Bias-Aware Inference in Regularized Regression Models*

Timothy B. Armstrong†
University of Southern California

Michal Kolesár‡
Princeton University

Soonwoo Kwon§
Brown University

August 10, 2023

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 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 an empirical application 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-22049356.

†email: timothy.armstrong@usc.edu
‡email: mkolesar@princeton.edu
§email: soonwoo_kwon@brown.edu

1 Introduction

We are interested in estimation and inference on a scalar coefficient β in a linear regression model

$$Y_i = w_i\beta + z_i'\gamma + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

when the k-vector z_i of controls is large. In such settings, the classic ordinary least squares (OLS) estimator is often uninformative, exhibiting variance that is too large; 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 is the lasso (Tibshirani, 1996) or other variants of ℓ_1 penalization (e.g. Candès and Tao, 2007; Belloni et al., 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 sparse. For inference, several papers have proposed CIs based on "double lasso" estimators (see, among others, Belloni et al., 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014), with asymptotic justification relying on rate conditions for the sparsity of γ. However, in many applications in economics, the sparsity assumption is not compelling and may be hard to motivate. Furthermore, it is unclear what sparsity level this approach implicitly imposes in a given finite sample.

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. Such restrictions arise naturally in a plethora of applications. For instance, the dimensionality of the control vector is often large due to the inclusion of additional controls that are collectively believed to only be weakly associated with the outcome, but are nonetheless included to purge any possible confounding. One can then take the penalty to be an ℓ_p norm for the additional controls. If $z_i'\gamma$ is a basis approximation to some smooth function, we can define $\text{Pen}(\gamma)$ to incorporate bounds on the derivatives of this function. The regularity parameter C plays a role analogous to a sparsity bound.

We obtain sharp finite-sample results deriving near-optimal estimators and CIs under this penalty constraint and the idealized assumption that the regression errors ε_i are Gaussian with a known homoskedastic variance. We show that the class of estimators that exactly resolves the trade-off between worst-case bias and variance can be obtained by (1) running a penalized propensity score regression of w_i on z_i using $\text{Pen}(\cdot)$ as the penalty function; and then (2) using the residuals from this regression 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 obtains automatically as a byproduct of the regularized regression in step (1). 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 choose the weight on the penalty function in step (1) to optimize the MSE of the resulting estimator, or the length of the resulting CI.

In the more realistic setting with heteroskedasticity and an unknown error distribution, bias-aware CIs can be formed using heteroskedasticity-robust variance estimators, and we give conditions for their asymptotic validity, allowing for high-dimensional asymptotics with $k \gg n$. Our setup can allow for the effect of w_i on the outcome to be heterogeneous, either by including interactions of the treatment and demeaned covariates among the controls, or by reinterpreting β in eq. (1) as a weighted average treatment effect. We show that the treatment weights solve a bias-variance tradeoff in a problem where we can pick the estimand to make the estimation problem as easy as possible.

We also employ the high-dimensional asymptotics to study rates of convergence of the bias-aware CIs when $\text{Pen}(\gamma)$ is an ℓ_p norm. We show that if $k \gg n$ and C does not shrink with n, the optimal CI shrinks more slowly than $n^{-1/2}$, so that the bias term asymptotically dominates; accounting for bias in CI construction thus cannot be avoided even in large samples. Furthermore, we show that, in the ℓ_1 case, this rate cannot be improved even if one additionally imposes the same ℓ_1 bound in the propensity score regression of w_i on z_i, as well as a certain degree of sparsity in both regressions.

Explicit specification of the regularity parameter C that bounds the magnitude of γ is a key input for our approach. Our efficiency bounds show that it is impossible to automate the choice of C when forming CIs. We discuss how relating the magnitude of γ to other quantities, such as the magnitude of the control coefficients in a short regression that only includes baseline controls, can help guide its choice. We develop a rule of thumb specification for C based on this idea that we use in our simulations and empirical application. Robustness of the results can be assessed by computing a breakdown value of C, its largest value such that the empirical finding of interest, such as rejecting a particular null hypothesis, holds. Selection of C cannot be automated due to the impossibility of getting a sufficiently informative data-driven upper bound for it. We show, however, that it is possible to obtain a *lower* CI for C, which can be used as a specification check to ensure that the chosen value is not too low.

The requirement to explicitly choose C may seem like a limitation of our approach relative to sparsity-based approaches, where the analogous tuning parameter, the degree of sparsity, does not need to be explicitly specified. However, good finite-sample performance of such methods relies on bounding these tuning parameters implicitly, and such implicit bounds are hard to calculate or evaluate in a given problem. We demonstrate this issue in

a Monte Carlo analysis, where we show that double-lasso CIs suffer from 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 transparent assumptions rather than "asymptotic promises" about tuning parameters that are hard to evaluate in a particular sample.

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, including Armstrong and Kolesár (2021a,b), Kolesár and Rothe (2018), Imbens and Wager (2019), Rambachan and Roth (2023), Noack and Rothe (2021), and Kwon and Kwon (2020). Muralidharan et al. (2023) 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 partly linear model. We provide a novel finite-sample justification for this idea, as well as an exact result giving the optimal penalization of this regression. Our setup allows for a general form of $\mathrm{Pen}(\cdot)$, and yields existing estimators in a few special cases; the bias-aware CIs to accompany such estimators are novel. First, we recover the optimal linear estimators in Heckman (1988), who considered the partly linear model with a penalty function bounding the first or second derivative of a univariate nonparametric regression function. Next, we reproduce the result in Li (1982) that the optimal estimator uses ridge regression when the penalty corresponds to an ℓ_2 norm. Finally, Li and Müller (2021) consider the weighted ℓ_2 norm $\mathrm{Pen}(\gamma) = (\sum_{i=1}^n (z_i'\gamma)^2)^{1/2}$. They develop bias-aware CIs under this penalty based on a likelihood ratio statistic, which are numerically shown to be close to optimal under homoskedasticity and a particular weighted average length criterion. However, unlike our CI, the Li and Müller CI may end up being longer than the long regression CI, as we illustrate in our empirical application in Section 7.

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 unknown error distribution. Section 4 derives rates of convergence under high-dimensional asymptotics and bounds on an ℓ_p norm. 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 an empirical application. Proofs and auxiliary results appear in appendices.

2 Finite-sample results

This section sets up an idealized version of our model with Gaussian homoskedastic errors. We then show how to construct estimators and CIs in this model that are near-optimal in finite samples.

2.1 Setup

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

$$Y = w\beta + Z\gamma + \varepsilon, \tag{2}$$

where $w = (w_1, \ldots, w_n)' \in \mathbb{R}^n$ is the variable of interest with coefficient $\beta \in \mathbb{R}$ and $Z = (z_1, \ldots, z_n)' \in \mathbb{R}^{n \times k}$ 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 k is large relative to n (including the case $k > n$), the researcher needs to make *a priori* restrictions on the control coefficients γ. We assume that these restrictions can be formalized by restricting the parameter space for $(\beta, \gamma')'$ to be $\mathbb{R} \times \Gamma$ where, for some linear subspace $\mathcal{G}$ of $\mathbb{R}^k$ and some seminorm $\text{Pen}(\cdot)$ on $\mathcal{G}$,

$$\Gamma = \Gamma(\text{Pen}; C) = \{\gamma \in \mathcal{G} \colon \text{Pen}(\gamma) \leq C\}. \tag{3}$$

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 cover settings where only a subset of the control coefficients is restricted.

A common class of restrictions arises when $\text{Pen}(\gamma)$ is a weighted ℓ_p norm on a subset of the coefficients. To describe two examples in this class of restrictions, 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, $Z = (Z_1, Z_2)$. Partition $\gamma = (\gamma_1', \gamma_2')'$ accordingly. Let H_A denote the projection matrix onto the column space of a matrix A. Let $\|\cdot\|_p$ denote the ℓ_p norm.

Example 2.1 (ℓ_2 penalty)**.** We specify the penalty as

$$\text{Pen}(\gamma) = \|M\gamma\|_2 = \sqrt{\gamma' M' M \gamma}, \tag{4}$$

where the $k_2 \times k$ matrix M incorporates scaling the variables and picking out which variables are to be constrained. If $M = (0, I_{k_2})$, then $\text{Pen}(\gamma) = \|\gamma_2\|_2$, with γ_1 unconstrained. Setting

$M = (0, (Z_2'(I - H_{Z_1})Z_2/n)^{1/2})$ corresponds to the specification considered in Li and Müller (2021), which restricts the average of the squared mean effects $z_{2i}'\gamma_2$ on Y_i, after controlling for the baseline controls z_{1i}. □

Example 2.2 (ℓ_1 penalty)**.** A weighted ℓ_1 penalty replaces the norm in eq. (4) with an ℓ_1 norm. We focus on the unweighted case for simplicity, setting $\text{Pen}(\gamma) = \|\gamma_2\|_1$. □

In addition to selecting the penalty, the specification of Γ also requires the researcher to pick the regularity parameter C; here we take it 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). Although we rule out non-convex constraints on Γ, such as sparsity, our results nonetheless have implications for such settings, as we discuss in Section 5.

Our goal is to construct estimators and CIs for β. To evaluate estimators $\hat{\beta}$ of β, we consider their worst-case performance over the parameter space $\mathbb{R} \times \Gamma$ under the MSE criterion,

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

where $E_{\beta,\gamma}$ denotes expectation under $(\beta, \gamma')'$. An interval $\{\hat{\beta} \pm \hat{\chi}\}$ with half-length $\hat{\chi} = \hat{\chi}(Y, X)$ is a CI with level $1 - \alpha$ if it satisfies the coverage requirement

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

where $P_{\beta,\gamma}$ denotes probability under $(\beta, \gamma')'$. To compare two CIs under a particular parameter vector $(\beta, \gamma')'$, we prefer the one with shorter expected length $E_{\beta,\gamma}[2\hat{\chi}]$. Note that optimizing expected length will not necessarily lead to CIs centered at an estimator $\hat{\beta}$ that is optimal under the MSE criterion.

2.2 Linear estimators and CIs

We start by considering estimators that are linear in the outcomes Y, $\hat{\beta} = a'Y$, and derive CIs based on such estimators. The n-vector of weights a may depend on the design matrix X or the known variance σ^2. In Section 2.3 below, we show how to choose the weights a optimally, and in Section 2.4 we show that when a is optimally chosen, the resulting estimators and CIs are optimal or near-optimal among all procedures, not just linear ones.

Under a given parameter vector $(\beta, \gamma')'$, the bias of $\hat{\beta} = a'Y$ is given by $a'(w\beta + Z\gamma) - \beta$. As $(\beta, \gamma')'$ ranges over the parameter space $\mathbb{R} \times \Gamma$, the bias ranges over the interval $[-\overline{\text{bias}}_\Gamma(\hat{\beta}), \overline{\text{bias}}_\Gamma(\hat{\beta})]$, where

$$\overline{\text{bias}}_\Gamma(\hat{\beta}) = \sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} a'(w\beta + Z\gamma) - \beta \tag{5}$$

denotes the worst-case bias. The variance of $\hat{\beta}$ does not depend on $(\beta, \gamma')'$, and is given by $\text{var}(\hat{\beta}) = \sigma^2 a'a$.

To form a CI centered at $\hat{\beta}$, note that the z-statistic $(\hat{\beta} - \beta)/\text{var}(\hat{\beta})^{1/2}$ follows a $\mathcal{N}(b, 1)$ distribution with mean bounded by $|b| \leq \overline{\text{bias}}_\Gamma(\hat{\beta})/\text{var}(\hat{\beta})^{1/2}$. Thus, a two-sided CI can be formed as

$$\hat{\beta} \pm \chi, \quad \text{where} \quad \chi = \text{var}(\hat{\beta})^{1/2} \cdot \text{cv}_\alpha\left(\overline{\text{bias}}_\Gamma(\hat{\beta})/\text{var}(\hat{\beta})^{1/2}\right), \tag{6}$$

and $\text{cv}_\alpha(B)$ denotes the $1-\alpha$ quantile of the folded normal distribution, $|\mathcal{N}(B, 1)|$.[1]

This CI is *bias-aware* in that the critical value $\text{cv}_\alpha(\cdot)$ reflects the potential finite-sample bias of $\hat{\beta}$. Following the terminology in Donoho (1994), we refer to the CI as a fixed-length confidence interval (FLCI), since its length 2χ is fixed: it depends only on the non-random design matrix X, and known variance σ^2, but not on Y or the parameter vector $(\beta, \gamma')'$.

2.3 Optimal weights

Both the MSE $R(\hat{\beta}; \Gamma) = \overline{\text{bias}}_\Gamma(\hat{\beta})^2 + \text{var}(\hat{\beta})$ and the FLCI length 2χ given in eq. (6) are increasing in the variance of $\hat{\beta}$ and in its worst-case bias $\overline{\text{bias}}_\Gamma(\hat{\beta})$. Therefore, to find the optimal weights, we first minimize variance subject to a bound B on worst-case bias,

$$\min_{a \in \mathbb{R}} a'a \quad \text{s.t.} \quad \sup_{\beta \in \mathbb{R}, \gamma \in \Gamma} a'(w\beta + Z\gamma) - \beta \leq B. \tag{7}$$

We then vary the bound B to find the optimal bias-variance tradeoff for a given criterion (MSE or FLCI length). Since this optimization does not depend on the outcome data Y, optimizing the weights does not affect the coverage properties of the resulting CI.

Our main computational result, in Theorem 2.1 below, shows that the estimator solving the optimization problem in eq. (7) is given by a simple two-step procedure. In the first step, we estimate a penalized regression of w on Z with penalty $\text{Pen}(\pi)$, so that the coefficient

[1] The critical value $\text{cv}_{1-\alpha}(B)$ 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.

estimate on Z, π_λ, solves

$$\min_{\pi} \|w - Z\pi\|_2^2 \quad \text{s.t.} \quad \text{Pen}(\pi) \leq t_\lambda, \tag{8}$$

where t_λ is a bound on the penalty term. We refer to eq. (8) as a (regularized) propensity score regression, even though we don't require w_i to be binary. In the second step, we use the residuals $\tilde{w}_\lambda := w - Z\pi_\lambda$ from the propensity score regression as instruments in a univariate regression of Y on w. The tuning parameter λ indexes the weight placed on the constraint in eq. (8), and its selection depends on the criterion we are optimizing. It may correspond to the Lagrange multiplier in a Lagrangian formulation of eq. (8), or, if we can solve eq. (8) directly, we may take $t_\lambda = \lambda$.

Theorem 2.1. *Let $\tilde{w}_\lambda = w - Z\pi_\lambda$, where π_λ solves eq. (8), and suppose that $\|\tilde{w}_\lambda\|_2 > 0$. Then $a_\lambda = \frac{\tilde{w}_\lambda}{\tilde{w}_\lambda' w}$ solves eq. (7) with the bound given by $B = \frac{C}{t_\lambda} \cdot a_\lambda' Z\pi_\lambda$. Consequently, the worst-case bias and variance of the estimator*

$$\hat{\beta}_\lambda = a_\lambda' Y = \frac{\tilde{w}_\lambda' Y}{\tilde{w}_\lambda' w} \tag{9}$$

are given by

$$\overline{\text{bias}}_\Gamma(\hat{\beta}_\lambda) = C\overline{B}_\lambda, \quad \textit{and} \quad V_\lambda = \sigma^2 \|a_\lambda\|_2^2, \quad \textit{where } \overline{B}_\lambda = \frac{a_\lambda' Z\pi_\lambda}{\text{Pen}(\pi_\lambda)}. \tag{10}$$

The result follows by applying the general theory of Ibragimov and Khas'minskii (1985), Donoho (1994), Low (1995), and Armstrong and Kolesár (2018) to our setting, which allows us to rewrite eq. (7) as a convex optimization problem. Solving it then yields the result.

With the solution to eq. (7) in hand, for estimation and CI construction, we select penalties λ^*_{MSE} and λ^*_{FLCI} that optimize the MSE and CI length, respectively. Specifically, the penalties solve the univariate optimization problems

$$\lambda^*_{\text{MSE}} = \operatorname*{argmin}_\lambda V_\lambda + (C\overline{B}_\lambda)^2, \quad \lambda^*_{\text{FLCI}} = \operatorname*{argmin}_\lambda \text{cv}_\alpha(C\overline{B}_\lambda / \sqrt{V_\lambda})\sqrt{V_\lambda}, \tag{11}$$

with V_λ and $\overline{B}_\lambda$ given in eq. (10). The optimal linear estimator is then given by $\hat{\beta}_{\lambda^*_{\text{MSE}}}$, and the optimal FLCI takes the form $\hat{\beta}_{\lambda^*_{\text{FLCI}}} \pm \sigma \|a_{\lambda^*_{\text{FLCI}}}\|_2 \cdot \text{cv}_\alpha\left(\frac{C}{\sigma} \frac{a_{\lambda^*_{\text{FLCI}}}' Z\pi_{\lambda^*_{\text{FLCI}}}}{\text{Pen}(\pi_{\lambda^*_{\text{FLCI}}})\|a_{\lambda^*_{\text{FLCI}}}\|_2}\right)$.

As $t_\lambda \to 0$, provided that $\text{Pen}(\cdot)$ is a norm on Z_2, $\hat{\beta}_\lambda$ 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 the other direction, as $t_\lambda \to \infty$, $\hat{\beta}_\lambda$ converges to the long regression estimate $\hat{\beta}_{\text{long}} = \frac{w'(I - H_Z)Y}{w'(I - H_Z)w}$, provided

that w is not in the column space of Z (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. In other words, the short and long regressions are corner solutions of the bias-variance tradeoff, in which weight is entirely placed on variance, or on bias.

Example 2.1 (ℓ_2 penalty, continued)**.** In this case, a convenient Lagrangian formulation for (8) is

$$\pi_\lambda = \underset{\pi}{\operatorname{argmin}} \|w - Z\pi\|_2^2 + \lambda \|M\pi\|_2^2.$$

Suppose $Z'Z + \lambda M'M$ is invertible.[2] Then the first order conditions immediately imply the closed form solution

$$\pi_\lambda = (Z'Z + \lambda M'M)^{-1} Z'w,$$

which is a generalized ridge regression estimator of the propensity score.[3] Plugging this expression for π_λ into eq. (9) yields

$$\hat{\beta}_\lambda = e_1' \left(X'X + \lambda \begin{pmatrix} 0 & 0 \\ 0 & M'M \end{pmatrix} \right)^{-1} X'Y,$$

where $e_1 = (1, 0, \ldots, 0)'$ is the first standard basis vector. Thus, the optimal estimate can also be obtained from a generalized ridge regression of Y onto X. The optimality of ridge regression in this setting was shown by Li (1982), and the above derivation gives this result as a special case of Theorem 2.1. Under the Li and Müller (2021) specification $M = (0, (Z_2'(I - H_{Z_1})Z_2/n)^{1/2})$, the estimator further simplifies to a weighted average of the short and long regression estimates,

$$\hat{\beta}_\lambda = \omega(\lambda)\hat{\beta}_{\text{short}} + (1 - \omega(\lambda))\hat{\beta}_{\text{long}}, \tag{12}$$

with weights

$$\omega(\lambda) = \frac{\lambda/n}{\lambda/n + \varsigma^2}, \qquad \varsigma^2 = \frac{w'(I - H_Z)w}{w'(I - H_{Z_1})w} = \frac{\text{var}(\hat{\beta}_{\text{short}})}{\text{var}(\hat{\beta}_{\text{long}})}.$$

The weight on the short regression increases with λ (as the relative weight on variance in the bias-variance tradeoff increases), and decreases with ς^2. □

[2] Invertibility holds so long as no element $\pi \neq 0$ satisfies $Z\pi = 0$ and $M\pi = 0$ simultaneously. Intuitively, if Z has rank less than k, then the data is not informative about certain directions π, and we require the matrix M to place sufficient restrictions on π in these directions.

[3] We reserve the term "ridge regression" without the qualifier "generalized" for the case $M'M = I_k$.

Example 2.2 (ℓ_1 penalty, continued)**.** In this case, the solution to (8) is given by a variant of the lasso estimate (Tibshirani, 1996) that only penalizes γ_2.

The resulting estimator $\hat{\beta}_\lambda$ is related to estimators proposed for constructing CIs using the lasso (see, among others, Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Belloni et al., 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 non-linear 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.

While under the ℓ_2 constraints with $M = (0, (Z_2'(I - H_{Z_1})Z_2/n)^{1/2})$, the class of optimal estimators in eq. (12) depends on the data only through the short and long regression estimates, $\hat{\beta}_\lambda$ under ℓ_1 constraints (or under ℓ_2 constraints with other choices of M) doesn't simply interpolate between these two extremes, and the optimal weights α_λ display much richer variation. This is analogous to the form of optimal weights in regression discontinuity designs (e.g. Armstrong and Kolesár, 2018; Imbens and Wager, 2019), where one needs to consider a range of bandwidths, rather than just interpolating between estimators that consider the maximal and minimal possible bandwidths. □

Example 2.3 (Partly linear model)**.** To flexibly control for a low-dimensional set of covariates $\tilde{z}_i$, one may specify a semiparametric model

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

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 of order q. 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 $\text{Pen}(\gamma) = \min_{h:\, h(\tilde{z}_i)=\gamma_i,\, i=1,\ldots n} \widetilde{\text{Pen}}(h)$ (assuming the minimum is taken). 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 $g_\lambda(\cdot)$ is analogous to the regularized regression estimate π_λ in (8): it solves

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

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. □

2.4 Efficiency among non-linear procedures

So far, we have restricted attention to procedures that are linear in the outcomes Y. We now show that the estimator $\hat{\beta}_{\lambda^*_{\text{MSE}}}$, and the CI based on the estimator $\hat{\beta}_{\lambda^*_{\text{FLCI}}}$ are in fact highly efficient among all procedures, not just linear ones. This is due to the convexity and symmetry of the parameter space Γ, and 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.

Corollary 2.1. *Let λ^*_{MSE} and λ^*_{FLCI} be given in eq. (11), where the optimization is over all λ with $t_\lambda > 0$ such that $\|w - Z\pi_\lambda\|_2 > 0$. Let $\hat{\beta}_\lambda$, $\overline{B}_\lambda$ and V_λ be given in eq. (10). Let $\tilde{\beta}$ and $\tilde{\beta} \pm \tilde{\chi}$ denote some other (possibly non-linear) estimator and some other (possibly non-linear, variable-length) CI.*

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

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

$$\frac{\text{R}_{\text{MSE}}(\tilde{\beta})}{\text{R}_{\text{MSE}}(\hat{\beta}_{\lambda^*_{\text{MSE}}})} \geq \kappa^*_{\text{MSE}}(X, \sigma, \Gamma) \geq 0.8,$$

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

(iii) The improvement of the expected length of the CI $\tilde{\beta} \pm \tilde{\chi}$ over the optimal linear FLCI $\hat{\beta}_{\lambda^*_{\text{FLCI}}} \pm \text{cv}_\alpha(C\overline{B}_{\lambda^*_{\text{FLCI}}}/V^{1/2}_{\lambda^*_{\text{FLCI}}})V^{1/2}_{\lambda^*_{\text{FLCI}}}$ at $\gamma = 0$ and any β is bounded by

$$\frac{E_{\beta,0}[\tilde{\chi}]}{\text{cv}_\alpha(C\overline{B}_{\lambda^*_{\text{FLCI}}}/V^{1/2}_{\lambda^*_{\text{FLCI}}})V^{1/2}_{\lambda^*_{\text{FLCI}}}} \geq \kappa^*_{\text{FLCI}}(X, \sigma, \Gamma),$$

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

By construction, the estimator $\hat{\beta}_\lambda$ minimizes variance among all linear estimators with a bound $C\overline{B}_\lambda$ on the bias (or equivalently, it minimizes bias among all linear estimators with a bound V_λ on the variance). Corollary 2.1(i) shows that this optimality property is retained if

we enlarge the class of estimators to all estimators, including non-linear ones. As a result, the minimax linear estimator $\hat{\beta}_{\lambda^*_{\text{MSE}}}$ (i.e. the estimator attaining the lowest worst-case MSE in the class of linear estimators) continues to perform well among all estimators, including non-linear ones: by Corollary 2.1(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 2.1(iii) shows that it is not possible to substantively improve upon the FLCI based on $\hat{\beta}_{\lambda^*_{\text{FLCI}}}$ in terms of expected length when $\gamma = 0$, even if we consider variable length CIs that "direct power" at $\gamma = 0$ (potentially at the expense of longer expected length when $\gamma \neq 0$). The construction of the FLCI may appear conservative: its length depends on the worst-case bias over the parameter space for $(\beta, \gamma')'$, which, as the proof of Theorem 2.1 shows, attains at $\gamma = Ct^{-1}_{\lambda^*_{\text{FLCI}}}\pi_{\lambda^*_{\text{FLCI}}}$, with $\text{Pen}(\gamma) = C$. Therefore, one may be concerned that when the magnitude of γ is much smaller than C, the FLCI is too long. Corollary 2.1(iii) shows that this is not the case, and the efficiency of the FLCI is at least 71.7% relative to variable-length CIs that optimize their expected length when $\gamma = 0$. The exact efficiency bound $\kappa^*_{\text{MSE}}(X, \sigma, \Gamma)$ can be computed explicitly in particular applications, and we have found that it is typically considerably higher than 71.7%.

A consequence of Corollary 2.1(iii) is that it is impossible to form a CI that is adaptive with respect to the regularity parameter C that bounds $\text{Pen}(\gamma)$. 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 2.1(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. In particular, it is impossible to automate the choice of the regularity parameter C when forming a CI. We therefore recommend varying C as a form of sensitivity analysis, or using auxiliary information to choose C; see Remark 3.3.

2.5 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 relax this assumption, we replace β in eq. (1) with a covariate-specific coefficient $\beta(z)$ that represents the conditional ATE for units i with $z_i = z$, obtaining the model

$$Y_i = w_i\beta(z_i) + z_i'\gamma + \varepsilon_i. \tag{13}$$

Suppose the parameter of interest takes the form $\int \beta(z)\, d\mu(w,z)$ 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 non-convex averages of $\beta(z)$. 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 2.1 can be applied to inference on this parameter under any convex and symmetric restriction on the function $\beta(\cdot)$ and the parameter γ—only the worst-case bias calculation in eq. (5) changes. We now discuss two particular specifications for μ and $\beta(\cdot)$.

For the first approach, let μ correspond to a weighted empirical measure with weights c_i that sum to one, so the parameter of interest is given by $\tilde{\beta} = \sum_i c_i \beta(z_i)$. For example, the unweighted case $c_i = 1/n$ gives the (conditional on the sample) ATE, while setting $c_i = w_i / \sum_j w_j$ gives the ATE for the treated. Assume also that the function $\beta(z)$ is linear, $\beta(z) = z'\delta$, and that the first element of z_i is a constant. Consider a parameter space for $(\delta', \gamma')'$ given by $\{(\delta', \gamma')' \in \mathcal{B} \colon \mathrm{Pen}(\delta, \gamma) \leq C\}$, where $\mathcal{B}$ is a subspace of a Euclidean space with a seminorm Pen. Allowing $\mathcal{B}$ to be a subspace allows us to restrict $\beta(z)$ to only depend on a subset of the controls. As noted by Imbens and Wooldridge (2009, Section 5.3) in the unpenalized case, we can map this problem back into the model in eqs. (1) and (3) by rewriting eq. (13) under these assumptions as

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

where $z_{i,-1}$ is the vector of controls excluding a constant and δ_{-1} is the corresponding subvector of δ. This is exactly our problem in eq. (1), with the control vector consisting of the original controls z_i as well as the interaction of the treatment with the demeaned controls, $w_i(z_{i,-1} - \sum_i c_i z_{i,-1})$. We use this approach in our empirical application in Section 7.

For the second approach, we compute the same linear estimator and bias-aware CIs as in the homogeneous TE model in eq. (1); we only change 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 $\hat{\beta}$ relative to an estimand $\int \beta(z) d\mu(w,z)$ when the heterogeneity is completely unrestricted by $\widetilde{\mathrm{bias}}_\Gamma(\hat{\beta}; \mu) = \sup_{\beta(\cdot),\gamma} \left[\sum_{i=1}^n a_i w_i \beta(z_i) + a'Z\gamma - \int \beta(z)\, d\mu(w,z)\right]$, where the supremum is over $\gamma \in \Gamma$ and all functions $\beta(\cdot)$. 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 $\tilde{\beta}_{a,w} = \sum_i a_i w_i \beta(z_i)$.

Theorem 2.2. *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* $\mu = \mu^*_{a,w}$, *then* $\widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta};\mu) = \overline{\text{bias}}_\Gamma(\hat{\beta})$, *with* $\overline{\text{bias}}_\Gamma(\hat{\beta})$ *given in eq.* (5). *If* $\mu \neq \mu^*_a$, *then* $\widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta};\mu) = \infty$. *Furthermore, the estimator* $\hat{\beta}_\lambda$ *given in eq.* (9) *solves*

$$\min_{\hat{\beta}=a'Y, a\in\mathbb{R}^n} \text{var}(\hat{\beta}) \quad s.t. \quad \min_{\mu} \widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}) \leq C\overline{B}_\lambda, \tag{14}$$

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

The theorem gives three results for inference when $\beta(\cdot)$ is unrestricted. First, given a linear estimator $a'Y$, the only estimand for which the bias is finite is $\tilde{\beta}_{a,w}$. Second, for this estimand, the bias is the same as in the homogeneous TE model with $\beta(z) = \beta$. Thus, assuming homogeneous TEs leads to valid inference for $\tilde{\beta}_{a,w}$ 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 $\tilde{\beta}_{a_\lambda,w} = \sum_i \tilde{w}_{\lambda,i} w_i \beta(z_i) / \sum_i \tilde{w}_{\lambda,i} w_i$. Observe that, when the treatment w_i is binary, the weights are non-negative if and only if the residual in the propensity score regression $\tilde{w}_{\lambda,i}$ is positive whenever $w_i = 1$—this can be easily verified in a given application. Equivalently, the weights are positive if the fitted values $z_i'\pi_\lambda$ are smaller than one: the fitted values in the propensity score regression must respect the population constraint that treatment probabilities must be smaller than one. This finding is a finite-sample analog of the identification result in Goldsmith-Pinkham et al. (2022), who show that the estimand in the partly 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 (14) is a finite-sample version of the "moving the goalposts" problem considered by Crump et al. (2006, Section 5.4).[4] 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 et al. (2022) 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.

[4]The optimization problem (14) 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.

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 $\text{Pen}(\cdot)$, regularity parameter C, and initial estimates of residuals $\hat{\varepsilon}_{\text{init},1}, \ldots, \hat{\varepsilon}_{\text{init},n}$.

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. Compute the solution path $\{\pi_\lambda\}_{\lambda>0}$ for the regularized propensity score regression in eq. (8), indexed by the penalty weight λ. For each λ, compute $\hat{\beta}_\lambda$ as in eq. (9), and $\overline{B}_\lambda$, and V_λ as in eq. (10), with $\hat{\sigma}^2$ in place of σ^2 in the formula for V_λ.
3. Compute λ^*_{MSE} and λ^*_{FLCI} as in eq. (11), and compute the robust variance estimate $\hat{V}_{\lambda,\text{rob}} = \sum_{i=1}^n {a_{\lambda,i}}^2 \hat{\varepsilon}_{\text{init},i}^2$, where $a_\lambda = \frac{w - Z\pi_\lambda}{(w-Z\pi_\lambda)'w}$.

Return the estimator $\hat{\beta}_{\lambda^*_{\text{MSE}}}$ and the CI $\hat{\beta}_{\lambda^*_{\text{FLCI}}} \pm \text{cv}_\alpha\left(C\overline{B}_{\lambda^*_{\text{FLCI}}}/\hat{V}^{1/2}_{\lambda^*_{\text{FLCI}},\text{rob}}\right)\cdot \hat{V}^{1/2}_{\lambda^*_{\text{FLCI}},\text{rob}}$. □

The following remarks discuss the implementation choices, and the optimality and validity of the baseline procedure.

Remark 3.1 (Validity)**.** As the initial residual estimates $\hat{\varepsilon}_{\text{init},i}$, we can take residuals from a regularized outcome regression of Y on X (see eq. (20) in Appendix B.1). We give conditions for asymptotic validity of the resulting CIs in Appendix B.2. The key requirement is that the maximal Lindeberg weight $\text{Lind}(a_\lambda) = \max_{1\le i\le n} {a_{\lambda,i}}^2 / \sum_{j=1}^n {a_{\lambda,j}}^2$ associated with the estimator $\hat{\beta}_\lambda$ shrink quickly enough relative to error in the estimator used to form the residuals. Ensuring that $\text{Lind}(a_\lambda)$ is small prevents the estimator from putting too much weight on a particular observation, so that the Lindeberg condition for the central limit theorem holds.

Whether these conditions hold for the optimal estimator will in general depend on the form of $\text{Pen}(\gamma)$ and on the magnitude of C relative to n. To ensure that $\text{Lind}(a_\lambda)$ is small enough in a particular sample for a normal approximation to work well, one may impose a bound on this term by only minimizing eq. (11) over λ such that $\text{Lind}(a_\lambda)$ is small enough when computing λ^*_{FLCI}. This is similar to proposals by Noack and Rothe (2021), and Javanmard and Montanari (2014) in other settings. As discussed further in Appendix B.2, under

mild regularity conditions, imposing such a bound doesn't affect the convergence rate of the resulting CI.

Remark 3.2 (Efficiency)**.** The weights $a_{\lambda^*_{\text{FLCI}}}$ and $a_{\lambda^*_{\text{MSE}}}$ are not optimal under heteroskedasticity. One could in principle generalize the feasible generalized least squares (FGLS) approach used for unconstrained estimation by deriving optimal weights under the assumption $\varepsilon \sim \mathcal{N}(0, \Sigma)$ (which simply follows the above analysis after pre-multiplying by $\Sigma^{-1/2}$), and derive conditions under which the estimator and CI that plug in an estimate of Σ are optimal asymptotically when the assumption of known variance and Gaussian errors is dropped. Instead of pursuing this generalization, our baseline implementation computes the weights a_λ under the assumption of homoskedasticity, but we use robust standard errors when computing the CI. Thus, analogous to the ubiquitous practice of reporting OLS with Eicker-Huber-White (EHW) standard errors in the unconstrained setting, our baseline implementation leverages homoskedasticity for efficiency, but the CIs remain valid when the homoskedasticity assumption is violated.

Remark 3.3 (Choice of C)**.** By Corollary 2.1(iii), one cannot use a data-driven rule to automate the choice of C when forming a CI. Therefore, plausible magnitudes of $\text{Pen}(\gamma)$ need to be assessed using prior knowledge.

Such assessments can be aided by relating the magnitude of $\text{Pen}(\gamma)$ to other quantities. Let us now describe an approach to calibrating C that we use in our numerical and empirical work in Sections 6 and 7. Let $z_{i1} = (1, \tilde{z}'_{i1})'$ denote a vector of baseline controls, believed to be important confounders, and let z_{i2} be a possibly high-dimensional vector of additional controls, believed to be less important. Suppose that $\text{Pen}(\gamma) = \|\gamma_2\|$ corresponds to some norm on the additional controls as in Examples 2.1 and 2.2. To formalize the belief that the baseline controls are more important, we use the norm of the population coefficient $\widetilde{\gamma}_{short}$ on $\tilde{z}_{i1}$ in the short regression of Y_i on a constant, w_i and $\tilde{z}_{i1}$ as a bound on $\|\gamma_2\|$. Since $\widetilde{\gamma}_{short}$ is unknown, we set $C^{rot} = \|\widehat{\widetilde{\gamma}}_{short}\|$ as a rule of thumb, where $\widehat{\widetilde{\gamma}}_{short}$ is an OLS estimate of $\widetilde{\gamma}_{short}$.[5]

Calibrations of the regularity parameter C should be complemented by varying C as a form of sensitivity analysis. Robustness of the results can also be assessed by computing two additional values of the regularity parameter. The first is a "breakdown value" C^*, the largest value of C such the empirical finding of interest holds. Second, by way of a specification check, one can form a lower CI $[\underline{\hat{C}}, \infty)$ for C to assess the plausibility of a given

[5]Formally, one should account for sampling uncertainty in $\widehat{\widetilde{\gamma}}_{short}$ to ensure validity of the CI under the assumption $\|\gamma_2\| \leq \|\widetilde{\gamma}_{short}\|$, such as by combining a first stage CI for $\widetilde{\gamma}_{short}$ with a Bonferroni correction. In our Monte Carlos in Section 6, however, we find that the C^{rot} leads to valid coverage when this assumption holds even without additional corrections for sampling uncertainty.

bound on $\text{Pen}(\gamma)$. We present such a CI in Appendix B.3 for the case where $\text{Pen}(\gamma)$ takes the form of an ℓ_p constraint.

Remark 3.4 (Computational issues)**.** Step 2 involves computing the solution path of a regularized regression estimator. Efficient algorithms exist for computing these paths under ℓ_1 penalties and its variants (Efron et al., 2004; Rosset and Zhu, 2007). Under ℓ_2 penalty, the regularized regression 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. (8) 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 Rates of convergence

We now derive the rates of convergence for the optimal linear FLCIs as $n \to \infty$. For ease of notation, we assume all coefficients are constrained, and focus on the case $\text{Pen}(\gamma) = \|\gamma\|_p$ for some $p \geq 1$, and the case $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$ (see Example 2.1). We allow the regularity parameter $C = C_n$ go to 0 or ∞ with the sample size, and consider high dimensional asymptotics where $k = k_n \gg n$. We consider a standard "high dimensional" setting, placing conditions on the design matrix X that hold with high probability when w_i, z_i are drawn i.i.d. over i, with the eigenvalues of $\text{var}((w_i, z_i')')$ bounded away from zero and infinity.

Let $q \in [0, \infty]$ denote the Hölder conjugate of p, satisfying $1/p + 1/q = 1$. We will show that when $\text{Pen}(\gamma) = \|\gamma\|_p$, the optimal linear FLCI shrinks at the rate

$$n^{-1/2} + Cr_q(k, n) \quad \text{where} \quad r_q(k, n) = \begin{cases} k^{1/q}/\sqrt{n} & \text{if } q < \infty, \\ \sqrt{\log k}/\sqrt{n} & \text{if } q = \infty. \end{cases} \tag{15}$$

Furthermore, for $p = 1$ and $p = 2$, we will show that no other CI can shrink at a faster rate. For $p = 1$, we will in fact prove a stronger result showing that imposing sparsity bounds on the outcome and propensity score regressions, in addition to the bound on $\text{Pen}(\gamma)$, does not help achieve a faster rate, unless one assumes sparsity of order greater than $C_n\sqrt{n/\log(k)}$ (termed the "ultra sparse" case in Cai and Guo (2017)). For the case $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$, we will show that the optimal rate is given by $n^{-1/2} + C$ when $k \gg n$.

If $k \gg n$ and $C = C_n$ 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. Since taking $C_n \to 0$ rules out even a single coefficient being bounded away from zero, these

bounds imply that taking $p < 2$ in "high dimensional" settings is necessary for consistency, with $p = 1$ offering the best rate conditions. It also follows from these rate results that if $C_n = C$ does not decrease to zero with n, the bias term can dominate asymptotically, making it necessary to explicitly account for bias in CI construction even in large samples.

4.1 Upper bounds

To state the result, given $\eta > 0$, let $\mathcal{E}_n(\eta)$ denote the set of design matrices X for which there exists $\delta \in \mathbb{R}^k$ such that

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

Let $R^*_{\text{FLCI}}(X, C) = 2\,\text{cv}_\alpha(C\overline{B}_{\lambda^*_{\text{FLCI}}}/V^{1/2}_{\lambda^*_{\text{FLCI}}}) \cdot V^{1/2}_{\lambda^*_{\text{FLCI}}}$ denote the length of the optimal linear FLCI.

Theorem 4.1. *(i) Suppose* $\text{Pen}(\gamma) = \|\gamma\|_p$. *There exists a finite constant* K_η *depending only on* η *such that* $R^*_{\text{FLCI}}(X, C) \leq K_\eta n^{-1/2}(1 + Ck^{1/q})$ *for* $p > 1$, *and* $R^*_{\text{FLCI}}(X, C) \leq K_\eta n^{-1/2}(1 + C\sqrt{\log k})$ *for* $p = 1$ *for any* $X \in \mathcal{E}_n(\eta)$. *(ii) Suppose* $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$. *There exists a finite constant* K_η *depending only on* η *such that* $R^*_{\text{FLCI}}(X, C) \leq K_\eta(n^{-1/2} + C)$ *for any* X *such that* $\eta \leq w'w/n$.

The second part of the theorem follows since the short regression without any controls achieves a bias that is of the order C. The first part shows that the upper bounds on the rate of convergence match those in eq. (15) if the high-level condition $X \in \mathcal{E}_n(\eta)$ holds. The next lemma shows that this high-level condition holds with high probability when w_i, Z_i are drawn i.i.d. from a distribution satisfying mild conditions on moments and covariances.

Lemma 4.1. *Suppose* w_i, z_i *are drawn i.i.d. over* i, *and let* $\delta = \text{argmin}_b E[(w_i - z_i'b)^2]$ *so that* $z_i'\delta$ *is the population best linear predictor error of* w_i. *Suppose that the linear prediction error* $E[(w_i - z_i'\delta)^2]$ *is bounded away from zero as* $k \to \infty$, $E[w_i^2] < \infty$, *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* $X \in \mathcal{E}_n(\eta)$ *with probability at least* $1 - \tilde{\eta}$ *for large enough* n.

4.2 Lower bounds

We now show that the rates in eq. (15) are sharp when $p = 2$, or $p = 1$.

4.2.1 $p = 2$

As with the upper bound in Section 4.1, we derive a bound that holds when the design matrix X is in some set, and then show that this set has high probability when w_i, z_i are drawn i.i.d. from a sequence of distributions satisfying certain conditions. We focus on the case $k \geq n$. Let $\widetilde{\mathcal{E}}_n(\eta)$ denote the set of design matrices X such that

$$\eta \leq \frac{1}{n} w'w \leq \eta^{-1}, \quad \min \text{eig}(ZZ'/k) \geq \eta,$$

where $\text{eig}(A)$ denotes the set of eigenvalues of a square matrix A.

Theorem 4.2. *Let $\hat{\beta} \pm \hat{\chi}$ be a CI with coverage at least $1 - \alpha$ under $\text{Pen}(\gamma) \leq C$. (i) If $\text{Pen}(\gamma) = \|\gamma\|_2$, there exists a constant $c_\eta > 0$ depending only on η such that the expected length under $\beta = 0$, $\gamma = 0$ satisfies $E_{0,0}[\hat{\chi}] \geq c_\eta n^{-1/2}(1 + Ck^{1/2})$ for any $X \in \widetilde{\mathcal{E}}_n(\eta)$. (ii) If $\text{Pen}(\gamma) = \|Z\gamma/\sqrt{n}\|_2$, there exists a constant $c_\eta > 0$ depending only on η such that the expected length under $\beta = 0$, $\gamma = 0$ satisfies $E_{0,0}[\hat{\chi}] \geq c_\eta(n^{-1/2} + C)$ for any $X \in \widetilde{\mathcal{E}}_n(\eta)$.*

If z_i is i.i.d. over i, then EZZ'/k is equal to the $n \times n$ identity matrix times the scalar $\frac{1}{k}\sum_{j=1}^k E[z_{ij}^2]$. Thus, the condition on the minimum eigenvalue of ZZ'/k will hold under concentration conditions on the matrix $Z'Z$ so long as the second moments of the covariates are bounded from below. Here, we state a result for a special case where the z_{ij}'s are i.i.d. normal, which is immediate from Donoho (2006, Lemma 3.4).

Lemma 4.2. *Suppose that w_i are i.i.d. over i and that z_{ij} are i.i.d. normal over i and j. Then, for any $\tilde{\eta} > 0$, there exists $\eta > 0$ such that $X \in \widetilde{\mathcal{E}}_n(\eta)$ with probability at least $1 - \tilde{\eta}$ once n and k/n are large enough.*

4.2.2 $p = 1$

We now consider the case where $p = 1$, as in Example 2.2. Rather than imposing conditions on X in a fixed design setting that hold with high probability (as in Section 4.1 and Section 4.2.1), 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 4.1 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.

We introduce some additional notation to cover the random design setting, which we use only in this section. We consider a random design model

$$Y = w\beta + Z\gamma + \varepsilon, \quad \varepsilon \mid Z, w \sim \mathcal{N}(0, \sigma^2 I_n),$$

$$w = Z\delta + v, \quad v \mid Z \sim \mathcal{N}(0, \sigma_v^2 I_n),$$
$$z_{ij} \sim \mathcal{N}(0,1) \quad \text{i.i.d. over } i, j.$$

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 4.3. *Let $\hat{\beta} \pm \hat{\chi}$ be a CI satisfying $P_\vartheta(\beta \in \{\hat{\beta} \pm \hat{\chi}\}) \geq 1 - \alpha$ for all ϑ in $\Theta(C_n, C_n \cdot K\sqrt{n/\log k}, \eta_n)$ where $\alpha < 1/2$. Suppose $k \to \infty$, $C_n\sqrt{\log k}/n \to 0$ and $C_n \leq \sqrt{k/n} \cdot k^{-\tilde{\eta}}$ for some $\tilde{\eta} > 0$. Then, there exists c such that, if K is large enough and $\eta_n \to 0$ 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 $E_{\vartheta^*}[\hat{\chi}] \geq c \cdot n^{-1/2}(1 + C_n\sqrt{\log k})$ once n is large enough.*

Theorem 4.3 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 4.3, imposing sparsity does not allow one to improve upon the CIs that uses only the ℓ_1 bound $\|\gamma\|_1 \leq C_n$ (thereby attaining the rate in Theorem 4.1), unless one imposes sparsity of order greater than $C_n\sqrt{n/\log k}$. We provide further comparison with CIs that impose sparsity in the next section.

5 Comparison with sparsity constraints

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

$$\widetilde{\Gamma}(s) = \{\gamma \colon \|\gamma\|_0 \leq s\}, \tag{16}$$

where $\|\gamma\|_0 = \#\{j \colon \gamma_j \neq 0\}$ 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 non-convex), this parameter space is not covered by our setup. Nonetheless, as we show in Section 5.1, 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. In Section 5.2, we provide a comparison of our approach to these double lasso CIs.

5.1 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. (8) is the lasso estimate in the propensity score regression of w on Z, and our estimator (9) 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 et al., 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014). For concreteness, we focus on the estimator in Zhang and Zhang (2014), which is given by

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

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

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

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

Remark 5.1. Note that $\hat{\beta}_{\text{ZZ}}$ is non-linear 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 non-convex parameter space (16). In contrast, Corollary 2.1 shows that under the convex parameter space $\Gamma = \{\gamma \colon \|\gamma\|_1 \leq C\}$, the estimator $\hat{\beta}_\lambda$ in (9) 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 use of other non-linear estimators.

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

$$\|\hat{\gamma}_{\text{lasso}} - \gamma\|_1 \leq \tilde{C} \quad \text{where} \quad \tilde{C} = \text{const.} \cdot s\sqrt{\log k}/\sqrt{n}, \tag{17}$$

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 $\tilde{Y} = Y - \hat{\beta}_{\text{lasso}} - Z\hat{\gamma}_{\text{lasso}}$. Heuristically, we can treat the bound in eq. (17) 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 Zhang and Zhang (2014). Whereas Zhang and Zhang (2014) motivate their approach as one possible way of correcting the initial estimate $\hat{\beta}_{\text{lasso}}$ using the bound in eq. (17), the above analysis shows that their correction is in fact identical to an approach in which one optimizes this correction numerically.[6]

Under the bound in eq. (17) 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, $|\tilde{b}| \leq \tilde{C}\overline{B}_\lambda$, with $\overline{B}_\lambda$ given in Theorem 2.1 and $\tilde{C}$ given in eq. (17), and the variance of the random term $a_\lambda'\varepsilon$ is given by V_λ in Theorem 2.1. Using arguments similar to those used to prove Theorem 4.1, it follows that $\tilde{C}\overline{B}_\lambda/\sqrt{V_\lambda}$ 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 Zhang and Zhang (2014), which takes the form

$$\{\hat{\beta}_{\text{ZZ}} \pm z_{1-\alpha/2}\hat{V}_\lambda^{1/2}\}, \tag{18}$$

where $\hat{V}_\lambda$ is an estimate of the variance V_λ. We use the term "double lasso CI" to refer to this CI, and to related CIs such as those proposed in Belloni et al. (2014); Javanmard and Montanari (2014); van de Geer et al. (2014).

Remark 5.2. To avoid the assumption that $s(\log k)/\sqrt{n} \to 0$ one could, in principle, extend our approach and the above analysis to form valid bias-aware CIs as $\{\hat{\beta}_{\text{ZZ}} \pm [\tilde{C}\overline{B}_\lambda + z_{1-\alpha/2}\hat{V}_\lambda^{1/2}]\}$.[7] Unfortunately, finding a computable constant $\tilde{C}$ in (17) that is sharp enough to yield useful bounds in practice appears to be difficult, although it is an interesting area for future research.

[6]The estimator proposed by Javanmard and Montanari (2014) performs a numerical optimization of this form, but with the constraint (17) 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 Zhang and Zhang (2014).

[7]We use the slightly more conservative approach of adding and subtracting the bound $\tilde{C}\overline{B}_\lambda$ rather than using the critical value $\text{cv}_\alpha(\tilde{C}\overline{B}_\lambda/\hat{V}_\lambda^{1/2})$ as in eq. (6), since the "bias" term for $\hat{\beta}_{\text{ZZ}}$ is correlated with ε through the first step estimates $\hat{\beta}_{\text{lasso}}, \hat{\gamma}_{\text{lasso}}$.

5.2 Comparison of our approach with CIs based on double lasso estimators

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.

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 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 regularity parameter 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 lead to the bias of $\hat{\beta}_{ZZ}$ decreasing more quickly than its standard deviation. Thus, the CI in eq. (18) is "asymptotically valid" without the need to explicitly specify the sparsity index s: one need only make an "asymptotic promise" that s increases slowly enough. However, such asymptotic promises are difficult to evaluate in a given finite-sample setting. Indeed, as shown by Wüthrich and Zhu (2021) and confirmed in our Monte Carlos 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. (18), 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. (17)), 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 with $s \gg \sqrt{n}/\log k$. Indeed, Theorem 4.3 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 et al., 2023). In contrast, our approach optimally incorporates the bound C regardless of the asymptotic regime.

6 Simulation results

We now illustrate the performance of our methods when the penalty takes the form of an ℓ_1 norm on a subset of k_2 controls, as in Example 2.2. We consider a design taken from Belloni et al. (2014), with data generated from a random regressor model that supplements eq. (2) with a propensity score regression

$$w = Z\pi + \sigma_{\tilde{w}}\tilde{w},$$

with $\tilde{w}_i$ and ε_i independent standard Gaussian, and independent of z_i, which are distributed i.i.d. $\mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 2^{-|i-j|}$. Similar to Wüthrich and Zhu (2021), we tweak the Belloni et al. (2014) design by considering regression coefficients that are of similar magnitude rather than decaying. This allows us to separately vary the degree of sparsity and the signal-to-noise ratio. Specifically, we set

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

We consider three methods for constructing CIs for β with nominal level 95%. The first two methods implement Algorithm 3.1, with the penalty given by the ℓ_1 norm of γ_2, the last k_2 regression coefficients. The first method, which we refer to as "oracle," sets the penalty parameter C to the actual value of $\|\gamma_2\|_1$, and uses knowledge of the variance of the error term ε_i. The second CI, termed "AKK," uses initial residual estimates based on the lasso estimator (that only penalizes γ_2), with the penalty chosen via 10-fold cross-validation. The CI uses the rule of thumb calibration $C^{rot} = \|\widehat{\gamma}_{short}\|_1$ from Remark 3.3, where $\widehat{\gamma}_{short}$ are OLS estimates from a short regression that only includes the first k_1 controls (the "baseline"

controls). The final method, termed "BCH", implements the double lasso procedure by Belloni et al. (2014), using the R package `hdm`, without penalizing the k_1 baseline controls and including an intercept.

The data generating process (DGP) in our random regressor model depends on 8 parameters: n, k_1, k_2, s, β, $\sigma_{\tilde{w}}$, c_1 and c_2. We consider $n \in \{500, 1000\}$, $k_1 \in \{5, 10\}$, $k_2 \in \{100, 200, 500, 1000\}$, $s \in \{10, 20, 100\}$, $\beta \in \{0, 2\}$, and $\sigma_{\tilde{w}} \in \{0.5, 1\}$. We calibrate c_1 and c_2 by fixing the population R^2 from the regression of Y on Z, and fixing the ratio $\nu_{rot} = \|\gamma_2\|_1 / \|\tilde{\gamma}_{short}\|_1$. This allows us to directly control the signal-to-noise ratio, and the validity of the rule-of-thumb calibration: if $\nu_{rot} \leq 1$, then the population restriction underlying our rule of thumb is valid. We consider 4 values for the population R^2, $\{0.01, 0.1, 0.25, 0.5\}$, and 12 values for ν_{rot}, $\{0.2, 0.4, \ldots, 2.4\}$. This gives a total of 9,216 DGPs.

Table 1 reports the simulation results for $n = 500$. The results for $n = 1000$ are reported in Table 2. In line with the theory, the coverage of the oracle CI is close to nominal across all designs.[8] When $\nu_{rot} \leq 1$, coverage of the AKK CI is likewise close to nominal. Under mild violations of the population constraint, the CIs display moderate undercoverage: when $\nu_{rot} \leq 1.5$, coverage remains over 86.6% across all designs, and over 90.7% when $k_2 \leq 200$. Only when $\nu_{rot} > 1.5$ and $k_2 \geq 500$, the undercoverage becomes more severe. In contrast, the BCH method displays moderate undercoverage even in sparse designs with $s = 10$, with coverage at about 85% when $k_2 = 1000$ and $n = 500$. The undercoverage gets more severe, with coverage dipping below 60% once $s = 20$, and the CIs almost entirely miss the true parameter in dense designs with $s = 100$. These results illustrate the concern discussed in Section 5.2 that asymptotic sparsity requirements may be difficult to evaluate in finite samples.

The favorable coverage of the AKK CIs relies heavily on using the bias-aware critical value. Unreported simulations show that the coverage of CIs constructed using the same estimators as the AKK CIs but with standard critical values (i.e., 1.96 for 95% coverage), rather than our bias-aware critical values, can be as low as 74.8% for DGPs with $\nu_{rot} \leq 1$.

The AKK CIs display a mild increase in average length relative to the oracle, with the length penalty ranging between 0 and 16%. The length penalty relative to the BCH method is also in this range for designs where both methods achieve good coverage. This is a bargain price to pay for the much more reliable and transparent coverage performance.

[8] The slight undercoverage reported in the tables is due to Monte Carlo error: with 1000 simulation draws, the expected worst-case coverage over 160 DGPs is 93% if the true coverage for each DGP is 95%.

	$k_2 = 100$			$k_2 = 200$			$k_2 = 500$			$k_2 = 1000$		
ν_{rot}	AKK	BCH	Or	AKK	BCH	Or	AKK	BCH	Or	AKK	BCH	Or
Panel A: Coverage Probability, minimum across DGPs												
						$s = 10$						
$[0, 1]$	92.6	91.3	93.1	92.4	88.8	93.2	93.4	88.4	93.6	93.3	85.4	93.8
$(1, 1.5]$	91.3	89.2	93.7	91.1	86.6	93.8	89.2	86.9	93.7	87.4	84.9	93.9
$(1.5, 2.5]$	85.8	90.2	93.1	83.3	88.9	93.3	78.2	86.2	93.5	68.0	83.0	93.7
						$s = 20$						
$[0, 1]$	92.1	80.1	93.2	92.9	75.5	93.7	93.3	68.1	93.5	93.7	62.2	93.9
$(1, 1.5]$	92.2	80.9	94.1	90.7	74.0	94.1	90.1	67.9	93.4	86.6	58.7	94.6
$(1.5, 2.5]$	85.6	79.2	93.2	82.9	72.4	92.9	74.4	68.3	93.8	65.5	59.1	93.4
						$s = 100$						
$[0, 1]$	92.9	40.6	93.8	93.1	35.7	93.3	94.5	33.7	93.6	94.5	32.5	93.4
$(1, 1.5]$	92.8	17.1	93.9	93.9	14.2	93.9	93.6	9.8	93.8	92.6	8.1	94.7
$(1.5, 2.5]$	90.8	2.5	93.9	88.6	1.6	94.3	80.5	0.7	95.0	70.7	0.3	94.3
Panel B: Relative length, average across DGPs												
						$s = 10$						
$[0, 1]$	1.01	0.96		1.04	0.94		1.10	0.91		1.16	0.89	
$(1, 1.5]$	0.98	0.95		0.99	0.92		0.99	0.86		1.00	0.82	
$(1.5, 2.5]$	0.97	0.95		0.96	0.91		0.95	0.85		0.94	0.80	
						$s = 20$						
$[0, 1]$	1.01	0.96		1.04	0.94		1.10	0.90		1.16	0.87	
$(1, 1.5]$	0.98	0.94		0.98	0.90		0.98	0.84		0.98	0.79	
$(1.5, 2.5]$	0.96	0.94		0.95	0.90		0.92	0.82		0.91	0.76	
						$s = 100$						
$[0, 1]$	1.01	0.95		1.04	0.92		1.10	0.88		1.16	0.85	
$(1, 1.5]$	0.98	0.92		0.98	0.87		0.97	0.79		0.96	0.73	
$(1.5, 2.5]$	0.96	0.91		0.95	0.85		0.90	0.74		0.86	0.67	

Notes: For each method, panel A reports the worst-case coverage probability of nominal 95% level CIs over 160 DGPs for $\nu_{rot} \in [0, 1]$ and $\nu_{rot} \in (1.5, 2.5]$, and 64 DGPs for $\nu_{rot} \in (1, 1.5]$, where each DGP averages across 1000 Monte Carlo draws. Panel B reports the average relative length across the DGPs. Relative length is defined as the average length of the AKK and BCH CIs, averaged over the Monte Carlo draws, divided by the average length of the oracle CI.

Table 1: Simulation results for $n = 500$.

	$k_2 = 100$			$k_2 = 200$			$k_2 = 500$			$k_2 = 1000$		
ν_{rot}	AKK	BCH	Or	AKK	BCH	Or	AKK	BCH	Or	AKK	BCH	Or
Panel A: Coverage Probability, minimum across DGPs												
						$s = 10$						
$[0, 1]$	92.8	91.6	93.3	93.1	91.1	93.2	92.3	90.6	93.1	92.8	90.1	92.6
$(1, 1.5]$	92.9	92.9	92.9	92.7	92.9	93.5	91.9	92.7	93.5	89.7	91.8	93.7
$(1.5, 2.5]$	90.0	91.7	93.0	88.6	92.2	92.9	84.0	91.5	92.6	79.1	91.4	93.5
						$s = 20$						
$[0, 1]$	92.8	86.1	93.0	93.0	84.3	92.8	93.6	79.2	93.4	93.6	74.9	93.6
$(1, 1.5]$	92.6	86.4	93.1	92.4	83.5	93.3	89.8	78.6	93.0	87.9	75.6	94.0
$(1.5, 2.5]$	89.4	86.0	92.9	89.0	82.4	93.5	85.0	78.2	93.2	76.9	70.1	92.9
						$s = 100$						
$[0, 1]$	92.9	27.1	93.2	93.2	18.6	93.8	94.1	13.3	93.7	94.2	10.3	93.7
$(1, 1.5]$	92.4	9.5	93.7	93.2	6.8	93.6	94.0	4.2	94.3	93.1	3.6	95.0
$(1.5, 2.5]$	92.0	2.9	93.0	91.1	0.8	92.7	85.8	0.1	94.2	75.4	0.0	94.7
Panel B: Relative length, average across DGPs												
						$s = 10$						
$[0, 1]$	1.00	0.98		1.02	0.96		1.05	0.94		1.09	0.91	
$(1, 1.5]$	0.99	0.97		0.99	0.95		0.98	0.91		0.98	0.86	
$(1.5, 2.5]$	0.98	0.97		0.97	0.95		0.95	0.90		0.93	0.84	
						$s = 20$						
$[0, 1]$	1.00	0.98		1.01	0.96		1.05	0.93		1.09	0.90	
$(1, 1.5]$	0.99	0.97		0.98	0.94		0.98	0.89		0.97	0.84	
$(1.5, 2.5]$	0.98	0.97		0.97	0.94		0.94	0.88		0.91	0.82	
						$s = 100$						
$[0, 1]$	1.00	0.97		1.01	0.95		1.05	0.92		1.09	0.88	
$(1, 1.5]$	0.99	0.95		0.99	0.92		0.98	0.86		0.96	0.79	
$(1.5, 2.5]$	0.98	0.95		0.97	0.91		0.94	0.83		0.89	0.74	

Notes: See Table 1.

Table 2: Simulation results for $n = 1000$

7 Empirical application

This section shows the performance of our methods using survey data on $n = 496$ winners of major and minor prizes in the Massachusetts lottery in 1984–88 from Imbens et al. (2001) to estimate the marginal propensity to earn (MPE) out of unearned income, a key structural parameter in labor and public economics. While unearned income is typically endogenous, Imbens et al. (2001) argue that in this sample, observable individual characteristics proxy well enough for the frequency of lottery ticket purchases that the magnitude of winnings is as good as random. 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, onto yearly lottery payments, X_i, and individual controls, the coefficient on X_i may be interpreted as the MPE.

We focus on a specification taken from Li and Müller (2021), who augment a baseline set of $k_1 = 7$ individual controls Z_1 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_2 that are constructed by taking demeaned cross-products of 4 the baseline binary controls and their interactions with X_i and dropping collinear terms. Both Z_1 and Z_2 are standardized. Following the discussion in Section 2.5, the coefficient on X_i in this specification can be interpreted as the average MPE, allowing for heterogeneity in the MPE with respect to the binary controls. In contrast, the short regression estimand in a regression that only includes Z_1 is biased for the average MPE in presence of such heterogeneity.

The MPE estimate in the long regression equals -0.049, close to the short regression estimate -0.052 that only includes the baseline controls Z_1 and corresponds to the specification in Table 4, column II row 1 in Imbens et al. (2001). However, the long regression estimate is very noisy: the 95% confidence interval $(-0.115, 0.016)$ includes positive values for the average MPE which economic theory rules out, and it is over 3 times longer than the short regression CI $(-0.073, -0.032)$. To increase precision of inference, Li and Müller (2021) restrict the average squared mean effects $z_{2i}'\gamma_2$ using an ℓ_2 penalty given in Example 2.1. Calibrating C to the rule of thumb value from Remark 3.3, $C^{rot} = 7.2$, yields the CI $(-0.116, 0.018)$ using the Li and Müller (2021) method, which is even longer than the long regression CI.[9] The CI constructed using our method, $(-0.114, 0.015)$, improves slightly upon the long CI, but it is still too wide to be informative.[10] The Li and Müller (2021) penalty

[9] 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.

[10] To make the methods more comparable and not conflate the comparison with differences in standard

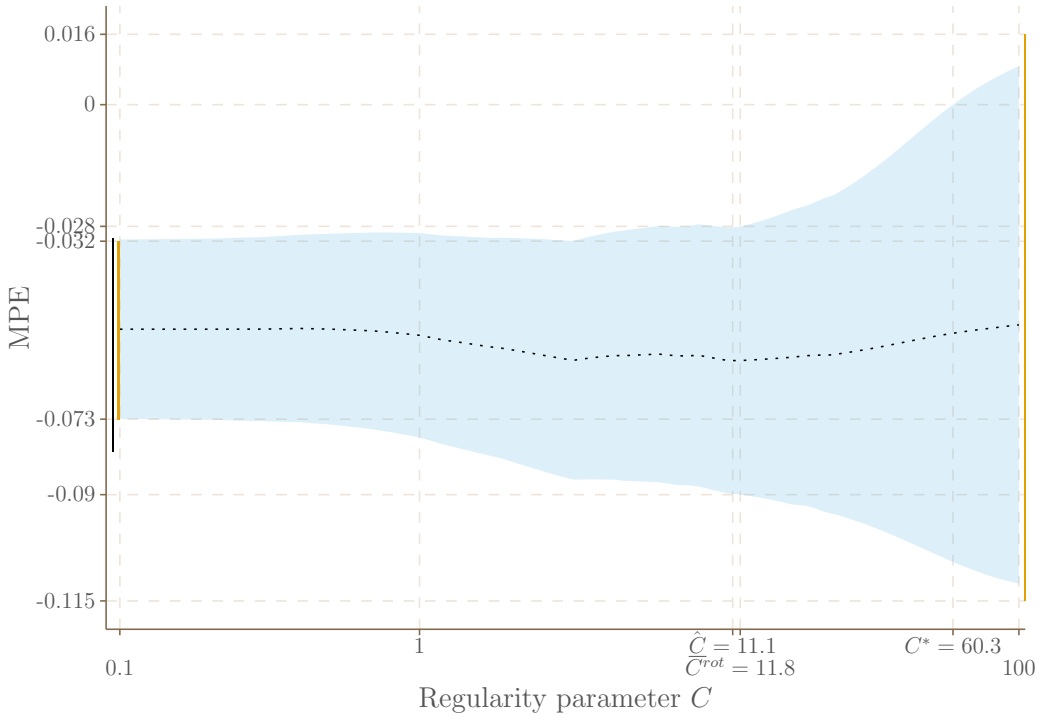

Figure 1: 95% CIs for the marginal propensity to earn out of unearned income under ℓ_1 penalty.

Notes: Orange solid vertical lines at the left and right endpoints of the x-axis mark CIs based on the short and long regression, respectively. Black solid vertical line at the left endpoint of the x-axis marks BCH CI. The blue shaded area depicts the bias-aware CIs, while the dotted black line shows the point estimates as a function of the regularity parameter C. The rule of thumb value of the regularity parameter, C^{rot}, its breakdown value C^*, and the endpoint $\hat{\underline{C}}$ of a lower CI for C, discussed in Remark 3.3, are all marked on the x-axis (log scale).

affords only marginal precision gains because the penalty limits the average influence of the additional regressors—but these regressors only marginally increase the regression R^2 in the long regression: the adjusted R^2 increases from 0.233 in the short regression to 0.236 in the long regression.

In contrast, limiting the total influence of the additional controls by imposing a bound on $\|\gamma_2\|_1$ as in Example 2.2 yields much more substantive precision gains. Figure 1 depicts the CIs constructed using the implementation in Algorithm 3.1 for a wide range of the penalty parameter. The rule of thumb calibration from Remark 3.3 yields $C^{rot} = 11.8$. At this calibration the point estimate is -0.059, with a CI given by $(-0.090, -0.028)$, about half as long as the long regression CI (depicted by an orange vertical line in the figure). In line with the simulation results in Section 6, the CI is also close to the double lasso CI of Belloni et al.

error construction, variance estimates underlying CIs for all methods use residual estimates based on a lasso estimator that penalizes only γ_2, with penalty chosen by 10-fold cross-validation.

(2014), given by $(-0.080, -0.031)$, (depicted in the figure by a black vertical line). Doubling the rule-of-thumb value of C changes the CI little, yielding $(-0.092, -0.026)$.

Appendix A Proofs

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

A.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

$$\omega(\delta) = \sup_{\beta,\pi} 2\beta \quad \text{s.t.} \quad \beta\|w - Z\pi\|_2 \leq \delta/2, \quad \beta\,\text{Pen}(\pi) \leq C. \tag{19}$$

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), $(\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}\prime})'$ plays the role of g_δ^*, and the solution (f_δ^*, g_δ^*) satisfies $f_\delta^* = -g_\delta^* = -(\beta_\delta^{\text{mod}}, \gamma_\delta^{\text{mod}\prime})'$ by centrosymmetry.

This optimization problem is clearly related to the problem in eq. (8): we want to make $\|w - Z\pi\|_2$ and $\text{Pen}(\pi)$ small so that large values of β satisfy the constraint in (19). The following lemma formalizes the connection.

Lemma A.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. (19) has a solution $\beta_\delta^{mod}, \pi_\delta^{mod}$ with $\beta_\delta^{mod} > 0$. For $t_\lambda = C/\beta_\delta^{mod} = 2C/\omega(\delta)$, this solution π_δ^{mod} is also a solution to the penalized regression (8) with optimized objective $\|w - Z\pi_\delta^{mod}\|_2 = \delta/(2\beta_\delta^{mod}) = \delta/\omega(\delta) > 0$; and (ii) for any $t_\lambda > 0$, the penalized regression problem (8) 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 (19) 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. (8) 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 (8) 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 (19) 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 $\beta, \tilde{\pi}^{(1)}$ 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, for small enough η, $\|w - Z\tilde{\pi}_\eta\|_2$ 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 (19) 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 $\text{Pen}(\pi_\lambda) \leq t_\lambda$. □

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 A.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 $a_\lambda = \frac{w - Z\pi_\delta}{(w - Z\pi_\delta)'w}$, 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).

A.2 Proof of Corollary 2.1

Part (i) of Corollary 2.1 follows from Low (1995). In particular, consider the one-dimensional submodel $\beta \in [-C/t_\lambda, C/t_\lambda]$, $\gamma = -\pi_\lambda \beta$. Let $b_\lambda = (w - Z\pi_\lambda)/\|w - Z\pi_\lambda\|_2^2$, and let $B \in \mathbb{R}^{(n-1)\times n}$ be an orthogonal matrix that's 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 $\sup_\beta \operatorname{var}(\delta(Y))$ among all estimators $\delta(Y)$ with $\sup_\beta |E_\beta[\delta(Y)] - \beta| \leq (1-\kappa)C/t_\lambda = C\overline{B}_\lambda$, and, likewise, it minimizes $\sup_\beta |E_\beta[\delta(Y)] - \beta|$ among all estimators with $\sup_\beta \operatorname{var}(\delta(Y)) \leq \kappa^2\sigma^2\|b_\lambda\|_2^2 = V_\lambda$. Since the worst-case bias $\overline{\operatorname{bias}}_\Gamma(\hat{\beta}_\lambda) \leq C\overline{B}_\lambda$ and variance $(\hat{\beta}_\lambda) = V_\lambda$ are the same in the full model by Theorem 2.1, the result follows.

Part (ii) of Corollary 2.1 is immediate from Donoho (1994). In particular, it holds with

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

where $\omega(\delta)$ is defined in eq. (19), 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 2.1 follows from Corollary 3.3 in Armstrong and Kolesár (2018), with

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

where $Z \sim \mathcal{N}(0,1)$, $\omega(\delta)$ is given in eq. (19), and by Lemma D.1 in Armstrong and Kolesár, since the problem is translation invariant with ι given by the parameter $\beta = 1$, $\gamma = 0$, $\omega'(\delta) = \delta/[w'(w - Z\pi_\delta^{\mathrm{mod}}) \cdot \omega(\delta)]$. The universal lower bound 0.717 when $\alpha = 0.05$ follows from Theorem 4.1 in Armstrong and Kolesár (2021b).

A.3 Proof of Theorem 2.2

Note that $\widetilde{\overline{\operatorname{bias}}}_\Gamma(\hat{\beta}; \mu^*_{a,w}) = \sup_{\gamma\in\Gamma} a'Z\gamma$. If $a'w = 1$, then it follows from eq. (5) that $\overline{\operatorname{bias}}_\Gamma(\hat{\beta}) = \sup_{\gamma\in\Gamma} a'Z\gamma$. This proves the first part of the theorem.

To prove the second part of the theorem, note that, if μ is any signed measure not equal to $\mu^*_{a,w}$, then we must have (i) $\mu(\{w_j, z_j\}) \neq \sum_{i:z_i=z_j} a_i w_i$ for some j or (ii) μ must place positive mass on some subset $\mathcal{Z}$ that does not intersect with $\{(w_i, z_i)\}_{i=1}^n$. If (i) holds, then

the bias $\sum_{i=1}^n a_i w_i \beta(z_i) + a'Z\gamma - \int \beta(z)\, d\mu(w,z)$ can be made arbitrarily large by making $\beta(z_j)$ large and setting $\beta(z) = 0$ for $z \neq z_j$. If (ii) holds, then the bias $\sum_{i=1}^n a_i w_i \beta(z_i) + a'Z\gamma - \int \beta(z)\, d\mu(w,z)$ can be made arbitrarily large by setting $\beta(z)$ to be constant on $z \in \mathcal{Z}$ and equal to a number that is set to be arbitrarily large, and setting $\beta(z) = 0$ elsewhere. Thus, $\widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu^*_{a,w}) = \infty$ if $\mu \neq \mu^*_{a,w}$.

To prove the final assertion, the weights a_λ that 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 fist assertion that $\overline{\text{bias}}_\Gamma(\hat{\beta}) = \min_\mu \widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu) = \widetilde{\overline{\text{bias}}}_\Gamma(\hat{\beta}; \mu^*_a)$, where the minimization is over all signed measures μ that integrate to one. Thus, under the heterogeneous TE model (13), the weights a^*_λ solve eq. (14).

A.4 Proof of Theorem 4.1

To prove that the claimed upper bound holds for $X \in \mathcal{E}_n(\eta)$, we first note that, since the FLCI based on $\hat{\beta}_{\lambda^*_{\text{FLCI}}}$ is shorter than the FLCI 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 $n^{-1/2}(1+Ck^{1/q})$ when $p > 1$ or $n^{-1/2}(1+C\sqrt{\log k})$ when $p = 1$. We consider 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 $\mathcal{E}(\eta)$. The variance of the estimator $\tilde{a}'Y$ is $\frac{\sum_{i=1}^n v_i^2}{\left(\sum_{i=1}^n v_i w_i\right)^2} \leq \eta^{-3}/n$. The worst-case bias is

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

where the first equality follows by Hölder's inequality, and the last quality follows by definition of $\mathcal{E}_n(\eta)$. This yields the convergence rate $n^{-1/2} + Cr_q(k,n)$, as claimed. For part (ii), by analogous reasoning, it suffices to consider the short regression estimator $\hat{\beta}_0 = w'Y/w'w$. The variance of this estimator is $\sigma^2/w'w \leq \eta^{-1}\sigma^2/n$. The bias of the estimator is $w'Z\gamma/w'w$. By the Cauchy-Schwarz inequality, this quantity is bounded in absolute value by $\|w/w'w\|_2\|Z\gamma\|_2 = \|Z\gamma/\sqrt{n}\|_2/\sqrt{w'w/n} \leq \eta^{-1/2}C$. This yields the desired convergence rate.

A.5 Proof of Lemma 4.1

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 $Ew_i^2 < \infty$, 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 large enough η by the law of large numbers for triangular

arrays.

For the last inequality in the definition of $\mathcal{E}_n(\eta)$, 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 $2k\exp\left(-K\cdot\eta^{-2}\log k\right) = 2k^{1-K\eta^{-2}}$ for some constant K by Hoeffding's inequality for sub-Gaussian random variables (Vershynin, 2018, Theorem 2.6.3). This can be made arbitrarily small uniformly in k by making η small, as required.

A.6 Proof of Theorem 4.2

By Corollary 2.1(iii), it suffices to show the bound for $R^*_{\text{FLCI}}(X, C)$. We first note that any estimator $a'Y$ that does not have infinite worst-case bias must satisfy $a'w = 1$, which implies $1 \leq \|a\|_2 \cdot \|w\|_2$ by the Cauchy-Schwarz inequality, so that the variance $\sigma^2 a'a$ is bounded by $\sigma^2/\|w\|_2^2 \leq \sigma^2\eta^{-1}/n$. It therefore suffices to show that the worst-case bias is bounded by a constant times $C\sqrt{k/n}$ (for (i)), or a constant times C (for (ii)).

For part (i), let $\tilde{\gamma} = -C\eta\sqrt{k/n}Z'(ZZ')^{-1}w$. Observe

$$\text{Pen}(\gamma) = C\eta\sqrt{k/n}\sqrt{w'(ZZ')^{-1}w} \leq C\eta\|w/\sqrt{n}\|_2 \max\text{eig}((ZZ'/k)^{-1})^{1/2} \leq C.$$

Let $\tilde{\beta} = C\eta\sqrt{k/n}$. Then $w\tilde{\beta} + Z\tilde{\gamma} = 0$. Thus, $\tilde{\beta}, \tilde{\gamma}$ is observationally equivalent to the parameter vector $\beta = 0$, $\gamma = 0$, which implies that the length of any CI must be at least $C\eta\sqrt{k/n}$.

Part (ii), follows by an analogous argument, with $\tilde{\gamma} = -C\eta^{1/2}Z'(ZZ')^{-1}w$ and $\tilde{\beta} = C\eta^{1/2}$.

A.7 Proof of Theorem 4.3

Since the lower bound $c\cdot n^{-1/2}$ follows from standard efficiency bounds with finite dimensional parameters (e.g. taking the submodel where $\delta = \gamma = 0$), we show the lower bound $E_{\vartheta^*}\hat{\chi} \geq C_n \cdot c \cdot \sqrt{\log k}/\sqrt{n}$. To show this, we follow essentially the same arguments as Cai and Guo (2017, Theorem 3) and Javanmard and Montanari (2018, Proposition 4.2), noting that the required bounds on $\|\delta\|$ and $\|\gamma\|$ hold for the distributions used in the lower bound. 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 over the parameter space $\Theta(C_n, C_n \cdot K\sqrt{n/\log k}, \eta_n)$ 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 $C_n\sqrt{\log k}/\sqrt{n}$ and such that $\chi^2(f_{\pi_0}, f_{\pi_1}) \to 0$, where π_0 is the distribution that places probability one on ϑ^* given in the statement of the theorem.

To this end, we first note that we can assume $\sigma_0^2 = \sigma_{v,0}^2 = 1$ without loss of generality, since dividing Y_i and w_i by σ_0 and $\sigma_{v,0}$ leads to the same model with parameters multiplied by constants that depend only on σ_0 and $\sigma_{v,0}$.

Let π_1 be defined by a uniform prior for δ over the set with $\|\delta\|_0 = s$ and each element $\delta_j \in \{0, \nu\}$, where s and ν will be determined below. We then set the remaining parameters as deterministic functions of δ: $\beta = -\|\delta\|_2^2/(1 - \|\delta\|_2^2)$, $\gamma = (1 - \beta)\delta$, $\sigma_v^2 = 1 - \|\delta\|_2^2$ and $\sigma^2 = (1 - 2\|\delta\|_2^2)/(1 - \|\delta\|_2^2)$. We note that $\|\delta\|_2$ is constant under this prior, so that β is a unit point mass as required. This leads to the variance matrix

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

for ϑ in the support of π_1, and $\Sigma_{\vartheta^*} = I_{k+2}$ 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

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

We set $\nu = (\sqrt{c_\nu}/2) \cdot \sqrt{\log k}/\sqrt{n}$ for some $c_\nu > 0$ so that $e^{4n\nu^2} = k^{c_\nu}$. We then set s to be the greatest integer less than $C_n/\nu = (2C_n/\sqrt{c_\nu}) \cdot (\sqrt{n}/\sqrt{\log k})$. The condition that $C_n \le \sqrt{k/n} \cdot k^{-\tilde{\eta}}$ for some $\tilde{\eta} > 0$ then guarantees that $s \le k^\psi$ for some $\psi < 1/2$, so that the

above display is bounded by

$$e^{k^{2\psi-1}(1-k^{\psi-1})^{-1}}\left(1+\frac{1}{s}k^{2\psi-1}(k^{c_\nu}-1)\right)^s-1.$$

This converges to zero as required if c_ν is chosen small enough so that $2\psi+c_\nu<1$.

Finally, we note that, under π_1, $\|\delta\|_2^2=(1+o(1))s\nu^2=(1+o(1))C_n\nu=(1+o(1))\cdot C_n(\sqrt{c_\nu}/2)\cdot\sqrt{\log k}/\sqrt{n}$ and $|\beta|=\|\delta\|_2^2(1+o(1))=(1+o(1))C_n(\sqrt{c_\nu}/2)\cdot\sqrt{\log k}/\sqrt{n}$. Thus, we obtain a lower bound of $C_n\cdot c\cdot\sqrt{\log k}/\sqrt{n}$ as required.

Appendix B 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 B.1 derives the properties of a regularized estimator of the regression function $w_i\beta+z_i'\gamma$. Appendix B.2 gives conditions under which this estimator can be used to construct initial estimates of residuals in Algorithm 3.1. Appendix B.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 $k_2\to\infty$ and $k_1/n\to 0$. 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 $H_{X_1}=X_1X_1^+$ 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 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 B.1. *There exists $\eta>0$ such that, for all i and n and all $Q\in\mathcal{Q}_n$,*

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

and $1/\eta<E_Q\varepsilon_i^2$. In addition, the elements of $M_{X_1}X_2$ are bounded by some constant K_X uniformly over n.

B.1 Estimating the regression function globally

Consider the regularized regression estimator of θ, given by

$$\hat{\theta} = \underset{\vartheta}{\operatorname{argmin}} \|Y - X\vartheta\|_2^2/n + \lambda\|\vartheta_2\|_p. \tag{20}$$

In order to derive the rate of convergence $\hat{\theta}$ in Theorem B.1 below, we first give an elementary property of this estimator, following standard arguments (see Bühlmann and van de Geer (2011, Section 6.2) and van de Geer (2000, Chapter 10.1)).

Lemma B.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,

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

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

$$\|X(\hat{\theta} - \theta)\|_2^2/n = \|H_{X_1}\varepsilon\|_2^2/n + \|M_{X_1}X_2(\hat{\theta}_2 - \theta_2)\|_2^2/n, \tag{21}$$

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 B.1 to derive rates of convergence for the regularized regression estimator in eq. (20) 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 B.1 in Appendix B.3 below. We leave the analysis of $\hat{\theta}$ under such choices of λ for future research.

Theorem B.1. *Suppose that Assumption B.1 holds. Let $\hat{\theta}$ be the penalized regression estimator defined in eq. (20) with $\lambda = K_n r_q(k_2, n)$, where $K_n \to \infty$ and $r_q(k, n)$ given in eq. (15). Then*

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

Proof. By Lemma B.2 below, if we set $\lambda_0 = \lambda = K_n r_q(k_2, n)$, the condition of Lemma B.1, and hence the conclusion that $\|M_{X_1} X_2(\hat{\theta} - \theta)\|_2^2/n \leq 2K_n \|\theta_2\|_p r_q(k_2, n)$, 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. (21). □

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

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 $2k_2 \exp(-\tilde{K} \cdot K_n^2 \log k_2) = 2k_2^{1-\tilde{K}\cdot K_n^2} \to 0$ by Hoeffding's inequality for sub-Gaussian random variables (Vershynin, 2018, Thm. 2.6.3). □

B.2 Feasible CIs with unknown error distribution

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

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

B.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 $a_1, \ldots, a_n$ 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 $V_Q = \mathrm{var}_Q(\hat{\beta}_a) = \sum_{i=1}^n a_i^2 E_Q \varepsilon_i^2$. We consider coverage of the feasible bias-aware CI

$$\hat{\beta}_a \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_\Gamma(\hat{\beta}_a)/\hat{V}_a^{1/2}) \cdot \hat{V}_a^{1/2}, \tag{22}$$

where $\overline{\mathrm{bias}}_\Gamma(\hat{\beta}_a)$ is the worst-case bias, given in eq. (5). Under non-Gaussian errors, valid coverage will require conditions on the quantity

$$\mathrm{Lind}(a) = \max_{1 \le i \le n} \frac{a_i^2}{\sum_{j=1}^n a_j^2}$$

in order to invoke a Lindeberg central limit theorem. This quantity, which we refer to as 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 $\mathrm{Lind}(a) \to 0$ and the ℓ_2 rate of convergence of the estimator $X\hat{\theta}$ of the regression function.

Theorem B.2. *Suppose that, for some $\eta > 0$, $\eta \le E_Q \varepsilon_i^2$ and $E_Q|\varepsilon_i|^{2+\eta} \le 1/\eta$ 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) $\max\{\sqrt{n}c_n, 1\} \cdot \mathrm{Lind}(a) \to 0$*; and*

(ii) $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_{\theta,Q}(\|X(\hat{\theta} - \theta)\|_2 \le c_n) \to 1$.

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

$$\liminf_n \inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q\left(\beta \in \left\{\hat{\beta}_a \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_\Gamma(\hat{\beta}_a)/\sqrt{\hat{V}_a}) \cdot \sqrt{\hat{V}_a}\right\}\right) \ge 1 - \alpha. \tag{23}$$

Proof. We have

$$\frac{\hat{V}_a - V_Q}{V_Q} = \frac{\sum_{i=1}^n a_i^2(\hat{\varepsilon}_i^2 - \varepsilon_i^2)}{V_Q} + \frac{\sum_{i=1}^n a_i^2(\varepsilon_i^2 - E_Q\varepsilon_i^2)}{V_Q}. \tag{24}$$

Let $\tilde{b}_i = a_i^2 / \sum_{j=1}^n a_j^2$ so that $\max_{1\le i\le n} \tilde{b}_i = \text{Lind}(a)$. The second term in eq. (24) is bounded by $|\sum_{i=1}^n \tilde{b}_i(\varepsilon_i^2 - E_Q\varepsilon_i^2)|/\eta$. The absolute $1+\eta$ moment of this quantity is bounded by a constant times $\sum_{i=1}^n \tilde{b}_i^{1+\eta} \cdot 1/\eta^{1+\eta}$ by von Bahr and Esseen (1965). This is bounded by $\max_{1\le i\le n} \tilde{b}_i^\eta \cdot \sum_{i=1}^n \tilde{b}_i/\eta^{1+\eta} = \max_{1\le i\le n} \tilde{b}_i^\eta/\eta^{1+\eta} = \text{Lind}(a)/\eta^{1+\eta} \to 0$. The first term in eq. (24) is bounded by $\text{Lind}(a)/\eta$ times

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

For some constant K that depends only on η, we have $2\|\varepsilon\|_2 \le K\sqrt{n}$ with probability approaching one uniformly over $Q \in \mathcal{Q}_n$. Since $\|\hat{\varepsilon} - \varepsilon\|_2 = \|X(\hat{\theta} - \theta)\|_2 \le c_n$ it follows that, with probability approaching one uniformly over $\theta \in \Theta$ and $Q \in \mathcal{Q}_n$, the first term in eq. (24) is bounded by $\text{Lind}(a) \cdot (K\sqrt{n} + c_n) \cdot c_n \to 0$. It follows that for any $\delta > 0$, $\inf_{\theta\in\Theta, Q\in\mathcal{Q}_n} P_Q\left(\left|(\hat{V}_a - V_Q)/V_Q\right| < \delta\right) \to 1$. 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)). □

For the setting in Theorem B.1, condition (ii) in Theorem B.2 will hold with $c_n = \sqrt{K_n n((k_1+1)/n + C_n r_q(k_2, n))}$ for a slowly increasing constant K_n. Condition (i) in Theorem B.2 will then hold so long as $\sqrt{K_n((k_1+1)/n + C_n r_q(k_2,n))} \cdot n\,\text{Lind}(a) \to 0$. This gives the following result.

Corollary B.1. *Suppose that Assumption B.1 holds. Let $\hat{\varepsilon}$ be the residuals from the regularized regression in eq. (20), with λ given in Theorem B.1 for some $K_n \to \infty$. Then, if $\sqrt{K_n((k_1+1)/n + C_n r_q(k_2,n))} \cdot n\,\text{Lind}(a) \to 0$, the coverage result in eq. (23) holds with $\Theta = \mathbb{R}^{k_1+1} \times \{\gamma_2 \colon \|\gamma_2\|_p \le C_n\}$.*

B.2.2 Optimized weights

We now apply the results in Appendix B.2.1 to the feasible CIs based on optimized weights in Algorithm 3.1. We make two modifications relative to the baseline algorithm. First, we impose a bound on the Lindeberg weight $\text{Lind}(a)$, as described in Remark 3.1. Second, we

compute the weights using some nonrandom initial guess $\tilde{\sigma}^2$ in Step 1 of the algorithm.[11]

In the homoskedastic model with error variance $\tilde{\sigma}^2$, a FLCI centered at the linear estimator $\hat{\beta}_a = a'Y$ has length

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

where $\overline{\mathrm{bias}}_\Gamma(\hat{\beta}_a)$ is the worst-case bias of the linear estimator $\hat{\beta}_a = a'Y$. Let the weights a_b^* minimize the above display subject to the constraint $\mathrm{Lind}(a) \leq b$.

It follows immediately from Corollary B.1 that a feasible CI centered at $\hat{\beta}_{a_b^*}$ will have asymptotic coverage so long as the constraint on b is chosen appropriately.

Theorem B.3. *Suppose that Assumption B.1 holds. Let $\hat{\varepsilon}$ be the residuals from the regularized regression in eq. (20), with λ given in Theorem B.1 for some $K_n \to \infty$. Let $\Gamma = \mathbb{R}^{k_1} \times \{\gamma_2 \colon \|\gamma_2\|_p \leq C_n\}$ so that $\Theta = \mathbb{R} \times \mathbb{R}^{k_1} \times \{\gamma_2 \colon \|\gamma_2\|_p \leq C_n\}$. Consider a sequence b_n such that $\sqrt{K_n((k_1+1)/n + C_n r_q(k_2,n))} \cdot n \cdot b_n \to 0$. Then the CI in eq. (22) with $a = a_{b_n}^*$ satisfies*

$$\liminf_n \inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q\left(\beta \in \{\hat{\beta}_a \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_\Gamma(\hat{\beta}_a)/\hat{V}_a^{1/2}) \cdot \hat{V}_a^{1/2}\}\right) \geq 1 - \alpha.$$

Imposing a condition on the Lindeberg weight can in general affect the performance of the CI. The following theorem shows that the optimal rate of convergence derived in Theorem 4.1 will still be obtained if the constraint b_n on the Lindeberg weight is chosen appropriately.

Theorem B.4. *Suppose the conditions of Theorem B.3 hold, with $k_1 = 0$ so that $\Gamma = \{\gamma : \|\gamma\|_p \leq C_n\}$. Suppose also that, for some $\eta > 0$, the design matrix X is in the set $\mathcal{E}_n(\eta)$ defined in Section 4.1 for large enough n, and that for some sequence b_n, $\max_{1 \leq i \leq n}(w_i - z_i'\delta)^2/n = o(b_n)$ 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 $\mathcal{E}_n(\eta)$ in Section 4.1. Then there exists a constant K such that the CI in eq. (22) with $a = a_{b_n}^*$ satisfies*

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

Proof. Let $v_i = w_i - z_i'\delta$ and $\tilde{a}_i = v_i / \sum_{j=1}^n v_i w_j$ where δ is given in the definition of $\mathcal{E}_n(\eta)$. Note that $\mathrm{Lind}(\tilde{a}) = \max_{1 \leq i \leq n}(w_i - z_i'\delta)^2 / \sum_{j=1}^n (w_j - z_j'\delta)^2$. 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

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

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 4.1 that the right-hand side of the above display is bounded by a constant times $n^{-1/2} + C_n r_q(k,n)$. Furthermore, by the uniform consistency of $\hat{V}$ (which follows from Theorem B.2) 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. □

While Theorem B.3 imposes an upper bound $b_n = o(n^{-1}(K_n((k_1+1)/n + C_n r_q(k_2,n)))^{-1/2})$ on b_n, Theorem B.4 imposes 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 the conditions of both Theorem B.3 and Theorem B.4. For example, if $\max_{1\leq i\leq n}(w_i - z_i'\delta)^2$ is bounded, then the conditions of Theorem B.4 will hold with $b_n = K_n/n$ for a slowly increasing sequence K_n. Taking the same sequence K_n in the choice of λ in Theorem B.1 for simplicity, the condition in Theorem B.3 becomes

$$\sqrt{K_n((k_1+1)/n + C_n r_q(k_2,n))} \cdot n \cdot b_n = \sqrt{K_n((k_1+1)/n + C_n r_q(k_2,n))} \cdot K_n \to 0.$$

Since $C_n r_q(k_2,n)$ 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 $n \to \infty$.

B.3 Lower CIs for C

We present a lower CI for the regularity parameter 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. (20), 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

$$\underline{\hat{C}} = \sup_{\lambda > \lambda^*_\alpha} \frac{\lambda - \lambda^*_\alpha}{\lambda + \lambda^*_\alpha} \|\hat{\theta}_2(\lambda)\|_p. \tag{25}$$

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 $\hat{\underline{C}}$ is feasible.

Theorem B.5. *Consider $\hat{\underline{C}}$ in eq. (25) 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 $P_{\beta,\gamma_1,\gamma_2}(C \in [\hat{\underline{C}}, \infty)) \geq 1-\alpha$.*

Proof. It follows from Lemma B.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. □

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 B.1 for some slowly increasing K_n. This leads to the moderate deviations critical value $\hat{\lambda}^*_\alpha$, which sets

$$\alpha = \sum_{j=1}^{k_2} 2\Phi(-\hat{\lambda}^*_\alpha/\hat{V}_j^{1/2}). \tag{26}$$

Remark B.1. The analysis in Theorem B.1 of the regularized regression estimator in eq. (20) 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. (26). 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. (20) after using some initial penalty choice satisfying the conditions of Theorem B.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 unknown error distribution. Belloni 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 et al. (2013).

Theorem B.6. *Suppose Assumption B.1 holds with $p = 1$ and that and $\frac{1}{n}\sum_{i=1}^n \tilde{x}_{ij}^2 \geq \eta$ for $j = 1, \ldots, k$ for all n, where $\tilde{x}_{ij} = (M_{X_1} X_2)_{ij}$. Let $\hat{\lambda}_\alpha^*$ be given in eq. (25) with $\hat{V}_j$ formed using residuals $\hat{\varepsilon}$ from the regularized regression in eq. (20) with penalty λ chosen as in Theorem B.1 for some $K_n \to \infty$ with $K_n(k_1/n + (C_n+1)\sqrt{\log k_2}/\sqrt{n}) \cdot (\log k_2)^2 \to 0$. Then*

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

In particular, letting $\hat{\underline{C}}$ be given in eq. (25) with λ_α^* given by $\hat{\lambda}_\alpha^*$, we have

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

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

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

which holds with probability approaching one uniformly over $Q \in \mathcal{Q}_n$ when $\tilde{K}$ is large enough, this is bounded by $(2K_X/n)^2(\tilde{K}\sqrt{n} + \sqrt{nK_n}(k_1/n + 2C_n\sqrt{\log k_2}/\sqrt{n})^{1/2}) \cdot \sqrt{nK_n}(k_1/n + 2C_n\sqrt{\log k_2}/\sqrt{n})^{1/2}$. 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, $\max_{1 \leq j \leq k_2} \left|\hat{V}_j - \tilde{V}_j\right| / V_{Q,j}$ is bounded by

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

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

In addition, note that $(\tilde{V}_j - V_{Q,j})/V_{Q,j} = \sum_{i=1}^n \tilde{a}_{ij}(\varepsilon_i - E_Q\varepsilon_i)/n$, where $\tilde{a}_{ij} = \tilde{x}_{ij}^2/(nV_{j,Q}) \leq K_X^2\tilde{\eta}^{-1}$ 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 $2\exp(-cn\delta^2)$ for some constant c that depends only on K_X, $\tilde{\eta}$ and η. Thus, for any sequence δ_n, we have $P_Q(\max_{1 \leq j \leq k_2} |\tilde{V}_j - V_{Q,j}|/V_{Q,j} \geq \delta) \leq 2k_2\exp(-cn\delta_n^2)$, 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 differentiability 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 $K_n^{1/2}(k_1/n + (C_n + 1)\sqrt{\log k_2}/\sqrt{n})^{1/2}$, 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 $\|\gamma_2\| \leq C_n$. 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 $\|\gamma_2\|_p \leq C_n$ and any sequence $Q_n \in \mathcal{Q}_n$ is bounded by a term that converges to zero plus

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

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

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

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

Using standard calculations and the fact that $nV_{Q_n,j}$ is uniformly bounded from above and below, we have $(\log k_2)/K \leq \lambda^2_{\alpha,n}/V_{Q_n,j} \leq K \log k_2$ for some constant K. Thus, the right-hand side of the above display converges to 1 uniformly over n and $1 \leq j \leq k$ so long as $c_n \log k_2 \to 0$, 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 $\bar{t}_n/n^{1/6} \to 0$, 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 $\limsup_{n\to\infty} \max_{1\leq j\leq k_2} B_{n,j} \leq 1$ so long as $(\log k_2)^{3/2}/\sqrt{n} \to 0$, which is guaranteed by the conditions of the theorem. □

References

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288.

Armstrong, T. and Kolesár, M. (2021a). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177.

Armstrong, T. and Kolesár, M. (2021b). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108.

Armstrong, T. B. and Chan, H. P. (2016). Multiscale adaptive inference on conditional moment inequalities. *Journal of Econometrics*, 194(1):24–43.

Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Armstrong, T. B. and Kolesár, M. (2016). Optimal inference in a class of regression models. arXiv:1511.06028v2.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 1(4):791–896.

Bühlmann, P. and van de Geer, S. A. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin, Heidelberg.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646.

Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Working Paper 330, National Bureau of Economic Research, Cambridge, MA.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829.

Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. J. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, New York, NY, third edition.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, NY.

Goldberg, M. (2017). Continuity of seminorms on finite-dimensional vector spaces. *Linear Algebra and its Applications*, 515:175–179.

Goldsmith-Pinkham, P., Hull, P., and Kolesár, M. (2022). Contamination bias in linear regressions.

Heckman, N. E. (1988). Minimax estimates in a semiparametric model. *Journal of the American Statistical Association*, 83(404):1090–1096.

Ibragimov, I. A. and Khas'minskii, R. Z. (1985). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32.

Imbens, G. and Wager, S. (2019). Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278.

Imbens, G. W., Rubin, D. B., and Sacerdote, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, 91(4):778–794.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *Annals of Statistics*, 46(6A):2593–2622.

Kolesár, M. and Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108(8):2277–2304.

Kwon, K. and Kwon, S. (2020). Inference in regression discontinuity designs under monotonicity. arXiv: 2011.14216.

Li, C. M. and Müller, U. K. (2021). Linear regression with many controls of limited explanatory power. *Quantitative Economics*, 12(2):405–442.

Li, K.-C. (1982). Minimaxity of the method of regularization of stochastic processes. *The Annals of Statistics*, 10(3):937–942.

Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *The Annals of Statistics*, 23(3):824–835.

Muralidharan, K., Romero, M., and Wüthrich, K. (2023). Factorial designs, model selection, and (incorrect) inference in randomized experiments. *The Review of Economics and Statistics*, forthcoming.

Noack, C. and Rothe, C. (2021). Bias-aware inference in fuzzy regression discontinuity designs. arXiv: 1906.04631.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, forthcoming.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK.

van de Geer, S. A., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, UK, first edition.

von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Annals of Mathematical Statistics*, 36(3):808–818.

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia, PA.

Wüthrich, K. and Zhu, Y. (2021). Omitted variable bias of Lasso-based inference methods: A finite sample analysis. *The Review of Economics and Statistics*, forthcoming.

Yosida, K. (1995). *Functional Analysis*. Springer-Verlag, Berlin, reprint of 6th edition.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.