

Bias-Aware Inference in Regularized Regression Models*

Timothy B. Armstrong[†]

University of Southern California

Michal Kolesár[‡]

Princeton University

Soonwoo Kwon[§]

Brown University

October 21, 2026

Abstract

We consider inference on a scalar regression coefficient under a constraint on the magnitude of the control coefficients. A class of estimators based on a regularized propensity score regression is shown to exactly solve a tradeoff between worst-case bias and variance. We derive confidence intervals (CIs) based on these estimators that are bias-aware: they account for the possible bias of the estimator. Under homoskedastic Gaussian errors, these estimators and CIs are near-optimal in finite samples for mean squared error and CI length. We also provide conditions for asymptotic validity of the CIs with an unknown and possibly heteroskedastic error distribution, and derive novel optimal rates of convergence under high-dimensional asymptotics that allow the number of regressors to increase more quickly than the number of observations. Extensive simulations and two empirical applications illustrate the performance of our methods.

*An earlier version of this paper was circulated under the title “Optimal Inference in Regularized Regression Models”. Parts of this paper incorporate material from Section 4 of the working paper Armstrong and Kolesár (2016), which was taken out in the published version (Armstrong and Kolesár 2018). We thank Victor Chernozhukov for helpful comments and discussion. We thank Mark Li and Ulrich Müller for sharing their code. Armstrong acknowledges support from National Science Foundation Grant SES-2049765. Kolesár acknowledges support by the Sloan Research Fellowship and from National Science Foundation Grant SES-2049356. Kwon acknowledges support from the Annenberg Institute. Juan Yamin Silva provided excellent research assistance.

[†]email: timothy.armstrong@usc.edu

[‡]email: mkolesar@princeton.edu

[§]email: soonwoo_kwon@brown.edu

1 Introduction

We study estimation and inference for a scalar coefficient β in a linear regression model

$$Y_i = w_i\beta + z_i'\gamma + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

when the dimension k of the controls z_i is large. In such settings, the ordinary least squares (OLS) estimator is often uninformative, exhibiting a large variance; the estimator is not even defined when $k > n$. This motivates modifying the OLS objective function to penalize large values of γ , thereby lowering variance at the cost of introducing bias.

The most popular of these approaches are the lasso (Tibshirani 1996) and other variants of ℓ_1 penalization (e.g., Candès and Tao 2007; Belloni, Chernozhukov, and Wang 2011). There is a large literature (see, e.g., Bühlmann and van de Geer 2011, for a review) showing favorable mean squared error (MSE) properties of these estimators when γ is (approximately) sparse. For inference, several papers have developed CIs based on “double lasso” estimators (see, among others, Belloni, Chernozhukov, and Hansen 2014; Javanmard and Montanari 2014; van de Geer et al. 2014; C.-H. Zhang and S. S. Zhang 2014). These methods come with large-sample coverage guarantees even when $k \gg n$, provided that the sparsity level of γ grows slowly enough relative to n . However, because the theory only restricts the rate of growth of the sparsity level, it is unclear what exact degree of sparsity this approach implicitly requires to guarantee good performance in a given sample. Furthermore, in many applications in economics, there are no compelling substantive arguments for why only a small number of the controls should be able to account for the bulk of the confounding.

We propose bounding the *magnitude* of the control coefficients, rather than their sparsity level, by assuming that $\text{Pen}(\gamma) \leq C$. The penalty function $\text{Pen}(\cdot)$ formalizes the notion of magnitude, and it can incorporate any restrictions on γ that place it in a convex symmetric set. In many applications, high-dimensional controls arise in robustness checks, where researchers examine the sensitivity of the baseline results to the inclusion of additional controls, such as squares, interactions, or other transformations of the initial controls, or additional covariates that are believed to be less important.¹ In such cases, one can take the penalty to be an ℓ_p norm of the additional controls. We focus on two specifications in particular. The first is an ℓ_1 penalty that can be interpreted as bounding the *total effect* of the additional controls (the maximal impact on the outcome of simultaneously shifting each of them by up to one standard deviation). The second is a weighted ℓ_2 penalty that bounds their *explanatory power* (their partial R^2). Following the proposals in the recent sensitivity analysis

¹Prominent examples include Jones, Molitor, and Reif (2019), Carneiro et al. (2021), Dhar, Jain, and Jayachandran (2022), and Boudreau (2024).

literature (Altonji, Elder, and Taber 2005; Oster 2019; Cinelli and Hazlett 2020; Diegert, Masten, and Poirier 2026), the penalty bound C can be calibrated by relating the penalty to the total effect or explanatory power of the baseline controls in the baseline specification.

We obtain sharp finite-sample results deriving near-optimal estimators and CIs under the chosen penalty constraint in the idealized scenario in which the regression errors ε_i are Gaussian with a known homoskedastic variance. The estimator that optimally trades off the finite-sample worst-case bias against variance takes a simple form, which can be understood as a penalized version of the Frisch-Waugh formula. First, one runs a penalized propensity score regression of w_i on z_i using $\text{Pen}(\cdot)$ as the penalty function. Second, one uses the propensity score residuals as an instrument in the univariate regression of Y_i on w_i . CIs based on these estimators can be constructed by using a critical value that incorporates the worst-case bias of the estimator, which we show is obtained automatically as a byproduct of the first-step propensity score regression. Because these CIs are *bias-aware*—they account for the potential finite-sample bias of the estimator—they are valid in finite samples in this idealized Gaussian setup. We show how to tune the penalty parameter in the propensity score regression to optimize the MSE of the resulting estimator, or the length of the resulting CI.

In the more realistic scenario with heteroskedasticity and an unknown error distribution, bias-aware CIs can be formed using heteroskedasticity-robust variance estimators. We give conditions for their asymptotic validity, allowing for high-dimensional asymptotics with $k \gg n$. We also study rates of convergence and the efficiency of the bias-aware CIs relative to the unpenalized regression. We show that when $k < n$, meaningfully improving upon the unpenalized regression by bounding the partial R^2 of the controls is only possible when the sample size is small to moderate. In the high-dimensional setting with $k \gg n$, consistency requires shrinking the bound on the partial R^2 to zero, which may be hard to motivate. By contrast, the ℓ_1 penalty bound that limits the total effect of the controls achieves fast convergence rates even when C does not shrink with n . Furthermore, we show that this convergence rate cannot be improved even if one additionally imposes the same ℓ_1 bound on the coefficients in the propensity score regression, as well as a certain degree of sparsity in both regressions.

Our approach requires an explicit specification of the penalty bound C . Our efficiency bounds show that it is impossible to automate the choice of C when forming CIs; fundamentally, one cannot obtain a sufficiently informative data-driven upper bound for C . We show, however, that it is possible to obtain a one-sided *lower* CI for C , which can be used as a specification check to ensure that the chosen value is not too low. We also discuss how a careful calibration of C can be supplemented by computing its breakdown value, the largest

value of C such that the empirical finding of interest, such as rejecting a particular null hypothesis, holds.

The requirement to explicitly choose C may seem like a limitation of our approach relative to sparsity-based approaches, which do not require explicitly bounding the degree of sparsity. However, good finite-sample performance of such methods *does* require bounding the degree of sparsity implicitly, and such implicit bounds are hard to calculate in a given problem. We demonstrate this issue in a Monte Carlo analysis, where we show that double-lasso CIs display moderate to severe undercoverage even in designs that are apparently sparse. Indeed, we view the explicit specification of C as an advantage of our approach, because our coverage guarantees and efficiency bounds are based on explicit assumptions rather than “asymptotic promises” that are hard to evaluate in a particular sample.

Two empirical applications illustrate the theory. The first application investigates the robustness of the analysis in Imbens, Rubin, and Sacerdote (2001) to treatment effect heterogeneity. The sample size is moderate, and in line with the theory, we find that bounding the partial R^2 of the additional controls yields nonnegligible efficiency gains over the unpenalized regression. Assuming an ℓ_1 penalty bound yields tighter inference. The point estimates remain stable over a range of specifications, indicating that the original analysis is not sensitive to dropping the assumption of constant treatment effect. The second application examines the sensitivity to linear functional form assumptions of the conclusions in Connors et al. (1996) and Hirano and Imbens (2001) that right heart catheterization (RHC) increases mortality. Due to the large sample size, the bound on partial R^2 has little bite; the ℓ_1 penalty bound suggests that the effects of RHC are likely smaller in magnitude than in the baseline regression, though they remain statistically significant.

Our results relate to several strands of literature. Our procedures and efficiency bounds apply the general theory of estimation and inference on linear functionals in convex Gaussian models developed in Ibragimov and Khas’minskii (1985), Donoho (1994), Low (1995), and Armstrong and Kolesár (2018), and add to a growing literature applying this approach to various settings (Armstrong and Kolesár 2021a; Armstrong and Kolesár 2021b; Kolesár and Rothe 2018; Imbens and Wager 2019; Rambachan and Roth 2023; Noack and Rothe 2024; K. Kwon and S. Kwon 2020). Muralidharan, Romero, and Wüthrich (2025) apply the approach in the present paper to experiments with factorial designs and bounds on interaction effects.

The idea of using propensity score residuals to estimate β goes back at least to the work of Robinson (1988) on the partially linear model. We provide a novel finite-sample justification for this idea, as well as an exact result giving the optimal penalization in this regression. Our setup allows for a general form of $\text{Pen}(\cdot)$, and yields existing estimators in a few special cases, including the linear estimator of Heckman (1988) in the partially linear

model, and the ridge regression estimator considered in K.-C. Li (1982) under an unweighted ℓ_2 penalty. Finally, C. M. Li and Müller (2021) consider the weighted ℓ_2 penalty that bounds the explanatory power of the controls, and develop a bias-aware CI based on a likelihood ratio statistic, which is numerically shown to be close to optimal under homoskedasticity and a particular weighted average length criterion. However, unlike our CI, the Li–Müller CI may end up being longer than the unpenalized regression CI, as one of our empirical applications in Section 7 illustrates.

The next section presents our finite-sample results in the idealized model with Gaussian errors. Section 3 discusses implementation in the more realistic setting with an unknown error distribution. Section 4 gives efficiency results and derives rates of convergence. Section 5 compares our approach to CIs motivated by sparsity constraints. The performance of our methods is evaluated in a Monte Carlo study in Section 6, while Section 7 illustrates them in two empirical applications. Proofs, extensions, and auxiliary results appear in appendices.

2 Method

To motivate our method, we set up an idealized version of our model with Gaussian homoskedastic errors. We then show how to construct estimators and CIs in this model that optimize the finite-sample tradeoff between bias and variance.

2.1 Setup

We write the model in eq. (1) in vector form as

$$Y = w\beta + Z\gamma + \varepsilon, \quad (2)$$

where $w = (w_1, \dots, w_n)' \in \mathbb{R}^n$ stacks the variable of interest, and $Z = (z_1, \dots, z_n)' \in \mathbb{R}^{n \times k}$ is a matrix of control variables. The design matrix $X = (w, Z)$ is treated as fixed. To obtain finite-sample results, we further assume that the errors are normal and homoskedastic, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, with σ^2 known. To ensure informative inference on β when the number of controls k is large relative to the sample size n (and possibly exceeds it), we assume the control coefficients γ are “small” in the sense that their parameter space is given by

$$\Gamma = \Gamma(\text{Pen}; C) = \{\gamma \in \mathcal{G}: \text{Pen}(\gamma) \leq C\} \quad (3)$$

for some seminorm $\text{Pen}(\cdot)$ on a linear subspace $\mathcal{G}$ of $\mathbb{R}^k$, and some penalty bound C . We do not restrict β , so that the model parameters satisfy $(\beta, \gamma)' \in \mathbb{R} \times \Gamma$. The requirement

that $\text{Pen}(\cdot)$ be a seminorm means that it satisfies the triangle inequality ($\text{Pen}(\gamma + \tilde{\gamma}) \leq \text{Pen}(\gamma) + \text{Pen}(\tilde{\gamma})$), and homogeneity ($\text{Pen}(c\gamma) = |c| \text{Pen}(\gamma)$ for any scalar c), but, unlike a norm, it is not necessarily positive definite ($\text{Pen}(\gamma) = 0$ does not imply $\gamma = 0$). This allows us to restrict only a subset of the control coefficients.

For the choice of $\text{Pen}(\gamma)$, we focus on two leading examples, both of which partition the controls into a set of $k_1 \geq 0$ unrestricted baseline controls and a set of $k_2 = k - k_1$ additional controls, $Z = (Z_1, Z_2)$. We partition $\gamma = (\gamma'_1, \gamma'_2)'$ accordingly. We denote the ℓ_p norm of a vector a by $\|a\|_p$, and the projection onto the column space of a matrix A by H_A .

Example 2.1 (ℓ_2 penalty). Let M be a $k_2 \times k_2$ matrix that scales the additional controls, and set $\text{Pen}(\gamma) = \|M\gamma_2\|_2$. Setting $M = I_{k_2}$ weights the coefficients equally, while setting $M = (Z'_2(I - H_{Z_1})Z_2/n)^{1/2}$ corresponds to the specification considered in C. M. Li and Müller (2021). The latter specification has an intuitive interpretation: the population partial R^2 from adding Z_2 to the regression of $Y - w\beta$ onto Z_1 is given by $R^2_{Y-w\beta} = \|M\gamma_2\|_2^2 / (\|M\gamma_2\|_2^2 + \sigma^2)$. If w is a treatment dummy, imposing eq. (3) thus amounts to bounding the partial R^2 from the regression of the untreated outcome on the additional controls by $C^2/(C^2 + \sigma^2)$. This specification for the penalty is therefore natural if the researcher wishes to limit the *explanatory power* of the additional controls. We refer to it as the $R^2_{Y-w\beta}$ penalty in the sequel. $\square$

Example 2.2 (ℓ_1 penalty). We focus on the unweighted ℓ_1 penalty, and after standardizing the controls to have zero mean and unit variance, set $\text{Pen}(\gamma) = \|\gamma_2\|_1$. This specification limits the *total effect* of the controls: $\|\gamma_2\|_1 = \sum_{j=k_1+1}^k |\gamma_j|$ is the maximal impact on the outcome by simultaneously shifting each control by up to one standard deviation. $\square$

Specification of the parameter space Γ requires selecting the penalty bound C , in addition to picking the penalty; here we take C as given, and defer a discussion of its choice to Section 3.

Formulating the parameter space Γ in terms of a seminorm is not restrictive in the sense that essentially any convex set Γ that is symmetric ($\gamma \in \Gamma$ implies $-\gamma \in \Gamma$) can be defined in this way (see Yosida 1995, Proposition 5, p. 26). Furthermore, the assumption that the controls enter linearly in eq. (2) is not essential: as we show in Appendix A.1, our setup also covers the partially linear model that only places smoothness restrictions on how the controls enter the outcome equation. Although we rule out nonconvex constraints on Γ , such as sparsity, our results nonetheless have implications for such settings, as we discuss in Section 5.

2.2 Linear estimators and CIs

We now show how to build finite-sample valid CIs based on linear estimators of the form $\hat{\beta} = a'Y$ for some n -vector of weights a that may depend on the design matrix X and the known variance σ^2 . We then show how to choose the weights a optimally. Section 4.1 below shows that when a is optimally chosen, the resulting estimators and CIs are optimal or near-optimal among all procedures, not just linear ones.

Since the estimator $\hat{\beta}$ is linear, under Gaussian errors, it is normally distributed with mean $a'(w\beta + Z\gamma)$ and known standard deviation $\text{sd}(\hat{\beta}) = \sigma\|a\|_2$. Consequently, the z -statistic $(\hat{\beta} - \beta)/\text{sd}(\hat{\beta})$ follows a $\mathcal{N}(b, 1)$ distribution, where $b = (a'(w\beta + Z\gamma) - \beta)/\text{sd}(\hat{\beta})$ denotes the bias-standard deviation ratio. While we cannot directly compute b because the bias depends on the unknown parameters β and γ , under the parameter space restriction in eq. (3), we know $|b|$ is bounded by $\overline{\text{bias}}_\Gamma(\hat{\beta})/\text{sd}(\hat{\beta})$, where $\overline{\text{bias}}_\Gamma(\hat{\beta})$ denotes the *worst-case bias* of $\hat{\beta}$ over the parameter space. To ensure that the worst-case bias is finite, it is necessary that $a'w = 1$; under this condition,

$$\overline{\text{bias}}_\Gamma(\hat{\beta}) = \sup_{\gamma \in \Gamma} a'Z\gamma. \quad (4)$$

Thus, we can form a two-sided CI for β by replacing the usual critical value based on quantiles of the standard normal distribution with a critical value that accounts for the non-zero mean of the z -statistic:

$$\hat{\beta} \pm \text{cv}_\alpha \left(\overline{\text{bias}}_\Gamma(\hat{\beta})/\text{sd}(\hat{\beta}) \right) \cdot \text{sd}(\hat{\beta}). \quad (5)$$

Here $\text{cv}_\alpha(B)$ denotes the $1 - \alpha$ quantile of the folded normal distribution, $|\mathcal{N}(B, 1)|$.² This CI is *bias-aware* in that the critical value $\text{cv}_\alpha(\cdot)$ reflects the potential finite-sample bias of $\hat{\beta}$.

In the terminology of Donoho (1994), the CI in eq. (5) is a *fixed-length* CI, since its length depends only on the non-random weights a and the known variance σ^2 , but not on the outcome data Y or the parameter vector $(\beta, \gamma)'$. This is useful, as it allows us to pick the weights a that minimize the length of the resulting CI without affecting its coverage properties. To do this, observe that the CI length depends on the n -vector of weights a only through the standard deviation and the worst-case bias, and it is increasing in both $\text{sd}(\hat{\beta})$ and $\overline{\text{bias}}_\Gamma(\hat{\beta})$. As a result, we can find the optimal weights by first minimizing variance subject to a bound B on the worst-case bias,

$$\min_{a \in \mathbb{R}^n} a'a \quad \text{s.t.} \quad \sup_{\gamma \in \Gamma} a'Z\gamma \leq B, \quad a'w = 1, \quad (6)$$

²The critical value $\text{cv}_\alpha(B)$ can be computed as the square root of the $1 - \alpha$ quantile of a non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter B^2 .

and then varying the bound B to find the bias-variance tradeoff yielding the shortest CI.

While solving eq. (6) for a range of bounds B may appear cumbersome, applying the general theory of Ibragimov and Khas'minskii (1985), Donoho (1994), Low (1995), and Armstrong and Kolesár (2018) to our setting allows us to rewrite it as a tractable convex optimization problem. Solving that problem yields our main computational result:

Theorem 2.1. *Let π_λ be a penalized propensity score regression estimator that solves*

$$\min_{\pi} \|w - Z\pi\|_2^2 \quad \text{s.t.} \quad \text{Pen}(\pi) \leq t, \quad (7)$$

for some strictly positive bound $t = t(\lambda)$. Suppose that the propensity score residual $\tilde{w}_\lambda = w - Z\pi_\lambda$ is not identically zero, and that $\text{Pen}(\pi_\lambda) > 0$. Then the worst-case bias and standard deviation of the estimator

$$\hat{\beta}_\lambda = a'_\lambda Y, \quad a_\lambda = \frac{\tilde{w}_\lambda}{\tilde{w}'_\lambda w}, \quad (8)$$

are given by

$$\overline{\text{bias}}_\Gamma(\hat{\beta}_\lambda) = C\overline{B}_\lambda, \quad \text{and} \quad \text{sd}_\lambda = \sigma \|a_\lambda\|_2, \quad \text{where} \quad \overline{B}_\lambda = \frac{a'_\lambda Z\pi_\lambda}{\text{Pen}(\pi_\lambda)}. \quad (9)$$

Furthermore, the weights a_λ solve eq. (6) with the bound $B = C\overline{B}_\lambda$.

According to Theorem 2.1, to find the optimal weights, we first run a penalized propensity score regression of w on Z (we use the term propensity score regression even though w may not be binary). It does not matter precisely how the tuning parameter λ maps to the penalty bound t . We may take λ to correspond to the Lagrange multiplier in a Lagrangian formulation of eq. (7), but we can also take $\lambda = t$ or other parametrizations depending on what is most convenient. We then use the residuals $\tilde{w}_\lambda$ from the penalized propensity score regression as instruments in the univariate regression of Y on w . The resulting estimator $\hat{\beta}_\lambda$ solves the bias-variance tradeoff in eq. (6) for a particular value of B that depends on the penalty weight λ .

Since eq. (9) gives explicit formulas for the worst-case bias and standard deviation of $\hat{\beta}_\lambda$, finding the penalty weight λ_{CI}^* that optimizes the length of the resulting CI is a simple univariate optimization problem. We can use the same approach to find the best estimator. This is because the worst-case MSE of any linear estimator is likewise increasing in its worst-case bias and its standard deviation. Specifically, the penalty weights λ_{MSE}^* and λ_{CI}^* that optimize the MSE and CI length, respectively, solve

$$\lambda_{\text{MSE}}^* = \arg \min_{\lambda} \text{sd}_\lambda^2 + (C\overline{B}_\lambda)^2, \quad \lambda_{\text{CI}}^* = \arg \min_{\lambda} \text{cv}_\alpha(C\overline{B}_\lambda / \text{sd}_\lambda) \text{sd}_\lambda, \quad (10)$$

with sd_λ and $\bar{B}_\lambda$ given in eq. (9), and the optimization is over penalty weights such that $\|w - Z\pi_\lambda\|_2 > 0$. The optimal linear estimator and the optimal linear CI are then given by

$$\hat{\beta}_{\lambda_{\text{MSE}}}^* \quad \text{and} \quad \hat{\beta}_{\lambda_{\text{CI}}}^* \pm \text{cv}_\alpha(C\bar{B}_{\lambda_{\text{CI}}}^*/\text{sd}_{\lambda_{\text{CI}}}^*)\text{sd}_{\lambda_{\text{CI}}}^*. \quad (11)$$

Using the *penalized* propensity score residuals as instruments for w to calculate $\hat{\beta}_\lambda$ can be understood as a penalized version of the Frisch-Waugh formula. Indeed, if we vary the penalty weight such that $t \rightarrow \infty$, $\hat{\beta}_\lambda$ converges to the long regression estimate $\hat{\beta}_{\text{long}} = \frac{w'(I-H_Z)Y}{w'(I-H_Z)w}$ that uses the unpenalized propensity score residuals, $(I - H_Z)w$, as instruments, provided they are not identically zero (which ensures that the condition $\|\tilde{w}_\lambda\|_2 > 0$ in Theorem 2.1 holds for all λ). This estimator minimizes variance among all linear estimators that are unbiased, so Theorem 2.1 reduces to the Gauss-Markov theorem in this case. If we vary the penalty weight such that $t \rightarrow 0$ and $\text{Pen}(\cdot)$ is a norm on γ_2 , then $\hat{\beta}_\lambda$ converges to the short regression estimate $\hat{\beta}_{\text{short}} = \frac{w'(I-H_{Z_1})Y}{w'(I-H_{Z_1})w}$ that only includes the unrestricted controls Z_1 . This estimator minimizes variance among all linear estimators with finite worst-case bias. In other words, the short and long regressions are corner solutions of the bias-variance tradeoff, in which weight is placed entirely on variance or on bias.

Example 2.1 (ℓ_2 penalty, continued). When $\text{Pen}(\pi) = \|M\pi_2\|_2$ for some matrix M , one can use the Lagrangian formulation for eq. (7) to show that $\hat{\beta}_\lambda$ is equivalent to a ridge regression of Y onto X that minimizes $\|Y - w\beta - Z_1\gamma_1 - Z_2\gamma_2\|_2^2 + \lambda\|M\gamma_2\|_2^2$. The optimality of ridge regression in this setting was shown by K.-C. Li (1982); we obtain this result as a special case of Theorem 2.1.

For the $R_{Y-w\beta}^2$ penalty that sets $M = (Z_2'(I - H_{Z_1})Z_2/n)^{1/2}$, the ridge estimator further simplifies to a weighted average of the long and short regression estimates. For ease of interpretation, we write the penalty bound as $C = \sigma\bar{R}/(1 - \bar{R}^2)^{1/2}$ where $\bar{R}^2$ is the implied bound on $R_{Y-w\beta}^2$, the partial R^2 from adding Z_2 to the regression of the untreated outcome on Z_1 . If the long regression estimator is infeasible (i.e., the unpenalized propensity score residual $(I - H_Z)w$ is identically zero), the short regression estimator is optimal. Otherwise, to find the optimal bias-variance tradeoff, it is convenient to set λ to correspond directly to the weight on the short regression,

$$\hat{\beta}_\lambda = \lambda\hat{\beta}_{\text{short}} + (1 - \lambda)\hat{\beta}_{\text{long}}. \quad (12)$$

Theorem 2.1 then implies this estimator has worst-case bias given by

$$C\bar{B}_\lambda = \lambda \cdot \sigma_{\text{short}} \sqrt{\frac{n\bar{R}^2 R_w^2}{1 - \bar{R}^2}}, \quad \text{and} \quad \text{sd}_\lambda = \sqrt{1 + (1 - \lambda)^2 R_w^2 / (1 - R_w^2)} \sigma_{\text{short}}, \quad (13)$$

where $\sigma_{\text{short}} = \sigma / \sqrt{w'(I - H_{Z_1})w}$ denotes the short regression standard deviation, and $R_w^2 = 1 - \frac{w'(I - H_Z)w}{w'(I - H_{Z_1})w}$ is the sample partial R^2 from adding Z_2 to the propensity score regression of w onto Z_1 . Plugging these expressions into eq. (10) then allows one to compute the optimal weight on the short regression. $\square$

Example 2.2 (ℓ_1 penalty, continued). In this case, the solution to (7) is given by a variant of the lasso estimate (Tibshirani 1996) that only penalizes π_2 .

The resulting estimator $\hat{\beta}_\lambda$ is related to estimators proposed for constructing CIs using the lasso (see, among others, C.-H. Zhang and S. S. Zhang 2014; Javanmard and Montanari 2014; van de Geer et al. 2014; Belloni, Chernozhukov, and Hansen 2014). These papers propose estimators for β that combine lasso estimates from the outcome regression of Y onto X with lasso estimates from the propensity score regression, which yields an estimate that is nonlinear in Y . In contrast, our estimator only uses lasso estimates for the propensity score regression, and is linear in Y . We give a detailed comparison between our estimator and this “double lasso” approach in Section 5. $\square$

Remark 2.1 (Heterogeneous treatment effects). If w_i is as good as randomly assigned conditional on z_i , the coefficient β in eq. (1) can be interpreted as the average treatment effect (ATE) of a one-unit increase in the variable w . This interpretation requires that the individual treatment effects (TEs) are mean independent of z_i . To allow for TE heterogeneity in our framework, the simplest approach is to consider the model with interactions, $Y_i = w_i z_i' \delta + z_i' \gamma + \varepsilon_i$, where $z_i' \delta$ now represents the conditional ATE for units with $z_i = z$. Suppose the parameter of interest is the weighted average of the conditional ATEs, $\beta = \sum_{i=1}^n c_i z_i' \delta$, where the weights c_i sum to one. With binary w_i , the unweighted case $c_i = 1/n$ gives the (conditional on the sample) ATE, while setting $c_i = w_i / \sum_{j=1}^n w_j$ gives the ATE for the treated. Then, as noted by Imbens and Wooldridge (2009, Section 5.3), we can rewrite the interacted model as

$$Y_i = w_i \beta + w_i (z_{i,-1} - \sum_{j=1}^n c_j z_{j,-1})' \delta_{-1} + z_i' \gamma + \varepsilon_i,$$

where we assume that the first element of z_i is constant, so that $z_{i,-1}$ is the vector of controls excluding a constant and δ_{-1} is the corresponding subvector of δ . This is exactly in the form of eq. (1), with the control vector consisting of *both* the original controls z_i and the interaction of the treatment with the demeaned controls, $w_i (z_{i,-1} - \sum_{j=1}^n c_j z_{j,-1})$. We use this approach in our first empirical application in Section 7. Theorem A.1 in Appendix A.2 shows that when TE are heterogeneous, and one uses our approach without including the interactions, the estimator $\hat{\beta}_\lambda$ provides inference on a weighted average of treatment effects; however,

some of the weights may be negative. Appendix A.2 also discusses how our approach can be adapted to allow the conditional ATEs to be nonlinear in z_i .

3 Implementation with non-Gaussian and heteroskedastic errors

We now discuss practical implementation issues, allowing ε to be non-Gaussian and heteroskedastic. As a baseline, we propose the following implementation:

Algorithm 3.1 (Baseline implementation).

Input Data (Y, X) , penalty $\text{Pen}(\cdot)$, penalty bound C , and initial estimates of residuals

$$\hat{\varepsilon}_{\text{init},1}, \dots, \hat{\varepsilon}_{\text{init},n}.$$

Output Estimator and CI for β .

1. Compute an initial variance estimator, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_{\text{init},i}^2$, assuming homoskedasticity.
2. For each λ , use eq. (8) to compute $\hat{\beta}_\lambda$, and use eq. (9) to compute $\bar{B}_\lambda$ and sd_λ , with $\hat{\sigma}$ in place of σ in the formula for sd_λ . For the $R_{Y-w\beta}^2$ penalty, use the closed-form expressions in eq. (12) and eq. (13).
3. Compute λ_{MSE}^* and λ_{CI}^* as in eq. (10), assuming homoskedasticity. Compute the robust standard error $\widehat{\text{se}}_{\lambda_{\text{CI}}^*, \text{rob}}^2 = \sum_{i=1}^n a_{\lambda_{\text{CI}}^*, i}^2 \hat{\varepsilon}_{\text{init},i}^2$.

Return the estimator $\hat{\beta}_{\lambda_{\text{MSE}}^*}$ and the CI $\hat{\beta}_{\lambda_{\text{CI}}^*} \pm \text{cv}_\alpha(C\bar{B}_{\lambda_{\text{CI}}^*}/\widehat{\text{se}}_{\lambda_{\text{CI}}^*, \text{rob}})\widehat{\text{se}}_{\lambda_{\text{CI}}^*, \text{rob}}$. $\square$

The baseline implementation assumes homoskedasticity for the purpose of computing the weights a_λ , but the CI is computed using robust standard errors to ensure asymptotic validity under heteroskedastic errors. Appendix C.2 gives formal asymptotic validity results. The key regularity condition is that the maximal Lindeberg weight $\text{Lind}(a_{\lambda_{\text{CI}}^*})$, where

$$\text{Lind}(a) = \frac{\max_{1 \leq i \leq n} a_i^2}{\sum_{j=1}^n a_j^2}, \quad (14)$$

vanishes fast enough relative to the estimation error in $\widehat{\text{se}}_{\lambda_{\text{CI}}^*, \text{rob}}$ (as measured by the ℓ_2 convergence rate of $\hat{\varepsilon}_{\text{init}}$). Appendix C.2 also discusses the Lindeberg weight diagnostics in our empirical applications (specifically, see Remark C.1), as well as a modification of the baseline algorithm that ensures this regularity condition holds. When $C = \infty$ or $C = 0$, the baseline implementation reduces to the ubiquitous practice of reporting (long or short

regression) OLS with heteroskedasticity-robust standard errors. While it is possible to exploit heteroskedasticity for computing the optimal weights a_λ , we do not pursue the generalization here.

The following remarks discuss the implementation choices in more detail.

Remark 3.1 (Choice of initial residuals). Following the analysis in Appendix C.2, we can take residuals from a regularized outcome regression of Y on X as the initial residual estimates $\hat{\varepsilon}_{\text{init},i}$. In our simulations and empirical applications, we use the lasso, with penalty chosen via 10-fold cross-validation.

Remark 3.2 (Choice of C). To help calibrate the penalty bound C , we follow the recent sensitivity analysis literature (Altonji, Elder, and Taber 2005; Oster 2019; Cinelli and Hazlett 2020; Diegert, Masten, and Poirier 2026) that advocates for calibrating the importance of unobservables by relating them to observable analogues. We focus on the setting in Examples 2.1 and 2.2 in which only a subset of the controls is penalized; the penalized controls Z_2 then play a role analogous to the unobserved confounders in the sensitivity analysis literature. For the ℓ_1 penalty, we calibrate C by setting it to $\rho \|\hat{\gamma}_{\text{short}}\|_1$ for some ρ , where $\hat{\gamma}_{\text{short}}$ denotes the OLS coefficient on the baseline controls Z_1 (with the intercept excluded).³ This bounds the total effect of the penalized controls on the outcome in the long regression to be at most ρ times the total effect of Z_1 in the short regression. As a conservative baseline specification capturing the belief that the baseline controls are more important, we propose setting $\rho = 1$. This is analogous to the common practice in the sensitivity analysis literature to assume the “effect” of unobservables is no larger than the “effect” of observables (e.g. Altonji, Elder, and Taber 2005; Oster 2019); here we formalize “effect” by the ℓ_1 norms in the short and long regression.⁴

For the $R_{Y-w\beta}^2$ penalty (or other versions of the ℓ_2 bound), our theoretical analysis in Section 4 implies that an analogous calibration will typically not yield substantive improvements over the long regression if the long regression is feasible. Specifically, consider implementing the $R_{Y-w\beta}^2$ penalty by bounding $R_{Y-w\beta}^2 \leq \bar{R}^2$ for some bound $\bar{R}^2$. We show in Section 4.2 that ensuring gains over the long regression requires setting $\bar{R}^2 \lesssim 1/n$, which has two implications. First, the implied values of $\bar{R}^2$ necessary for substantive improvements over the

³Formally, one should account for sampling uncertainty in $\hat{\gamma}_{\text{short}}$ to ensure validity of the CI under the assumption $\|\gamma_2\|_1 \leq \rho \|\gamma_{\text{short}}\|_1$, where γ_{short} denotes the population short regression coefficient. For instance, one can combine a first-stage CI for $\|\gamma_{\text{short}}\|_1$ with a Bonferroni correction. In our simulations in Section 6, however, we find that our calibration leads to valid coverage when this assumption holds even without additional corrections for sampling uncertainty.

⁴If one wishes to penalize all controls, one can select the ℓ_1 penalty bound by selecting a subset of controls Z_1 that one deems most important, and set $C = \rho \|\hat{\gamma}_{\text{short}}\|_1$ for some ρ . A natural default calibration is $\rho = 2$, so that the total effect of *all* controls in the long regression is at most twice as large as the effect of the important controls in the short regression.

long regression are unreasonably small once the sample size is in the thousands, as is the case in our second empirical application in Section 7.2; we therefore do not recommend using $R_{Y-w\beta}^2$ bounds or other versions of the ℓ_2 penalty in such cases. Second, setting $\bar{R}^2$ to, say, the short regression analogue will have little bite even when the sample size is only in the hundreds. Nonetheless, in some applications with moderate sample sizes, it may be possible to use direct arguments for setting $\bar{R}^2$ to smaller values, allowing for nonnegligible gains over the long regression. For example, if the additional controls are transformations of the baseline controls as in functional form robustness checks, one may set $\bar{R}^2$ to a fraction of the short regression R^2 , formalizing the empirical regularity that higher-order terms have less explanatory power than main effects. We illustrate this in our first empirical application in Section 7.1.

To complement the analysis under default choices of ρ or $\bar{R}^2$, we recommend varying them by way of sensitivity analysis, and reporting a “breakdown value”, the largest value such that the empirical finding of interest holds. This is analogous to the ubiquitous practice of reporting breakdown values for sensitivity parameters in the sensitivity analysis section of empirical papers.

Careful *a priori* calibration of the penalty bound is necessary: by Corollary 4.1(iii) in Section 4.1 below, one cannot start with an initial conservative choice for C and hope to substantively improve upon the CI based on this conservative choice by using a data-driven rule for picking C , without sacrificing coverage. However, it is possible to form a lower CI $[\hat{C}, \infty)$ for C as a specification check to assess the plausibility of a given penalty bound. We present such a CI in Appendix C.3 for the case where $\text{Pen}(\gamma)$ takes the form of an ℓ_p constraint.

Remark 3.3 (Computational issues). Step 2 involves computing the solution path of a regularized regression estimator. Efficient algorithms exist for computing these paths under the ℓ_1 penalty and its variants (Efron et al. 2004; Rosset and J. Zhu 2007). Under the $R_{Y-w\beta}^2$ penalty, the estimator $\hat{\beta}_\lambda$ has a closed form, so that our algorithm can again be implemented in a computationally efficient manner. For other types of penalties, the convexity of the optimization problem in eq. (7) can be exploited to yield efficient implementation. We also note that since the solution path π_λ does not depend on C , it only needs to be computed once, even when multiple choices of C are considered in a sensitivity analysis.

4 Theoretical properties

The optimal linear estimator and CI in eq. (11) trade off bias and variance optimally among procedures that are linear in the outcomes Y . Section 4.1 shows that this estimator and

CI are in fact highly efficient among all procedures, not just linear ones. Section 4.2 shows that reasonable calibrations of the penalty $R_{Y-w\beta}^2$ yield slow convergence rates when the regression is high-dimensional, and that achieving meaningful improvements over the long regression requires imposing $R_{Y-w\beta}^2 \leq \bar{R}^2$ for some $\bar{R}^2 \lesssim 1/n$. Section 4.3 derives analogous results for the unweighted ℓ_2 penalty and shows that, by contrast, the rate results under the ℓ_1 penalty are much better. As in Section 2, this section assumes that the errors ε are Gaussian and homoskedastic. We use the notation $R_{\text{CI}}(C) = \text{cv}_\alpha(C\bar{B}_{\lambda_{\text{CI}}^*}/\text{sd}_{\lambda_{\text{CI}}^*})\text{sd}_{\lambda_{\text{CI}}^*}$ to denote the half-length of the optimal linear CI.

4.1 Efficiency among nonlinear procedures

To state the efficiency result for the optimal linear estimator and CI in eq. (11), let $\text{var}_{\beta,\gamma}(\hat{\beta})$ denote the variance of any estimator $\hat{\beta} = \hat{\beta}(X, Y)$, and let

$$R_{\text{MSE}}(\hat{\beta}; \Gamma) = \sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} E_{\beta,\gamma}[(\hat{\beta} - \beta)^2]$$

denote its worst-case MSE performance over the parameter space, where $E_{\beta,\gamma}$ denotes expectation under $(\beta, \gamma)'$. We refer to the interval $[\hat{\beta} - \hat{\chi}, \hat{\beta} + \hat{\chi}]$ with half-length $\hat{\chi} = \hat{\chi}(X, Y)$ as a $1 - \alpha$ CI for β if it satisfies the coverage requirement

$$\inf_{\beta \in \mathbb{R}, \gamma \in \Gamma} P_{\beta,\gamma}(|\hat{\beta} - \beta| \leq \hat{\chi}) \geq 1 - \alpha,$$

where $P_{\beta,\gamma}$ denotes probability under $(\beta, \gamma)'$.

Corollary 4.1. *Let the optimal linear estimator and the optimal linear CI be given in eq. (11). Define $\hat{\beta}_\lambda$, $\bar{B}_\lambda$ and sd_λ as in Theorem 2.1. Let $\tilde{\beta}$ and $\tilde{\beta} \pm \tilde{\chi}$ denote some other, possibly nonlinear, estimator and some other $1 - \alpha$ CI, also possibly nonlinear in Y . Let $\overline{\text{bias}}_\Gamma(\tilde{\beta}) = \sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} |E[\tilde{\beta} - \beta]|$ denote its worst-case bias.*

(i) *For any λ , $\sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} \text{var}_{\beta,\gamma}(\tilde{\beta}) \leq \text{sd}_\lambda^2$ implies $\overline{\text{bias}}_\Gamma(\tilde{\beta}) \geq C\bar{B}_\lambda$, and $\overline{\text{bias}}_\Gamma(\tilde{\beta}) \leq C\bar{B}_\lambda$ implies $\sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} \text{var}_{\beta,\gamma}(\tilde{\beta}) \geq \text{sd}_\lambda^2$.*

(ii) *The worst-case MSE improvement of $\tilde{\beta}$ over $\hat{\beta}_{\lambda_{\text{MSE}}^*}$ is bounded by*

$$\frac{R_{\text{MSE}}(\tilde{\beta})}{R_{\text{MSE}}(\hat{\beta}_{\lambda_{\text{MSE}}^*})} \geq \kappa_{\text{MSE}}^*(X, \sigma, \Gamma) \geq 0.8,$$

where $\kappa_{\text{MSE}}^*(X, \sigma, \Gamma)$ is given in Appendix B.2.

(iii) The ratio of the expected length of the CI $\tilde{\beta} \pm \tilde{\chi}$ to that of the optimal linear CI at $\gamma = 0$ and any β is bounded below by

$$\frac{E_{\beta,0}[\tilde{\chi}]}{R_{\text{CI}}(C)} \geq \kappa_{\text{CI}}^*(X, \sigma, \Gamma, \alpha),$$

where $\kappa_{\text{CI}}^*(X, \sigma, \Gamma, \alpha)$ is given in Appendix B.2 and is at least 0.717 when $\alpha = 0.05$.

Corollary 4.1 follows from the general results in Donoho (1994), Low (1995) and Armstrong and Kolesár (2018) for estimation of linear functionals in Gaussian models with convex parameter spaces. By construction, the estimator $\hat{\beta}_\lambda$ minimizes variance among all linear estimators with a bound $C\bar{B}_\lambda$ on the bias (or equivalently, it minimizes bias among all linear estimators with a bound sd_λ^2 on the variance). Corollary 4.1(i) shows that this optimality property is retained if we enlarge the class of estimators to all estimators, including nonlinear ones. As a result, the optimal linear estimator $\hat{\beta}_{\lambda_{\text{MSE}}^*}$ continues to perform well among all estimators, including nonlinear ones: by Corollary 4.1(ii), its worst-case MSE efficiency is at least 80%. The exact efficiency bound $\kappa_{\text{MSE}}^*(X, \sigma, \Gamma)$ depends on the design matrix, noise level, and particular choice of the parameter space, and can be computed explicitly in particular applications. We have found that typically the efficiency is considerably higher than the 80% lower bound.

Finally, Corollary 4.1(iii) shows that it is not possible to substantively improve upon the optimal linear CI in terms of expected length when $\gamma = 0$, even if we “direct power” at $\gamma = 0$ (potentially at the expense of longer expected length when $\gamma \neq 0$). The construction of the CI may appear conservative: its length depends on the worst-case bias over the parameter space for γ , which, as the proof of Theorem 2.1 shows, is attained when $\text{Pen}(\gamma) = C$. Therefore, one may think that it may be possible to improve the CI’s performance when in fact $\text{Pen}(\gamma) \ll C$. Corollary 4.1(iii) shows that this is not the case, even in the best-case scenario in which $\gamma = 0$. This generalizes the classic result that when the long regression is feasible and we set $C = \infty$, one-sided t -tests based on the long regression are uniformly most powerful. The exact efficiency bound $\kappa_{\text{CI}}^*(X, \sigma, \Gamma, \alpha)$ can be computed explicitly in particular applications, and we have found that it is typically considerably higher than 71.7%.

A consequence of Corollary 4.1(iii) is that it is impossible to form a CI that is adaptive with respect to the penalty bound C . In the present setting, an adaptive CI would have length that automatically reflects the true regularity $\text{Pen}(\gamma)$ while maintaining coverage under a conservative *a priori* bound on $\text{Pen}(\gamma)$. However, according to Corollary 4.1(iii), any CI must have expected width that reflects the conservative *a priori* bound C rather than the true regularity $\text{Pen}(\gamma)$, even when $\text{Pen}(\gamma)$ is much smaller than the conservative *a priori* bound C . This is why careful *a priori* calibration of the penalty bound C is necessary,

as discussed in Remark 3.2.

4.2 Efficiency and rates of convergence under the $R_{Y-w\beta}^2$ penalty

We now give two results that limit the scope of applicability of the $R_{Y-w\beta}^2$ penalty. First, we show that when the long regression is feasible, meaningful improvements are only possible for small sample sizes. Second, we show that in high-dimensional settings, reasonable calibrations of the penalty bound yield slow convergence rates. As in Remark 3.2, for ease of interpretation, rather than bounding C , we bound $R_{Y-w\beta}^2 \leq \bar{R}^2$ for some bound $\bar{R}^2$.

For the first result, assume that R_w^2 , the partial R^2 from adding Z_2 to the propensity score regression, does not equal 1, so that the long regression is feasible. It then follows from eq. (13) and the definition of the optimal linear CI in eq. (11) that its length relative to the long regression CI $\hat{\beta}_{\text{long}} \pm z_{1-\alpha/2}\sigma_{\text{short}}/\sqrt{1-R_w^2}$ is given by

$$\text{RE}(\kappa, R_w^2; \alpha) = \min_{\lambda \in [0,1]} \text{cv}_\alpha \left(\kappa \frac{\lambda \cdot \sqrt{R_w^2(1-R_w^2)}}{\sqrt{R_w^2(1-\lambda)^2 + 1 - R_w^2}} \right) \frac{\sqrt{R_w^2(1-\lambda)^2 + 1 - R_w^2}}{z_{1-\alpha/2}}, \quad (15)$$

where $\kappa^2 = \kappa^2(n, \bar{R}^2) = n\bar{R}^2/(1-\bar{R}^2)$ and $z_{1-\alpha/2}$ is the usual critical value corresponding to the $1-\alpha/2$ quantile of a standard normal distribution. In particular, the relative efficiency depends on the data only through R_w^2 , the partial R^2 from adding Z_2 to the propensity score regression. Note the maximal possible efficiency gain is capped at the ratio of short and long regression standard errors, $\text{RE}(0, R_w^2; \alpha) = \sqrt{1-R_w^2}$. It is achieved when we set $\bar{R}^2 = 0$ (which amounts to assuming that $\gamma_2 = 0$), in which case the short regression is optimal and unbiased.

Before discussing large-sample implications of the formula in eq. (15), we first use it to derive a simple bound on how small κ and hence the sample size needs to be in order to achieve a meaningful improvement over the long regression. In particular, suppose we wish to improve over the long regression by at least 100 $r\%$, so that $\text{RE}(\kappa, R_w^2; \alpha) \leq 1-r$. Using the fact that $\text{cv}_\alpha(B) \geq B + z_{1-\alpha}$ to lower bound the left-hand side of this inequality and solving for κ then yields, for $r > 1 - z_{1-\alpha}/z_{1-\alpha/2}$ ($\approx 16\%$ when $\alpha = 0.05$)

$$\kappa \leq (1-r)z_{1-\alpha/2} \sqrt{\frac{R_w^2}{1-R_w^2}} - \sqrt{z_{1-\alpha}^2 - (1-r)^2 z_{1-\alpha/2}^2}. \quad (16)$$

Consider a favorable scenario in which $R_w^2 = 0.96$, so that the long regression standard errors are 5 times as big as the short regression standard errors. Then, to achieve 20% improvement over the long regression for $\alpha = 0.05$, by eq. (16), we require $\kappa \leq 7.2$, which requires $n \leq 980$

for $\bar{R}^2 = 0.05$ and $n \leq 5,100$ for $\bar{R}^2 = 0.01$. This suggests that once the sample size is in the thousands, reasonable calibrations for $\bar{R}^2$ will not yield appreciable efficiency gains. For example, in our second empirical application in Section 7.2, $R_w^2 \approx 0.71$ and $n = 5,735$, so eq. (16) implies that a necessary condition for a 20% efficiency gain is $\bar{R}^2 \leq 0.0007$.

For modest sample sizes, meaningful efficiency gains are possible if the long regression is appreciably noisier than the short regression. In our first empirical example in Section 7.1, $R_w^2 \approx 0.76$ and $n = 496$, yielding the corresponding necessary condition $\bar{R}^2 \leq 0.011$. Thus, the moderate sample size leaves more scope for gains than in the second application.

More generally, eq. (16) implies that improving upon the long regression requires setting $\bar{R}^2 \lesssim 1/(n(1 - R_w^2))$, or equivalently $C \lesssim 1/\sqrt{n(1 - R_w^2)}$. This yields the rate $\bar{R}^2 \lesssim 1/n$ when R_w^2 is bounded away from one. Under asymptotics where $R_w^2 \rightarrow 1$ (such as when $k \rightarrow \infty$, and the variance of the propensity score residual shrinks as we add more controls, which captures the notion that the long regression is feasible, but noisy), $\bar{R}^2$ may shrink more slowly, though reasonable calibrations of $\bar{R}^2$ will typically still be uninformative as the finite-sample bounds for κ above illustrate.

Combining the relative efficiency calculations above with the observation that when $R_w^2 = 1$, the optimal linear CI is based on the short regression alone then yields the second result, which gives the rate of convergence of the optimal linear estimator.

Theorem 4.1. *Suppose that $\eta/n \leq \sigma^2/w'(I - H_{Z_1})w \leq 1/(n\eta)$ for some small constant $\eta > 0$, and that $R_{Y-w\beta}^2 \leq \bar{R}^2$ for some $\bar{R}^2 < 1$. Let $\kappa^2 = n\bar{R}^2/(1 - \bar{R}^2)$, and define $C\bar{B}_\lambda$ and sd_λ as in eq. (13). Then there exists a finite constant $K_{\eta,\alpha}$ depending only on η and α such that*

$$\frac{1}{K_{\eta,\alpha}} \frac{1 + \min\{\kappa, 1/\sqrt{1 - R_w^2}\}}{\sqrt{n}} \leq R_{\text{CI}}(C) \leq K_{\eta,\alpha} \frac{1 + \min\{\kappa, 1/\sqrt{1 - R_w^2}\}}{\sqrt{n}}.$$

We note that Theorem 4.1 does not restrict R_w^2 (and thus accommodates scenarios where R_w^2 converges to, or equals one), or how fast k may grow relative to n . In the case where $R_w^2 = 1$ (which holds generically when $k \geq n$), the term $\min\{\kappa, 1/\sqrt{1 - R_w^2}\}$ reduces to $\kappa = \sqrt{n\bar{R}^2/(1 - \bar{R}^2)}$ and the CI length is proportional to $\sqrt{\bar{R}^2/(1 - \bar{R}^2)}$ times a constant that does not depend on n . Thus, consistent estimation requires taking the bound $\bar{R}^2$ to zero with the sample size. This slow rate of convergence is not a feature of the particular way the R^2 penalty normalizes the ℓ_2 norm of γ —in the next subsection we give matching rates of convergence for the optimal linear CI under the unweighted version of the ℓ_2 penalty.

4.3 Rates of convergence under unweighted ℓ_p norms

For ease of notation, we assume all coefficients are constrained, so that $\text{Pen}(\gamma) = \|\gamma\|_p$. We consider asymptotics under which k grows at least as fast as n ; if $k/n \rightarrow 0$ then the optimal rate is $1/\sqrt{n}$ and it can be achieved by the long regression.

Let $q \in [1, \infty]$ denote the Hölder conjugate of p , satisfying $1/p + 1/q = 1$. We will show that when $\text{Pen}(\gamma) = \|\gamma\|_p$, the rate of convergence for the optimal linear CI is at least

$$R_p^*(k, C, R_w^2) = \frac{1 + \min\{Cr_p(k), 1/\sqrt{1 - R_w^2}\}}{\sqrt{n}}, \quad \text{where} \quad r_p(k) = \begin{cases} k^{1/q} & \text{if } p > 1, \\ \sqrt{\log k} & \text{if } p = 1. \end{cases} \quad (17)$$

Furthermore, for $p = 1$ and $p = 2$, we will give lower bounds that match this rate for a range of conditions on how (C, k) changes with the sample size. If $k \gg n$ and C does not decrease to zero with n , these rates require $p < 2$ (so that $q > 2$) for consistency. When $p = 1$, we can then allow k to grow exponentially with n , whereas setting $1 < p < 2$ allows for k to grow at a polynomial rate in n that depends on p . We note that when the optimal rate is slower than $n^{-1/2}$, the bias term is not asymptotically negligible, making it necessary to use our bias-aware critical values even in large samples.

Remark 4.1. These stark differences in convergence rates can be explained by the fact that the constraint $\|\gamma\|_p \leq C$ becomes more stringent with smaller p , since $\|\gamma\|_p \leq \|\gamma\|_r$ for $r < p$. Specifically, consider an ℓ_p constraint $\{\gamma \in \mathbb{R}^k: \|\gamma\|_p \leq C_p\}$, for some $p > 1$, but with the bound C_p calibrated so that the parameter space has the same volume as the parameter space under the ℓ_1 constraint, $\{\gamma \in \mathbb{R}^k: \|\gamma\|_1 \leq C_1\}$, for some fixed bound C_1 . This ensures that the strength of the parameter space restriction is the same under the ℓ_p as under the ℓ_1 constraint. As $k \rightarrow \infty$ this requires setting $C_p \asymp k^{1/p-1}$.⁵ In the high-dimensional setting with $R_w^2 = 1$, this implies the convergence rate $R_p^*(k, C_p, 1) \asymp C_p r_p(k)/\sqrt{n} \asymp 1/\sqrt{n}$, which is the same rate as the convergence rate under the ℓ_1 constraint, up to log factors. In other words, if C_p is calibrated so that the strength of the ℓ_p constraint is the same across different values of p , the convergence rates are also the same. While these calculations are useful for understanding the differences in convergence rates, we do not mean to suggest that the volume of a parameter space should be used as a metric to gauge its plausibility in practice: since taking $C \rightarrow 0$ rules out even a single coefficient being bounded away from zero, imposing an ℓ_1 constraint with a fixed C appears more defensible than imposing an ℓ_2 constraint, but taking $C \rightarrow 0$.

⁵To see this, note the volume of a unit ℓ_p ball is given by $(2\Gamma(1 + 1/p))^k / \Gamma(1 + k/p)$, so that $C_p = \frac{C_1}{\Gamma(1+1/p)} \left(\frac{\Gamma(1+k/p)}{\Gamma(1+k)} \right)^{1/k} = \frac{C_1}{\Gamma(1+1/p)} \frac{e^{1-1/p}}{p^{1/p}} k^{1/p-1} (1 + O(1/k))$.

To derive the upper bound, we assume that the design matrix X is well-behaved in that $X \in \mathcal{E}_n^p(\eta)$, where $\mathcal{E}_n^p(\eta)$ denotes the set of design matrices X for which there exists $\delta \in \mathbb{R}^k$ such that

$$\frac{1}{n}\|w - Z\delta\|_2^2 \leq \frac{1}{\eta}, \quad \frac{1}{n}w'(w - Z\delta) \geq \eta, \quad \frac{1}{n}\|Z'(w - Z\delta)\|_q \leq \frac{r_p(k)}{\sqrt{n\eta}}.$$

Lemma A.1 in Appendix A.4 shows that $X \in \mathcal{E}_n^p(\eta)$ holds with high probability for $\eta > 0$ small enough when w_i, Z_i are drawn i.i.d. from a distribution satisfying mild conditions on moments and covariances if δ is defined as the coefficient from the population version of the propensity score regression, so long as the population propensity score residual is not degenerate. For the condition $X \in \mathcal{E}_n^2(\eta)$, a second sufficient condition that allows for a population propensity score residual that is asymptotically degenerate is that (w_i, Z_i) are drawn i.i.d. from a distribution such that w_i is not degenerate and has bounded second moments, and Z_i is sub-Gaussian. Then $X \in \mathcal{E}_n^2(\eta)$ with high probability for $\eta > 0$ small enough if we let $\delta = 0$. This follows by the operator norm bound, $\|Z'w\|_2 \leq \|Z\|_2\|w\|_2$, and the standard result in random matrix theory that $\|Z\|_2 \lesssim \sqrt{k}$ (e.g., Vershynin 2012, Theorems 5.39 and 5.58).

Theorem 4.2. *Suppose that for some constant $\eta > 0$, $\eta \leq w'w/n \leq 1/\eta$, $\eta \leq \sigma \leq 1/\eta$, and $X \in \mathcal{E}_n^p(\eta)$. Set $\text{Pen}(\gamma) = \|\gamma\|_p$. Then there exists a finite constant $K_{\eta,\alpha}$ depending only on η and α such that $R_{\text{CI}}(C) \leq K_{\eta,\alpha}R_p^*(k, C, R_w^2)$.*

We now give a matching lower bound for $p = 2$. To state the result, let $d_{\max}$ denote the largest singular value of Z , and let $d_{\min}$ denote its smallest non-zero singular value.

Theorem 4.3. *Suppose that for some constant $\eta > 0$, $\eta \leq w'w/n \leq 1/\eta$, $\eta \leq \sigma \leq 1/\eta$, $d_{\max} \geq \sqrt{k}\eta$ and $d_{\max}/d_{\min} \leq 1/\eta$. Set $\text{Pen}(\gamma) = \|\gamma\|_2$. Then there exists a finite constant $K_{\eta,\alpha}$ depending only on η and α such that*

$$R_{\text{CI}}(C) \geq \frac{R_2^*(k, C, R_w^2)}{K_{\eta,\alpha}}.$$

If Z has independent sub-Gaussian rows, standard results in random matrix theory (e.g., Vershynin 2012, Theorems 5.39 and 5.58) imply that the conditions on the singular values of Z hold with probability approaching one provided that the limit of k/n lies outside the interval $[1/\tilde{\eta}, \tilde{\eta}]$, where $\tilde{\eta}$ depends only on the sub-Gaussian norm of the rows. If Z has i.i.d. Gaussian rows, it is sufficient to take any $\tilde{\eta} > 1$ (e.g., Vershynin 2012, Corollary 5.35). Furthermore, these conditions have also been shown to hold for more general distributions and

$\tilde{\eta}$ large enough, provided that Z has finite fourth moments (e.g., Srivastava and Vershynin 2013).

Hence, for a large class of design matrices X , the convergence rate of the optimal linear CI under the unweighted ℓ_2 penalty matches that under the $R_{Y-w\beta}^2$ penalty. In particular, rate improvements upon the long regression require setting $C \lesssim 1/\sqrt{k(1-R_w^2)}$. If $R_w^2 = 1$, consistency requires taking $C \rightarrow 0$, which rules out even a single coefficient being bounded away from zero.

We next consider a lower bound for the case $p = 1$. Rather than imposing conditions on X in a fixed design setting that hold with high probability under random design, we directly consider a random design setting, and we do not condition on X when requiring coverage of CIs. This allows us to strengthen the conclusion of our theorem by showing that the rate in Theorem 4.2 is sharp even if one imposes a linear model for w_i given z_i along with sparsity and ℓ_1 bounds on the coefficients in this model. The random design model will also imply that $R_w^2 = 1$ for n large enough when $k \geq n$. In particular, we assume that

$$\begin{aligned} Y &= w\beta + Z\gamma + \varepsilon, \quad \varepsilon \mid Z, w \sim \mathcal{N}(0, \sigma^2 I_n), \\ w &= Z\delta + v, \quad v \mid Z \sim \mathcal{N}(0, \sigma_v^2 I_n), \\ z_{ij} &\sim \mathcal{N}(0, 1) \quad \text{i.i.d. over } i, j. \end{aligned}$$

We use P_ϑ and E_ϑ for probability and expectation when Y, X follow this model with parameters $\vartheta = (\beta, \gamma', \delta', \sigma^2, \sigma_v^2)'$. Let $\sigma_0^2 > 0$ and $\sigma_{v,0}^2 > 0$ be given and let $\Theta(C, s, \eta)$ denote the set of parameters $\vartheta = (\beta, \gamma', \delta', \sigma^2, \sigma_v^2)$ where $|\sigma^2 - \sigma_0^2| \leq \eta$, $|\sigma_v^2 - \sigma_{v,0}^2| \leq \eta$, $\|\gamma\|_1 \leq C$, $\|\delta\|_1 \leq C$, $\|\gamma\|_0 \leq s$ and $\|\delta\|_0 \leq s$.

Theorem 4.4. *Let $\hat{\beta} \pm \hat{\chi}$ be a CI satisfying $P_\vartheta(|\beta - \hat{\beta}| \leq \hat{\chi}) \geq 1 - \alpha$ for all ϑ in $\Theta(C, C \cdot K\sqrt{n/\log k}, \eta)$ where $\alpha < 1/2$. Suppose $k \rightarrow \infty$, $C\sqrt{(\log k)/n} \rightarrow 0$, and $c_0\sqrt{(\log k)/n} \leq C \leq \sqrt{k/n} \cdot k^{-\tilde{\eta}}$ for some $c_0 > 0$ and $\tilde{\eta} > 0$. Then, there exists c such that, if K is large enough and $\eta \rightarrow 0$ slowly enough, the expected length of this CI under the parameter vector ϑ^* given by $\beta = 0$, $\gamma = 0$, $\delta = 0$, $\sigma^2 = \sigma_0^2$, $\sigma_v^2 = \sigma_{v,0}^2$ satisfies $E_{\vartheta^*}[\hat{\chi}] \geq c \cdot n^{-1/2}(1 + C\sqrt{\log k})$ once n is large enough.*

Theorem 4.4 follows from similar arguments to Cai and Guo (2017) and Javanmard and Montanari (2018), who provide similar bounds for the case where only a sparsity bound is imposed. According to Theorem 4.4, imposing sparsity does not allow one to improve upon the CI that uses only the ℓ_1 bound $\|\gamma\|_1 \leq C$ (thereby attaining the rate in Theorem 4.2), unless one imposes a sparsity bound that is smaller than $C\sqrt{n/\log k}$. We provide further comparison with CIs that impose sparsity in the next section.

The condition $C \leq \sqrt{k/n} \cdot k^{-\tilde{\eta}}$ in Theorem 4.4 places a lower bound on the rate at which k must increase. If C is a positive constant, it requires that we take $\tilde{\eta} < 1/2$ and that $k \gtrsim n^{1/(1-2\tilde{\eta})}$, so that k must grow faster than n . If k/n converges to a constant that is strictly smaller than one, the optimal convergence rate is $1/\sqrt{n}$ and it can be achieved by the long regression. We conjecture that if k/n converges to a constant greater than one, or if k/n diverges more slowly than any positive power of n , the optimal rate is between $1/\sqrt{n}$ and $\sqrt{\log k}/\sqrt{n}$. However, showing this would require a different argument. To our knowledge, this case is also not covered in the literature on sparse high-dimensional regression.

5 Comparison with sparsity constraints

Several authors have considered CIs for β using “double lasso” estimators (see, among others, Belloni, Chernozhukov, and Hansen 2014; Javanmard and Montanari 2014; van de Geer et al. 2014; C.-H. Zhang and S. S. Zhang 2014). These CIs are valid under the parameter space

$$\tilde{\Gamma}(s) = \{\gamma: \|\gamma\|_0 \leq s\}, \quad (18)$$

where $\|\gamma\|_0 = \#\{j: \gamma_j \neq 0\}$ is the ℓ_0 “norm,” which indexes the sparsity of γ , and with s increasing slowly enough relative to n and k . Since $\|\gamma\|_0$ is not a true norm or seminorm (it is nonconvex), this parameter space is not covered by our setup. Nonetheless, as we show in Appendix A.3, if the sparsity assumption is used to bound the ℓ_1 loss of a preliminary lasso estimator, arguments from Section 2 lead to estimators and CIs that are analogous to those proposed in the double lasso literature.

When should one use a double lasso CI, and when should one use the approach in the present paper? In principle, this depends on the *a priori* assumptions one is willing to make, and whether they are best captured by a sparsity bound or a bound on convex penalty function, such as the ℓ_1 or ℓ_2 norm. In many settings, it may be difficult to motivate the assumption that a regression function has a sparse approximation, whereas upper bounds on the magnitude of the coefficients may be more plausible.⁶

A key advantage of the CIs and estimators we propose is that they have sharp finite-sample optimality properties and coverage guarantees in the fixed design Gaussian model with known error variance. While this is an idealized setting, the worst-case bias calculations do not depend on the error distribution, and remain the same under non-Gaussian, heteroskedastic errors. Our approach directly accounts for the potential finite-sample bias

⁶One may also formalize bounds on the magnitude of the coefficients by imposing a nonconvex ℓ_p penalty with $p \in (0, 1)$. However, such constraints are harder to interpret than the ℓ_1 and ℓ_2 bounds that are the focus of this paper.

of the estimator, rather than relying on “asymptotic promises” about rates at which certain constants involved in bias terms converge to zero.

On the flip side, our CIs require an explicit choice of the penalty bound C in order to form a “bias-aware” CI. In contrast, CIs based on double lasso estimators do not require explicitly choosing the regularity (in this case, the sparsity s), since they ignore bias. This is justified under asymptotics in which s increases more slowly than $\sqrt{n}/\log k$, which ensures that the double lasso bias is asymptotically negligible. However, such asymptotic promises are difficult to evaluate in a given finite-sample setting. Indeed, as shown by Wüthrich and Y. Zhu (2023) and confirmed in our Monte Carlo simulations in Section 6 below, the double lasso CI leads to undercoverage in finite samples even in relatively sparse settings. To ensure good finite-sample coverage of the CI in eq. (22), one needs to ensure that the actual finite-sample bias is negligible relative to the standard deviation of the estimator. But since any bias bound depends on the sparsity index s (as in the bound in eq. (21)), this gets us back to having to explicitly specify s .

Thus, CIs that ignore bias such as conventional CIs based on double lasso estimators do not avoid the problem of specifying s or C : they merely make such choices implicit in their asymptotic promises. These issues show up formally in the asymptotic analysis of such CIs. In particular, double lasso CIs require the “ultra sparse” asymptotic regime $s = o(\sqrt{n}/\log k)$, and they undercover asymptotically in the “moderately sparse” regime where s increases more slowly than $n/\log k$ with $s \gg \sqrt{n}/\log k$. Indeed, Theorem 4.4 above, as well as the results of Cai and Guo (2017) and Javanmard and Montanari (2018) show that it is impossible to avoid explicitly specifying s if one allows for the moderately sparse regime.

On the other end of the spectrum, in the “low dimensional” regime where $k \ll n$, the double lasso CI is asymptotically equivalent to the usual CI based on the long regression. Thus, the double lasso CI cannot be used when the goal is to use *a priori* information on γ to improve upon the CI based on the long regression (as in, for example, Muralidharan, Romero, and Wüthrich 2025). In contrast, our approach optimally incorporates the bound C regardless of the asymptotic regime.

We emphasize that the discussion above focuses on CIs rather than estimation. Debiased lasso estimators may perform well in terms of mean squared error across a range of regimes. In particular, they may perform well when the ℓ_1 norm of γ is small (but γ is not sparse), even if they do not match the performance of the optimal linear estimator we propose that explicitly targets this case.

6 Simulation results

We now illustrate the performance of our methods in simulations, focusing on the ℓ_1 penalty from Example 2.2 and $R_{Y-w\beta}^2$ penalty from Example 2.1.

We consider a design taken from Belloni, Chernozhukov, and Hansen (2014), with data generated from a random regressor model that supplements eq. (1) with a propensity score regression

$$w_i = z_i' \pi + \sigma_{\tilde{w}} \tilde{w}_i,$$

with $\tilde{w}_i$ and ε_i independent standard Gaussian. The design matrix Z is random, with i.i.d. rows $z_i \sim \mathcal{N}(0, \Sigma)$ and $\Sigma_{ij} = 2^{-|i-j|}$. Similar to Wüthrich and Y. Zhu (2023), we tweak the Belloni, Chernozhukov, and Hansen (2014) design by considering regression coefficients that are of similar magnitude rather than decaying, so that we can separately vary the degree of sparsity and the signal-to-noise ratio. Specifically, we set

$$\gamma_j = \pi_j = \begin{cases} c_1 & \text{if } j \leq k_1, \\ c_2 & \text{if } k_1 < j \leq k_1 + s, \\ 0 & \text{otherwise.} \end{cases}$$

The data generating process (DGP) is therefore indexed by 9 parameters: n , k_1 , k_2 , s , β , σ , $\sigma_{\tilde{w}}$, c_1 and c_2 . We fix $k_1 = 5$, $\sigma = 1$, and $\beta = 0$, and consider $n \in \{500, 1000\}$, $k_2 \in \{0.2n, 0.8n\}$, $s \in \{10, 20, 100\}$, and $\sigma_{\tilde{w}} \in \{0.5, 1\}$. We calibrate c_1 and c_2 by fixing R_Z^2 , the population R^2 from the regression of Y on Z , and fixing the ratio $\rho = \|\gamma_2\|_1 / \|\gamma_{\text{short}}\|_1$, where γ_{short} is the population short regression coefficient. This allows us to directly control the signal-to-noise ratio, and the validity of the rule-of-thumb calibration that sets $\rho = 1$ and replaces γ_{short} with an estimate. We consider $R_Z^2 \in \{0.1, 0.5\}$ and $\rho \in \{0.2, 1, 1.4, 2\}$. This gives a total of 192 DGPs.

We consider the following six methods for constructing nominal 95% CIs for β :

AKK1 Implements Algorithm 3.1 with the penalty given by the ℓ_1 norm of γ_2 , the last k_2 regression coefficients. Following Remark 3.1, the initial residual estimates use the lasso estimator (that only penalizes γ_2), with the penalty chosen via 10-fold cross-validation. Following Remark 3.2, we set $C = \rho \|\hat{\gamma}_{\text{short}}\|_1$ with $\rho = 1$, where $\hat{\gamma}_{\text{short}}$ are OLS estimates from a short regression that only includes the first k_1 controls.

BCH The double lasso procedure of Belloni, Chernozhukov, and Hansen (2014), implemented via the R package `hdn` without penalizing the k_1 baseline controls.

LR The OLS estimator based on the long regression with HC1 standard errors.

Oracle AKK1 The oracle version of AKK1, which sets the penalty bound C to the actual value of $\|\gamma_2\|_1$, and uses knowledge of the variance of the error term ε_i .

Oracle AKK2 Implements Algorithm 3.1 with the $R_{Y-w\beta}^2$ penalty, using knowledge of the actual value of $R_{Y-w\beta}^2$ to set the bound $\bar{R}^2$, and the variance of the error term ε_i .

Oracle LM The procedure proposed by C. M. Li and Müller (2021), implemented using knowledge of the actual value of $R_{Y-w\beta}^2$ and the variance of the error term.

We focus on the oracle versions of the $R_{Y-w\beta}^2$ penalty and the C. M. Li and Müller (2021) procedure for two reasons. First, it ensures the comparison between these two methods is not clouded by tuning parameter choices; second, it allows us to illustrate the theoretical results from Section 4.2 that unless the sample size is moderate, efficiency gains relative to the long regression are modest even under the oracle calibration.

Table 1 reports the simulation results for DGPs with $n = 500$, and Table 2 reports them for DGPs with $n = 1000$. In line with the theory, all oracle CIs have coverage close to nominal across the board, as does the long regression. For AKK1, the coverage is also close to nominal so long as $\rho \leq 1$, so that the population restriction underlying the rule of thumb is satisfied. Furthermore, in these designs, the method is robust to mild violations of the population constraint, with coverage remaining above 90% when $\rho = 1.4$. Only when $\rho = 2$ do we start to see mild undercoverage in select designs, as the red cells, which highlight coverage below 90%, indicate. By contrast, the BCH method displays mild undercoverage even in sparse designs with $s = 10$, and coverage dips below 75% once $s = 20$. Coverage drops off to almost zero in the non-sparse designs with $s = 100$. These results illustrate the concern discussed in Section 5 that asymptotic sparsity requirements may be difficult to evaluate in finite samples.

Panel B shows that AKK1 regularization yields substantial length reductions over the long regression, especially when k_2 is large relative to n : with $k_2 = .8n$, the AKK1 CIs are roughly half as long as the long regression CI. With $k_2 = .2n$, the gains are more modest, with the AKK1 CIs at 91–96% of the long regression length. By contrast, the oracle AKK2 and LM procedures that bound $R_{Y-w\beta}^2$ display much more modest efficiency gains, even though they use the true value of $R_{Y-w\beta}^2$. To help put the gains in context, the tables report the value of $R_{Y-w\beta}^2$ in the second column. In line with the theory in Section 4.2, the percentage gains in length are typically in the single digits unless $R_{Y-w\beta}^2$ is less than about 3%; both methods perform similarly.

		$k_2 = 100$						$k_2 = 400$					
		Feasible methods			Oracles			Feasible methods			Oracles		
ρ	$R_{Y-w\beta}^2$	AKK1	BCH	LR	AKK1	AKK2	LM	AKK1	BCH	LR	AKK1	AKK2	LM
Panel A: Coverage Probability, minimum across DGPs													
$s = 10$													
0.2	0.011	94.5	92.5	94.3	94.6	94.4	94.3	96.6	90.8	94.0	94.6	98.1	96.9
1.0	0.106	92.8	92.4	94.2	94.8	94.8	94.5	93.4	90.3	94.1	95.1	94.9	95.2
1.4	0.139	92.0	92.2	92.7	93.6	93.1	93.2	90.4	87.9	93.5	94.7	94.6	95.3
2.0	0.175	88.9	91.5	93.7	94.7	94.6	94.5	84.8	88.0	94.0	94.3	95.3	94.8
$s = 20$													
0.2	0.006	94.9	92.9	94.8	95.0	95.3	94.8	95.8	91.5	94.4	93.9	97.5	96.4
1.0	0.078	93.0	85.2	94.2	94.2	94.4	94.7	92.7	74.4	93.7	93.9	94.4	94.7
1.4	0.109	90.8	88.3	94.3	94.1	93.7	94.6	91.4	83.6	93.7	95.3	94.2	94.2
2.0	0.145	89.0	85.0	94.1	94.0	94.1	95.2	82.3	81.1	93.4	93.9	94.2	94.4
$s = 100$													
0.2	0.001	94.8	94.5	94.4	95.0	94.9	95.2	95.1	93.2	93.3	94.1	95.3	95.1
1.0	0.028	94.2	59.5	94.3	95.1	94.9	94.6	95.8	59.8	94.8	96.1	96.9	96.0
1.4	0.045	93.8	37.5	94.2	94.7	94.6	94.4	94.7	27.9	94.6	97.0	96.4	96.0
2.0	0.071	91.7	12.5	93.7	94.3	94.2	94.6	90.9	5.9	93.8	97.0	95.6	95.5
Panel B: Relative Length, mean across DGPs													
$s = 10$													
0.2	0.011	0.96	0.89		0.92	0.97	0.97	0.55	0.43		0.46	0.70	0.70
1.0	0.106	0.94	0.90		0.95	1.00	0.97	0.51	0.44		0.52	0.93	0.90
1.4	0.139	0.92	0.90		0.96	1.00	0.97	0.49	0.44		0.53	0.95	0.91
2.0	0.175	0.91	0.90		0.96	1.00	0.96	0.47	0.44		0.54	0.96	0.90
$s = 20$													
0.2	0.006	0.96	0.89		0.92	0.96	0.96	0.56	0.43		0.46	0.64	0.64
1.0	0.078	0.94	0.89		0.96	0.99	0.98	0.52	0.43		0.53	0.91	0.88
1.4	0.109	0.93	0.90		0.97	1.00	0.97	0.50	0.43		0.55	0.94	0.90
2.0	0.145	0.92	0.90		0.97	1.00	0.97	0.49	0.44		0.56	0.95	0.91
$s = 100$													
0.2	0.001	0.96	0.89		0.92	0.93	0.94	0.56	0.43		0.46	0.52	0.52
1.0	0.028	0.95	0.88		0.97	0.98	0.99	0.55	0.43		0.55	0.80	0.79
1.4	0.045	0.95	0.87		0.98	0.99	0.98	0.54	0.42		0.58	0.85	0.84
2.0	0.071	0.94	0.87		0.98	0.99	0.98	0.52	0.42		0.61	0.90	0.87

Notes: Panel A reports, for each method, the minimum coverage probability of nominal 95% CIs over $\sigma_{\tilde{w}} \in \{0.5, 1\}$ and $R_Z^2 \in \{0.1, 0.5\}$, the four DGPs that vary. Panel B reports mean relative length across the four DGPs. Within each DGP, relative length is the method's average CI length across simulation draws divided by the average length of the long regression CI. 1,000 Monte Carlo draws. Cells with coverage below 90% highlighted in red.

Table 1: Finite-sample coverage of nominal 95% confidence intervals and their average length relative to the long regression ($n = 500$).

		$k_2 = 200$						$k_2 = 800$					
		Feasible methods			Oracles			Feasible methods			Oracles		
ρ	$R_{Y-w\beta}^2$	AKK1	BCH	LR	AKK1	AKK2	LM	AKK1	BCH	LR	AKK1	AKK2	LM
Panel A: Coverage Probability, minimum across DGPs													
$s = 10$													
0.2	0.011	94.7	92.8	94.8	95.1	95.0	94.6	95.9	90.2	93.4	95.3	95.9	95.1
1.0	0.106	94.3	94.1	94.5	94.7	94.2	95.2	93.4	91.4	94.5	94.2	95.0	94.6
1.4	0.139	92.1	93.7	94.8	93.5	95.0	95.0	91.2	93.0	93.8	94.4	93.6	94.2
2.0	0.175	90.5	94.0	93.9	93.3	93.7	94.1	84.7	93.8	94.3	94.6	94.7	95.0
$s = 20$													
0.2	0.006	94.7	91.3	94.5	95.2	95.3	94.9	94.7	91.4	93.7	94.2	97.6	96.2
1.0	0.078	93.3	89.6	94.4	94.4	94.3	94.8	92.9	86.7	93.7	93.7	94.4	94.0
1.4	0.109	92.6	85.0	94.5	94.4	94.7	94.4	90.2	83.8	94.0	94.4	94.1	94.1
2.0	0.145	90.3	83.7	93.8	93.9	94.1	94.9	84.0	75.7	94.7	94.4	94.6	95.3
$s = 100$													
0.2	0.001	93.5	92.8	92.7	93.8	93.5	92.8	95.5	94.5	95.0	95.5	96.3	96.0
1.0	0.028	93.6	38.1	93.5	93.9	93.9	93.7	95.4	31.7	94.0	95.7	95.3	94.9
1.4	0.045	94.6	16.5	94.8	94.8	95.3	94.7	94.8	7.7	94.2	96.4	95.6	95.4
2.0	0.071	93.5	3.8	93.7	94.2	94.4	94.2	88.3	1.5	94.4	96.6	95.0	95.3
Panel B: Relative Length, mean across DGPs													
$s = 10$													
0.2	0.011	0.96	0.89		0.92	0.98	0.99	0.56	0.44		0.46	0.78	0.78
1.0	0.106	0.94	0.90		0.95	1.00	0.98	0.52	0.44		0.52	0.96	0.96
1.4	0.139	0.93	0.90		0.96	1.00	0.98	0.50	0.44		0.54	0.97	0.96
2.0	0.175	0.92	0.90		0.96	1.00	0.97	0.49	0.44		0.55	0.98	0.95
$s = 20$													
0.2	0.006	0.96	0.89		0.92	0.97	0.98	0.57	0.44		0.46	0.72	0.72
1.0	0.078	0.95	0.90		0.96	1.00	0.99	0.53	0.44		0.54	0.95	0.95
1.4	0.109	0.94	0.90		0.97	1.00	0.98	0.52	0.44		0.56	0.97	0.95
2.0	0.145	0.93	0.90		0.97	1.00	0.97	0.50	0.44		0.57	0.97	0.96
$s = 100$													
0.2	0.001	0.96	0.89		0.92	0.94	0.95	0.57	0.44		0.47	0.57	0.58
1.0	0.028	0.96	0.88		0.97	0.99	0.99	0.56	0.43		0.56	0.87	0.87
1.4	0.045	0.95	0.88		0.98	0.99	0.99	0.55	0.43		0.59	0.92	0.92
2.0	0.071	0.94	0.88		0.98	1.00	0.99	0.53	0.43		0.62	0.95	0.94

Notes: See Table 1.

Table 2: Finite-sample coverage of nominal 95% confidence intervals and their average length relative to the long regression ($n = 1000$).

7 Empirical applications

We revisit two empirical applications to illustrate the performance of our methods using real data. The first application conducts inference on the marginal propensity to earn (MPE) out of unearned income using data from Imbens, Rubin, and Sacerdote (2001), and the second reanalyzes the effectiveness of RHC, using data from Connors et al. (1996).

7.1 Effect of unearned income on labor earnings

The MPE out of unearned income is a key structural parameter in labor and public economics. While unearned income is typically endogenous, Imbens, Rubin, and Sacerdote (2001) argue that in their sample of $n = 496$ winners of major and minor prizes in the Massachusetts lottery in 1984–88, the magnitude of winnings is as good as random conditional on individual characteristics that proxy for the frequency of lottery ticket purchases. The lottery winnings are paid out over 20 years, so that in a regression of the average social security earnings in the 6 years after the lottery, Y_i , on yearly lottery payments, w_i , and individual controls, the coefficient on w_i may be interpreted as the MPE, under the assumption that the MPE is constant.

We relax the assumption that the MPE is constant using a specification taken from C. M. Li and Müller (2021): we augment the baseline set of $k_1 = 7$ individual controls z_{i1} consisting of the intercept, two continuous controls (years of education and age), and 4 binary controls (indicators for male, college, age over 55, and age over 65) with $k_2 = 25$ additional controls z_{i2} that are constructed by taking demeaned cross-products of the four baseline binary controls and their interactions with w_i and dropping collinear terms. All non-binary covariates contained in z_{i1} and z_{i2} are standardized. Following the discussion in Remark 2.1, the coefficient on w_i in this specification can be interpreted as the average MPE, allowing for heterogeneity in the MPE with respect to the binary controls. By contrast, the short regression estimand in a regression that only includes z_{i1} does not yield the average MPE in the presence of such heterogeneity.

We consider the same five methods as in the simulations: AKK1, BCH, long regression, AKK2, and the C. M. Li and Müller (2021) method. In addition, we report the short regression that omits the interactions with the baseline controls.⁷ The results are reported in Panel A of Table 3.

The MPE estimate in the long regression is close to the short regression estimate -0.052

⁷To make the methods more comparable and not conflate the comparison with differences in standard error construction, for variance calculations, all methods use residual estimates based on a lasso estimator that penalizes only γ_2 , with penalty chosen by 10-fold cross-validation.

Method	Point estimate	95% Confidence interval	CI length
Panel A: Imbens, Rubin, and Sacerdote (2001)			
Short regression	−0.0522	[−0.0736, −0.0309]	0.0427
Long regression	−0.0493	[−0.1158, 0.0171]	0.1329
AKK1 ($\rho = 1$)	−0.0600	[−0.0982, −0.0218]	0.0764
AKK1 ($\rho = 2$)	−0.0567	[−0.1065, −0.0068]	0.0997
BCH	−0.0557	[−0.0804, −0.0310]	0.0494
AKK2 ($\bar{R}^2 = .05$)	−0.0497	[−0.1099, 0.0104]	0.1203
Li–Müller ($\bar{R}^2 = .05$)	−0.0504	[−0.1168, 0.0161]	0.1329
Panel B: Connors et al. (1996)			
Short regression	0.0635	[0.0385, 0.0885]	0.0500
Long regression	0.0482	[0.0026, 0.0938]	0.0912
AKK1 ($\rho = 1$)	0.0410	[0.0053, 0.0768]	0.0715
AKK1 ($\rho = 2$)	0.0444	[0.0060, 0.0828]	0.0768
BCH	0.0562	[0.0297, 0.0827]	0.0530
AKK2 ($\bar{R}^2 = .05$)	0.0484	[0.0030, 0.0938]	0.0908
Li–Müller ($\bar{R}^2 = .05$)	0.0482	[0.0025, 0.0940]	0.0915

Notes: AKK1 uses an ℓ_1 penalty bound with the rule-of-thumb calibration $\rho = 1$; $\rho = 2$ is reported as a robustness check. AKK2 and the C. M. Li and Müller (2021) method impose $R_{Y-w\beta}^2 \leq 0.05$. For variance calculations, all methods use residual estimates based on a lasso estimator that penalizes only γ_2 , with penalty chosen by 10-fold cross-validation.

Table 3: Performance of penalized regression methods in empirical applications.

that only includes the baseline controls z_{i1} and corresponds to the specification in Table 4, column II, row 1 in Imbens, Rubin, and Sacerdote (2001). However, the long regression estimate is very noisy: the 95% confidence interval includes positive values for the average MPE, which economic theory rules out, and it is over three times as long as the short regression CI. To increase precision, C. M. Li and Müller (2021) consider restricting the sum of the squared effects, $\sum_{i=1}^n (\tilde{z}'_{i2} \gamma_2)^2$, where $\tilde{z}_{i2}$ are the residuals from projecting z_{i2} onto z_{i1} . To make this restriction more interpretable, we implement this penalty by restricting $R_{Y-w\beta}^2 \leq \bar{R}^2$ as in Example 2.1, and calibrate the bound to $\bar{R}^2 = 0.05$. Given that the R^2 in the short regression equals 0.244, this choice of $\bar{R}^2$ limits the explanatory power of the interaction terms to be about a fifth of the main effects, formalizing the empirical regularity that main effects are more important. Ex post, this regularity turns out to hold in this dataset as well, given the long regression R^2 is only 0.285. This choice of $\bar{R}^2$ yields the C. M. Li and Müller (2021) CI (−0.117, 0.016), with essentially the same length as the long regression

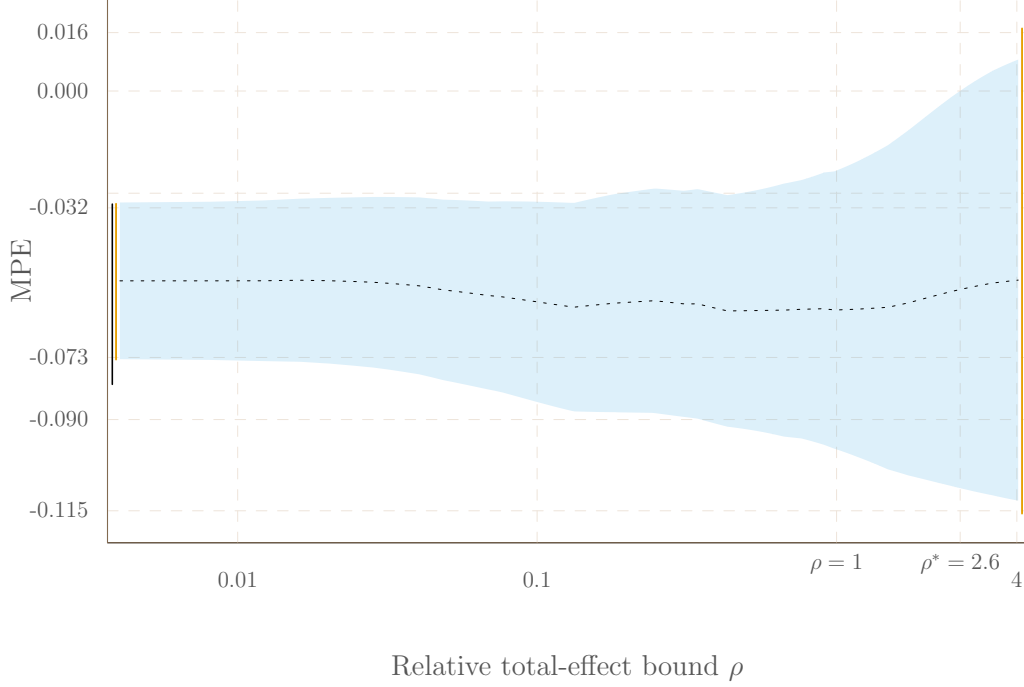


Figure 1: 95% CIs for the MPE out of unearned income under the ℓ_1 penalty.

Notes: Orange solid vertical lines at the left and right endpoints of the x -axis mark CIs based on the short and long regressions, respectively. The black solid vertical line at the left endpoint of the x -axis marks the BCH CI. The blue shaded area depicts the bias-aware CIs, while the dotted black line shows the point estimates as a function of $\rho = C/\|\hat{\gamma}_{\text{short}}\|_1$, the relative total-effect bound. The default $\rho = 1$ and breakdown value ρ^* are marked on the x -axis (log scale).

CI.⁸ By contrast, the AKK2 CI, $(-0.110, 0.010)$, is 9.5% shorter than the long regression CI, although it still includes zero. This moderate improvement over the long regression is consistent with the theory in Section 4.2, which under a homoskedastic benchmark, predicts about a 5.5% improvement.

Limiting the total effect of the additional controls by imposing an ℓ_1 penalty bound as in Example 2.2 yields more substantive precision gains. At the baseline calibration $\rho = 1$, the point estimate is similar to the long regression estimate, but the CI is only about 60% as long, and excludes zero. At $\rho = 2$, the CI widens a bit, yielding $(-0.107, -0.007)$. Figure 1 examines the sensitivity of the method to the choice of ρ more thoroughly, showing that moderate changes in ρ have little impact. The breakdown value for statistical significance is

⁸C. M. Li and Müller (2021) show that their method is close to optimal in terms of weighted average length under a homoskedastic benchmark. This may no longer be the case under heteroskedasticity. Their CI is variable-length, and may be longer than the long regression CI in some samples even in the homoskedastic case. In contrast, our construction guarantees length improvements over the long regression in all samples under homoskedasticity.

$\rho^* = 2.59$: the total effect of the additional controls would need to be more than 2.6 times that of the baseline controls in order for the CI to include zero.⁹ The double lasso CI of Belloni, Chernozhukov, and Hansen (2014), depicted by a black vertical line in the figure, is very short. However, as the simulation results in Section 6 demonstrate, its coverage guarantees may be unreliable when the underlying sparsity conditions are not met, and in this application, it is not clear why the coefficients on the additional controls should admit a sparse approximation.

The point estimates are remarkably stable across specifications. Thus, while the penalized regression results indicate that there is more uncertainty about the precise magnitude of the average MPE than the short regression suggests, we find little evidence that the short regression is biased.

7.2 Right heart catheterization

We revisit the observational study of Connors et al. (1996), who found that RHC led to an increase in 30-day mortality among critically ill patients. We follow the empirical design in Hirano and Imbens (2001). After removing observations with missing values, the final analysis sample consists of $n = 5,735$ patients. The treatment is an indicator for RHC, and the outcome is a binary indicator for 30-day mortality. The short regression includes the treatment indicator and the original 72 baseline covariates used in Hirano and Imbens (2001). Here we examine the robustness of their results to functional form misspecification by including second- and third-order interactions of the baseline controls.¹⁰ The long regression thus includes $k_1 = 72$ baseline controls and $k_2 = 4,078$ higher-order terms.¹¹

We consider the same six methods as in the previous application. Panel B of Table 3 reports the results.¹² All methods yield positive point estimates and CIs, consistent with the original finding that RHC is associated with higher mortality. However, there is some disagreement between the methods about the magnitudes. The long regression estimate that RHC leads to a 4.8 percentage point increase in 30-day mortality is close to the lower endpoint of the short regression CI. At the same time, the long regression CI is over 80%

⁹Remark C.1 in Appendix C examines the sensitivity of the results to imposing a bound on the maximal Lindeberg weight, defined in eq. (14). The AKK1 and AKK2 CIs widen somewhat, as does the short regression CI, but the qualitative conclusions remain unchanged. However, the long regression CI becomes uninformative, and so does the Li-Müller procedure.

¹⁰Specifically, the additional covariates consist of squares and cubes of all continuous baseline covariates, all pairwise interactions (continuous×continuous, continuous×dummy, and dummy×dummy), and higher-order cross-terms (products of polynomial transformations up to third order), after removing constant columns.

¹¹The long regression drops 200 covariates due to collinearity. Since our procedure does not require the design matrix to be nonsingular, we keep all covariates when applying our methods.

¹²The maximal Lindeberg weights for all methods are small; see Remark C.1.

wider than the short regression CI. The short regression R^2 is 0.223. We again calibrate the $\bar{R}^2$ bound for AKK2 and the C. M. Li and Müller (2021) method to $\bar{R}^2 = 0.05$ to capture the notion that the interactions have about a fifth of the explanatory power of the main effects. Consistent with the theoretical calculations under the homoskedastic benchmark in Section 4.2, however, due to the large sample size, both methods yield CIs that are very close to the long regression. The BCH CI is very short, and close to the short regression. However, it is not clear why the long regression should admit a sparse approximation. The ℓ_1 penalty calibrated with $\rho = 1$ yields a 22% precision gain relative to the long regression, with the CI given by (0.005, 0.077); the point estimate, 0.041, implies a treatment effect that is about two percentage points lower than the short regression estimate. At $\rho = 2$, the CI changes little. Overall, the analysis indicates that the short regression result that RHC increases mortality is robust to functional form assumptions, though the magnitude of the effect is likely about two percentage points lower than the short regression estimate suggests.

Appendix A Extensions and additional results

A.1 Partially linear model

To flexibly control for a low-dimensional set of covariates $\tilde{z}_i$, one may specify a semiparametric model

$$Y_i = w_i\beta + h(\tilde{z}_i) + \varepsilon_i, \quad \widetilde{\text{Pen}}(h) \leq \tilde{C},$$

where the penalty $\widetilde{\text{Pen}}(h)$ is a seminorm on functions $h(\cdot)$ that penalizes the “roughness” of h , such as the Hölder or Sobolev seminorm. Minimax linear estimation in this model for particular choices of $\widetilde{\text{Pen}}(h)$ has been considered in Heckman (1988). This setting is covered by our setup if we define $Z = I_n$, $\gamma_i = h(\tilde{z}_i)$, and $\text{Pen}(\gamma) = \min_{h: h(\tilde{z}_i) = \gamma_i, i=1, \dots, n} \widetilde{\text{Pen}}(h)$ (assuming the minimum is attained). Note, in particular, that the assumption that $\widetilde{\text{Pen}}(\cdot)$ is a seminorm on the space of functions h guarantees that $\text{Pen}(\cdot)$ is a seminorm on $\mathbb{R}^n$. Theorem 2.1 then implies that the optimal estimator takes the form

$$\hat{\beta}_\lambda = \frac{\sum_{i=1}^n (w_i - g_\lambda(\tilde{z}_i)) Y_i}{\sum_{i=1}^n (w_i - g_\lambda(\tilde{z}_i)) w_i},$$

where $g_\lambda(\cdot)$ solves

$$\min_g \sum_{i=1}^n (w_i - g(\tilde{z}_i))^2 \quad \text{s.t.} \quad \widetilde{\text{Pen}}(g) \leq t$$

for some value of t ; $g_\lambda(\cdot)$ is thus analogous to the regularized regression estimate π_λ in eq. (7).

When $\widetilde{\text{Pen}}$ is the Sobolev seminorm, this yields a spline estimate g_λ (see, for example

Wahba 1990). Interestingly, the estimator proposed in the seminal work by Robinson (1988) takes a similar form to the estimator $\hat{\beta}_\lambda$, involving residuals from a nonparametric regression of w on $\tilde{z}_i$. While the analysis in Robinson (1988) is asymptotic, our results imply that a version of this estimator has sharp finite-sample optimality properties.

A.2 Heterogeneous treatment effects

A more general approach to handling treatment effect heterogeneity relative to the one discussed in Remark 2.1 is to consider the model

$$Y_i = w_i\tau(z_i) + z_i'\gamma + \varepsilon_i, \quad (19)$$

where $\tau(z)$ represents the (possibly nonlinear in z) conditional ATE for units i with $z_i = z$. The parameter of interest is given by $\beta = \int \tau(z) d\mu(w, z)$, where μ is a signed measure defined on some set that includes the empirical support $\{(w_i, z_i)\}_{i=1}^n$. Allowing μ to be signed allows for inference on nonconvex averages of $\tau(z)$. We allow μ to place nonzero mass outside the empirical support $\{(w_i, z_i)\}_{i=1}^n$, thereby allowing for extrapolation. The general theory of estimation and inference on linear functionals developed in Donoho (1994) and Armstrong and Kolesár (2018) underlying Theorem 2.1 and corollary 4.1 can be applied to inference on this parameter under any convex and symmetric restriction on the function $\tau(\cdot)$ and the parameter γ —only the worst-case bias calculation in eq. (4) changes.

Assuming $\tau(z)$ is linear and setting μ to be a weighted empirical measure with weights c_i that sum to one yields the specification discussed in Remark 2.1. Another case of interest arises when we compute the same linear estimator and bias-aware CIs as in the homogeneous TE model in eq. (1). Allowing for TE heterogeneity then requires changing the interpretation of the estimand as targeting a particular weighted average of TEs given in the next theorem. To set the stage for the theorem, let us denote the worst-case bias of a linear estimator $\hat{\beta} = a'Y$ relative to an estimand $\int \tau(z) d\mu(w, z)$ when the heterogeneity is completely unrestricted by $\widetilde{\text{bias}}_\Gamma(\hat{\beta}; \mu) = \sup_{\tau(\cdot), \gamma} [\sum_{i=1}^n a_i w_i \tau(z_i) + a'Z\gamma - \int \tau(z) d\mu(w, z)]$, where the supremum is over $\gamma \in \Gamma$ and all measurable functions $\tau(\cdot)$. For an n -vector a , let $\mu_{a,w}^*(w, z)$ denote a weighted empirical measure with (possibly negative) weights $a_i w_i$, so that the parameter of interest becomes $\beta_{a,w} = \sum_{i=1}^n a_i w_i \tau(z_i)$.

For a signed measure μ on (w, z) , let μ_z denote its marginal on z , $\mu_z(A) = \mu(\{(w, z) : z \in A\})$. Since $\tau(\cdot)$ depends only on z , the estimand depends on μ only through μ_z .

Theorem A.1. *Let μ be a signed measure with $\int d\mu(w, z) = 1$, and let $\hat{\beta} = a'Y$ be an estimator with $a'w = 1$. If $\mu_z = (\mu_{a,w}^*)_z$, so that $\int \tau(z) d\mu(w, z) = \beta_{a,w}$, then $\widetilde{\text{bias}}_\Gamma(\hat{\beta}; \mu) =$*

$\overline{\text{bias}}_\Gamma(\hat{\beta})$, with $\overline{\text{bias}}_\Gamma(\hat{\beta})$ given in eq. (4). If $\mu_z \neq (\mu_{a,w}^*)_z$, then $\widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu) = \infty$. Furthermore, the estimator $\hat{\beta}_\lambda$ given in eq. (8) solves

$$\min_{\hat{\beta}=a'Y, a \in \mathbb{R}^n} \text{var}(\hat{\beta}) \quad \text{s.t.} \quad \min_{\mu} \widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu) \leq C\overline{B}_\lambda, \quad (20)$$

where the minimization over μ is over all signed measures such that $\int d\mu(w, z) = 1$.

Proof. Since the bias depends on μ only through μ_z , it suffices to consider $\mu = \mu_{a,w}^*$. Note that $\widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu_{a,w}^*) = \sup_{\gamma \in \Gamma} a'Z\gamma$, which proves the first part of the theorem.

To prove the second part of the theorem, let $\nu = (\mu_{a,w}^*)_z - \mu_z$, so that the bias equals $\int \tau d\nu + a'Z\gamma$. If $\nu \neq 0$, there is a measurable set A with $\nu(A) \neq 0$. Setting $\gamma = 0 \in \Gamma$, and $\tau(z) = c$ for $z \in A$ and $\tau(z) = 0$ otherwise, gives bias $c\nu(A)$, which can be made arbitrarily large by taking c to have the same sign as $\nu(A)$ and $|c|$ large. Thus $\widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu) = \infty$ if $\mu_z \neq (\mu_{a,w}^*)_z$.

To prove the final assertion, recall that the weights a_λ minimize the variance of the linear estimator $\hat{\beta}$ subject to the bound $C\overline{B}_\lambda$ on worst-case bias $\overline{\text{bias}}_\Gamma(\hat{\beta})$ for β in the constant treatment effects model in eq. (1). It follows from the first two assertions that $\overline{\text{bias}}_\Gamma(\hat{\beta}) = \min_{\mu} \widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu) = \widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu_{a,w}^*)$, where the minimization is over all signed measures μ that integrate to one. Thus, under the heterogeneous TE model (19), the same weights a_λ solve eq. (20). $\square$

The theorem gives three results for inference when $\tau(\cdot)$ is unrestricted. First, given a linear estimator $a'Y$, the only estimand for which the bias is finite is $\beta_{a,w}$. Second, for this estimand, the bias is the same as in the homogeneous TE model with $\tau(z) = \beta$. Thus, assuming homogeneous TEs leads to valid inference for $\beta_{a,w}$ when TEs are in fact heterogeneous. In particular, the bias-aware CI based on the estimator $\hat{\beta}_\lambda$ provides inference on the weighted average $\beta_{a_\lambda, w} = \sum_{i=1}^n \tilde{w}_{\lambda, i} w_i \tau(z_i) / \sum_{i=1}^n \tilde{w}_{\lambda, i} w_i$. Observe that, when the treatment w_i is binary, the weights are nonnegative if and only if the residual in the propensity score regression $\tilde{w}_{\lambda, i}$ is nonnegative for all treated units—this can be easily verified in a given application. Equivalently, the weights are nonnegative if and only if $z_i' \pi_\lambda \leq 1$ for all treated units: the fitted values in the propensity score regression must respect the population constraint that treatment probabilities cannot exceed one. This finding is a finite-sample analog of the identification result in Goldsmith-Pinkham, Hull, and Kolesár (2024), who show that the estimand in the partially linear model has an analogous weighted ATE interpretation under heterogeneous TEs. An analogous identification result in a random design setting dates to at least Angrist (1998), who gave a weighted ATE interpretation to the OLS estimand.

Third, the estimator $\hat{\beta}_\lambda$ remains optimal in the heterogeneous TE model in that it solves

a bias-variance tradeoff in a problem where we can pick the estimand to make the estimation problem as easy as possible. The problem (20) is a finite-sample version of the “moving the goalposts” problem considered by Crump et al. (2006, Section 5.4).¹³ Crump et al. (2006) derived the measure μ that minimizes the asymptotic variance of a particular class of inverse propensity score weighted estimators of weighted ATEs in a random design setup. Goldsmith-Pinkham, Hull, and Kolesár (2024) show this measure in fact minimizes the variance among all regular estimators; the measure coincides with the weighting of the treatment effects given in Angrist (1998) under homoskedasticity, giving an optimality property to the OLS estimator.

A.3 Connection between double lasso and optimal estimator under ℓ_1 constraints

When $\text{Pen}(\gamma) = \|\gamma\|_1$ (example 2.2), the solution π_λ to eq. (7) is the lasso estimate in the propensity score regression of w on Z , and our estimator (8) uses residuals from this lasso regression. This is related to “double lasso” estimators used to form CIs for β under sparsity constraints on γ (see, among others, Belloni, Chernozhukov, and Hansen 2014; Javanmard and Montanari 2014; van de Geer et al. 2014; C.-H. Zhang and S. S. Zhang 2014). For concreteness, we focus on the estimator in C.-H. Zhang and S. S. Zhang (2014), which is given by

$$\hat{\beta}_{ZZ} = \hat{\beta}_{\text{lasso}} + \frac{\tilde{w}'_\lambda(Y - w\hat{\beta}_{\text{lasso}} - Z\hat{\gamma}_{\text{lasso}})}{\tilde{w}'_\lambda w},$$

where $\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}}$ are the lasso estimates from regressing Y on X :

$$\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}} = \arg \min_{\beta, \gamma} \|Y - w\beta - Z\gamma\|_2^2 + \tilde{\lambda}(|\beta| + \|\gamma\|_1)$$

for some penalty parameter $\tilde{\lambda} > 0$.

Remark A.1. Note that $\hat{\beta}_{ZZ}$ is nonlinear in Y , due to nonlinearity of the lasso estimates $\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}}$, which is consistent with the goal of efficiency in the nonconvex parameter space (18). In contrast, Corollary 4.1 shows that under the convex parameter space $\Gamma = \{\gamma: \|\gamma\|_1 \leq C\}$, the estimator $\hat{\beta}_\lambda$ in (8) which only uses lasso in the propensity score regression of w on Z , is already highly efficient among all estimators, so that there is no further role for substantive efficiency gains from the lasso regression of Y on X , or from the

¹³The optimization problem (20) was also used by Imbens and Wager (2019) in the context of a regression discontinuity design with multiple cutoffs, although they did not explicitly note the optimality properties of the resulting estimator.

use of other nonlinear estimators.

To further understand the connection between these estimators, we note that C.-H. Zhang and S. S. Zhang (2014) motivate their approach by bounds of the form

$$\|\hat{\gamma}_{\text{lasso}} - \gamma\|_1 \leq \tilde{C} \quad \text{where} \quad \tilde{C} = \text{const.} \cdot s\sqrt{\log k}/\sqrt{n}, \quad (21)$$

which hold with high probability with the constant depending on certain “compatibility constants” that describe the regularity of the design matrix X (see Bühlmann and van de Geer 2011, Theorem 6.1, and references in the surrounding discussion). This suggests correcting the initial estimate $\hat{\beta}_{\text{lasso}}$ by estimating $\tilde{\beta} = \beta - \hat{\beta}_{\text{lasso}}$ in the regression

$$\tilde{Y} = w(\beta - \hat{\beta}_{\text{lasso}}) + Z(\gamma - \hat{\gamma}_{\text{lasso}}) + \varepsilon = w\tilde{\beta} + Z\tilde{\gamma} + \varepsilon,$$

where $\tilde{Y} = Y - w\hat{\beta}_{\text{lasso}} - Z\hat{\gamma}_{\text{lasso}}$. Heuristically, we can treat the bound in eq. (21) as a constraint $\|\tilde{\gamma}\|_1 \leq \tilde{C}$ on the unknown parameter $\tilde{\gamma} = \gamma - \hat{\gamma}_{\text{lasso}}$ and search for an optimal estimator of $\tilde{\beta} = \beta - \hat{\beta}_{\text{lasso}}$ under this constraint. Applying the optimal estimator derived in Theorem 2.1 then suggests estimating $\beta - \hat{\beta}_{\text{lasso}}$ with $\frac{\tilde{w}'_{\lambda}\tilde{Y}}{\tilde{w}'_{\lambda}w}$. Adding this estimate to $\hat{\beta}_{\text{lasso}}$ gives the estimate $\hat{\beta}_{\text{ZZ}}$ proposed by C.-H. Zhang and S. S. Zhang (2014). Whereas C.-H. Zhang and S. S. Zhang (2014) motivate their approach as one possible way of correcting the initial estimate $\hat{\beta}_{\text{lasso}}$ using the bound in eq. (21), the above analysis shows that their correction is in fact identical to an approach in which one optimizes this correction numerically.¹⁴

Under the bound in eq. (21) it follows that $\hat{\beta}_{\text{ZZ}} - \beta = \tilde{b} + a_{\lambda}'\varepsilon$ where $a_{\lambda} = \frac{\tilde{w}_{\lambda}}{\tilde{w}'_{\lambda}w}$ are the optimal weights under the ℓ_1 constraint $\|\tilde{\gamma}\|_1 \leq \tilde{C}$, given in Theorem 2.1. Furthermore, $|\tilde{b}| \leq \tilde{C}\bar{B}_{\lambda}$, with $\bar{B}_{\lambda}$ given in Theorem 2.1 and $\tilde{C}$ given in eq. (21), and the variance of the random term $a_{\lambda}'\varepsilon$ is given by sd_{λ}^2 in Theorem 2.1. Using arguments similar to those used to prove Theorem 4.2, it follows that $\tilde{C}\bar{B}_{\lambda}/\text{sd}_{\lambda}$ is bounded by a constant times $s(\log k)/\sqrt{n}$, so that one can ignore bias in large samples as long as this term converges to zero. This leads to the CI proposed by C.-H. Zhang and S. S. Zhang (2014), which takes the form

$$\{\hat{\beta}_{\text{ZZ}} \pm z_{1-\alpha/2}\hat{\text{se}}_{\lambda}\}, \quad (22)$$

where $\hat{\text{se}}_{\lambda}$ is an estimate of sd_{λ} .

Remark A.2. To avoid the assumption that $s(\log k)/\sqrt{n} \rightarrow 0$ one could, in principle,

¹⁴The estimator proposed by Javanmard and Montanari (2014) performs a numerical optimization of this form, but with the constraint (21) replaced by a constraint on $|\hat{\beta}_{\text{lasso}} - \beta| + \|\hat{\gamma}_{\text{lasso}} - \gamma\|_1$. Thus, Theorem 2.1 shows that a modification of the constraint used in Javanmard and Montanari (2014) yields the same estimator as C.-H. Zhang and S. S. Zhang (2014).

extend our approach and the above analysis to form valid bias-aware CIs as $\{\hat{\beta}_{ZZ} \pm [\tilde{C}\bar{B}_\lambda + z_{1-\alpha/2}\hat{s}\hat{e}_\lambda]\}$.¹⁵ Unfortunately, finding a computable constant $\tilde{C}$ in (21) that is sharp enough to yield useful bounds in practice appears to be difficult, although it is an interesting area for future research.

A.4 Verification of high-level conditions

The next result gives a set of low-level sufficient conditions under which $X \in \mathcal{E}_n^p(\eta)$ with high probability.

Lemma A.1. *Suppose w_i, z_i are drawn i.i.d. over i , and let $\delta = \arg \min_b E[(w_i - z'_i b)^2]$ so that $z'_i \delta$ is the population best linear predictor of w_i . Suppose that the linear prediction error $E[(w_i - z'_i \delta)^2]$ is bounded away from zero as $k \rightarrow \infty$, $E[w_i^2] < \infty$, and that $\sup_j E[|(w_i - z'_i \delta)z_{ij}|^{\max\{2, q\}}] < \infty$ when $p > 1$, and, for some $c > 0$, $P(|(w_i - z'_i \delta)z_{ij}| \geq t) \leq 2 \exp(-ct^2)$ for all j when $p = 1$. Then, for any $\tilde{\eta} > 0$, there exists η such that $X \in \mathcal{E}_n^p(\eta)$ with probability at least $1 - \tilde{\eta}$ for large enough n .*

Proof. By the orthogonality condition for the best linear predictor, we have $E[w_i v_i] = E[v_i^2]$, where $v_i = w_i - z'_i \delta$, which is bounded from below uniformly over k by assumption. Since $E[w_i v_i]$ is bounded from above by $E w_i^2 < \infty$, it follows from the law of large numbers for triangular arrays that $\frac{1}{n} \sum_{i=1}^n w_i v_i \geq \eta$ with probability approaching one once η is small enough. Similarly, $\frac{1}{n} \sum_{i=1}^n v_i^2 \leq 1/\eta$ for small enough η by the law of large numbers for triangular arrays.

For the last inequality in the definition of $\mathcal{E}_n^p(\eta)$, first consider the case $p > 1$ so that $q < \infty$. We then have $E\|\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i v_i\|_q^q = E \sum_{j=1}^k |\sum_{i=1}^n v_i z_{ij} / \sqrt{n}|^q \leq k \cdot K$ by von Bahr (1965), where K is a constant that depends only on an upper bound for $\max_j E[|v_i z_{ij}|^{\max\{q, 2\}}]$. Applying Markov's inequality gives the required bound. When $p = 1$, then $q = \infty$ so that

$$P\left(\left\|\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i v_i\right\|_q \geq \eta^{-1} \sqrt{\log k}\right) \leq \sum_{j=1}^k P\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n v_i z_{ij}\right| > \eta^{-1} \sqrt{\log k}\right),$$

which is bounded by $2k \exp(-K \cdot \eta^{-2} \log k) = 2k^{1-K\eta^{-2}}$ for some constant K by Hoeffding's inequality for sub-Gaussian random variables (Vershynin 2026, Theorem 2.7.3). This can be made arbitrarily small uniformly in k by making η small, as required. $\square$

¹⁵We use the slightly more conservative approach of adding and subtracting the bound $\tilde{C}\bar{B}_\lambda$ rather than using the critical value $\text{cv}_\alpha(\tilde{C}\bar{B}_\lambda/\hat{s}\hat{e}_\lambda)$ as in eq. (5), since the “bias” term for $\hat{\beta}_{ZZ}$ is correlated with ε through the first step estimates $\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}}$.

Lemma A.1 requires that w_i and z_i be drawn from a distribution such that the population prediction error from regressing w_i on z_i is bounded away from zero as $k \rightarrow \infty$. This condition will hold, for example, if $w_i = z_i' \delta + u_i$ with u_i a nondegenerate random variable that is independent of z_i . Note that we do not impose any conditions on the norm or sparsity level of δ in this regression.

Appendix B Proofs

This appendix gives proofs for all results in the main text.

B.1 Proof of Theorem 2.1

To prove Theorem 2.1, we first explain how our results fall into the general setup used in Donoho (1994), Low (1995) and Armstrong and Kolesár (2018). In the notation of Armstrong and Kolesár (2018), $(\beta, \gamma)'$ plays the role of the parameter f , the functional of interest is given by $L(\beta, \gamma) = \beta$ and $K(\beta, \gamma) = w\beta + Z\gamma$. The parameter space $\mathbb{R} \times \Gamma$ is centrosymmetric, so that the modulus of continuity (eq. (25) in Armstrong and Kolesár 2018) is given by

$$\omega(\delta) = \sup_{\beta, \gamma} 2\beta \quad \text{s.t.} \quad \|w\beta + Z\gamma\|_2 \leq \delta/2, \quad \text{Pen}(\gamma) \leq C.$$

Using the substitution $\pi = -\gamma/\beta$, we can write this as

$$\omega(\delta) = \sup_{\beta, \pi} 2\beta \quad \text{s.t.} \quad \beta \|w - Z\pi\|_2 \leq \delta/2, \quad \beta \text{Pen}(\pi) \leq C. \quad (23)$$

Let $\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}}$ and $\pi_\delta^{\text{mod}} = -\gamma_\delta^{\text{mod}}/\beta_\delta^{\text{mod}}$ denote a solution to this problem when it exists. In the notation of Armstrong and Kolesár (2018), $(\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}})'$ plays the role of g_δ^* , and the solution (f_δ^*, g_δ^*) satisfies $f_\delta^* = -g_\delta^* = -(\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}})'$ by centrosymmetry.

This optimization problem is clearly related to the problem in eq. (7): we want to make $\|w - Z\pi\|_2$ and $\text{Pen}(\pi)$ small so that large values of β satisfy the constraint in (23). The following lemma formalizes the connection.

Lemma B.1. *If there exists $\pi \in \mathcal{G}$ such that $w = Z\pi$ and $\text{Pen}(\pi) = 0$, then $\omega(\delta) = \infty$ for all $\delta \geq 0$. Otherwise, (i) for any $\delta > 0$, the modulus problem in eq. (23) has a solution $\beta_\delta^{\text{mod}}, \pi_\delta^{\text{mod}}$ with $\beta_\delta^{\text{mod}} > 0$. For $t = C/\beta_\delta^{\text{mod}} = 2C/\omega(\delta)$, this solution π_δ^{mod} is also a solution to the penalized regression (7) with optimized objective $\|w - Z\pi_\delta^{\text{mod}}\|_2 = \delta/\omega(\delta) > 0$; and (ii) for any $t = t_\lambda > 0$, the penalized regression problem (7) has a solution π_λ . Setting $\beta_\lambda = C/t_\lambda$ and*

$\delta_\lambda = 2\beta_\lambda \|w - Z\pi_\lambda\|_2 = (2C/t_\lambda) \|w - Z\pi_\lambda\|_2$, the pair $\beta_\lambda, \pi_\lambda$ solves the modulus problem (23) at $\delta = \delta_\lambda$, with optimized objective $\omega(\delta_\lambda) = 2C/t_\lambda$, so long as $\|w - Z\pi_\lambda\|_2 > 0$.

Proof. If there exists $\pi \in \mathcal{G}$ such that $w = Z\pi$ and $\text{Pen}(\pi) = 0$, then the result is immediate. Suppose there does not exist such a π .

First, we show that the problem in eq. (7) has a solution. Let $\mathcal{G}^{(0)}$ denote the linear subspace of vectors $\pi \in \mathcal{G}$ such that $Z\pi = 0$ and $\text{Pen}(\pi) = 0$, and let $\mathcal{G}^{(1)}$ be a subspace such that $\mathcal{G} = \mathcal{G}^{(0)} \oplus \mathcal{G}^{(1)}$, so that we can write $\pi \in \mathcal{G}$ uniquely as $\pi = \pi^{(0)} + \pi^{(1)}$ where $\pi^{(0)} \in \mathcal{G}^{(0)}$ and $\pi^{(1)} \in \mathcal{G}^{(1)}$. Note that $Z\pi = Z\pi^{(1)}$ and, applying the triangle inequality twice, $\text{Pen}(\pi^{(1)}) = \text{Pen}(\pi^{(1)}) - \text{Pen}(-\pi^{(0)}) \leq \text{Pen}(\pi) \leq \text{Pen}(\pi^{(0)}) + \text{Pen}(\pi^{(1)}) = \text{Pen}(\pi^{(1)})$ so that $\text{Pen}(\pi) = \text{Pen}(\pi^{(1)})$. Thus, the problem (7) can be written in terms of $\pi^{(1)} \in \mathcal{G}^{(1)}$ only. The level sets of this optimization problem are bounded and are closed by continuity of the seminorm $\text{Pen}(\cdot)$ (Goldberg 2017), and so it has a solution, which is also a solution in the original problem. Similarly, to show that the problem (23) has a solution, note that feasible values of β are bounded by a constant times the inverse of the minimum of $\max\{\|w - Z\pi\|_2, \text{Pen}(\pi)\}$ over π , which is strictly positive by continuity of $\text{Pen}(\pi)$ and the fact that there does not exist π with $\max\{\|w - Z\pi\|_2, \text{Pen}(\pi)\} = 0$. Thus, we can restrict $\beta, \pi^{(1)}$ to a compact set without changing the optimization problem.

To show the first statement in the lemma, note that $\beta_\delta^{\text{mod}} > 0$, since it is feasible to set $\pi = 0$ and $\beta = \delta/(2\|w\|_2)$, and that $\|w - Z\pi_\delta^{\text{mod}}\|_2 > 0$, since otherwise a strictly larger value of β could be achieved by multiplying π_δ^{mod} by $1 - \eta$ for $\eta > 0$ small enough. Now, if the first statement did not hold, there would exist a $\tilde{\pi}$ with $\text{Pen}(\tilde{\pi}) \leq C/\beta_\delta^{\text{mod}}$ such that $\|w - Z\tilde{\pi}\|_2 \leq \|w - Z\pi_\delta^{\text{mod}}\|_2 - \nu$ for small enough $\nu > 0$. Then, letting $\tilde{\pi}_\eta = (1 - \eta)\tilde{\pi}$, we would have $\|w - Z\tilde{\pi}_\eta\|_2 \leq \|w - Z\tilde{\pi}\|_2 + \eta\|Z\tilde{\pi}\|_2 \leq \|w - Z\pi_\delta^{\text{mod}}\|_2 - \nu + \eta\|Z\tilde{\pi}\|_2 \leq \delta/(2\beta_\delta^{\text{mod}}) - \nu + \eta\|Z\tilde{\pi}\|_2$. Thus, $\|w - Z\tilde{\pi}_\eta\|_2$ will be strictly less than $\delta/(2\beta_\delta^{\text{mod}})$ for small enough η and $\text{Pen}(\tilde{\pi}_\eta) \leq (1 - \eta)C/\beta_\delta^{\text{mod}} < C/\beta_\delta^{\text{mod}}$. This is a contradiction, since it would allow a strictly larger value of β by setting $\pi = \tilde{\pi}_\eta$.

The second statement follows immediately, since any pair $\tilde{\beta}, \tilde{\pi}$ satisfying the constraints in the modulus (23) for $\delta = \delta_\lambda$ with $\tilde{\beta} > \beta_\lambda$ would have to have $\|w - Z\tilde{\pi}\|_2 < \|w - Z\pi_\lambda\|_2$ while maintaining the constraint $\text{Pen}(\tilde{\pi}) \leq t_\lambda$. $\square$

We now prove Theorem 2.1. The class of bias-variance optimizing estimators, $\hat{L}_\delta$ in the notation of Armstrong and Kolesár (2018), is given by $\frac{(w\beta_\delta^{\text{mod}} + Z\gamma_\delta^{\text{mod}})'Y}{(w\beta_\delta^{\text{mod}} + Z\gamma_\delta^{\text{mod}})'w}$, where we use eq. (26) in Armstrong and Kolesár (2018) to compute the form of this estimator under centrosymmetry, and Lemma D.1 in Armstrong and Kolesár (2018) to calculate the derivative $\omega'(\delta)$, since the problem is translation invariant with ι given by the parameter $\beta = 1, \gamma = 0$. Given λ with $\|w - Z\pi_\lambda\|_2 > 0$, it follows from Lemma B.1 that, for δ_λ given in the lemma,

this estimator $\hat{L}_{\delta_\lambda}$ is equal to $\hat{\beta}_\lambda = a'_\lambda Y$ where $a_\lambda = \frac{w - Z\pi_\lambda}{(w - Z\pi_\lambda)'w}$, as defined in Theorem 2.1. The worst-case bias formula in Theorem 2.1 then follows from the fact that the maximum bias is attained at $\gamma = -\gamma_{\delta_\lambda}^{\text{mod}} = Ct_\lambda^{-1}\pi_\lambda$ by Lemma A.1 in Armstrong and Kolesár (2018) (or Lemma 4 in Donoho 1994).

B.2 Proof of Corollary 4.1

Part (i) of Corollary 4.1 follows from Low (1995). In particular, consider the one-dimensional submodel $\beta \in [-C/\text{Pen}(\pi_\lambda), C/\text{Pen}(\pi_\lambda)]$, $\gamma = -\pi_\lambda\beta$. Let $b_\lambda = (w - Z\pi_\lambda)/\|w - Z\pi_\lambda\|_2^2$, and let $B \in \mathbb{R}^{n \times (n-1)}$ be a matrix with orthonormal columns that is orthogonal to b_λ . Note that in this submodel, $B'Y = B'(w - Z\pi_\lambda)\beta + B'\varepsilon = B'\varepsilon$, which does not depend on the unknown parameter β , and is independent of $b'_\lambda Y$. Therefore, $b'_\lambda Y \sim \mathcal{N}(\beta, \|b_\lambda\|_2^2 \sigma^2)$ is a sufficient statistic in this submodel. By Theorem 1 in Low (1995), in this submodel, the estimator $\hat{\beta}_\lambda = a'_\lambda Y = \kappa b'_\lambda Y$, where $\kappa = \|w - Z\pi_\lambda\|_2^2 / (w - Z\pi_\lambda)'w$ minimizes $\sup_\beta \text{var}(\delta(Y))$ among all estimators $\delta(Y)$ with $\sup_\beta |E_\beta[\delta(Y)] - \beta| \leq (1 - \kappa)C/\text{Pen}(\pi_\lambda) = C\bar{B}_\lambda$, and, likewise, it minimizes $\sup_\beta |E_\beta[\delta(Y)] - \beta|$ among all estimators with $\sup_\beta \text{var}(\delta(Y)) \leq \kappa^2 \sigma^2 \|b_\lambda\|_2^2 = \text{sd}_\lambda^2$. Since the worst-case bias $\overline{\text{bias}}_\Gamma(\hat{\beta}_\lambda) = C\bar{B}_\lambda$ and standard deviation $\text{sd}(\hat{\beta}_\lambda) = \text{sd}_\lambda$ are the same in the full model by Theorem 2.1, the result follows.

Part (ii) of Corollary 4.1 is immediate from Donoho (1994). In particular, it holds with

$$\kappa_{\text{MSE}}^*(X, \sigma, \Gamma) = \frac{\sup_{\delta > 0} (\omega(\delta)/\delta)^2 \rho_N(\delta/2, \sigma)}{\sup_{\delta > 0} (\omega(\delta)/\delta)^2 \rho_A(\delta/2, \sigma)} \geq 0.8,$$

where $\omega(\delta)$ is defined in eq. (23), and ρ_A and ρ_N are the minimax risk among affine estimators, and among all estimators, respectively, in the bounded normal means problem $Y \sim \mathcal{N}(\theta, \sigma^2)$, $|\theta| \leq \tau$, defined in Donoho (1994), and the last inequality follows from eq. (4) in Donoho (1994).

Finally, Part (iii) of Corollary 4.1 follows from Corollary 3.3 in Armstrong and Kolesár (2018), with

$$\kappa_{\text{CI}}^*(X, \sigma, \Gamma, \alpha) = \frac{(1 - \alpha)E[\omega(2\sigma(z_{1-\alpha} - Z)) \mid Z \leq z_{1-\alpha}]}{2 \min_\delta \text{cv}_\alpha \left(\frac{\omega(\delta)}{2\sigma\omega'(\delta)} - \frac{\delta}{2\sigma} \right) \cdot \sigma\omega'(\delta)},$$

where $Z \sim \mathcal{N}(0, 1)$, $\omega(\delta)$ is given in eq. (23), and by Lemma D.1 in Armstrong and Kolesár, since the problem is translation invariant with ι given by the parameter $\beta = 1$, $\gamma = 0$, $\omega'(\delta) = \delta/[w'(w - Z\pi_\delta^{\text{mod}}) \cdot \omega(\delta)]$. The universal lower bound 0.717 when $\alpha = 0.05$ follows from Theorem 4.1 in Armstrong and Kolesár (2021b).

B.3 Proof of Theorem 4.1

Let $L_\lambda = \lambda R_w \kappa + \sqrt{1 + \frac{R_w^2}{1-R_w^2}(1-\lambda)^2}$. Since $b + z_{1-\alpha} \leq \text{cv}_\alpha(b) \leq b + z_{1-\alpha/2}$, it follows that the ratio $\text{cv}_\alpha(b_\lambda) \text{sd}_\lambda / \sigma_{\text{short}} L_\lambda$ is bounded above and below by positive constants. Hence, it suffices to show that $\min_{\lambda \in [0,1]} L_\lambda \asymp 1 + \min\{\kappa, 1/\sqrt{1-R_w^2}\}$.

Suppose first that $\kappa^2/(1+\kappa^2) > R_w^2$. Then the result follows from $\min_\lambda L_\lambda = L_0 = 1/\sqrt{1-R_w^2}$. Since $1 + 1/\sqrt{1-R_w^2} \leq 2/\sqrt{1-R_w^2} \leq 2\sqrt{1+\kappa^2} \leq 2(1+\kappa)$, it follows that $1/\sqrt{1-R_w^2} \asymp 1 + \min\{\kappa, 1/\sqrt{1-R_w^2}\}$, which yields the claim.

Consider now the case $\kappa^2/(1+\kappa^2) \leq R_w^2$. Then L_λ is minimized at $\lambda = 1 - \sqrt{\frac{\kappa^2(1-R_w^2)^2}{R_w^2 - \kappa^2 R_w^2(1-R_w^2)}}$, with $\min_\lambda L_\lambda = \ell(R_w)$, where $\ell(R_w) = R_w \kappa + \sqrt{1 - \kappa^2(1-R_w^2)}$. We claim that $\ell(R_w) \asymp \kappa + 1$. The upper bound follows from $\ell(R_w) \leq \kappa + 1$ and the lower bound from the fact that $\ell(R_w)$ is increasing in R_w so that

$$\ell(R_w) \geq \min_{R_w: \kappa^2/(1+\kappa^2) \leq R_w^2} \ell(R_w) = \ell(\kappa/\sqrt{1+\kappa^2}) = \sqrt{1+\kappa^2} \geq (1+\kappa)/\sqrt{2}.$$

Hence, $\min_\lambda L_\lambda \asymp 1 + \kappa$. Since $\kappa \leq R_w/\sqrt{1-R_w^2} \leq 1/\sqrt{1-R_w^2}$, this yields the claim.

B.4 Proof of Theorem 4.2

To prove that the claimed upper bound holds for $X \in \mathcal{E}_n^p(\eta)$, we first note that, since the CI based on $\hat{\beta}_{\lambda_{\text{CI}}^*}$ is shorter than the CI based on any linear estimator $a'Y$, it suffices to show that there exists a sequence of weight vectors a such that the worst-case bias and standard deviation are bounded by constants times $R_p^*(k, C, R_w^2)$. Consider first the weights $\tilde{a}_i = \frac{v_i}{\sum_{j=1}^n v_j w_j}$, where $v_i = w_i - z_i' \delta$, with δ given in the definition of $\mathcal{E}_n^p(\eta)$. The variance of the estimator $\tilde{a}'Y$ is $\frac{\sigma^2 \sum_{i=1}^n v_i^2}{(\sum_{i=1}^n v_i w_i)^2} \leq \sigma^2 \eta^{-3}/n \leq \eta^{-5}/n$. The worst-case bias is

$$\sup_{\gamma: \|\gamma\|_p \leq C} \tilde{a}'Z\gamma = C \|Z'\tilde{a}\|_q = n^{-1/2} C \frac{n^{-1/2} \|Z'(w - Z\delta)\|_q}{n^{-1} |w'(w - Z\delta)|} \leq C \frac{r_p(k)/\sqrt{n}}{\eta^2},$$

where the first equality follows by Hölder's inequality, and the last inequality follows by definition of $\mathcal{E}_n^p(\eta)$. Since $z_{1-\alpha} + b \leq \text{cv}_\alpha(b) \leq z_{1-\alpha/2} + b$, it follows that choosing these weights yields the convergence rate $(1 + Cr_p(k))/\sqrt{n}$.

Next, consider the long regression estimator, which has no bias and standard deviation given by

$$\frac{\sigma}{\|\tilde{w}\|_2} \asymp n^{-1/2} \frac{1}{\sqrt{1-R_w^2}} \asymp n^{-1/2} \left(1 + \frac{1}{\sqrt{1-R_w^2}} \right),$$

where $\tilde{w} = (I - ZZ^+)w$, and we use the fact that $\|\tilde{w}\|_2^2 = (1 - R_w^2)\|w\|_2^2 \asymp (1 - R_w^2)n$. Hence,

the claimed upper bound is achieved by running the long regression if $Cr_p(k) > 1/\sqrt{1 - R_w^2}$, and using the weights $\tilde{a}$ otherwise.

B.5 Proof of Theorem 4.3

Step 1: Closed-form bias and variance expressions Solving the minimization problem yields $\pi_\lambda = (Z'Z + \lambda I_k)^{-1}Z'w$, so that letting $H_\lambda = Z(Z'Z + \lambda I_k)^{-1}Z'$, we have

$$a_\lambda = \frac{(I_n - H_\lambda)w}{w'(I_n - H_\lambda)w},$$

and

$$\bar{B}_\lambda = \frac{w'(H_\lambda - H_\lambda^2)w}{w'(I_n - H_\lambda)w \cdot \|(Z'Z + \lambda I_k)^{-1}Z'w\|_2} = \frac{\lambda \|(Z'Z + \lambda I_k)^{-1}Z'w\|_2}{w'(I_n - H_\lambda)w}.$$

where the second equality uses the observation that

$$H_\lambda - H_\lambda^2 = H_\lambda - Z(Z'Z + \lambda I_k)^{-1}Z'Z(Z'Z + \lambda I)^{-1}Z' = \lambda Z(Z'Z + \lambda I_k)^{-2}Z'.$$

Let $Z = UDV'$ denote the singular value decomposition of Z , where D is an $r \times r$ matrix collecting its singular values, $r = \text{rank}(Z)$, U is $n \times r$, and V is $k \times r$. Then we may write

$$(Z'Z + \lambda I_k)^{-1} = V(D^2 + \lambda I_r)^{-1}V' + (I_k - VV')/\lambda,$$

so that $H_\lambda = U \text{diag}(d_j^2/(\lambda + d_j^2))U'$. Hence, $(I_n - H_\lambda)w = \tilde{w} + U \text{diag}(\lambda/(\lambda + d_j^2))U'w$, so that

$$\|a_\lambda\|_2 = \frac{\left(\|\tilde{w}\|_2^2 + \lambda^2 \sum_{j=1}^r \frac{(U'w)_j^2}{(\lambda + d_j^2)^2}\right)^{1/2}}{\|\tilde{w}\|_2^2 + \lambda \mathcal{D}_\lambda}, \quad \mathcal{D}_\lambda = \sum_{j=1}^r \frac{(U'w)_j^2}{\lambda + d_j^2}.$$

Similarly, using the fact that $\|(Z'Z + \lambda I_k)^{-1}Z'w\|_2^2 = \|V(D^2 + \lambda I_r)^{-1}DU'w\|_2^2 = \sum_{j=1}^r \frac{d_j^2 (U'w)_j^2}{(d_j^2 + \lambda)^2}$, we obtain

$$\bar{B}_\lambda = \frac{\lambda \|(Z'Z + \lambda I_k)^{-1}Z'w\|_2}{\|\tilde{w}\|_2^2 + \lambda \mathcal{D}_\lambda} = \frac{\lambda \left(\sum_{j=1}^r \frac{d_j^2 (U'w)_j^2}{(d_j^2 + \lambda)^2}\right)^{1/2}}{\|\tilde{w}\|_2^2 + \lambda \mathcal{D}_\lambda}$$

Combining the expressions for $\|a_\lambda\|_2$ and $\bar{B}_\lambda$ gives

$$\text{sd}_\lambda + C\bar{B}_\lambda \asymp f(\lambda), \quad f(\lambda) = \frac{\|\tilde{w}\|_2 + \lambda \left(\sum_{j=1}^r \frac{(1 + C^2 d_j^2)(U'w)_j^2}{(\lambda + d_j^2)^2}\right)^{1/2}}{\|\tilde{w}\|_2^2 + \lambda \mathcal{D}_\lambda}$$

Therefore, to prove the lower bound assertion, it suffices to show that $\min_\lambda f(\lambda)$ matches the claimed rates.

Step 2: lower bounds Observe that

$$\frac{R_w^2 n}{\lambda + d_{\max}^2} \lesssim \frac{R_w^2 \|w\|_2^2}{\lambda + d_{\max}^2} \leq \mathcal{D}_\lambda \leq \frac{R_w^2 \|w\|_2^2}{\lambda + d_{\min}^2} \lesssim \frac{R_w^2 n}{\lambda + d_{\min}^2},$$

Using these bounds, we obtain the lower bound

$$\begin{aligned} f(\lambda) &\geq \frac{\|\tilde{w}\|_2 + \lambda \left(\min_j \frac{1+C^2 d_j^2}{\lambda + d_j^2} \right)^{1/2} \mathcal{D}_\lambda^{1/2}}{\|\tilde{w}\|_2^2 + \lambda \mathcal{D}_\lambda} \gtrsim \frac{\|\tilde{w}\|_2 + \lambda \left(\min_j \frac{1+C^2 d_j^2}{\lambda + d_j^2} \right)^{1/2} \frac{R_w \sqrt{n}}{\sqrt{\lambda + d_{\max}^2}}}{\|\tilde{w}\|_2^2 + \frac{\lambda R_w^2 n}{\lambda + d_{\min}^2}} \\ &\asymp \frac{\|\tilde{w}\|_2 + \lambda \left(\min_j \frac{1+C^2 d_j^2}{\lambda + d_j^2} \right)^{1/2} \frac{R_w \sqrt{n}}{\sqrt{\lambda + d_{\max}^2}}}{\|\tilde{w}\|_2^2 + \frac{\lambda R_w^2 n}{\lambda + d_{\max}^2}} \end{aligned}$$

where the second line uses the fact that since $d_{\min}/d_{\max} \asymp 1$, $(\lambda + d_{\min}^2)/(\lambda + d_{\max}^2) \asymp 1$ uniformly over $\lambda \geq 0$.

Since the function $(1 + C^2 x)/(x + \lambda)$ is decreasing if $\lambda C^2 \leq 1$, it follows that if $\lambda C^2 \leq 1$, $\min_j \frac{1+C^2 d_j^2}{\lambda + d_j^2} = \frac{1+C^2 d_{\max}^2}{\lambda + d_{\max}^2}$, so that

$$f(\lambda) \gtrsim \frac{\|\tilde{w}\|_2(\lambda + d_{\max}^2) + \lambda(1 + C^2 d_{\max}^2)^{1/2} R_w \sqrt{n}}{\|\tilde{w}\|_2^2(\lambda + d_{\max}^2) + \lambda R_w^2 n}.$$

Since this is a ratio of two affine functions, it is monotone in λ , so that the minimum over $\lambda \in [0, 1/C^2]$ is taken at one of the extremes. Thus, letting $h(C) = \frac{R_w^2 n}{1+C^2 d_{\max}^2}$,

$$\begin{aligned} \min_{\lambda \in [0, 1/C^2]} f(\lambda) &\gtrsim \min \left\{ \frac{1}{\|\tilde{w}\|_2}, \frac{\|\tilde{w}\|_2 + h(C)^{1/2}}{\|\tilde{w}\|_2^2 + h(C)} \right\} \asymp \min \left\{ \frac{1}{\|\tilde{w}\|_2}, \frac{1}{\|\tilde{w}\|_2 + h(C)^{1/2}} \right\} \\ &= \frac{1}{\|\tilde{w}\|_2 + h(C)^{1/2}} \asymp \frac{1}{n^{1/2}} \frac{1 + C d_{\max}}{\sqrt{1 - R_w^2 C d_{\max}} + 1}, \quad (24) \end{aligned}$$

where we use $\sqrt{a+b} \asymp \sqrt{a} + \sqrt{b}$, and $\|\tilde{w}\|_2^2 = (1 - R_w^2) \|w\|_2^2 \asymp (1 - R_w^2)n$. Similarly, if $\lambda C^2 \geq 1$, $\min_j \frac{1+C^2 d_j^2}{\lambda + d_j^2} = \frac{1+C^2 d_{\min}^2}{\lambda + d_{\min}^2}$, so that

$$f(\lambda) \gtrsim \frac{\|\tilde{w}\|_2 + \lambda \left(\frac{1+C^2 d_{\min}^2}{\lambda + d_{\min}^2} \right)^{1/2} \frac{R_w \sqrt{n}}{\sqrt{\lambda + d_{\max}^2}}}{\|\tilde{w}\|_2^2 + \frac{\lambda R_w^2 n}{\lambda + d_{\max}^2}} \asymp \frac{\|\tilde{w}\|_2(\lambda + d_{\max}^2) + \lambda(1 + C^2 d_{\min}^2)^{1/2} R_w \sqrt{n}}{\|\tilde{w}\|_2^2(\lambda + d_{\max}^2) + \lambda R_w^2 n},$$

which is again a ratio of affine functions, so that the minimum over $\lambda \in [1/C^2, \infty]$ is taken at one of the extremes. Using the same arguments as in eq. (24), we therefore obtain

$$\begin{aligned} \min_{\lambda \in [1/C^2, \infty]} f(\lambda) &\gtrsim \min \left\{ \frac{1}{n^{1/2}} \frac{1 + Cd_{\min}}{\sqrt{1 - R_w^2} Cd_{\max} + 1}, \frac{\|\tilde{w}\|_2 + (1 + C^2 d_{\min}^2)^{1/2} R_w \sqrt{n}}{\|\tilde{w}\|_2^2 + R_w^2 n} \right\} \\ &\asymp \min \left\{ \frac{1}{n^{1/2}} \frac{1 + Cd_{\min}}{\sqrt{1 - R_w^2} Cd_{\max} + 1}, \frac{1 + Cd_{\min} R_w}{\sqrt{n}} \right\}. \end{aligned}$$

Since this minimum is smaller than the right-hand side of eq. (24), it follows that

$$\begin{aligned} \min_{\lambda \geq 0} f(\lambda) &\gtrsim \frac{1}{\sqrt{n}} \min \left\{ \frac{1 + Cd_{\min}}{\sqrt{1 - R_w^2} Cd_{\max} + 1}, 1 + Cd_{\min} R_w \right\} \\ &\asymp \frac{1}{\sqrt{n}} \min \left\{ \frac{1 + Cd_{\max}}{\sqrt{1 - R_w^2} Cd_{\max} + 1}, 1 + Cd_{\max} R_w \right\}. \end{aligned}$$

Observe that for $a \in [0, 1]$ and $x \geq 0$, $\frac{1+x}{1+ax} \geq \frac{1+\min\{x, 1/a\}}{4}$. Indeed, if $ax \leq 1$, then $\frac{1+x}{1+ax} \geq \frac{1+x}{2} = \frac{1+\min\{x, 1/a\}}{2}$, while if $ax > 1$, $\frac{1+x}{1+ax} \geq \frac{1+x}{2ax} \geq \frac{1}{2a} \geq \frac{1}{4} \left(1 + \frac{1}{a}\right) = \frac{1}{4} (1 + \min\{x, 1/a\})$. Applying this inequality to the first minimand in the preceding display then yields that it is bounded below by $\frac{1+\min\{Cd_{\max}, 1/\sqrt{1-R_w^2}\}}{4}$. The result then follows from observing that $1 + \min\left\{\frac{1}{\sqrt{1-R_w^2}}, R_w Cd_{\max}\right\} \asymp 1 + \min\left\{\frac{1}{\sqrt{1-R_w^2}}, Cd_{\max}\right\}$.

B.6 Proof of Theorem 4.4

Since the lower bound $c \cdot n^{-1/2}$ follows from standard efficiency bounds with finite dimensional parameters (e.g., taking the submodel where $\delta = \gamma = 0$), it suffices to show that $E_{\vartheta^*} \hat{\chi} \geq c \cdot C \sqrt{\log k} / \sqrt{n}$ for some $c > 0$ once n is large enough. To show this, we follow essentially the same arguments as Cai and Guo (2017, Theorem 3) and Javanmard and Montanari (2018, Proposition 4.2), checking that the distributions used in the lower bound satisfy the constraints that define $\Theta(C, C \cdot K \sqrt{n/\log k}, \eta)$. Under a given parameter vector $\vartheta = (\beta, \gamma', \delta', \sigma^2, \sigma_v^2)$, the data $(Y_i, w_i, z_i)'$ are i.i.d. normal with mean zero and variance matrix

$$\Sigma_{\vartheta} = \begin{pmatrix} \sigma^2 + \beta^2(\sigma_v^2 + \|\delta\|_2^2) + 2\beta\delta'\gamma + \|\gamma\|_2^2 & \beta(\sigma_v^2 + \|\delta\|_2^2) + \gamma'\delta & \beta\delta' + \gamma' \\ \beta(\sigma_v^2 + \|\delta\|_2^2) + \gamma'\delta & \sigma_v^2 + \|\delta\|_2^2 & \delta' \\ \beta\delta + \gamma & \delta & I_k \end{pmatrix}.$$

Let f_{π} denote the distribution of the data $\{Y_i, w_i, z_i\}_{i=1}^n$ when the parameters follow a prior distribution π , and let $\chi^2(f_{\pi_0}, f_{\pi_1})$ denote the chi-square distance between these distributions

for prior distributions π_0 and π_1 . By Lemma 1 in Cai and Guo (2017), it suffices to find a prior distribution π_1 supported on the parameter space $\Theta(C, C \cdot K \sqrt{n/\log k}, \eta)$ such that π_1 places probability one on $\beta = \beta_{1,n}$ for some sequence with $|\beta_{1,n}|$ bounded from below by a constant times $C\sqrt{\log k}/\sqrt{n}$, and such that $\chi^2(f_{\pi_0}, f_{\pi_1}) \rightarrow 0$, where π_0 is the distribution that places probability one on ϑ^* .

We first note that we can assume $\sigma_0^2 = \sigma_{v,0}^2 = 1$ without loss of generality. Dividing Y_i by σ_0 and w_i by $\sigma_{v,0}$ leads to the same model with β , γ , δ , σ^2 and σ_v^2 multiplied by constants that depend only on σ_0 and $\sigma_{v,0}$.

We may also assume $\tilde{\eta} < 1/2$, since decreasing $\tilde{\eta}$ only weakens the condition $C \leq \sqrt{k/n} \cdot k^{-\tilde{\eta}}$. Fix a constant $c_\nu \in (0, \min\{2\tilde{\eta}, c_0^2\})$, and set $\nu = \frac{\sqrt{c_\nu}}{2} \cdot \frac{\sqrt{\log k}}{\sqrt{n}}$ and $s = \lfloor \frac{C}{2\nu} \rfloor$. Let π_1 be defined by a uniform prior for δ over the set of vectors with $\|\delta\|_0 = s$ and each element $\delta_j \in \{0, \nu\}$. We set the remaining parameters as deterministic functions of δ :

$$\beta_{1,n} = -\frac{\|\delta\|_2^2}{1 - \|\delta\|_2^2}, \quad \gamma = (1 - \beta_{1,n})\delta = \frac{\delta}{1 - \|\delta\|_2^2}, \quad \sigma_v^2 = 1 - \|\delta\|_2^2, \quad \sigma^2 = \frac{1 - 2\|\delta\|_2^2}{1 - \|\delta\|_2^2}.$$

Note that $\|\delta\|_2^2 = s\nu^2$ is constant under this prior, so the prior for $\beta_{1,n}$ is a unit point mass as required. Also, by the lower bound on C , and $c_\nu \leq c_0^2$, $C/(2\nu) \geq 1$, so that $s \geq 1$. Furthermore, $\|\delta\|_2^2 = s\nu^2 \leq C\nu/2$, with $C\nu = \frac{\sqrt{c_\nu}}{2} \cdot C\sqrt{\frac{\log k}{n}} \rightarrow 0$, which implies that $\|\delta\|_2^2 < 1/2$ for n large enough, so that the variance parameters are positive. We now verify that π_1 is supported on $\Theta(C, C \cdot K \sqrt{n/\log k}, \eta)$ for n large enough. Since $s \leq C/(2\nu)$, $\|\delta\|_1 = s\nu \leq C/2$ and, once $\|\delta\|_2^2 \leq 1/2$, $\|\gamma\|_1 = s\nu/(1 - \|\delta\|_2^2) \leq C$. Also, $\|\gamma\|_0 = \|\delta\|_0 = s \leq C/(2\nu) = c_\nu^{-1/2} C \sqrt{n/\log k} \leq C \cdot K \sqrt{n/\log k}$ once $K \geq c_\nu^{-1/2}$. Finally, $|\sigma_v^2 - 1| = \|\delta\|_2^2$ and $|\sigma^2 - 1| = \|\delta\|_2^2/(1 - \|\delta\|_2^2) \leq C\nu$, which is at most η for n large provided $\eta \rightarrow 0$ more slowly than $C\sqrt{(\log k)/n}$.

For ϑ in the support of π_1 , these choices lead to the variance matrix

$$\Sigma_\vartheta = \begin{pmatrix} 1 & 0 & \delta' \\ 0 & 1 & \delta' \\ \delta & \delta & I_k \end{pmatrix},$$

while $\Sigma_{\vartheta^*} = I_{k+2}$ under the point mass π_0 . It now follows from eqs. (118) and (119) in Javanmard and Montanari (2018) (which are applications of Lemmas 2 and 3 in Cai and Guo (2017)) that

$$\chi^2(f_{\pi_0}, f_{\pi_1}) \leq e^{\frac{s^2}{k-s}} \left(1 + \frac{s}{k}(e^{4n\nu^2} - 1)\right)^s - 1 \leq \exp\left(\frac{s^2}{k-s} + \frac{s^2 k^{c_\nu}}{k}\right) - 1,$$

where the second inequality uses $e^{4n\nu^2} = k^{c\nu}$ and $1 + x \leq e^x$. Let $\psi = 1/2 - \tilde{\eta} \in (0, 1/2)$. Since $s \leq k^{1/2-\tilde{\eta}}/\sqrt{c\nu \log(k)} \leq k^\psi$ once $\log(k) \geq 1/c\nu$, it follows that $\frac{s^2}{k-s} + \frac{s^2 k^{c\nu}}{k} \leq \frac{k^{2\psi-1}}{1-k^{\psi-1}} + k^{2\psi+c\nu-1} \rightarrow 0$, so that the right-hand side of the preceding display converges to zero.

Finally, we note that under π_1 , $|\beta_{1,n}| = \|\delta\|_2^2/(1 - \|\delta\|_2^2) \geq \|\delta\|_2^2 \geq C\nu/4$, where the last inequality uses the fact that since $\lfloor x \rfloor \geq x/2$ for $x \geq 1$, $s\nu^2 \geq C\nu/4$. Hence, $|\beta_{1,n}| \geq C\sqrt{c\nu}/8 \cdot \sqrt{\log(k)/n}$, which yields the result.

Appendix C Additional results

This appendix presents additional results that are useful for implementing Algorithm 3.1, and for assessing the plausibility of the assumption $\text{Pen}(\gamma) \leq C$. Appendix C.1 derives the properties of a regularized estimator of the regression function $w_i\beta + z'_i\gamma$. Appendix C.2 gives conditions under which this estimator can be used to construct initial estimates of residuals in Algorithm 3.1. Appendix C.3 presents a lower CI for C that can be used to assess the plausibility of the assumption $\text{Pen}(\gamma) \leq C$.

In this appendix, we focus primarily on the ℓ_p penalty $\text{Pen}(\gamma) = \|\gamma_2\|_p$, with $k_2 \rightarrow \infty$ and $k_1/n \rightarrow 0$. To state the results concisely, we use the notation $\theta = (\beta, \gamma')'$ and let $\Theta = \mathbb{R} \times \Gamma$ denote its parameter space. Let $X = (X_1, X_2)$, where $X_1 = (w, Z_1)$, and $X_2 = Z_2$. We partition θ accordingly, with $\theta_1 = (\beta, \gamma'_1)'$, and $\theta_2 = \gamma_2$. Let $H_{X_1} = X_1 X_1^+$ and $M_{X_1} = I - H_{X_1}$ denote projections onto the column space of X_1 and its orthogonal complement, where X_1^+ denotes the pseudo-inverse (so that $X_1^+ = (X_1' X_1)^{-1} X_1'$ if X_1 is full rank).

We allow the distribution Q of ε to be unknown and possibly non-Gaussian, and only maintain the assumption that ε_i is mean zero and independent across i . The class of possible distributions for Q is denoted by $\mathcal{Q}_n$. We use $P_{\theta,Q}$ and $E_{\theta,Q}$ to denote probability and expectation when Y is drawn according to $Q \in \mathcal{Q}_n$ and $\theta \in \Theta$, and we use the notation P_Q and E_Q for expressions that depend on Q only and not on θ .

We use the following assumption repeatedly throughout this appendix.

Assumption C.1. *There exists $\eta > 0$ such that, for all i and n and all $Q \in \mathcal{Q}_n$,*

$$P_Q(|\varepsilon_i| > t) \leq 2 \exp(-\eta t^2) \quad \text{when } p = 1, \quad E_Q[|\varepsilon_i|^{\max\{2+\eta, q\}}] < 1/\eta \quad \text{when } p > 1$$

and $\eta < E_Q \varepsilon_i^2$. In addition, the elements of $M_{X_1} X_2$ are bounded by some constant K_X uniformly over n .

C.1 Estimating the regression function globally

Consider the regularized regression estimator of θ , given by

$$\hat{\theta} = \arg \min_{\vartheta} \|Y - X\vartheta\|_2^2/n + \lambda \|\vartheta_2\|_p. \quad (25)$$

In order to derive the rate of convergence of $\hat{\theta}$ in Theorem C.1 below, we first give an elementary property of this estimator, following standard arguments (e.g. Bühlmann and van de Geer 2011, Section 6.2).

Lemma C.1. *If $\|2X_2' M_{X_1} \varepsilon\|_q/n \leq \lambda_0$, then $\|M_{X_1} X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n + (\lambda - \lambda_0)\|\hat{\theta}_2\|_p \leq (\lambda + \lambda_0)\|\theta_2\|_p$.*

Proof. Write the objective function as

$$\|H_{X_1}(Y - X_2\vartheta_2) - X_1\vartheta_1\|_2^2/n + \|M_{X_1}Y - M_{X_1}X_2\vartheta_2\|_2^2/n + \lambda\|\vartheta_2\|_p.$$

The first summand can be set to zero for any ϑ_2 by taking $\vartheta_1 = X_1^+(Y - X_2\vartheta_2)$. Therefore,

$$\hat{\theta}_2 = \arg \min_{\vartheta} \|M_{X_1}Y - M_{X_1}X_2\vartheta_2\|_2^2/n + \lambda\|\vartheta_2\|_p,$$

with $\hat{\theta}_1 = X_1^+(Y - X_2\hat{\theta}_2)$. This implies $H_{X_1}\varepsilon = H_{X_1}Y - H_{X_1}X\theta = H_{X_1}X(\hat{\theta} - \theta)$, so that

$$\|X(\hat{\theta} - \theta)\|_2^2/n = \|H_{X_1}\varepsilon\|_2^2/n + \|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n, \quad (26)$$

Using the fact that $\hat{\theta}_2$ attains a lower value of the objective than the true parameter value θ_2 , we obtain an ℓ_p version of what in the ℓ_1 case Bühlmann and van de Geer (2011, Lemma 6.1) term “the Basic Inequality”,

$$\|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n + \lambda\|\hat{\theta}_2\|_p \leq 2\varepsilon' M_{X_1}X_2(\hat{\theta}_2 - \theta_2)/n + \lambda\|\theta_2\|_p.$$

By Hölder’s inequality, $2\varepsilon' M_{X_1}X_2(\hat{\theta}_2 - \theta_2) \leq \|2X_2' M_{X_1} \varepsilon\|_q \|\hat{\theta}_2 - \theta_2\|_p$ so that, on the event $\|2X_2' M_{X_1} \varepsilon\|_q/n \leq \lambda_0$, we have

$$\|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n + \lambda\|\hat{\theta}_2\|_p \leq \lambda_0\|\hat{\theta}_2 - \theta_2\|_p + \lambda\|\theta_2\|_p \leq \lambda_0\|\hat{\theta}_2\|_p + (\lambda + \lambda_0)\|\theta_2\|_p,$$

which implies the result. □

We now use Lemma C.1 to derive rates of convergence for the regularized regression estimator in eq. (25) for estimating the regression function in ℓ_2 loss. For simplicity, we

use a fixed sequence for the penalty parameter λ satisfying certain rate conditions. This yields simple sufficient conditions that allow $\hat{\theta}$ to be used for auxiliary purposes such as standard error construction. In practice, data-driven methods such as cross-validation may be appealing. We discuss another possible choice based on moderate deviations bounds in Remark C.2 in Appendix C.3 below. We leave the analysis of $\hat{\theta}$ under such choices of λ for future research.

Theorem C.1. *Suppose that Assumption C.1 holds. Let $\hat{\theta}$ be the penalized regression estimator defined in eq. (25) with $\lambda = K_n r_p(k_2)/\sqrt{n}$, where $K_n \rightarrow \infty$ and $r_p(k)$ given in eq. (17). Then*

$$\sup_{\theta \in \mathbb{R}^{k+1}} \sup_{Q \in \mathcal{Q}_n} P_{\theta, Q} \left(\|X(\hat{\theta} - \theta)\|_2^2/n > K_n((k_1 + 1)/n + 2\|\theta_2\|_p r_p(k_2)/\sqrt{n}) \right) \rightarrow 0,$$

Proof. By Lemma C.2 below, if we set $\lambda_0 = \lambda = K_n r_p(k_2)/\sqrt{n}$, the condition of Lemma C.1, and hence the conclusion that $\|M_{X_1} X_2(\hat{\theta} - \theta)\|_2^2/n \leq 2K_n \|\theta_2\|_p r_p(k_2)/\sqrt{n}$, holds with probability approaching one uniformly over $\theta \in \mathbb{R}^{k+1}$ and $Q \in \mathcal{Q}_n$. In addition, since H_{X_1} is idempotent with rank at most $k_1 + 1$ and $E_Q \varepsilon \varepsilon'$ is diagonal with elements bounded uniformly over $Q \in \mathcal{Q}_n$, we have $E_Q \|H_{X_1} \varepsilon\|_2^2/n \leq \tilde{K}(k_1 + 1)/n$ for some constant $\tilde{K}$. The result follows by Markov's inequality and eq. (26). $\square$

Lemma C.2. *Under Assumption C.1, $\inf_{Q \in \mathcal{Q}_n} P_Q(\|2X_2' M_{X_1} \varepsilon\|_q/n \leq K_n r_p(k_2)/\sqrt{n}) \rightarrow 1$.*

Proof. Let $\tilde{x}_{ij} = (2M_{X_1} X_2)_{ij}$. For $q < \infty$, we have

$$E_Q \|2X_2' M_{X_1} \varepsilon\|_q^q = E_Q \sum_{j=1}^{k_2} \left| \sum_{i=1}^n \tilde{x}_{ij} \varepsilon_i \right|^q \leq k_2 \cdot K \cdot n^{q/2}$$

for some constant K that depends only on η , q and K_X , by von Bahr (1965). The result then follows by Markov's inequality. For $q = \infty$, we have

$$P_Q \left(\|2X_2' M_{X_1} \varepsilon\|_q/n > K_n \sqrt{\log k_2}/\sqrt{n} \right) = P_Q \left(\max_j \left| \sum_{i=1}^n \tilde{x}_{ij} \varepsilon_i \right|/n > K_n \sqrt{\log k_2}/\sqrt{n} \right),$$

which, for some $\tilde{K} > 0$, is bounded by $2k_2 \exp(-\tilde{K} \cdot K_n^2 \log k_2) = 2k_2^{1-\tilde{K} \cdot K_n^2} \rightarrow 0$ by Hoeffding's inequality for sub-Gaussian random variables (Vershynin 2026, Theorem 2.7.3). $\square$

C.2 Feasible CIs with an unknown error distribution

This appendix presents formal results for feasible CIs when the error distribution is unknown. Appendix C.2.1 presents general results for feasible CIs for linear estimators in our setting.

Appendix C.2.2 specializes these results to the feasible CIs in Section 3, with some technical modifications.

C.2.1 General results

We consider standard errors for linear estimators $\hat{\beta}_a = a'Y$, deviating slightly from the notation in the main text by making the dependence on the weights explicit with the subscript a . As in the main text, the weights a are nonrandom: they can depend on X but not on Y . We consider asymptotics where the weights a are allowed to depend on n so that $a_1, \dots, a_n$ is a triangular array rather than a sequence, but we leave this implicit in the notation.

Let $\hat{\theta}$ be an estimate of θ , and let $\hat{\varepsilon} = Y - X\hat{\theta}$. Consider the estimator $\hat{V}_a = \sum_{i=1}^n a_i^2 \hat{\varepsilon}_i^2$ of $V_Q = \text{var}_Q(\hat{\beta}_a) = \sum_{i=1}^n a_i^2 E_Q \varepsilon_i^2$. We consider coverage of the feasible bias-aware CI

$$\hat{\beta}_a \pm \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}_a)/\hat{V}_a^{1/2}) \cdot \hat{V}_a^{1/2}, \quad (27)$$

where $\overline{\text{bias}}_\Gamma(\hat{\beta}_a)$ is the worst-case bias, given in eq. (4). Under non-Gaussian errors, valid coverage will require the maximal Lindeberg weight defined in eq. (14) to converge to zero to ensure asymptotic normality of the estimator. The maximal Lindeberg weight turns out to also be relevant for controlling the contribution of estimation error in $\hat{\theta}$ in the variance estimate $\hat{V}_a$. In particular, in the following theorem, there is a tradeoff between the rate at which $\text{Lind}(a) \rightarrow 0$ and the ℓ_2 rate of convergence of the estimator $X\hat{\theta}$ of the regression function.

Theorem C.2. *Suppose that, for some $\eta > 0$, $\eta \leq E_Q \varepsilon_i^2$ and $E_Q |\varepsilon_i|^{2+\eta} \leq 1/\eta$ for all i and all $Q \in \mathcal{Q}_n$. Suppose also that, for some sequence c_n with $c_n = \mathcal{O}(\sqrt{n})$, we have*

- (i) $\max\{\sqrt{n}c_n, 1\} \cdot \text{Lind}(a) \rightarrow 0$; and
- (ii) $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_{\theta, Q}(\|X(\hat{\theta} - \theta)\|_2 \leq c_n) \rightarrow 1$.

Then, for any $\delta > 0$, $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q(|(\hat{V}_a - V_Q)/V_Q| < \delta) \rightarrow 1$. Furthermore,

$$\liminf_n \inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q\left(\beta \in \left\{\hat{\beta}_a \pm \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}_a)/\sqrt{\hat{V}_a}) \cdot \sqrt{\hat{V}_a}\right\}\right) \geq 1 - \alpha. \quad (28)$$

Proof. We have

$$\frac{\hat{V}_a - V_Q}{V_Q} = \frac{\sum_{i=1}^n a_i^2 (\hat{\varepsilon}_i^2 - \varepsilon_i^2)}{V_Q} + \frac{\sum_{i=1}^n a_i^2 (\varepsilon_i^2 - E_Q \varepsilon_i^2)}{V_Q}. \quad (29)$$

Let $\tilde{b}_i = a_i^2 / \sum_{j=1}^n a_j^2$ so that $\max_{1 \leq i \leq n} \tilde{b}_i = \text{Lind}(a)$. The second term in eq. (29) is bounded by $|\sum_{i=1}^n \tilde{b}_i (\varepsilon_i^2 - E_Q \varepsilon_i^2)| / \eta$. The absolute $1 + \eta/2$ moment of this quantity is bounded by a constant times $\sum_{i=1}^n \tilde{b}_i^{1+\eta/2} \cdot 1/\eta^{1+\eta/2}$ by von Bahr and Esseen (1965). This is bounded by $\max_{1 \leq i \leq n} \tilde{b}_i^{\eta/2} \cdot \sum_{i=1}^n \tilde{b}_i / \eta^{1+\eta/2} = \max_{1 \leq i \leq n} \tilde{b}_i^{\eta/2} / \eta^{1+\eta/2} = \text{Lind}(a)^{\eta/2} / \eta^{1+\eta/2} \rightarrow 0$. The first term in eq. (29) is bounded by $\text{Lind}(a)/\eta$ times

$$\sum_{i=1}^n |\hat{\varepsilon}_i^2 - \varepsilon_i^2| = \sum_{i=1}^n |\hat{\varepsilon}_i + \varepsilon_i| \cdot |\hat{\varepsilon}_i - \varepsilon_i| \leq \|\hat{\varepsilon} + \varepsilon\|_2 \|\hat{\varepsilon} - \varepsilon\|_2 \leq (\|\hat{\varepsilon} - \varepsilon\|_2 + 2\|\varepsilon\|_2) \|\hat{\varepsilon} - \varepsilon\|_2.$$

For some constant K that depends only on η , we have $2\|\varepsilon\|_2 \leq K\sqrt{n}$ with probability approaching one uniformly over $Q \in \mathcal{Q}_n$. Since $\|\hat{\varepsilon} - \varepsilon\|_2 = \|X(\hat{\theta} - \theta)\|_2 \leq c_n$ it follows that, with probability approaching one uniformly over $\theta \in \Theta$ and $Q \in \mathcal{Q}_n$, the first term in eq. (29) is bounded by $\text{Lind}(a) \cdot (K\sqrt{n} + c_n) \cdot c_n \rightarrow 0$. It follows that for any $\delta > 0$, $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q \left(\left| (\hat{V}_a - V_Q)/V_Q \right| < \delta \right) \rightarrow 1$. Coverage of the CI then follows from Theorem F.1 in Armstrong and Kolesár (2018), with the central limit theorem condition following by using the weights and moment bounds to verify the Lindeberg condition (see Lemma F.1 in Armstrong and Kolesár (2018)). $\square$

If $\hat{\theta}$ is based on the regularized regression in eq. (25), it follows from Theorem C.1 that (ii) in Theorem C.2 will hold with $c_n = \sqrt{K_n(k_1 + 1 + Cn^{1/2}r_p(k_2))}$ for a slowly increasing constant K_n . Hence, condition (i) in Theorem C.2 will then hold so long as

$$\sqrt{K_n n \cdot (k_1 + 1 + Cn^{1/2}r_p(k_2))} \cdot \text{Lind}(a) \rightarrow 0. \quad (30)$$

The following corollary records this result.

Corollary C.1. *Suppose that Assumption C.1 holds. Let $\hat{\varepsilon}$ be the residuals from the regularized regression in eq. (25), with λ given in Theorem C.1 for some $K_n \rightarrow \infty$, and suppose eq. (30) holds. Then eq. (28) holds with $\Theta = \mathbb{R}^{k_1+1} \times \{\gamma_2 : \|\gamma_2\|_p \leq C\}$.*

C.2.2 Optimized weights

We now apply the results in Appendix C.2.1 to the feasible CIs based on optimized weights in Algorithm 3.1. For technical reasons, we assume that the initial variance estimator in Step 1 of the algorithm is replaced by a nonrandom guess $\tilde{\sigma}^2$.¹⁶

It follows directly from Corollary C.1 that the feasible CI will be asymptotically valid provided that the maximal Lindeberg weight shrinks quickly enough relative to error in

¹⁶Alternatively, one could use sample-splitting or cross-fitting. In our Monte Carlo simulations, we find that the feasible CIs have good coverage without imposing these technical modifications.

the estimator used to form the residuals, so that $\text{Lind}(a_{\lambda_{\text{CI}}^*})$ satisfies eq. (30). This is equivalent to checking that $\text{Lind}(a_{\lambda_{\text{CI}}^*}) \leq b_n$ for some sequence of constants b_n such that $\sqrt{K_n n(k_1 + 1 + Cn^{1/2} r_p(k_2))} \cdot b_n \rightarrow 0$. Since the Lindeberg weight is observable, this condition can be checked directly in a given application. Since the theory only gives a rate condition on b_n , operationalizing the check requires one to take a stance on the constants. In our empirical work, we follow the proposal in Noack and Rothe (2024, Section 6.2) and use $b_n = 0.075$ as a simple benchmark for the maximal Lindeberg weight, leaving a more principled implementation to future work.

If the optimal weights violate the constraint $\text{Lind}(a_{\lambda_{\text{CI}}^*}) \leq b_n$, then the feasible CI based on the optimized weights may asymptotically undercover. To prevent this scenario, a principled approach is to constrain the bias-variance tradeoff in eq. (6), and only optimize it over a set of weights a that satisfy $\text{Lind}(a) \leq b_n$. Constraining the optimization problem to guarantee that the maximal Lindeberg weight of the solution is not too large is analogous to proposals by Noack and Rothe (2024), and Javanmard and Montanari (2014) in other settings. For our setting, given an initial guess $\tilde{\sigma}^2$ for the variance, the optimal bias-variance tradeoff is then given by minimizing

$$2\tilde{\sigma}\|a\|_2 \cdot \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}_a)/(\tilde{\sigma}\|a\|_2))$$

over weights a that satisfy $\text{Lind}(a) \leq b_n$. We denote the solution to this minimization problem by $a_{b_n}^*$. Imposing the Lindeberg weight condition can in general affect the performance of the CI. We show in Theorem C.3 below, however, that the optimal rate of convergence derived in Theorem 4.2 still obtains for a range of choices of b_n .

As a robustness check in our empirical work, we use a simpler approach that can be viewed as approximating the optimal weights $a_{b_n}^*$. In particular, if the constraint $\text{Lind}(a_{\lambda_{\text{CI}}^*}) \leq b_n$ is violated, we sequentially delete the observation with the largest Lindeberg weight until the constraint holds. This has the advantage that it still allows us to use existing computational algorithms discussed in Remark 3.3. Furthermore, as we discuss in Remark C.1 below, in practice the number of deletions is very small, suggesting that the efficiency cost relative to using the weights $a_{b_n}^*$ is minimal. This approach can also be viewed as a more principled version of the common empirical practice of running robustness checks using a sample that drops extreme values of X ; we tie the definition of “extreme” directly to the Lindeberg weight, rather than using an ad-hoc definition.

Remark C.1 (Lindeberg weight diagnostics in the empirical applications). In the lottery application, the full sample values for $\text{Lind}(a)$ are 0.149 for the short regression, 0.114 for AKK1, 0.349 for AKK2, and 0.376 for the long regression. For each method, we sequen-

tially remove the observation with the largest individual Lindeberg weight until $\text{Lind}(a)$ lies below the $b_n = 0.075$ benchmark. After each deletion, we re-estimate using the same procedure, including reoptimizing λ for AKK1 and AKK2. The short regression CI changes from $(-0.074, -0.031)$ to $(-0.086, -0.033)$ after three deletions; the AKK1 CI changes from $(-0.098, -0.022)$ to $(-0.108, -0.029)$ after one deletion; and the AKK2 CI changes from $(-0.110, 0.010)$ to $(-0.156, 0.035)$ after six deletions. While all CIs widen somewhat, the qualitative conclusions are largely unchanged.

By contrast, when applied to the long regression, the procedure yields, after three deletions, a CI that is too wide to be informative. Correspondingly, the C. M. Li and Müller (2021) procedure, which requires asymptotically normal short and long regression estimators, also yields uninformative inference. The fact that both the long regression and the C. M. Li and Müller (2021) procedure are sensitive to outliers provides an additional motivation for using our procedure.

In the RHC application, the corresponding values for the short regression, long regression, AKK1, and AKK2 are 0.0013, 0.0024, 0.0023, and 0.0024, respectively, all well below the benchmark $b_n = 0.075$.

We now show that the optimal rate of convergence derived in Theorem 4.2 is unaffected by imposing a Lindeberg weight constraint in the optimization.

Theorem C.3. *Suppose that Assumption C.1 holds. Let $\hat{\varepsilon}$ be the residuals from the regularized regression in eq. (25), with λ given in Theorem C.1 for some $K_n \rightarrow \infty$. Let b_n be a sequence satisfying $b_n = o(1/\sqrt{K_n n(1 + Cn^{1/2}r_p(k_2))})$. Suppose $k_1 = 0$, so that $\Gamma = \{\gamma: \|\gamma\|_p \leq C\}$. Suppose also that, for some $\eta > 0$, the design matrix X is in the set $\mathcal{E}_n^p(\eta)$ defined in Section 4.3 for large enough n , and that $\max_{1 \leq i \leq n} (w_i - z_i' \delta)^2/n = o(b_n)$ and $\frac{1}{n} \sum_{i=1}^n (w_i - z_i' \delta)^2$ is bounded away from zero where δ is given in the definition of $\mathcal{E}_n^p(\eta)$ in Section 4.3. Then there exists a constant K such that the CI in eq. (27) with $a = a_{b_n}^*$ satisfies*

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q \left(2 \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}_a)/\hat{V}_a^{1/2}) \cdot \hat{V}_a^{1/2} \geq K n^{-1/2} (1 + C r_p(k)) \right) = 0.$$

Proof. Let $v_i = w_i - z_i' \delta$ and $\tilde{a}_i = v_i / \sum_{j=1}^n v_j w_j$ where δ is given in the definition of $\mathcal{E}_n^p(\eta)$. Note that $\text{Lind}(\tilde{a}) = \max_{1 \leq i \leq n} (w_i - z_i' \delta)^2 / \sum_{j=1}^n (w_j - z_j' \delta)^2$. Under the assumptions of the theorem, this is bounded by a constant times $\max_{1 \leq i \leq n} (w_i - z_i' \delta)^2/n = o(b_n)$. Thus, the weights $\tilde{a}$ are feasible for the constrained optimization problem that defines $a_{b_n}^*$. It follows that

$$2\tilde{\sigma} \|a_{b_n}^*\|_2 \cdot \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}_{a_{b_n}^*})/(\tilde{\sigma} \|a_{b_n}^*\|_2)) \leq 2\tilde{\sigma} \|\tilde{a}\|_2 \cdot \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}_{\tilde{a}})/(\tilde{\sigma} \|\tilde{a}\|_2)).$$

It follows from the proof of Theorem 4.2 that the right-hand side of the above display is

bounded by a constant times $n^{-1/2}(1 + Cr_p(k))$. Furthermore, by the uniform consistency of $\hat{V}$ (which follows from Theorem C.2, where we note that the conditions of Theorem C.2 follow by the discussion leading up to Corollary C.1) and the fact that the variance of $\hat{\beta}_{a_{b_n}^*}$ is bounded from above uniformly over $\mathcal{Q}_n$, the width $2 \cdot \text{cv}_\alpha(\overline{\text{bias}}_\Gamma(\hat{\beta}_a)/\sqrt{\hat{V}}) \cdot \sqrt{\hat{V}}$ is bounded by a constant times the left-hand side of the above display with probability approaching one uniformly over $\theta \in \Theta$ and $Q \in \mathcal{Q}_n$. The result follows. $\square$

Theorem C.3 imposes both an upper bound on b_n (it must decrease at least as quickly as $1/\sqrt{K_n n(1 + Cn^{1/2}r_p(k_2))}$) as well as a lower bound on b_n (it must decrease more slowly than $\max_{1 \leq i \leq n}(w_i - z'_i \delta)^2/n$). To interpret the latter condition, note that $w_i - z'_i \delta$ plays the role of the residual in a best linear predictor regression of w_i on z_i in a random design setting. Thus, the condition $\max_{1 \leq i \leq n}(w_i - z'_i \delta)^2/n = o(b_n)$ is a tail condition on this best linear predictor error.

Depending on how quickly $\max_{1 \leq i \leq n}(w_i - z'_i \delta)^2$ increases, there will be a range of choices of b_n that satisfy both conditions. For example, if $\max_{1 \leq i \leq n}(w_i - z'_i \delta)^2$ is bounded, then the lower bound on b_n will hold if b_n is equal to $1/n$ times a slowly increasing sequence. Suppose, for simplicity, that we take $b_n = K_n/n$ where K_n is the same sequence in the upper bound. Then the upper bound $b_n = o(1/\sqrt{K_n n(1 + Cn^{1/2}r_p(k_2))})$ will then hold so long as

$$\sqrt{K_n n(1 + Cn^{1/2}r_p(k_2))} \cdot (K_n/n) = K_n \cdot \sqrt{K_n(1/n + Cn^{-1/2}r_p(k_2))} \rightarrow 0.$$

Since $Cn^{-1/2}r_p(k_2)$ is the rate at which the optimal CI shrinks, this condition is essentially the same as requiring that the optimal CI shrinks towards zero as $n \rightarrow \infty$.

C.3 Lower CIs for C

We present a lower CI for the penalty bound C , which can be used to assess the plausibility of the assumption $\text{Pen}(\gamma_2) \leq C$. Let $\hat{\theta}_2(\lambda)$ denote the regularized regression estimator of γ_2 , given in eq. (25), with penalty λ . Let λ_α^* denote an upper bound for the $1 - \alpha$ quantile of $\|2X_2' M_{X_1} \varepsilon\|_q/n$. Let

$$\underline{\hat{C}} = \sup_{\lambda > \lambda_\alpha^*} \frac{\lambda - \lambda_\alpha^*}{\lambda + \lambda_\alpha^*} \|\hat{\theta}_2(\lambda)\|_p. \quad (31)$$

In the idealized finite sample setting with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ with σ^2 known, λ_α^* can be computed exactly, so that $\underline{\hat{C}}$ is feasible.

Theorem C.4. *Consider $\underline{\hat{C}}$ in eq. (31) with λ_α^* given by the $1 - \alpha$ quantile of $\|2X_2' M_{X_1} \varepsilon\|_q/n$. Then, for any $\beta, \gamma_1, \gamma_2$ with $\|\gamma_2\|_p \leq C$, we have $P_{\beta, \gamma_1, \gamma_2}(C \in [\underline{\hat{C}}, \infty)) \geq 1 - \alpha$.*

Proof. It follows from Lemma C.1 that, on the event $\|2X_2' M_{X_1} \varepsilon\|_q/n \leq \lambda_\alpha^*$ (which holds with probability at least $1 - \alpha$ by assumption), we have $\frac{\lambda - \lambda_\alpha^*}{\lambda + \lambda_\alpha^*} \|\hat{\theta}_2(\lambda)\|_p \leq \|\gamma_2\|_p \leq C$ for all $\lambda > \lambda_\alpha^*$. Thus, the supremum of this quantity over λ in this set is also no greater than C on this event. $\square$

We now present a feasible version of this CI when the error distribution is unknown and possibly heteroskedastic in the case where $p = 1$. Let $\tilde{x}_{ij} = (M_{X_1} X_2)_{ij}$. Since $q = \infty$ in this case, we need to choose $\hat{\lambda}_\alpha^*$ such that

$$2\|X_2' M_{X_1} \varepsilon\|_\infty/n = \max_{1 \leq j \leq k_2} \left| \sum_{i=1}^n 2\tilde{x}_{ij} \varepsilon_i/n \right| \leq \hat{\lambda}_\alpha^*$$

with probability at least $1 - \alpha$ asymptotically. Let $\hat{V}_j = \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 \hat{\varepsilon}_i^2$, where $\hat{\varepsilon}_i$ is the residual from an initial regularized regression with λ chosen as in Theorem C.1 for some slowly increasing K_n . This leads to the moderate deviations critical value $\hat{\lambda}_\alpha^*$, which sets

$$\alpha = \sum_{j=1}^{k_2} 2\Phi(-\hat{\lambda}_\alpha^*/\hat{V}_j^{1/2}). \quad (32)$$

Remark C.2. The analysis in Theorem C.1 of the regularized regression estimator in eq. (25) relies on choosing a penalty parameter greater than $2\|X_2' M_{X_1} \varepsilon\|_\infty/n$ with high probability, which is precisely the goal of the critical value $\hat{\lambda}_\alpha^*$ given in eq. (32). This suggests an iterative procedure in which one uses $\hat{\lambda}_\alpha^*$ (perhaps with some sequence α_n converging slowly to zero) as a data-driven penalty parameter in the regression in eq. (25) after using some initial penalty choice satisfying the conditions of Theorem C.1 to form the residuals used to compute $\hat{\lambda}_\alpha^*$.

The penalty choice $\hat{\lambda}_\alpha^*$ is related to data-driven choices of the lasso penalty in the case with an unknown error distribution. Belloni, Chen, et al. (2012) use similar ideas to choose the penalty parameter in this setting under ℓ_0 constraints, although our implementation is somewhat different, since our parameter space constrains the penalty loadings we place on each parameter. While $\hat{\lambda}_\alpha^*$ does not take into account correlations between the moments, one could take into account these correlations using a bootstrap implementation, as suggested by Chernozhukov, Chetverikov, and Kato (2013).

Theorem C.5. Suppose Assumption C.1 holds with $p = 1$ and that $\frac{1}{n} \sum_{i=1}^n \tilde{x}_{ij}^2 \geq \eta$ for $j = 1, \dots, k_2$ for all n , where $\tilde{x}_{ij} = (M_{X_1} X_2)_{ij}$. Let $\hat{\lambda}_\alpha^*$ be given in eq. (32) with $\hat{V}_j$ formed using residuals $\hat{\varepsilon}$ from the regularized regression in eq. (25) with penalty λ chosen as in

Theorem C.1 for some $K_n \rightarrow \infty$ with $K_n(k_1/n + (C+1)\sqrt{\log k_2}/\sqrt{n}) \cdot (\log k_2)^2 \rightarrow 0$. Then

$$\limsup_n \sup_{\beta, \gamma: \|\gamma_2\|_1 \leq C} \sup_{Q \in \mathcal{Q}_n} P_{\theta, Q} \left(\max_{1 \leq j \leq k_2} \left| \sum_{i=1}^n 2\tilde{x}_{ij}\varepsilon_i/n \right| > \hat{\lambda}_\alpha^* \right) \leq \alpha.$$

In particular, define $\hat{C}$ as in eq. (31), but with λ_α^* replaced with $\hat{\lambda}_\alpha^*$. We then have

$$\liminf_n \inf_{\beta, \gamma: \|\gamma_2\|_1 \leq C} \inf_{Q \in \mathcal{Q}_n} P_{\theta, Q} \left(C \in [\hat{C}, \infty) \right) \geq 1 - \alpha.$$

Proof. Let $\tilde{V}_j = \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 \varepsilon_i^2$ and let $V_{Q,j} = \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 E_Q \varepsilon_i^2$. Note that

$$\begin{aligned} |\hat{V}_j - \tilde{V}_j| &= \left| \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 (\hat{\varepsilon}_i^2 - \varepsilon_i^2) \right| = \left| \sum_{i=1}^n (2\tilde{x}_{ij}/n)^2 (\hat{\varepsilon}_i + \varepsilon_i)(\hat{\varepsilon}_i - \varepsilon_i) \right| \\ &\leq (2K_X/n)^2 \|\hat{\varepsilon} + \varepsilon\|_2 \|\hat{\varepsilon} - \varepsilon\|_2 \leq (2K_X/n)^2 (2\|\varepsilon\|_2 + \|\hat{\varepsilon} - \varepsilon\|_2) \|\hat{\varepsilon} - \varepsilon\|_2. \end{aligned}$$

On the event that $2\|\varepsilon\|_2 \leq \sqrt{n}\tilde{K}$ and

$$\|\hat{\varepsilon} - \varepsilon\|_2 = \|X(\hat{\theta} - \theta)\|_2 \leq \sqrt{nK_n} \cdot (k_1/n + 2C\sqrt{\log k_2}/\sqrt{n})^{1/2},$$

which holds with probability approaching one uniformly over $Q \in \mathcal{Q}_n$ when $\tilde{K}$ is large enough (by Chebyshev's inequality), this is bounded by $(2K_X/n)^2 (\tilde{K}\sqrt{n} + \sqrt{nK_n}(k_1/n + 2C\sqrt{\log k_2}/\sqrt{n})^{1/2}) \cdot \sqrt{nK_n}(k_1/n + 2C\sqrt{\log k_2}/\sqrt{n})^{1/2}$. Since $V_{Q,j} \geq \tilde{\eta}/n$ uniformly over j and over n for some $\tilde{\eta} > 0$, this implies that, on this event, $\max_{1 \leq j \leq k_2} |\hat{V}_j - \tilde{V}_j|/V_{Q,j}$ is bounded by

$$4\tilde{\eta}^{-1} (K_X^2/n) (\tilde{K}\sqrt{n} + \sqrt{nK_n}(k_1/n + 2C\sqrt{\log k_2}/\sqrt{n})^{1/2}) \cdot \sqrt{nK_n}(k_1/n + 2C\sqrt{\log k_2}/\sqrt{n})^{1/2},$$

which in turn is bounded by a constant times $K_n^{1/2}(k_1/n + 2C\sqrt{\log k_2}/\sqrt{n})^{1/2}$ so long as this quantity converges to zero.

In addition, note that $(\tilde{V}_j - V_{Q,j})/V_{Q,j} = \sum_{i=1}^n \tilde{a}_{ij}(\varepsilon_i^2 - E_Q \varepsilon_i^2)/n$, where $\tilde{a}_{ij} = \frac{4\tilde{x}_{ij}^2}{nV_{Q,j}} \leq 4K_X^2\tilde{\eta}^{-1}$ and $\tilde{\eta}$ is a lower bound for $nV_{Q,j}$. Using this bound on $\tilde{a}_{ij}$ and the tail bound on ε_i , it follows from Bernstein's inequality for sub-exponential random variables that, for $\delta < 1$, $P_Q(|\tilde{V}_j - V_{Q,j}|/V_{Q,j} \geq \delta)$ is bounded from above by $2\exp(-c n \delta^2)$ for some constant c that depends only on K_X , $\tilde{\eta}$ and η . Thus, for any sequence δ_n , we have $P_Q(\max_{1 \leq j \leq k_2} |\tilde{V}_j - V_{Q,j}|/V_{Q,j} \geq \delta_n) \leq 2k_2 \exp(-c n \delta_n^2)$, which converges to zero so long as δ_n is bounded from below by a large enough constant times $\sqrt{\log k_2}/\sqrt{n}$.

This gives the rate of convergence for $\hat{V}_j/V_{Q,j}$ to one which, by continuous differentia-

bility of $t \mapsto \sqrt{t}$ at $t = 1$, gives the same rates for $\sqrt{\hat{V}_j}/\sqrt{V_{Q,j}}$. In particular, letting c_n be given by a large enough constant times $K_n^{1/2}(k_1/n + (C+1)\sqrt{\log k_2}/\sqrt{n})^{1/2}$, the event $\max_{1 \leq j \leq k_2} \left| \sqrt{\hat{V}_j}/\sqrt{V_{Q,j}} - 1 \right| \leq c_n$ holds with probability approaching one uniformly over $Q \in \mathcal{Q}_n$ and β, γ with $\|\gamma_2\| \leq C$. On this event, we have

$$\alpha = \sum_{j=1}^{k_2} 2\Phi(-\hat{\lambda}_\alpha^*/\sqrt{\hat{V}_j}) \geq \sum_{j=1}^{k_2} 2\Phi(-\hat{\lambda}_\alpha^*/(\sqrt{V_{Q_n,j}}(1-c_n))).$$

Thus, letting $\lambda_{\alpha,n}$ solve $\alpha = \sum_{j=1}^{k_2} 2\Phi(-\lambda_{\alpha,n}/(\sqrt{V_{Q_n,j}}))$, we have $\hat{\lambda}_\alpha^*/(1-c_n) = \lambda_{\tilde{\alpha},n}$ for some $\tilde{\alpha} \leq \alpha$, so that $\hat{\lambda}_\alpha^*/(1-c_n) \geq \lambda_{\alpha,n}$. It follows that the non-coverage probability under any sequence of parameters with $\|\gamma_2\|_p \leq C$ and any sequence $Q_n \in \mathcal{Q}_n$ is bounded by a term that converges to zero plus

$$\begin{aligned} P_{Q_n} \left(\max_{1 \leq j \leq k_2} \left| \frac{1}{n} \sum_{i=1}^n 2\tilde{x}_{ij}\varepsilon_i \right| > (1-c_n)\lambda_{\alpha,n} \right) &\leq \sum_{j=1}^{k_2} F_{n,j}((1-c_n)\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}}) \\ &= \sum_{j=1}^{k_2} 2\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}}) \cdot A_{n,j} \cdot B_{n,j}, \end{aligned}$$

where $F_{n,j}(t) = P_{Q_n}(|n^{-1} \sum_{i=1}^n 2\tilde{x}_{ij}\varepsilon_i/\sqrt{V_{Q_n,j}}| > t)$, $A_{n,j} = \frac{\Phi(-(1-c_n)\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}})}{\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}})}$ and $B_{n,j} = \frac{F_{n,j}((1-c_n)\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}})}{2\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}})}$. Since $\sum_{j=1}^{k_2} 2\Phi(-\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}}) = \alpha$ by definition, it suffices to show that $\limsup_{n \rightarrow \infty} \max_{1 \leq j \leq k_2} \max\{A_{n,j}, B_{n,j}\} \leq 1$.

For $A_{n,j}$, we use the bound $\Phi(-s)/\Phi(-t) \leq [s^{-1}/(t^{-1}-t^{-3})] \exp((t^2-s^2)/2)$ (this follows from the bound $(t^{-1}-t^{-3}) \exp(-t^2/2)/\sqrt{2\pi} \leq \Phi(-t) \leq t^{-1} \exp(-t^2/2)/\sqrt{2\pi}$ given in Lemma 2, Section 7.1 in Feller (1968)), which gives

$$A_{n,j} \leq \frac{(1-c_n)^{-1}}{1 - (\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}})^{-2}} \exp([1 - (1-c_n)^2]\lambda_{\alpha,n}^2/(2V_{Q_n,j})).$$

Using standard calculations and the fact that $nV_{Q_n,j}$ is uniformly bounded from above and below, we have $(\log k_2)/K \leq \lambda_{\alpha,n}^2/V_{Q_n,j} \leq K \log k_2$ for some constant K . Thus, the right-hand side of the above display converges to 1 uniformly over n and $1 \leq j \leq k_2$ so long as $c_n \log k_2 \rightarrow 0$, which is guaranteed by the assumptions of the theorem.

For $B_{n,j}$, we use a moderate deviations bound as in Feller (1971, Chapter 16.7). In particular, the bound $|F_{n,j}(t)/(2\Phi(-t)) - 1| \leq \tilde{K}t^3/\sqrt{n}$ holds for all $1 \leq t < \bar{t}_n$, where $\bar{t}_n$ is any sequence with $\bar{t}_n/n^{1/6} \rightarrow 0$, and $\tilde{K}$ depends only on $\bar{t}_n$ and the moment conditions and

tail bounds on ε_i (Armstrong and Chan 2016, Lemma B.5). Using the fact that $\lambda_{\alpha,n}/\sqrt{V_{Q_n,j}}$ is bounded by a constant times $\sqrt{\log k_2}$, it follows that $\limsup_{n \rightarrow \infty} \max_{1 \leq j \leq k_2} B_{n,j} \leq 1$ so long as $(\log k_2)^{3/2}/\sqrt{n} \rightarrow 0$, which is guaranteed by the conditions of the theorem. $\square$

References

- Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (Feb. 2005). “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools”. In: *Journal of Political Economy* 113.1, pp. 151–184. DOI: [10.1086/426036](#).
- Angrist, Joshua D. (Mar. 1998). “Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants”. In: *Econometrica* 66.2, pp. 249–288. DOI: [10.2307/2998558](#).
- Armstrong, Timothy B. and Hock Peng Chan (Sept. 2016). “Multiscale adaptive inference on conditional moment inequalities”. In: *Journal of Econometrics* 194.1, pp. 24–43. DOI: [10.1016/j.jeconom.2016.04.001](#).
- Armstrong, Timothy B. and Michal Kolesár (May 2016). *Optimal inference in a class of regression models*. arXiv: [1511.06028v2](#).
- (Mar. 2018). “Optimal Inference in a Class of Regression Models”. In: *Econometrica* 86.2, pp. 655–683. DOI: [10.3982/ECTA14434](#).
- (May 2021a). “Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness”. In: *Econometrica* 89.3, pp. 1141–1177. DOI: [10.3982/ECTA16907](#).
- (Jan. 2021b). “Sensitivity Analysis Using Approximate Moment Condition Models”. In: *Quantitative Economics* 12.1, pp. 77–108. DOI: [10.3982/QE1609](#).
- Von Bahr, Bengt (June 1965). “On the Convergence of Moments in the Central Limit Theorem”. In: *Annals of Mathematical Statistics* 36.3, pp. 808–818. DOI: [10.1214/aoms/1177700055](#).
- Von Bahr, Bengt and Carl-Gustav Esseen (Feb. 1965). “Inequalities for the r th Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$ ”. In: *The Annals of Mathematical Statistics* 36.1, pp. 299–303. DOI: [10.1214/aoms/1177700291](#).
- Belloni, Alexandre, Daniel L. Chen, Victor Chernozhukov, and Christian B. Hansen (Nov. 2012). “Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain”. In: *Econometrica* 80.6, pp. 2369–2429. DOI: [10.3982/ECTA9626](#).
- Belloni, Alexandre, Victor Chernozhukov, and Christian B. Hansen (Apr. 2014). “Inference on Treatment Effects after Selection among High-Dimensional Controls”. In: *The Review of Economic Studies* 81.2, pp. 608–650. DOI: [10.1093/restud/rdt044](#).

- Belloni, Alexandre, Victor Chernozhukov, and Lie Wang (Dec. 2011). “Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming”. In: *Biometrika* 98.4, pp. 791–806. DOI: [10.1093/biomet/asr043](https://doi.org/10.1093/biomet/asr043).
- Boudreau, Laura (2024). “Multinational Enforcement of Labor Law: Experimental Evidence on Strengthening Occupational Safety and Health Committees”. In: *Econometrica* 92.4, pp. 1269–1308. DOI: [10.3982/ECTA19408](https://doi.org/10.3982/ECTA19408).
- Bühlmann, Peter and Sara A. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Heidelberg: Springer. DOI: [10.1007/978-3-642-20192-9](https://doi.org/10.1007/978-3-642-20192-9).
- Cai, T. Tony and Zijian Guo (Apr. 2017). “Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity”. In: *The Annals of Statistics* 45.2, pp. 615–646. DOI: [10.1214/16-AOS1461](https://doi.org/10.1214/16-AOS1461).
- Candès, Emmanuel and Terence Tao (Dec. 2007). “The Dantzig Selector: Statistical Estimation When p Is Much Larger than n ”. In: *The Annals of Statistics* 35.6, pp. 2313–2351. DOI: [10.1214/009053606000001523](https://doi.org/10.1214/009053606000001523).
- Carneiro, Pedro, Lucy Kraftman, Giacomo Mason, Lucie Moore, Imran Rasul, and Molly Scott (Aug. 2021). “The Impacts of a Multifaceted Prenatal Intervention on Human Capital Accumulation in Early Life”. In: *American Economic Review* 111.8, pp. 2506–2549. DOI: [10.1257/aer.20191726](https://doi.org/10.1257/aer.20191726).
- Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (Dec. 2013). “Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors”. In: *The Annals of Statistics* 41.6, pp. 2786–2819. DOI: [10.1214/13-AOS1161](https://doi.org/10.1214/13-AOS1161).
- Cinelli, Carlos and Chad Hazlett (Feb. 2020). “Making Sense of Sensitivity: Extending Omitted Variable Bias”. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 82.1, pp. 39–67. DOI: [10.1111/rssb.12348](https://doi.org/10.1111/rssb.12348).
- Connors Jr., Alfred F., Theodore Speroff, Neal V. Dawson, Charles Thomas, et al. (1996). “The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients”. In: *The Journal of the American Medical Association* 276.11, pp. 889–897. DOI: [10.1001/jama.1996.03540110043030](https://doi.org/10.1001/jama.1996.03540110043030).
- Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik (Sept. 2006). *Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand*. Working Paper 330. Cambridge, MA: National Bureau of Economic Research. DOI: [10.3386/t0330](https://doi.org/10.3386/t0330).
- Dhar, Diva, Tarun Jain, and Seema Jayachandran (Mar. 2022). “Reshaping Adolescents’ Gender Attitudes: Evidence from a School-Based Experiment in India”. In: *American Economic Review* 112.3, pp. 899–927. DOI: [10.1257/aer.20201112](https://doi.org/10.1257/aer.20201112).

- Diegert, Paul, Matthew A. Masten, and Alexandre Poirier (Feb. 2026). *Assessing Omitted Variable Bias When the Controls Are Endogenous*. arXiv: [2206.02303](#).
- Donoho, David L. (Mar. 1994). “Statistical Estimation and Optimal Recovery”. In: *The Annals of Statistics* 22.1, pp. 238–270. DOI: [10.1214/aos/1176325367](#).
- Efron, Bradley, Trevor Hastie, Iain M. Johnstone, and Robert J. Tibshirani (Apr. 2004). “Least Angle Regression”. In: *The Annals of Statistics* 32.2, pp. 407–451. DOI: [10.1214/009053604000000067](#).
- Feller, William (Jan. 1968). *An Introduction to Probability Theory and Its Applications*. 3rd ed. Vol. 1. New York, NY: Wiley.
- (1971). *An Introduction to Probability Theory and Its Applications*. 2nd ed. Vol. 2. New York, NY: Wiley.
- Van de Geer, Sara A., Peter Bühlmann, Ya’acov Ritov, and Ruben Dezeure (June 2014). “On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models”. In: *The Annals of Statistics* 42.3, pp. 1166–1202. DOI: [10.1214/14-AOS1221](#).
- Goldberg, Moshe (Feb. 2017). “Continuity of Seminorms on Finite-Dimensional Vector Spaces”. In: *Linear Algebra and its Applications* 515, pp. 175–179. DOI: [10.1016/j.laa.2016.11.013](#).
- Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (Dec. 2024). “Contamination Bias in Linear Regressions”. In: *American Economic Review* 114.12, pp. 4015–4051. DOI: [10.1257/aer.20221116](#).
- Heckman, Nancy E. (Dec. 1988). “Minimax Estimates in a Semiparametric Model”. In: *Journal of the American Statistical Association* 83.404, pp. 1090–1096. DOI: [10.1080/01621459.1988.10478706](#).
- Hirano, Keisuke and Guido W. Imbens (Dec. 2001). “Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization”. In: *Health Services and Outcomes Research Methodology* 2.3–4, pp. 259–278. DOI: [10.1023/A:1020371312283](#).
- Ibragimov, Ildar A. and Rafail Z. Khas’minskii (1985). “On Nonparametric Estimation of the Value of a Linear Functional in Gaussian White Noise”. In: *Theory of Probability & Its Applications* 29.1, pp. 18–32. DOI: [10.1137/1129002](#).
- Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote (Sept. 2001). “Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players”. In: *American Economic Review* 91.4, pp. 778–794. DOI: [10.1257/aer.91.4.778](#).

- Imbens, Guido W. and Stefan Wager (May 2019). “Optimized Regression Discontinuity Designs”. In: *Review of Economics and Statistics* 101.2, pp. 264–278. DOI: [10.1162/rest_a_00793](#).
- Imbens, Guido W. and Jeffrey M. Wooldridge (Mar. 2009). “Recent Developments in the Econometrics of Program Evaluation”. In: *Journal of Economic Literature* 47.1, pp. 5–86. DOI: [10.1257/jel.47.1.5](#).
- Javanmard, Adel and Andrea Montanari (2014). “Confidence Intervals and Hypothesis Testing for High-Dimensional Regression”. In: *Journal of Machine Learning Research* 15.82, pp. 2869–2909. URL: <http://jmlr.org/papers/v15/javanmard14a.html>.
- (Dec. 2018). “Debiasing the lasso: Optimal sample size for Gaussian designs”. In: *Annals of Statistics* 46.6A, pp. 2593–2622. DOI: [10.1214/17-AOS1630](#).
- Jones, Damon, David Molitor, and Julian Reif (Nov. 2019). “What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study”. In: *The Quarterly Journal of Economics* 134.4, pp. 1747–1791. DOI: [10.1093/qje/qjz023](#).
- Kolesár, Michal and Christoph Rothe (Aug. 2018). “Inference in Regression Discontinuity Designs with a Discrete Running Variable”. In: *American Economic Review* 108.8, pp. 2277–2304. DOI: [10.1257/aer.20160945](#).
- Kwon, Koohyun and Soonwoo Kwon (Nov. 2020). *Inference in Regression Discontinuity Designs under Monotonicity*. arXiv: [2011.14216](#).
- Li, Chenchuan Mark and Ulrich K. Müller (May 2021). “Linear Regression with Many Controls of Limited Explanatory Power”. In: *Quantitative Economics* 12.2, pp. 405–442. DOI: [10.3982/QE1577](#).
- Li, Ker-Chau (Sept. 1982). “Minimaxity of the Method of Regularization of Stochastic Processes”. In: *The Annals of Statistics* 10.3, pp. 937–942. DOI: [10.1214/aos/1176345883](#).
- Low, Mark G. (June 1995). “Bias-Variance Tradeoffs in Functional Estimation Problems”. In: *Annals of Statistics* 23.3, pp. 824–835. DOI: [10.1214/aos/1176324624](#).
- Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich (May 2025). “Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments”. In: *The Review of Economics and Statistics* 107.3, pp. 589–604. DOI: [10.1162/rest_a_01317](#).
- Noack, Claudia and Christoph Rothe (May 2024). “Bias-Aware Inference in Fuzzy Regression Discontinuity Designs”. In: *Econometrica* 92.3, pp. 687–711. DOI: [10.3982/ECTA19466](#).
- Oster, Emily (Apr. 2019). “Unobservable Selection and Coefficient Stability: Theory and Evidence”. In: *Journal of Business & Economic Statistics* 37.2, pp. 187–204. DOI: [10.1080/07350015.2016.1227711](#).

- Rambachan, Ashesh and Jonathan Roth (Feb. 2023). “A More Credible Approach to Parallel Trends”. In: *Review of Economic Studies* 90.5, pp. 2555–2591. DOI: [10.1093/restud/rdad018](#).
- Robinson, Peter M. (July 1988). “Root- N -Consistent Semiparametric Regression”. In: *Econometrica* 56.4, pp. 931–954. DOI: [10.2307/1912705](#).
- Rosset, Saharon and Ji Zhu (July 2007). “Piecewise Linear Regularized Solution Paths”. In: *The Annals of Statistics* 35.3, pp. 1012–1030. DOI: [10.1214/009053606000001370](#).
- Srivastava, Nikhil and Roman Vershynin (Sept. 2013). “Covariance Estimation for Distributions with $2 + \varepsilon$ Moments”. In: *The Annals of Probability* 41.5, pp. 3081–3111. DOI: [10.1214/12-AOP760](#).
- Tibshirani, Robert J. (1996). “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288. DOI: [10.1111/j.2517-6161.1996.tb02080.x](#).
- Vershynin, Roman (2012). “Introduction to the Non-Asymptotic Analysis of Random Matrices”. In: *Compressed Sensing: Theory and Applications*. Ed. by Yonina C. Eldar and Gitta Kutyniok. Cambridge: Cambridge University Press, pp. 210–268. DOI: [10.1017/CB09780511794308.006](#).
- (2026). *High-Dimensional Probability: An Introduction with Applications in Data Science*. 2nd ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press. DOI: [10.1017/9781009490672](#).
- Wahba, Grace (1990). *Spline models for observational data*. Philadelphia, PA: SIAM. DOI: [10.1137/1.9781611970128](#).
- Wüthrich, Kaspar and Ying Zhu (July 2023). “Omitted Variable Bias of Lasso-Based Inference Methods: A Finite Sample Analysis”. In: *The Review of Economics and Statistics* 105.4, pp. 982–997. DOI: [10.1162/rest_a_01128](#).
- Yosida, Kôzoku (1995). *Functional Analysis*. 6th ed. Berlin: Springer. DOI: [10.1007/978-3-642-61859-8](#).
- Zhang, Cun-Hui and Stephanie S. Zhang (Jan. 2014). “Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models”. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76.1, pp. 217–242. DOI: [10.1111/rssb.12026](#).