrument can remain informative even when exact instrument exogeneity is not credible.

3.2 ROBUST POLICY IMPROVEMENT

Since the true demand function f_0 cannot be point-identified under an invalid instrumental variable, we account for the worst-case scenario over all plausible demand functions in $\mathcal{F}(\rho; \mathcal{S}_K)$. Under this

framework, we aim for a robust policy improvement over a deployable baseline policy π_0. Formally, the optimal robust policy $\pi^\star(\rho)$ is defined as the solution to the following max-min problem:

$$\pi^\star(\rho) \in \arg\max_{\pi\in\Pi} \min_{f\in\mathcal{F}(\rho;\mathcal{S}_K)} \{V_f(\pi) - V_f(\pi_0)\}. \tag{3}$$

where ρ can be specified via sensitivity analysis or domain knowledge. Solving the max-min problem in Equation (3) directly over the function class $\mathcal{S}_K$ can be computationally challenging. The primary difficulty arises from the constraint defined in $\mathcal{F}(\rho;\mathcal{S}_K)$:

$$\mathbb{E}\left[\{\mathbb{E}[Y - f_{\boldsymbol{\beta}}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right] \le \rho^2. \tag{4}$$

Directly enforcing the constraint in Equation (4) would require computing a conditional regression for $f_{\boldsymbol{\beta}}(\boldsymbol{X}, P)$. The following variational identity bypasses this computational bottleneck.

Proposition 2. *Suppose that $f_{\boldsymbol{\beta}} \in \mathcal{S}_K$ is a measurable function such that $\mathbb{E}\left[\{Y - f_{\boldsymbol{\beta}}(\boldsymbol{X}, P)\}^2\right] < \infty$. Then*

$$\begin{aligned}&\mathbb{E}\left[\{\mathbb{E}[Y - f_{\boldsymbol{\beta}}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right]\\ =&\max_{g\in L^2(P_{\boldsymbol{X},Z})} \left\{2\mathbb{E}[g(\boldsymbol{X}, Z)\{Y - f_{\boldsymbol{\beta}}(\boldsymbol{X}, P)\}] - \mathbb{E}\left[g(\boldsymbol{X}, Z)^2\right]\right\},\end{aligned} \tag{5}$$

where $L^2(P_{\boldsymbol{X},Z})$ denotes the space of square-integrable functions with respect to distribution $P_{\boldsymbol{X},Z}$. Moreover, the maximum is attained at

$$g^\star_{f_{\boldsymbol{\beta}}}(\boldsymbol{x}, z) = \mathbb{E}[Y - f_{\boldsymbol{\beta}}(\boldsymbol{X}, P) \mid \boldsymbol{X} = \boldsymbol{x}, Z = z].$$

Proposition 2 rewrites the criterion in Equation (4) using unconditional expectations. To obtain an estimable sieve approximation, we replace $L^2(P_{\boldsymbol{X},Z})$ with

$$\mathcal{T}_L = \left\{g_{\boldsymbol{\gamma}}(\boldsymbol{x}, z) = \boldsymbol{\psi}_L(\boldsymbol{x}, z)^\top\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathbb{R}^L\right\}, \tag{6}$$

where $\boldsymbol{\psi}_L = (\psi_1, \ldots, \psi_L)^\top$ is a vector of prespecified basis functions of $(\boldsymbol{X}, Z)$, such as polynomials, splines, wavelets, or tensor products.

Restricting the variational function to $\mathcal{T}_L$ leads to the following double-sieve ambiguity set:

$$\mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L) = \left\{f_{\boldsymbol{\beta}} \in \mathcal{S}_K : \max_{g\in\mathcal{T}_L}\left[2\mathbb{E}[g(\boldsymbol{X}, Z)\{Y - f_{\boldsymbol{\beta}}(\boldsymbol{X}, P)\}] - \mathbb{E}\{g(\boldsymbol{X}, Z)^2\}\right] \le \rho^2\right\}. \tag{7}$$

Remark 1. In $\mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L)$, the dimension K determines the finite-dimensional approximation to the demand-function space $\mathcal{S}$, whereas L determines the finite-dimensional approximation to the variational function space.

To derive a closed-form solution to the inner minimization, we define the following population quantities:

$$\begin{aligned}G_{K,L} =&\mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top], \quad \bar{\boldsymbol{\beta}}_{K,L} = G_{K,L}^{-1}\mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}[\boldsymbol{\psi}_L Y],\\ q_{K,L} =&\mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\bar{\boldsymbol{\beta}}_{K,L}\}\right]^\top\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\bar{\boldsymbol{\beta}}_{K,L}\}\right],\end{aligned}$$

where the arguments of the basis vectors $\boldsymbol{\psi}_L$ and $\boldsymbol{\phi}_K$ are omitted for brevity.

This yields the following closed-form solution.

Theorem 1 (Closed-form worst-case policy improvement)**.** *Suppose that $\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]$ is positive definite and $\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]$ has full column rank. If $\rho^2 \ge q_{K,L}$, then*

$$\begin{aligned}&\min_{f_{\boldsymbol{\beta}}\in\mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L)} \left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}\\ =&V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi_0) - \sqrt{\rho^2 - q_{K,L}}\left\|G_{K,L}^{-1/2}\nabla_{\boldsymbol{\beta}}\left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}\right\|_2,\end{aligned} \tag{8}$$

where $\nabla_{\boldsymbol{\beta}} V_{f_{\boldsymbol{\beta}}}(\pi) = \mathbb{E}[\pi(\boldsymbol{X})\boldsymbol{\phi}_K\{\boldsymbol{X}, \pi(\boldsymbol{X})\}]$ *and* $\nabla_{\boldsymbol{\beta}} V_{f_{\boldsymbol{\beta}}}(\pi_0) = \mathbb{E}[\pi_0(\boldsymbol{X})\boldsymbol{\phi}_K\{\boldsymbol{X}, \pi_0(\boldsymbol{X})\}]$. *Furthermore, if* $\nabla_{\boldsymbol{\beta}}\left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\} \neq 0$, *the minimum in (8) is attained at*

$$\boldsymbol{\beta}_\pi^{\mathrm{wc}} = \bar{\boldsymbol{\beta}}_{K,L} - \sqrt{\rho^2 - q_{K,L}}\, \frac{G_{K,L}^{-1}\nabla_{\boldsymbol{\beta}}\left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}}{\left\|G_{K,L}^{-1/2}\nabla_{\boldsymbol{\beta}}\left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}\right\|_2}; \tag{9}$$

if $\nabla_{\boldsymbol{\beta}}\left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\} = 0$, *the minimum in (8) is attained at* $\boldsymbol{\beta}_\pi^{\mathrm{wc}} = \bar{\boldsymbol{\beta}}_{K,L}$.

The condition $\rho^2 \geq q_{K,L}$ ensures that $\mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)$ is nonempty. Theorem 1 converts the max–min problem into a single maximization over the policy class. Specifically, the optimal robust policy can be obtained by solving

$$\max_{\pi\in\Pi} V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi_0) - \sqrt{\rho^2 - q_{K,L}}\left\|G_{K,L}^{-1/2}\nabla_{\boldsymbol{\beta}}\left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}\right\|_2. \tag{10}$$

Figure 1 provides a geometric interpretation of Equation (8). The difference $V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi_0)$ is the policy improvement evaluated at the center $\bar{\boldsymbol{\beta}}_{K,L}$. The factor $\sqrt{\rho^2 - q_{K,L}}$ determines the radius of $\mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)$. Geometrically, the lowest attainable contour of the policy improvement is tangent to the ellipsoid at $\boldsymbol{\beta}_\pi^{\mathrm{wc}}$.

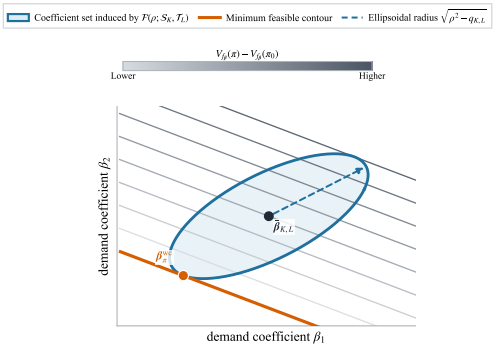


Figure 1: A geometric illustration of Theorem 1. The blue ellipse represents the ambiguity set of $\mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)$. The gray lines are contours of the policy improvement $V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)$.

3.3 Trade-off between policy improvement and robustness

The robust policy problem in Equation (3) induces a direct trade-off. A larger sensitivity radius allows for a greater violation of instrument exogeneity, but enlarges the ambiguity set and consequently weakens the guaranteed policy improvement. For every radius ρ, we define the maximum worst-case improvement in policy value as

$$\max_{\pi\in\Pi} \min_{f\in\mathcal{F}(\rho;\mathcal{S}_K)} \{V_f(\pi) - V_f(\pi_0)\}.$$

Thus, a decision maker can prespecify a minimum economically worthwhile target improvement $\delta > 0$. We define the maximum allowable sensitivity radius $\rho_{\mathrm{gain,K}}(\delta)$ subject to this target δ by solving

$$\max_{\rho\geq 0} \rho \quad \text{subject to} \quad \max_{\pi\in\Pi} \min_{f\in\mathcal{F}(\rho;\mathcal{S}_K)} \{V_f(\pi) - V_f(\pi_0)\} \geq \delta.$$

The radius $\rho_{\mathrm{gain,K}}(\delta)$ measures the largest IV violation that the target improvement can tolerate. Correspondingly, the safe policy is denoted as $\pi^\star\left(\rho_{\mathrm{gain,K}}(\delta)\right)$. This policy guarantees a worst-case policy value improvement of at least δ while allowing the largest possible degree of instrument invalidity. To obtain an estimable approximation, we replace $\mathcal{F}(\rho; \mathcal{S}_K)$ with its double-sieve counterpart $\mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)$ and define $M_{K,L}(\rho) = \max_{\pi\in\Pi} \min_{f_{\boldsymbol{\beta}}\in\mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L)} \left\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}$. Let $\pi^\star_{K,L}(\rho)$ denote the policy attaining this maximum. Then, we solve the following optimization problem:

$$\rho_{\mathrm{gain},K,L}(\delta) = \max_{\rho\geq\sqrt{q_{K,L}}} \rho \quad \text{subject to} \quad M_{K,L}(\rho) \geq \delta. \tag{11}$$

Theorem 2 (Largest radius supporting a prespecified improvement)**.** *Suppose that* $\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]$ *is positive definite,* $\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]$ *has full column rank,* Π *is compact,* $\pi_0 \in \Pi$, *and the two terms in Equation (8) are continuous in* π. *Let* $\delta > 0$ *and suppose that* $\max_{\pi\in\Pi}\left\{V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi_0)\right\} > \delta$. *Then there exists a unique* $\rho_{\mathrm{gain},K,L}(\delta) > \sqrt{q_{K,L}}$ *satisfying*

$$M_{K,L}\{\rho_{\mathrm{gain},K,L}(\delta)\} = \delta. \tag{12}$$

Theorem 2 reduces the optimization problem in (11) to a one-dimensional root-finding problem. We summarize the complete estimation procedure in Algorithm 1.

Algorithm 1 Robust policy estimation

Require: An i.i.d. sample of observations $\{(Y_i, P_i, \boldsymbol{X}_i, Z_i)\}_{i=1}^n$, $\boldsymbol{\phi}_K$, $\boldsymbol{\psi}_L$, policy class Π, fixed radius ρ, and target $\delta > 0$

1: Set $\boldsymbol{\phi}_{K,i} = \boldsymbol{\phi}_K(\boldsymbol{X}_i, P_i)$ and $\boldsymbol{\psi}_{L,i} = \boldsymbol{\psi}_L(\boldsymbol{X}_i, Z_i)$.

2: Compute $\widehat{G}_{K,L}$, $\widehat{\widetilde{\boldsymbol{\beta}}}_{K,L}$, and $\widehat{q}_{K,L}$ by replacing population expectations with empirical averages.

3: For $f_{\boldsymbol{\beta}} \in \mathcal{S}_K$, set $\widehat{V}_{f_{\boldsymbol{\beta}}}(\pi) = E_n[\pi(\boldsymbol{X}_i) f_{\boldsymbol{\beta}}\{\boldsymbol{X}_i, \pi(\boldsymbol{X}_i)\}]$.

4: If $\rho^2 \geq \widehat{q}_{K,L}$, compute

$$\widehat{M}_{K,L}(\rho) = \max_{\pi \in \Pi} \left\{ \widehat{V}_{f_{\widehat{\boldsymbol{\beta}}_{K,L}}}(\pi) - \widehat{V}_{f_{\widehat{\boldsymbol{\beta}}_{K,L}}}(\pi_0) - \sqrt{\rho^2 - \widehat{q}_{K,L}} \left\| \widehat{G}_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{\widehat{V}_{f_{\boldsymbol{\beta}}}(\pi) - \widehat{V}_{f_{\boldsymbol{\beta}}}(\pi_0)\} \right\|_2 \right\}. \tag{13}$$

Solve the optimization in Equation (13) and denote its solution by $\widehat{\pi}^\star_{K,L}(\rho)$. Otherwise, set $\widehat{\pi}^\star_{K,L}(\rho) = \pi_0$ and report that the empirical ambiguity set at radius ρ is empty.

5: If $\widehat{M}_{K,L}(\sqrt{\widehat{q}_{K,L}}) \geq \delta$, solve $\widehat{\rho}_{\mathrm{gain},K,L}(\delta)$ such that $\widehat{M}_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\} = \delta$ over $\left[\sqrt{\widehat{q}_{K,L}}, \sqrt{E_n[Y^2]}\right]$ by the bisection method, and solve Equation (13) at this radius to obtain $\widehat{\pi}^\star_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}$. Otherwise, retain π_0 and report that the target is unsupported.

6: **return** (i) $\widehat{\pi}^\star_{K,L}(\rho)$ if $\rho^2 \geq \widehat{q}_{K,L}$; (ii) $\widehat{\pi}^\star_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}$ and $\widehat{\rho}_{\mathrm{gain},K,L}(\delta)$ if $\widehat{M}_{K,L}(\sqrt{\widehat{q}_{K,L}}) \geq \delta$.

4 Theoretical Results

In this section, we establish the convergence rate of the worst-case regret $\mathrm{Reg}_{\mathrm{mm}}\{\widehat{\pi}^\star_{K,L}(\rho); \rho\}$, which is given by

$$\mathrm{Reg}_{\mathrm{mm}}(\widehat{\pi}^\star_{K,L}(\rho); \rho) = \min_{f \in \mathcal{F}(\rho; \mathcal{S}_K)} \{V_f(\pi^\star(\rho)) - V_f(\pi_0)\} - \min_{f \in \mathcal{F}(\rho; \mathcal{S}_K)} \{V_f(\widehat{\pi}^\star_{K,L}(\rho)) - V_f(\pi_0)\}.$$

To establish the theoretical results, we need the following assumption.

Assumption 2. (i) Assume that $(\boldsymbol{X}, Z)$ is supported on a compact rectangular domain $\mathcal{D}_g \subset \mathbb{R}^{d_g}$. Suppose that $g^\star_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}} \in \mathcal{H}^{s_g}(\mathcal{D}_g, B_g)$ for a fixed smoothness parameter $s_g > 0$ and radius $B_g < \infty$. Assume that the approximation error satisfies $\inf_{g \in \mathcal{T}_L} \|g^\star_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}} - g\|_\infty \leq b_L \to 0$ uniformly over $\pi \in \Pi$, wherein $\boldsymbol{\beta}^{\mathrm{wc}}_\pi$ is evaluated at ρ for Theorem 3, and at $\rho_{\mathrm{gain,K}}(\delta)$ for Theorem 4.

(ii) Let $Q_K = \mathbb{E}[\boldsymbol{\phi}_K \boldsymbol{\phi}_K^\top]$ and $Q_L = \mathbb{E}[\boldsymbol{\psi}_L \boldsymbol{\psi}_L^\top]$. Assume that $cI_K \leq Q_K \leq CI_K$ and $cI_L \leq Q_L \leq CI_L$, where c and C ($0 < c < C < \infty$) are independent of n, K, L. Suppose that $G_{K,L}$ is positive definite with $\lambda_{\min}(G_{K,L}) = \tau_{K,L}^{-2} > 0$, where $\tau_{K,L}$ may diverge as $K, L \to \infty$.

(iii) Define the envelope as $\zeta_{K,L} = \left\{ \sup_{(\boldsymbol{x},z) \in \mathcal{D}_g} \|\boldsymbol{\psi}_L(\boldsymbol{x}, z)\|_2^2 + \sup_{(\boldsymbol{x},p) \in \mathcal{D}_f} \|\boldsymbol{\phi}_K(\boldsymbol{x}, p)\|_2^2 \right\}^{1/2}$. Assume that $\tau_{K,L}^2 \zeta_{K,L}^2 \sqrt{\frac{\log n}{n}} \longrightarrow 0$, $\tau_{K,L} \zeta_{K,L} b_L^2 \longrightarrow 0$ and $\mathbb{E}[Y^2] \leq C$.

In practice, we can construct the sieve space $\mathcal{T}_L$ using polynomials, splines, orthogonal wavelets, Fourier series, or tensor-product basis functions (Newey, 1997; Chen, 2007; Belloni et al., 2015). Under suitable smoothness and boundary conditions (Chen, 2007; Chen & Christensen, 2015), we have $b_L = O(L^{-s_g/d_g})$ for these basis functions. The uniform eigenvalue bounds on Q_K and Q_L in Assumption 2(ii) are standard regularity conditions satisfied by normalized B-splines, orthonormal wavelets, and Fourier bases (Chen & Christensen, 2015). We allow $\lambda_{\min}(G_{K,L})$ to approach zero to accommodate the ill-posedness caused by limited information from the instruments (Chen & Christensen, 2018). Assumption 2(iii) restricts the complexity of these sieve spaces. For the B-splines, orthogonal wavelets and Fourier series basis functions, the envelope satisfies

$\zeta_{K,L} = O(\sqrt{K+L})$. In well-posed settings (see Blundell et al. (2007) for details), we have $\tau_{K,L} = O(1)$ and $\tau_{K,L}^2\zeta_{K,L}^2\sqrt{\frac{\log n}{n}} = O((K+L)\sqrt{\frac{\log n}{n}})$.

Theorem 3. *Suppose that Assumptions 2 and 3 (see Appendix B.6) hold. Let* $\mathcal{E}(n,K,L) = \tau_{K,L}^2\zeta_{K,L}^2\sqrt{\frac{\log n}{n}} + \tau_{K,L}\zeta_{K,L}b_L^2$. *We have* $\mathrm{Reg}_{\mathrm{mm}}\{\widehat{\pi}^\star_{K,L}(\rho);\rho\} = O_p(\mathcal{E}(n,K,L))$.

Theorem 3 separates the statistical error $\tau_{K,L}^2\zeta_{K,L}^2\sqrt{\log n/n}$ from the approximation error $\tau_{K,L}\zeta_{K,L}b_L^2$. The approximation error is quadratic in b_L. This allows for a more flexible choice of the dimension L. Specifically, in the well-posed setting with $\tau_{K,L} = O(1)$, suppose that $K \asymp L$, $\zeta_{K,L} = O(\sqrt{K+L})$, and $b_L = O(L^{-s_g/d_g})$. Then $\mathrm{Reg}_{\mathrm{mm}}\{\widehat{\pi}^\star_{K,L}(\rho);\rho\} = O_p\Big(L\sqrt{\log n/n} + L^{1/2-2s_g/d_g}\Big)$. If $s_g > d_g/4$, choosing $L \asymp \big(n^{d_g/(4s_g+d_g)}(\log n)^{-d_g/(4s_g+d_g)}\big)$ yields $\mathrm{Reg}_{\mathrm{mm}}\{\widehat{\pi}^\star_{K,L}(\rho);\rho\} = O_p\left[\left(\frac{\log n}{n}\right)^{\frac{4s_g-d_g}{8s_g+2d_g}}\right]$.

Theorem 4. *Suppose that the conditions (i) and (ii) of Proposition 1 and the assumptions of Theorem 3 hold. Assume that* $\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2\right] \le \rho^2_{\mathrm{gain,K}}(\delta)$ *and* $\tau_{K,L}\zeta_{K,L}\alpha_K \longrightarrow 0$, *then*

$$V_{f_0}\big[\widehat{\pi}^\star_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}\big] - V_{f_0}(\pi_0) \ge \delta - O_p\left(\tau_{K,L}^2\zeta_{K,L}^2\sqrt{\frac{\log n}{n}} + \tau_{K,L}\zeta_{K,L}(b_L^2+\alpha_K)\right) \quad (14)$$

Theorem 4 demonstrates that the proposed safe policy $\widehat{\pi}^\star_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}$ achieves a true policy value improvement of δ, up to a vanishing term. Specifically, in the well-posed setting with $\tau_{K,L} = O(1)$, $\zeta_{K,L} = O(\sqrt{K+L})$, $b_L = O(L^{-s_g/d_g})$, and $\alpha_K = O(K^{-s_f/d_f})$, the remainder in Equation (14) satisfies $\tau_{K,L}^2\zeta_{K,L}^2\sqrt{\frac{\log n}{n}} + \tau_{K,L}\zeta_{K,L}(b_L^2+\alpha_K) = O\left[(K+L)\sqrt{\frac{\log n}{n}} + \sqrt{K+L}\left(L^{-2s_g/d_g} + K^{-s_f/d_f}\right)\right]$. If $s_g > d_g/4$ and $s_f > d_f/2$, setting $K \asymp L \asymp \left(\frac{n}{\log n}\right)^{\max\left\{\frac{d_g}{4s_g+d_g}, \frac{d_f}{2s_f+d_f}\right\}}$, we have

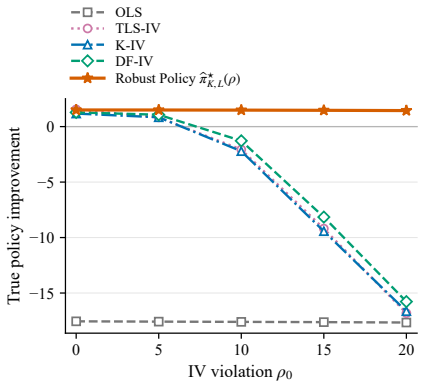
$$V_{f_0}\big[\widehat{\pi}^\star_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}\big] - V_{f_0}(\pi_0) \ge \delta - O_p\left(\left(\frac{\log n}{n}\right)^{\min\left\{\frac{4s_g-d_g}{8s_g+2d_g}, \frac{2s_f-d_f}{4s_f+2d_f}\right\}}\right).$$

5 Simulation Studies

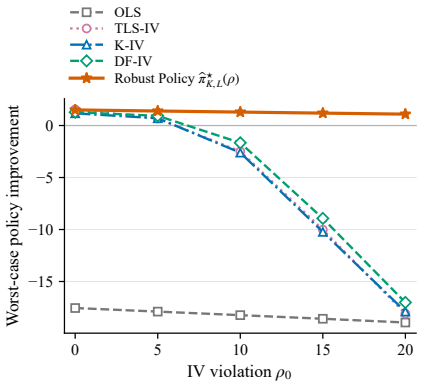
5.1 Data-generating process

Let covariates $\boldsymbol{X} \sim \mathrm{Unif}[-1,1]^5$. We generate the instrument Z and price P as: $Z = \frac{\rho_0}{100}U + \sqrt{1-\left(\frac{\rho_0}{100}\right)^2}V$ and $P = 0.40 + 1.40\Phi\left(\frac{Z+0.50U}{\sqrt{1.25+\rho_0/100}}\right)$, where $U, V, \xi \sim \mathcal{N}(0,1)$ and Φ is the standard normal cumulative distribution function. Observed prices are supported on $[0.40, 1.80]$. Demand is generated by $Y = 120(2.20 + 0.10\sum_{j=1}^5 X_j - P) + 100U + 0.50\xi$. The unobserved confounder U correlates with Z, invalidating the instrument. Specifically, the violation magnitude is $[\mathbb{E}\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)^2\}]^{1/2} = \rho_0$. We set the baseline policy to the constant pricing rule $\pi_0(\boldsymbol{x}) \equiv 1$ for all $\boldsymbol{x} \in [-1,1]^5$.

We compare the proposed robust policy against policies that maximize empirical revenue using demand functions estimated via ordinary least squares (OLS), two-stage least squares (Newey & Powell, 2003) (TLS-IV), kernel instrumental variable regression (Singh et al., 2019) (K-IV), and deep feature instrumental variable regression (Xu et al., 2021) (DF-IV). We use n observations for both demand and policy estimation. For each method, we report the true improvement of policy value $V_{f_0}(\pi) - V_{f_0}(\pi_0)$ under the demand f_0 and the worst-case improvement of policy value $\min_{f\in\mathcal{F}(\rho;\mathcal{S}_K)}\{V_f(\pi) - V_f(\pi_0)\}$. The implementation details are shown in Appendix C.



(a) True policy value improvement.

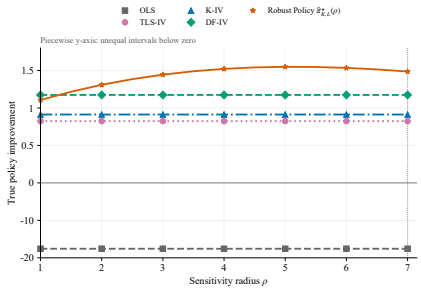


(b) Worst-case policy value improvement (evaluated at $\rho = \rho_0$).

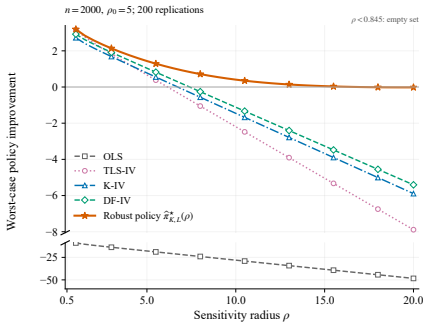
Figure 2: Mean true and worst-case policy value improvements over 200 replications with $n = 2000$. Here, we consider a setting in which domain experts can use prior knowledge to provide a conservative upper bound on instrument exogeneity violations. Thus, we set $\rho = \max\{\rho_0 + 2, \sqrt{\widehat{q}_{K,L}}\}$ to construct a conservative sensitivity radius.

5.2 Results

Figure 2(a) shows that the proposed robust policy maintains positive true revenue improvements as the instrument exogeneity violation level ρ_0 increases. Figure 2(b) demonstrates that the worst-case improvements, evaluated at $\rho = \rho_0$, remain consistently positive, further highlighting the robustness of our approach. OLS incurs substantial losses even when $\rho_0 = 0$, reflecting its failure to account for price endogeneity. While TLS-IV, K-IV, and DF-IV perform comparably to the robust policy when $\rho_0 = 0$, their true and worst-case improvements decline sharply as ρ_0 increases, as these methods rely on the valid instrument assumption.



(a) True policy value improvement.



(b) Worst-case policy value improvement (evaluated at ρ).

Figure 3: Mean true and worst-case policy value improvement across sensitivity radii ρ over 200 replications. The data generating process uses $\rho_0 = 5$ with a sample size of $n = 2000$.

Figure 3(a) shows that the proposed robust policy achieves higher true policy value improvement than all competing methods across the sensitivity radii $\rho \in [1.5, 7]$. These results suggest that choosing an appropriate sensitivity radius, guided by sensitivity analysis and domain knowledge, can improve the policy value. Figure 3(b) compares the mean worst-case policy value improvements over ambiguity sets with varying sensitivity radii ρ. All policies are evaluated over the same ambiguity set, and the proposed robust policy achieves the largest mean worst-case improvement among the methods. The proposed policy maintains this advantage even under larger ambiguity sets, which correspond to more severe violations of instrument exogeneity.

Additional experiments are reported in Appendix A.

AI USE STATEMENT

We used generative AI tools to assist with several tasks: brainstorming, sourcing and searching for information, identifying relevant literature, suggesting the structure of the paper, and proposing titles and keywords. Additionally, we used these tools to edit the manuscript to improve readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Chunrong Ai, Yue Fang, and Haitian Xie. Data-driven policy learning for continuous treatments. *Journal of Econometrics*, 253:106170, 2026.

Timothy B. Armstrong and Michal Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021. doi: 10.3982/QE1609.

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

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186 (2):345–366, 2015. doi: 10.1016/j.jeconom.2015.02.014.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

Hengrui Cai, Chengchun Shi, Rui Song, and Wenbin Lu. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34: 15285–15300, 2021.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521, 2016.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, volume 6B, pp. 5549–5632. Elsevier, 2007. doi: 10.1016/S1573-4412(07) 06076-X.

Xiaohong Chen and Timothy M. Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015. doi: 10.1016/j.jeconom.2015.03.010.

Xiaohong Chen and Timothy M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84, 2018. doi: 10.3982/QE722.

Yifan Cui and Eric Tchetgen Tchetgen. A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*, 116(533): 162–173, 2021.

Yuri R Fonseca, Caio F Peixoto, and Yuri F Saporito. Nonparametric instrumental variable regression through stochastic approximate gradients. *Advances in Neural Information Processing Systems*, 37:131756–131785, 2024.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Efficient and sharp off-policy learning under unobserved confounding. In *International Conference on Learning Representations*, volume 2026, pp. 57552–57587, 2026.

Shalmali Joshi, Junzhe Zhang, and Elias Bareinboim. Towards safe policy learning under partial identifiability: A causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13004–13012, 2024.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL https://proceedings.neurips.cc/paper/2018/hash/3a09a524440d44d7f19870070a5ad42f-Abstract.html.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pp. 1243–1251. PMLR, 2018b.

Juno Kim, Dimitri Meunier, Arthur Gretton, Taiji Suzuki, and Zhu Li. Optimality and adaptivity of deep neural features for instrumental variable regression. In *International Conference on Learning Representations*, volume 2025, pp. 94163–94206, 2025.

Benjamin Kompa, David R. Bellamy, Thomas Kolokotrones, James M. Robins, and Andrew L. Beam. Deep learning methods for proximal inference via maximum moment restriction. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 11189–11201. Curran Associates, Inc., 2022. doi: 10.52202/068431-0813. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/487c9d6ef55e73aa9dfd4b48fe3713a6-Abstract.html.

Cheuk Hang Leung, Yiyan Huang, Yijun Li, and Qi Wu. Distributionally robust policy evaluation and learning for continuous treatment with observational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18209–18217, 2025.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, 2004.

Matthew A. Masten and Alexandre Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020. doi: 10.3982/QE1288. URL https://doi.org/10.3982/QE1288.

Rui Miao, Zhengling Qi, Cong Shi, and Lin Lin. Personalized pricing with invalid instrumental variables: Identification, estimation, and policy learning. *arXiv preprint arXiv:2302.12670*, 2023.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997. doi: 10.1016/S0304-4076(97)00011-0.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003. doi: 10.1111/1468-0262.00459. URL https://doi.org/10.1111/1468-0262.00459.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Daqian Shao, Ashkan Soleymani, Francesco Quinzan, and Marta Kwiatkowska. Learning decision policies with instrumental variables through double machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44489–44514. PMLR, 2024. URL https://proceedings.mlr.press/v235/shao24d.html.

Tao Shen and Yifan Cui. Proxy-aided demand learning with an application to various pricing problems. *Operations Research*, 74(2):770–787, 2026. doi: 10.1287/opre.2025.1793. URL https://pubsonline.informs.org/doi/10.1287/opre.2025.1793. Published online November 13, 2025.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/17b3c7061788dbe82de5abe9f6fe22b3-Paper.pdf.

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

Jingyuan Wang, Zhimei Ren, Ruohan Zhan, and Zhengyuan Zhou. Distributionally robust policy learning under concept drifts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267. PMLR, 2025.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=sy4Kg_ZQmS7.

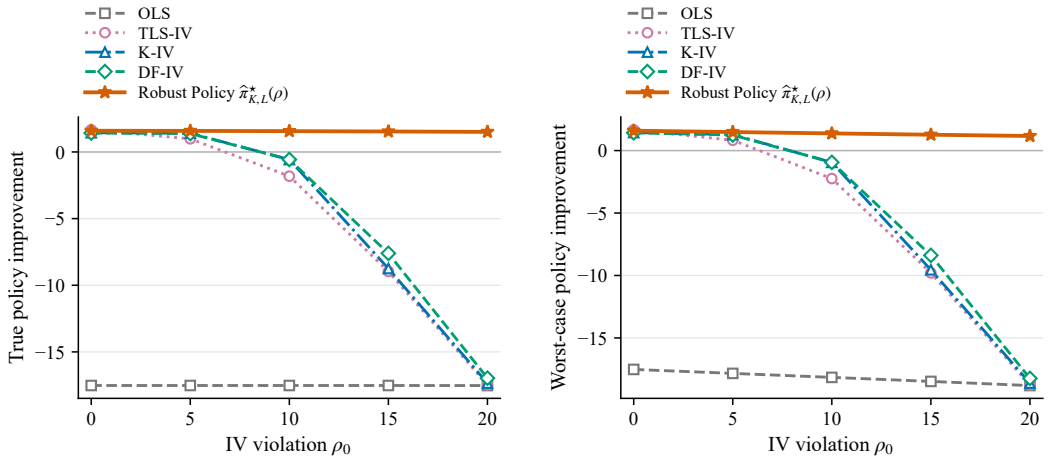
Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

A Additional Simulation Studies

A.1 Supplementary Results for Section 5

Figures 4 and 5 report additional simulation results with $n = 8000$ in Section 5. These results show patterns consistent with those reported in Section 5.2.



(a) True policy value improvement

(b) Worst-case policy value improvement (evaluated at ρ_0)

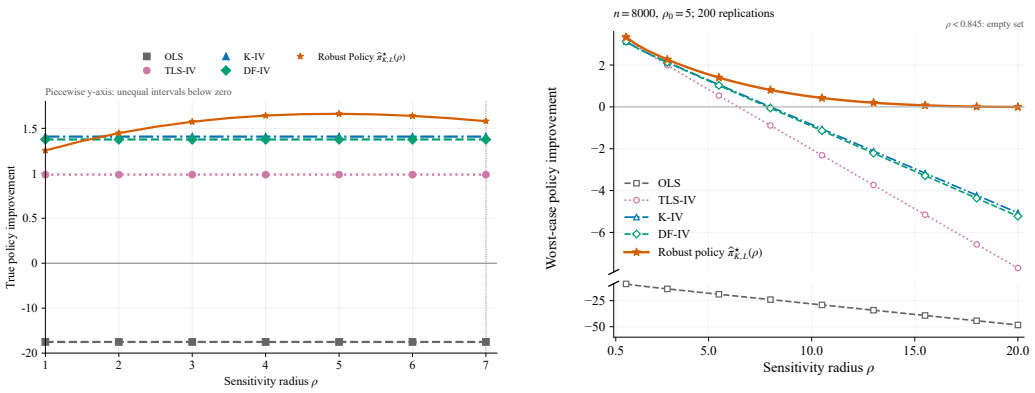
Figure 4: Mean true and worst-case improvements of policy value over 200 replications with $n = 8000$.

As shown in Figure 6(a), the mean true policy value improvement consistently exceeds the target δ, with the 2.5th percentile remaining strictly above it. These results provide strong empirical support for the theoretical guarantees in Theorem 4. Furthermore, Figure 6(b) demonstrates that the mean worst-case policy value improvement closely aligns with the target δ.

A.2 Nonlinear Demand

In the section, we consider nonlinear demand function. Specifically, observed demand is generated by

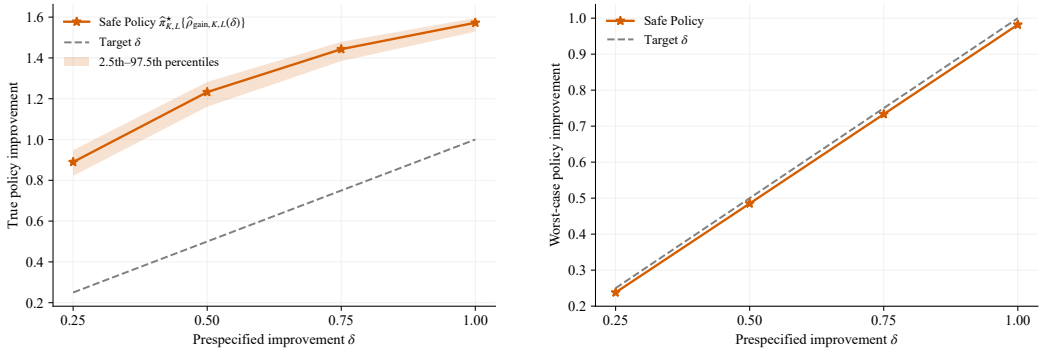
$$Y = 120\left(2.20 + 0.10\sum_{j=1}^{5} x_j - p\right) + 5(1 + 0.5x_1)\sin\left(\frac{2\pi(p - 0.65)}{1.4}\right) + 100U + 0.50\xi.$$



(a) True policy value improvement.

(b) Worst-case policy value improvement (evaluated at ρ).

Figure 5: Mean true and worst-case policy value improvement across sensitivity radii ρ over 200 replications. The data generating process uses $\rho_0 = 5$ with a sample size of $n = 8000$.



(a) True policy value improvement.

(b) Worst-case policy value improvement (evaluated at $\widehat{\rho}_{\text{gain},K,L}(\delta)$)

Figure 6: Mean true and worst-case policy value improvements of the safe policy $\widehat{\pi}^{\star}_{K,L}\{\widehat{\rho}_{\text{gain},K,L}(\delta)\}$ over 200 replications with $n = 8000$ and $\rho_0 = 5$. The shaded band shows the 2.5th–97.5th percentiles of true policy value improvement across 200 replications. The dashed line shows the target δ.

We retain the covariate, instrument, price, and baseline-policy settings of Section 5 and use the sieve spaces

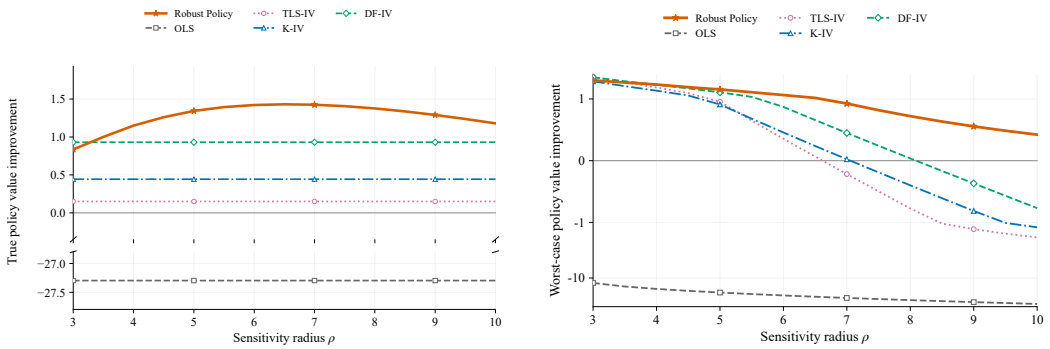
$$\mathcal{S}_7 = \text{span}\{1, x_1, \ldots, x_5, p\}, \qquad \mathcal{T}_8 = \text{span}\{1, x_1, \ldots, x_5, z, z^2\}.$$

We set $n = 8000$ and $\rho_0 = 5$.

Figure 7(a) shows that the robust policy has the highest mean true improvement for $\rho \geq 3.5$ among the methods compared. In Figure 7(b), it has the highest mean worst-case improvement for $\rho \geq 4$. These results suggest that the robust policy largely retains its advantage when the true demand function lies outside $\mathcal{S}_7$.

A.3 Synthetic airline-ticket pricing study

Following standard practice in causal inference (Hess et al., 2026; Shen & Cui, 2026), we evaluate the performance of the proposed method using synthetic data in order to ensure access to ground-truth counterfactuals. Airline-ticket pricing provides a natural setting for studying policy learning under endogeneity. Airlines adjust prices in response to demand conditions, some of which are unobserved in historical data and also affect sales. Motivated by the synthetic airline-ticket pricing examples of Kompa et al. (2022) and Shen & Cui (2026), we consider a setting with webpage views



(a) True policy value improvement.

(b) Worst-case policy value improvement (evaluated at ρ).

Figure 7: Mean policy value improvements across sensitivity radii ρ in the nonlinear-demand setting over 200 replications with $n = 8000$ and $\rho_0 = 5$.

as an observed customer-interest covariate and a fuel-cost index as a candidate instrument. We use this setting to evaluate pricing policy improvement when historical prices are endogenous and the candidate instrument is potentially invalid.

Data-generating process. Let $W \sim \mathrm{Unif}(-1, 1)$ represent a scaled webpage-view covariate. Independently of W, generate $U, Z, \epsilon_P, \epsilon_Y$ independently from $\mathrm{Unif}(-\sqrt{3}, \sqrt{3})$, where U is an unobserved confounder (seasonal fluctuations) and Z is a centered fuel-cost index used as an instrumental variable. Historical prices and sales amounts follow

$$\begin{aligned} P &= 27 + W + 4Z + 2U + \epsilon_P, \\ S &= 100 + (2 + W)(P - 27) - \frac{1}{4}(P - 27)^2 + 2U + 4Z + \epsilon_Y, \end{aligned} \tag{15}$$

Following Shen & Cui (2026), we interpret S as sales amount and directly optimize its expectation, without multiplying by price again. Under an intervention setting price to p, the conditional mean response is

$$f_0(w, p) = \mathbb{E}\{S(p) \mid W = w\} = 100 + (2 + w)(p - 27) - \frac{1}{4}(p - 27)^2.$$

Accordingly, the policy value in this application is $V_f(\pi) = \mathbb{E}[f\{W, \pi(W)\}]$. The shared component U makes historical prices endogenous. The direct term $4Z$ violates instrument exogeneity:

$$\mathbb{E}[S - f_0(W, P) \mid W, Z] = 4Z.$$

Since $\mathbb{E}[Z^2] = 1$, the population violation magnitude is $\left[\mathbb{E}\left\{\mathbb{E}[S - f_0(W, P) \mid W, Z]^2\right\}\right]^{1/2} = 4$. We use quadratic polynomial sieve spaces

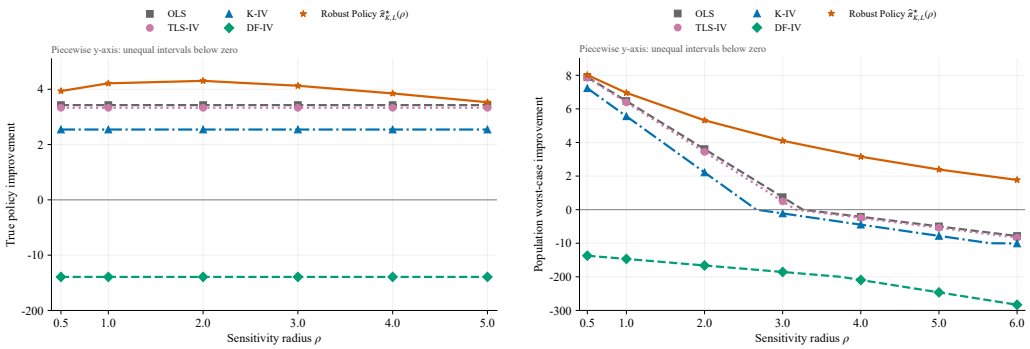
$$\begin{aligned} \mathcal{S}_6 &= \mathrm{span}\{1, w, w^2, p - 27, w(p - 27), (p - 27)^2\}, \\ \mathcal{T}_6 &= \mathrm{span}\{1, w, w^2, z, wz, z^2\}. \end{aligned}$$

We use 200 replications with $n = 2{,}000$ observations each and no additional reference sample. The baseline is the fixed-price policy $\pi_0(w) = 27$. All methods optimize over the same policy class,

$$\Pi = \left\{ \pi_{\boldsymbol{\theta}}(w) = \mathrm{clip}_{[0,50]}\{27 + \theta_0 + \theta_1 w\} : \boldsymbol{\theta} \in \mathbb{R}^2 \right\}.$$

Thus, $\pi_0 \in \Pi$ at $\boldsymbol{\theta} = \mathbf{0}$. The coefficients are unrestricted, while recommended prices lie in $[0, 50]$. We compare Robust Policy with OLS, TLS-IV, K-IV, and DF-IV. OLS and TLS-IV use $\mathcal{S}_6$, with TLS-IV using instruments from $\mathcal{T}_6$. K-IV uses a Nyström approximation with 256 landmarks. Each DF-IV feature network has two hidden layers of 10 units. All policy searches use L-BFGS-B.

We evaluate each learned policy using its true improvement $V_{f_0}(\widehat{\pi}) - V_{f_0}(\pi_0)$ and its population worst-case improvement $\min_{f \in \mathcal{F}(\rho; \mathcal{S}_6)} \{V_f(\widehat{\pi}) - V_f(\pi_0)\}$.

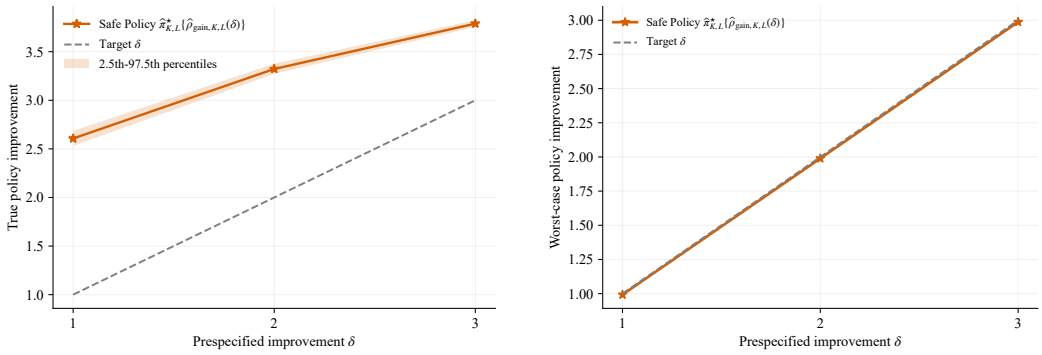


(a) True policy value improvement.

(b) Worst-case policy value improvement (evaluated at ρ).

Figure 8: Mean true and worst-case policy value improvement across sensitivity radii ρ over 200 replications with a sample size of $n = 2000$.

Results. Figure 8(a) shows that the proposed robust policy achieves the largest mean true policy value improvement from $\rho = 0.5$ to $\rho = 5$. Its average improvement peaks at 4.30 when $\rho = 2$, compared with 3.43 for OLS, the best-performing comparison method under this criterion. Figure 8(b) illustrates the trade-off between robustness, controlled by ρ, and worst-case policy improvement. The proposed robust policy achieves the largest mean worst-case improvement among the compared methods at every evaluated radius ρ.



(a) True policy value improvement.

(b) Worst-case policy value improvement (evaluated at $\widehat{\rho}_{\text{gain},K,L}(\delta)$)

Figure 9: Mean true and worst-case policy value improvements of the safe policy $\widehat{\pi}^{\star}_{K,L}\{\widehat{\rho}_{\text{gain},K,L}(\delta)\}$ over 200 replications with $n = 2000$ on airline-ticket pricing data. The shaded band shows the 2.5th–97.5th percentiles of true policy value improvement across 200 replications. The dashed line shows the target δ.

Finally, Figure 9(a) shows mean true policy improvements of 2.607, 3.322, and 3.789 for $\delta = 1, 2, 3$, respectively, with the empirical 2.5th percentiles also exceeding the corresponding targets. Panel (b) shows mean population worst-case improvements of 0.992, 1.989, and 2.987, closely matching these targets. Together, these results suggest that the worst-case targets provide conservative benchmarks for true policy gains in this setting.

B PROOFS

B.1 PROOF OF PROPOSITION 1

Proof. Since $f_0 \in \mathcal{S}$, there exists $f_{0,K} \in \mathcal{S}_K$ such that

$$\|f_{0,K} - f_0\|_\infty = \inf_{\widetilde{f} \in \mathcal{S}_K} \|\widetilde{f} - f_0\|_\infty \leq \alpha_K.$$

By the triangle inequality and the contraction property of conditional expectations,

$$\begin{aligned}
&\left[\mathbb{E}\left[\{\mathbb{E}[Y - f_{0,K}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right]\right]^{1/2} \\
&\leq \left[\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2\right]\right]^{1/2} + \left[\mathbb{E}\left[\{\mathbb{E}[f_0(\boldsymbol{X}, P) - f_{0,K}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right]\right]^{1/2} \\
&\leq \left[\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2\right]\right]^{1/2} + \|f_{0,K} - f_0\|_\infty \\
&\leq \left[\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, Z)\}^2\right]\right]^{1/2} + \alpha_K < \rho.
\end{aligned}$$

Thus, $f_{0,K} \in \mathcal{F}(\rho; \mathcal{S}_K)$. For any $f \in \mathcal{F}(\rho; \mathcal{S}_K)$, the triangle inequality gives

$$\begin{aligned}
&\left[\mathbb{E}\left[\{\mathbb{E}[f(\boldsymbol{X}, P) - f_{0,K}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right]\right]^{1/2} \\
&\quad \leq \left[\mathbb{E}\left[\{\mathbb{E}[Y - f(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right]\right]^{1/2} + \left[\mathbb{E}\left[\{\mathbb{E}[Y - f_{0,K}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right]\right]^{1/2} \leq 2\rho.
\end{aligned} \tag{16}$$

Since $f, f_{0,K} \in \mathcal{S}_K$, there exist $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{0,K}$ such that $f = f_{\boldsymbol{\beta}}$ and $f_{0,K} = f_{\boldsymbol{\beta}_{0,K}}$. Since $f - f_{0,K} = \boldsymbol{\phi}_K^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})$, Equation (16) implies that

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})^\top G_K (\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K}) \leq 4\rho^2.$$

By the definition of $\lambda_{\min}$,

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})^\top G_K (\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K}) \geq \lambda_{\min} (\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})^\top Q_K (\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K}).$$

Combining the preceding two inequalities gives

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})^\top Q_K (\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K}) \leq \frac{4\rho^2}{\lambda_{\min}}.$$

Moreover, by the definition of Q_K and the identity $f - f_{0,K} = \boldsymbol{\phi}_K^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})$, we have

$$\begin{aligned}
&\mathbb{E}\left[\{f(\boldsymbol{X}, P) - f_{0,K}(\boldsymbol{X}, P)\}^2\right] \\
=&\mathbb{E}\left[\left\{\boldsymbol{\phi}_K(\boldsymbol{X}, P)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})\right\}^2\right] \\
=&(\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K})^\top Q_K (\boldsymbol{\beta} - \boldsymbol{\beta}_{0,K}) \\
\leq&\frac{4\rho^2}{\lambda_{\min}}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\left[\mathbb{E}\left[\{f(\boldsymbol{X}, P) - f_0(\boldsymbol{X}, P)\}^2\right]\right]^{1/2} \\
\leq&\left[\mathbb{E}\left[\{f(\boldsymbol{X}, P) - f_{0,K}(\boldsymbol{X}, P)\}^2\right]\right]^{1/2} + \left[\mathbb{E}\left[\{f_{0,K}(\boldsymbol{X}, P) - f_0(\boldsymbol{X}, P)\}^2\right]\right]^{1/2} \\
\leq&\frac{2\rho}{\sqrt{\lambda_{\min}}} + \|f_{0,K} - f_0\|_\infty \\
\leq&\frac{2\rho}{\sqrt{\lambda_{\min}}} + \alpha_K \\
<&\left[\mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, P)\}^2\right]\right]^{1/2},
\end{aligned}$$

where the last inequality follows from the upper bound on ρ. Therefore, for any $f \in \mathcal{F}(\rho; \mathcal{S}_K)$

$$\mathbb{E}\left[\{f(\boldsymbol{X}, P) - f_0(\boldsymbol{X}, P)\}^2\right] < \mathbb{E}\left[\{\mathbb{E}(\epsilon \mid \boldsymbol{X}, P)\}^2\right] = \mathbb{E}\left[\{f_{\text{ols}}(\boldsymbol{X}, P) - f_0(\boldsymbol{X}, P)\}^2\right].$$

□

B.2 Proof of Proposition 2

Proof. Conditional Jensen's inequality and the moment condition in the proposition imply that $g^\star_{f_{\boldsymbol{\beta}}} \in L^2(P_{\boldsymbol{X},Z})$. For every $g \in L^2(P_{\boldsymbol{X},Z})$, iterated expectations give

$$\mathbb{E}\left[g(\boldsymbol{X},Z)\{Y - f_{\boldsymbol{\beta}}(\boldsymbol{X},P)\}\right] = \mathbb{E}\{g(\boldsymbol{X},Z)g^\star_{f_{\boldsymbol{\beta}}}(\boldsymbol{X},Z)\}.$$

Completing the square yields

$$\begin{aligned}
&2\mathbb{E}\left[g(\boldsymbol{X},Z)\{Y - f_{\boldsymbol{\beta}}(\boldsymbol{X},P)\}\right] - \mathbb{E}\{g(\boldsymbol{X},Z)^2\} \\
&\quad = \mathbb{E}\{g^\star_{f_{\boldsymbol{\beta}}}(\boldsymbol{X},Z)^2\} - \mathbb{E}\left[\{g(\boldsymbol{X},Z) - g^\star_{f_{\boldsymbol{\beta}}}(\boldsymbol{X},Z)\}^2\right] \\
&\quad \le \mathbb{E}\{g^\star_{f_{\boldsymbol{\beta}}}(\boldsymbol{X},Z)^2\}.
\end{aligned}$$

Equality holds at $g = g^\star_{f_{\boldsymbol{\beta}}}$. This proves Equation (5) and the stated maximizer. □

B.3 Proof of Theorem 1

Proof. Substituting $g(\boldsymbol{X},Z) = \boldsymbol{\psi}_L(\boldsymbol{X},Z)^\top\boldsymbol{\gamma}$ into the variational criterion gives

$$2\boldsymbol{\gamma}^\top\mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\boldsymbol{\beta}\}\right] - \boldsymbol{\gamma}^\top\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\boldsymbol{\gamma}.$$

The maximum over $\boldsymbol{\gamma}$ is attained at

$$\boldsymbol{\gamma}_{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\boldsymbol{\beta}\}\right],$$

and the maximum equals

$$\mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\boldsymbol{\beta}\}\right]^\top\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\boldsymbol{\beta}\}\right]. \tag{17}$$

The three population quantities used in the theorem are

$$G_{K,L} = \mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top], \tag{18}$$

$$\bar{\boldsymbol{\beta}}_{K,L} = G_{K,L}^{-1}\mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}[\boldsymbol{\psi}_L Y], \tag{19}$$

$$\begin{aligned}
q_{K,L} = {}& \mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\bar{\boldsymbol{\beta}}_{K,L}\}\right]^\top\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1} \\
&\times \mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\bar{\boldsymbol{\beta}}_{K,L}\}\right].
\end{aligned} \tag{20}$$

The first-order condition defining $\bar{\boldsymbol{\beta}}_{K,L}$ is

$$\mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]^{-1}\mathbb{E}\left[\boldsymbol{\psi}_L\{Y - \boldsymbol{\phi}_K^\top\bar{\boldsymbol{\beta}}_{K,L}\}\right] = \mathbf{0}.$$

Expanding Equation (17) around $\bar{\boldsymbol{\beta}}_{K,L}$ therefore gives

$$\begin{aligned}
&\max_{g\in\mathcal{T}_L}\left\{2\mathbb{E}\left[g(\boldsymbol{X},Z)\{Y - f_{\boldsymbol{\beta}}(\boldsymbol{X},P)\}\right] - \mathbb{E}\{g(\boldsymbol{X},Z)^2\}\right\} \\
&\quad = q_{K,L} + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_{K,L})^\top G_{K,L}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_{K,L}).
\end{aligned} \tag{21}$$

Consequently, the double-sieve ambiguity set is nonempty when $\rho^2 \ge q_{K,L}$, and its coefficient vectors satisfy

$$\left\|G_{K,L}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_{K,L})\right\|_2 \le \sqrt{\rho^2 - q_{K,L}}.$$

The policy improvement is linear in $\boldsymbol{\beta}$, with gradient $\nabla_{\boldsymbol{\beta}}\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}$. For every feasible coefficient vector, Cauchy–Schwarz gives

$$\begin{aligned}
&\nabla_{\boldsymbol{\beta}}\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}^\top(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_{K,L}) \\
&\quad \ge -\sqrt{\rho^2 - q_{K,L}}\left\|G_{K,L}^{-1/2}\nabla_{\boldsymbol{\beta}}\{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2.
\end{aligned}$$

This proves Equation (8). If the policy-value gradient is nonzero, equality is attained at the coefficient vector in Equation (9). If the gradient is zero, the policy improvement is constant over the ambiguity set, so every feasible coefficient is a minimizer; in particular, one may take $\boldsymbol{\beta}_\pi^{\mathrm{wc}} = \bar{\boldsymbol{\beta}}_{K,L}$. □

B.4 Proof of Theorem 2

Proof. By Theorem 1, the policy-specific worst-case improvement is

$$V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi_0) - \sqrt{\rho^2 - q_{K,L}} \left\| G_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\} \right\|_2 . \tag{22}$$

Each policy-specific curve is continuous and nonincreasing in ρ. Compactness of Π and continuity in π imply that the maximum is attained and that $M_{K,L}$ is continuous.

Whenever the first line of Equation (22) is positive, its gradient norm in the second line is also positive. Indeed, a zero gradient would make the linear policy improvement equal to zero at every coefficient vector, including $\bar{\boldsymbol{\beta}}_{K,L}$. Now take $\rho_1 < \rho_2$ such that $M_{K,L}(\rho_2) \geq \delta$, and let a policy attain the maximum at ρ_2. Its fitted improvement exceeds zero and its gradient norm is positive. Evaluating the same policy at ρ_1 shows that $M_{K,L}(\rho_1) > M_{K,L}(\rho_2)$. Thus, $M_{K,L}$ is strictly decreasing wherever it is at least δ.

At $\rho = \sqrt{q_{K,L}}$, the assumed fitted improvement gives $M_{K,L}(\sqrt{q_{K,L}}) > \delta$. On the compact set of policies whose fitted improvement is at least δ, the gradient norm in Equation (22) is continuous and bounded away from zero. Consequently, $M_{K,L}(\rho) < \delta$ for all sufficiently large ρ. Continuity and strict monotonicity establish existence and uniqueness of the solution in Equation (12). □

B.5 Definition of Hölder class

Write $s = r + \nu$, where r is a nonnegative integer and $\nu \in (0, 1]$. For a compact rectangular domain $\mathcal{D} \subset \mathbb{R}^d$, define the Hölder norm by

$$\|g\|_{C^{r,\nu}(\mathcal{D})} = \max_{|\boldsymbol{a}|_1 \leq r} \|D^{\boldsymbol{a}} g\|_\infty + \max_{|\boldsymbol{a}|_1 = r} \sup_{\substack{\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{D} \\ \boldsymbol{z}_1 \neq \boldsymbol{z}_2}} \frac{|D^{\boldsymbol{a}} g(\boldsymbol{z}_1) - D^{\boldsymbol{a}} g(\boldsymbol{z}_2)|}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2^\nu}.$$

Here $\boldsymbol{a}$ is a multi-index, and $D^{\boldsymbol{a}} g$ denotes the corresponding mixed derivative. The Hölder space $C^{r,\nu}(\mathcal{D})$ consists of functions $g \in C^r(\mathcal{D})$ for which this norm is finite. The corresponding Hölder ball is

$$\mathcal{H}^s(\mathcal{D}, B) = \left\{ g \in C^{r,\nu}(\mathcal{D}) : \|g\|_{C^{r,\nu}(\mathcal{D})} \leq B \right\}.$$

Integer smoothness s is covered by taking $\nu = 1$ and $r = s - 1$.

B.6 Additional regularity conditions

The observations $\{(Y_i, P_i, \boldsymbol{X}_i, Z_i)\}_{i=1}^n$ are independent and identically distributed, and the basis functions are deterministic. All uniform constants c and C below are independent of n, K, L.

Assumption 3 (Regularity conditions)**.** (i) Assume the policy class $\Pi = \{\pi_{\boldsymbol{\vartheta}} : \boldsymbol{\vartheta} \in \Theta\}$ contains the baseline π_0, where $\Theta \subset \mathbb{R}^d$ is compact with fixed dimension d and bounded diameter. For all $\pi \in \Pi$, $(\boldsymbol{X}, \pi(\boldsymbol{X})) \in \mathcal{D}_f$ almost surely. There exist fixed constants $a \geq 0$ and $C < \infty$, independent of n, K, L, such that for all $\boldsymbol{\vartheta}, \boldsymbol{\vartheta}' \in \Theta$,

$$\left\| \pi_{\boldsymbol{\vartheta}}(\boldsymbol{X}) \boldsymbol{\phi}_K\{\boldsymbol{X}, \pi_{\boldsymbol{\vartheta}}(\boldsymbol{X})\} - \pi_{\boldsymbol{\vartheta}'}(\boldsymbol{X}) \boldsymbol{\phi}_K\{\boldsymbol{X}, \pi_{\boldsymbol{\vartheta}'}(\boldsymbol{X})\} \right\|_2 \leq C K^a \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2.$$

(ii) There exists a fixed constant $c_0 > 0$ such that both $\rho^2 - q_{K,L} \geq c_0$ and $\rho_{\mathrm{gain,K}}(\delta)^2 - q_{K,L} \geq c_0$ for the target $\delta > 0$.

Assumption 3(i) allows the Lipschitz constant to grow polynomially with K, a condition commonly imposed in the sieve estimation literature (Belloni et al., 2015). Assumption 3(ii) ensures that the squared sensitivity radii exceed their minimum feasible values by a uniform positive margin. Thus, $\mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)$ and $\mathcal{F}(\rho_{\mathrm{gain,K}}(\delta); \mathcal{S}_K, \mathcal{T}_L)$ are nonempty.

B.7 Proof of Theorem 3

Define

$$\boldsymbol{h}_\pi(\boldsymbol{X}) = \pi(\boldsymbol{X}) \boldsymbol{\phi}_K\{\boldsymbol{X}, \pi(\boldsymbol{X})\} - \pi_0(\boldsymbol{X}) \boldsymbol{\phi}_K\{\boldsymbol{X}, \pi_0(\boldsymbol{X})\}.$$

We first prove the following lemma.

Lemma 1 (Stability of the sieve value)**.** *Suppose Assumptions 2 and 3(i) hold. Define*

$$\Delta_n = \|(E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\psi}_L^\top]\|_{\mathrm{op}} + \|(E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\phi}_K^\top]\|_{\mathrm{op}} + \|(E_n - E)[\boldsymbol{\psi}_L Y]\|_2 + \sup_{\pi \in \Pi} \|(E_n - E)\boldsymbol{h}_\pi\|_2,$$

and write

$$m_{K,L}(\pi, r) = \min_{f \in \mathcal{F}(r; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\},$$

whenever the feasible set is nonempty. Let $\widehat{m}_{K,L}$ denote the same optimization problem with E replaced by E_n.

For fixed $R, c_* > 0$, there exist constants $\epsilon_0, C_0 > 0$, independent of n, K, L, such that, whenever $\tau_{K,L}\Delta_n \le \epsilon_0$, the matrices $\widehat{Q}_L$ and $\widehat{G}_{K,L}$ are positive definite and

$$|\widehat{q}_{K,L} - q_{K,L}| \le C_0 \tau_{K,L} \Delta_n. \tag{23}$$

Moreover, uniformly over $\pi \in \Pi$ and $0 \le r \le R$ satisfying $r^2 - q_{K,L} \ge c_*$, all minima below are well-defined and attained, and

$$\begin{aligned} & m_{K,L}\left(\pi, \sqrt{r^2 + C_0 \tau_{K,L} \Delta_n}\right) - C_0 \tau_{K,L} \Delta_n \\ & \quad \le \widehat{m}_{K,L}(\pi, r) \\ & \quad \le m_{K,L}\left(\pi, \sqrt{r^2 - C_0 \tau_{K,L} \Delta_n}\right) + C_0 \tau_{K,L} \Delta_n, \end{aligned} \tag{24}$$

$$|\widehat{m}_{K,L}(\pi, r) - m_{K,L}(\pi, r)| \le C_0 \tau_{K,L}^2 \zeta_{K,L} \Delta_n. \tag{25}$$

Proof of Lemma 1.

Recall the population and sample constraint functions

$$J_{K,L}(\boldsymbol{\beta}) = \sup_{g \in \mathcal{T}_L} \{2\mathbb{E}[g(Y - f_{\boldsymbol{\beta}})] - \mathbb{E}[g^2]\},$$
$$\widehat{J}_{K,L}(\boldsymbol{\beta}) = \sup_{g \in \mathcal{T}_L} \{2E_n[g(Y - f_{\boldsymbol{\beta}})] - E_n[g^2]\}.$$

The population and sample policy objectives satisfy

$$V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0) = \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{h}_\pi],$$
$$\widehat{V}_{f_{\boldsymbol{\beta}}}(\pi) - \widehat{V}_{f_{\boldsymbol{\beta}}}(\pi_0) = \boldsymbol{\beta}^\top E_n[\boldsymbol{h}_\pi].$$

Hence, whenever the feasible coefficient sets are nonempty and compact, adding and subtracting the minimum of the population objective over the sample feasible set gives

$$\begin{aligned} & |\widehat{m}_{K,L}(\pi, r) - m_{K,L}(\pi, r)| \\ & \le \underbrace{\left| \min_{\boldsymbol{\beta}: \widehat{J}_{K,L}(\boldsymbol{\beta}) \le r^2} \boldsymbol{\beta}^\top E_n[\boldsymbol{h}_\pi] - \min_{\boldsymbol{\beta}: \widehat{J}_{K,L}(\boldsymbol{\beta}) \le r^2} \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{h}_\pi] \right|}_{\mathrm{I}_\pi} \\ & \quad + \underbrace{\left| \min_{\boldsymbol{\beta}: \widehat{J}_{K,L}(\boldsymbol{\beta}) \le r^2} \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{h}_\pi] - \min_{\boldsymbol{\beta}: J_{K,L}(\boldsymbol{\beta}) \le r^2} \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{h}_\pi] \right|}_{\mathrm{II}_\pi}. \end{aligned} \tag{26}$$

Parts (i) and (ii) establish compactness and nonemptiness. Part (iii) bounds I_π and II_π, corresponding to estimation of the objective and the constraint, respectively. All bounds are uniform over $\pi \in \Pi$ and $0 \le r \le R$ satisfying $r^2 - q_{K,L} \ge c_*$.

(i) Compactness of the feasible coefficient sets.

Write

$$A_{K,L} = \mathbb{E}[\psi_L \phi_K^\top], \quad \widehat{A}_{K,L} = E_n[\psi_L \phi_K^\top],$$
$$Q_L = \mathbb{E}[\psi_L \psi_L^\top], \quad \widehat{Q}_L = E_n[\psi_L \psi_L^\top].$$

We first justify the quadratic representation

$$\widehat{J}_{K,L}(\boldsymbol{\beta}) = E_n[\psi_L(Y - f_{\boldsymbol{\beta}})]^\top \widehat{Q}_L^{-1} E_n[\psi_L(Y - f_{\boldsymbol{\beta}})]. \tag{27}$$

By $G_{K,L} \preceq Q_K \preceq C I_K$,

$$\tau_{K,L} \geq C^{-1/2}, \qquad \Delta_n \leq C^{1/2} \tau_{K,L} \Delta_n.$$

Weyl's inequality gives

$$\lambda_{\min}(\widehat{Q}_L) \geq c - \Delta_n, \qquad \lambda_{\max}(\widehat{Q}_L) \leq C + \Delta_n.$$

Whenever $\tau_{K,L}\Delta_n \leq \epsilon_0$, choose $C^{1/2}\epsilon_0 \leq \min\{c/2, 1\}$ to obtain

$$\frac{c}{2} I_L \preceq \widehat{Q}_L \preceq (C+1) I_L, \qquad (C+1)^{-1} I_L \preceq \widehat{Q}_L^{-1} \preceq \frac{2}{c} I_L. \tag{28}$$

For $g_{\boldsymbol{\gamma}} = \psi_L^\top \boldsymbol{\gamma}$, completing the square gives

$$\begin{aligned}
&2E_n[g_{\boldsymbol{\gamma}}(Y - f_{\boldsymbol{\beta}})] - E_n[g_{\boldsymbol{\gamma}}^2] \\
&\quad = 2\boldsymbol{\gamma}^\top E_n[\psi_L(Y - f_{\boldsymbol{\beta}})] - \boldsymbol{\gamma}^\top \widehat{Q}_L \boldsymbol{\gamma} \\
&\quad = E_n[\psi_L(Y - f_{\boldsymbol{\beta}})]^\top \widehat{Q}_L^{-1} E_n[\psi_L(Y - f_{\boldsymbol{\beta}})] \\
&\qquad - \{\boldsymbol{\gamma} - \widehat{Q}_L^{-1} E_n[\psi_L(Y - f_{\boldsymbol{\beta}})]\}^\top \widehat{Q}_L \{\boldsymbol{\gamma} - \widehat{Q}_L^{-1} E_n[\psi_L(Y - f_{\boldsymbol{\beta}})]\}.
\end{aligned}$$

The last quadratic term is nonnegative and equals zero at $\boldsymbol{\gamma} = \widehat{Q}_L^{-1} E_n[\psi_L(Y - f_{\boldsymbol{\beta}})]$. Taking the supremum proves Equation (27). The same calculation gives

$$J_{K,L}(\boldsymbol{\beta}) = \mathbb{E}[\psi_L(Y - f_{\boldsymbol{\beta}})]^\top Q_L^{-1} \mathbb{E}[\psi_L(Y - f_{\boldsymbol{\beta}})].$$

Both criteria are continuous, so their sublevel sets are closed. To show boundedness, we establish a coefficient bound on the slightly larger range

$$\min\{J_{K,L}(\boldsymbol{\beta}), \widehat{J}_{K,L}(\boldsymbol{\beta})\} \leq R^2 + \mathbb{E}[Y^2] + 1. \tag{29}$$

This range contains all feasible coefficients at $r \leq R$. It will also contain the criterion minimizers and the coefficients at the perturbed radii used below.

First, Cauchy–Schwarz and the definition of Δ_n give

$$\|\mathbb{E}[\psi_L Y]\|_2 \leq \|Q_L\|_{\mathrm{op}}^{1/2} \sqrt{\mathbb{E}[Y^2]} \leq C,$$
$$\|E_n[\psi_L Y]\|_2 \leq \|\mathbb{E}[\psi_L Y]\|_2 + \Delta_n \leq C.$$

If $J_{K,L}(\boldsymbol{\beta}) \leq R^2 + \mathbb{E}[Y^2] + 1$, then

$$\begin{aligned}
\|A_{K,L}\boldsymbol{\beta}\|_2 &\leq \|\mathbb{E}[\psi_L Y]\|_2 + \|\mathbb{E}[\psi_L(Y - f_{\boldsymbol{\beta}})]\|_2 \\
&\leq \|\mathbb{E}[\psi_L Y]\|_2 + \|Q_L\|_{\mathrm{op}}^{1/2} \sqrt{J_{K,L}(\boldsymbol{\beta})} \leq C.
\end{aligned}$$

Since $G_{K,L} = A_{K,L}^\top Q_L^{-1} A_{K,L}$,

$$\|A_{K,L}\boldsymbol{\beta}\|_2^2 \geq c\boldsymbol{\beta}^\top G_{K,L} \boldsymbol{\beta} \geq c\tau_{K,L}^{-2} \|\boldsymbol{\beta}\|_2^2. \tag{30}$$

Thus $\|\boldsymbol{\beta}\|_2 \leq C\tau_{K,L}$ for these population coefficients.

For the sample coefficients, the triangle inequality gives

$$\begin{aligned}
\|\widehat{A}_{K,L}\boldsymbol{\beta}\|_2 &\geq \|A_{K,L}\boldsymbol{\beta}\|_2 - \|\widehat{A}_{K,L} - A_{K,L}\|_{\mathrm{op}} \|\boldsymbol{\beta}\|_2 \\
&\geq \left(\frac{\sqrt{c}}{\tau_{K,L}} - \Delta_n\right) \|\boldsymbol{\beta}\|_2 \geq \frac{\sqrt{c}}{2\tau_{K,L}} \|\boldsymbol{\beta}\|_2,
\end{aligned}$$

where the last inequality holds when $\tau_{K,L}\Delta_n \le \epsilon_0 \le \sqrt{c}/2$. Combining this with Equation (28) yields

$$\boldsymbol{\beta}^\top \widehat{G}_{K,L}\boldsymbol{\beta} = \|\widehat{Q}_L^{-1/2}\widehat{A}_{K,L}\boldsymbol{\beta}\|_2^2 \ge \frac{c}{4(C+1)}\tau_{K,L}^{-2}\|\boldsymbol{\beta}\|_2^2. \tag{31}$$

In particular, $\widehat{G}_{K,L}$ is positive definite. If $\widehat{J}_{K,L}(\boldsymbol{\beta}) \le R^2 + \mathbb{E}[Y^2] + 1$, then

$$\begin{aligned}\|\widehat{A}_{K,L}\boldsymbol{\beta}\|_2 &\le \|E_n[\boldsymbol{\psi}_L Y]\|_2 + \|E_n[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 \\ &\le \|E_n[\boldsymbol{\psi}_L Y]\|_2 + \|\widehat{Q}_L\|_{\mathrm{op}}^{1/2}\sqrt{\widehat{J}_{K,L}(\boldsymbol{\beta})} \le C.\end{aligned}$$

Therefore, uniformly over the coefficient vectors in Equation (29),

$$\|\boldsymbol{\beta}\|_2 \le C\tau_{K,L}. \tag{32}$$

For each finite (K, L), both feasible coefficient sets at $r \le R$ are thus closed and bounded, and hence compact.

(ii) Nonemptiness of the feasible coefficient sets.

The population center satisfies

$$J_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) = q_{K,L}, \qquad q_{K,L} \le J_{K,L}(0) \le \mathbb{E}[Y^2].$$

Hence Equation (32) gives $\|\bar{\boldsymbol{\beta}}_{K,L}\|_2 \le C\tau_{K,L}$. For every radius in the lemma, $q_{K,L} \le r^2 - c_*$, so the population feasible set is nonempty. To show that the same center is also sample-feasible, we bound $|\widehat{J}_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) - q_{K,L}|$.

First, the moment difference satisfies

$$\begin{aligned}&\|(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2 \\ &\quad= \|(E_n - E)[\boldsymbol{\psi}_L Y] - (\widehat{A}_{K,L} - A_{K,L})\bar{\boldsymbol{\beta}}_{K,L}\|_2 \\ &\quad\le \|(E_n - E)[\boldsymbol{\psi}_L Y]\|_2 + \|\widehat{A}_{K,L} - A_{K,L}\|_{\mathrm{op}}\|\bar{\boldsymbol{\beta}}_{K,L}\|_2 \\ &\quad\le \Delta_n(1 + C\tau_{K,L}) \le C\tau_{K,L}\Delta_n.\end{aligned}$$

Second, the population residual moment obeys

$$\|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2^2 \le \|Q_L\|_{\mathrm{op}} q_{K,L} \le C.$$

Consequently, when $\tau_{K,L}\Delta_n \le \epsilon_0 \le 1$,

$$\begin{aligned}&\|E_n[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2 \\ &\quad\le \|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2 + \|(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2 \\ &\quad\le C + C\tau_{K,L}\Delta_n \le C.\end{aligned}$$

Finally, Equation (28) and the inverse identity imply

$$\begin{aligned}\|\widehat{Q}_L^{-1} - Q_L^{-1}\|_{\mathrm{op}} &= \|\widehat{Q}_L^{-1}(Q_L - \widehat{Q}_L)Q_L^{-1}\|_{\mathrm{op}} \\ &\le \|\widehat{Q}_L^{-1}\|_{\mathrm{op}}\|\widehat{Q}_L - Q_L\|_{\mathrm{op}}\|Q_L^{-1}\|_{\mathrm{op}} \le \frac{2}{c^2}\Delta_n.\end{aligned} \tag{33}$$

Subtracting the two quadratic criteria at the population center and substituting these bounds yields

$$\begin{aligned}&|\widehat{J}_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) - q_{K,L}| \\ &\quad\le \|\widehat{Q}_L^{-1}\|_{\mathrm{op}}\|(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2 \\ &\qquad\times \left\{\|E_n[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2 + \|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2\right\} \\ &\qquad+ \|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\bar{\boldsymbol{\beta}}_{K,L}})]\|_2^2\|\widehat{Q}_L^{-1} - Q_L^{-1}\|_{\mathrm{op}} \\ &\quad\le C\tau_{K,L}\Delta_n + C\Delta_n \le C\tau_{K,L}\Delta_n.\end{aligned} \tag{34}$$

For ϵ_0 small enough that $C\epsilon_0 \le c_*/2$, whenever $\tau_{K,L}\Delta_n \le \epsilon_0$,

$$\begin{aligned}\widehat{J}_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) &\le q_{K,L} + C\tau_{K,L}\Delta_n \\ &\le r^2 - c_* + C\epsilon_0 \le r^2 - c_*/2 < r^2.\end{aligned}$$

Thus the population center belongs to both feasible coefficient sets. Their compactness from part (i) ensures that the minima in Equation (26) are finite and attained.

(iii) Bounds for the two estimation errors.

(a) Bounding I_π.

By Equation (26),

$$\begin{aligned}\mathrm{I}_\pi &\le \sup_{\boldsymbol{\beta}:\widehat{J}_{K,L}(\boldsymbol{\beta})\le r^2} |\boldsymbol{\beta}^\top E_n[\boldsymbol{h}_\pi] - \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{h}_\pi]| \\ &= \sup_{\boldsymbol{\beta}:\widehat{J}_{K,L}(\boldsymbol{\beta})\le r^2} |\boldsymbol{\beta}^\top (E_n - E)\boldsymbol{h}_\pi|.\end{aligned}$$

For every $\boldsymbol{\beta}$ and $\pi \in \Pi$,

$$|\boldsymbol{\beta}^\top (E_n - E)\boldsymbol{h}_\pi| \le \|\boldsymbol{\beta}\|_2 \|(E_n - E)\boldsymbol{h}_\pi\|_2 \le \|\boldsymbol{\beta}\|_2 \Delta_n. \tag{35}$$

Since $r \le R$, all sample-feasible coefficients satisfy Equation (29). Combining Equations (32) and (35),

$$\sup_{\pi\in\Pi} \mathrm{I}_\pi \le \sup_{\substack{\pi\in\Pi \\ \boldsymbol{\beta}:\widehat{J}_{K,L}(\boldsymbol{\beta})\le r^2}} |\boldsymbol{\beta}^\top (E_n - E)\boldsymbol{h}_\pi| \le C\tau_{K,L}\Delta_n. \tag{36}$$

(b) Bounding II_π.

We will bound II_π by the change in the population minimum when the squared radius changes by $C\tau_{K,L}\Delta_n$:

$$\begin{aligned}\mathrm{II}_\pi \le \max\Big\{ & m_{K,L}(\pi, \sqrt{r^2 - C\tau_{K,L}\Delta_n}) - m_{K,L}(\pi, r), \\ & m_{K,L}(\pi, r) - m_{K,L}(\pi, \sqrt{r^2 + C\tau_{K,L}\Delta_n})\Big\}.\end{aligned} \tag{37}$$

For Equation (37), it suffices to show

$$\begin{aligned}m_{K,L}(\pi, \sqrt{r^2 + C\tau_{K,L}\Delta_n}) &\le \min_{\boldsymbol{\beta}:\widehat{J}_{K,L}(\boldsymbol{\beta})\le r^2} \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{h}_\pi] \\ &\le m_{K,L}(\pi, \sqrt{r^2 - C\tau_{K,L}\Delta_n}).\end{aligned} \tag{38}$$

Indeed, $m_{K,L}(\pi, r)$ also lies between these two bounds. The inequalities in Equation (38) follow from the set inclusions

$$\begin{aligned}&\{\boldsymbol{\beta} : J_{K,L}(\boldsymbol{\beta}) \le r^2 - C\tau_{K,L}\Delta_n\} \\ &\quad\subseteq \{\boldsymbol{\beta} : \widehat{J}_{K,L}(\boldsymbol{\beta}) \le r^2\} \\ &\quad\subseteq \{\boldsymbol{\beta} : J_{K,L}(\boldsymbol{\beta}) \le r^2 + C\tau_{K,L}\Delta_n\}.\end{aligned} \tag{39}$$

We establish these inclusions by proving

$$\sup_{\boldsymbol{\beta}:\min\{J_{K,L}(\boldsymbol{\beta}),\widehat{J}_{K,L}(\boldsymbol{\beta})\}\le R^2+\mathbb{E}[Y^2]+1} |\widehat{J}_{K,L}(\boldsymbol{\beta}) - J_{K,L}(\boldsymbol{\beta})| \le C\tau_{K,L}\Delta_n. \tag{40}$$

The same bound will also control $|\widehat{q}_{K,L} - q_{K,L}|$.

Subtracting the population criterion from the sample criterion,

$$\begin{aligned}&\widehat{J}_{K,L}(\boldsymbol{\beta}) - J_{K,L}(\boldsymbol{\beta}) \\ &\quad= \{(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\}^\top \widehat{Q}_L^{-1} E_n[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})] \\ &\qquad+ \mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]^\top \widehat{Q}_L^{-1} (E_n - E)[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})] \\ &\qquad+ \mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]^\top (\widehat{Q}_L^{-1} - Q_L^{-1}) \mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})].\end{aligned}$$

Applying $|\boldsymbol{a}^\top M \boldsymbol{b}| \le \|\boldsymbol{a}\|_2 \|M\|_{\mathrm{op}} \|\boldsymbol{b}\|_2$,

$$\begin{aligned}
&|\widehat{J}_{K,L}(\boldsymbol{\beta}) - J_{K,L}(\boldsymbol{\beta})| \\
&\quad \le \underbrace{\|\widehat{Q}_L^{-1}\|_{\mathrm{op}}}_{(J_1)} \underbrace{\|(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2}_{(J_2)} \\
&\qquad \times \left(\underbrace{\|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2}_{(J_3)} + \underbrace{\|E_n[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2}_{(J_4)} \right) \\
&\qquad + \underbrace{\|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2}_{(J_3)} \underbrace{\|\widehat{Q}_L^{-1} - Q_L^{-1}\|_{\mathrm{op}}}_{(J_5)}.
\end{aligned}$$

We bound these five factors uniformly over the coefficient vectors in Equation (29).

Bound for (J_1). Equation (28) gives

$$\|\widehat{Q}_L^{-1}\|_{\mathrm{op}} \le 2/c.$$

Bound for (J_2). By linearity of $f_{\boldsymbol{\beta}}$ and Equation (32),

$$\begin{aligned}
&\|(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 \\
&\quad = \|(E_n - E)[\boldsymbol{\psi}_L Y] - (E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\phi}_K^\top]\boldsymbol{\beta}\|_2 \\
&\quad \le \|(E_n - E)[\boldsymbol{\psi}_L Y]\|_2 + \|(E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\phi}_K^\top]\|_{\mathrm{op}} \|\boldsymbol{\beta}\|_2 \\
&\quad \le \Delta_n(1 + C\tau_{K,L}) \le C\tau_{K,L}\Delta_n.
\end{aligned}$$

Bound for (J_3). If $J_{K,L}(\boldsymbol{\beta}) \le R^2 + \mathbb{E}[Y^2] + 1$, then

$$\begin{aligned}
\|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2^2 &\le \|Q_L\|_{\mathrm{op}} J_{K,L}(\boldsymbol{\beta}) \\
&\le C\{R^2 + \mathbb{E}[Y^2] + 1\} \le C.
\end{aligned}$$

If $\widehat{J}_{K,L}(\boldsymbol{\beta}) \le R^2 + \mathbb{E}[Y^2] + 1$, the bound for (J_2) instead gives

$$\begin{aligned}
&\|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 \\
&\quad \le \|E_n[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 + \|(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 \\
&\quad \le \|\widehat{Q}_L\|_{\mathrm{op}}^{1/2} \sqrt{\widehat{J}_{K,L}(\boldsymbol{\beta})} + C\tau_{K,L}\Delta_n \\
&\quad \le \sqrt{(C+1)\{R^2 + \mathbb{E}[Y^2] + 1\}} + C\tau_{K,L}\Delta_n \le C,
\end{aligned}$$

where the last inequality uses $\tau_{K,L}\Delta_n \le \epsilon_0 \le 1$. Equation (29) ensures that at least one of these two cases applies. Hence $(J_3) \le C$.

Bound for (J_4). By the triangle inequality and the bounds for (J_2) and (J_3),

$$\begin{aligned}
&\|E_n[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 \\
&\quad \le \|\mathbb{E}[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 + \|(E_n - E)[\boldsymbol{\psi}_L(Y - f_{\boldsymbol{\beta}})]\|_2 \\
&\quad \le C + C\tau_{K,L}\Delta_n \le C.
\end{aligned}$$

Bound for (J_5). Using Equation (33),

$$\begin{aligned}
\|\widehat{Q}_L^{-1} - Q_L^{-1}\|_{\mathrm{op}} &\le \|\widehat{Q}_L^{-1}\|_{\mathrm{op}} \|\widehat{Q}_L - Q_L\|_{\mathrm{op}} \|Q_L^{-1}\|_{\mathrm{op}} \\
&\le (2/c)\Delta_n(1/c) = 2\Delta_n/c^2.
\end{aligned}$$

Substituting the five bounds into the criterion-error inequality,

$$\begin{aligned}
|\widehat{J}_{K,L}(\boldsymbol{\beta}) - J_{K,L}(\boldsymbol{\beta})| &\le (2/c)(C\tau_{K,L}\Delta_n)(C + C) + C^2(2\Delta_n/c^2) \\
&\le C\tau_{K,L}\Delta_n,
\end{aligned}$$

where the last inequality uses $\tau_{K,L} \geq C^{-1/2}$. The bound is uniform on the range in Equation (29), which proves Equation (40).

We first apply this bound at the two criterion minimizers. By Equation (34),

$$\begin{aligned} q_{K,L} &= J_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) \leq \mathbb{E}[Y^2], \\ \widehat{q}_{K,L} &= \widehat{J}_{K,L}(\widehat{\bar{\boldsymbol{\beta}}}_{K,L}) \leq \widehat{J}_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) \\ &\leq q_{K,L} + C\tau_{K,L}\Delta_n \leq \mathbb{E}[Y^2] + 1, \end{aligned}$$

provided $C\tau_{K,L}\Delta_n \leq 1$. Both minimizers therefore satisfy Equation (29). Their minimizing properties and Equation (40) imply

$$\begin{aligned} \widehat{q}_{K,L} - q_{K,L} &\leq \widehat{J}_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) - J_{K,L}(\bar{\boldsymbol{\beta}}_{K,L}) \leq C\tau_{K,L}\Delta_n, \\ q_{K,L} - \widehat{q}_{K,L} &\leq J_{K,L}(\widehat{\bar{\boldsymbol{\beta}}}_{K,L}) - \widehat{J}_{K,L}(\widehat{\bar{\boldsymbol{\beta}}}_{K,L}) \leq C\tau_{K,L}\Delta_n. \end{aligned}$$

Consequently,

$$|\widehat{q}_{K,L} - q_{K,L}| \leq C\tau_{K,L}\Delta_n. \tag{41}$$

Before comparing the perturbed radii, we also bound their policy-value directions. The bounded price interval and the definition of $\zeta_{K,L}$ give

$$\begin{aligned} \sup_{\pi\in\Pi} \|\mathbb{E}[\boldsymbol{h}_\pi]\|_2 &\leq \sup_{\pi\in\Pi} E\|\boldsymbol{h}_\pi\|_2 \\ &\leq \sup_{\pi\in\Pi} E\big[|\pi(\boldsymbol{X})|\|\boldsymbol{\phi}_K\{\boldsymbol{X},\pi(\boldsymbol{X})\}\|_2 \\ &\qquad + |\pi_0(\boldsymbol{X})|\|\boldsymbol{\phi}_K\{\boldsymbol{X},\pi_0(\boldsymbol{X})\}\|_2\big] \\ &\leq C\zeta_{K,L}. \end{aligned}$$

Thus

$$\sup_{\pi\in\Pi} \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2 \leq \|G_{K,L}^{-1/2}\|_{\mathrm{op}} \sup_{\pi\in\Pi} \|\mathbb{E}[\boldsymbol{h}_\pi]\|_2 \leq C\tau_{K,L}\zeta_{K,L}. \tag{42}$$

Also, $\zeta_{K,L}^2 \geq \mathrm{tr}(Q_K) \geq cK$ and $\tau_{K,L} \geq C^{-1/2}$ imply that $\tau_{K,L}\zeta_{K,L}$ is bounded away from zero. For the remaining comparisons, fix a common $C \geq 1$ large enough for Equations (36), (40), (41), and (42), and such that $\tau_{K,L}\zeta_{K,L} \geq C^{-1}$.

If $C\tau_{K,L}\Delta_n \leq \min\{1, c_*/2\}$, then

$$r^2 - C\tau_{K,L}\Delta_n - q_{K,L} \geq c_*/2.$$

The population feasible sets at both perturbed radii are therefore nonempty. They are compact because $G_{K,L} \succ 0$. Moreover, $r^2 + C\tau_{K,L}\Delta_n \leq R^2 + 1$, so every coefficient in the three sets of Equation (39) satisfies Equation (29). If $J_{K,L}(\boldsymbol{\beta}) \leq r^2 - C\tau_{K,L}\Delta_n$, Equation (40) gives

$$\widehat{J}_{K,L}(\boldsymbol{\beta}) \leq J_{K,L}(\boldsymbol{\beta}) + C\tau_{K,L}\Delta_n \leq r^2.$$

If $\widehat{J}_{K,L}(\boldsymbol{\beta}) \leq r^2$, the same bound gives

$$J_{K,L}(\boldsymbol{\beta}) \leq \widehat{J}_{K,L}(\boldsymbol{\beta}) + C\tau_{K,L}\Delta_n \leq r^2 + C\tau_{K,L}\Delta_n.$$

This proves Equation (39), and hence Equations (38) and (37). Combining the population-objective comparison with Equation (36) also yields

$$\begin{aligned} &m_{K,L}(\pi, \sqrt{r^2 + C\tau_{K,L}\Delta_n}) - C\tau_{K,L}\Delta_n \\ &\quad\leq \widehat{m}_{K,L}(\pi, r) \\ &\quad\leq m_{K,L}(\pi, \sqrt{r^2 - C\tau_{K,L}\Delta_n}) + C\tau_{K,L}\Delta_n. \end{aligned} \tag{43}$$

It remains to bound the two radius changes in Equation (37). The closed-form population minimum is

$$m_{K,L}(\pi, r) = \bar{\boldsymbol{\beta}}_{K,L}^\top \mathbb{E}[\boldsymbol{h}_\pi] - \sqrt{r^2 - q_{K,L}}\, \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2.$$

Consequently,

$$\begin{aligned}0 &\le m_{K,L}(\pi, \sqrt{r^2 - C\tau_{K,L}\Delta_n}) - m_{K,L}(\pi, r)\\ &= \{\sqrt{r^2 - q_{K,L}} - \sqrt{r^2 - C\tau_{K,L}\Delta_n - q_{K,L}}\}\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2\\ &= \frac{C\tau_{K,L}\Delta_n \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2}{\sqrt{r^2 - q_{K,L}} + \sqrt{r^2 - C\tau_{K,L}\Delta_n - q_{K,L}}},\end{aligned}$$

and

$$\begin{aligned}0 &\le m_{K,L}(\pi, r) - m_{K,L}(\pi, \sqrt{r^2 + C\tau_{K,L}\Delta_n})\\ &= \{\sqrt{r^2 + C\tau_{K,L}\Delta_n - q_{K,L}} - \sqrt{r^2 - q_{K,L}}\}\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2\\ &= \frac{C\tau_{K,L}\Delta_n \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2}{\sqrt{r^2 + C\tau_{K,L}\Delta_n - q_{K,L}} + \sqrt{r^2 - q_{K,L}}}.\end{aligned}$$

Each denominator is at least $2\sqrt{c_*/2} = \sqrt{2c_*}$. Equation (42) therefore bounds each radius change by $C^2\tau_{K,L}^2\zeta_{K,L}\Delta_n/\sqrt{2c_*}$. Substituting into Equation (37) gives

$$\sup_{\pi\in\Pi} \mathrm{II}_\pi \le \frac{C^2}{\sqrt{2c_*}}\tau_{K,L}^2\zeta_{K,L}\Delta_n. \tag{44}$$

(c) Combining the two bounds.

Equations (26), (36), and (44) imply

$$\begin{aligned}\sup_{\pi\in\Pi} |\widehat{m}_{K,L}(\pi, r) - m_{K,L}(\pi, r)| &\le \sup_{\pi\in\Pi} \mathrm{I}_\pi + \sup_{\pi\in\Pi} \mathrm{II}_\pi\\ &\le C\tau_{K,L}\Delta_n + \frac{C^2}{\sqrt{2c_*}}\tau_{K,L}^2\zeta_{K,L}\Delta_n\\ &\le C^2\left(1 + \frac{1}{\sqrt{2c_*}}\right)\tau_{K,L}^2\zeta_{K,L}\Delta_n,\end{aligned}$$

where the last inequality uses $\tau_{K,L}\zeta_{K,L} \ge C^{-1}$. Choose

$$C_0 \ge \max\left\{C,\, C^2\left(1 + \frac{1}{\sqrt{2c_*}}\right)\right\}, \qquad C_0\epsilon_0 \le \min\{1, c_*/2\},$$

and retain the earlier restrictions on ϵ_0. Equation (41) proves the criterion-minimum bound. Since $m_{K,L}(\pi, \cdot)$ is nonincreasing, Equation (43) remains valid with C_0 in place of C. The preceding uniform bound proves the final assertion. □

Proof of Theorem 3. Note that

$$\begin{aligned}0 \le\ & \mathrm{Reg}_{\mathrm{mm}}\{\widehat{\pi}_{K,L}^\star(\rho); \rho\}\\ \le\ & \sup_{\pi\in\Pi}\left[\min_{f\in\mathcal{F}(\rho;\mathcal{S}_K)}\{V_f(\pi) - V_f(\pi_0)\} - \min_{f\in\mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L)}\{V_f(\pi) - V_f(\pi_0)\}\right]\\ &+ M_{K,L}(\rho) - \min_{f\in\mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L)}\{V_f(\widehat{\pi}_{K,L}^\star(\rho)) - V_f(\pi_0)\}.\end{aligned} \tag{45}$$

Step 1: Approximation error. By Equation (5), for $f \in \mathcal{F}(\rho; \mathcal{S}_K)$,

$$\begin{aligned}&\max_{g\in\mathcal{T}_L}\left\{2\mathbb{E}[g(\boldsymbol{X}, Z)\{Y - f(\boldsymbol{X}, P)\}] - \mathbb{E}[g(\boldsymbol{X}, Z)^2]\right\}\\ &= \|g_f^\star\|_2^2 - \inf_{g\in\mathcal{T}_L}\|g - g_f^\star\|_2^2\\ &\le \|g_f^\star\|_2^2\\ &= \mathbb{E}\left[\{\mathbb{E}[Y - f(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\}^2\right]\\ &\le \rho^2,\end{aligned} \tag{46}$$

where the last inequality follows from the definition of $\mathcal{F}(\rho;\mathcal{S}_K)$. It follows that

$$\mathcal{F}(\rho;\mathcal{S}_K) \subseteq \mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L). \tag{47}$$

When $\mathcal{F}(\rho;\mathcal{S}_K)$ is nonempty, as verified below, this inclusion implies

$$\min_{f\in\mathcal{F}(\rho;\mathcal{S}_K,\mathcal{T}_L)}\{V_f(\pi)-V_f(\pi_0)\} \le \min_{f\in\mathcal{F}(\rho;\mathcal{S}_K)}\{V_f(\pi)-V_f(\pi_0)\} \tag{48}$$

By the bounded price interval, Assumption 3(i), and the definition of $\zeta_{K,L}$ in Assumption 2(iii),

$$\begin{aligned}
&\sup_{\pi\in\Pi}\left\|\nabla_{\boldsymbol{\beta}}\{V_{f_{\boldsymbol{\beta}}}(\pi)-V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2 \\
&\quad=\sup_{\pi\in\Pi}\left\|\mathbb{E}\left[\pi(\boldsymbol{X})\boldsymbol{\phi}_K\{\boldsymbol{X},\pi(\boldsymbol{X})\}-\pi_0(\boldsymbol{X})\boldsymbol{\phi}_K\{\boldsymbol{X},\pi_0(\boldsymbol{X})\}\right]\right\|_2 \\
&\quad\le\sup_{\pi\in\Pi}\mathbb{E}\left[|\pi(\boldsymbol{X})|\|\boldsymbol{\phi}_K\{\boldsymbol{X},\pi(\boldsymbol{X})\}\|_2+|\pi_0(\boldsymbol{X})|\|\boldsymbol{\phi}_K\{\boldsymbol{X},\pi_0(\boldsymbol{X})\}\|_2\right] \\
&\quad\le C\sup_{(\boldsymbol{x},p)\in\mathcal{D}_f}\|\boldsymbol{\phi}_K(\boldsymbol{x},p)\|_2 \le C\zeta_{K,L}.
\end{aligned} \tag{49}$$

By Assumption 3(ii), $\rho^2-q_{K,L}\ge c_0$. Since $b_L\to 0$, for sufficiently large L,

$$\rho^2-b_L^2-q_{K,L}\ge c_0-b_L^2\ge c_0/2. \tag{50}$$

The projected feasible set at $\sqrt{\rho^2-b_L^2}$ is therefore nonempty. Its coefficient set is compact because $G_{K,L}\succ 0$.

Fix $\pi\in\Pi$, and let $\boldsymbol{\beta}_\pi^{\mathrm{wc}}$ denote a worst-case coefficient at the radius $\sqrt{\rho^2-b_L^2}$; that is,

$$f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}\in\operatorname*{arg\,min}_{f\in\mathcal{F}(\sqrt{\rho^2-b_L^2};\mathcal{S}_K,\mathcal{T}_L)}\{V_f(\pi)-V_f(\pi_0)\}.$$

The approximation condition in Assumption 2(i) is imposed at ρ. We first show that its bound also holds for the coefficient $\boldsymbol{\beta}_\pi^{\mathrm{wc}}$ just selected at $\sqrt{\rho^2-b_L^2}$. Since $\pi_0\in\Pi$ and its worst-case coefficient is selected to be $\bar{\boldsymbol{\beta}}_{K,L}$, the same assumption gives

$$\inf_{g\in\mathcal{T}_L}\|g-g^\star_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}\|_\infty\le b_L.$$

For a nonzero value gradient, Equation (9) expresses $\boldsymbol{\beta}_\pi^{\mathrm{wc}}$ as a convex combination of $\bar{\boldsymbol{\beta}}_{K,L}$ and the worst-case coefficient at ρ, with weight $\sqrt{(\rho^2-b_L^2-q_{K,L})/(\rho^2-q_{K,L})}$ on the latter. For a zero value gradient, select $\bar{\boldsymbol{\beta}}_{K,L}$ at both radii, so the same relation holds. The conditional mean $g^\star_{f_{\boldsymbol{\beta}}}$ is affine in $\boldsymbol{\beta}$. Since $\mathcal{T}_L$ is linear, the triangle inequality and the approximation bound at ρ imply

$$\begin{aligned}
&\inf_{g\in\mathcal{T}_L}\|g-g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\|_\infty \\
&\quad\le\left(1-\sqrt{\frac{\rho^2-b_L^2-q_{K,L}}{\rho^2-q_{K,L}}}\right)\inf_{g\in\mathcal{T}_L}\|g-g^\star_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}\|_\infty+\sqrt{\frac{\rho^2-b_L^2-q_{K,L}}{\rho^2-q_{K,L}}}\,b_L \\
&\quad\le\left(1-\sqrt{\frac{\rho^2-b_L^2-q_{K,L}}{\rho^2-q_{K,L}}}\right)b_L+\sqrt{\frac{\rho^2-b_L^2-q_{K,L}}{\rho^2-q_{K,L}}}\,b_L=b_L.
\end{aligned} \tag{51}$$

In particular,

$$\inf_{g\in\mathcal{T}_L}\|g-g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\|_2^2\le\left(\inf_{g\in\mathcal{T}_L}\|g-g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\|_\infty\right)^2\le b_L^2.$$

By the choice of $\boldsymbol{\beta}_\pi^{\mathrm{wc}}$ at $\sqrt{\rho^2-b_L^2}$, we have

$$\max_{g\in\mathcal{T}_L}\left\{2\mathbb{E}[g(\boldsymbol{X},Z)(Y-f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}})]-\mathbb{E}[g(\boldsymbol{X},Z)^2]\right\}\le\rho^2-b_L^2.$$

Applying Equation (5) to $f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}$ gives

$$\begin{aligned}
\left\|g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\right\|_2^2&=\max_{g\in\mathcal{T}_L}\left\{2\mathbb{E}[g(\boldsymbol{X},Z)(Y-f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}})]-\mathbb{E}[g(\boldsymbol{X},Z)^2]\right\}+\inf_{g\in\mathcal{T}_L}\left\|g-g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\right\|_2^2 \\
&\le\rho^2-b_L^2+\inf_{g\in\mathcal{T}_L}\left\|g-g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\right\|_2^2.
\end{aligned}$$

Using Equation (51) and $\|\cdot\|_2 \le \|\cdot\|_\infty$, we obtain

$$\left\|g^\star_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}}\right\|_2^2 \le \rho^2 - b_L^2 + b_L^2 = \rho^2.$$

By the definition of $g^\star_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}}$, $f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi} \in \mathcal{F}(\rho; \mathcal{S}_K)$. This proves that $\mathcal{F}(\rho; \mathcal{S}_K)$ is nonempty. Its coefficient set is a closed subset of the compact projected feasible set, so the minimum in Equation (48) is attained. Since $\boldsymbol{\beta}^{\mathrm{wc}}_\pi$ is chosen to attain the double-sieve minimum at radius $\sqrt{\rho^2 - b_L^2}$, we have

$$V_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}}(\pi) - V_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}}(\pi_0) = \min_{f \in \mathcal{F}(\sqrt{\rho^2 - b_L^2}; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\}.$$

Therefore,

$$\min_{f \in \mathcal{F}(\rho; \mathcal{S}_K)} \{V_f(\pi) - V_f(\pi_0)\} \le V_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}}(\pi) - V_{f_{\boldsymbol{\beta}^{\mathrm{wc}}_\pi}}(\pi_0) = \min_{f \in \mathcal{F}(\sqrt{\rho^2 - b_L^2}; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\}. \tag{52}$$

By Equation (8), we have

$$\min_{f \in \mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\} = V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi_0) - \sqrt{\rho^2 - q_{K,L}} \left\|G_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2,$$

and

$$\min_{f \in \mathcal{F}(\sqrt{\rho^2 - b_L^2}; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\} = V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}(\pi_0) - \sqrt{\rho^2 - b_L^2 - q_{K,L}} \left\|G_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2.$$

By Equations (48) and (52), we have

$$\begin{aligned}
0 &\le \min_{f \in \mathcal{F}(\sqrt{\rho^2 - b_L^2}; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\} - \min_{f \in \mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\} \\
&= \left\{\sqrt{\rho^2 - q_{K,L}} - \sqrt{\rho^2 - b_L^2 - q_{K,L}}\right\} \left\|G_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2 \\
&= \frac{b_L^2}{\sqrt{\rho^2 - q_{K,L}} + \sqrt{\rho^2 - b_L^2 - q_{K,L}}} \left\|G_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2.
\end{aligned}$$

By the definition of $\tau_{K,L}$ and Equation (49),

$$\begin{aligned}
&\sup_{\pi \in \Pi} \left\|G_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2 \\
&\quad \le \|G_{K,L}^{-1/2}\|_{\mathrm{op}} \sup_{\pi \in \Pi} \left\|\nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2 \\
&\quad \le C \tau_{K,L} \zeta_{K,L}.
\end{aligned} \tag{53}$$

Combining Equations (53) and (50),

$$\frac{b_L^2}{\sqrt{\rho^2 - q_{K,L}} + \sqrt{\rho^2 - b_L^2 - q_{K,L}}} \sup_{\pi \in \Pi} \left\|G_{K,L}^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\}\right\|_2 \le C \tau_{K,L} \zeta_{K,L} b_L^2.$$

It follows that

$$0 \le \sup_{\pi \in \Pi} \left[\min_{f \in \mathcal{F}(\rho; \mathcal{S}_K)} \{V_f(\pi) - V_f(\pi_0)\} - \min_{f \in \mathcal{F}(\rho; \mathcal{S}_K, \mathcal{T}_L)} \{V_f(\pi) - V_f(\pi_0)\}\right] \le C \tau_{K,L} \zeta_{K,L} b_L^2. \tag{54}$$

Step 2: Statistical error. Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm, i.e., $\|A\|_{\mathrm{op}} = \sup_{\substack{\boldsymbol{u} \in \mathbb{R}^k \\ \|\boldsymbol{u}\|_2 \le 1}} \|A\boldsymbol{u}\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$. Since $G_{K,L}$ is positive definite, the cross-moment matrix $\mathbb{E}[\boldsymbol{\psi}_L \boldsymbol{\phi}_K^\top]$ has full column rank, which implies $K \le L$. Furthermore, the trace bound $cL \le \operatorname{tr} \mathbb{E}[\boldsymbol{\psi}_L \boldsymbol{\psi}_L^\top] = E\|\boldsymbol{\psi}_L\|_2^2 \le \zeta_{K,L}^2$, combined with Assumption 2(iii), ensures that $K + L = o(n)$ and $\zeta_{K,L}\sqrt{\log n / n} = o(1)$.

Step 2.1: We aim to bound $\left\|(E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\psi}_L^\top]\right\|_{\text{op}} + \left\|(E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\phi}_K^\top]\right\|_{\text{op}} + \|(E_n - E)[\boldsymbol{\psi}_L Y]\|_2$.

Note that

$$\begin{aligned}\left\|\mathbb{E}\left[\left\{\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top - \mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\right\}^2\right]\right\|_{\text{op}} &= \left\|\mathbb{E}[\|\boldsymbol{\psi}_L\|_2^2\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top] - \{\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\}^2\right\|_{\text{op}} \\ &\le \left\|\mathbb{E}[\|\boldsymbol{\psi}_L\|_2^2\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\right\|_{\text{op}} \\ &\le \zeta_{K,L}^2 \left\|\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\right\|_{\text{op}} \\ &\le C\zeta_{K,L}^2,\end{aligned}$$

where the last inequality follows from $\lambda_{\max}\left(\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\right) \le C$.

Following the same arguments, we have

$$\begin{aligned}&\left\|\mathbb{E}\left[\left\{\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top - \mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]\right\}\left\{\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top - \mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\right\}\right]\right\|_{\text{op}} \\ \le &\left\|\mathbb{E}\left[\|\boldsymbol{\phi}_K\|_2^2\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top\right]\right\|_{\text{op}} \\ \le &\zeta_{K,L}^2\left\|\mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\right\|_{\text{op}} \le C\zeta_{K,L}^2,\end{aligned}$$

and

$$\begin{aligned}&\left\|\mathbb{E}\left[\left\{\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top - \mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\right\}\left\{\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top - \mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]\right\}\right]\right\|_{\text{op}} \\ \le &\left\|\mathbb{E}\left[\|\boldsymbol{\psi}_L\|_2^2\boldsymbol{\phi}_K\boldsymbol{\phi}_K^\top\right]\right\|_{\text{op}} \\ \le &\zeta_{K,L}^2\left\|\mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\phi}_K^\top]\right\|_{\text{op}} \le C\zeta_{K,L}^2.\end{aligned}$$

Thus, we have

$$\sigma_\psi^2 := \left\|\sum_{i=1}^n \mathbb{E}\left[\left\{\frac{\boldsymbol{\psi}_{L,i}\boldsymbol{\psi}_{L,i}^\top - \mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]}{n}\right\}^2\right]\right\|_{\text{op}} \le \frac{C\zeta_{K,L}^2}{n}$$

and

$$\begin{aligned}\sigma_{\psi\phi}^2 := \max\Bigg\{&\left\|\sum_{i=1}^n E\Big[\frac{\{\boldsymbol{\psi}_{L,i}\boldsymbol{\phi}_{K,i}^\top - \mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]\}\{\boldsymbol{\phi}_{K,i}\boldsymbol{\psi}_{L,i}^\top - \mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\}}{n^2}\Big]\right\|_{\text{op}}, \\ &\left\|\sum_{i=1}^n E\Big[\frac{\{\boldsymbol{\phi}_{K,i}\boldsymbol{\psi}_{L,i}^\top - \mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\psi}_L^\top]\}\{\boldsymbol{\psi}_{L,i}\boldsymbol{\phi}_{K,i}^\top - \mathbb{E}[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]\}}{n^2}\Big]\right\|_{\text{op}}\Bigg\} \le \frac{C\zeta_{K,L}^2}{n}.\end{aligned}$$

Let E_n denote the empirical average. Applying the matrix Bernstein inequality (Tropp, 2012, Theorems 1.4 and 1.6), we obtain

$$\begin{aligned}&\left\|(E_n - E)[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\right\|_{\text{op}} + \left\|(E_n - E)[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]\right\|_{\text{op}} \\ &= O_p\left(\zeta_{K,L}\sqrt{\frac{\log n}{n}} + \frac{\zeta_{K,L}^2\log n}{n}\right).\end{aligned}$$

By Assumption 2(iii),

$$\zeta_{K,L}\sqrt{\frac{\log n}{n}} = o(1),$$

and therefore

$$\frac{\zeta_{K,L}^2\log n}{n} = o\left(\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right).$$

Hence,

$$\left\|(E_n - E)[\boldsymbol{\psi}_L\boldsymbol{\psi}_L^\top]\right\|_{\text{op}} + \left\|(E_n - E)[\boldsymbol{\psi}_L\boldsymbol{\phi}_K^\top]\right\|_{\text{op}} = O_p\left(\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right).$$

For the response moment, independence across observations and a standard second-moment calculation give $E\|(E_n - E)[\boldsymbol{\psi}_L Y]\|_2^2 \leq C\zeta_{K,L}^2/n$. By Markov's inequality, this translates to $\|(E_n - E)[\boldsymbol{\psi}_L Y]\|_2 = O_p(\zeta_{K,L} n^{-1/2})$.

Combining these empirical bounds, we conclude that

$$\left\|(E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\psi}_L^\top]\right\|_{\text{op}} + \left\|(E_n - E)[\boldsymbol{\psi}_L \boldsymbol{\phi}_K^\top]\right\|_{\text{op}} + \|(E_n - E)[\boldsymbol{\psi}_L Y]\|_2 = O_p\left(\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right). \tag{55}$$

Step 2.2: We next bound $\sup_{\pi\in\Pi}\left\|\nabla_{\boldsymbol{\beta}}\left\{\widehat{V}_{f_{\boldsymbol{\beta}}}(\pi) - \widehat{V}_{f_{\boldsymbol{\beta}}}(\pi_0) - V_{f_{\boldsymbol{\beta}}}(\pi) + V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}\right\|_2$.

Define

$$\boldsymbol{h}_\pi(\boldsymbol{X}) = \pi(\boldsymbol{X})\boldsymbol{\phi}_K\{\boldsymbol{X}, \pi(\boldsymbol{X})\} - \pi_0(\boldsymbol{X})\boldsymbol{\phi}_K\{\boldsymbol{X}, \pi_0(\boldsymbol{X})\}.$$

We have

$$\sup_{\pi\in\Pi}\left\|\nabla_{\boldsymbol{\beta}}\left\{\widehat{V}_{f_{\boldsymbol{\beta}}}(\pi) - \widehat{V}_{f_{\boldsymbol{\beta}}}(\pi_0) - V_{f_{\boldsymbol{\beta}}}(\pi) + V_{f_{\boldsymbol{\beta}}}(\pi_0)\right\}\right\|_2 = \sup_{\pi\in\Pi}\|(E_n - E)\boldsymbol{h}_\pi\|_2.$$

By Assumption 2(iii),

$$\sup_{\pi\in\Pi}\|\boldsymbol{h}_\pi(\boldsymbol{X})\|_2 \leq C\zeta_{K,L}$$

almost surely. For every fixed π, we have

$$\begin{aligned} E\|(E_n - E)\boldsymbol{h}_\pi\|_2 &\leq \left\{E\|(E_n - E)\boldsymbol{h}_\pi\|_2^2\right\}^{1/2} \\ &= \frac{1}{\sqrt{n}}\left\{E\|\boldsymbol{h}_\pi - E\boldsymbol{h}_\pi\|_2^2\right\}^{1/2} \leq \frac{C\zeta_{K,L}}{\sqrt{n}}. \end{aligned}$$

For a fixed policy $\pi \in \Pi$, define

$$F_\pi(X_1, \ldots, X_n) := \|(E_n - E)\boldsymbol{h}_\pi\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n \{\boldsymbol{h}_\pi(\boldsymbol{X}_i) - \mathbb{E}[\boldsymbol{h}_\pi(\boldsymbol{X})]\}\right\|_2.$$

Suppose that the ith observation $\boldsymbol{X}_i$ is replaced by an independent copy $\boldsymbol{X}_i'$, while all other observations are kept fixed. By the reverse triangle inequality,

$$\begin{aligned} &|F_\pi(X_1, \ldots, X_i, \ldots, \boldsymbol{X}_n) - F_\pi(X_1, \ldots, X_i', \ldots, \boldsymbol{X}_n)| \\ &\leq \frac{1}{n}\|\boldsymbol{h}_\pi(\boldsymbol{X}_i) - \boldsymbol{h}_\pi(\boldsymbol{X}_i')\|_2 \\ &\leq \frac{1}{n}\{\|\boldsymbol{h}_\pi(\boldsymbol{X}_i)\|_2 + \|\boldsymbol{h}_\pi(\boldsymbol{X}_i')\|_2\} \\ &\leq \frac{C\zeta_{K,L}}{n}. \end{aligned}$$

Therefore, by McDiarmid's inequality, for every $t > 0$,

$$P\{F_\pi - \mathbb{E}[F_\pi] > t\} \leq \exp\left(-\frac{Cnt^2}{\zeta_{K,L}^2}\right).$$

Since $\mathbb{E}[F_\pi] = \mathbb{E}\left[\|(E_n - E)\boldsymbol{h}_\pi\|_2\right] \leq \frac{C\zeta_{K,L}}{\sqrt{n}}$, we have

$$P\left\{\|(E_n - E)\boldsymbol{h}_\pi\|_2 > \frac{C\zeta_{K,L}}{\sqrt{n}} + t\right\} \leq \exp\left(-\frac{Cnt^2}{\zeta_{K,L}^2}\right).$$

By Assumption 3(i), the Lipschitz constant grows at a rate of K^a for some fixed $a \geq 0$. We set the covering radius to $\varepsilon_n = n^{-(a+2)}$. Since $K \leq n$ for sufficiently large n, we have $K^a\varepsilon_n \leq n^{-2}$.

Because the parameter space $\Theta \subset \mathbb{R}^d$ is compact with a fixed dimension d, there exists an ε_n-cover $\{\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_{N_n}\} \subset \Theta$ such that $\Theta \subset \bigcup_{j=1}^{N_n} B(\boldsymbol{\vartheta}_j, \varepsilon_n)$. This covering number satisfies $N_n \leq C\varepsilon_n^{-d} = O(n^{d(a+2)})$.

Hence, for every $\boldsymbol{\vartheta} \in \Theta$, there exists an index $j(\boldsymbol{\vartheta}) \in \{1, \ldots, N_n\}$ such that

$$\left\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}\right\|_2 \leq \varepsilon_n.$$

For each covering point $\boldsymbol{\vartheta}_j$, we have, for every $t > 0$,

$$P\left\{\left\|(E_n - E)\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_j}}\right\|_2 > \frac{C\zeta_{K,L}}{\sqrt{n}} + t\right\} \leq \exp\left(-\frac{Cnt^2}{\zeta_{K,L}^2}\right).$$

Taking $t = A\zeta_{K,L}\sqrt{\frac{\log n}{n}}$, for a sufficiently large constant $A > 0$, yields

$$P\left\{\left\|(E_n - E)\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_j}}\right\|_2 > C\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right\} \leq n^{-CA^2}.$$

Thus, we have

$$\begin{aligned}
&P\left\{\max_{1\leq j\leq N_n}\left\|(E_n - E)\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_j}}\right\|_2 > C\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right\} \\
&\leq N_n n^{-CA^2} \\
&\leq Cn^{d(a+2)-CA^2}.
\end{aligned}$$

By choosing A sufficiently large such that $CA^2 > d(a+2) + 2$, we obtain

$$P\left\{\max_{1\leq j\leq N_n}\left\|(E_n - E)\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_j}}\right\|_2 > C\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right\} \leq Cn^{-2}. \tag{56}$$

We next extend this bound to an arbitrary $\boldsymbol{\vartheta} \in \Theta$. Let $\boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}$ be a covering point satisfying $\left\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}\right\|_2 \leq \varepsilon_n$. Then

$$\begin{aligned}
&\|(E_n - E)\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}}\|_2 \\
&\quad\leq \left\|(E_n - E)\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}}}\right\|_2 \\
&\qquad + \left\|(E_n - E)\left(\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}} - \boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}}}\right)\right\|_2.
\end{aligned}$$

By the triangle inequality,

$$\left\|(E_n - E)\left(\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}} - \boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}}}\right)\right\|_2 \leq E_n\left\|\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}} - \boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}}}\right\|_2 + E\left\|\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}} - \boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}}}\right\|_2.$$

Assumption 3(i) implies

$$\left\|\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}} - \boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}'}}\right\|_2 \leq CK^a\left\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\right\|_2.$$

By Assumptions 2 (ii) and (iii),

$$\zeta_{K,L}^2 \geq E\|\boldsymbol{\phi}_K\|_2^2 = \operatorname{tr}\mathbb{E}[\boldsymbol{\phi}_K\boldsymbol{\phi}_K^\top] \geq cK.$$

Hence,

$$\frac{K\log n}{n} \leq \frac{1}{c}\frac{\zeta_{K,L}^2\log n}{n} \to 0,$$

so that $K \leq n$ for all sufficiently large n. Therefore,

$$K^a n^{-(a+2)} \leq n^a n^{-(a+2)} = n^{-2}.$$

Therefore,

$$\left\|(E_n - E)\left(\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}} - \boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}_{j(\boldsymbol{\vartheta})}}}\right)\right\|_2 \leq 2CK^a\varepsilon_n \leq 2Cn^{-2}.$$

By Equation (56), we obtain, with probability at least $1 - Cn^{-2}$,

$$\sup_{\boldsymbol{\vartheta}\in\Theta} \|(E_n - E)\boldsymbol{h}_{\pi_{\boldsymbol{\vartheta}}}\|_2 \le C\zeta_{K,L}\sqrt{\frac{\log n}{n}} + Cn^{-2}.$$

It follows that

$$\sup_{\pi\in\Pi} \|(E_n - E)\boldsymbol{h}_{\pi}\|_2 = O_p\left(\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right). \tag{57}$$

Step 2.3: Statistical error and policy optimization. Let $m_{K,L}$, $\widehat{m}_{K,L}$, and Δ_n be defined as in Lemma 1. For radii at which the respective feasible coefficient sets are nonempty and compact, write

$$M_{K,L}(r) = \max_{\pi\in\Pi} m_{K,L}(\pi, r), \qquad \widehat{M}_{K,L}(r) = \max_{\pi\in\Pi} \widehat{m}_{K,L}(\pi, r).$$

The Lipschitz condition in Assumption 3(i) and boundedness of the feasible coefficient sets imply that both value functions are continuous in $\boldsymbol{\vartheta}$. Compactness of Θ therefore ensures that these maxima are attained. If the sample feasible coefficient set at ρ is empty or not compact, set $\widehat{\pi}^{\star}_{K,L}(\rho) = \pi_0$.

By Equations (45) and (54),

$$\begin{aligned} 0 \le\ & \mathrm{Reg}_{\mathrm{mm}}\{\widehat{\pi}^{\star}_{K,L}(\rho); \rho\} \\ & \le C\tau_{K,L}\zeta_{K,L}b_L^2 + M_{K,L}(\rho) - m_{K,L}\{\widehat{\pi}^{\star}_{K,L}(\rho), \rho\}. \end{aligned} \tag{58}$$

It remains to bound $M_{K,L}(\rho) - m_{K,L}\{\widehat{\pi}^{\star}_{K,L}(\rho), \rho\}$.

Let $\omega_n = \zeta_{K,L}\sqrt{\log n/n}$. Equations (55) and (57), together with Assumption 2(iii), imply

$$\Delta_n = O_p(\omega_n), \qquad \tau_{K,L}\Delta_n = O_p(\tau_{K,L}\omega_n) = o_p(1). \tag{59}$$

Since ρ is fixed and $\rho^2 - q_{K,L} \ge c_0$ by Assumption 3(ii), apply Lemma 1 with $r = R = \rho$ and $c_* = c_0$. Whenever $\tau_{K,L}\Delta_n \le \epsilon_0$, the lemma ensures that the sample feasible coefficient set is nonempty and compact, and Equation (25) gives

$$\sup_{\pi\in\Pi} |\widehat{m}_{K,L}(\pi, \rho) - m_{K,L}(\pi, \rho)| \le C\tau_{K,L}^2\zeta_{K,L}\Delta_n. \tag{60}$$

Under the same condition, the sample optimizer satisfies

$$\widehat{M}_{K,L}(\rho) = \widehat{m}_{K,L}\{\widehat{\pi}^{\star}_{K,L}(\rho), \rho\}.$$

For every $\pi \in \Pi$,

$$\begin{aligned} m_{K,L}(\pi, \rho) &\le \widehat{m}_{K,L}(\pi, \rho) + \sup_{\pi'\in\Pi} |\widehat{m}_{K,L}(\pi', \rho) - m_{K,L}(\pi', \rho)|, \\ \widehat{m}_{K,L}(\pi, \rho) &\le m_{K,L}(\pi, \rho) + \sup_{\pi'\in\Pi} |\widehat{m}_{K,L}(\pi', \rho) - m_{K,L}(\pi', \rho)|. \end{aligned}$$

Taking the maximum over π in both inequalities gives

$$|M_{K,L}(\rho) - \widehat{M}_{K,L}(\rho)| \le \sup_{\pi\in\Pi} |\widehat{m}_{K,L}(\pi, \rho) - m_{K,L}(\pi, \rho)|. \tag{61}$$

Consequently,

$$\begin{aligned} 0 &\le M_{K,L}(\rho) - m_{K,L}\{\widehat{\pi}^{\star}_{K,L}(\rho), \rho\} \\ &= M_{K,L}(\rho) - \widehat{M}_{K,L}(\rho) + \widehat{m}_{K,L}\{\widehat{\pi}^{\star}_{K,L}(\rho), \rho\} - m_{K,L}\{\widehat{\pi}^{\star}_{K,L}(\rho), \rho\} \\ &\le |M_{K,L}(\rho) - \widehat{M}_{K,L}(\rho)| + \sup_{\pi\in\Pi} |\widehat{m}_{K,L}(\pi, \rho) - m_{K,L}(\pi, \rho)| \\ &\le 2\sup_{\pi\in\Pi} |\widehat{m}_{K,L}(\pi, \rho) - m_{K,L}(\pi, \rho)| \\ &\le C\tau_{K,L}^2\zeta_{K,L}\Delta_n, \end{aligned} \tag{62}$$

where the last inequality uses Equation (60).

Equation (59) gives

$$P\{\tau_{K,L}\Delta_n > \epsilon_0\} \longrightarrow 0.$$

Thus the preceding optimization-error bound holds with probability tending to one. Since $\Delta_n = O_p(\omega_n)$,

$$\begin{aligned} M_{K,L}(\rho) - m_{K,L}\{\widehat{\pi}^\star_{K,L}(\rho), \rho\} &= O_p\big(\tau^2_{K,L}\zeta_{K,L}\omega_n\big) \\ &= O_p\left(\tau^2_{K,L}\zeta^2_{K,L}\sqrt{\frac{\log n}{n}}\right). \end{aligned} \tag{63}$$

Step 2.4: Conclusion. Combining Equations (58) and (63), we obtain

$$\begin{aligned} 0 \le\ & \mathrm{Reg}_{\mathrm{mm}}\{\widehat{\pi}^\star_{K,L}(\rho); \rho\} \\ &\le C\tau_{K,L}\zeta_{K,L}b_L^2 + M_{K,L}(\rho) - m_{K,L}\{\widehat{\pi}^\star_{K,L}(\rho), \rho\} \\ &= O_p\left(\tau^2_{K,L}\zeta^2_{K,L}\sqrt{\frac{\log n}{n}} + \tau_{K,L}\zeta_{K,L}b_L^2\right) \\ &= O_p\{\mathcal{E}(n,K,L)\}. \end{aligned}$$

□

B.8 Proof of Theorem 4

We first establish three lemmas for the population radius, the estimated radius, and the true policy value. Use $m_{K,L}$, $\widehat{m}_{K,L}$, and Δ_n from Lemma 1, and recall

$$M_K(r) = \max_{\pi\in\Pi}\min_{f\in\mathcal{F}(r;\mathcal{S}_K)}\{V_f(\pi) - V_f(\pi_0)\},$$

$$M_{K,L}(r) = \max_{\pi\in\Pi} m_{K,L}(\pi, r), \qquad \widehat{M}_{K,L}(r) = \max_{\pi\in\Pi}\widehat{m}_{K,L}(\pi, r).$$

Lemma 2 (Approximation of the population gain radius)**.** *Fix $\delta > 0$ such that the population gain-constrained problem has a maximizer $\rho_{\mathrm{gain,K}}(\delta)$. Suppose Assumptions 2 and 3 hold at this radius, with $\boldsymbol{\beta}^{\mathrm{wc}}_\pi = \bar{\boldsymbol{\beta}}_{K,L}$ when $\mathbb{E}[\boldsymbol{h}_\pi] = 0$. Then, for sufficiently large L,*

$$\sqrt{c_0} \le \rho_{\mathrm{gain,K}}(\delta) \le \sqrt{\mathbb{E}[Y^2]} \le C,$$

$$\sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2} \le \rho_{\mathrm{gain},K,L}(\delta) \le \rho_{\mathrm{gain,K}}(\delta),$$

$$M_{K,L}\{\rho_{\mathrm{gain},K,L}(\delta)\} = \delta, \qquad 0 \le \rho_{\mathrm{gain,K}}(\delta) - \rho_{\mathrm{gain},K,L}(\delta) \le Cb_L^2,$$

where C is independent of n, K, L.

Proof.

Step 1: Approximation at the contracted radius. By Assumption 3(ii),

$$\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2 - q_{K,L} \ge c_0 - b_L^2 \ge c_0/2$$

for sufficiently large L. Since $G_{K,L} \succ 0$, the projected feasible coefficient set at $\sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2}$ is nonempty and compact. For each $\pi \in \Pi$, let $\boldsymbol{\beta}^{\mathrm{wc}}_\pi$ denote a worst-case coefficient at this contracted radius:

$$\boldsymbol{\beta}^{\mathrm{wc}}_\pi \in \operatorname*{arg\,min}_{\boldsymbol{\beta}: J_{K,L}(\boldsymbol{\beta}) \le \rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2} \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{h}_\pi].$$

The approximation condition in Assumption 2(i) is imposed at $\rho_{\mathrm{gain,K}}(\delta)$. We first verify that the same bound holds for the coefficients selected above. Since $\pi_0 \in \Pi$ and its worst-case coefficient is selected to be $\bar{\boldsymbol{\beta}}_{K,L}$, the assumption gives

$$\inf_{g\in\mathcal{T}_L} \|g - g^\star_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}\|_\infty \le b_L.$$

When $\mathbb{E}[\boldsymbol{h}_\pi] \neq 0$, Equation (9) expresses $\boldsymbol{\beta}_\pi^{\mathrm{wc}}$ as a convex combination of $\bar{\boldsymbol{\beta}}_{K,L}$ and the worst-case coefficient at $\rho_{\mathrm{gain,K}}(\delta)$, with weight

$$\sqrt{\frac{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2 - q_{K,L}}{\rho_{\mathrm{gain,K}}(\delta)^2 - q_{K,L}}} \in [0,1]$$

on the latter. When $\mathbb{E}[\boldsymbol{h}_\pi] = 0$, select $\bar{\boldsymbol{\beta}}_{K,L}$ at both radii, so the same relation holds. The conditional mean $g^\star_{f_{\boldsymbol{\beta}}}$ is affine in $\boldsymbol{\beta}$. Since $\mathcal{T}_L$ is linear, the triangle inequality and the approximation bound at $\rho_{\mathrm{gain,K}}(\delta)$ imply

$$\begin{aligned}
&\inf_{g\in\mathcal{T}_L} \|g - g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\|_\infty \\
&\quad\le \left(1 - \sqrt{\frac{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2 - q_{K,L}}{\rho_{\mathrm{gain,K}}(\delta)^2 - q_{K,L}}}\right) \inf_{g\in\mathcal{T}_L} \|g - g^\star_{f_{\bar{\boldsymbol{\beta}}_{K,L}}}\|_\infty \\
&\qquad + \sqrt{\frac{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2 - q_{K,L}}{\rho_{\mathrm{gain,K}}(\delta)^2 - q_{K,L}}}\, b_L \le b_L.
\end{aligned}$$

Consequently,

$$\sup_{\pi\in\Pi} \inf_{g\in\mathcal{T}_L} \|g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}} - g\|_\infty \le b_L. \tag{64}$$

Step 2: Comparison of the population value functions. By Equations (5) and (64),

$$\begin{aligned}
\|g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\|_2^2 &= J_{K,L}(\boldsymbol{\beta}_\pi^{\mathrm{wc}}) + \inf_{g\in\mathcal{T}_L} \|g - g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\|_2^2 \\
&\le \rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2 + \left(\inf_{g\in\mathcal{T}_L} \|g - g^\star_{f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}}}\|_\infty\right)^2 \\
&\le \rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2 + b_L^2 = \rho_{\mathrm{gain,K}}(\delta)^2.
\end{aligned}$$

Thus $f_{\boldsymbol{\beta}_\pi^{\mathrm{wc}}} \in \mathcal{F}(\rho_{\mathrm{gain,K}}(\delta); \mathcal{S}_K)$. In particular, the full conditional-moment feasible set at $\rho_{\mathrm{gain,K}}(\delta)$ is nonempty. Its coefficient set is closed and contained in the compact projected feasible coefficient set, so its minimum is attained for every $\pi \in \Pi$. Equation (47) and the choice of $\boldsymbol{\beta}_\pi^{\mathrm{wc}}$ give

$$\begin{aligned}
m_{K,L}(\pi, \rho_{\mathrm{gain,K}}(\delta)) &\le \min_{f\in\mathcal{F}(\rho_{\mathrm{gain,K}}(\delta);\mathcal{S}_K)} \{V_f(\pi) - V_f(\pi_0)\} \\
&\le \boldsymbol{\beta}_\pi^{\mathrm{wc}\top} \mathbb{E}[\boldsymbol{h}_\pi] \\
&= m_{K,L}\left(\pi, \sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2}\right).
\end{aligned}$$

The policy Lipschitz condition in Assumption 3(i) and boundedness of each feasible coefficient set imply continuity of these value functions in $\boldsymbol{\vartheta}$. Compactness of Θ ensures that the policy maxima are attained. Taking maxima in the preceding inequalities yields

$$\begin{aligned}
M_{K,L}(\rho_{\mathrm{gain,K}}(\delta)) &\le M_K(\rho_{\mathrm{gain,K}}(\delta)) \\
&\le M_{K,L}\left(\sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2}\right).
\end{aligned} \tag{65}$$

Step 3: Compare the gain radii. Choose a policy $\pi \in \Pi$ attaining $M_{K,L}(\sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2})$. By the definition of $\rho_{\mathrm{gain,K}}(\delta)$ and Equation (65),

$$\delta \le M_K(\rho_{\mathrm{gain,K}}(\delta)) \le m_{K,L}\left(\pi, \sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2}\right).$$

Since $\delta > 0$, Equation (8) implies

$$\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2 > 0;$$

otherwise $\mathbb{E}[\boldsymbol{h}_\pi] = 0$ and the preceding policy value would equal zero. Applying Equation (8) at the two radii gives

$$\begin{aligned} M_{K,L}(\sqrt{q_{K,L}}) &\geq m_{K,L}(\pi, \sqrt{q_{K,L}}) \\ &= m_{K,L}\left(\pi, \sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2}\right) \\ &\quad + \sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2 - q_{K,L}}\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2 \\ &> \delta. \end{aligned}$$

Theorem 2 therefore identifies $\rho_{\mathrm{gain},K,L}(\delta)$ as the unique solution of $M_{K,L}(r) = \delta$. Since $M_{K,L}$ is nonincreasing,

$$M_{K,L}\{\rho_{\mathrm{gain},K,L}(\delta)\} = \delta, \qquad \sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2} \leq \rho_{\mathrm{gain},K,L}(\delta).$$

For the upper bound, the full conditional-moment feasible sets are nonempty at every radius at least $\rho_{\mathrm{gain,K}}(\delta)$ by nesting. Equation (47) implies $M_{K,L}(r) \leq M_K(r)$ at these radii. If $\rho_{\mathrm{gain},K,L}(\delta) > \rho_{\mathrm{gain,K}}(\delta)$, the definition of the latter as the largest feasible gain radius would give

$$\delta = M_{K,L}\{\rho_{\mathrm{gain},K,L}(\delta)\} \leq M_K\{\rho_{\mathrm{gain},K,L}(\delta)\} < \delta,$$

a contradiction. Therefore,

$$\sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2} \leq \rho_{\mathrm{gain},K,L}(\delta) \leq \rho_{\mathrm{gain,K}}(\delta).$$

Since $q_{K,L} \geq 0$, the radius gap gives $\rho_{\mathrm{gain,K}}(\delta) \geq \sqrt{c_0}$. Also,

$$\|\mathbb{E}[Y \mid \boldsymbol{X}, Z]\|_2^2 \leq \mathbb{E}[Y^2], \qquad r \geq \sqrt{\mathbb{E}[Y^2]} \implies 0 \in \mathcal{F}(r; \mathcal{S}_K) \implies M_K(r) \leq 0 < \delta.$$

It follows that $\rho_{\mathrm{gain,K}}(\delta) \leq \sqrt{\mathbb{E}[Y^2]} \leq C$. Finally,

$$\begin{aligned} 0 &\leq \rho_{\mathrm{gain,K}}(\delta) - \rho_{\mathrm{gain},K,L}(\delta) \\ &\leq \rho_{\mathrm{gain,K}}(\delta) - \sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2} \\ &= \frac{b_L^2}{\rho_{\mathrm{gain,K}}(\delta) + \sqrt{\rho_{\mathrm{gain,K}}(\delta)^2 - b_L^2}} \leq C b_L^2. \end{aligned} \tag{66}$$

□

Lemma 3 (Statistical error of the gain radius)**.** *Under the conditions of Lemma 2, there exist constants $\epsilon_0, C > 0$, independent of n, K, L, such that, for sufficiently large L, whenever $\tau_{K,L}\Delta_n \leq \epsilon_0$, the equation $\widehat{M}_{K,L}(r) = \delta$ has a unique solution $\widehat{\rho}_{\mathrm{gain},K,L}(\delta)$, equal to the largest radius supporting δ, and*

$$|\widehat{\rho}_{\mathrm{gain},K,L}(\delta) - \rho_{\mathrm{gain},K,L}(\delta)| \leq C\tau_{K,L}\Delta_n. \tag{67}$$

Proof.

Step 1: Compare the population and sample values near the target radius. Lemma 2 gives

$$\sqrt{c_0} \leq \rho_{\mathrm{gain,K}}(\delta) \leq \sqrt{\mathbb{E}[Y^2]} \leq C.$$

Choose fixed $\eta, R > 0$ such that, for $|r - \rho_{\mathrm{gain,K}}(\delta)| \leq 2\eta$,

$$\sqrt{c_0}/2 \leq r \leq R, \qquad r^2 - q_{K,L} \geq c_0/2. \tag{68}$$

Such constants exist because the target radius is uniformly bounded and its squared gap is at least c_0. Lemma 1, with $c_* = c_0/2$, implies that whenever $\tau_{K,L}\Delta_n \leq \epsilon_0$,

$$\begin{aligned} &m_{K,L}\left(\pi, \sqrt{r^2 + C_0\tau_{K,L}\Delta_n}\right) - C_0\tau_{K,L}\Delta_n \\ &\leq \widehat{m}_{K,L}(\pi, r) \\ &\leq m_{K,L}\left(\pi, \sqrt{r^2 - C_0\tau_{K,L}\Delta_n}\right) + C_0\tau_{K,L}\Delta_n \end{aligned} \tag{69}$$

uniformly over these radii and $\pi \in \Pi$. The feasible coefficient sets are nonempty and compact. The policy Lipschitz condition and compactness of Θ therefore ensure that the policy maxima are attained. Taking maxima in Equation (69) gives

$$\begin{aligned} & M_{K,L}\left(\sqrt{r^2 + C_0\tau_{K,L}\Delta_n}\right) - C_0\tau_{K,L}\Delta_n \\ & \quad \le \widehat{M}_{K,L}(r) \\ & \quad \le M_{K,L}\left(\sqrt{r^2 - C_0\tau_{K,L}\Delta_n}\right) + C_0\tau_{K,L}\Delta_n. \end{aligned} \tag{70}$$

We will use these radius shifts to locate the estimated gain radius. The same stability lemma also gives

$$\sup_{\substack{\pi\in\Pi \\ |r-\rho_{\mathrm{gain,K}}(\delta)|\le 2\eta}} |\widehat{m}_{K,L}(\pi, r) - m_{K,L}(\pi, r)| \le C_0\tau_{K,L}^2\zeta_{K,L}\Delta_n, \tag{71}$$

and

$$|\widehat{q}_{K,L} - q_{K,L}| \le C_0\tau_{K,L}\Delta_n.$$

Step 2: Bound the change in the population value. The projection identity gives

$$\bar{\boldsymbol{\beta}}_{K,L}^\top G_{K,L}\bar{\boldsymbol{\beta}}_{K,L} + q_{K,L} = \mathbb{E}[\boldsymbol{\psi}_L Y]^\top Q_L^{-1}\mathbb{E}[\boldsymbol{\psi}_L Y] \le \mathbb{E}[Y^2].$$

Thus, for a fixed $C_1 \ge 1$,

$$\begin{aligned} |\bar{\boldsymbol{\beta}}_{K,L}^\top \mathbb{E}[\boldsymbol{h}_\pi]| &\le \|G_{K,L}^{1/2}\bar{\boldsymbol{\beta}}_{K,L}\|_2 \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2 \\ &\le C_1\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2. \end{aligned} \tag{72}$$

For $\sqrt{q_{K,L}} \le r < r'$, choose π' attaining $M_{K,L}(r')$. Since $\pi_0 \in \Pi$, Equation (8) implies

$$0 \le M_{K,L}(r') \le \bar{\boldsymbol{\beta}}_{K,L}^\top \mathbb{E}[\boldsymbol{h}_{\pi'}] \le C_1\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_{\pi'}]\|_2.$$

Consequently,

$$\begin{aligned} M_{K,L}(r) - M_{K,L}(r') &\ge m_{K,L}(\pi', r) - m_{K,L}(\pi', r') \\ &= \frac{(r'-r)(r'+r)\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_{\pi'}]\|_2}{\sqrt{r'^2 - q_{K,L}} + \sqrt{r^2 - q_{K,L}}} \\ &\ge (r'-r)\|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_{\pi'}]\|_2 \\ &\ge \frac{r'-r}{C_1} M_{K,L}(r'). \end{aligned} \tag{73}$$

The second inequality uses $q_{K,L} \ge 0$. By Lemma 2, $M_{K,L}\{\rho_{\mathrm{gain},K,L}(\delta)\} = \delta$. Applying Equation (73) on the two sides of this radius gives

$$\begin{aligned} M_{K,L}(r) - \delta &\ge \frac{\delta}{C_1}\{\rho_{\mathrm{gain},K,L}(\delta) - r\}, \quad \sqrt{q_{K,L}} \le r \le \rho_{\mathrm{gain},K,L}(\delta), \\ \delta - M_{K,L}(r) &\ge \frac{\delta\{r - \rho_{\mathrm{gain},K,L}(\delta)\}}{C_1 + r - \rho_{\mathrm{gain},K,L}(\delta)}, \quad r \ge \rho_{\mathrm{gain},K,L}(\delta). \end{aligned}$$

For the second inequality, Equation (73) first gives

$$\delta - M_{K,L}(r) \ge \frac{r - \rho_{\mathrm{gain},K,L}(\delta)}{C_1} M_{K,L}(r),$$

and rearrangement yields the stated bound. With $\kappa = \delta/(C_1 + \eta) > 0$, it follows that

$$|M_{K,L}(r) - \delta| \ge \kappa|r - \rho_{\mathrm{gain},K,L}(\delta)| \quad \text{if } |r - \rho_{\mathrm{gain},K,L}(\delta)| \le \eta. \tag{74}$$

All radii in this last interval are population-feasible for sufficiently large L: Lemma 2 gives $|\rho_{\mathrm{gain},K,L}(\delta) - \rho_{\mathrm{gain,K}}(\delta)| \le Cb_L^2 \le \eta/2$, and Equation (68) then applies. The constant κ is independent of n, K, L.

Step 3: Locate the estimated gain radius. For the radii in Equation (68), if $C_0\tau_{K,L}\Delta_n \le c_0/4$, then

$$\left|\sqrt{r^2 \pm C_0\tau_{K,L}\Delta_n} - r\right| = \frac{C_0\tau_{K,L}\Delta_n}{\sqrt{r^2 \pm C_0\tau_{K,L}\Delta_n} + r} \le \frac{2C_0}{\sqrt{c_0}}\tau_{K,L}\Delta_n.$$

Choose a fixed C_2 satisfying

$$C_2 \ge \frac{2C_0}{\sqrt{c_0}} + \frac{2C_0}{\kappa}.$$

Reduce ϵ_0, if necessary, so that

$$C_0\epsilon_0 \le c_0/4, \qquad \left(C_2 + \frac{2C_0}{\sqrt{c_0}}\right)\epsilon_0 \le \eta/2.$$

For sufficiently large L and $\tau_{K,L}\Delta_n \le \epsilon_0$, the radii $\rho_{\mathrm{gain},K,L}(\delta) \pm C_2\tau_{K,L}\Delta_n$ and their shifted versions below are within η of $\rho_{\mathrm{gain},K,L}(\delta)$. The unshifted radii are also sample-feasible, because

$$r^2 - \widehat{q}_{K,L} \ge r^2 - q_{K,L} - |\widehat{q}_{K,L} - q_{K,L}| \ge c_0/2 - C_0\tau_{K,L}\Delta_n \ge c_0/4.$$

For the lower comparison, the square-root bound implies

$$\begin{aligned}&\sqrt{\{\rho_{\mathrm{gain},K,L}(\delta) - C_2\tau_{K,L}\Delta_n\}^2 + C_0\tau_{K,L}\Delta_n} \\ &\quad\le \rho_{\mathrm{gain},K,L}(\delta) - \left(C_2 - \frac{2C_0}{\sqrt{c_0}}\right)\tau_{K,L}\Delta_n \\ &\quad\le \rho_{\mathrm{gain},K,L}(\delta) - \frac{2C_0}{\kappa}\tau_{K,L}\Delta_n.\end{aligned}$$

Equations (70) and (74) therefore give

$$\begin{aligned}&\widehat{M}_{K,L}\{\rho_{\mathrm{gain},K,L}(\delta) - C_2\tau_{K,L}\Delta_n\} \\ &\quad\ge M_{K,L}\left(\sqrt{\{\rho_{\mathrm{gain},K,L}(\delta) - C_2\tau_{K,L}\Delta_n\}^2 + C_0\tau_{K,L}\Delta_n}\right) - C_0\tau_{K,L}\Delta_n \\ &\quad\ge \delta + 2C_0\tau_{K,L}\Delta_n - C_0\tau_{K,L}\Delta_n = \delta + C_0\tau_{K,L}\Delta_n.\end{aligned}$$

For the upper comparison,

$$\begin{aligned}&\sqrt{\{\rho_{\mathrm{gain},K,L}(\delta) + C_2\tau_{K,L}\Delta_n\}^2 - C_0\tau_{K,L}\Delta_n} \\ &\quad\ge \rho_{\mathrm{gain},K,L}(\delta) + \left(C_2 - \frac{2C_0}{\sqrt{c_0}}\right)\tau_{K,L}\Delta_n \\ &\quad\ge \rho_{\mathrm{gain},K,L}(\delta) + \frac{2C_0}{\kappa}\tau_{K,L}\Delta_n,\end{aligned}$$

so the same two equations imply

$$\begin{aligned}&\widehat{M}_{K,L}\{\rho_{\mathrm{gain},K,L}(\delta) + C_2\tau_{K,L}\Delta_n\} \\ &\quad\le M_{K,L}\left(\sqrt{\{\rho_{\mathrm{gain},K,L}(\delta) + C_2\tau_{K,L}\Delta_n\}^2 - C_0\tau_{K,L}\Delta_n}\right) + C_0\tau_{K,L}\Delta_n \\ &\quad\le \delta - 2C_0\tau_{K,L}\Delta_n + C_0\tau_{K,L}\Delta_n = \delta - C_0\tau_{K,L}\Delta_n.\end{aligned}$$

If $\Delta_n = 0$, the sample and population moments defining the criteria and policy objectives coincide, so the two gain radii are equal. Suppose now that $\Delta_n > 0$. The preceding comparisons are then strict at level δ. By Equation (13) and compactness of Θ, $\widehat{M}_{K,L}$ is continuous and nonincreasing on $r \ge \sqrt{\widehat{q}_{K,L}}$. For $\sqrt{\widehat{q}_{K,L}} \le r < r'$ and a policy π' attaining $\widehat{M}_{K,L}(r') > 0$, we have

$$\|\widehat{G}_{K,L}^{-1/2} E_n[\boldsymbol{h}_{\pi'}]\|_2 > 0,$$

since a zero sample value direction would give zero improvement. The sample closed form yields

$$\begin{aligned}\widehat{M}_{K,L}(r) - \widehat{M}_{K,L}(r') &\geq \widehat{m}_{K,L}(\pi', r) - \widehat{m}_{K,L}(\pi', r') \\ &= \{\sqrt{r'^2 - \widehat{q}_{K,L}} - \sqrt{r^2 - \widehat{q}_{K,L}}\}\|\widehat{G}_{K,L}^{-1/2} E_n[\boldsymbol{h}_{\pi'}]\|_2 > 0.\end{aligned}$$

Thus $\widehat{M}_{K,L}$ crosses the positive level δ exactly once, and this crossing is the largest radius supporting δ. In particular,

$$\widehat{M}_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\} = \delta,$$
$$\rho_{\mathrm{gain},K,L}(\delta) - C_2\tau_{K,L}\Delta_n \leq \widehat{\rho}_{\mathrm{gain},K,L}(\delta) \leq \rho_{\mathrm{gain},K,L}(\delta) + C_2\tau_{K,L}\Delta_n.$$

This proves Equation (67). The comparisons in Step 1 hold simultaneously over all radii in the neighborhood specified there, so their use at the sample-dependent comparison radii above is valid. □

Lemma 4 (True value of a policy)**.** *Fix $\delta > 0$ such that the population gain-constrained problem has a maximizer $\rho_{\mathrm{gain,K}}(\delta)$. Suppose Assumptions 2 and 3 hold at this radius. Assume, additionally, that*

$$\inf_{\widetilde{f}\in\mathcal{S}_K} \|\widetilde{f} - f_0\|_{\infty,\mathcal{D}_f} \leq \alpha_K \longrightarrow 0, \qquad \|\mathbb{E}[\epsilon \mid \boldsymbol{X}, Z]\|_2 \leq \rho_{\mathrm{gain,K}}(\delta).$$

Let Δ_n be defined as in Lemma 1. There exist fixed constants $\eta, \epsilon_0, C > 0$, independent of n, K, L, such that, for sufficiently large n, K, L, whenever $\tau_{K,L}\Delta_n \leq \epsilon_0$,

$$\begin{aligned}V_{f_0}(\pi) - V_{f_0}(\pi_0) \geq \widehat{m}_{K,L}(\pi, r) &- C\tau_{K,L}^2\zeta_{K,L}\Delta_n \\ &- C\tau_{K,L}\zeta_{K,L}\{|r - \rho_{\mathrm{gain,K}}(\delta)| + \alpha_K + n^{-1}\}\end{aligned} \tag{75}$$

uniformly over $\pi \in \Pi$ and $|r - \rho_{\mathrm{gain,K}}(\delta)| \leq \eta$. All sample minima in this inequality are well-defined and attained.

Proof.

Step 1: Compare the true value with the population sieve value. Choose $\widetilde{f} \in \mathcal{S}_K$ such that

$$\|\widetilde{f} - f_0\|_{\infty,\mathcal{D}_f} \leq \alpha_K + n^{-1}.$$

Since $Y = f_0(\boldsymbol{X}, P) + \epsilon$ and $(\boldsymbol{X}, P) \in \mathcal{D}_f$ almost surely,

$$\begin{aligned}\|\mathbb{E}[Y - \widetilde{f}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\|_2 &\leq \|\mathbb{E}[\epsilon \mid \boldsymbol{X}, Z]\|_2 + \|\mathbb{E}[f_0(\boldsymbol{X}, P) - \widetilde{f}(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]\|_2 \\ &\leq \rho_{\mathrm{gain,K}}(\delta) + \|f_0 - \widetilde{f}\|_{\infty,\mathcal{D}_f} \\ &\leq \rho_{\mathrm{gain,K}}(\delta) + \alpha_K + n^{-1}.\end{aligned}$$

Consequently,

$$\begin{aligned}\widetilde{f} &\in \mathcal{F}\big(\rho_{\mathrm{gain,K}}(\delta) + \alpha_K + n^{-1}; \mathcal{S}_K\big) \\ &\subseteq \mathcal{F}\big(\rho_{\mathrm{gain,K}}(\delta) + \alpha_K + n^{-1}; \mathcal{S}_K, \mathcal{T}_L\big),\end{aligned}$$

where the inclusion follows from Equation (5). The bounded price interval and Assumption 3(i) give

$$\begin{aligned}&\sup_{\pi\in\Pi}\Big|\{V_{\widetilde{f}}(\pi) - V_{\widetilde{f}}(\pi_0)\} - \{V_{f_0}(\pi) - V_{f_0}(\pi_0)\}\Big| \\ &\quad\leq \sup_{\pi\in\Pi} E\big[|\pi(\boldsymbol{X})| + |\pi_0(\boldsymbol{X})|\big]\|\widetilde{f} - f_0\|_{\infty,\mathcal{D}_f} \\ &\quad\leq C(\alpha_K + n^{-1}).\end{aligned}$$

It follows that, uniformly over $\pi \in \Pi$,

$$\begin{aligned}V_{f_0}(\pi) - V_{f_0}(\pi_0) &\geq V_{\widetilde{f}}(\pi) - V_{\widetilde{f}}(\pi_0) - C(\alpha_K + n^{-1}) \\ &\geq m_{K,L}\big(\pi, \rho_{\mathrm{gain,K}}(\delta) + \alpha_K + n^{-1}\big) - C(\alpha_K + n^{-1}).\end{aligned} \tag{76}$$

Step 2: Bound the effect of changing the radius. Lemma 2 gives

$$\sqrt{c_0} \le \rho_{\mathrm{gain,K}}(\delta) \le \sqrt{\mathbb{E}[Y^2]} \le \sqrt{C}.$$

Choose a fixed $\eta > 0$ sufficiently small that

$$0 < \rho_{\mathrm{gain,K}}(\delta) - \eta, \qquad \rho_{\mathrm{gain,K}}(\delta) + \eta \le \sqrt{C} + \eta,$$
$$\inf_{|r - \rho_{\mathrm{gain,K}}(\delta)| \le \eta} (r^2 - q_{K,L}) \ge c_0 - 2\sqrt{C}\,\eta - \eta^2 \ge c_0/2.$$

By the bounded price interval and the basis envelope,

$$\begin{aligned}\sup_{\pi\in\Pi} \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2 &\le \tau_{K,L} \sup_{\pi\in\Pi} \|\mathbb{E}[\boldsymbol{h}_\pi]\|_2 \\ &\le \tau_{K,L} \sup_{\pi\in\Pi} E\Big[|\pi(\boldsymbol{X})|\|\boldsymbol{\phi}_K\{\boldsymbol{X}, \pi(\boldsymbol{X})\}\|_2 \\ &\qquad + |\pi_0(\boldsymbol{X})|\|\boldsymbol{\phi}_K\{\boldsymbol{X}, \pi_0(\boldsymbol{X})\}\|_2\Big] \\ &\le C\tau_{K,L}\zeta_{K,L}.\end{aligned}$$

For $|r - \rho_{\mathrm{gain,K}}(\delta)| \le \eta$ and $|r' - \rho_{\mathrm{gain,K}}(\delta)| \le \eta$, Equation (8) therefore gives

$$\begin{aligned}|m_{K,L}(\pi, r) - m_{K,L}(\pi, r')| &= \left|\sqrt{r^2 - q_{K,L}} - \sqrt{r'^2 - q_{K,L}}\right| \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2 \\ &= \frac{|r - r'|(r + r')}{\sqrt{r^2 - q_{K,L}} + \sqrt{r'^2 - q_{K,L}}} \|G_{K,L}^{-1/2}\mathbb{E}[\boldsymbol{h}_\pi]\|_2 \\ &\le C\tau_{K,L}\zeta_{K,L}|r - r'|,\end{aligned} \tag{77}$$

uniformly over $\pi \in \Pi$. Since $\alpha_K + n^{-1} \to 0$, $\rho_{\mathrm{gain,K}}(\delta) + \alpha_K + n^{-1}$ belongs to this interval for sufficiently large n, K. Equations (76) and (77) imply

$$\begin{aligned}V_{f_0}(\pi) - V_{f_0}(\pi_0) &\ge m_{K,L}(\pi, r) - C\tau_{K,L}\zeta_{K,L}\left|\rho_{\mathrm{gain,K}}(\delta) + \alpha_K + n^{-1} - r\right| - C(\alpha_K + n^{-1}) \\ &\ge m_{K,L}(\pi, r) - C\tau_{K,L}\zeta_{K,L}\big\{|r - \rho_{\mathrm{gain,K}}(\delta)| + \alpha_K + n^{-1}\big\} - C(\alpha_K + n^{-1}).\end{aligned}$$

Moreover, $G_{K,L} \preceq Q_K \preceq CI_K$ and $\zeta_{K,L}^2 \ge E\|\boldsymbol{\phi}_K\|_2^2 = \mathrm{tr}(Q_K) \ge cK \ge c$ give

$$\tau_{K,L}\zeta_{K,L} \ge \sqrt{c/C} > 0.$$

Absorbing $C(\alpha_K + n^{-1})$ into the preceding bound yields

$$\begin{aligned}V_{f_0}(\pi) - V_{f_0}(\pi_0) &\ge m_{K,L}(\pi, r) \\ &\quad - C\tau_{K,L}\zeta_{K,L}\big\{|r - \rho_{\mathrm{gain,K}}(\delta)| + \alpha_K + n^{-1}\big\}.\end{aligned}$$

Step 3: Replace the population sieve value by its sample counterpart. The radii under consideration are uniformly bounded and satisfy $r^2 - q_{K,L} \ge c_0/2$. Whenever $\tau_{K,L}\Delta_n \le \epsilon_0$ for sufficiently small ϵ_0, Lemma 1 ensures that the sample minima are well-defined and attained. Equation (25) gives

$$\begin{aligned}m_{K,L}(\pi, r) &\ge \widehat{m}_{K,L}(\pi, r) - |\widehat{m}_{K,L}(\pi, r) - m_{K,L}(\pi, r)| \\ &\ge \widehat{m}_{K,L}(\pi, r) - C\tau_{K,L}^2\zeta_{K,L}\Delta_n,\end{aligned}$$

uniformly over the stated policies and radii. Substitution into the bound from Step 2 proves Equation (75). □

Proof of Theorem 4.

Step 1: Combine the two radius errors. By Equation (59) and Assumption 2(iii),

$$\Delta_n = O_p\left(\zeta_{K,L}\sqrt{\frac{\log n}{n}}\right), \qquad \tau_{K,L}\Delta_n = o_p(1).$$

Hence the condition $\tau_{K,L}\Delta_n \le \epsilon_0$ in Lemma 3 holds with probability tending to one. Combining this lemma with Lemma 2,

$$\begin{aligned}
&|\widehat{\rho}_{\mathrm{gain},K,L}(\delta) - \rho_{\mathrm{gain,K}}(\delta)| \\
&\quad \le \underbrace{|\widehat{\rho}_{\mathrm{gain},K,L}(\delta) - \rho_{\mathrm{gain},K,L}(\delta)|}_{\text{statistical error}} \\
&\qquad + \underbrace{|\rho_{\mathrm{gain},K,L}(\delta) - \rho_{\mathrm{gain,K}}(\delta)|}_{\text{approximation error}} \\
&\quad \le C\tau_{K,L}\Delta_n + Cb_L^2
\end{aligned} \tag{78}$$

whenever $\tau_{K,L}\Delta_n \le \epsilon_0$, for sufficiently large L. It follows that

$$|\widehat{\rho}_{\mathrm{gain},K,L}(\delta) - \rho_{\mathrm{gain,K}}(\delta)| = O_p\left(\tau_{K,L}\zeta_{K,L}\sqrt{\frac{\log n}{n}} + b_L^2\right) = o_p(1). \tag{79}$$

The samples on which the gain problem is undefined or the target is unsupported have probability tending to zero by Lemma 3.

Step 2: Evaluate the true-value bound at the selected policy. By Equation (79), the estimated radius lies in the neighborhood of Lemma 4 with probability tending to one. Choose a measurable maximizing policy for the sample problem. Lemma 3 then gives, with probability tending to one,

$$\begin{aligned}
&\widehat{m}_{K,L}\big(\widehat{\pi}^{\star}_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}, \widehat{\rho}_{\mathrm{gain},K,L}(\delta)\big) \\
&\quad = \widehat{M}_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\} = \delta.
\end{aligned}$$

The bound in Lemma 4 is uniform in both the policy and radius, so it applies to these sample-selected arguments. Using Equation (75),

$$\begin{aligned}
&V_{f_0}\left[\widehat{\pi}^{\star}_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}\right] - V_{f_0}(\pi_0) \\
&\quad \ge \delta - C\tau_{K,L}^2\zeta_{K,L}\Delta_n \\
&\qquad - C\tau_{K,L}\zeta_{K,L}\left\{|\widehat{\rho}_{\mathrm{gain},K,L}(\delta) - \rho_{\mathrm{gain,K}}(\delta)| + \alpha_K + n^{-1}\right\} \\
&\quad \ge \delta - C\tau_{K,L}^2\zeta_{K,L}\Delta_n - C\tau_{K,L}\zeta_{K,L}\{\tau_{K,L}\Delta_n + b_L^2 + \alpha_K + n^{-1}\} \\
&\quad \ge \delta - C\left\{\tau_{K,L}^2\zeta_{K,L}\Delta_n + \tau_{K,L}\zeta_{K,L}(b_L^2 + \alpha_K + n^{-1})\right\}
\end{aligned}$$

on the same events, whose probabilities tend to one. Moreover, $G_{K,L} \preceq Q_K \preceq CI_K$ and $\zeta_{K,L}^2 \ge \mathrm{tr}(Q_K) \ge cK$ imply $\tau_{K,L}\zeta_{K,L} \ge \sqrt{c/C} > 0$. Hence

$$\frac{\tau_{K,L}\zeta_{K,L}n^{-1}}{\tau_{K,L}^2\zeta_{K,L}^2\sqrt{\log n/n}} = \frac{1}{\tau_{K,L}\zeta_{K,L}\sqrt{n\log n}} \longrightarrow 0.$$

Using the stochastic order of Δ_n from Step 1, we obtain

$$\begin{aligned}
&V_{f_0}\left[\widehat{\pi}^{\star}_{K,L}\{\widehat{\rho}_{\mathrm{gain},K,L}(\delta)\}\right] - V_{f_0}(\pi_0) \\
&\quad \ge \delta - O_p\left(\tau_{K,L}^2\zeta_{K,L}^2\sqrt{\frac{\log n}{n}} + \tau_{K,L}\zeta_{K,L}(b_L^2 + \alpha_K)\right).
\end{aligned}$$

□

C Additional simulation details for section 5

Common policy class and basis functions. All five methods optimize $\pi_{\boldsymbol{\theta}}(\boldsymbol{X}) = \theta_0 + \sum_{j=1}^{5}\theta_j X_j$, with $\theta_0 \in [0.7, 1.5]$ and $\theta_j \in [-0.06, 0.06]$. These common bounds ensure that prices lie in $[0.4, 1.8]$ for every $\boldsymbol{X} \in [-1, 1]^5$; no clipping is needed. OLS, TLS-IV, and the robust estimator use the same 7-dimensional sieve space $\mathcal{S}_K = \mathrm{span}\{1, X_1, \ldots, X_5, P\}$. We set $\mathcal{T}_L = \mathrm{span}\{1, X_1, \ldots, X_5, Z\}$.

Demand estimation and tuning. TLS-IV is series two-stage least squares (Newey & Powell, 2003) with the same bases as the robust estimator. K-IV (Singh et al., 2019) uses exact Gaussian kernels on standardized $(\boldsymbol{X}, P)$ and $(\boldsymbol{X}, Z)$. Bandwidth multipliers are $0.5, 1, 2$ times the median-distance bandwidth; first-stage ridge values are $10^{-3}, 10^{-2}$, and second-stage values are $10^{-4}, 10^{-3}, 10^{-2}$. No random-feature or Nyström approximation is used.

DF-IV (Xu et al., 2021) uses two networks of widths $6, 32, 16, 8$ with two `tanh` hidden layers and an appended intercept. Both ridge stages are differentiated analytically. Adam uses learning rate 10^{-3} and weight decay 10^{-5}, with at most 500 epochs. We compare common ridge values 10^{-3} and 10^{-2}. Validation is checked every ten epochs, with a 60-epoch patience after epoch 100; the best checkpoint is retained. Computation uses double precision on CPU.

For K-IV and DF-IV, observations are randomly divided into 40% first-stage, 40% second-stage, and 20% validation samples. Hyperparameters minimize the held-out weighted moment criterion using the fixed 7-dimensional instrument basis. The selected K-IV model is refitted on two disjoint halves of all observations. For DF-IV, the selected features are frozen and the final two-stage heads are refitted on these halves. Neither true demand nor population policy improvement is used in this tuning.

Policy optimization. All five methods use L-BFGS-B: OLS, TLS-IV, K-IV, and DF-IV methods maximize fitted empirical revenue, while the robust method maximizes empirical worst-case policy improvement in Equation (13).

Population evaluation. Under the known data-generating process, the true policy improvement is calculated by

$$V_{f_0}(\pi_{\boldsymbol{\theta}}) - V_{f_0}(\pi_0) = 1.70 - 120 \left\{ (\theta_0 - 1.10)^2 + \frac{1}{3} \sum_{j=1}^{5} (\theta_j - 0.05)^2 \right\}. \tag{80}$$

We evaluate all learned policies over the same population ambiguity set $\mathcal{F}(\rho; \mathcal{S}_K)$ with various ρ. Let $m(\boldsymbol{X}, Z) = \mathbb{E}[Y \mid \boldsymbol{X}, Z]$ and $\boldsymbol{h}_K(\boldsymbol{X}, Z) = \mathbb{E}[\boldsymbol{\phi}_K(\boldsymbol{X}, P) \mid \boldsymbol{X}, Z]$. Define

$$\begin{aligned} G_K &= \mathbb{E}[\boldsymbol{h}_K \boldsymbol{h}_K^\top], \\ \bar{\boldsymbol{\beta}}_K &= G_K^{-1} \mathbb{E}[\boldsymbol{h}_K m], \\ q_K &= \mathbb{E}[(m - \boldsymbol{h}_K^\top \bar{\boldsymbol{\beta}}_K)^2]. \end{aligned}$$

Since the conditional-moment constraint is quadratic in $\boldsymbol{\beta}$ and the policy improvement is linear in $\boldsymbol{\beta}$, the population worst-case improvement has the closed-form expression

$$\min_{f \in \mathcal{F}(\rho; \mathcal{S}_K)} \{V_f(\pi) - V_f(\pi_0)\} = V_{f_{\bar{\boldsymbol{\beta}}_K}}(\pi) - V_{f_{\bar{\boldsymbol{\beta}}_K}}(\pi_0) - \sqrt{\rho^2 - q_K} \left\| G_K^{-1/2} \nabla_{\boldsymbol{\beta}} \{V_{f_{\boldsymbol{\beta}}}(\pi) - V_{f_{\boldsymbol{\beta}}}(\pi_0)\} \right\|_2. \tag{81}$$

We use the known data-generating model to calculate m, $\boldsymbol{h}_K$, and the expectations defining G_K, $\bar{\boldsymbol{\beta}}_K$, and q_K by numerical integration.