Cross-Validation for Ridge Regularization in Debiased Machine Learning*

Ka Yan Cheng
Department of Economics
Emory University
ka.yan.cheng@emory.edu

Last updated: November 9, 2026

Please click here for the latest version.

Abstract

This paper studies cross-validation in debiased machine learning (DML) for linear functionals of a regression function γ_0, where γ_0 and a debiasing nuisance α_0 are estimated with ridge regularization. To ensure valid inference, DML places a joint rate requirement on the estimation errors for γ_0 and α_0. The feasibility of attaining the required rate depends on the joint complexity of the two. We study cross-validation for ridge-regularized estimators of γ_0 and α_0 on a shared, ill-conditioned dictionary of growing size. Based on our learning-rate bounds, we characterize the boundary at which the rate condition is achievable in terms of how fast the dictionary can approximate the nuisances. We then propose theory-driven feasible candidate sets for both the dictionary dimensions and the ridge penalties, and formally justify the use of separate cross-validation in DML based on these sets. In particular, we develop an oracle inequality for the resulting nuisance estimation risk. Our simulation results confirm these findings.

KEYWORDS: Debiased machine learning, Riesz regression, ridge regularization, cross-validation, data-driven tuning parameter selection, oracle inequality.

*I am grateful to Zheng Fang and Alejandro Sánchez-Becerra for their continuous guidance and support. I would also like to thank Young Ahn, Ivan Fernandez-Val, Florian F. Gunsilius, David A. Hirshberg, Yiqi Liu, Esfandiar Maasoumi, Pedro H. C. Sant'Anna, and Rahul Singh for helpful comments and discussions. All errors are mine.

1 Introduction

In semiparametric models, moment conditions defining the parameter of interest involve nonparametric nuisance functions. Estimation of these functions is therefore integral to inference on the target parameter. Without appropriate bias correction, errors arising from these initial steps can propagate into the moment condition, bias the resulting estimator, and render standard confidence intervals invalid (Chernozhukov et al., 2018). By explicitly accounting for the impact of first-stage estimation error on the moment equation, debiased machine learning (DML) provides a principled framework for conducting estimation and inference on parameters that depend on high-dimensional or nonparametric nuisance functions.

When the target parameter is a bounded linear functional of the regression function, such impact can be characterized by the product of the regression residual with the functional's associated Riesz representer (Chernozhukov et al., 2022a). The Riesz representer captures the sensitivity of the functional to perturbations in the regression function, thereby serves as an important ingredient for bias correction. When the Riesz representer is unknown, directly estimating it based on its analytical form can be unstable, as this might involve inverting unobserved quantities that can get close to zero (e.g., the propensity scores in Example 2.1 below). Worse yet, this may even be impossible as the representer need not admit a closed-form expression at all. A substantial body of recent work studies estimation of the Riesz representer without requiring knowledge on its form. This line of research has been conducted under various frameworks, including Chernozhukov et al. (2022b,c, 2023, 2026); Singh (2024). All these approaches rely on the idea of Riesz regression proposed in Chernozhukov et al. (2024). By exploiting the Riesz representation form of the moment condition, Riesz regression allows us to estimate the Riesz representer by utilizing only the moment equation that defines the target parameter.[1] In this paper, we adopt Riesz regression and model both nuisances (the regression function γ_0 and the Riesz representer α_0) as linear combinations of a common, growing set of feature functions.

Aligned with the emphasis on predictive performance in the field of machine learning, DML demands only fast enough L^2-convergence rates on both nuisances. As a result, with less concern about collinearity, it encourages practitioners to use a richer dictionary of functions. In that case, it is reasonable to believe that the Gram matrix will become increasingly ill-conditioned as the dimension of the dictionary grows.[2] This therefore motivates the use of ridge penalties to stabilize estimation on both sides. In this paper, we work within the DML framework for bounded linear functional of a regression function, and estimate the nuisance functions using two ridge-regularized sieve estimators

[1] Building on the same idea, instead of estimating the Riesz representer, Hirshberg and Wager (2021) considered a minimax linear estimator for the entire bias correction term, which also does not require the explicit form of the Riesz representer.

[2] The Gram matrix is the expected outer product of the dictionary vector.

on a shared, ill-conditioned dictionary of growing size. Focusing on such case, we ask two important questions: (i) under what regularity conditions can the two nuisances be estimated well enough, jointly, to warrant valid $\sqrt{n}$-inference on the parameter of interest, and (ii) in practice, how should we choose the sieve size and the ridge penalty for their estimator.

DML delivers valid inference provided the product of the two nuisance estimators' L^2-rates is $o_p(n^{-1/2})$. As our first contribution, we characterize the boundary at which this condition is attainable according to our learning-rate bounds. The characterization is given in terms of how fast the dictionary can approximate the two nuisances as its dimension grows. Such speed is governed by both the underlying data-generating process and the choice of dictionary. Approximation theory quantifies it by relating the decay of the approximation error to the regularity of the target and the richness of the dictionary. Specifically, it delivers rates of the form $\delta(p) \lesssim p^{-\kappa}$, where $\delta(p)$ is the approximation error and p is the dictionary dimension. See, for example, Timan (1963) for a classical treatment, and Section 2.3 of Chen (2007) for an overview of the results that are most relevant in economic settings. Writing κ_γ and κ_α for the approximation exponents of the outcome regression and the Riesz representer, we trace out a product-rate feasibility frontier in the $(\kappa_\gamma, \kappa_\alpha)$ space. Throughout, feasibility refers to certification of the product-rate condition that is backed by our learning-rate bounds. This frontier separates the pairs for which our learning-rate bounds certify the product-rate requirement for some tuning choices from those for which no tuning choice can provide such a guarantee. This is important because, for problems associated with a pair of $(\kappa_\gamma, \kappa_\alpha)$ above the frontier, our analysis provides a route to justifying separate cross-validation of the two nuisances based on their respective prediction risks. Outside this region, based on our learning-rate bounds, the product-rate requirement cannot be met by any choice of tuning, and therefore valid inference, if available at all, may call for tuning that depart from predictive optimality, as undersmoothing does in kernel density estimation.

We must emphasize that our frontier is not a statement about the possibility in achieving valid inference in general. Rather, for average treatment effect (our Example 2.1 below), the minimal Hölder smoothness conditions for $\sqrt{n}$-inference has already been established by Robins et al. (2009).[3] And, as in the present paper, Bradic et al. (2025) work directly with the approximation exponent. With a well-conditioned Gram matrix, they show that $\sqrt{n}$-inference on the coefficient in a partially linear model is attainable on a strictly larger region than the product-rate condition permits. That said, valid inference remains possible without the product-rate condition, though the cost must be

[3]Related results in the paper are stated for an example they termed missing at random (MAR), which corresponds to the treated arm in our Example 2.1. The analogous result for the untreated arm can be obtained by replacing the missingness indicator A (or, in our notation, the treatment indicator D) with $1 - A$ (or $1 - D$), respectively.

paid elsewhere. It can be paid, for example, by: (i) a more complicated sample-splitting scheme: Newey and Robins (2018) employ three-way cross-fitting,[4] estimating the two nuisances and the target estimator on three disjoint subsamples; or (ii) a higher order correction: Robins et al. (2017) proposes a higher-order influence function estimator, which attains the minimal requirement stated in Robins et al. (2009). We focus on the product-rate condition because it reduces the inference validity problem to the predictive accuracy of each nuisance, which is exactly what most off-the-shelf machine learning methods and existing tuning approaches are built to target.

While our frontier says when the product-rate condition can be met, meeting it still requires appropriately chosen tuning parameters. Based on one of Chernozhukov et al. (2018)'s data-generating process, Bach et al. (2024, Appendix A) works with random forest to illustrate the importance of proper tuning in DML. With a one thousand repetitions simulation, they showed that the studentized DML estimates are approximately normal when the nuisances are well tuned, whereas under underfitting learners the empirical distributions of these estimates is centered away from zero, despite the use of an orthogonal moment and cross-fitting.

By working within the region where separate prediction-based tuning of the two nuisances aligns with the goal of providing valid inference, we can then draw on a substantial existing literature on tuning for predictive performance. For a review of tuning methods that are commonly used in econometrics, we refer the reader to Chetverikov (2024) and references therein. Among these methods, cross-validation is one of the most commonly used, and its theoretical properties in nonparametric estimation has been shown in different settings. For penalty selection, see Li (1987) and Andrews (1991) for linear estimators (including ridge), and Chetverikov et al. (2021) for the lasso. For the choice of sieve dimensions in nonparametric regression, see Hansen (2014, Section 8.9). In a general setup of choosing an estimator from a given collection, Vaart et al. (2006) establishes oracle inequalities for the cross-validation procedure. The work most closely related to ours in this area is Cetin (2026), which provide results of this type in the context of debiased machine learning, for both the estimator of the outcome regression and the Riesz representer.

All these results, except for Chetverikov et al. (2021), share a common form, that is, they guarantee that the selected tuning performs nearly as well as the best one in a given candidate set. Specifically, to support the inference result, Cetin (2026) directly assumes the sets that they choose from contains a pair of nuisance estimators satisfying the product-rate condition, but guidance on designing such a candidate set is still lacking. Therefore, our second contribution is to fill in this gap: we provide theory-driven candidate sets for both the dictionary dimensions and the ridge penalties. Through an oracle inequality argument, we establish L^2-convergence rates for the nuisance estima-

[4] We borrow the term three-way cross-fitting from Hirshberg and Wager (2021).

tors whose tuning parameters are selected from these sets by separate cross-validation. Expressed in terms of the approximation exponents κ_γ and κ_α, these rates bring our analysis full circle: they satisfy the product-rate requirement whenever the problem's associated exponent pair lies above our feasibility frontier.

The remainder of this paper is structured as follows. Section 2 introduces the setup and reviews the standard DML arguments. Section 3 describes the ridge-regularized sieve estimators for the first steps. Section 4 discusses the boundary on the nuisance problems in which guaranteed valid inference is possible with our learning-rate bounds. Section 5 proposes candidate sets for the tuning parameters and develop theoretical guarantee for separate cross-validation based on these sets. Section 6 provides simulation studies and finally, Section 7 concludes.

Notation. Throughout this paper, we denote by P_V the marginal distribution of any random variable $V \in \mathcal{V} \subset \mathbb{R}^{d_V}$. For $1 \leq p < \infty$, we define the space $L^p(P_V) \equiv \{f : \mathcal{V} \to \mathbb{R} : \|f\|_{L^p} < \infty\}$ with $\|f\|_{L^p} \equiv (\int_{\mathcal{V}} |f(v)|^p dP_V(v))^{1/p}$. In particular, when $p = 2$, $L^2(P_V)$ is a Hilbert space with inner product $\langle f, g\rangle \equiv \int_{\mathcal{V}} f(v)g(v)dP_V(v)$ and induced norm $\|f\|_{L^2} \equiv \langle f, f\rangle^{1/2}$. For any set $\Upsilon \subset L^2(P_V)$, $\overline{\Upsilon}$ represents the mean-square closure of Υ. Similarly, we define $L^\infty(P_V) \equiv \{f : \mathcal{V} \to \mathbb{R} : \|f\|_\infty < \infty\}$ with $\|f\|_\infty \equiv \operatorname{ess\,sup}_{v\in\mathcal{V}} |f(v)|$. Whenever we write expression in form of $\|\hat{\gamma}_{-k} - \gamma_0\|_{L^2}$, the estimator $\hat{\gamma}_{-k}$ is constructed by training samples $\{W_i : i \in I_{-k}\}$ only and we interpret these norms as being conditional on the training samples. In other words, the estimators are treated as fixed when evaluating the norm. For any matrix A, we denote its transpose by $A^\intercal$, and its inverse by A^{-1}. Moreover, with $\|v\|_2 \equiv \{v^\intercal v\}^{1/2}$ and $|v|$ indicating the Euclidean norm and the cardinality of a vector $v \in \mathbb{R}^{d_v}$, the spectral (ℓ_2-operator) norm of any matrix A is written as $\|A\|_{\text{op}} \equiv \sup_{x\neq 0}\{\|Ax\|_2/\|x\|_2\}$ and its Frobenius norm $\|A\|_{\text{F}}$. Lastly, for $a, b \in \mathbb{R}$, $a \lesssim b$ is used to denote that $a \leq cb$ for some constant $c > 0$ and $a \vee b \equiv \max\{a, b\}$.

2 Setup

In this section, we begin by reviewing the standard DML arguments. Let P_0 denote the true distribution of the observables $W \equiv (Y, X^\intercal)^\intercal \in \mathcal{Y}\times\mathcal{X} \subset \mathbb{R}^{1+d_X}$, and let $P_{0,X}$ denote the true marginal distribution of X. The parameter of interest $\theta_0 \in \mathbb{R}$ is characterized by the moment condition in which

$$\theta = \mathbb{E}[m(W, \gamma_0)] \tag{1}$$

admits a unique solution in $\mathbb{R}$ at θ_0. Here, $m : \mathcal{Y} \times \mathcal{X} \times L^2(P_{0,X}) \to \mathbb{R}$ is a real-valued, predefined function of the data W and of the unknown nuisance regression function

$\gamma_0 : \mathcal{X} \to \mathbb{R}$. Specifically, we have

$$\gamma_0(x) = \mathbb{E}[Y|X = x] \ , \quad x \in \mathcal{X} \ ,$$

and we assume $\gamma_0 \in L^2(P_{0,X})$.

2.1 Orthogonal Moment and the Auto-DML Estimator

Let $\theta(\gamma) \equiv \mathbb{E}[m(W, \gamma)]$ so that $\theta_0 = \theta(\gamma_0)$. Rather than relying on the plug-in moment $m(W, \hat{\gamma})$, the DML literature suggests to use Neyman orthogonal score which accounts for the influence of first-step estimation error in $\hat{\gamma}$ on $\theta(\gamma)$. Following Chernozhukov et al. (2022c), the orthogonal score takes the form

$$\psi(W, \theta, \gamma, \alpha) \equiv m(W, \gamma) - \theta + \alpha(X) \left(Y - \gamma(X)\right) \ , \tag{2}$$

which the function α_0 represents the Riesz representer of functional $\theta(\gamma)$. An extended discussion on orthogonality is included in Appendix C. For the Riesz representer to exist, we impose the following restrictions on the moment function.

Assumption 2.1 (Moment function)**.** *For the moment function m in (1), the map $\gamma \mapsto m(W, \gamma)$ satisfies: (i) well-definedness: $\|\mathbb{E}[m(W, \gamma)]\| < \infty$ for all $\gamma \in L^2(P_{0,X})$; (ii) linear in mean: $\gamma \mapsto \mathbb{E}[m(W, \gamma)]$ is linear in $L^2(P_{0,X})$; and (iii) local mean-square continuity at γ_0: $\mathbb{E}[\{m(W, \gamma) - m(W, \gamma_0)\}^2] \to 0$ as $\|\gamma - \gamma_0\|_{L^2} \to 0$ for all $\gamma \in L^2(P_{0,X})$.*

Assumption 2.1(i) ensures that the moment condition in (1) is well-defined. Assumptions 2.1(ii)(iii) specialize the framework to functionals $\theta(\gamma)$ which admit a special inner-product form, thereby facilitates a simple construction of the orthogonal score as suggested in Chernozhukov et al. (2022c). In particular, if Assumption 2.1 is satisfied, then according to the Riesz representation theorem there exist a unique $\alpha_0 \in L^2(P_{0,X})$ in which

$$\mathbb{E}[m(W, \gamma)] = \mathbb{E}[\alpha_0(X)\gamma(X)] \ , \quad \forall \gamma \in L^2(P_{0,X}) \ . \tag{3}$$

We refer to this α_0 as the Riesz representer (of the functional $\gamma \mapsto \mathbb{E}[m(W, \gamma)]$).

The class of functionals implied by Assumptions 2.1(ii)(iii) has been widely studied in the DML literature under the name automatic debiased machine learning (Auto-DML), including in Chernozhukov et al. (2022b,c, 2024); van der Laan et al. (2025). This framework covers a number of parameters that are important in economic applications (see, e.g., Example 1-4 in Chernozhukov et al. (2022c)). To motivate our setup, we provide one of the examples here.

Example 2.1 (Average treatment effect)**.** *Let $X = (D, Z^\intercal)^\intercal$, in which $D \in \{0, 1\}$ being the binary treatment indicator, and $Z \in \mathbb{R}^{d_Z}$ are covariates. The parameter of interest*

is

$$\theta_0 = \mathbb{E}\left[\gamma_0(1, Z) - \gamma_0(0, Z)\right] \tag{4}$$

with $\gamma_0(d, z) = \mathbb{E}[Y|D = d, Z = z]$. *Under standard assumptions,*[5] θ_0 *is the average treatment effect (Rosenbaum and Rubin, 1983). The expectation in* (4) *can be written into an inner-product form so that*

$$\theta_0 = \mathbb{E}[\alpha_0(X)\gamma_0(X)] \ , \quad \textit{with}$$

$$\alpha_0(X) = \left\{ \frac{D}{Pr(D = 1|Z)} - \frac{1 - D}{1 - Pr(D = 1|Z)} \right\} , \tag{5}$$

in which the Riesz representer α_0 *is the difference between two Horvitz-Thompson weights (Horvitz and Thompson, 1952).*

Let $X_i \equiv (X_{i1}, \ldots, X_{id_X})^\intercal \in \mathcal{X} \subset \mathbb{R}^{d_X}$ and $W_i \equiv (Y_i, X_i^\intercal)^\intercal \in \mathcal{Y} \times \mathcal{X} \subset \mathbb{R}^{1+d_X}$. In accordance with Chernozhukov et al. (2022c), the parameter of interest θ_0 defined in (1) can be estimated by the Auto-DML estimator

$$\hat{\theta}_n = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \{ m(W_i, \hat{\gamma}_{-k}) + \hat{\alpha}_{-k}(X_i) \left(Y_i - \hat{\gamma}_{-k}(X_i) \right) \} \ , \tag{6}$$

where with some positive integer constant $K > 1$, $\bigcup_{k=1}^{K} I_k$ is a partition of the sample index set $\{1, \ldots, n\}$ and we denote $n_k \equiv |I_k|$. As in the literature, for each fold $k \in \{1, \ldots, K\}$, the first-step learners $\hat{\alpha}_{-k} \in \mathcal{A}_{n_{-k}}$ and $\hat{\gamma}_{-k} \in \mathcal{G}_{n_{-k}}$ are trained only on $\{W_i : i \in I_{-k}\}$, with $I_{-k} \equiv \{1, \ldots, n\} \setminus I_k$, with $n_{-k} \equiv |I_{-k}|$, and evaluated only on $\{W_i : i \in I_k\}$. This implies that during estimation, at each fold k, the orthogonal moment function is averaged over samples that are different from those used to construct $\hat{\alpha}_{-k}$ and $\hat{\gamma}_{-k}$. This procedure, known as cross-fitting, is a key component of DML, though it is not the focus of this paper. Further discussion on cross-fitting can be found in Chernozhukov et al. (2018).

2.2 Asymptotic Normality under High-Level Conditions

High-level conditions for asymptotic normality of the Auto-DML estimator are well established in the literature, with Chernozhukov et al. (2022c) providing the canonical treatment. We restate them in here for completeness.

Assumption 2.2 (Sampling and sample-splitting)**.** *The Auto-DML estimator* $\hat{\theta}_n : \prod_{k=1}^{K} \mathcal{A}_{n_{-k}} \times \prod_{k=1}^{K} \mathcal{G}_{n_{-k}} \times \{W_i\}_{i=1}^{n} \to \mathbb{R}$ *is constructed as in* (6) *with: (i)* $\{W_i\}_{i=1}^{n}$ *being an i.i.d. sample from the law of* W*; (ii) the total number of folds* K *is fixed;*

[5] These are, with $Y(1)$ and $Y(0)$ denoting potential outcomes with and without treatment, (i) the stable unit treatment value assumption (SUTVA): $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$; (ii) overlap: $0 < Pr(D = 1|z) < 1$, for all z in the support of Z; and (iii) selection on observables: $Y(0), Y(1) \perp\!\!\!\perp D|Z$.

(iii) the data split $\{I_k\}_{k=1}^K$ *is random and independent of the sample* $\{W_i\}_{i=1}^n$*; and (iv)* $\min_{1\leq k\leq K} n_k \to \infty$ *with* $\max_{1\leq k\leq K}\{n_k/n_{-k}\} \to c < \infty$ *as* $n \to \infty$.

Assumption 2.3 (Bounded second moment for the moment function)**.** $\mathbb{E}[m(W,\gamma)^2] < \infty$ *for all* $\gamma \in L^2(P_{0,X})$*.*

Assumption 2.4 (Bounded conditional variance and bounded Riesz representer)**.** *(i)* $\mathbb{E}[\{Y-\gamma_0(X)\}^2|X] \in L^\infty(P_{0,X})$*; and (ii)* $\alpha_0 \in L^\infty(P_{0,X})$*.*

Assumption 2.2 restricts the data-generating process and the cross-fitting scheme. Part (i) of Assumption 2.2 rules out dependence across observations within and across folds, while part(iii) of Assumption 2.2 ensures that data used to evaluate the estimator is independent of those used to construct it. Together, Assumptions 2.2(i)(iii) form the backbone of cross-fitting. On the other hand, part(ii) and (iv) of Assumption 2.2 force the training sample size to grow with the overall sample size, which is important for consistent estimation of the nuisances.

Assumptions 2.3 and 2.4 are standard regularity assumptions in the Auto-DML literature. Assumption 2.3 is needed to keep the asymptotic variance finite. In Assumption 2.4, essential boundedness is imposed on the conditional variance and the Riesz representer. This allows the rate requirements on the nuisance estimators to be stated in the L^2-norm (see Assumption 2.5 below), which facilitates verifications on machine learning first steps, for which mean-square convergence rates are commonly derived in the literature.

Assumption 2.5 (L^2-convergence rate on first-step estimators)**.** *For all* $k = 1,\ldots,K$*, the first-step estimators satisfy: (i) consistency:* $\|\hat{\alpha}_{-k}-\alpha_0\|_{L^2} = o_p(1)$ *with* $\|\hat{\gamma}_{-k}-\gamma_0\|_{L^2} = o_p(1)$*; and (ii) the product rate:* $\|\hat{\alpha}_{-k}-\alpha_0\|_{L^2}\|\hat{\gamma}_{-k}-\gamma_0\|_{L^2} = o_p(n_{-k}^{-1/2})$ *as* $n_{-k} \to \infty$*.*

We separate Assumption 2.5 from the above assumptions to underscore its importance in this paper — one of our paper's focus is to provide valid tuning in the sense that the product rate in Assumption 2.5(ii) can be satisfied. Also note that in Chernozhukov et al. (2018), when they considered a more general case of linear scores, both $\hat{\alpha}_{-k}$ and $\hat{\gamma}_{-k}$ are required to converge faster than $n_{-k}^{-1/4}$. However, under our specific setup pinned down by (1) and Assumption 2.1, a product rate of $n_{-k}^{-1/2}$ (Assumption 2.5(ii)) is sufficient.

Lemma 2.1 (Asymptotic linearization and normality)**.** *Suppose for every fold* $k \in \{1,\ldots,K\}$*,* $\hat{\alpha}_{-k},\hat{\gamma}_{-k} \in L^2(P_{0,X})$*, and Assumptions 2.1, 2.2, 2.4, and 2.5 hold. Then,*

$$\sqrt{n}(\hat{\theta}_n-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\{m(W_i,\gamma_0)-\theta_0+\alpha_0(X_i)(Y_i-\gamma_0(X_i))\}+o_p(1)\ , \tag{7}$$

Additionally, if Assumption 2.3 is also satisfied, then if $\sigma \equiv \{\mathbb{E}[\{m(W,\gamma_0) - \theta_0 + \alpha_0(X)(Y - \gamma_0(X))\}^2]\}^{1/2} > 0$,

$$\sigma^{-1}\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0,1) \ .$$

3 Ridge-Regularized Sieve First Steps

We now turn to the nuisance components, which will be the focus of this and the following sections. Here we detail the linear sieve estimators for α_0 and γ_0, and establish their convergence rates. We state the conditions underlying these results in a way that makes explicit how the convergence rates depend on: (i) the sieve dimensions and (ii) the ridge penalties, as these tuning parameters are the objects of analysis in this paper. This section is not among the paper's contributions. Instead, results established below serve as inputs to the analysis that follows.

3.1 Dictionary and the Approximation Space

Let $(b_j)_{j\geq 0}$ be a predetermined dictionary of real-valued functions defined on $\mathcal{X}$, with each dictionary component satisfying $b_j \in L_2(P_{0,X})$ and we fix $b_0(x) \equiv 1$ for all $x \in \mathcal{X}$. For $p = 1, 2, \ldots$, define

$$\Gamma^{(p)} \equiv \text{span}\{b_0, b_1, \ldots, b_p\} \ . \tag{8}$$

We let the dimension p grow with the sample size and define the limiting approximation space to be

$$\Gamma \equiv \overline{\bigcup_{p=0}^{\infty} \Gamma^{(p)}} = \lim_{n\to\infty} \Gamma^{(p_n)} \ . \tag{9}$$

Finally, define

$$\bar{\gamma} \equiv \arg\min_{\gamma\in\Gamma} \mathbb{E}\big[(Y - \gamma(X))^2\big] \quad \text{and} \quad \bar{\alpha} \equiv \arg\min_{\alpha\in\Gamma} \mathbb{E}\big[(\alpha_0(X) - \alpha(X))^2\big]$$

to be the $L_2(P_{0,X})$ projections of the outcome Y and the Riesz representer α_0 onto Γ, respectively. If $(b_j)_{j\geq 0}$ is chosen in a way such that Γ is rich enough to cover the truth (i.e., $\gamma_0, \alpha_0 \in \Gamma \subset L^2(P_{0,X})$), then

$$\bar{\gamma} = \gamma_0 \quad \text{and} \quad \bar{\alpha} = \alpha_0 \ . \tag{10}$$

In other words, if $(b_j)_{j\geq 0}$ is complete in $L^2(P_{0,X})$, then $\Gamma = L^2(P_{0,X})$ and so (10) holds. We introduce below a set of conditions on the dictionary.

Assumption 3.1 (Dictionary)**.** *The shared dictionary of functions used for estimating both* γ_0 *and* α_0 *satisfies: (i) for all indices* $j \geq 1$, $b_j \in L^2(P_{0,X})$*; (ii) for each integer* $j \geq$

1, $\mathbb{E}[b_j(X)] = 0$*; and (iii)* $\gamma_0, \alpha_0 \in \Gamma \subset L^2(P_{0,X})$*, with* Γ *being the limiting approximation space generated by* $(b_j)_{j\geq 0}$ *as defined in* (9)*.*

In here, Assumption 3.1(iii) ensures that the identification equalities in (10) holds. On the other hand, Assumption 3.1(i) is to make sure that the limiting approximation space Γ is a subspace of $L_2(P_{0,X})$, which is a necessary condition for Assumption 3.1(iii) since $\gamma_0, \alpha_0 \in L^2(P_{0,X})$. Finally, Assumption 3.1(ii) is simply a normalization to separate the population intercept. In this paper, we focus on the case in which a common dictionary is employed for both estimators. This is indeed without loss of generality since the common dictionary can be the union of two different ones.

3.2 Riesz Regression

The orthogonal moment introduced in Section 2.1 performs bias correction through an essential element — the Riesz representer α_0. Having discussed the sieve space in Section 3.1, we are now ready to characterize the finite-dimensional approximation of α_0 in $\Gamma^{(p)}$. Specifically, in this subsection we focus on

$$\bar{\alpha}^{(p)} \equiv \underset{\alpha\in\Gamma^{(p)}}{\arg\min}\, \mathbb{E}\big[(\alpha_0(X) - \alpha(X))^2\big] \,. \tag{11}$$

Define $b^{(p)} \equiv (b_1, \ldots, b_p)^\intercal : \mathcal{X} \to \mathbb{R}^p$, where $(b_1, \ldots, b_p)$ are the first p functions in the dictionary. Since $\Gamma^{(p)}$ is the linear span of $\{b_0, b_1, \ldots, b_p\}$, the minimization problem in (11) is equivalent to finding $(\bar{\rho}_0, \bar{\rho}^{(p)})$ in which

$$(\bar{\rho}_0, \bar{\rho}^{(p)}) = \underset{\rho_0\in\mathbb{R},\, \rho^{(p)}\in\mathbb{R}^p}{\arg\min}\, \mathbb{E}\left[\left(\alpha_0(X) - \rho_0 - b^{(p)}(X)^\intercal\rho^{(p)}\right)^2\right] \,. \tag{12}$$

Based on (3), observe that

$$\begin{pmatrix} \mathbb{E}[\alpha_0(X)] \\ \mathbb{E}[\alpha_0(X)b_1(X)] \\ \vdots \\ \mathbb{E}[\alpha_0(X)b_p(X)] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[m\,(W, 1)] \\ \mathbb{E}[m\,(W, b_1(X))] \\ \vdots \\ \mathbb{E}[m\,(W, b_p(X))] \end{pmatrix} \,.$$

Therefore, first-order condition of (12) allows us to characterize $\bar{\alpha}^{(p)}$ through the linear system

$$\begin{pmatrix} M^0 \\ M^{(p)} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \Sigma^{(p)} \end{pmatrix} \begin{pmatrix} \bar{\rho}_0 \\ \bar{\rho}^{(p)} \end{pmatrix} \,, \tag{13}$$

in which

$$\Sigma^{(p)} \equiv \mathbb{E}[b^{(p)}(X)b^{(p)}(X)^\intercal]$$

is the Gram matrix while the moment vectors are defined through $M^0 \equiv \mathbb{E}[m(W,1)]$ and

$$M^{(p)} \equiv \mathbb{E}\left[m_{b^{(p)}}(W)\right], \quad \text{with} \quad m_{b^{(p)}}(W) \equiv \begin{pmatrix} m\left(W, b_1(X)\right) \\ \vdots \\ m\left(W, b_p(X)\right) \end{pmatrix}. \tag{14}$$

We refer to (13) as the Riesz equation, and our goal is to solve for $(\bar{\rho}_0, \bar{\rho}^{(p)})$. By exploiting the special structure in (3), Riesz regression allows us to estimate the Riesz representer without knowing its analytical form.

3.3 Spectral Properties of the Gram Matrix

Solving for the coefficients in both the outcome and the Riesz regressions requires inverting the Gram matrix. We therefore devote this subsection to describe the conditions we maintain on the Gram matrix's spectral properties. Since $\Sigma^{(p)}$ is a real symmetric matrix, it admits an orthonormal basis of eigenvectors $\{\phi_j^{(p)}\}_{j=1}^p$ in $\mathbb{R}^p$. Together with the corresponding set of eigenvalues $\{\nu_j^{(p)}\}_{j=1}^p$, we can write

$$\Sigma^{(p)} = \sum_{j=1}^{p} \nu_j^{(p)} \phi_j^{(p)} \left(\phi_j^{(p)}\right)^{\mathsf{T}} .$$

Without loss of generality, assume that the eigenpairs are ordered so that $\nu_1^{(p)} \geq \nu_2^{(p)} \geq \ldots \geq \nu_p^{(p)}$. Note that since $\Sigma^{(p)}$ is positive semidefinite, its spectral norm equals its largest eigenvalue. That is, $\|\Sigma^{(p)}\|_{\text{op}} = \nu_1^{(p)}$.

Assumption 3.2 (Spectral properties)**.** *(i) There exist a non-increasing function $\varphi : \mathbb{N} \to (0,\infty)$ with $\varphi(j) \to 0$ as $j \to \infty$, such that for every $p \geq 1$ and $1 \leq j \leq p$, $\nu_j^{(p)} \lesssim \varphi(j)$; (ii) $\nu_1^{(p)} \gtrsim 1$; and (iii) for every finite p, $\nu_p^{(p)} > 0$.*

In machine learning, predictive performance is often prioritized over interpretability. Consistent with this objective, our primary concern on the nuisances in this paper is the estimator's L^2-convergence rate rather than inference on individual coefficients. Consequently, collinearity is less of a concern when deciding whether to include additional dictionary elements, and practitioners may therefore be inclined to use a richer dictionary. Nevertheless, as the dictionary expands, successive elements are likely to contribute progressively less independent variation. Assumption 3.2(i) formalizes this phenomenon. Such kind of decay induces ill-conditioning when p gets large, which motivates the use of ridge penalty. To keep the analysis general, we use an upper envelope function rather than committing to a specific form of spectral decay, such as polynomial or exponential decay.

Assumption 3.2(ii), in turn, is a standard regularity condition ensuring that the Gram matrix does not degenerate. Finally, Assumption 3.2(iii) is imposed to simplify

the notation and analysis, it can be dropped by replacing the inverse on the population Gram matrix by its Moore-Penrose inverse.

3.4 Smoothness of the Nuisances Relative to the Gram Matrix

Having specified the spectral properties of the Gram matrix in Section 3.3, we now complete the picture of the regression problems by stating how the nuisances' coefficients align with the eigenvectors of the Gram matrix. With $M^{(p)}$ as defined in (14), and defining the cross-moment

$$C^{(p)} \equiv \mathbb{E}\left[b^{(p)}(X)Y\right] \in \mathbb{R}^{p\times 1} ,$$

we impose the following source conditions.

Assumption 3.3 (Source condition)**.** *There exist constants $s_\alpha, s_\gamma \geq 1/2$ and vectors $h_\alpha^{(p)}, h_\gamma^{(p)} \in \mathbb{R}^{p\times 1}$ such that,*

$$(i) \qquad M^{(p)} = \left(\Sigma^{(p)}\right)^{s_\alpha+1} h_\alpha^{(p)} \quad \forall p \geq 1 , \quad \textit{with} \quad \sup_{p\geq 1} \left\| h_\alpha^{(p)} \right\|_2 < \infty ,$$

$$(ii) \qquad C^{(p)} = \left(\Sigma^{(p)}\right)^{s_\gamma+1} h_\gamma^{(p)} \quad \forall p \geq 1 , \quad \textit{with} \quad \sup_{p\geq 1} \left\| h_\gamma^{(p)} \right\|_2 < \infty .$$

Assumption 3.3 states that the nuisances' coefficients,

$$\bar{\rho}^{(p)} = \left(\Sigma^{(p)}\right)^{-1} M^{(p)} = \left(\Sigma^{(p)}\right)^{s_\alpha} h_\alpha^{(p)} \quad \text{(for Riesz regression)} \quad \text{and}$$

$$\bar{\beta}^{(p)} = \left(\Sigma^{(p)}\right)^{-1} C^{(p)} = \left(\Sigma^{(p)}\right)^{s_\gamma} h_\gamma^{(p)} \quad \text{(for outcome regression)},$$

are given by powers of $\Sigma^{(p)}$ applied to bounded vectors, with s_α and s_γ setting the exponents. Paired with the non-increasing upper envelope function φ in Assumption 3.2, this translates into the coefficients' decay pattern along higher-order eigendirections. At the same time, the uniform bound imposed on the $h^{(p)}$ vectors is essential for bounding the convergence rate of the estimator uniformly in p.

The source condition is a standard regularity condition when analyzing ill-posed inverse problems, which is often linked to the smoothness of the target function. Related discussions can be found in Chen and Pouzo (2012), pp. 301–302, in which they described how the source condition can be translated to a Sobolev-type restriction. In Carrasco et al. (2007), the source condition is termed as β-regularity in their Definition 3.4. In order to accommodate the sieve setup, our version of source condition differs from them by considering it at each $p \geq 1$. Besides, we also choose to restrict our attention to $s_\alpha, s_\gamma \geq 1/2$, so that the ridge estimators are saturated in the sense that further smoothness cannot improve their convergence rates. Within the saturated regime $s \geq$

1/2, the convergence rate no longer varies across s. This determines the optimal penalty rates without referencing to s, and the only thing left for our data-driven process to choose is the constant.

Instead of imposing the source condition, one can also choose to put assumptions on the effective dimensions. With penalty parameter λ, the effective dimension at $p \geq 1$ is

$$\mathcal{N}^{(p)}(\lambda) = \sum_{j=1}^{p} \frac{\nu_j^{(p)}}{\nu_j^{(p)} + \lambda} \ .$$

We deliberately avoid it because conditions imposed on the effective dimensions place implicit restrictions on the penalty parameters. By leaving them unrestricted, we can make the dependence of the resulting convergence rates on the penalty parameters explicit.

3.5 Learning Rates

When $\hat{\alpha}_{-k}$ and $\hat{\gamma}_{-k}$ are estimated through ridge regressions, estimator of the Riesz representer takes the following forms. With

$$\tilde{b}_{-k}^{(p)}(x) \equiv b^{(p)}(x) - \frac{1}{n_{-k}} \sum_{i \in I_{-k}} b^{(p)}(X_i) \ ,$$

we have

$$\begin{aligned} \hat{\alpha}_{-k}(x; p, \zeta) &= \frac{1}{n_{-k}} \sum_{i \in I_{-k}} m(W_i, 1) + \tilde{b}_{-k}^{(p)}(x)^\intercal \hat{\rho}_{-k}(b, p, \zeta) \ , \\ \hat{\rho}_{-k}(p, \zeta) &= \left(\frac{1}{n_{-k}} \sum_{i \in I_{-k}} \tilde{b}_{-k}^{(p)}(X_i) \tilde{b}_{-k}^{(p)}(X_i)^\intercal + \zeta I_p \right)^{-1} \left(\frac{1}{n_{-k}} \sum_{i \in I_{-k}} m_{\tilde{b}_{-k}^{(p)}}(W_i) \right) \ , \end{aligned} \tag{15}$$

while the regression learner can be expressed as

$$\begin{aligned} \hat{\gamma}_{-k}(x; p, \lambda) &= \frac{1}{n_{-k}} \sum_{i \in I_{-k}} Y_i + \tilde{b}_{-k}^{(p)}(x)^\intercal \hat{\beta}_{-k}(b, p, \lambda) \ , \\ \hat{\beta}_{-k}(p, \lambda) &= \left(\frac{1}{n_{-k}} \sum_{i \in I_{-k}} \tilde{b}_{-k}^{(p)}(X_i) \tilde{b}_{-k}^{(p)}(X_i)^\intercal + \lambda I_p \right)^{-1} \left(\frac{1}{n_{-k}} \sum_{i \in I_{-k}} \tilde{b}_{-k}^{(p)}(X_i) Y_i \right) \ . \end{aligned} \tag{16}$$

In the sense of first-step estimations, $\{Y_i, m(W_i, 1), \tilde{b}_{-k}^{(p)}(X_i) : i \in I_{-k}\}$ serves as the training set, whereas $\tilde{b}_{-k}^{(p)}(x)$ is the test vector. Throughout this paper, we take dictionary b as exogenously specified, and focus our attention on the choice of penalty parameters $\zeta, \lambda \in \mathbb{R}_{>0}$, and sieve dimensions $p_\gamma, p_\alpha \in \mathbb{Z}_{>0}$. Therefore, we suppress the estimators'

dependence on b and write simply $\hat{\alpha}_{-k}(x; p_\alpha, \zeta)$, and $\hat{\gamma}_{-k}(x; p_\gamma, \lambda)$.

After defining the ridge estimators in (15) and (16), we introduce an additional set of assumptions needed to establish their learning rates.

Assumption 3.4 (Pointwise linearity)**.** *The map $\gamma \mapsto m(w, \gamma)$ is linear on $L^2(P_{0,X})$ for each $w \in \mathcal{Y} \times \mathcal{X}$.*

Assumption 3.5 (Expected maximal row-norm bound)**.** *For each fold $k \in \{1, \ldots, K\}$, we have $\mathbb{E}\left[\max_{1 \le i \le n_{-k}} \|b^{(p)}(X_i)\|_2^2\right] \lesssim p$ for each $p \ge 1$.*

Assumption 3.6 (Uniformly bounded second moments)**.**

$$(i) \quad \sup_{j \ge 1} \mathbb{E}\left[(b_j(X)Y)^2\right] < \infty \ ; \ \textit{and} \quad (ii) \quad \sup_{j \ge 1} \mathbb{E}\left[m(W, b_j)^2\right] < \infty \ .$$

Assumption 3.4 is used in the learning-rate analysis of the Riesz representer estimator, in which dictionary in the empirical moment vector is centered using training-fold means. It is satisfied by our Example 2.1. Combining with Assumption 2.1(iii), it implies L^2-Lipschitz continuity of the moment function, which its statement and proof is included in Appendix B. Assumption 3.5 is imposed so that we can make use of the Rudelson's inequality (Rudelson, 1999) to establish a concentration bound for the Gram matrix. This allows us to avoid imposing strong tail conditions, such as sub-Gaussianity, on the features. These conditions can be restrictive in economic applications. For example, income data are typically heavy-tailed. The Rudelson's inequality is also used in Newey and Robins (2018) for their spline regression estimator. Later when we develop the cross-validation results in Section 5, this assumption will be strengthened to uniformly bounded dictionary-row norm. This stronger assumption, however, is not needed for learning-rate result below. Assumption 3.6 is a regularity assumption needed to bound the sum of the variance terms for the vector concentration in Lemma B.1 (ii)(iii). Next, with $\Gamma^{(p)}$ defined in (8), recall

$$\bar{\alpha}^{(p)} \equiv \underset{\alpha \in \Gamma^{(p)}}{\arg\min}\ \mathbb{E}\big[(\alpha_0(X) - \alpha(X))^2\big] \ ,$$

and let

$$\bar{\gamma}^{(p)} \equiv \underset{\gamma \in \Gamma^{(p)}}{\arg\min}\ \mathbb{E}\big[(\gamma_0(X) - \gamma(X))^2\big] \ ,$$

which represent the best L^2-approximations of α_0 and γ_0 in the p-dimensional approximation space.

Assumption 3.7 (Vanishing approximation error)**.** *There exist vanishing sequences $\delta_\alpha(p)$ and $\delta_\gamma(p)$ such that $\delta_\alpha(p), \delta_\gamma(p) \to 0$ as $p \to \infty$, and: (i) $\|\alpha_0 - \bar{\alpha}^{(p)}\|_{L^2} \lesssim \delta_\alpha(p)$; and (ii) $\|\gamma_0 - \bar{\gamma}^{(p)}\|_{L^2} \lesssim \delta_\gamma(p)$.*

Note that the conditions stated in Assumption 3.7 are directly implied by Assumption 3.1(iii). We state them separately here to introduce the vanishing sequences $\delta_\alpha(p)$

and $\delta_\gamma(p)$, which will then enter the learning rates to be established in Lemma 3.1. Later in Assumption 4.2, we will characterize how fast these sequences decay as a function of the sieve size.

Assumption 3.8 (The asymptotic regime)**.** *The sieve size grows with the training sample size. That is, $p_{\gamma,n_{-k}}, p_{\alpha,n_{-k}} \to \infty$ as $n_{-k} \to \infty$. Specifically, for $j \in \{\alpha, \gamma\}$ and $k \in \{1, \dots, K\}$, there exist $\tau_j \in (0,1)$ such that $p_{j,n_{-k}} \log p_{j,n_{-k}} \asymp n_{-k}^{\tau_j}$.*

Assumption 3.9 (Vanishing penalties)**.** *For each $k \in \{1, \dots, K\}$, the penalty parameters are vanishing such that $\zeta_{n_{-k}}, \lambda_{n_{-k}} \to 0$ as $n_{-k} \to \infty$. Specifically, as $n_{-k} \to \infty$, we have:*

$$(i) \quad \sqrt{\frac{p_{\alpha,n_{-k}} \log p_{\alpha,n_{-k}}}{n_{-k}}} = o\left(\zeta_{n_{-k}}\right); \quad \textit{and} \qquad (ii) \quad \sqrt{\frac{p_{\gamma,n_{-k}} \log p_{\gamma,n_{-k}}}{n_{-k}}} = o\left(\lambda_{n_{-k}}\right) .$$

Assumption 3.8 demands for $j \in \{\alpha, \gamma\}$,

$$\frac{p_{j,n_{-k}} \log p_{j,n_{-k}}}{n_{-k}} \to 0 \quad \text{as} \quad n_{-k} \to \infty .$$

Assumptions 3.8 (combined with the concentration results established in Lemma B.1) and 3.9 are important ingredients for establishing the consistency of the ridge estimators, which is one of the high-level requirements for the asymptotic normality of the Auto-DML estimator stated in Assumption 2.5(i).

Lemma 3.1 (Learning rates of ridge estimated first steps)**.** *Suppose Assumptions 2.2(i)(iii), 2.4(i), 3.1, 3.2, 3.3(ii), 3.5, 3.6(i), 3.7(ii), 3.8, and 3.9(ii) hold. For each $k \in \{1, \dots, K\}$, as $n_{-k} \to \infty$, we have*

$$\left\|\hat{\gamma}_{-k}(p_{\gamma,n_{-k}}, \lambda_{n_{-k}}) - \gamma_0\right\|_{L^2}^2 = O_p\left(\frac{1}{\lambda_{n_{-k}}} \cdot \frac{p_{\gamma,n_{-k}} \log p_{\gamma,n_{-k}}}{n_{-k}} + \lambda_{n_{-k}}^2 + \delta_\gamma^2(p_{\gamma,n_{-k}})\right) . \tag{17}$$

If, in addition, Assumptions 2.1 and 3.4 holds and Assumptions 2.4(i), 3.3(ii), 3.6(i), 3.7(ii), and 3.9(ii) are replaced by Assumptions 2.3, 3.3(i), 3.6(ii), 3.7(i), and 3.9(i) respectively, then for each fold $k \in \{1, \dots, K\}$, as $n_{-k} \to \infty$,

$$\left\|\hat{\alpha}_{-k}(p_{\alpha,n_{-k}}, \zeta_{n_{-k}}) - \alpha_0\right\|_{L^2}^2 = O_p\left(\frac{1}{\zeta_{n_{-k}}} \cdot \frac{p_{\alpha,n_{-k}} \log p_{\alpha,n_{-k}}}{n_{-k}} + \zeta_{n_{-k}}^2 + \delta_\alpha^2(p_{\alpha,n_{-k}})\right) . \tag{18}$$

4 Feasibility in Satisfying the Product-Rate Requirement

Recall that in order to have valid inference backed by Lemma 2.1, one of the important ingredients is to have the first-step estimators satisfying a product rate stated in

Assumption 2.5(ii). That is, as $n_{-k} \to \infty$,

$$\|\hat{\alpha}_{-k} - \alpha_0\|_{L^2} \|\hat{\gamma}_{-k} - \gamma_0\|_{L^2} = o_p\left(n_{-k}^{-1/2}\right) . \tag{19}$$

This, depends on both (i) how well the estimator performs and (ii) how difficult the underlying problem is. In this section, we discuss these two factors by considering the stochastic and deterministic components of the L^2-rates in (17)-(18), respectively.

4.1 Parametrizing the components of the learning rates

Building on Lemma 3.1, we can now determine the optimal rate at which the two penalty parameters should converge to zero. By matching the orders of the first two terms in (17) and (18), which correspond to the squared bias and the variance of the estimator, we obtain the following optimal rates for the penalty parameters.

Assumption 4.1 (Optimal decay rate for penalty parameters)**.**

$$\zeta_{n_{-k}} \asymp \left(\frac{p_{\alpha,n_{-k}} \log p_{\alpha,n_{-k}}}{n_{-k}}\right)^{1/3} , \text{ and } \quad \lambda_{n_{-k}} \asymp \left(\frac{p_{\gamma,n_{-k}} \log p_{\gamma,n_{-k}}}{n_{-k}}\right)^{1/3} .$$

Again, by confining our attention to the saturated ridge regime (by restricting $s_\alpha, s_\gamma \geq 1/2$ in Assumption 3.3), the optimal rates in here are independent of unknown parameters s_α and s_γ. With Assumption 4.1, orders of the stochastic terms in (17) and (18) are now pinned down at

$$O_p\left(\left(\frac{p_{f,n_{-k}} \log p_{f,n_{-k}}}{n_{-k}}\right)^{2/3}\right), \quad f \in \{\alpha, \gamma\} .$$

Before proceeding, let us also parametrize the vanishing rates for the sieve approximation errors, which constitute the deterministic part of the convergence rates.

Assumption 4.2 (Vanishing rate for sieve approximation error)**.** *The vanishing sequences $\delta_\alpha^2(p)$ and $\delta_\gamma^2(p)$ in Assumption 3.7 satisfy:*

$$(i) \quad \delta_\alpha^2(p) = O\left((p \log p)^{-2\kappa_\alpha}\right) ,$$

$$(ii) \quad \delta_\gamma^2(p) = O\left((p \log p)^{-2\kappa_\gamma}\right) ,$$

as $p \to \infty$, with constants $\kappa_\alpha, \kappa_\gamma > 0$.

Assumption 4.2 is in the same spirit as Assumption 3 of Newey (1997), which requires the approximation error of the sieve to vanish at a certain rate as its dimension grows. We state it in the L^2-norm rather than the sup-norm, since our analysis concerns only the L^2-convergence of the first-step estimators. Besides, in contrast to the conventional

practice of formulating the restriction in form of $O(p^{-2\kappa_0})$, our assumption comes with an extra factor of $(\log p)^{-2\kappa_0}$. This allows us to have a more direct comparison between the rates of the deterministic and the stochastic parts of the estimation error, which the latter comes with an additional log factor arising from the Gram matrix's concentration result (Lemma B.1(i)). Note, however, that for any $0 < \kappa < \kappa_0$, as $p \to \infty$,

$$\frac{p^{-2\kappa_0}}{(p \log p)^{-2\kappa}} = p^{-2(\kappa_0 - \kappa)} (\log p)^{2\kappa} \to 0 \ .$$

Therefore, we have for all $0 < \kappa < \kappa_0$,

$$\delta^2(p) = O\left(p^{-2\kappa_0}\right) \implies \delta^2(p) = o\left((p \log p)^{-2\kappa}\right) \ ,$$

as $p \to \infty$. As a result, our formulation in Assumption 4.2 does not prevent us from using existing results on κ_0 in the literature.

4.2 The Product-Rate Warrant Set

To determine the requirement for the product rate in (19) to hold, we shall focus on the exponents governing the convergence rates. Let

$$\|\hat{\alpha}_{-k} - \alpha_0\|_{L^2} = O_p\left(n_{-k}^{-e_\alpha}\right) \ , \text{ and } \|\hat{\gamma}_{-k} - \gamma_0\|_{L^2} = O_p\left(n_{-k}^{-e_\gamma}\right) \ .$$

The exponents e_α and e_γ will be the main objects of interest throughout the rest of this section.

Lemma 4.1 (Exponents of the learning rates)**.** *If Assumptions 3.8, 4.1, and 4.2 hold, then for $j \in \{\alpha, \gamma\}$,*

$$e_j\left(\tau_j; \kappa_j\right) = \min\left(\tau_j \kappa_j, \frac{1}{3}(1 - \tau_j)\right) \ .$$

It leads us to the below assumption, which guarantees the product rate in (19).

Assumption 4.3 (Product-rate warrant set)**.** *For parameters $\tau_\alpha, \tau_\gamma \in (0,1)$ in Assumption 3.8, and $\kappa_\alpha, \kappa_\gamma > 0$ in Assumption 4.2, $(\kappa_\alpha, \kappa_\gamma, \tau_\alpha, \tau_\gamma) \in \mathcal{W}$, with*

$$\mathcal{W} \equiv \Bigg\{ (\kappa_\alpha, \kappa_\gamma, \tau_\alpha, \tau_\gamma) \in \mathbb{R}^2_{>0} \times (0,1)^2 : \\ \min\left(\tau_\alpha \kappa_\alpha, \frac{1}{3}(1 - \tau_\alpha)\right) + \min\left(\tau_\gamma \kappa_\gamma, \frac{1}{3}(1 - \tau_\gamma)\right) > \frac{1}{2} \Bigg\} \ .$$

Intuitively, κ_α and κ_γ characterize how well the Riesz representer and the outcome regression function can be approximated by a simple dictionary, respectively. On the other hand, τ_α and τ_γ govern the rate at which the corresponding sieve dimension

increases with the training sample size. It is important to emphasize that $(\kappa_\alpha, \kappa_\gamma)$ are structural parameters determined by the intrinsic characteristics of the nuisance components in the problem. Specifically, κ_α depends on the complexity of the target functional, whereas κ_γ depends on the complexity of the outcome model. By contrast, τ_α and τ_γ are tuning parameters set as part of the sieve design.

The product-rate warrant set in Assumption 4.3 formalizes the trade-offs both within and across these nuisance components. For instance, when the Riesz representer is relatively hard to be approximated, the imposed requirement on the outcome regression becomes more stringent. At the same time, the growth rate of the sieve dimension should be chosen in order to balance the ridge estimation error against the sieve approximation error.

There are two points that we would need to emphasize in here. First, the product-rate condition in (19) is sufficient but not necessary to guarantee valid inference (see Remark 4.1 and related discussions in Section 1). Second, the product-rate warrant set $\mathcal{W}$ in Assumption 4.3 is derived from the upper bounds in (17) and (18), which might not be sharp. Therefore, Assumption 4.3 should not be interpreted as a necessary condition for either the product-rate condition or valid inference.

Lemma 4.2 ($\sqrt{n}$-consistent and asymptotic normality under lower-level conditions)**.** *Suppose Assumptions 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 4.1, 4.2 and 4.3 are satisfied. Then with $\sigma > 0$, $\sigma^{-1}\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0,1)$.*

4.3 The Cost of Restricting Common Sieve Growth

In this subsection, we illustrate the importance to respect the distinctions between the two nuisance components when constructing the respective sieve spaces.

Proposition 4.1 (Stronger necessary conditions under same sieve-size growth rate)**.** *To satisfy Assumption 4.3, we need*

$$\min\{\kappa_\alpha, \kappa_\gamma\} > \frac{1}{3} \ , \ \textit{and} \quad \tau_\alpha + \tau_\gamma < \frac{1}{2}. \tag{20}$$

Suppose we restrict $\tau_\alpha = \tau_\gamma = \tau$, then the necessary conditions for Assumption 4.3 to hold are tightened to

$$\min\{\kappa_\alpha, \kappa_\gamma\} > 1 \ , \ \textit{and} \quad \tau < \frac{1}{4} \ .$$

Intuitively, requiring all sieve dimensions to grow at the same rate prevents the estimator from fully leveraging the benefit of "budget sharing" in Auto-DML. Restricting $\tau_\alpha = \tau_\gamma = \tau$ therefore rules out every data-generating process with

$$\frac{1}{3} < \min\{\kappa_\alpha, \kappa_\gamma\} \leq 1 \ .$$

Figure 2 illustrates this idea with the example of $(\kappa_\alpha, \kappa_\gamma) = (3/4, 4)$, in which $\min\{\kappa_\alpha, \kappa_\gamma\} = 3/4 < 1$. In this case, no choice of $\tau = \tau_\alpha = \tau_\gamma$ can reach the warrant area that is established from our stated bounds in (17) and (18).

4.4 Feasibility Frontier for the Product-Rate Warrant Set

Taking one step further, we now characterize the optimal sieve growth rates $(\tau_\alpha^*, \tau_\gamma^*)$ under known $(\kappa_\alpha, \kappa_\gamma)$. Individually, each one of them gives the fastest convergence rate of its per-nuisance L^2-risk. Together, they also give the nuisance-error product its optimal rate.

Lemma 4.3 (Optimal dictionary growth rates)**.** *For each nuisance $j \in \{\alpha, \gamma\}$, given κ_j,*

$$\tau_j^*(\kappa_j) \equiv \arg\max_{\tau_j} \left(e_j\left(\tau_j; \kappa_j\right)\right) = \frac{1}{3\kappa_j + 1} \ , \tag{21}$$

$$e_j^*(\kappa_j) \equiv \max_{\tau_j} \left(e_j\left(\tau_j; \kappa_j\right)\right) = \frac{\kappa_j}{3\kappa_j + 1} \ .$$

Meanwhile, τ_α^* and τ_γ^* in (21) also maximize the product-rate exponent collectively. That is,

$$\left(\tau_\alpha^*(\kappa_\alpha), \tau_\gamma^*(\kappa_\gamma)\right) = \arg\max_{\tau_\alpha, \tau_\gamma} \left(e_\alpha(\tau_\alpha; \kappa_\alpha) + e_\gamma(\tau_\gamma; \kappa_\gamma)\right) \ ,$$

$$e^*(\kappa_\alpha, \kappa_\gamma) \equiv \max_{\tau_\alpha, \tau_\gamma} \left(e_\alpha(\tau_\alpha; \kappa_\alpha) + e_\gamma(\tau_\gamma; \kappa_\gamma)\right) = \sum_{j \in \{\alpha, \gamma\}} \frac{\kappa_j}{3\kappa_j + 1} \ . \tag{22}$$

This benchmark is important, as it traces out a frontier in $(\kappa_\alpha, \kappa_\gamma)$ below which the sliced plane $\mathbf{S}(\kappa_\alpha, \kappa_\gamma) \equiv \{(\tau_\alpha, \tau_\gamma) : (\kappa_\alpha, \kappa_\gamma, \tau_\alpha, \tau_\gamma) \in \mathcal{W}\}$, for $\mathcal{W}$ in Assumption 4.3, is empty. Since our asymptotic theory is backed by the product-rate condition in (19), it is therefore important to account for such limit to avoid invalid inference. Based on the result in (22), we now define the product-rate feasibility frontier

$$f \equiv \left\{(\kappa_\alpha, \kappa_\gamma) \in \mathbb{R}^2_{>0} : \frac{\kappa_\alpha}{3\kappa_\alpha + 1} + \frac{\kappa_\gamma}{3\kappa_\gamma + 1} = \frac{1}{2}\right\} \ . \tag{23}$$

For any problem associated with $(\kappa_\alpha, \kappa_\gamma)$ that sits on or below the frontier, $\mathbf{S}(\kappa_\alpha, \kappa_\gamma)$ is empty. As a result, Assumption 4.3 cannot be satisfied and Lemma 4.2 does not apply. Figure 1 provides a graphical illustration of the frontier.

Remark 4.1 (Product-rate feasibility versus valid-inference feasibility)**.** *We stress that the frontier defined above concerns the feasibility in achieving the product-rate condition used to guarantee valid inference, but not the feasibility in achieving valid inference itself. The product-rate condition, being a sufficient condition to provide valid inference under Lemma 2.1, need not to be a necessary one. Some other estimators that deliver $\sqrt{n}$-inference without requiring such condition are discussed in Section 1. We focus on this*

route as this is one of the key appeals of the DML framework: it allows valid inference to depend solely on the L^2-rates of the nuisances.

Remark 4.2 (Sharpness of the warrant set boundary)**.** *Our product-rate warrant set is established based on our stated bounds in* (17) *and* (18)*, which might not be sharp. Therefore, under the stated assumptions, the condition $(\kappa_\alpha, \kappa_\gamma, \tau_\alpha, \tau_\gamma) \in \mathcal{W}$ is sufficient for the product-rate condition in* (19)*, but we do not claim that it is necessary. As a result, the orange-shaded area in Figure 2 and the combined orange- and blue-shaded area in Figure 1 merely depict the warrant set established from our stated bounds in* (17) *and* (18)*. Boundaries of these areas therefore shall not be interpreted as thresholds beyond which the product-rate condition necessarily fails.*

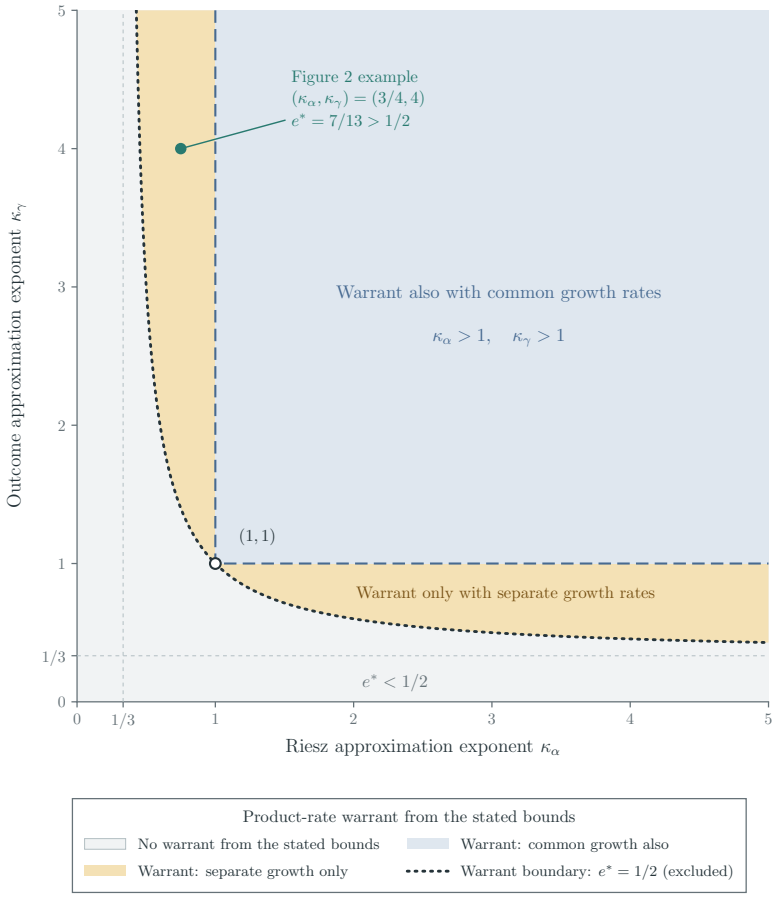


Figure 1: Feasibility Frontier for the Product-Rate Warrant Set

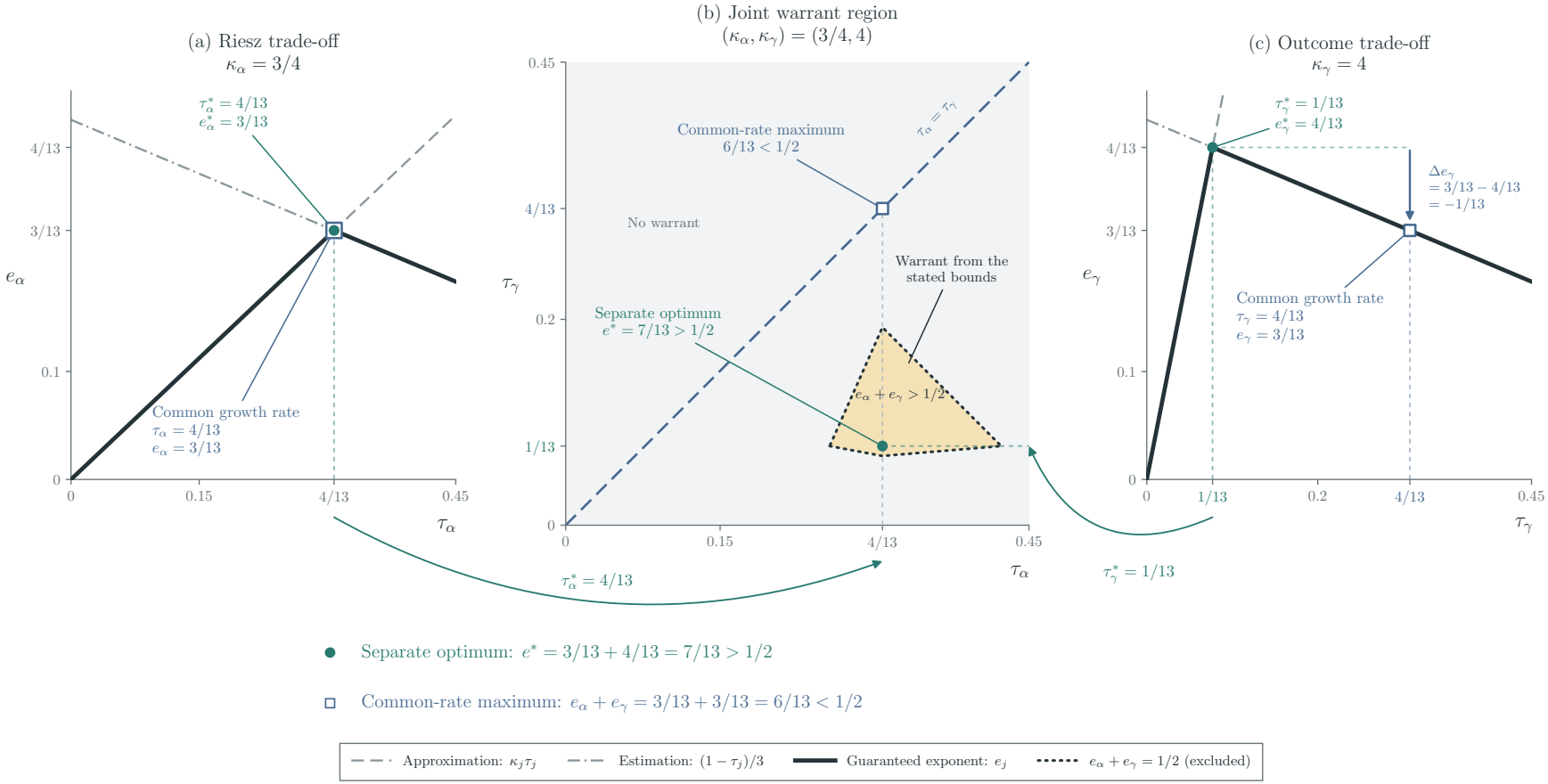


Figure 2: Cost of Restricting Common Sieve Growth Illustrated at $(\kappa_\alpha, \kappa_\gamma) = (3/4, 4)$

As noted in the paragraph following Assumption 4.2, we can use existing results in the literature to make statements about $(\kappa_\alpha, \kappa_\gamma)$, and thus identify the region containing the problem's associated point in Figure 1. However, applying these results requires imposing additional conditions on the dictionary and assumptions about the smoothness of the nuisance functions, which is generally unknown in empirical applications. A formal test that allows practitioners to assess whether a problem lies above the feasibility frontier would therefore be helpful, but developing such a test is beyond the scope of this paper.

5 Prediction-Based Tuning with Cross-Validation

Prediction-based tuning sometimes can conflict with the goal of providing valid inference. This phenomenon occurs in a number of settings, the most famous example being undersmoothing when selecting the bandwidth in kernel density estimation. In contrast to that case, as we can see from Lemma 4.3, while each nuisance's predictive criterion maximizes the absolute value of its exponent to give the fastest rate, the product-rate exponent requirement — the condition needed to support valid inference in the Auto-DML framework — is additively separable in each of these individual parts, and therefore the two goals aligned. In this section, we investigate the use of per-nuisance prediction-based tuning, specifically, the popular cross-validation method, in Auto-DML.

Recall that the first-step learners $\hat{\alpha}_{-k}$ and $\hat{\gamma}_{-k}$ are trained only on $\{W_i : i \in I_{-k}\}$. Therefore, their tuning process should likewise rely solely on the same set of training data. This gives a nested structure in which a V-fold cross-validation procedure is performed within each of the training sample partition built for cross-fitting. From this point onward, for notation simplicity, we assume the total sample size n is divisible by the number of folds K for cross-fitting, and each fold is of equal size. That is, for every fold $k \in \{1, \ldots, K\}$,

$$n^o_{t'} \equiv n_k = n/K, \quad \text{and} \quad n^o_t \equiv n_{-k} = n(K-1)/K \ .^{6}$$

Furthermore, we denote by $\mathbf{T}^{\gamma}_{n^o_t}$ and $\mathbf{T}^{\alpha}_{n^o_t}$ the candidate sets over which the tuning procedure searches.

5.1 The Cross-Validation Procedure

The complete cross-validation procedure considered in this section is summarized in Algorithm 1, which the involved candidate sets are to be discussed later in Section 5.2. Within each cross-fitting fold k, based on the cross-validation partition $\bigcup_{v=1}^{V} I_{-k,v} =$

[6]The notation is to be understood with 'o' as outer (cross-fitting) layer, and 't' as training fold. Counterparts are thus inner (cross-validation) layer, and validation/evaluation fold respectively.

I_{-k}, Algorithm 1 performs V-fold cross-validation separately for each of two first-step estimators. Similar to the sample-splitting rules placed in Assumption 2.2 on the cross-fitting folds, we impose the following conditions on the cross-validation sample partitions. Let $I_{-k,-v} \equiv I_{-k} \setminus I_{-k,v}$.

Assumption 5.1 (Sample partitions for cross-validation)**.** *For each $k \in \{1, \ldots, K\}$, the cross-validation sample partition $\{I_{-k,v}\}_{v=1}^{V}$ of $\{i : i \in I_{-k}\}$ satisfies: (i) the total number of folds V is fixed; (ii) the data split $\{I_{-k,v}\}_{v=1}^{V}$ is random and independent of the sample $\{W_i\}_{i \in I_{-k}}$; and (iii) $\min_{1 \leq v \leq V} |I_{-k,v}| \to \infty$ with $\max_{1 \leq v \leq V} \{|I_{-k,v}|/|I_{-k,-v}|\} \to c < \infty$ as $n \to \infty$.*

Since each cross-validation procedure is conducted within a fixed cross-fitting fold k, all related quantities are computed from the same training sample $\{i : i \in I_{-k}\}$. Given that the procedure does not vary with k, for the remainder of this subsection and the subsections that follow, we suppress the $-k$ subscript and maintain it only when the outer fold matters. For example, notation I_v refers to $I_{-k,v}$, with I_{-v} refers to $I_{-k,-v}$. Similar to the outer cross-fitting folds, in order to simplify our notations, we also assume a shared sample size for all the inner folds. For $k \in \{1, \ldots, K\}$ and $v \in \{1, \ldots, V\}$, we denote

$$n_{t'}^{o'} \equiv |I_{-k,v}| = \frac{n(K-1)}{V \cdot K}, \quad \text{and} \quad n_{t}^{o'} \equiv |I_{-k,-v}| = \frac{n(K-1)(V-1)}{V \cdot K} .$$

While it is intuitive to implement cross-validation in penalized outcome regression, which gives choices

$$\hat{\mathbf{t}}_\gamma \equiv (\hat{p}_\gamma, \hat{\lambda}) = \underset{\mathbf{t}_\gamma \in \mathbf{T}_{n_t^o}^\gamma}{\arg\min} \frac{1}{V} \sum_{v=1}^{V} \frac{1}{n_{t'}^{o'}} \sum_{i \in I_v} \left\{ \left(\frac{1}{n_t^{o'}} \sum_{j \in I_{-v}} Y_j + \tilde{b}^{(p_\gamma)}(X_i)^\intercal \hat{\beta}_{-v}(p_\gamma, \lambda) \right) - Y_i \right\}^2 ,$$

this becomes less trivial on the Riesz representer side. Recall that under the Auto-DML framework, the Riesz representer is estimated through Riesz regression, which obviates the need to specify its analytical form. This poses a problem since it prevents us from evaluating the estimator directly with α_0. In view of this, we follow the practice suggested in Appendix A.1.2 of Chernozhukov et al. (2022c) to tune according to the Riesz loss. The Riesz loss

$$(\alpha; W) \mapsto \alpha(X)^2 - 2m(W, \alpha)$$

comes from the extremum characterization of α_0 provided in Chernozhukov et al. (2024),

which lays the foundation of Riesz Regression. In our case, that is,

$$\hat{\mathbf{t}}_\alpha \equiv (\hat{p}_\alpha, \hat{\zeta}) = \underset{\mathbf{t}_\alpha \in \mathbf{T}^\alpha_{n^o_t}}{\arg\min} \frac{1}{V} \sum_{v=1}^{V} \frac{1}{n^{o'}_{t'}} \sum_{i \in I_v} \left\{ \left(\frac{1}{n^{o'}_t} \sum_{j \in I_{-v}} m(W_j, 1) + \tilde{b}^{(p_\alpha)}(X_i)^\intercal \hat{\rho}_{-v} \right)^2 \right.$$
$$\left. - 2m\left(W_i \, , \, \frac{1}{n^{o'}_t} \sum_{j \in I_{-v}} m(W_j, 1) + \tilde{b}^{(p_\alpha)}(X_i)^\intercal \hat{\rho}_{-v} \right) \right\} .$$

With cross-validated choices $\hat{\mathbf{t}}_\gamma \equiv (\hat{p}_\gamma, \hat{\lambda})$ and $\hat{\mathbf{t}}_\alpha \equiv (\hat{p}_\gamma, \hat{\zeta})$, below we present the oracle inequalities for their resulting estimators' expected fold-average estimation error. To do this, we introduce an additional set of assumptions.

Assumption 5.2 (Bounded dictionary)**.** *There exists a common constant $B > 0$ such that, for every $p \geq 1$,*

$$\operatorname*{ess\,sup}_{x \in \mathcal{X}} \|b^{(p)}(X)\|_2^2 \leq B \ .$$

Assumption 5.3 (Conditional sub-exponential regression error)**.** *There exists $\eta > 0$ such that*

$$\left\| \mathbb{E}\left[e^{\eta |Y - \gamma_0(X)|} | X \right] \right\|_\infty < \infty$$

Assumption 5.4 (Moment bound)**.** *There exists constant $C_\infty \in (0, \infty)$ such that for all $h \in L^\infty(P_{0,X})$,*

$$\|m(W, h)\|_\infty \leq C_\infty \|h\|_\infty \ .$$

Our oracle inequalities build on results in Vaart et al. (2006), which bound the cross-validation remainder using Bernstein numbers of the centered loss functions

$$L^\gamma(W, \gamma) \equiv \{Y - \gamma\}^2 - \{Y - \gamma_0\}^2 \ ,$$

$$L^\alpha(W, \alpha) \equiv \left\{\alpha^2 - 2m(W, \alpha)\right\} - \left\{\alpha_0^2 - 2m(W, \alpha_0)\right\} \ .$$

The definition of a Bernstein pair is provided in Definition B.1 within Appendix B. To ensure that the Bernstein pair of $w \mapsto L^\gamma(w, \gamma)$ exists, together with Assumptions 3.1(ii), 3.2(i), 3.3(ii), Assumption 5.2 strengthens Assumption 3.5 to give $\gamma_0 \in L^\infty(P_{0,X})$ (see Lemma B.6). Assumptions 5.3 and 5.4 help us to pin down a Bernstein pair of $w \mapsto L^\gamma(w, \gamma)$ and $w \mapsto L^\alpha(w, \alpha)$ respectively, which the calculations are provided in Lemma B.4 of Appendix B. For the rest of this section, $\mathbb{E}_{W_{-k}}$ denotes expectation over the outer training sample.

Lemma 5.1 (Oracle inequalities for the nuisance estimators)**.** *If Assumptions 2.1(iii), 2.2(i)(iii), 2.4(ii), 3.1(ii)(iii), 3.2(i), 3.3(ii), 3.4, 5.1(ii), 5.2, 5.3, and 5.4 hold, and for every $v \in \{1, \ldots, V\}$ and $f \in \{\gamma, \alpha\}$,*

$$\hat{f}_{-v}(\mathbf{t}_f) \in L^\infty(P_{0,X}) \qquad \forall \mathbf{t}_f \in \mathbf{T}^f_{n^o_t}$$

then for any $\delta \in (0,1)$,

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\hat{\mathbf{t}}_f)-f_0\right\|_{L^2}^2\right]$$

$$\leq (1+2\delta)\inf_{\mathbf{t}_f\in\mathbf{T}^f_{n_t^o}}\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\mathbf{t}_f)-f_0\right\|_{L^2}^2\right]$$

$$+(1+\delta)\frac{16\log\left(1+\left|\mathbf{T}^f_{n_t^o}\right|\right)}{n^{o\prime}_{t\prime}}\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\max_{\mathbf{t}_f\in\mathbf{T}^f_{n_t^o}}\left[\left(1+\left\|\hat{f}_{-v}(\mathbf{t}_f)-f_0\right\|_{\infty}^2\right)\left(\frac{C(1+\delta)}{\delta}\right)\right]\right]$$

with some constant $C<\infty$.

The result in here tells us that, up to constant $(1+2\delta)$ and an additional remainder term, the cross-validation procedure can provide us an expected fold-average estimation error that is, at least as good as the best one within the provided candidate set $\mathbf{T}^f_{n_t^o}$.

The work most closely related to our result in here is Cetin (2026), which has also provided theoretical guarantees for cross-validated selection based on the Riesz loss through an oracle inequality argument. It allows selection across both estimation methods and tuning parameters. Working specifically with ridge-regularized estimation over a growing and ill-conditioned dictionary, we address the candidate coverage requirement in their Assumption 2 by constructing candidate sets that contain a pair of tuning choices whose corresponding nuisance estimators satisfy the product-rate condition. We devote the rest of this section to present the candidate-set construction and to establish the sets' theoretical guarantees.

5.2 Candidate Sets

This section provides a candidate-set construction for the tuning parameters. Our construction is feasible in the sense that it does not require knowledge on the approximation exponents $(\kappa_\alpha, \kappa_\gamma)$. Such design will be formally justified in Section 5.3 below under the stated assumptions. We note, however, that we do not claim its superiority over other candidate sets or its uniqueness in providing the suggested theoretical guarantees.

For the sieve size of nuisance $f \in \{\alpha, \gamma\}$, let the corresponding candidate set be

$$\mathcal{S}^f_{n_t^o} \equiv \left\{2, 3, \ldots, \bar{p}^f_{n_t^o}\right\} \ , \tag{24}$$

$$\text{with} \quad \bar{p}^f_{n_t^o} \equiv \max\left\{p\in\mathbb{Z}:\ p\geq 3 \quad \text{and} \quad p\log p < \frac{n_t^o}{\log n_t^o}\right\} \ . \tag{25}$$

The design makes sure the upper bound $\bar{p}_{n_t^o}$ goes to infinity with n_t^o, while respecting

Algorithm 1. Cross-Validated Tuning for First Steps within Fold k

Input: $b_1, b_2, b_3, \ldots$ `# Predetermined real-valued functions`
$\{\{W_i\}_{i \in I_v}\}_{v=1}^{V}$ `# Partitioned data`

Output: $(\hat{p}_\alpha, \hat{p}_\gamma, \hat{\zeta}, \hat{\lambda})$ `# Selected sieve sizes & penalty parameters`

$n_t^o \leftarrow n(K-1)/K$ `# For each candidate dimension` p `and inner fold` v

$\tilde{b}_{-v}^{(p)}(x) \leftarrow b^{(p)}(x) - \frac{1}{|I_{-v}|} \sum_{i \in I_{-v}} b^{(p)}(X_i)$

$(\mathcal{S}^\alpha, \mathcal{S}^\gamma) \leftarrow (\mathcal{S}_{n_t^o}^\alpha, \mathcal{S}_{n_t^o}^\gamma)$ `# Sieve sizes candidate sets based on (24)`

$\mathcal{D}^\alpha \leftarrow \{\}, \quad \mathcal{D}^\gamma \leftarrow \{\}$ `# Create empty dictionaries to hold the results`

for $p_\alpha \in \mathcal{S}^\alpha$ **do**

  $\mathcal{Z} \leftarrow \mathcal{Z}_{n_t^o}(p_\alpha)$ `# Penalty candidate set based on (26)`

  for $\zeta \in \mathcal{Z}$ **do**

    for $v \leftarrow 1$ **to** V **do**

      `# Fit` $\hat{\alpha}_{-v}$ `using` $\{W_i\}_{i \in I_{-v}}$

      $\bar{\alpha}_{-v} \leftarrow \sum_{i \in I_{-v}} m(W_i, 1)/|I_{-v}|$ `# Intercept for` $\hat{\alpha}_{-v}$

      `# Vector of coefficients for` $\hat{\alpha}_{-v}$ `based on (15)`

      $\hat{\rho}_{-v} \leftarrow \hat{\rho}_{-v}(p_\alpha, \zeta)$

      `# Evaluate the fitted first steps using` $\{W_i\}_{i \in I_v}$

      `# Riesz Loss`

      $\hat{R}_{-v}^\alpha \leftarrow$
      $\sum_{i \in I_v} \{(\bar{\alpha}_{-v} + \tilde{b}_{-v}^{(p_\alpha)}(X_i)^\intercal \hat{\rho}_{-v})^2 - 2m(W_i, \bar{\alpha}_{-v} + \tilde{b}_{-v}^{(p_\alpha)}(X_i)^\intercal \hat{\rho}_{-v})\}/|I_v|$

    $\mathcal{D}^\alpha[(p_\alpha, \zeta)] \leftarrow \sum_{v=1}^{V} \hat{R}_{-v}^\alpha / V$

for $p_\gamma \in \mathcal{S}^\gamma$ **do**

  $\Lambda \leftarrow \Lambda_{n_t^o}(p_\gamma)$ `# Penalty candidate set based on (27)`

  for $\lambda \in \Lambda$ **do**

    for $v \leftarrow 1$ **to** V **do**

      `# Fit` $\hat{\gamma}_{-v}$ `using` $\{W_i\}_{i \in I_{-v}}$

      $\bar{Y}_{-v} \leftarrow \sum_{i \in I_{-v}} Y_i/|I_{-v}|$ `# Intercept for` $\hat{\gamma}_{-v}$

      `# Vector of coefficients for` $\hat{\gamma}_{-v}$ `based on (16)`

      $\hat{\beta}_{-v} \leftarrow \hat{\beta}_{-v}(p_\gamma, \lambda)$

      `# Evaluate the fitted first steps using` $\{W_i\}_{i \in I_v}$

      $\hat{R}_{-v}^\gamma \leftarrow \sum_{i \in I_v} \{\bar{Y}_{-v} + \tilde{b}_{-v}^{(p_\gamma)}(X_i)^\intercal \hat{\beta}_{-v} - Y_i\}^2/|I_v|$ `# MSE`

    `# Sum across folds and insert into the dictionaries`

    $\mathcal{D}^\gamma[(p_\gamma, \lambda)] \leftarrow \sum_{v=1}^{V} \hat{R}_{-v}^\gamma / V$

`# Select key with minimum validation loss`

$(\hat{p}_\alpha, \hat{\zeta}) \leftarrow \arg\min_{(p_\alpha, \zeta) \in \mathrm{keys}(\mathcal{D}^\alpha)} \mathcal{D}^\alpha[(p_\alpha, \zeta)]$

$(\hat{p}_\gamma, \hat{\lambda}) \leftarrow \arg\min_{(p_\gamma, \lambda) \in \mathrm{keys}(\mathcal{D}^\gamma)} \mathcal{D}^\gamma[(p_\gamma, \lambda)]$

$\bar{p}_{n_t^o} \log \bar{p}_{n_t^o} = o(n_t^o)$. For the ridge penalties, based on Assumption 4.1, we have

$$\mathcal{Z}_{n_t^o}(p_\alpha) \equiv \left\{ \zeta \in \mathbb{R}_{>0} : \zeta = 2^{-\ell} \cdot r_{n_t^o}(p_\alpha), \quad \ell = 0, \ldots, M_{n_t^o} \right\} , \tag{26}$$

$$\Lambda_{n_t^o}(p_\gamma) \equiv \left\{ \lambda \in \mathbb{R}_{>0} : \lambda = 2^{-\ell} \cdot r_{n_t^o}(p_\gamma), \quad \ell = 0, \ldots, M_{n_t^o} \right\}, \tag{27}$$

with

$$r_{n_t^o}(p) \equiv \left(\frac{p \log p}{n_t^o} \right)^{1/3} , \quad M_{n_t^o} \equiv \left\lfloor \frac{\log \left[16 (n_t^o)^{1/9} \left\{ \log(e(n_t^o)) \right\}^{1/3} \right]}{\log 2} \right\rfloor .$$

The somewhat unusual form of $M_{n_t^o}$ is designed to allow the candidate grid to include progressively smaller penalty multiplier as the sample size grows, while still preserve the requirements for our theoretical guarantees.

5.3 Theoretical Guarantees

With cross-validated choices $\hat{\mathbf{t}}_\gamma$ and $\hat{\mathbf{t}}_\alpha$, we refit each nuisance on the full training part I_{-k} to obtain estimators $\hat{\gamma}_{-k}(\hat{\mathbf{t}}_\gamma)$ and $\hat{\alpha}_{-k}(\hat{\mathbf{t}}_\alpha)$. In this section, we aim to study the convergence rates of

$$\left\| \hat{\alpha}_{-k}\left(\hat{p}_\alpha, \hat{\zeta}\right) - \alpha_0 \right\|_{L^2} \text{ and } \left\| \hat{\gamma}_{-k}\left(\hat{p}_\gamma, \hat{\lambda}\right) - \gamma_0 \right\|_{L^2} . \tag{28}$$

The end goal is to verify that, within the warrant set delineated by the possibility frontier in Section 4, the product-rate requirement in (19) can be achieved by using separate cross-validation in Auto-DML with our proposed candidate sets. According to Section 5.2, we now have

$$\mathbf{T}_{n_t^o}^\gamma = \left\{ (p, \lambda) : p \in \mathcal{S}_{n_t^o}^\gamma, \ \lambda \in \Lambda_{n_t^o}(p) \right\} ,$$

$$\mathbf{T}_{n_t^o}^\alpha = \left\{ (p, \zeta) : p \in \mathcal{S}_{n_t^o}^\alpha, \ \zeta \in \mathcal{Z}_{n_t^o}(p) \right\} ,$$

with $\mathcal{S}_{n_t^o}^\gamma, \mathcal{S}_{n_t^o}^\alpha, \Lambda_{n_t^o}(p), \mathcal{Z}_{n_t^o}(p)$ defined in (24), (26), (27). Recall that earlier in Lemma 5.1, for each nuisance $f \in \{\alpha, \gamma\}$, we have already established an oracle inequality for its expected fold-average estimation error

$$\mathbb{E}_{W_{-k}} \left[\frac{1}{V} \sum_{v=1}^{V} \left\| \hat{f}_{-v}(\hat{\mathbf{t}}_f) - f_0 \right\|_{L^2}^2 \right] , \tag{29}$$

under cross-validated tuning $\mathbf{t}_f$ within an arbitrary candidate set.

With that in hand, to achieve our goal, what remains is to show that our candidate sets $\mathbf{T}_{n_t^o}^\gamma$ and $\mathbf{T}_{n_t^o}^\alpha$ exhibit three important properties. First, to make full use of the oracle inequalities in Lemma 5.1, we must control the remainder term. That being said, for each nuisance $f \in \{\gamma, \alpha\}$, the designed candidate set $\mathbf{T}_{n_t^o}^f$ should be able to bound

the following quantity:

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\max_{\mathbf{t}_f\in\mathbf{T}^f_{n_t^o}}\left[\left(1+\left\|\hat{f}_{-v}(\mathbf{t}_f)-f_0\right\|^2_\infty\right)\right]\right] .$$

Additionally, for the remainder term to vanish, cardinality of the set $|\mathbf{T}^f_{n_t^o}|$ must be controlled so that $\log(1+|\mathbf{T}^f_{n_t^o}|) = o(n_{t'}^{o'})$. Finally, the oracle inequality can at most tells us that, up to a constant and an additional remainder term, the cross-validation procedure produces an estimator that performs at least as good as the best one within the provided candidate set $\mathbf{T}^f_{n_t^o}$. Therefore, it is important to show that within our candidate sets, there exists elements which the desired rate can be accomplished. We now establish each of these properties.

Lemma 5.2 (Properties of the candidate sets)**.** *For each nuisance $f \in \{\gamma, \alpha\}$, consider the candidate set $\mathbf{T}^f_{n_t^o}$ defined above.*

(i) Expected maximum bound.

If Assumptions 2.2(i)(iii), 2.4(i), 3.1(ii)(iii), 3.2(i), 3.3(ii), 5.1(ii), and 5.2 hold, then

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\max_{\mathbf{t}_\gamma\in\mathbf{T}^\gamma_{n_t^o}}\left[\left(1+\|\hat{\gamma}_{-v}(\mathbf{t}_\gamma)-\gamma_0\|^2_\infty\right)\right]\right] = O(1) .$$

If Assumptions above hold, with: Assumptions 2.1(iii), 3.4 replacing 2.4(i), Assumption 3.3(i) replacing 3.3(ii), and in addition, Assumption 2.4(ii) also hold, then

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\max_{\mathbf{t}_\alpha\in\mathbf{T}^\alpha_{n_t^o}}\left[\left(1+\|\hat{\alpha}_{-v}(\mathbf{t}_\alpha)-\alpha_0\|^2_\infty\right)\right]\right] = O(1) .$$

(ii) Cardinality control.

We have for $f \in \{\gamma, \alpha\}$

$$\log(1+|\mathbf{T}^f_{n_t^o}|) = O(\log n_t^o) = o(n_{t'}^{o'}) .$$

(iii) Covering a rate-attaining choice.

Suppose Assumptions 5.1 and 5.2 hold. If, in addition, assumptions in Lemma 3.1 for the result on $\hat{\gamma}_{-k}$ hold with Assumptions 4.2(ii), then there exists $\mathbf{t}^*_\gamma \in \mathbf{T}^\gamma_{n_t^o}$ such that

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{\gamma}_{-v}(\mathbf{t}^*_\gamma)-\gamma_0\right\|^2_{L^2}\right] = O\left((n_t^o)^{-\frac{2\kappa_\gamma}{3\kappa_\gamma+1}}\right) .$$

Instead, if assumptions in Lemma 3.1 for the result on $\hat{\alpha}_{-k}$ hold with Assumptions

2.4(ii) and 4.2(i), then there exists $\mathbf{t}_\alpha^* \in \mathbf{T}_{n_t^o}^\alpha$ *such that*

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\|\hat{\alpha}_{-v}(\mathbf{t}_\alpha^*)-\alpha_0\|_{L^2}^2\right] = O\left((n_t^o)^{-\frac{2\kappa_\alpha}{3\kappa_\alpha+1}}\right) .$$

Combining properties (i)-(iii) above with the established oracle inequality would allow us to draw conclusion on the expected cross-validation fold-average estimation error

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\hat{\mathbf{t}}_f)-f_0\right\|_{L^2}^2\right] = O\left((n_t^o)^{-\frac{2\kappa_f}{3\kappa_f+1}}\right) .$$

To translate the bound in (29) to our refitted target in (28), we impose the following refit stability condition, which guarantees that for any $\kappa_f > 0$,

$$\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_f\in\mathbf{T}_{n_t^o}^f}\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\mathbf{t}_f)-\hat{f}_{-k}(\mathbf{t}_f)\right\|_{L^2}^2\right] = O\left((n_t^o)^{-\frac{2\kappa_f}{3\kappa_f+1}}\right) .$$

Assumption 5.5 (Refit stability)**.** *For each nuisance* $f \in \{\gamma, \alpha\}$,

$$\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_f\in\mathbf{T}_{n_t^o}^f}\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\mathbf{t}_f)-\hat{f}_{-k}(\mathbf{t}_f)\right\|_{L^2}^2\right] = O\left((n_t^o)^{-2/3}\right) .$$

Theorem 5.1 (Risk at the selected tuning)**.** *Suppose Assumption 5.5 is satisfied. If Assumptions in Lemmas 5.1 and 5.2 hold, then for* $f \in \{\gamma, \alpha\}$

$$\|\hat{f}_{-k}(\hat{\mathbf{t}}_f)-f_0\|_{L^2} = O_p\left((n_t^o)^{-\frac{\kappa_f}{3\kappa_f+1}}\right) .$$

Finally, Theorem 5.1 brings our analysis full circle. Under the stated assumptions, separate cross-validation for the two nuisance estimators over our proposed candidate sets attains the convergence rates established above without requiring knowledge of κ_γ and κ_α. These rates satisfy the product-rate requirement whenever the problem's associated exponent pair lies strictly above our feasibility frontier, thereby supporting valid inference in the Auto-DML framework.

Lemma 5.3 (Asymptotic normality with separately cross-validated nuisances)**.** *Suppose Assumptions 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 4.2, 5.1, 5.2, 5.3, 5.4, and 5.5 are satisfied. Furthermore, provided that the approximation exponent pair* $(\kappa_\alpha, \kappa_\gamma)$ *lies strictly above the feasibility frontier. That is,*

$$\frac{\kappa_\alpha}{3\kappa_\alpha+1}+\frac{\kappa_\gamma}{3\kappa_\gamma+1} > \frac{1}{2}.$$

Then with $\sigma > 0$, $\sigma^{-1}\sqrt{n}(\hat{\theta}_n(\hat{\mathbf{t}}_\gamma, \hat{\mathbf{t}}_\alpha)-\theta_0) \xrightarrow{L} N(0,1)$.

6 Simulation Studies

In this section, we use simulations to illustrate our theoretical findings. We study the Auto-DML estimator with tuning parameters selected by our proposed procedure. Across data-generating designs lying above and below the feasibility frontier for our product-rate warrant set, we examine: (i) the finite-sample coverage of its confidence intervals and (ii) the empirical distribution of its studentized estimation error.

6.1 Simulation Design

We consider the parameter studied in Example 1 of Chernozhukov et al. (2022a). With observables $W \equiv (Y, D, X^\intercal)^\intercal$, the parameter of interest is

$$\theta_0 = \mathbb{E}\left[D \cdot \gamma_0(X)\right] = \mathbb{E}\left[\mathbb{E}[D|X] \cdot \gamma_0(X)\right] \ .$$

Here, $m(W, \gamma) = D \cdot \gamma(X)$, and the Riesz representer is $\alpha_0(X) = \mathbb{E}[D|X]$. As noted in the paper, this parameter of interest is a component of the expected conditional covariance

$$\mathbb{E}\left[\mathrm{Cov}(D, Y|X)\right] = \mathbb{E}[DY] - \theta_0 \ ,$$

which is therefore important to both covariance analysis and the estimation of a partially linear model. Below, we provide the detail of our data-generating process. It is designed with a goal to satisfy our maintained regularity assumptions while allowing us to vary the approximation exponents $(\kappa_\gamma, \kappa_\alpha)$ for experimenting.

DGP. The data-generating process is constructed as follows:

$$Y = f_{\kappa_\gamma}(X) + U \ , \qquad D = f_{\kappa_\alpha}(X) + UV \ .$$

where $U, V \in \{-1, 1\}$ with $P(U = 1) = 0.5$ and $P(V = 1) = 0.65$. Variables U, V involved in the error terms are mutually independent, and each of them is also independent of X. This gives $\mathbb{E}[U] = \mathbb{E}[UV] = 0$ and $\mathrm{Cov}(U, UV) = 0.3$.

$$f_\kappa(X) = \frac{X_1}{\sqrt{2}} + \frac{1}{\sqrt{2S_\kappa}} \sum_{\ell=1}^{\infty} (2^\ell \cdot \ell)^{-\kappa} X_{2^\ell} \quad \text{with} \quad S_\kappa = \sum_{\ell=1}^{\infty} (2^\ell \cdot \ell)^{-2\kappa} \ ,$$

which gives $\mathbb{E}[f_\kappa] = 0$ and $\mathbb{E}[f_\kappa^2] = 1$. As we can see from the design, both nuisance functions depend only on coordinates $1, 2, 4, 8, \ldots$ of X. We refer to the remaining coordinates as inactive coordinates. In addition to the nuisance function, the approximation exponent also depends on the dictionary used to approximate it. We next specify the

dictionary and its covariance structure. Let

$$b_j(X) = \sqrt{v_j} X_j \ , \qquad v_j = \begin{cases} 1, & j = 1 \\ \frac{1}{4} j^{-1/10}, & j \in \{2, 4, 8, \ldots\} \ , \\ \frac{1}{100 j^2} + \frac{2}{d_{irr}} \mathbf{1}\{j \in \mathcal{I}_{d_{irr}}\} \ , & otherwise \ , \end{cases}$$

and $\mathcal{I}_{d_{irr}}$ denotes the first d_{irr} inactive coordinates (not powers of two). For example, $\mathcal{I}_4 = \{3, 5, 6, 7\}$. We use $d_{irr} \in \{16, 128, 1024\}$. Combining the design of the data-generating process and the dictionary we have

$$\inf_{a,c} \mathbb{E}[\{f_\kappa(X) - a - b^{(p)}(X)'c\}^2] = \frac{1}{2S_\kappa} \sum_{\ell > \lfloor \log_2 p \rfloor} (2^\ell \cdot \ell)^{-2\kappa} \asymp (p \log p)^{-2\kappa} \ ,$$

which is exactly what we want.

For each combination of $(\kappa_\alpha, \kappa_\gamma)$, d_{irr}, and sample size $n \in \{2048, 4096, 8192\}$, we conduct 2000 Monte Carlo replications. Within each repetition, we use two-fold outer-cross-fitting, with three-fold inner cross-validation.

6.2 Simulation Results

The simulation results are presented in Figure 4. We conduct experiments at three different points as indicated in Figure 3, which are:

- Case A (Deep inside the frontier): $(\kappa_\gamma, \kappa_\alpha) = (3, 3)$
- Case B (Marginally inside the frontier): $(\kappa_\gamma, \kappa_\alpha) = (1.1, 1.1)$
- Case C (Outside the frontier): $(\kappa_\gamma, \kappa_\alpha) = (0.3, 0.3)$

Consistent with our theory, in Cases A and B, whose approximation-exponent pairs lie above the feasibility frontier, the Auto-DML estimator with tuning parameters selected by our proposed procedure yields confidence intervals with coverage close to the nominal 95%. The empirical distributions of its studentized estimation error also closely resemble the standard normal distribution. In contrast, Case C exhibits undercoverage, accompanied by studentized error distributions shifted to the left hand side of zero. This illustrates that the core components of DML (the use of Neyman orthogonal moment functions and cross-fitting) do not by themselves guarantee accurate finite-sample coverage, even when the nuisance tuning parameters are selected separately by cross-validation, as is common in practice. Practitioners should therefore recognize that the inferential guarantees of DML remain conditional on sufficiently accurate nuisance estimation. Orthogonality, cross-fitting, and cross-validation do not eliminate this requirement.

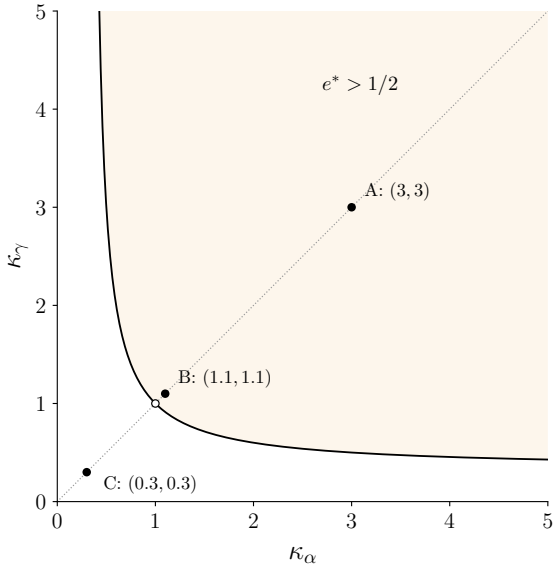


Figure 3: Simulation Designs Relative to the Frontier

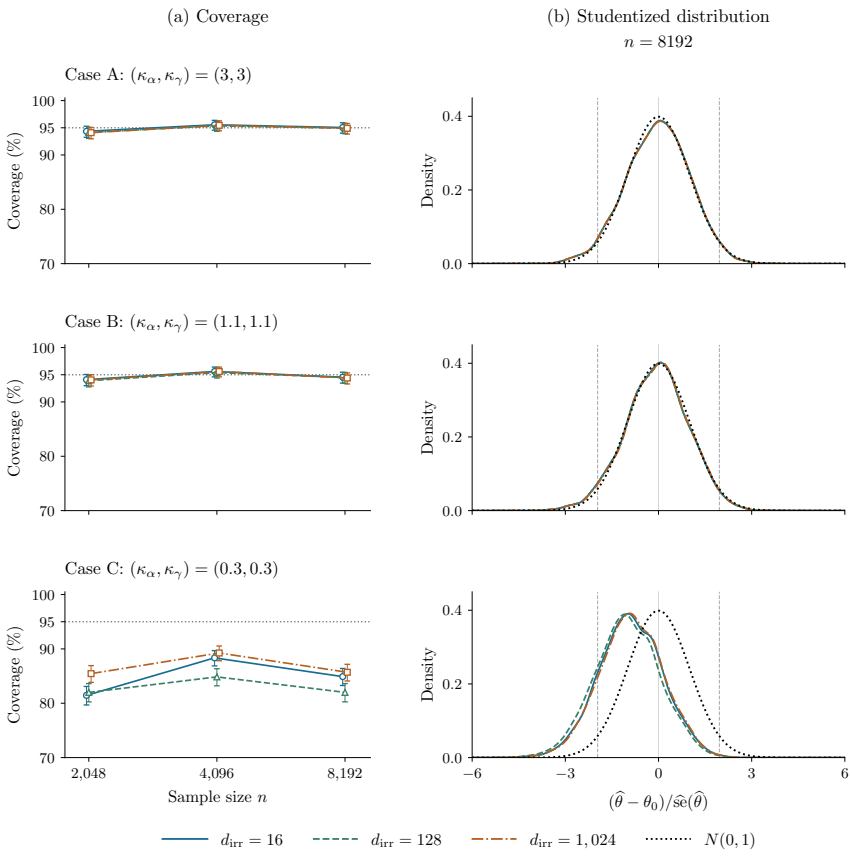


Figure 4: Coverage and Studentized Error Distributions

To assess the robustness of these patterns, we considered additional asymmetric designs, as reported in Appendix D. Results there are consistent with those presented here. One observation worth highlighting, however, is that the symmetric reference case R exhibits coverage close to the nominal level despite lying below the feasibility frontier. This result is consistent with the possibility that our bounds on the learning rates in (17) and (18) are rather conservative and that a sharper analysis could potentially yield a larger product-rate warrant set.

7 Conclusion

In this paper, we study cross-validation in Auto-DML from an inference point of view. We focus on the ridge-regularized sieve estimators for both first steps and characterize a feasibility frontier in terms of the approximation exponents of the two nuisances. Such frontier describes when our learning-rate bounds warrant the product-rate condition. Building on this characterization, we construct candidate sets for both the sieve dimensions and the ridge penalties, which require no knowledge on the approximation exponents. Finally, we justify separate cross-validation over these sets. These results connect the product-rate requirement underlying DML's inference to an implementable procedure for tuning each nuisance separately. We view these results as a small step towards answering the difficult question of how to choose tuning parameters in DML. An interesting next step is to investigate whether coupling the candidate sets and jointly selecting the nuisance tuning parameters can improve the finite-sample accuracy of the target estimator while preserving asymptotically valid inference.

References

Andrews, D. W. (1991): "Asymptotic Optimality of Generalized C, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

Bach, P., O. Schacht, V. Chernozhukov, S. Klaassen, and M. Spindler (2024): "Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study," *arXiv preprint arXiv:2402.04674*.

Bradic, J., V. Chernozhukov, W. K. Newey, and Y. Zhu (2025): "Minimax Semiparametric Learning With Approximate Sparsity," *arXiv preprint arXiv:1912.12213*.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Chapter 77 Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, Elsevier, vol. 6, 5633–5751.

Cetin, T. (2026): "Adaptive Double Machine Learning via Riesz-Risk Cross-Validation," *SSRN preprint SSRN:6215762*.

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Elsevier, vol. 6, 5549–5632.

Chen, X. and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80, 277–321.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022a): "Locally Robust Semiparametric Estimation," *Econometrica*, 90, 1501–1535.

Chernozhukov, V., W. Newey, V. M. Quintas-Martínez, and V. Syrgkanis (2022b): "RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 3901–3914.

Chernozhukov, V., W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis (2024): "Automatic Debiased Machine Learning via Riesz Regression," *arXiv preprint arXiv:2104.14737*.

Chernozhukov, V., W. K. Newey, and R. Singh (2022c): "Automatic Debiased Machine Learning of Causal and Structural Effects," *Econometrica*, 90, 967–1027.

——— (2023): "A Simple and General Debiased Machine Learning Theorem with Finite-Sample Guarantees," *Biometrika*, 110, 257–264.

Chernozhukov, V., W. K. Newey, R. Singh, and V. Syrgkanis (2026): "Adversarial Estimation of Riesz Representers," *Journal of the American Statistical Association*, 1–12.

Chetverikov, D. (2024): "Tuning Parameter Selection in Econometrics," *arXiv preprint arXiv:2405.03021*.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021): "On Cross-Validated Lasso in High Dimensions," *The Annals of Statistics*, 49, 1300–1317.

Hansen, B. E. (2014): "Nonparametric Sieve Regression," in *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, ed. by J. S. Racine, L. Su, and A. Ullah, Oxford University Press, 214–248.

Hirshberg, D. A. and S. Wager (2021): "Augmented Minimax Linear Estimation," *The Annals of Statistics*, 49.

Horvitz, D. G. and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Kato, K. (2013): "Quasi-Bayesian Analysis of Nonparametric Instrumental Variables Models," *The Annals of Statistics*, 41.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K. and J. R. Robins (2018): "Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation," *arXiv preprint arXiv:1801.09138*.

Robins, J., E. Tchetgen Tchetgen, L. Li, and A. Van Der Vaart (2009): "Semiparametric Minimax Rates," *Electronic Journal of Statistics*, 3.

Robins, J. M., L. Li, R. Mukherjee, E. T. Tchetgen, and A. Van Der Vaart (2017): "Minimax Estimation of a Functional on a Structured High-Dimensional Model," *The Annals of Statistics*, 45.

Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rudelson, M. (1999): "Random Vectors in the Isotropic Position," *Journal of Functional Analysis*, 164, 60–72.

Singh, R. (2024): "Kernel Ridge Riesz Representers: Generalization, Mis-specification, and the Counterfactual Effective Dimension," *arXiv preprint arXiv:2102.11076*.

Timan, A. F. (1963): *Theory of Approximation of Functions of a Real Variable*, Elsevier.

Vaart, A. W. V. D., S. Dudoit, and M. J. V. D. Laan (2006): "Oracle Inequalities for Multi-Fold Cross Validation," *Statistics & Decisions*, 24, 351–371.

van der Laan, L., A. Bibaut, N. Kallus, and A. Luedtke (2025): "Automatic Debiased Machine Learning for Smooth Functionals of Nonparametric M-Estimands," *arXiv preprint arXiv:2501.11868*.

Appendix A Proof of Main Result

Proof of Lemma 2.1 Let $\hat{\psi}_k(W) \equiv m(W, \hat{\gamma}_{-k}) - \theta_0 + \hat{\alpha}_{-k}(X)(Y - \hat{\gamma}_{-k}(X))$ and $\psi_0(W) \equiv m(W, \gamma_0) - \theta_0 + \alpha_0(X)(Y - \gamma_0(X))$. From the definition of $\hat{\theta}_n$ in (6),

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} \{m(W_i, \hat{\gamma}_{-k}) - \theta_0 + \hat{\alpha}_{-k}(X_i)(Y_i - \hat{\gamma}_{-k}(X_i))\} \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} \hat{\psi}_k(W_i) \ .\end{aligned}$$

Subtracting $\sum_{i=1}^{n} \psi_0(W_i)/\sqrt{n}$ and rearranging, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} \hat{\psi}_k(W_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_0(W_i) \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} [\{m(W_i, \hat{\gamma}_{-k}) - m(W_i, \gamma_0)\} - \{\alpha_0(X_i)(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i))\}] \\ &+ \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} \{(\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i))(Y_i - \gamma_0(X_i))\} \\ &+ \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} \{(\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i))(\gamma_0(X_i) - \hat{\gamma}_{-k}(X_i))\} \ .\end{aligned} \tag{A.30}$$

We now show that each of the three terms on the right-hand side is $o_p(1)$, which establishes the first part of the lemma. Before proceeding, define the terms

$$\begin{aligned}Z_{1i,k} &\equiv \{m(W_i, \hat{\gamma}_{-k}) - m(W_i, \gamma_0)\} - \{\alpha_0(X_i)(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i))\} \ , \\ Z_{2i,k} &\equiv (\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i))(Y_i - \gamma_0(X_i)) \ , \\ Z_{3i,k} &\equiv (\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i))(\gamma_0(X_i) - \hat{\gamma}_{-k}(X_i)) \ .\end{aligned}$$

First, consider the leading term in (A.30). Fix $\epsilon > 0$ and $k \in \{1, \ldots, K\}$. By conditional Markov's inequality,

$$P\left(\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{1i,k} \right\| > \epsilon \ \middle| \ \{W_j : j \notin I_k\} \right) \leq \frac{\mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{1i,k} \right\|^2 \ \middle| \ \{W_j : j \notin I_k\} \right]}{\epsilon^2} \ . \tag{A.31}$$

Now fix $k \in \{1, \ldots, K\}$. Notice that conditional on the first-step training sample $\{W_j : j \notin I_k\}$, $\hat{\gamma}_{-k}$ is fixed and by Assumption 2.2(i)(iii), $\{Z_{1i,k} : i \in I_k\}$ are i.i.d. Since

$0 \leq n_k \leq n$, together with the triangle inequality, we obtain with Assumptions 2.1,

$$
\begin{aligned}
&\Big\|\mathbb{E}\Big[\frac{1}{\sqrt{n}}\sum_{i\in I_k} Z_{1i,k} \mid \{W_j : j \notin I_k\}\Big]\Big\| \\
&= \Big\|\mathbb{E}\Big[\frac{1}{\sqrt{n}}\sum_{i\in I_k}\big(m(W_i,\hat{\gamma}_{-k}) - m(W_i,\gamma_0) - \alpha_0(X_i)\big(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i)\big)\big) \;\Big|\; \{W_j : j \notin I_k\}\Big]\Big\| \\
&\leq \sqrt{n_k}\cdot\Big\|\mathbb{E}\left[m(W_i,\hat{\gamma}_{-k}) - m(W_i,\gamma_0) - \alpha_0(X_i)\left(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i)\right) \;\middle|\; \{W_j : j \notin I_k\}\right]\Big\| \\
&= \sqrt{n_k}\cdot\Big\|\mathbb{E}\left[\alpha_0(X_i)\left(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i)\right) - \alpha_0(X_i)\left(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i)\right) \;\middle|\; \{W_j : j \notin I_k\}\right]\Big\| \\
&= 0
\end{aligned}
\tag{A.32}
$$

where the second last step follows by (3). Moreover, again since $0 \leq n_k \leq n$, combining result in (A.32), Assumptions 2.1, 2.2(i)(iii), and 2.4(ii) give us

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i\in I_k} Z_{1i,k}\Big\|^2 \;\Big|\; \{W_j : j \notin I_k\}\Big] \\
&\leq \mathbb{E}[Z_{1i,k}^2 \mid \{W_j : j \notin I_k\}] \\
&= \mathbb{E}[\big(m(W_i,\hat{\gamma}_{-k}) - m(W_i,\gamma_0) - \alpha_0(X_i)\big(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i)\big)\big)^2 \mid \{W_j : j \notin I_k\}] \\
&\leq 2\Big(\mathbb{E}[\big(m(W_i,\hat{\gamma}_{-k}) - m(W_i,\gamma_0)\big)^2 \mid \{W_j : j \notin I_k\}] \\
&\qquad + \mathbb{E}\left[\big(\alpha_0(X_i)\big(\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i)\big)\big)^2 \mid \{W_j : j \notin I_k\}\right]\Big) \\
&\leq 2\Big(\mathbb{E}[\big(m(W_i,\hat{\gamma}_{-k}) - m(W_i,\gamma_0)\big)^2 \mid \{W_j : j \notin I_k\}] \\
&\qquad + \|\alpha_0\|_\infty^2 \|\hat{\gamma}_{-k}(X_i) - \gamma_0(X_i)\|_{L^2}^2\Big) = o_p(1)\ .
\end{aligned}
\tag{A.33}
$$

with the conclusion drawn by Assumptions 2.5(i), 2.2(iv). With (A.31) and (A.33), we have for every $\epsilon > 0$ and $k \in \{1,\ldots,K\}$,

$$
P\left(\Big\|\frac{1}{\sqrt{n}}\sum_{i\in I_k} Z_{1i,k}\Big\| > \epsilon \;\middle|\; \{W_j : j \notin I_k\}\right) = o_p(1)\ . \tag{A.34}
$$

Next, notice that with the law of iterated expectation, for all $\epsilon > 0$ and $k \in \{1, \ldots, K\}$,

$$\begin{aligned} P\left(\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{1i,k} \right\| > \epsilon \right) &= \mathbb{E}\left[\mathbb{1}_{\{\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{1i,k} \| > \epsilon\}} \right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{1i,k} \| > \epsilon\}} \,\middle|\, \{W_j : j \notin I_k\} \right] \right] \\ &= o(1) \ , \end{aligned} \tag{A.35}$$

in which the conclusion follows from dominated convergence theorem with result in (A.34), given the fact that

$$\left| \mathbb{E}\left[\mathbb{1}_{\{\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{1i,k} \| > \epsilon\}} \,\middle|\, \{W_j : j \notin I_k\} \right] \right| \leq 1 \ .$$

Since K is fixed (Assumption 2.2(ii)), in addition with Assumption 2.2(iii), results from (A.35) give us

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} Z_{1i,k} = o_p(1) \ . \tag{A.36}$$

Repeating the argument leading to (A.36), to show that the second term in (A.30) is $o_p(1)$, it suffices to verify

$$\sqrt{n_k} \cdot \| \mathbb{E}\left[Z_{2i,k} \mid \{W_j : j \notin I_k\} \right] \| = o_p(1), \quad \mathbb{E}\left[Z_{2i,k}^2 \mid \{W_j : j \notin I_k\} \right] = o_p(1) \ .$$

uniformly across $k \in \{1, \ldots K\}$. Recall that

$$Z_{2i,k} \equiv (\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i))(Y_i - \gamma_0(X_i)) \ .$$

Therefore, by using the law of iterated expectation, we have for every $k \in \{1, \ldots, K\}$,

$$\begin{aligned} &\sqrt{n_k} \cdot \| \mathbb{E}[Z_{2i,k} | \{W_j : j \notin I_k\}] \| \\ &= \sqrt{n_k} \cdot \| \mathbb{E}\left[(\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i)) \mathbb{E}[(Y_i - \gamma_0(X_i)) \mid X] \mid \{W_j : j \notin I_k\} \right] \| \\ &= 0 \ . \end{aligned}$$

Meanwhile by the law of iterated expectation with Assumption 2.4(i), for each $k \in \{1, \ldots, K\}$,

$$\begin{aligned}
&\mathbb{E}[Z_{2i,k}^2 \mid \{W_j : j \notin I_k\}] \\
&= \mathbb{E}[(\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i))^2 (Y_i - \gamma_0(X_i))^2 \mid \{W_j : j \notin I_k\}] \\
&\leq \mathbb{E}[(\hat{\alpha}_{-k}(X_i) - \alpha_0(X_i))^2 \mathbb{E}[(Y_i - \gamma_0(X_i))^2 \mid X] \mid \{W_j : j \notin I_k\}] \\
&\lesssim \|\hat{\alpha}_{-k} - \alpha_0\|_{L^2}^2 = o_p(1) \ .
\end{aligned}$$

Conclusion is again due to Assumptions 2.5(i), 2.2(iv). Next, for $\ell = 3$, we obtain from the Cauchy-Schwarz inequality with Assumptions 2.5(ii), 2.2(iv),

$$\begin{aligned}
&\sqrt{n_k} \ \cdot \mathbb{E}[\|Z_{3i,k}\| | \{W_j : j \notin I_k\}] \\
&\leq \sqrt{n_k} \ \cdot \|\hat{\alpha}_{-k} - \alpha_0\|_{L^2} \|\hat{\gamma}_{-k} - \gamma_0\|_{L^2} = o_p(1) \ ,
\end{aligned}$$

which allows us to conclude

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} Z_{3i,k} = o_p(1)$$

by following the same logic above, with the conditional Markov's inequality replaced by one that is denoted in terms of absolute first moment. That is,

$$P\left(\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{3i,k} \right\| > \epsilon \ \middle| \ \{W_j : j \notin I_k\} \right) \leq \frac{\mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} Z_{3i,k} \right\| \ \middle| \ \{W_j : j \notin I_k\} \right]}{\epsilon} \ .$$

Therefore, we have proved

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} Z_{\ell i,k} = o_p(1) \ , \quad \forall \ell \in \{1, 2, 3\} \ . \tag{A.37}$$

Combining result from (A.30) with (A.37), we arrive at

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_0(W_i) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(W_i, \gamma_0) - \theta_0 + \alpha_0(X_i)(Y_i - \gamma_0(X_i))\} + o_p(1) \ ,
\end{aligned}$$

which concludes the first part of the lemma. For the second part of the lemma, observe that by the law of iterated expectation,

$$\begin{aligned}\mathbb{E}[\psi_0(W_i)] &= \mathbb{E}[m(W,\gamma_0) - \theta_0 + \alpha_0(X)(Y - \gamma_0(X))] \\ &= (\mathbb{E}[m(W,\gamma_0)] - \theta_0) + \mathbb{E}[\alpha_0(X)\mathbb{E}[Y - \gamma_0(X)|X]] \\ &= 0 \ ,\end{aligned}$$

while for the second moment by the triangle inequality and Assumptions 2.4(i)(ii), 2.3,

$$\begin{aligned}\mathbb{E}[\psi_0(W_i)^2] &= \mathbb{E}[\|m(W,\gamma_0) - \theta_0 + \alpha_0(X)(Y - \gamma_0(X))\|^2] \\ &\leq \mathbb{E}[(\|m(W,\gamma_0) - \theta_0\| + \|\alpha_0(X)(Y - \gamma_0(X))\|)^2] \\ &\leq 2\left(\mathbb{E}\left[\|m(W,\gamma_0) - \mathbb{E}[m(W,\gamma_0)]\,\|^2\right] + \mathbb{E}\left[\|\alpha_0(X)\|^2\mathbb{E}\left[\|(Y - \gamma_0(X))\|^2|X\right]\right]\right) \\ &\leq 2\left(\mathbb{E}\left[\|m(W,\gamma_0)\|^2\right] + \|\alpha_0\|_\infty^2\mathbb{E}\left[\mathbb{E}[\|(Y - \gamma_0(X))\|^2|X]\right]\right) \\ &= O(1) \ .\end{aligned}$$

The second conclusion then follow by the CLT. ■

For the proof of Lemma 3.1, recall that $\Sigma^{(p)} \equiv \mathbb{E}[b^{(p)}(X)b^{(p)}(X)^\intercal]$ and $M^{(p)} \equiv \mathbb{E}\left[m_{b^{(p)}}(W)\right]$. We now define their sample analogs

$$\hat{\Sigma}_{-k}^{(p)} \equiv \frac{1}{n_{-k}}\sum_{i\in I_{-k}} b^{(p)}(X_i)b^{(p)}(X_i)^\intercal \quad \text{and} \quad \hat{M}_{-k}^{(p)} \equiv \frac{1}{n_{-k}}\sum_{i\in I_{-k}} m_{b^{(p)}}(W_i) \ ,$$

$$\tilde{\Sigma}_{-k}^{(p)} \equiv \frac{1}{n_{-k}}\sum_{i\in I_{-k}} \tilde{b}_{-k}^{(p)}(X_i)\tilde{b}_{-k}^{(p)}(X_i)^\intercal \quad \text{and} \quad \tilde{M}_{-k}^{(p)} \equiv \frac{1}{n_{-k}}\sum_{i\in I_{-k}} m_{\tilde{b}_{-k}^{(p)}}(W_i) \ .$$

Besides, define

$$u_{-k} = \frac{1}{n_{-k}}\sum_{i\in I_{-k}} b^{(p)}(X_i) \ ,$$

so that we can write

$$\tilde{\Sigma}_{-k}^{(p)} \equiv \hat{\Sigma}_{-k}^{(p)} - u_{-k}u_{-k}^\intercal \ ,$$

and with Assumption 3.4,

$$\tilde{M}_{-k}^{(p)} \equiv \hat{M}_{-k}^{(p)} - u_{-k}\cdot\left(\frac{1}{n_{-k}}\sum_{i\in I_{-k}} m(W_i,1)\right) \ .$$

Proof of Lemma 3.1 We prove part (ii), while part (i) can be obtained in a similar

way. To simplify notation, we write $p_{n_{-k}}$ for $p_{\alpha,n_{-k}}$ and s for s_α in this proof. Before getting into the proof, observe that for each $k \geq 1$, by the CLT,

$$\|u_{-k}\|_2^2 = \sum_{j=1}^{p} \left(\frac{1}{n_{-k}} \sum_{i \in I_{-k}} b_j(X_i) \right)^2 = O_p\left(\frac{p}{n_{-k}} \right) .$$

Combining with Lemma B.1(i),(iii), we have

$$\begin{aligned} \left\| \tilde{\Sigma}_{-k}^{(p)} - \Sigma^{(p)} \right\|_{\mathrm{op}} &\leq \left\| \hat{\Sigma}_{-k}^{(p)} - \Sigma^{(p)} \right\|_{\mathrm{op}} + \|u_{-k}\|_2^2 \\ &= O_p\left(\sqrt{\frac{p \log p}{n_{-k}}} \right) . \end{aligned}$$

On the other hand, Assumption 2.3 and the CLT give

$$\left\| \frac{1}{n_{-k}} \sum_{i \in I_{-k}} m(W_i, 1) \right\|_2 = O_p(1) .$$

Therefore,

$$\begin{aligned} \left\| \tilde{M}_{-k}^{(p)} - M^{(p)} \right\|_{\mathrm{op}} &\leq \left\| \hat{M}_{-k}^{(p)} - M^{(p)} \right\|_{\mathrm{op}} + \|u_{-k}\|_2 \left\| \frac{1}{n_{-k}} \sum_{i \in I_{-k}} m(W_i, 1) \right\|_2 \\ &= O_p\left(\sqrt{\frac{p}{n_{-k}}} \right) . \end{aligned}$$

Now we begin the main proof. By Assumption 3.7(i), there exist constant $C_0 > 0$ such that

$$\begin{aligned} \|\hat{\alpha}_{-k} - \alpha_0\|_{L^2}^2 &\leq 2 \left(\|\hat{\alpha}_{-k} - \bar{\alpha}^{(p_{n_{-k}})}\|_{L^2}^2 + \|\bar{\alpha}^{(p_{n_{-k}})} - \alpha_0\|_{L^2}^2 \right) \\ &= 2 \left(\|\hat{\alpha}_{-k} - \bar{\alpha}^{(p_{n_{-k}})}\|_{L^2}^2 + C_0 \delta_\alpha^2(p_{n_{-k}}) \right) . \end{aligned} \tag{A.38}$$

Next, we inspect the leading term in (A.38). To begin, we fix $(b_j)_{j \geq 1}$ such that it satisfies Assumption 3.1. Since (i) $\bar{\alpha}^{(p_{n_{-k}})} \in \Gamma^{(p_{n_{-k}})}$ and (ii) $\{\tilde{b}_{j,-k}(X_i)\}_{i \in I_{-k}}$ are centered, by definition there exist $\bar{\rho} \in \mathbb{R}^{p_{n_{-k}}}$ such that

$$\bar{\alpha}^{(p_{n_{-k}})} = \frac{1}{n_{-k}} \sum_{i \in I_{-k}} \bar{\alpha}^{(p_{n_{-k}})}(X_i) + \tilde{b}_{-k}^{(p_{n_{-k}})}(X)^\intercal \bar{\rho} .$$

Therefore, we can write

$$
\begin{aligned}
&\|\hat{\alpha}_{-k} - \bar{\alpha}^{(p_{n_{-k}})}\|_{L^2}^2 \\
&= \mathbb{E}\left[\left\{\hat{\alpha}_{-k}(X) - \bar{\alpha}^{(p_{n_{-k}})}(X)\right\}^2 \,\middle|\, \{W_j : j \in I_{-k}\}\right] \\
&= \mathbb{E}\Bigg[\Bigg\{\left(\tilde{b}_{-k}^{(p_{n_{-k}})}(X)^\intercal \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right)\right) \\
&\qquad\qquad + \left(\frac{1}{n_{-k}} \sum_{i \in I_{-k}} m(W_i, 1) - \frac{1}{n_{-k}} \sum_{i \in I_{-k}} \bar{\alpha}^{(p_{n_{-k}})}(X_i)\right)\Bigg\}^2 \,\Bigg|\, \{W_j : j \in I_{-k}\}\Bigg] \\
&\leq 2\Bigg\{ \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right)^\intercal \left(\Sigma^{(p_{n_{-k}})}\right) \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right) \\
&\qquad\qquad + \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right)^\intercal \left(u_{-k} u_{-k}^\intercal\right) \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right) \\
&\qquad\qquad + \left(\frac{1}{n_{-k}} \sum_{i \in I_{-k}} m(W_i, 1) - \frac{1}{n_{-k}} \sum_{i \in I_{-k}} \bar{\alpha}^{(p_{n_{-k}})}(X_i)\right)^2 \Bigg\} \\
&\leq 2\left(\|(\Sigma^{(p_{n_{-k}})})^{1/2} \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right)\|_2^2 + \|\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\|_2^2 \|u_{-k}\|_2^2 + O_p\left(n_{-k}^{-1}\right)\right) ,
\end{aligned}
\tag{A.39}
$$

where the last line is due to applying the CLT given that (3) allows us to write

$$
\mathbb{E}[m(W, 1)] = \mathbb{E}[\alpha_0(X)] = \mathbb{E}[\bar{\alpha}^{(p_{n_{-k}})}(X)]
$$

with $\alpha_0 \in L^2(P_{0,X})$ under Assumption 2.3. The remainder of the proof concerns the leading two terms. We first focus on the middle term in (A.39), which is

$$
\|\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\|_2^2 \|u_{-k}\|_2^2 .
$$

From Assumptions 3.2(i) and 3.3, we have

$$
\|\bar{\rho}\|_2 = \left\|\left(\Sigma^{(p_{n_{-k}})}\right)^s h_\alpha^{(p_{n_{-k}})}\right\|_2 \leq \left\|\left(\Sigma^{(p_{n_{-k}})}\right)^s\right\|_{\text{op}} \left\|h^{(p_{n_{-k}})}\right\|_2 = O(1) . \tag{A.40}
$$

Next, see that from the normal equation $M^{(p)} = \Sigma^{(p)}\bar{\rho}^{(p)}$,

$$\begin{aligned}
&\left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)\left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right) \\
&= \tilde{M}_{-k}^{(p_{n_{-k}})} - \left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)\bar{\rho} \\
&= \left(\tilde{M}_{-k}^{(p_{n_{-k}})} - M^{(p_{n_{-k}})}\right) + \left(\Sigma^{(p_{n_{-k}})} - \tilde{\Sigma}_{-k}^{(p_{n_{-k}})}\right)\bar{\rho} - \zeta_{n_{-k}}\bar{\rho}\ .
\end{aligned}$$

Therefore,

$$\left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right) = \left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1}\left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)\left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right)$$

and

$$\begin{aligned}
\left\|\left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right)\right\|_2 &\leq \frac{\left\|\tilde{M}_{-k}^{(p_{n_{-k}})} - M^{(p_{n_{-k}})}\right\|_2 + \left\|\Sigma^{(p_{n_{-k}})} - \tilde{\Sigma}_{-k}^{(p_{n_{-k}})}\right\|_{\mathrm{op}} \|\bar{\rho}\|_2}{\zeta_{n_{-k}}} + \|\bar{\rho}\|_2\ , \qquad \text{(A.41)} \\
&= O_p\left(\sqrt{\frac{p\log p}{n_{-k}}} \cdot \frac{1}{\zeta_{n_{-k}}}\right) + O(1) = O_p(1)\ ,
\end{aligned}$$

where the conclusion is drawn by Assumption 3.9 (i). Therefore,

$$\|\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\|_2^2 \|u_{-k}\|_2^2 = O_p\left(\frac{p}{n_{-k}}\right)\ .$$

Now we move on to the leading term. Focus again on $\left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right)$. By subtracting and then adding back the expression

$$\tilde{\rho}(\zeta_{n_{-k}}) \equiv (\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} M^{(p_{n_{-k}})}\ ,$$

we can decompose the difference into two parts. Specifically,

$$\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho} = \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \tilde{\rho}(\zeta_{n_{-k}})\right) + \left(\tilde{\rho}(\zeta_{n_{-k}}) - \bar{\rho}\right)\ , \qquad \text{(A.42)}$$

where

$$\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \tilde{\rho}(\zeta_{n_{-k}}) = \left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} \tilde{M}_{-k}^{(p)} - \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} M^{(p_{n_{-k}})}$$

$$= \left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} \left(\tilde{M}_{-k}^{(p)} - M^{(p_{n_{-k}})}\right)$$

$$+ \left(\left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} - \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1}\right) M^{(p_{n_{-k}})} , \tag{A.43}$$

and

$$\tilde{\rho}(\zeta_{n_{-k}}) - \bar{\rho} = \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} M^{(p_{n_{-k}})} - \left(\Sigma^{(p_{n_{-k}})}\right)^{-1} M^{(p_{n_{-k}})}$$

$$= \left[\left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} - \left(\Sigma^{(p_{n_{-k}})}\right)^{-1}\right] M^{(p_{n_{-k}})} . \tag{A.44}$$

By using results in (A.42), (A.43), and (A.44), we have

$$\left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \bar{\rho}\right) \right\|_2^2$$

$$\leq 2\left(\left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\hat{\rho}_{-k}(\zeta_{n_{-k}}) - \tilde{\rho}(\zeta_{n_{-k}})\right) \right\|_2^2 + \left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\tilde{\rho}(\zeta_{n_{-k}}) - \bar{\rho}\right) \right\|_2^2 \right)$$

$$\lesssim \left\{ \left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} \left(\tilde{M}_{-k}^{(p)} - M^{(p_{n_{-k}})}\right) \right\|_2^2 \right.$$

$$\left. + \left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} - \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} \right) M^{(p_{n_{-k}})} \right\|_2^2 \right\}$$

$$+ \left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left[\left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} - \left(\Sigma^{(p_{n_{-k}})}\right)^{-1} \right] M^{(p_{n_{-k}})} \right\|_2^2 . \tag{A.45}$$

The proof is structured to inspect each of the three terms in (A.45), where the first two terms can be understood as the variance related terms and the last one is a bias related term. To begin the analysis on the first term, see that

$$\left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} = \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1/2} \left(I - \Delta_{n_{-k}}\right)^{-1} \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1/2} , \tag{A.46}$$

with

$$\left(I-\Delta_{n_{-k}}\right)^{-1} \equiv \left[I-\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}\left(\Sigma^{(p_{n_{-k}})}-\tilde{\Sigma}_{-k}^{(p_{n_{-k}})}\right)\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}\right]^{-1} ,$$

and given Assumption 3.9,

$$\begin{aligned}
\|\Delta_{n_{-k}}\|_{\text{op}} &= \left\|\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}\left(\Sigma^{(p_{n_{-k}})}-\tilde{\Sigma}_{-k}^{(p_{n_{-k}})}\right)\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}\right\|_{\text{op}} \\
&\leq \left\|\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1}\right\|_{\text{op}}\left\|\Sigma^{(p_{n_{-k}})}-\tilde{\Sigma}_{-k}^{(p_{n_{-k}})}\right\|_{\text{op}} \\
&= O_p\left(\frac{1}{\zeta_{n_{-k}}}\sqrt{\frac{p_{n_{-k}}\log p_{n_{-k}}}{n_{-k}}}\right) = o_p(1) .
\end{aligned}$$

Whenever $\|\Delta_{n_{-k}}\|_{\text{op}} = \epsilon$ for any arbitrarily small $\epsilon \in (0,1)$, $\left(I-\Delta_{n_{-k}}\right)^{-1}$ is a Neumann series and

$$\left\|\left(I-\Delta_{n_{-k}}\right)^{-1}\right\|_{\text{op}} \leq \frac{1}{1-\epsilon} .$$

Therefore,

$$\left\|\left(I-\Delta_{n_{-k}}\right)^{-1}\right\|_{\text{op}} = O_p(1) . \tag{A.47}$$

By using results in (A.46) and (A.47),

$$\begin{aligned}
&\left\|\left(\Sigma^{(p_{n_{-k}})}\right)^{1/2}(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})}+\zeta_{n_{-k}}I)^{-1}\right\|_{\text{op}} \\
&\quad= \left\|\left(\Sigma^{(p_{n_{-k}})}\right)^{1/2}\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}(I-\Delta)^{-1}\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}\right\|_{\text{op}} \\
&\quad\leq \left\|\left(\Sigma^{(p_{n_{-k}})}\right)^{1/2}\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}\right\|_{\text{op}}\|(I-\Delta)^{-1}\|_{\text{op}}\left\|\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1/2}\right\|_{\text{op}} \\
&\quad= \left(\sup_j\left(\frac{\nu_j^{(p_{n_{-k}})}}{\nu_j^{(p_{n_{-k}})}+\zeta_{n_{-k}}}\right)^{1/2}\right)\|(I-\Delta)^{-1}\|_{\text{op}}\left(\sup_j\frac{1}{\sqrt{\nu_j^{(p_{n_{-k}})}+\zeta_{n_{-k}}}}\right) \\
&\quad= \left(\frac{1}{1+\zeta_{n_{-k}}/\nu_1^{(p_{n_{-k}})}}\right)^{1/2}\|(I-\Delta)^{-1}\|_{\text{op}}\left(\frac{1}{\sqrt{\zeta_{n_{-k}}}\cdot\sqrt{\nu_{p_{n_{-k}}}^{(p_{n_{-k}})}/\zeta_{n_{-k}}+1}}\right) \\
&\quad= O_p\left(\zeta_{n_{-k}}^{-1/2}\right) ,
\end{aligned} \tag{A.48}$$

where the last two steps is due to Assumptions 3.2 and 3.9. Therefore,

$$\begin{aligned}
&\left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} \left(\tilde{M}_{-k}^{(p)} - M^{(p_{n_{-k}})}\right) \right\|_2^2 \\
&\quad \leq \left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} (\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} \right\|_{\text{op}}^2 \left\| \tilde{M}_{-k}^{(p)} - M \right\|_2^2 \\
&\quad = O_p\left(\zeta_{n_{-k}}^{-1} \cdot \frac{p_{n_{-k}}}{n_{-k}}\right) . \qquad \text{(A.49)}
\end{aligned}$$

Now move on to the second term in (A.45). First observe that

$$\begin{aligned}
&\left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left((\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} - (\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} \right) \\
&= \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} (\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} \left((\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I) - (\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I) \right) (\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} \\
&= \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} (\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} \left(\Sigma^{(p_{n_{-k}})} - \tilde{\Sigma}_{-k}^{(p_{n_{-k}})} \right) (\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} .
\end{aligned}$$

Therefore, we can bound the second term by

$$\begin{aligned}
&\left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} - \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} \right) M^{(p_{n_{-k}})} \right\|_2 \\
&= \left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} (\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} \left(\Sigma^{(p_{n_{-k}})} - \tilde{\Sigma}_{-k}^{(p_{n_{-k}})} \right) (\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} M^{(p_{n_{-k}})} \right\|_2 \\
&\leq \left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} (\tilde{\Sigma}_{-k}^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} \right\|_{\text{op}} \left\| \Sigma^{(p_{n_{-k}})} - \tilde{\Sigma}_{-k}^{(p_{n_{-k}})} \right\|_{\text{op}} \left\| (\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I)^{-1} M^{(p_{n_{-k}})} \right\|_2 .
\end{aligned}$$

With source condition stated in Assumption 3.3, together with Assumption 3.2, we have

$$\begin{aligned}
\left\|(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I)^{-1}M^{(p_{n_{-k}})}\right\|_2 &= \left\|(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I)^{-1}\left(\Sigma^{(p_{n_{-k}})}\right)^{s+1}h^{(p_{n_{-k}})}\right\|_2 \\
&\le \left\|(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I)^{-1}\left(\Sigma^{(p_{n_{-k}})}\right)^{s+1}\right\|_{\text{op}}\left\|h^{(p_{n_{-k}})}\right\|_2 \\
&= \left(\sup_j \frac{\left(\nu_j^{(p_{n_{-k}})}\right)^{s+1}}{\nu_j^{(p_{n_{-k}})}+\zeta_{n_{-k}}}\right)\left\|h^{(p_{n_{-k}})}\right\|_2 \\
&= \left(\sup_j \frac{1}{\left(\nu_j^{(p_{n_{-k}})}\right)^{-s}\left[1+\zeta_{n_{-k}}/\nu_j^{(p_{n_{-k}})}\right]}\right)\left\|h^{(p_{n_{-k}})}\right\|_2 \\
&= \left(\left(\nu_1^{(p_{n_{-k}})}\right)^{s}\frac{1}{1+\zeta_{n_{-k}}/\nu_1^{(p_{n_{-k}})}}\right)\left\|h^{(p_{n_{-k}})}\right\|_2 \\
&= O_p(1)\ .
\end{aligned}$$

Combined with the result stated in Lemma B.1(i) and (A.48), we have for the second term

$$\begin{aligned}
&\left\|\left(\Sigma^{(p_{n_{-k}})}\right)^{1/2}\left(\left(\tilde{\Sigma}_{-k}^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1}-\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1}\right)M^{(p_{n_{-k}})}\right\|_2^2 \\
&= O_p\left(\zeta_{n_{-k}}^{-1}\cdot\frac{p_{n_{-k}}\log p_{n_{-k}}}{n_{-k}}\right)\ . \qquad \text{(A.50)}
\end{aligned}$$

Finally for the third term, since

$$\begin{aligned}
&\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1}-\left(\Sigma^{(p_{n_{-k}})}\right)^{-1} \\
&= \left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1}\left(\Sigma^{(p_{n_{-k}})}-\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)\right)\left(\Sigma^{(p_{n_{-k}})}\right)^{-1} \\
&= -\zeta_{n_{-k}}\left(\Sigma^{(p_{n_{-k}})}+\zeta_{n_{-k}}I\right)^{-1}\left(\Sigma^{(p_{n_{-k}})}\right)^{-1}\ ,
\end{aligned}$$

again with Assumptions 3.2 and 3.3 ,

$$\begin{aligned}
&\left\| \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left[\left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} - \left(\Sigma^{(p_{n_{-k}})}\right)^{-1} \right] M^{(p_{n_{-k}})} \right\|_2 \\
&= \left\| -\zeta_{n_{-k}} \left(\Sigma^{(p_{n_{-k}})}\right)^{1/2} \left(\Sigma^{(p_{n_{-k}})} + \zeta_{n_{-k}} I\right)^{-1} \left(\Sigma^{(p_{n_{-k}})}\right)^{-1} \left(\Sigma^{(p_{n_{-k}})}\right)^{s+1} h^{(p_{n_{-k}})} \right\|_2 \\
&\leq \zeta_{n_{-k}} \left(\sup_j \left(\nu_j^{(p_{n_{-k}})}\right)^{s-1/2} \frac{1}{1 + \zeta_{n_{-k}}/\nu_j^{(p_{n_{-k}})}} \right) \left\| h^{(p_{n_{-k}})} \right\|_2 \\
&= \zeta_{n_{-k}} \left(\left(\nu_1^{(p_{n_{-k}})}\right)^{s-1/2} \frac{1}{1 + \zeta_{n_{-k}}/\nu_1^{(p_{n_{-k}})}} \right) \left\| h^{(p_{n_{-k}})} \right\|_2 \\
&= O_p(\zeta_{n_{-k}}) \ . \qquad \text{(A.51)}
\end{aligned}$$

Finally combining (A.45), (A.49), (A.50) and (A.51), we arrive at the stated result for the Riesz representer. ■

Proof of Lemma 4.1 Following Assumptions 3.8 and 4.1, we have for $\tau_\alpha, \tau_\gamma \in (0,1)$

$$p_{f,n_{-k}} \log p_{f,n_{-k}} \asymp n_{-k}^{\tau_f} \ , \quad f \in \{\alpha, \gamma\} \ ,$$

and the penalties

$$\zeta_{n_{-k}} \asymp n_{-k}^{(\tau_\alpha - 1)/3} \quad \text{with} \quad \lambda_{n_{-k}} \asymp n_{-k}^{(\tau_\gamma - 1)/3} \ .$$

Finally, combining with Assumption 4.2, we have for each nuisance

$$e_f\left(\tau_f; \kappa_f\right) = \min\left(\tau_f \kappa_f, \frac{1}{3}(1 - \tau_f)\right) \ , \quad f \in \{\alpha, \gamma\} \ .$$

■

Proof of Proposition 4.1 Let $\underline{\kappa} \equiv \min\{\kappa_\alpha, \kappa_\gamma\}$, $\bar{\kappa} \equiv \max\{\kappa_\alpha, \kappa_\gamma\}$, and assume Assumption 4.3 holds. We rewrite the inequality that define the product-rate warrant set as

$$\min\left(\tau_\alpha \kappa_\alpha, \frac{1}{3}(1 - \tau_\alpha)\right) + \min\left(\tau_\gamma \kappa_\gamma, \frac{1}{3}(1 - \tau_\gamma)\right) > \frac{1}{2} \ . \qquad \text{(A.52)}$$

To obtain the necessary condition on $(\tau_\alpha, \tau_\gamma)$, first consider

$$\min\left(\tau_f\kappa_f, \frac{1}{3}(1-\tau_f)\right) \leq \frac{1}{3}(1-\tau_f) \quad \text{for } f \in \{\alpha, \gamma\} \ . \tag{A.53}$$

Combining with (A.53), (A.52) becomes

$$\frac{1}{3}(1-\tau_\alpha) + \frac{1}{3}(1-\tau_\gamma) > \frac{1}{2} \iff \tau_\alpha + \tau_\gamma < \frac{1}{2} \ , \tag{A.54}$$

which gives the first necessary condition. Next, again for $f \in \{\alpha, \gamma\}$, given κ_f,

$$\tau_f^*(\kappa_f) \equiv \arg\max_{\tau_f}\left(\min\left(\tau_f\kappa_f, \frac{1}{3}(1-\tau_f)\right)\right) = \frac{1}{3\kappa_f + 1} \ ,$$

and therefore

$$\max_{\tau_f}\left(\min\left(\tau_f\kappa_f, \frac{1}{3}(1-\tau_f)\right)\right) = \frac{\kappa_f}{3\kappa_f + 1} \in (0, 1/3) \ .$$

To satisfy (A.52), the other necessary condition is thus

$$\frac{\underline{\kappa}}{3\underline{\kappa} + 1} + \frac{1}{3} > \frac{1}{2} \iff \underline{\kappa} > \frac{1}{3} \ . \tag{A.55}$$

Next, suppose we restrict $\tau_\alpha = \tau_\gamma = \tau$. Note that the above necessary conditions will still apply under this symmetric case, as the product-rate warrant set is now a subset of the previous one. Following the same logic that leads to (A.54), the first necessary condition becomes

$$\tau < \frac{1}{4} \ .$$

Meanwhile, let

$$G(\tau) \equiv \min\left(\tau\kappa_\alpha, \frac{1}{3}(1-\tau)\right) + \min\left(\tau\kappa_\gamma, \frac{1}{3}(1-\tau)\right) \ ,$$

which is concave and piecewise-linear function with kinks at

$$\tau^*(\underline{\kappa}) = \frac{1}{3\underline{\kappa} + 1} \ , \quad \tau^*(\bar{\kappa}) = \frac{1}{3\bar{\kappa} + 1} \ ,$$

and gives

$$G(\tau^*(\underline{\kappa})) = \frac{2\underline{\kappa}}{3\underline{\kappa} + 1} \ , \quad G(\tau^*(\bar{\kappa})) = \frac{\bar{\kappa} + \underline{\kappa}}{3\bar{\kappa} + 1} \ .$$

Next, notice that

$$G(\tau^*(\underline{\kappa})) - G(\tau^*(\bar{\kappa})) = \frac{2\underline{\kappa}(3\bar{\kappa}+1) - (\bar{\kappa}+\underline{\kappa})(3\underline{\kappa}+1)}{(3\underline{\kappa}+1)(3\bar{\kappa}+1)}$$
$$= \frac{(\bar{\kappa}-\underline{\kappa})(3\underline{\kappa}-1)}{(3\underline{\kappa}+1)(3\bar{\kappa}+1)} \geq 0 \ ,$$

where the conclusion follows from $\bar{\kappa} \geq \underline{\kappa}$, and $\underline{\kappa} > 1/3$ by (A.55). Therefore,

$$\tau^*(\kappa_\alpha, \kappa_\gamma) \equiv \arg\max_\tau \left(\min\left(\tau\kappa_\alpha, \frac{1}{3}(1-\tau)\right) + \min\left(\tau\kappa_\gamma, \frac{1}{3}(1-\tau)\right)\right) = \frac{1}{3\underline{\kappa}+1} \ ,$$

$$\max_\tau \left(\min\left(\tau\kappa_\alpha, \frac{1}{3}(1-\tau)\right) + \min\left(\tau\kappa_\gamma, \frac{1}{3}(1-\tau)\right)\right) = 2\left(\frac{\underline{\kappa}}{3\underline{\kappa}+1}\right) \ .$$

Therefore the other necessary condition becomes

$$2\left(\frac{\underline{\kappa}}{3\underline{\kappa}+1}\right) > \frac{1}{2} \iff \underline{\kappa} > 1 \ .$$

■

Proof of Lemma 4.2 The proof proceeds by showing that Assumption 2.5 in Lemma 2.1 can be replaced by Assumptions 3.1, 3.2, 3.3, 3.6, 3.7, 3.8, 3.9, 4.1, 4.2 and 4.3. We first combine results from Lemma 3.1 with Assumptions 4.1 and 4.2 to get for each $k \in \{1, \ldots, K\}$,

$$\|\hat{\gamma}_{-k} - \gamma_0\|_{L^2} = O_p\left(\max\left\{\left(\frac{p_{\gamma,n_{-k}} \log p_{\gamma,n_{-k}}}{n_{-k}}\right)^{1/3}, (p_{\gamma,n_{-k}} \log p_{\gamma,n_{-k}})^{-\kappa_\gamma}\right\}\right) \ , \quad \text{(A.56)}$$

$$\|\hat{\alpha}_{-k} - \alpha_0\|_{L^2} = O_p\left(\max\left\{\left(\frac{p_{\alpha,n_{-k}} \log p_{\alpha,n_{-k}}}{n_{-k}}\right)^{1/3}, (p_{\alpha,n_{-k}} \log p_{\alpha,n_{-k}})^{-\kappa_\alpha}\right\}\right) \ , \quad \text{(A.57)}$$

as $n_{-k} \to \infty$. Next, by plugging in the growth rates of the sieves as stated in Assumption 3.8, we have for both $f \in \{\alpha, \gamma\}$,

$$\left(\frac{p_{f,n_{-k}} \log p_{f,n_{-k}}}{n_{-k}}\right)^{1/3} \asymp n_{-k}^{(\tau_f - 1)/3} \ , \quad \text{(A.58)}$$

$$(p_{f,n_{-k}} \log p_{f,n_{-k}})^{-\kappa_f} \asymp n_{-k}^{-\tau_f \kappa_f} \ . \quad \text{(A.59)}$$

Finally, by putting (A.56), (A.57), (A.58), and (A.59) together, we can see that if

Assumption 4.3 holds, then as $n_{-k} \to \infty$,

$$\|\hat{\alpha}_{-k} - \alpha_0\|_{L^2} \|\hat{\gamma}_{-k} - \gamma_0\|_{L^2} = o_p\left(n_{-k}^{-1/2}\right) .$$

■

Proof of Lemma 4.3 For each nuisance $f \in \{\alpha, \gamma\}$, its convergence-rate exponent is a concave and piecewise-linear function with kink at

$$\tau_f \kappa_f = \frac{1}{3}(1 - \tau_f) \iff \tau_f = \frac{1}{3\kappa_f + 1} .$$

Therefore,

$$\tau_f^*(\kappa_f) \equiv \arg\max_{\tau_f} \left(e_f\left(\tau_f; \kappa_f\right)\right) = \frac{1}{3\kappa_f + 1} ,$$

and $e_f^*(\kappa_f)$ can then be obtained by $e_f(\tau_f^*(\kappa_f); \kappa_f)$. The conclusion on the product-rate exponent follows immediately as it is additively separable in τ_α and τ_γ. ■

Proof of Lemma 5.1 The proof combines Theorem 2.3 of Vaart et al. (2006) with the results in Lemma B.4 for the corresponding Bernstein pair (M^γ, v^γ) and (M^α, v^α) of the centered loss functions

$$L^\gamma(W, \gamma) \equiv \{Y - \gamma\}^2 - \{Y - \gamma_0\}^2 ,$$

$$L^\alpha(W, \alpha) \equiv \left\{\alpha^2 - 2m(W, \alpha)\right\} - \left\{\alpha_0^2 - 2m(W, \alpha_0)\right\} .$$

First notice that

$$\mathbb{E}[L^\gamma\left(W, \hat{\gamma}_{-v}\right) | \{W_j : j \in I_{-v}\}] = \|\hat{\gamma}_{-v} - \gamma_0\|_{L^2}^2 ,$$

$$\mathbb{E}[L^\alpha\left(W, \hat{\alpha}_{-v}\right) | \{W_j : j \in I_{-v}\}] = \|\hat{\alpha}_{-v} - \alpha_0\|_{L^2}^2 .$$

Writing in our notation, with the conditional version of Theorem 2.3 in Vaart et al. (2006) (which the last term of the statement is replaced following the convention stated in B.5 in Appendix B), set $p = 1$, and assume deterministic split, we have for $f \in \{\gamma, \alpha\}$,

$\mathcal{F}^f_{-v} = \{\hat{f}_{-v}(\mathbf{t}_f) : \mathbf{t}_f \in \mathbf{T}^f_{n^o_t}\}$, that the following is true for any $\delta > 0$:

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\hat{\mathbf{t}}_f) - f_0\right\|^2_{L^2}\right]$$

$$\leq (1+2\delta)\inf_{\mathbf{t}_f \in \mathbf{T}^f_{n^o_t}} \mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\mathbf{t}_f) - f_0\right\|^2_{L^2}\right] + (1+\delta)\frac{16\log\left(1+\left|\mathbf{T}^f_{n^o_t}\right|\right)}{n^{o\prime}_{t\prime}}$$

$$\times\, \mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\sup_{g\in\mathcal{F}^f_{-v}}\left[M^f(g) + \left(\frac{v^f(g)}{\|g-f_0\|^2_{L^2}}\right)\left(\frac{1+\delta}{\delta}\right)\right]\right] .$$

With Lemma B.6 and Assumption 2.4(ii), we have $\gamma_0, \alpha_0 \in L^\infty(P_{0,X})$. Therefore, with the additional assumption of for all $\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n^o_t}$ and $\mathbf{t}_\alpha \in \mathbf{T}^\alpha_{n^o_t}$, $\hat{\gamma}_{-v}(\mathbf{t}_\gamma), \hat{\alpha}_{-v}(\mathbf{t}_\alpha) \in L^\infty(P_{0,X})$, we can use Lemma B.4 to obtain for $f \in \{\gamma, \alpha\}$,

$$M^f \lesssim \|f - f_0\|^2_\infty + 1 ,$$

$$\frac{v^f}{\|f-f_0\|^2_{L^2}} \lesssim \|f - f_0\|^2_\infty + 1 .$$

Therefore,

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\hat{\mathbf{t}}_f) - f_0\right\|^2_{L^2}\right]$$

$$\leq (1+2\delta)\inf_{\mathbf{t}_f \in \mathbf{T}^f_{n^o_t}} \mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\mathbf{t}_f) - f_0\right\|^2_{L^2}\right]$$

$$+ (1+\delta)\frac{16\log\left(1+\left|\mathbf{T}^f_{n^o_t}\right|\right)}{n^{o\prime}_{t\prime}} \times \mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\max_{\mathbf{t}_f\in\mathbf{T}^f_{n^o_t}}\left[\left(1+\left\|\hat{f}_{-v}(\mathbf{t}_f) - f_0\right\|^2_\infty\right)\left(\frac{C(1+\delta)}{\delta}\right)\right]\right] .$$

■

Proof of Lemma 5.2 **Proof of (i): Expected maximum bound.** We begin by property (i), and prove it for γ. To simplify notation, we use p to denote p_γ in here. First of all, under Assumptions 3.1(ii)(iii), 3.2(i), 3.3(ii), and 5.2, we have from Lemma B.6 that $\gamma_0 \in L^\infty(P_{0,X})$, therefore for any $\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n^o_t}$,

$$\|\hat{\gamma}_{-v}(\mathbf{t}_\gamma) - \gamma_0\|^2_\infty \leq 2\left(\|\hat{\gamma}_{-v}(\mathbf{t}_\gamma)\|^2_\infty + \|\gamma_0\|^2_\infty\right) \lesssim 1 + \|\hat{\gamma}_{-v}(\mathbf{t}_\gamma)\|^2_\infty ,$$

Together with the independent fold construction in Assumptions 2.2(i)(iii) and 5.1(ii),

the proof proceed by bounding the following quantity:

$$\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \|\hat{\gamma}_{-v}(\mathbf{t}_\gamma)\|_\infty^2\right]$$

Next, see that

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \|\hat{\gamma}_{-v}(\mathbf{t}_\gamma)\|_\infty^2\right] \\
&\leq 2\mathbb{E}_{W_{-k}}\left[\left(\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}} Y_i\right)^2\right] + 2\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \left\|\tilde{b}^{(p)}_{-v}(X_i)^\intercal \hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_\infty^2\right] \\
&\leq 2\mathbb{E}\left[Y^2\right] + 2\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \left\|\tilde{b}^{(p)}_{-v}(X_i)^\intercal \hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_\infty^2\right] \ .
\end{aligned}$$

Since from Assumption 2.4(i) we have

$$\mathbb{E}[\{Y-\gamma_0(X)\}^2|X] \in L^\infty(P_{0,X}) \ ,$$

this gives

$$\mathbb{E}\left[Y^2\right] = C_Y < \infty \ .$$

For the second term, from Assumption 5.2 and recall that

$$\tilde{b}^{(p)}_{-v}(x) \equiv b^{(p)}(x) - \frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}} b^{(p)}(X_i) \ ,$$

therefore

$$\left\|\tilde{b}^{(p)}_{-v}(x)\right\|_2^2 \leq 4B$$

and thus

$$\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \left\|\tilde{b}^{(p)}(X_i)^\intercal \hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_\infty^2\right] \leq 4B\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_2^2\right] \ .$$

Combining everything we have until now, we get

$$\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \|\hat{\gamma}_{-v}(\mathbf{t}_\gamma)\|_\infty^2\right] \leq 2C_Y + 8B\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_2^2\right] \ .$$

Therefore, the last thing we need to consider is

$$\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}} \left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_2^2\right] .$$

Consider for all $\mathbf{t}_\gamma \in \mathbf{T}^\gamma_{n_t^o}$,

$$\begin{aligned}\left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_2^2 &= \left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma) - \bar{\beta}^{(p)} + \bar{\beta}^{(p)}\right\|_2^2 \\ &\leq 2\left(\left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma) - \bar{\beta}^{(p)}\right\|_2^2 + \left\|\bar{\beta}^{(p)}\right\|_2^2\right) , \end{aligned} \tag{A.60}$$

in which by the source condition and the spectral envelope in Assumption 3.3(ii) and 3.2(i), we know that

$$\sup_{p\geq 1}\left\|\bar{\beta}^{(p)}\right\|_2^2 = C_\beta < \infty$$

Applying the same algebra operation used to derive our learning-rate step in (A.41), and adapting it for the regression estimator, we obtain

$$\left\|\left(\hat{\beta}_{-v}(\lambda) - \bar{\beta}\right)\right\|_2 \leq \frac{\left\|\tilde{C}^{(p)}_{-v} - C^{(p)}\right\|_2 + \left\|\Sigma^{(p)} - \tilde{\Sigma}^{(p)}_{-v}\right\|_{\text{op}}\sqrt{C_\beta}}{\lambda} + \sqrt{C_\beta} ,$$

and therefore

$$\begin{aligned}&\left\|\left(\hat{\beta}_{-v}(\lambda) - \bar{\beta}\right)\right\|_2^2 \\ &\leq 2\left(\left\{\frac{\left\|\tilde{C}^{(p)}_{-v} - C^{(p)}\right\|_2 + \left\|\Sigma^{(p)} - \tilde{\Sigma}^{(p)}_{-v}\right\|_{\text{op}}\sqrt{C_\beta}}{\lambda}\right\}^2 + C_\beta\right) , \\ &\leq 2\left(\left\{2\frac{\left\|\tilde{C}^{(p)}_{-v} - C^{(p)}\right\|_2^2 + \left\|\Sigma^{(p)} - \tilde{\Sigma}^{(p)}_{-v}\right\|_{\text{op}}^2 C_\beta}{\lambda^2}\right\} + C_\beta\right) \end{aligned} \tag{A.61}$$

Combining (A.60) and (A.61), we have

$$\begin{aligned}&\left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_2^2 \\ &\leq 6\frac{\left\|\tilde{C}^{(p)}_{-v} - C^{(p)}\right\|_2^2 + \left\|\Sigma^{(p)} - \tilde{\Sigma}^{(p)}_{-v}\right\|_{\text{op}}^2 C_\beta}{\lambda^2} + 6C_\beta .\end{aligned}$$

Now we consider the worst case. Since from Lemma B.1,

$$\left\|\hat{C}_{-v}^{(p)} - C^{(p)}\right\|_2^2 = O_p\left(\frac{p}{n_t^{o\prime}}\right) , \quad \left\|\Sigma^{(p)} - \tilde{\Sigma}_{-v}^{(p)}\right\|_{\text{op}}^2 = O_p\left(\frac{p\log p}{n_t^{o\prime}}\right) ,$$

which increases with p, therefore

$$\arg\max_{\mathbf{t}_\gamma \in \mathbf{T}_{n_t^o}^\gamma} \left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_2^2 = \left(\bar{p} = \bar{p}_{n_t^o}, \quad \lambda_{\min} = 2^{-M_{n_t^o}} \cdot \left(\frac{2\log 2}{n_t^o}\right)^{1/3}\right) ,$$

and

$$\mathbb{E}_{W_{-k}}\left[\max_{\mathbf{t}_\gamma \in \mathbf{T}_{n_t^o}^\gamma} \left\|\hat{\beta}_{-v}(\mathbf{t}_\gamma)\right\|_2^2\right] \leq \mathbb{E}_{W_{-k}}\left[6\frac{\left\|\tilde{C}_{-v}^{(\bar{p})} - C^{(\bar{p})}\right\|_2^2 + \left\|\Sigma^{(\bar{p})} - \tilde{\Sigma}_{-v}^{(\bar{p})}\right\|_{\text{op}}^2 C_\beta}{\lambda_{\min}^2} + 6C_\beta\right] . \tag{A.62}$$

Let us look at the non constant term one by one. First, from Assumption 3.6,

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left[\left\|\tilde{C}_{-v}^{(\bar{p})} - C^{(\bar{p})}\right\|_2^2\right] \\
&\leq 2\mathbb{E}_{W_{-k}}\left[\left\|\hat{C}_{-k}^{(p)} - C^{(p)}\right\|_2^2\right] + 2\mathbb{E}_{W_{-k}}\left[\left\|\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}} b^{(\bar{p})}(X_i)\right\|_2^2 \left\|\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}} Y_i\right\|^2\right] \\
&\leq \frac{2}{n_t^{o\prime}}\text{Var}(b^{(\bar{p})}(X)Y) + 2\mathbb{E}_{W_{-k}}\left[\left\|\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}} b^{(\bar{p})}(X_i)\right\|_2^2 \left\|\mathbb{E}[Y] + \frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}} Y_i - \mathbb{E}[Y]\right\|^2\right] \\
&\leq \frac{2}{n_t^{o\prime}}\text{Var}(b^{(\bar{p})}(X)Y) \\
&\quad + 4\left(\mathbb{E}_{W_{-k}}\left[\left\|\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}} b^{(\bar{p})}(X_i)\right\|_2^2\right]\mathbb{E}[Y]^2 + B\cdot\mathbb{E}_{W_{-k}}\left[\frac{1}{n_t^{o\prime}}\left(\sum_{i\in I_{-v}} Y_i - \mathbb{E}[Y]\right)^2\right]\right) \\
&\leq \frac{6}{n_t^{o\prime}}BC_Y .
\end{aligned} \tag{A.63}$$

Next, we turn to

$$\begin{aligned}
\mathbb{E}_{W_{-k}} \left\| \Sigma^{(p)} - \tilde{\Sigma}_{-v}^{(\bar{p})} \right\|_{\mathrm{op}}^2 &\leq \mathbb{E}_{W_{-k}} \left\| \Sigma^{(\bar{p})} - \tilde{\Sigma}_{-v}^{(\bar{p})} \right\|_{\mathrm{F}}^2 \\
&\leq \left(\mathbb{E}_{W_{-k}} \left\| \Sigma^{(\bar{p})} - \hat{\Sigma}_{-v}^{(\bar{p})} \right\|_{\mathrm{F}} + \mathbb{E}_{W_{-k}} \left\| u_{-v} \right\|_2 \right)^2 \\
&\leq 2 \left(\mathbb{E}_{W_{-k}} \left\| \Sigma^{(\bar{p})} - \hat{\Sigma}_{-v}^{(\bar{p})} \right\|_{\mathrm{F}}^2 + \mathbb{E}_{W_{-k}} \left\| u_{-v} \right\|_2^2 \right) \\
&= 2(\frac{2B^2}{n_t^{o\prime}}) \\
&= \frac{4B^2}{n_t^{o\prime}} \ . \qquad \text{(A.64)}
\end{aligned}$$

Combining (A.62), (A.63) and (A.64), we arrive at

$$\mathbb{E}_{W_{-k}} \left[\max_{\mathbf{t}_\gamma \in \mathbf{T}_{n_t^o}^\gamma} \left\| \hat{\beta}_{-v}(\mathbf{t}_\gamma) \right\|_2^2 \right] \lesssim 1 + \frac{1}{n_t^{o\prime} \lambda_{\min}^2} \ .$$

Recall that $n_t^o = V n_t^{o\prime}/(V-1)$, and

$$\lambda_{\min} = 2^{-M_{n_t^o}} \cdot \left(\frac{2 \log 2}{n_t^o} \right)^{1/3} = 2^{-M_{n_t^o}} \cdot \left(\frac{V \cdot 2 \log 2}{(V-1) \cdot n_t^{o\prime}} \right)^{1/3}$$

which gives

$$\begin{aligned}
\frac{1}{n_t^{o\prime} \lambda_{\min}^2} &= 256 (n_t^o)^{2/9} \left\{ \log(e(n_t^o)) \right\}^{2/3} \cdot \left(\frac{V \cdot 2 \log 2}{(V-1)} \right)^{-2/3} (n_t^{o\prime})^{-1/3} \\
&= 256 \left(\frac{V}{(V-1)} \right) (2 \log 2)^{-2/3} \left\{ \log(e(n_t^o)) \right\}^{2/3} (n_t^o)^{-1/9} \\
&= o(1) \ .
\end{aligned}$$

The proof for α follows the same argument, with the assumptions replaced as stated. ■

Proof of (ii): Cardinality control. For each nuisance $f \in \{\gamma, \alpha\}$, with

$$\bar{p}_{n_t^o}^f \asymp \frac{n_t^o}{(\log n_t^o)^2} \ ,$$

$$M_{n_t^o} \asymp \log n_t^o \ ,$$

we have,

$$|\mathbf{T}_{n_t^o}^f| = (\bar{p}_{n_t^o}^f - 1) \cdot (M_{n_t^o} + 1) = O\left(\frac{n_t^o}{\log n_t^o} \right) \ ,$$

and thus

$$\log(1+|\mathbf{T}^f_{n^o_t}|) = O(\log n^o_t) = o(n^{o'}_{t'}) \ . \quad \blacksquare$$

Proof of (iii): Covering a rate-attaining choice. Consider $f \in \{\gamma, \alpha\}$. According to the result on learning rates in Lemma 3.1, when penalty terms follow the optimal decay rate of $(p_f \log p_f / n^o_t)^{1/3}$, convergence rate of its stochastic component is pinned down at

$$O_p\left(\left(\frac{p_f \log p_f}{n^o_t}\right)^{2/3}\right) \ .$$

Therefore, as we recall that the optimal exponent is

$$\frac{2\kappa_f}{3\kappa_f+1} \ ,$$

the optimal choice for p_f should follow

$$\left(\frac{p_f \log p_f}{n^o_t}\right)^{2/3} \asymp (n^o_t)^{-\frac{2\kappa_f}{3\kappa_f+1}} \quad \Leftrightarrow \quad p_f \log p_f \asymp (n^o_t)^{\frac{1}{3\kappa_f+1}} \ .$$

Thus, we have $\mathbf{t}^*_\gamma = (p^*_\gamma, \lambda^*)$ and $\mathbf{t}^*_\alpha = (p^*_\alpha, \zeta^*)$, such that

$$p^*_f \log p^*_f \asymp (n^o_t)^{\frac{1}{3\kappa_f+1}}, \quad \forall f \in \{\gamma, \alpha\} \ , \tag{A.65}$$

$$\lambda^* \asymp (p_\gamma \log p_\gamma / n^o_t)^{1/3} \, , \qquad \zeta^* \asymp (p_\alpha \log p_\alpha / n^o_t)^{1/3} \ .$$

It is trivial to see that such penalty choices are within our candidate sets $\Lambda_{n^o_t}$ and $\mathcal{Z}_{n^o_t}$, as each element in these sets is designed to grow at this specific rate. For the sieve dimension, as we described, the cap is designed to satisfy

$$\bar{p}_f \log \bar{p}_f = o(n^o_t) \ .$$

As $\kappa_f > 0$, we know that p^*_f will also sit within our set. Given such choice, if assumptions in Lemma 3.1 hold, together with Assumption 3.7, we have

$$\left\| \hat{f}_{-v}(\mathbf{t}^*_f) - f_0 \right\|^2_{L^2} = \ O_p\left((n^o_t)^{-\frac{2\kappa_f}{3\kappa_f+1}}\right).$$

Next, we need to translate this into an expectation statement. As we will borrow some intermediate steps from the proof of Lemma 3.1 (with k now switch to v as we are inside the cross-validation layer), below we present the proof for α, while the same argument applies for γ with the stated assumption replacement. To simplify notation, in what follows, we use p^* to denote p^*_α, and $\hat{\alpha}^*_{-v}$ to denote $\hat{\alpha}_{-v}(\mathbf{t}^*_\alpha)$. To begin, from (A.38) and

Assumption 4.2(i), we have

$$\|\hat{\alpha}^*_{-v} - \alpha_0\|^2_{L^2} \leq 2\left(\|\hat{\alpha}^*_{-v} - \bar{\alpha}^{(p^*)}\|^2_{L^2} + C_0\delta^2_\alpha(p^*)\right) , \tag{A.66}$$

$$\delta^2_\alpha(p) = O\left((p\log p)^{-2\kappa_\alpha}\right) . \tag{A.67}$$

Combining (A.65) with (A.67) gives

$$\delta^2_\alpha(p^*) = O\left((p^*\log p^*)^{-2\kappa_\alpha}\right) \asymp (n^o_t)^{-\frac{2\kappa_\alpha}{3\kappa_\alpha+1}} . \tag{A.68}$$

Now we look at the leading term in (A.66), which we have from (A.39),

$$\begin{aligned}
&\|\hat{\alpha}^*_{-v} - \bar{\alpha}^{(p^*)}\|^2_{L^2} \\
&\leq 2\Bigg\{ \left(\hat{\rho}^*_{-v} - \bar{\rho}\right)^\intercal \left(\Sigma^{(p^*)}\right) \left(\hat{\rho}^*_{-v} - \bar{\rho}\right) \\
&\qquad + \left(\hat{\rho}^*_{-v} - \bar{\rho}\right)^\intercal \left(u_{-v}u^\intercal_{-v}\right) \left(\hat{\rho}^*_{-v} - \bar{\rho}\right) \\
&\qquad + \left(\frac{1}{n^{o\prime}_t}\sum_{i\in I_{-v}} m(W_i,1) - \frac{1}{n^{o\prime}_t}\sum_{i\in I_{-v}} \bar{\alpha}^{(p^*)}(X_i)\right)^2 \Bigg\} .
\end{aligned} \tag{A.69}$$

We first concern the last term, taking expectation, we have

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left[\left(\frac{1}{n^{o\prime}_t}\sum_{i\in I_{-v}} m(W_i,1) - \frac{1}{n^{o\prime}_t}\sum_{i\in I_{-v}} \bar{\alpha}^{(p^*)}(X_i)\right)^2\right] \\
&\lesssim \frac{1}{n^{o\prime}_t}\left(\frac{1}{n^{o\prime}_t}\sum_{i\in I_{-v}} \mathbb{E}_{W_{-k}}\left[m(W_i,1)^2\right] + \frac{1}{n^{o\prime}_t}\sum_{i\in I_{-v}} \mathbb{E}_{W_{-k}}\left[\bar{\alpha}^{(p^*)}(X_i)^2\right]\right) \\
&\lesssim (n^{o\prime}_t)^{-1} ,
\end{aligned} \tag{A.70}$$

here the conclusion is drawn by the fact that (i) Assumptions 2.1(iii) and 3.4 give

$$\mathbb{E}_{W_{-k}}\left[m(W_i,1)^2\right] = O(1) ,$$

and (ii) with Assumptions 2.1 and 3.7, we have for $\delta_\alpha(p^*) \to 0$ as p^* grows with $n^{o\prime}_t$ and

therefore

$$\begin{aligned}\mathbb{E}_{W_{-k}}\left[\bar{\alpha}^{(p^*)}(X_i)^2\right] &= \left\|\bar{\alpha}^{(p^*)}\right\|_{L^2}^2 \\ &\lesssim \|\alpha_0\|_{L^2}^2 + \left\|\bar{\alpha}^{(p^*)} - \alpha_0\right\|_{L^2}^2 \\ &\lesssim 1 + \delta_\alpha(p^*)^2 = O(1)\ .\end{aligned}$$

Next, we move to the middle term of (A.69),

$$\begin{aligned}\mathbb{E}_{W_{-k}}\left[\left(\hat{\rho}^*_{-v} - \bar{\rho}\right)^\intercal \left(u_{-v}u_{-v}^\intercal\right)\left(\hat{\rho}^*_{-v} - \bar{\rho}\right)\right] &= \mathbb{E}_{W_{-k}}\| \left(\hat{\rho}^*_{-v} - \bar{\rho}\right)^\intercal u_{-v}\|_2^2 \\ &\leq \mathbb{E}_{W_{-k}}\|\hat{\rho}^*_{-v} - \bar{\rho}\|_2^2\|u_{-v}\|_2^2\ .\end{aligned}$$

With (A.40) and (A.41), we have

$$\|\bar{\rho}\|_2^2 = C_\rho < \infty\ , \tag{A.71}$$

and

$$\begin{aligned}&\mathbb{E}_{W_{-k}}\|\hat{\rho}^*_{-v} - \bar{\rho}\|_2^2\|u_{-v}\|_2^2 \\ &\lesssim \mathbb{E}_{W_{-k}}\left[\frac{\left\|\tilde{M}^{(p^*)}_{-v} - M^{(p^*)}\right\|_2^2\|u_{-v}\|_2^2 + \left\|\Sigma^{(p^*)} - \tilde{\Sigma}^{(p^*)}_{-v}\right\|_{\text{op}}^2\|\bar{\rho}\|_2^2\|u_{-v}\|_2^2}{(\zeta^*)^2} + C_\rho\|u_{-v}\|_2^2\right]\end{aligned}$$

For the last term,

$$\begin{aligned}\mathbb{E}_{W_{-k}}C_\rho\|u_{-v}\|_2^2 &= C_\rho\mathbb{E}_{W_{-k}}\|u_{-v}\|_2^2 \\ &= C_\rho\mathbb{E}_{W_{-k}}\sum_{j=1}^{p^*}\left(\frac{1}{n_t^{o'}}\sum_{i\in I_{-v}} b_j(X_i)\right)^2 \\ &\lesssim \frac{1}{n_t^{o'}}\left(\mathbb{E}_{W_{-k}}\sum_{j=1}^{p^*} b_j(X_i)^2\right) \\ &= \frac{1}{n_t^{o'}}\left(\mathbb{E}_{W_{-k}}\|b^{(p^*)}(X)\|_2^2\right) \\ &\lesssim (n_t^{o'})^{-1}\end{aligned}$$

By arguments in proof of Lemma B.1,

$$
\begin{aligned}
&\frac{1}{(\zeta^*)^2}\mathbb{E}_{W_{-k}}\left[\left\|\tilde{M}_{-v}^{(p^*)}-M^{(p^*)}\right\|_2^2\|u_{-v}\|_2^2\right]\\
&=\frac{1}{(\zeta^*)^2}\mathbb{E}_{W_{-k}}\left[\left\|\hat{M}_{-v}^{(p^*)}-u_{-v}\cdot\left(\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}}m(W_i,1)\right)-M^{(p^*)}\right\|_2^2\|u_{-v}\|_2^2\right]\\
&\lesssim\frac{1}{(\zeta^*)^2}\mathbb{E}_{W_{-k}}\left[\left\{\left\|\hat{M}_{-v}^{(p^*)}-M^{(p^*)}\right\|_2^2+\left\|u_{-v}\cdot\left(\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}}m(W_i,1)\right)\right\|_2^2\right\}\|u_{-v}\|_2^2\right]\\
&\lesssim\frac{1}{(\zeta^*)^2}\Bigg\{\mathbb{E}_{W_{-k}}\left[\left\|\hat{M}_{-v}^{(p^*)}-M^{(p^*)}\right\|_2^2\|u_{-v}\|_2^2\right]\\
&\qquad+\mathbb{E}_{W_{-k}}\left[\left\|u_{-v}\cdot\left(\frac{1}{n_t^{o\prime}}\sum_{i\in I_{-v}}m(W_i,1)\right)\right\|_2^2\|u_{-v}\|_2^2\right]\Bigg\}.
\end{aligned}
$$

From the leading term,

$$
\begin{aligned}
&\mathbb{E}_{W_{-k}}\left[\left\|\hat{M}_{-v}^{(p^*)}-M^{(p^*)}\right\|_2^2\|u_{-v}\|_2^2\right]\\
&=\frac{1}{(n_t^{o\prime})^4}\mathbb{E}_{W_{-k}}\left[\left(\left\|\sum_{i\in I_{-v}}\{m_{b^{(p^*)}}(W_i)-\mathbb{E}[m_{b^{(p^*)}}]\}\right\|_2^2\left\|\sum_{i\in I_{-v}}b^{(p^*)}(X_i)\right\|_2^2\right)\right]\\
&=\frac{1}{(n_t^{o\prime})^4}\sum_{i_1,i_2,i_3,i_4\in I_{-v}}\mathbb{E}_{W_{-k}}\Bigg[\Bigg(\{m_{b^{(p^*)}}(W_{i_1})-\mathbb{E}[m_{b^{(p^*)}}]\}^\intercal\{m_{b^{(p^*)}}(W_{i_2})-\mathbb{E}[m_{b^{(p^*)}}]\}\\
&\qquad\qquad\times b^{(p^*)}(X_{i_3})^\intercal b^{(p^*)}(X_{i_4})\Bigg)\Bigg]\\
&=\frac{(n_t^{o\prime})}{(n_t^{o\prime})^4}\mathbb{E}_{W_{-k}}\left[\left(\|m_{b^{(p^*)}}(W_{i_1})-\mathbb{E}[m_{b^{(p^*)}}]\|_2^2\left\|b^{(p^*)}(X_{i_1})\right\|_2^2\right)\right]\\
&\quad+\frac{(n_t^{o\prime})^2-n_t^{o\prime}}{(n_t^{o\prime})^4}\mathbb{E}_{W_{-k}}\left[\left(\|m_{b^{(p^*)}}(W_{i_1})-\mathbb{E}[m_{b^{(p^*)}}]\|_2^2\left\|b^{(p^*)}(X_{i_2})\right\|_2^2\right)\right]\\
&\quad+\frac{2((n_t^{o\prime})^2-n_t^{o\prime})}{(n_t^{o\prime})^4}\mathbb{E}_{W_{-k}}\Bigg[\{m_{b^{(p^*)}}(W_{i_1})-\mathbb{E}[m_{b^{(p^*)}}]\}^\intercal\{m_{b^{(p^*)}}(W_{i_2})-\mathbb{E}[m_{b^{(p^*)}}]\}\\
&\qquad\qquad\times b^{(p^*)}(X_{i_1})^\intercal b^{(p^*)}(X_{i_2})\Bigg)\Bigg]
\end{aligned}
\tag{A.72}
$$

All other pairs gives zero since $\mathbb{E}[b(X)] = 0$. Now notice that by Assumption 5.2,

$$\operatorname*{ess\,sup}_{x \in \mathcal{X}} \|b^{(p)}(X)\|_2^2 \leq B \ ,$$

Furthermore,

$$\begin{aligned}\mathbb{E}_{W_{-k}}\left[\|m_{b^{(p^*)}}(W_{i_1}) - \mathbb{E}[m_{b^{(p^*)}}]\|_2^2 \right] &\leq \mathbb{E}_{W_{-k}}\left[\|m_{b^{(p^*)}}(W_{i_1})\|_2^2 \right] \\ &\lesssim B.\end{aligned}$$

Therefore,

$$\mathbb{E}_{W_{-k}}\left[\left\| \hat{M}_{-v}^{(p^*)} - M^{(p^*)} \right\|_2^2 \|u_{-v}\|_2^2 \right] \lesssim (n_t^{o'})^{-2} \ .$$

We then look at the next term, which gives

$$\begin{aligned}&\mathbb{E}_{W_{-k}} \left\| \Sigma^{(p^*)} - \tilde{\Sigma}_{-v}^{(p^*)} \right\|_{\text{op}}^2 \|\bar{\rho}\|_2^2 \|u_{-v}\|_2^2 \\ &\lesssim \mathbb{E}_{W_{-k}} \left(\left\| \hat{\Sigma}_{-v}^{(p)} - \Sigma^{(p)} \right\|_{\text{op}}^2 + \|u_{-v}\|_2^4 \right) \|\bar{\rho}\|_2^2 \|u_{-v}\|_2^2 \\ &= \mathbb{E}_{W_{-k}} \left\| \hat{\Sigma}_{-v}^{(p)} - \Sigma^{(p)} \right\|_{\text{op}}^2 \|\bar{\rho}\|_2^2 \|u_{-v}\|_2^2 + \mathbb{E}_{W_{-k}} \|\bar{\rho}\|_2^2 \|u_{-v}\|_2^6\end{aligned}$$

Applying (A.71),

$$\begin{aligned}&\mathbb{E}_{W_{-k}} \left\| \Sigma^{(p^*)} - \tilde{\Sigma}_{-v}^{(p^*)} \right\|_{\text{op}}^2 \|\bar{\rho}\|_2^2 \|u_{-v}\|_2^2 \\ &\lesssim \mathbb{E}_{W_{-k}} \left\| \hat{\Sigma}_{-v}^{(p^*)} - \Sigma^{(p^*)} \right\|_{\text{op}}^2 \|u_{-v}\|_2^2 + \mathbb{E}_{W_{-k}} \|u_{-v}\|_2^6 \\ &\lesssim \mathbb{E}_{W_{-v}} \left\| \hat{\Sigma}_{-v}^{(p^*)} - \Sigma^{(p^*)} \right\|_{\text{op}}^2 \|u_{-v}\|_2^2 + B\mathbb{E}_{W_{-k}} \|u_{-v}\|_2^4 \\ &\lesssim (n_t^{o'})^{-2},\end{aligned}$$

by following the same argument before and using Frobenius norm to bound the operator norm. Therefore,

$$\begin{aligned}\mathbb{E}_{W_{-k}} \|\hat{\rho}_{-v}^* - \bar{\rho}\|_2^2 \|u_{-v}\|_2^2 &\lesssim (n_t^{o'})^{-2} (\zeta^*)^{-2} \\ &\asymp (p^* \log p^* / n_t^o)^{-2/3} (n_t^{o'})^{-2} \\ &\asymp (n_t^o)^{-\frac{4\kappa_\alpha + 2}{3\kappa_\alpha + 1}} \ .\end{aligned}$$

Finally, we turn to first term of (A.69),

$$\mathbb{E}_{W_{-k}}\left[\left(\hat{\rho}^*_{-v}-\bar{\rho}\right)^\intercal\left(\Sigma^{(p^*)}\right)\left(\hat{\rho}^*_{-v}-\bar{\rho}\right)\right]=\mathbb{E}_{W_{-k}}\left\|\left(\Sigma^{(p^*)}\right)^{1/2}\left(\hat{\rho}^*_{-v}-\bar{\rho}\right)^\intercal\right\|_2^2 .$$

From intermediate steps in the proof of Lemma 3.1,

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left\|\left(\Sigma^{(p^*)}\right)^{1/2}\left(\hat{\rho}^*_{-v}-\bar{\rho}\right)^\intercal\right\|_2^2\\
&\lesssim \mathbb{E}_{W_{-k}}\left\{\left\|\left(\Sigma^{(p^*)}\right)^{1/2}\left(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I\right)^{-1}\left(\tilde{M}^{(p^*)}_{-v}-M^{(p^*)}\right)\right\|_2^2\right.\\
&\qquad\left.+\left\|\left(\Sigma^{(p^*)}\right)^{1/2}\left(\left(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I\right)^{-1}-\left(\Sigma^{(p^*)}+\zeta^* I\right)^{-1}\right)M^{(p^*)}\right\|_2^2\right\}\\
&\qquad+\mathbb{E}_{W_{-k}}\left\|\left(\Sigma^{(p^*)}\right)^{1/2}\left[\left(\Sigma^{(p^*)}+\zeta^* I\right)^{-1}-\left(\Sigma^{(p^*)}\right)^{-1}\right]M^{(p^*)}\right\|_2^2\\
&\le \mathbb{E}_{W_{-k}}\left\|\left(\Sigma^{(p^*)}\right)^{1/2}(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\right\|_{\mathrm{op}}^2\left\|\tilde{M}^{(p^*)}_{-v}-M\right\|_2^2\\
&\qquad+\mathbb{E}_{W_{-k}}\left\|\left(\Sigma^{(p^*)}\right)^{1/2}(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\right\|_{\mathrm{op}}^2\left\|\Sigma^{(p^*)}-\tilde{\Sigma}^{(p^*)}_{-v}\right\|_{\mathrm{op}}^2\left\|(\Sigma^{(p^*)}+\zeta^* I)^{-1}M^{(p^*)}\right\|_2^2\\
&\qquad+(\zeta^*)^2\cdot\mathbb{E}_{W_{-k}}\left[\left(\sup_j\left(\nu_j^{(p^*)}\right)^{s-1/2}\frac{1}{1+\zeta^*/\nu_j^{(p^*)}}\right)^2\left\|h^{(p^*)}\right\|_2^2\right] .
\end{aligned}$$

The last term is $O\left((\zeta^*)^2\right)$. The first term and the second term share the same element

$$\begin{aligned}
&\left\|\left(\Sigma^{(p^*)}\right)^{1/2}(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\right\|_{\mathrm{op}}^2\\
&=\left\|(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\left(\Sigma^{(p^*)}\right)(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\right\|_{\mathrm{op}}\\
&\le\left\|(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\left(\Sigma^{(p^*)}-\tilde{\Sigma}^{(p^*)}_{-v}\right)(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\right\|_{\mathrm{op}}\\
&\qquad+\left\|(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\left(\tilde{\Sigma}^{(p^*)}_{-v}\right)(\tilde{\Sigma}^{(p^*)}_{-v}+\zeta^* I)^{-1}\right\|_{\mathrm{op}}\\
&\le\frac{1}{(\zeta^*)^2}\|\Sigma^{(p^*)}-\tilde{\Sigma}^{(p^*)}_{-v}\|_{\mathrm{op}}+\frac{1}{4\zeta^*} ,
\end{aligned}$$

since $t/(t+\zeta)^2\le 1/(4\zeta)$ for $t\ge 0$ and $\zeta>0$. Also notice that by Assumption 3.3 we

can rewrite

$$\left\|(\Sigma^{(p^*)}+\zeta^* I)^{-1}M^{(p^*)}\right\|_2^2=\left\|(\Sigma^{(p^*)}+\zeta^* I)^{-1}\left(\Sigma^{(p^*)}\right)^{s_\alpha+1}h_\alpha^{(p^*)}\right\|_2^2=O_p(1)$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left\|\left(\Sigma^{(p^*)}\right)^{1/2}\left(\hat{\rho}_{-v}^*-\bar{\rho}\right)\right\|_2^2\\
&\lesssim\frac{1}{(\zeta^*)^2}\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}\left\|\tilde{M}_{-v}^{(p^*)}-M\right\|_2^2\right]\\
&\quad+\frac{1}{\zeta^*}\mathbb{E}_{W_{-k}}\left[\left\|\tilde{M}_{-v}^{(p^*)}-M\right\|_2^2\right]\\
&\quad+\frac{1}{(\zeta^*)^2}\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}\right]\\
&\quad+\frac{1}{\zeta^*}\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2\right]+(\zeta^*)^2\\
&=\frac{1}{\zeta^*}\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2+\left\|\tilde{M}_{-v}^{(p^*)}-M\right\|_2^2\right]\\
&\quad+\frac{1}{(\zeta^*)^2}\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}\left\{\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2+\left\|\tilde{M}_{-v}^{(p^*)}-M\right\|_2^2\right\}\right]+(\zeta^*)^2
\end{aligned}$$

For the first term,

$$\frac{1}{\zeta^*}\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2+\left\|\tilde{M}_{-v}^{(p^*)}-M\right\|_2^2\right]\lesssim\frac{1}{\zeta^*}\left(\frac{1}{n_t^{o'}}+\frac{1}{n_t^{o'}}\right)\lesssim\frac{1}{\zeta^* n_t^{o'}}$$

For the second term,

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}\left\{\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2+\left\|\tilde{M}_{-v}^{(p^*)}-M\right\|_2^2\right\}\right]\\
&\leq\mathbb{E}_{W_{-k}}\left[\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2\right]^{1/2}\mathbb{E}_{W_{-k}}\left[\left\{\|\Sigma^{(p^*)}-\tilde{\Sigma}_{-v}^{(p^*)}\|_{\mathrm{op}}^2+\left\|\tilde{M}_{-v}^{(p^*)}-M\right\|_2^2\right\}^2\right]^{1/2}\\
&\lesssim\frac{1}{(n_t^{o'})^{1/2}}\left((n_t^{o'})^{-2}+(n_t^{o'})^{-2}\right)^{1/2}\\
&\lesssim(n_t^{o'})^{-3/2}\ ,
\end{aligned}$$

where the second last line is due to the index counting as in (A.72). Finally, we can

conclude for first term of (A.69),

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left[\left(\hat{\rho}^*_{-v}-\bar{\rho}\right)^\intercal\left(\Sigma^{(p^*)}\right)\left(\hat{\rho}^*_{-v}-\bar{\rho}\right)\right] \\
&\lesssim \frac{1}{\zeta^*(n_t^{o\prime})}+\frac{1}{(\zeta^*)^2(n_t^{o\prime})^{3/2}}+(\zeta^*)^2 \\
&\asymp (n_t^o)^{-\frac{2\kappa_\alpha+1}{3\kappa_\alpha+1}}+(n_t^o)^{-\frac{2.5\kappa_\alpha+3/2}{3\kappa_\alpha+1}}+(n_t^o)^{-\frac{2\kappa_\alpha}{3\kappa_\alpha+1}} \\
&\asymp (n_t^o)^{-\frac{2\kappa_\alpha}{3\kappa_\alpha+1}} \ .
\end{aligned}$$

Finally,

$$\mathbb{E}_{W_{-k}}\|\hat{\alpha}^*_{-v}-\alpha_0\|^2_{L^2}\lesssim (n_t^o)^{-\frac{2\kappa_\alpha}{3\kappa_\alpha+1}}+(n_t^{o\prime})^{-1}+(n_t^o)^{-\frac{4\kappa_\alpha+2}{3\kappa_\alpha+1}}+(n_t^o)^{-\frac{2\kappa_\alpha}{3\kappa_\alpha+1}}\asymp (n_t^o)^{-\frac{2\kappa_\alpha}{3\kappa_\alpha+1}} \ ,$$

and since the data split is random, adding across the cross-validation folds and then averaging gives the desired result. ■

Proof of Theorem 5.1 Fix $f\in\{\gamma,\alpha\}$. Applying the oracle inequality in Lemma 5.1 with a fixed $\delta\in(0,1)$, together with properties (i)-(iii) established in Lemma 5.2, gives

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\hat{\mathbf{t}}_f)-f_0\right\|^2_{L^2}\right]=O\left((n_t^o)^{-\frac{2\kappa_f}{3\kappa_f+1}}\right) \ .$$

Next, see that

$$\begin{aligned}
&\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-k}(\hat{\mathbf{t}}_f)-f_0\right\|^2_{L^2}\right] \\
&=\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-k}(\hat{\mathbf{t}}_f)-\hat{f}_{-v}(\hat{\mathbf{t}}_f)+\hat{f}_{-v}(\hat{\mathbf{t}}_f)-f_0\right\|^2_{L^2}\right] \\
&\lesssim \mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-k}(\hat{\mathbf{t}}_f)-\hat{f}_{-v}(\hat{\mathbf{t}}_f)\right\|^2_{L^2}\right]+\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-v}(\hat{\mathbf{t}}_f)-f_0\right\|^2_{L^2}\right] \\
&\leq \mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\max_{\mathbf{t}_f\in\mathbf{T}^f_{n_t^o}}\left\|\hat{f}_{-k}(\mathbf{t}_f)-\hat{f}_{-v}(\mathbf{t}_f)\right\|^2_{L^2}\right]+O\left((n_t^o)^{-\frac{2\kappa_f}{3\kappa_f+1}}\right) \ .
\end{aligned}$$

Combining this with the refit stability in Assumption 5.5 thus gives

$$\mathbb{E}_{W_{-k}}\left[\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{f}_{-k}(\hat{\mathbf{t}}_f)-f_0\right\|^2_{L^2}\right]=O\left((n_t^o)^{-\frac{2\kappa_f}{3\kappa_f+1}}\right) \ .$$

Finally, Markov's inequality delivers the stated result. ■

Proof of Lemma 5.3 Theorem 5.1 and the strict frontier condition imply assumption 2.5 for the separately cross-validated nuisance estimators. The conclusion therefore follows from Lemma 2.1. ■

Appendix B Supporting Results

To derive convergence rates for the ridge estimators, we need to apply matrix concentration inequalities. Formal concentration results are stated in Lemma B.1 below.

Lemma B.1 (Matrix and vector concentration)**.** *Define, for each fold $k \in \{1, \dots, K\}$ and $p \geq 1$, the following key matrices and vectors:*

1. *Gram matrix*

$$\hat{\Sigma}_{-k}^{(p)} \equiv \frac{1}{n_{-k}} \sum_{i \in I_{-k}} b^{(p)}(X_i) b^{(p)}(X_i)^\intercal \in \mathbb{R}^{(p)\times(p)} \ ,$$

$$\Sigma^{(p)} \equiv \mathbb{E}\left[b^{(p)}(X) b^{(p)}(X)^\intercal \right] \in \mathbb{R}^{(p)\times(p)} \ .$$

2. *Cross moment*

$$\hat{C}_{-k}^{(p)} \equiv \frac{1}{n_{-k}} \sum_{i \in I_{-k}} b^{(p)}(X_i) Y_i \in \mathbb{R}^{(p)\times 1} \ ,$$

$$C^{(p)} \equiv \mathbb{E}\left[b^{(p)}(X_i) Y_i \right] \in \mathbb{R}^{(p)\times 1} \ .$$

3. *Average moment*

$$\hat{M}_{-k}^{(p)} \equiv \frac{1}{n_{-k}} \sum_{i \in I_{-k}} m_{b^{(p)}}(W_i) \in \mathbb{R}^{(p)\times 1} \ ,$$

$$M^{(p)} \equiv \mathbb{E}\left[m_{b^{(p)}}(W) \right] \in \mathbb{R}^{(p)\times 1} \ .$$

Suppose Assumptions 2.2(i)(iii) hold. If Assumptions 3.1(iv), 3.2(i)(ii), and 3.8 hold, then as $n_{-k} \to \infty$,

$$(i) \qquad \left\| \hat{\Sigma}_{-k}^{(p)} - \Sigma^{(p)} \right\|_{\text{op}} = O_p\left(\sqrt{\frac{p \log p}{n_{-k}}} \right) \ ,$$

On the other hand, if Assumption 3.6(i) is satisfied, then

$$(ii) \qquad \left\| \hat{C}_{-k}^{(p)} - C^{(p)} \right\|_2 = O_p\left(\sqrt{\frac{p}{n_{-k}}} \right) \ .$$

Lastly, if Assumption 3.6(ii) is true, then

$$(iii) \qquad \left\| \hat{M}_{-k}^{(p)} - M^{(p)} \right\|_2 = O_p\left(\sqrt{\frac{p}{n_{-k}}} \right) \ .$$

Proof of Lemma B.1 Fix $k \in \{1, \dots, K\}$. For the first part, with Assumptions 2.2(i)(iii), we utilize Corollary 1 from Appendix E of Kato (2013), which extends Rudelson's inequality (Rudelson, 1999) to establish non-asymptotic bounds for the centered sample average of random matrices. To begin, Assumption 3.5 gives

$\mathbb{E}\left[\max_{1\le i\le n_{-k}}\|b^{(p)}(X_i)\|_2^2\right] \lesssim p$. Take $m = C_0 p$, and thus $\delta = C_1\sqrt{p\log(2p)/n_{-k}}$, for some constants $C_0, C_1 > 0$. Assumption 3.8 guarantees that even though p grows with n_{-k}, it grows at a relatively slow rate of $p = o(n_{-k}/\log n_{-k})$. Therefore, combining with Assumption 3.2(ii) we have $\|\Sigma^{(p)}\|_{\text{op}}^{1/2} > \delta$ for n_{-k} sufficiently large. As a result, the term $\|\Sigma^{(p)}\|_{\text{op}}^{1/2}\cdot\delta$ dominates in the bound, yielding

$$\mathbb{E}\left[\left\|\hat{\Sigma}_{-k}^{(p)} - \Sigma^{(p)}\right\|_{\text{op}}\right] \le \|\Sigma^{(p)}\|_{\text{op}}^{1/2}\cdot\delta \le C_3\sqrt{\frac{p\log(2p+2)}{n_{-k}}}\ ,$$

for some constant $C_3 > 0$ due to Assumption 3.2(i). Finally, by Markov's inequality, for any $\epsilon > 0$,

$$P\left(\left\|\hat{\Sigma}_{-k}^{(p)} - \Sigma^{(p)}\right\|_{\text{op}} > \epsilon\sqrt{\frac{p\log(2p+2)}{n_{-k}}}\right) \le \frac{C_3}{\epsilon}\ .$$

Since $O_p(\sqrt{p\log(2p+2)/n_{-k}}) = O_p(\sqrt{p\log p/n_{-k}})$, we have established the stated rate. Next, for parts (ii)-(iii), we have from Assumption 3.6(i)(ii), for some finite constant $C_4 > 0$ and over coordinate index $j \in \{1,\dots,p\}$,

$$\sup_{j\ge1}\mathbb{E}\left[(b_j(X)Y)^2\right] \le C_4, \quad \text{and} \quad \sup_{j\ge1}\mathbb{E}\left[m(W,b_j)^2\right] \le C_4\ . \tag{B.1}$$

Now we proceed with the same argument applying to vectors of interest $\hat{C}_{-k}^{(p)} - C^{(p)}$ and $\hat{M}_{-k}^{(p)} - M^{(p)}$. Below we prove it for $\hat{C}_{-k}^{(p)} - C^{(p)}$, while the one for $\hat{M}_{-k}^{(p)} - M^{(p)}$ can be obtained in same way. First notice that for each coordinate index $j \in \{1,\dots,p\}$, we have from Assumptions 2.2(i)(iii),

$$\mathbb{E}\left[\left[\hat{C}_{-k}^{(p)} - C^{(p)}\right]_j^2\right] = \text{Var}\left(\left[\hat{C}_{-k}^{(p)}\right]_j\right) = \frac{1}{n_{-k}}\text{Var}\left(b_j(X)Y\right) \le \frac{1}{n_{-k}}\mathbb{E}\left[(b_j(X)Y)^2\right]$$

and therefore according to (B.1),

$$\mathbb{E}\left[\left\|\hat{C}_{-k}^{(p)} - C^{(p)}\right\|_2^2\right] = \sum_{j=1}^{p}\mathbb{E}\left[\left[\hat{C}_{-k}^{(p)} - C^{(p)}\right]_j^2\right] < \frac{p}{n_{-k}}\cdot C_4\ .$$

By Markov's inequality, for any $\epsilon > 0$

$$P\left(\frac{\left\|\hat{C}_{-k}^{(p)} - C^{(p)}\right\|_2^2}{p/n_{-k}} > \epsilon\right) \le \frac{C_4}{\epsilon}\ .$$

Therefore,

$$\left\|\hat{C}_{-k}^{(p)} - C^{(p)}\right\|_2^2 = O_p\left(\frac{p}{n_{-k}}\right)\ .$$

■

Definition B.1 (Bernstein pair, as in Vaart et al. (2006)). *The Bernstein pair of a function $f : X \to \mathbb{R}$ is defined as (M, v) that satisfies*

$$M^2\mathbb{E}\left[e^{|f|/M} - 1 - \frac{|f|}{M}\right] \leq \frac{1}{2}v \ .$$

Lemma B.2 (Enlargement of the Bernstein pairs). *If (M_0, v_0) is a Bernstein pair of $f : X \to \mathbb{R}$, then for any $M > M_0$ and $v > v_0$, (M, v) is also a Bernstein pair of f.*

Proof of Lemma B.2 Suppose (M_0, v_0) is a Bernstein pair of $f : X \to \mathbb{R}$, then according to the definition of the Bernstein pair in B.1,

$$M_0^2\mathbb{E}\left[e^{|f|/M_0} - 1 - \frac{|f|}{M_0}\right] \leq \frac{1}{2}v_0 \ .$$

First consider the left hand side of the inequality, for any $M > M_0$,

$$\begin{aligned} M^2\mathbb{E}\left[e^{|f|/M} - 1 - \frac{|f|}{M}\right] &= M^2\mathbb{E}\left[\sum_{r=2}^{\infty}\frac{|f|^r}{M^r \cdot r!}\right] \\ &= \mathbb{E}\left[\sum_{r=2}^{\infty}\frac{|f|^r}{M^{r-2} \cdot r!}\right] \\ &\leq \mathbb{E}\left[\sum_{r=2}^{\infty}\frac{|f|^r}{M_0^{r-2} \cdot r!}\right] \ . \end{aligned}$$

Therefore, left hand side of the inequality is decreasing in M, and it is easy to see that the right hand side of it is increasing in v. The conclusion then follows immediately. ■

Lemma B.3 (L^2-Lipschitz continuity of the moment function). *If Assumptions 2.1(iii) and 3.4 hold, then for any $h, h' \in L^2(P_{0,X})$, there exists $C_m \in (0, \infty)$ such that*

$$\mathbb{E}[\{m(W, h) - m(W, h')\}^2] \leq C_m\|h - h'\|_{L^2}^2 \ .$$

Proof of Lemma B.3 First, for any $t > 0$, since $h, h' \in L^2(P_{0,X})$, we have

$$\gamma_0 + t(h - h') \in L^2(P_{0,X})$$

Under Assumption 2.1(iii), there exist $\delta > 0$ such that

$$\|\gamma_0 + t(h - h') - \gamma_0\|_{L^2} < \delta \implies \mathbb{E}[\{m(W, \gamma_0 + t(h - h')) - m(W, \gamma_0)\}^2] < 1 \ .$$

Based on Assumption 3.4, rearranging gives

$$\frac{1}{t} > \frac{\|h - h'\|_{L^2}}{\delta} \implies \mathbb{E}[\{m(W, h - h')\}^2] < \frac{1}{t^2} .$$

Now choose any $C > 1$ and

$$t = \frac{1}{C} \frac{\delta}{\|h - h'\|_{L^2}} ,$$

we have

$$\mathbb{E}[\{m(W, h - h')\}^2] < \frac{C^2}{\delta^2} \|h - h'\|_{L^2}^2 .$$

The statement thus hold for $C_m = C^2/\delta^2$. The equality case hold when $\|h - h'\|_{L^2}^2 = 0$. ■

Lemma B.4 (Bernstein pair for the centered loss)**.** *Suppose $\gamma_0, \alpha_0 \in L^\infty(P_{0,X})$ and Assumptions 2.1(iii), 3.4, 5.3, and 5.4 hold. For any fixed $\gamma, \alpha \in L^\infty(P_{0,X})$, define the centered loss functions*

$$L^\gamma(W, \gamma) \equiv \{Y - \gamma\}^2 - \{Y - \gamma_0\}^2 ,$$

$$L^\alpha(W, \alpha) \equiv \left\{\alpha^2 - 2m(W, \alpha)\right\} - \left\{\alpha_0^2 - 2m(W, \alpha_0)\right\} .$$

A Bernstein pair for $w \mapsto L^\gamma(w, \gamma)$ is (M^γ, v^γ), with

$$M^\gamma = 4(\eta^{-1} \vee 1)(\|\gamma - \gamma_0\|_\infty^2 \vee 1) ,$$

$$v^\gamma = 2\|\gamma - \gamma_0\|_{L^2}^2 (e\|\gamma - \gamma_0\|_\infty^2 + 8\eta^{-2} \|\mathbb{E}[e^{\eta|Y - \gamma_0(X)|} | X]\|_\infty) .$$

On the other hand, a Bernstein pair for $w \mapsto L^\alpha(w, \alpha)$ is (M^α, v^α), with

$$M^\alpha = \|(\alpha - \alpha_0)\|_\infty \{\|(\alpha - \alpha_0)\|_\infty + 2(\|\alpha_0\|_\infty + C_\infty)\} ,$$

$$v^\alpha = \frac{3}{2} \left\{ \|\alpha - \alpha_0\|_{L^2}^2 \left(\|(\alpha - \alpha_0)\|_\infty + 2 \left(\|\alpha_0\|_\infty + \sqrt{C_m} \right) \right)^2 \right\} .$$

Proof of Lemma B.4 The proof is a direct application of results in Vaart et al. (2006). The first part concerning the regression function follows immediately from Lemma 4.1 therein. For the claim regarding the Riesz function, we make use of property (i) of the Bernstein number provided in p.7 of their paper. That is, if $f : \mathcal{X} \to \mathbb{R}$ is uniformly bounded, then a Bernstein pair for f is $(\|f\|_\infty, 1.5\|f\|_{L^2}^2)$. We begin the proof by showing that $w \mapsto L^\alpha(w, \alpha)$ is uniformly bounded. First note that since $\alpha, \alpha_0 \in L^\infty(P_{0,X})$,

$\|(\alpha - \alpha_0)\|_\infty < \infty$. Therefore,

$$\begin{aligned}
&\| \left\{\alpha^2 - 2m(W,\alpha)\right\} - \left\{\alpha_0^2 - 2m(W,\alpha_0)\right\} \|_\infty \\
&= \|(\alpha^2 - \alpha_0^2) + 2m(W, \alpha_0 - \alpha)\|_\infty \\
&= \|(\alpha - \alpha_0)(\alpha + \alpha_0) + 2m(W, \alpha_0 - \alpha)\|_\infty \\
&= \|(\alpha - \alpha_0)^2 + (\alpha - \alpha_0)(2\alpha_0) + 2m(W, \alpha_0 - \alpha)\|_\infty \\
&\leq \|(\alpha - \alpha_0)\|_\infty^2 + 2\|\alpha_0\|_\infty\|(\alpha - \alpha_0)\|_\infty + 2C_\infty\|(\alpha - \alpha_0)\|_\infty \\
&= \|(\alpha - \alpha_0)\|_\infty \left\{\|(\alpha - \alpha_0)\|_\infty + 2(\|\alpha_0\|_\infty + C_\infty)\right\} < \infty \, . \qquad \text{(B.2)}
\end{aligned}$$

Then, from Lemma B.2, we can conclude M^α from (B.2). Next, with Lemma B.3

$$\begin{aligned}
&\| \left\{\alpha^2 - 2m(W,\alpha)\right\} - \left\{\alpha_0^2 - 2m(W,\alpha_0)\right\} \|_{L^2} \\
&= \|(\alpha^2 - \alpha_0^2) + 2m(W, \alpha_0 - \alpha)\|_{L^2} \\
&= \|(\alpha - \alpha_0)^2 + (\alpha - \alpha_0)(2\alpha_0) + 2m(W, \alpha_0 - \alpha)\|_{L^2} \\
&\leq \|(\alpha - \alpha_0)\|_\infty\|(\alpha - \alpha_0)\|_{L^2} + 2\|\alpha_0\|_\infty\|\alpha - \alpha_0\|_{L^2} + 2\sqrt{C_m}\|\alpha - \alpha_0\|_{L^2} \, . \\
&= \|\alpha - \alpha_0\|_{L^2}(\|(\alpha - \alpha_0)\|_\infty + 2(\|\alpha_0\|_\infty + \sqrt{C_m})) \, .
\end{aligned}$$

Therefore,

$$v^\alpha = \frac{3}{2}\left\{\|\alpha - \alpha_0\|_{L^2}^2 \left(\|(\alpha - \alpha_0)\|_\infty + 2\left(\|\alpha_0\|_\infty + \sqrt{C_m}\right)\right)^2\right\} \, .$$

■

Lemma B.5 (Conditional version of Lemma 2.2 and Theorem 2.3 in Vaart et al. (2006))**.** *In here, we follow all the notations in Vaart et al. (2006). Conditioning on training sample, $\mathbb{G}_S^1$ is an empirical process of an i.i.d. sample of size n^1, and suppose we have $Pf \geq 0$ for all $f \in \mathcal{F}(\mathbb{P}_S^0) \equiv \{x \mapsto L(x, \theta_k(\mathbb{P}_S^0)) : k \in \mathcal{K}\}$. Then, for any Bernstein pairs $(M(f), v(f))$, and for any $\delta > 0$, $1 \leq p \leq 2$,*

$$\begin{aligned}
&\mathbb{E}_{S,\mathbb{P}_S^0}\mathbb{E}_{\mathbb{P}_S^1} \max_{f \in \mathcal{F}(\mathbb{P}_S^0)} (\mathbb{G}_S^1 - \delta\sqrt{n^1}P)f \\
&\leq \frac{8}{(n^1)^{1/p-1/2}} \log(1 + |\mathcal{F}(\mathbb{P}_S^0)|)\mathbb{E}_{S,\mathbb{P}_S^0} \max_{f \in \mathcal{F}(\mathbb{P}_S^0)} \left[\frac{M(f)}{(n^1)^{1-1/p}} + \left(\frac{v(f)}{(\delta Pf)^{2-p}}\right)^{1/p}\right] \, .
\end{aligned}$$

Based on this, the last term of the original Theorem 2.3 now becomes

$$(1+\delta)\frac{16\log(1+|\mathcal{F}(\mathbb{P}_S^0)|)}{(n^1)^{1/p}}\mathbb{E}_{S,\mathbb{P}_S^0}\sup_{f\in\mathcal{F}(\mathbb{P}_S^0)}\left[\frac{M(f)}{(n^1)^{1-1/p}}+\left(\frac{v(f)}{(Pf)^{2-p}}\right)^{1/p}\left(\frac{1+\delta}{\delta}\right)^{2/p-1}\right] .$$

Proof of Lemma B.5 The first result follows by applying the original Lemma 2.2 conditional on the split S and the training sample $\mathbb{P}_S^0$, and then taking expectation over $(S,\mathbb{P}_S^0)$. The second result follows the same proof of Theorem 2.3 in Vaart et al. (2006), with result in their Lemma 2.2 replaced by our first result. ■

Lemma B.6 (Uniformly bounded outcome regression)**.** *If Assumptions 3.1(ii)(iii), 3.2(i), 3.3(ii), and 5.2 hold, then* $\gamma_0 \in L^\infty(P_{0,X})$.

Proof of Lemma B.6 Recall that the definition of $\bar{\gamma}^{(p)}$ is

$$\bar{\gamma}^{(p)} \equiv \arg\min_{\gamma\in\Gamma^{(p)}} \mathbb{E}\big[(\gamma_0(X)-\gamma(X))^2\big] .$$

Therefore, there exists $\bar{\beta}_0 \in \mathbb{R}$ and $\bar{\beta}^{(p)} \in \mathbb{R}^{(p)}$ such that

$$\bar{\gamma}^{(p)}(X) = \bar{\beta}_0 + b^{(p)}(X)^\intercal \bar{\beta}^{(p)} . \tag{B.3}$$

Furthermore, since from Assumptions 3.1(ii), we have for each integer $j \geq 1$, $\mathbb{E}[b_j(X)] = 0$, which allows us to rewrite (B.3) as

$$\bar{\gamma}^{(p)}(X) = \mathbb{E}[Y] + b^{(p)}(X)^\intercal \bar{\beta}^{(p)} . \tag{B.4}$$

Next, combining the source condition in Assumption 3.3(ii) and the spectral envelope in Assumption 3.2(i), we get for some $s_\gamma \geq 1/2$,

$$\sup_{p\geq 1}\left\|\bar{\beta}^{(p)}\right\|_2 \leq \sup_{p\geq 1}\left\|\left(\Sigma^{(p)}\right)^{s_\gamma}\right\|_{\text{op}} \cdot \sup_{p\geq 1}\left\|h_\gamma^{(p)}\right\|_2 < \infty . \tag{B.5}$$

Since $\gamma_0 \in L^2(P_{0,X})$, it must be also true that $|\mathbb{E}[Y]| < \infty$, together with Assumption 5.2 and (B.5), we can draw on (B.4) that

$$\sup_{p\geq 1}\left\|\bar{\gamma}^{(p)}\right\|_\infty \leq |\mathbb{E}[Y]| + \sqrt{B}\cdot\sup_{p\geq 1}\left\|\left(\Sigma^{(p)}\right)^{s_\gamma}\right\|_{\text{op}} \cdot \sup_{p\geq 1}\left\|h_\gamma^{(p)}\right\|_2 = C < \infty .$$

Furthermore, with Assumption 3.1(iii), there must exists a subsequence p_k such that

$$\bar{\gamma}^{(p_k)} \to \gamma_0 , \qquad \text{almost everywhere} .$$

Finally, since

$$\sup_{p\geq 1}\left\|\bar{\gamma}^{(p)}\right\|_{\infty}\leq C\ ,$$

for every k we have $|\bar{\gamma}^{(p_k)}|\leq C$ almost everywhere, and

$$|\gamma_0|=\lim_{k\to\infty}|\bar{\gamma}^{(p_k)}|\leq C\ ,\quad \text{almost everywhere}\ .$$

■

Appendix C Neyman Orthogonality via Directional Derivatives

The idea to add the correction term[7] $\alpha(X)(Y-\gamma(X))$ to the original moment function is motivated by Newey (1994). Based on pathwise derivative calculation, Newey (1994) suggested that the map

$$W \mapsto \psi(W, \theta_0, \gamma_0, \alpha_0) \ ,$$

with ψ defined as in (2), corresponds to the influence function of the estimator that is based on the original moment function m under general misspecification. Building on this characterization, Chernozhukov et al. (2022a) connect this framework to orthogonality and use it to eliminate the first-order effects of first-step estimation errors on the moment function. Below we list the primitive conditions for the augmented version of the moment function stated in (2) to be Neyman orthogonal.

Let $\mathbf{P}$ be a statistical model for the distribution of W with true distribution $P_0 \in \mathbf{P}$. For each $P \in \mathbf{P}$, we let γ_P denote the probability limit of $\hat{\gamma}$ under i.i.d. sampling $W \sim P$. In particular, $\gamma_0 = \gamma_{P_0}$.

Assumption C.1 (Square-integrability of the nuisance)**.** *For every $P \in \mathbf{P}$, $\gamma_P \in L^2(P_{0,X})$.*

Lemma C.1 (Neyman orthogonality)**.** *Suppose Assumptions 2.1 and C.1 hold. Let $\psi(W, \theta, \gamma, \alpha)$ be the augmented moment function defined in (2). Then ψ is Neyman orthogonal at $(\theta_0, \gamma_0, \alpha_0)$, i.e.*

$$\frac{d}{dt}\,\mathbb{E}\big[\psi(W, \theta_0, \gamma_0 + t(\gamma_P - \gamma_0), \alpha_0)\big]\Big|_{t=0+} = 0, \qquad \forall P \in \mathbf{P} \ . \tag{C.1}$$

Proof of Lemma C.1 Fix $P \in \mathbf{P}$, in which $\mathbf{P}$ satisfies Assumption C.1. Let $R(h) \equiv \mathbb{E}[\alpha_0(X)h(X)]$. Recall from (3), we showed that Assumption 2.1 implies

$$\mathbb{E}[m(W, \gamma)] = R(\gamma), \quad \forall \gamma \in L^2(P_{0,X}) \ .$$

Let $\gamma_t \equiv \gamma_0 + t(\gamma_P - \gamma_0)$. By Assumption C.1, $\gamma_t \in L^2(P_{0,X})$. Then

$$\begin{aligned}
\mathbb{E}\left[\psi(W, \theta_0, \gamma_t, \alpha_0)\right] &= \mathbb{E}[m(W, \gamma_t)] - \theta_0 + \mathbb{E}[\alpha_0(X)(Y - \gamma_t(X))] \\
&= R(\gamma_t) - \theta_0 + \mathbb{E}[\alpha_0(X)Y] - R(\gamma_t) \\
&= \mathbb{E}[\alpha_0(X)Y] - \theta_0 \ ,
\end{aligned}$$

[7] It is termed as the first-step influence function in Chernozhukov et al. (2022a).

which is constant in t. Therefore,

$$\frac{d}{dt}\,\mathbb{E}\big[\psi(W,\theta_0,\gamma_t,\alpha_0)\big]\Big|_{t=0+} = 0, \qquad \forall P \in \mathbf{P}\ .$$

■

Note that the notion of Neyman orthogonality stated in (C.1) is stronger than the one used in Chernozhukov et al. (2022a). In Chernozhukov et al. (2022a), the orthogonality condition is only required locally along directions in the tangent set (i.e. the closure in $L^2(P_{0,X})$ of all valid perturbation directions of γ induced by convex mixture paths of the underlying distribution). In contrast, the result above shows that orthogonality holds globally, for all directions of the form $\{\gamma_P - \gamma_0 : P \in \mathbf{P}\}$. This stronger conclusion follows from the special structure implied by Assumption 2.1, which ensures that the expectation of the augmented moment function does not depend on γ. As a consequence, the derivative with respect to γ vanishes everywhere, not just tangentially.

Recall Assumption 2.1 was imposed to establish the existence of the Riesz representer. This helps to relate the L^2-convergence rate of $\gamma \mapsto \mathbb{E}[m(W,\gamma)]$ to those of the first-step estimators. By the Cauchy–Schwarz inequality, we have

$$\|\mathbb{E}[m(W,\gamma-\gamma_0)]\| \lesssim \|\gamma-\gamma_0\|_{L^2}\ .$$

This, however, is not enough to establish $\sqrt{n}$-consistency of an estimator that is based directly on the original moment function *when the first-step estimators converge slower than the parametric $n^{1/2}$ rate in L^2-norm.* This motivates the use of the augmented moment function, which has the nice Neyman orthogonality property.

Neyman orthogonality plays an important role in establishing the *second order* local smoothness property of the expected augmented moment function. Specifically, it ensures that the moment function is first order insensitive to errors in the nuisance function, yielding

$$\|\mathbb{E}[\psi(W,\theta_0,\gamma,\alpha_0)-\psi(W,\theta_0,\gamma_0,\alpha_0)]\| \lesssim \|\gamma-\gamma_0\|_{L^2}^2\ .$$

In a general GMM framework, it is an immediate result from a Taylor expansion, provided that the map from γ to the expected augmented moment function satisfies: (i) twice continuously Frechet differentiability in a neighborhood of γ_0; and (ii) the Neyman orthogonality property (see Chernozhukov et al. (2022a) Theorem 3).

In our setting, the augmented moment function has a special linear structure. By

Assumption 2.1, we have for $\mathbb{E}[\alpha_0(X)Y] < \infty$,

$$\mathbb{E}[\psi(W, \theta_0, \gamma, \alpha_0)] = \mathbb{E}[\alpha_0(X)Y] - \theta_0 \qquad \forall \gamma \in L^2(P_{0,X}) \ ,$$

which is constant in γ. Therefore,

$$\|\mathbb{E}[\psi(W, \theta_0, \gamma, \alpha_0) - \psi(W, \theta_0, \gamma_0, \alpha_0)]\| = 0 \qquad \forall \gamma \in L^2(P_{0,X}) \ ,$$

and the second-order local smoothness condition is satisfied automatically.

Appendix D Additional Simulation Results

This appendix collects additional simulation results.

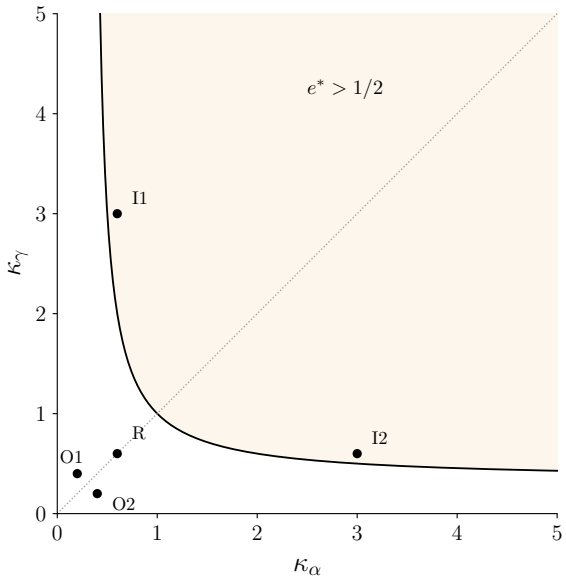


Figure 5: Additional Simulation Designs

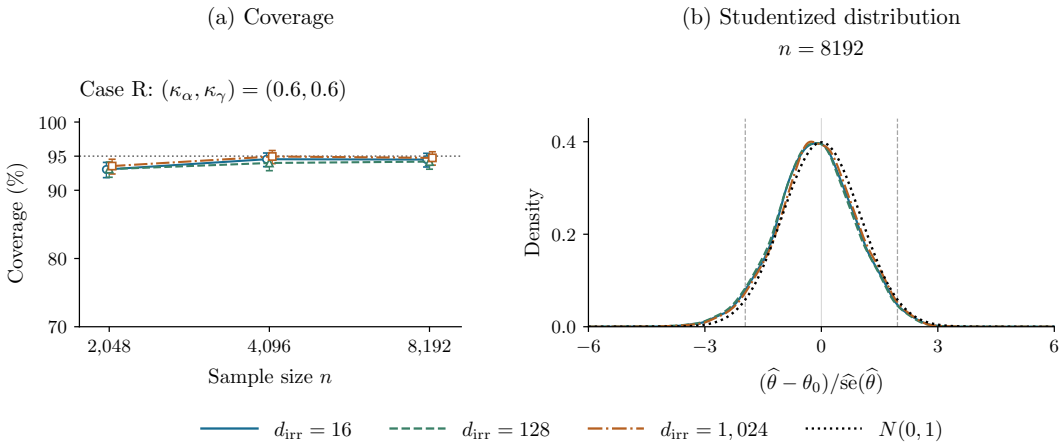


Figure 6: Simulation Designs Relative to the Frontier (Reference Case)

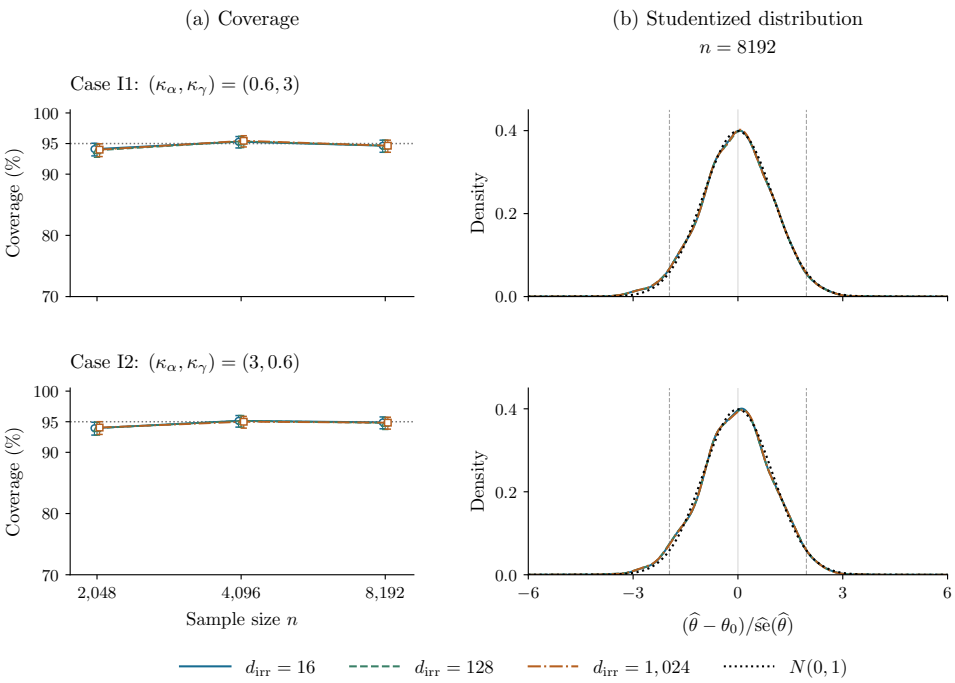


Figure 7: Coverage of Nominal 95% Confidence Interval (Inside Cases)

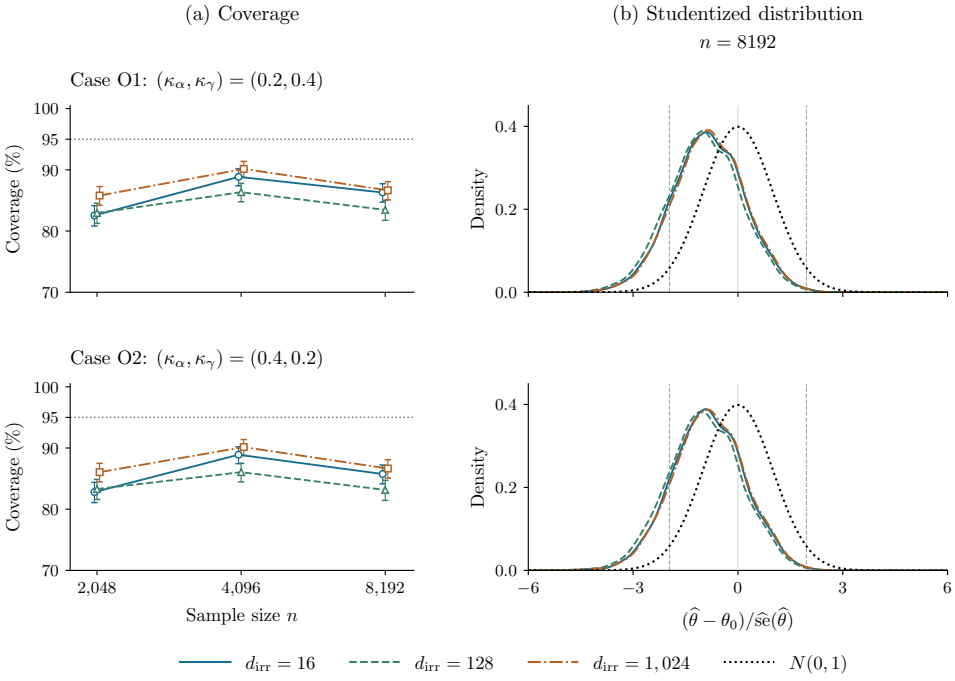


Figure 8: Coverage of Nominal 95% Confidence Interval (Outside Cases)