Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of high-dimensional M-estimators

Denis Chetverikov
Jesper R.-V. Sørensen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP03/22

Analytic and Bootstrap-after-Cross-Validation Methods for Selecting Penalty Parameters of High-Dimensional M-Estimators[*]

Denis Chetverikov[†] Jesper R.-V. Sørensen[‡]

January 11, 2022

Abstract

We develop two new methods for selecting the penalty parameter for the ℓ^1-penalized high-dimensional M-estimator, which we refer to as the analytic and bootstrap-after-cross-validation methods. For both methods, we derive nonasymptotic error bounds for the corresponding ℓ^1-penalized M-estimator and show that the bounds converge to zero under mild conditions, thus providing a theoretical justification for these methods. We demonstrate via simulations that the finite-sample performance of our methods is much better than that of previously available and theoretically justified methods.

Keywords: Penalty parameter selection, penalized M-estimation, high-dimensional models, sparsity, cross-validation, bootstrap.

1 Introduction

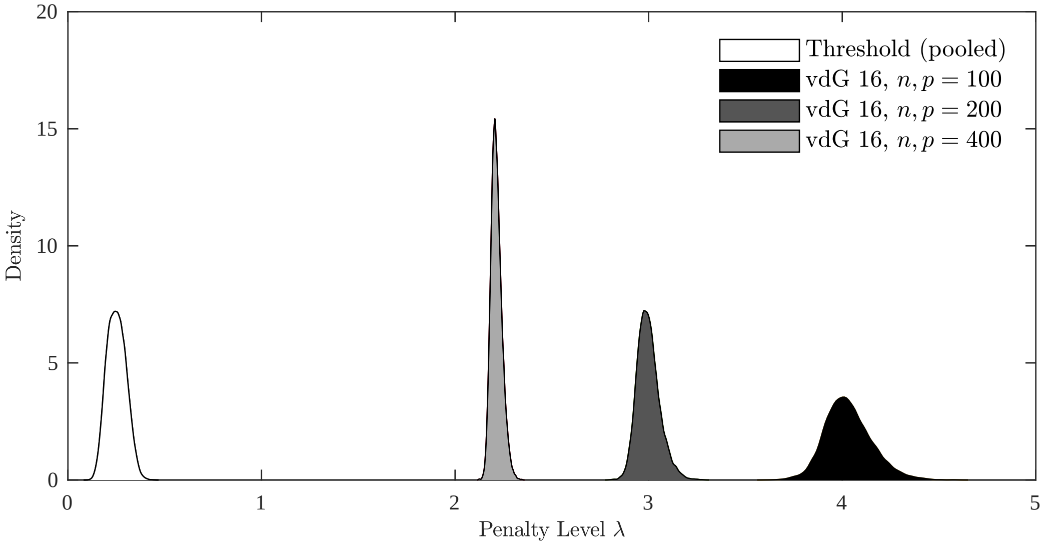
High-dimensional models have attracted substantial attention both in the econometrics and in the statistics/machine learning literature, e.g. see Belloni et al. (2018a) and Hastie et al. (2015), and ℓ^1-penalized estimators have emerged among the most useful methods for learning parameters of such models. However, implementing these estimators requires a choice of the penalty parameter and with few notable exceptions, e.g. ℓ^1-penalized linear mean and

[*] Date: January 2022. We thank Richard Blundell, Victor Chernozhukov, Bo Honoré, Whitney Newey, Joris Pinkse, Azeem Shaikh and Sara van de Geer for their insightful comments and discussions. Chetverikov's work was supported by NSF Grant SES - 1628889.

[†] Department of Economics, UCLA; e-mail: chetverikov@econ.ucla.edu.

[‡] Department of Economics, University of Copenhagen; e-mail: jrvs@econ.ku.dk.

Figure 1.1: Probability density functions of the smallest value (threshold) of the penalty parameter leading to all-zero estimated parameters and of the value of the penalty parameter obtained from the van de Geer (2016) method (vdG 16) in the setting of the ℓ^1-penalized logit estimator. The figure demonstrates that the van de Geer penalty parameter value substantially exceeds the threshold value for the samples considered and thus yields the trivial, all-zero, estimates; see Section 7 for details.



quantile regression estimators, the choice of this penalty parameter in practice often remains unclear. Some methods, such as cross-validation and related sample splitting methods, tend to perform well in simulations but, as we discuss below, generally lack a sufficient theoretical justification. Other methods, such as those discussed in van de Geer (2016), are supported by a sound asymptotic theory but tend to perform poorly in moderate samples of practical relevance, often leading to trivial estimates, with all estimated parameters being exactly zero; see Figure 1.1 for a demonstration in the case of the ℓ^1-penalized logit estimator. In this paper, we deal with these problems and (i) propose two new methods for choosing penalty parameters in the context of ℓ^1-penalized M-estimation, (ii) derive the supporting asymptotic theory, and (iii) demonstrate that our methods perform well in moderate samples.

We consider a model where the true value θ_0 of some parameter θ is given by the solution to an optimization problem

$$\theta_0 = \operatorname*{argmin}_{\theta \in \Theta} \mathrm{E}[m(X^\top \theta, Y)], \tag{1.1}$$

where $m : \mathbf{R} \times \mathcal{Y} \to \mathbf{R}$ is a known (potentially nonsmooth) loss function that is convex in its first argument, $X = (X_1, \dots, X_p)^\top \in \mathcal{X} \subseteq \mathbf{R}^p$ a vector of candidate regressors, $Y \in \mathcal{Y}$ one or more outcome variables, and $\Theta \subseteq \mathbf{R}^p$ a convex parameter space. Prototypical loss functions are square-error loss and negative log-likelihood but the framework (1.1) also covers many other cross-sectional models and associated modern as well as classical estimation approaches including logit and probit models, logistic calibration (Tan, 2017) covariate balancing (Imai and Ratkovic, 2014), and expectile regression (Newey and Powell, 1987). It also subsumes approaches to estimation of panel-data models such as the fixed-effects/conditional logit (Rasch, 1960) and trimmed least-absolute-deviations and least-squares methods for censored regression (Honoré, 1992), and partial likelihood estimation of heterogeneous panel models for duration (Chamberlain, 1985). We provide details on these examples in Section 2.[1]

For the purpose of estimation, we assume access to a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent observations from the distribution of the pair (X, Y), where the number p of candidate regressors in each $X_i = (X_{i1}, \ldots, X_{ip})^\top$ may be (potentially much) larger than the sample size n, meaning that we cover high-dimensional models. Following the literature on high-dimensional models, we also assume that the vector $\theta_0 = (\theta_{01}, \ldots, \theta_{0p})^\top$ is sparse in the sense that the number $s := \sum_{j=1}^p \mathbf{1}(\theta_{0j} \neq 0)$ of relevant regressors is much smaller than n.[2] With this sparsity assumption in mind, we study the ℓ^1-penalized M-estimator

$$\widehat{\theta}(\lambda) \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n m(X_i^\top \theta, Y_i) + \lambda \|\theta\|_1 \right\}, \tag{1.2}$$

where $\|\theta\|_1 = \sum_{j=1}^p |\theta_j|$ denotes the ℓ^1-norm of θ, and $\lambda \geqslant 0$ is a penalty parameter.

Implementing the estimator $\widehat{\theta}(\lambda)$ requires us to choose λ. To do so, we first extend the deterministic bound from Belloni and Chernozhukov (2011b) obtained for ℓ^1-penalized quantile regression to the general setting of ℓ^1-penalized M-estimators (1.2). In particular, we show that for an arbitrary choice of $c_0 > 1$, there exists a constant C, depending on the distribution of the pair (X, Y) and c_0, such that under mild regularity conditions, the event

$$\lambda \geqslant c_0 \max_{1 \leqslant j \leqslant p} \left| \frac{1}{n} \sum_{i=1}^n m_1'(X_i^\top \theta_0, Y_i) X_{ij} \right| \tag{1.3}$$

implies both

$$\|\widehat{\theta}(\lambda) - \theta_0\|_2 \leqslant C\sqrt{s} \left(\lambda + \sqrt{\frac{\ln(pn)}{n}} \right) \quad \text{and} \quad \|\widehat{\theta}(\lambda) - \theta_0\|_1 \leqslant Cs \left(\lambda + \sqrt{\frac{\ln(pn)}{n}} \right), \tag{1.4}$$

where $m_1'(t, y) := (\partial/\partial t)\, m(t, y)$ denotes the derivative of the loss function m with respect to its first argument (or a subgradient, if m is not differentiable). These bounds suggest the following principle: choose λ as small as possible subject to the event (1.3) occurring with

[1] We consider the single-index setup primarily for notational convenience. Section 6 discusses changes needed to accommodate richer modeling frameworks, including settings with multiple indices. Multiple indices occur naturally in, e.g., multinomial models such as the multinomial and conditional logit models.

[2] We take $s \geqslant 1$ throughout. This assumption is innocuous as we may always redefine s as $\max\{1, s\}$.

high probability. We therefore wish to set $\lambda = c_0 q(1-\alpha)$, where

$$q\left(1-\alpha\right) := \left(1-\alpha\right)\text{-quantile of } \max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n} m_1'(X_i^\top\theta_0, Y_i)X_{ij}\right|, \tag{1.5}$$

for some small user-specified probability tolerance level $\alpha \in (0,1)$, e.g. $\alpha = .1$. This choice, however, is typically infeasible since the random variable in (1.5) depends on the unknown θ_0. We thus have a vicious circle: to choose λ, we need an estimator of θ_0, but to estimate θ_0, we need to choose λ. In this paper, we offer two solutions to this problem, which constitute our key contributions.

To obtain our first solution, we show that whenever the loss function m is Lipschitz continuous with respect to its first argument, we can apply results from high-dimensional probability theory to derive an upper bound, say $\overline{q}(1-\alpha)$, on $q(1-\alpha)$ that does not depend on θ_0 and can be computed analytically from the available dataset. We can then set $\lambda = c_0\overline{q}(1-\alpha)$, which we refer to as the *analytic method*. This method is computationally straightforward and, as we demonstrate by means of example, has several applications. On the other hand, it is not universally applicable as the loss function may or may not be Lipschitz continuous. For example, it works for the logit model but not for the probit model. Moreover, this method is somewhat conservative, in the sense that it yields a penalty satisfying $\lambda > c_0 q\left(1-\alpha\right)$.

To obtain our second solution, we show that even though the estimator $\widehat{\theta}(\lambda)$ based on λ chosen by cross-validation or its variants is generally difficult to analyze, it can be used to construct provably good (in a sense to be made clear later) estimators of the random vectors $m_1'(X_i^\top\theta_0, Y_i)X_i$. We are then able to derive an estimator, say $\widehat{q}(1-\alpha)$, of $q(1-\alpha)$ via bootstrapping, as discussed in Belloni et al. (2018a), and to set $\lambda = c_0\widehat{q}(1-\alpha)$, which we refer to as the *bootstrap-after-cross-validation method*. This method is computationally somewhat more demanding than the analytic method, but it is generally much more widely applicable and nonconservative in the sense that it gives λ such that $\lambda \approx c_0 q(1-\alpha)$.

Drawing on simulations from a simple logit model, we illustrate the potential of our analytic and bootstrap-after-cross-validation methods. Our simulations indicate that, while both methods lead to useful estimates of θ_0 in the model (1.1) even in moderate samples, there may be significant gains from using the bootstrap-after-cross-validation method, even if the analytic method is also available. Both our methods substantially outperform the theoretically supported choice of λ discussed in van de Geer (2016). Moreover, our penalty selection methods are not dominated by cross-validation, which is popular in practice.

A key feature of our methods is that they yield bounds on both ℓ^1 and ℓ^2 estimation errors. In contrast, sample splitting methods typically yield bounds only for the excess risk

$\mathrm{E}_{X,Y}[m(X^\top\widehat{\theta}(\lambda), Y) - m(X^\top\theta_0, Y)]$, e.g. see Lecue and Mitchell (2012). These bounds can be translated into the ℓ^2 estimation error $\|\widehat{\theta}(\lambda) - \theta_0\|_2$, but it is not clear how to convert them into bounds on the ℓ^1 estimation error $\|\widehat{\theta}(\lambda) - \theta_0\|_1$.[3] A bound of the ℓ^1 type is crucial when we are interested in estimating dense functionals $a'\theta_0$ of θ_0 with $a \in \mathbf{R}^p$ being a vector of loadings with many nonzero components; see Belloni et al. (2018a) for details. Moreover, bounds on the ℓ^1 estimation error are needed to perform inference on components of θ_0 via double machine learning, as in Belloni et al. (2018b). When λ is selected using cross-validation, ℓ^1 and ℓ^2 estimation error bounds are typically both unknown. The only exception we are aware of is the linear mean regression model estimated using the LASSO. The bounds have for this special case been derived in Chetverikov et al. (2016) and Miolane and Montanari (2018), but their bounds are less sharp than those provided here.

The literature on learning parameters of high-dimensional models via ℓ^1-penalized M-estimation is large. Instead of listing all existing papers, we therefore refer the interested reader to the excellent textbook treatment in Wainwright (2019) and focus here on only a few key references. van de Geer (2008, 2016) derives bounds on the estimation errors of general ℓ^1-penalized M-estimators (1.2) and provides some choices of the penalty parameter λ. As discussed above, however, her penalty formulae give values of λ that are so large that the resulting estimators are typically trivial in moderate samples, with all coefficients being exactly zero (cf. Figure 1.1). Because of this issue, van de Geer (2008) remarks that her results should only be seen as an indication that her theory has something to say about finite sample sizes, and that other methods to choose λ should be used in practice. Negahban et al. (2012) develop error guarantees in a very general setting, and when specialized to our setting (1.2) their results become quite similar to our statement that the bounds (1.4) hold under the event (1.3). The same authors also note that a challenge to using these results in practice is that the random variable in (1.3) is usually impossible to compute because it depends on the unknown vector θ_0. It is exactly this challenge that we overcome in this paper. Belloni and Chernozhukov (2011b) study high-dimensional quantile regression and note that the distribution of the random variable in (1.3) is in this case pivotal, making the choice of the penalty parameter simple. However, quantile regression is the only setting we are aware of in which the distribution of the random variable in (1.3) is pivotal.[4] Finally, Ninomiya and Kawano (2016) consider information criteria for the choice of the penalty parameter λ but

[3] Any two norms on a finite-dimensional space are equivalent. However, the equivalence constants generally depend on the dimension (here p), which makes translation of error bounds for one norm into another a nontrivial manner when the dimension is growing.

[4] With a known censoring propensity, the linear programming estimator of Buchinsky and Hahn (1998) for censored quantile regression boils down to a variant of quantile regression and, therefore, leads to pivotality of the right-hand side of (1.3). However, known censoring propensity seems like a very special case.

focus on fixed-p asymptotics, thus precluding high-dimensional models.

The rest of the paper is organized as follows. In Section 2 we provide a portfolio of examples that constitute possible applications of our methods. We refer to several of these examples in later sections. In Section 3 we develop bounds on the estimation error of the ℓ^1-penalized M-estimator, which motivate our methods to choose the penalty parameter. We introduce and analyze the analytic method in Section 4 and the bootstrap-after-cross-validation method in Section 5. We discuss how these methods generalize to modeling frameworks richer than (1.1) in Section 6. In Section 7 we illustrate our methods via a simulation study and compare them with existing methods. We defer all proofs to the appendix, where we also provide implementation details and low-level conditions sufficient for assumptions made in the main text.

Notation

Throughout $W_i := (X_i, Y_i), i \in \{1, \ldots, n\}$, denotes n independent copies of a random vector $W := (X, Y) \in \mathcal{W}$. The distribution P of W, as well as the dimension p of the vector X and the number of nonzero components of the vector θ_0 may change with the sample size n, but we suppress this potential dependence. $\mathrm{E}[f(W)]$ denotes the expectation of a function f of W computed with respect to P, and $\mathbb{E}_n[f(W_i)] := n^{-1} \sum_{i=1}^n f(W_i)$ abbreviates the sample average. When only a nonempty subset $I \subsetneq \{1, \ldots, n\}$ is in use, we write $\mathbb{E}_I[f(W_i)] := |I|^{-1} \sum_{i \in I} f(W_i)$ for the subsample average. For a set of indices $I \subseteq \{1, \ldots, n\}$, I^c denotes the elements of $\{1, \ldots, n\}$ not in I. Given a vector $\delta \in \mathbf{R}^p$ and a nonempty set of indices $J \subseteq \{1, \ldots, p\}$, we let δ_J denote the vector in $\mathbf{R}^p$ with coordinates given by $\delta_{Jj} = \delta_j$ if $j \in J$ and zero otherwise. We denote its ℓ^q norms, $q \in [1, \infty]$, by $\|\delta\|_q$. The nonnegative and strictly positive reals are denoted $\mathbf{R}_+$ and $\mathbf{R}_{++}$, respectively. We abbreviate $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$, and take $n \geqslant 3$, $p \geqslant 2$, and $s \geqslant 1$ throughout. We introduce more notation as needed in the appendix.

2 Examples

In this section we discuss a variety of models that fit into the M-estimation framework (1.1) with the loss function $m(t, y)$ being convex in its first argument. (See also Section 6.2 for examples involving multiple indices.) We include models for cross-sectional data (Examples 1–5), panel data (Examples 6 and 7) and panel data for duration (Example 8). The examples cover both discrete and continuous outcomes in likelihood and nonlikelihood settings with smooth as well as kinked loss functions.

Example 1 (**Binary Response Model**)**.** A relatively simple model fitting our framework is the *binary response model*, i.e. a model for an outcome $Y \in \{0,1\}$ with

$$\mathrm{P}(Y=1|X) = F(X^\top \theta_0),$$

for a known cumulative distribution function (CDF) $F : \mathbf{R} \to [0,1]$. The log-likelihood of this model yields the following loss function:

$$m(t,y) = -y \ln F(t) - (1-y)\ln(1-F(t)). \tag{2.1}$$

The *logit* model arises here by setting $F(t) = 1/(1+\mathrm{e}^{-t}) =: \Lambda(t)$, the standard logistic CDF, and the loss function reduces in this case to

$$m(t,y) = \ln\left(1+\mathrm{e}^t\right) - yt. \tag{2.2}$$

The *probit* model arises by setting $F(t) = \int_{-\infty}^t (2\pi)^{-1/2}\, \mathrm{e}^{-u^2/2} \mathrm{d}u =: \Phi(t)$, the standard normal CDF, and the loss function in this case becomes

$$m(t,y) = -y \ln \Phi(t) - (1-y)\ln(1-\Phi(t)). \tag{2.3}$$

Note that the loss functions in both (2.2) and (2.3) are convex in t.

More generally, any binary response model with both F and $1-F$ being log-concave leads to a loss (2.1) that is convex in t. For these log-concavities it suffices that F admits a probability density function (PDF) $f = F'$, which is itself log-concave (Pratt, 1981, Section 5). Both the standard logistic and standard normal PDFs are log-concave. Also, $\ln f$ is concave whenever $f(t) \propto \mathrm{e}^{-|t|^a}$ for some $a \geqslant 1$ or $f(t) \propto t^{b-1}\mathrm{e}^{-t}$ for $t \geqslant 0$ and some $b \geqslant 1$, the extreme cases being the Laplace and exponential distributions, respectively. Other examples of distributions for which f is log-concave can be found in the Gumbel, Weibull, Pareto and beta families (Pratt, 1981, Section 6). A t-distribution with $0 < \nu < \infty$ degrees of freedom (the standard Cauchy arising from $\nu = 1$) does not have a log-concave density. However, both its CDF and complementary CDF are log-concave (*ibid.*). □

Example 2 (**Ordered Response Model**)**.** Consider the *ordered response model*, i.e. a model for an outcome $Y \in \{0,1,\ldots,J\}$ with

$$\mathrm{P}(Y=j|X) = F(\alpha_{j+1} - X^\top\theta_0) - F(\alpha_j - X^\top\theta_0), \quad j \in \{0,1,\ldots,J\},$$

for a known CDF $F : \mathbf{R} \to [0,1]$ and known cut-off points $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_J <$

$\alpha_{J+1} = +\infty$. (We here interpret $F(-\infty)$ as zero and $F(+\infty)$ as one.) The log-likelihood of this model yields the loss function

$$m(t, y) = -\sum_{j=0}^{J} \mathbf{1}(y = j) \ln (F(\alpha_{j+1} - t) - F(\alpha_j - t)), \tag{2.4}$$

that is convex in t for any distribution F admitting a log-concave PDF $f = F'$ (Pratt, 1981, Section 3). See Example 1 for specific distributions satisfying this criterion. □

Example 3 (**Logistic Calibration**). In the setting of average treatment effect estimation under a conditional independence assumption with a high-dimensional vector of controls, consider the *logit propensity score model*

$$\mathrm{P}(Y = 1|X) = \Lambda(X^\top \theta_0), \tag{2.5}$$

where $Y \in \{0, 1\}$ is a treatment indicator, X a vector of controls, and Λ the logistic CDF. Using (1.1), θ_0 can be identified with the logistic loss function in (2.2). However, as shown by Tan (2017), θ_0 can also be identified using (1.1) with the logistic *calibration* loss

$$m(t, y) = y\mathrm{e}^{-t} + (1 - y)t, \tag{2.6}$$

which is convex in t as well. As demonstrated by Tan (2017), using this alternative loss function gives substantial advantages: it leads to average treatment effect estimators that enjoy particularly nice robustness properties. Specifically, under some conditions, these treatment effect estimators remain root-n consistent and asymptotically normal even if the model for the outcome regression function is misspecified (*ibid.*). □

Example 4 (**Logistic Balancing**). In the same setting as that of the previous example, the *covariate balancing approach* (Imai and Ratkovic, 2014) amounts to specifying a parametric model for the treatment indicator $Y \in \{0, 1\}$,

$$\mathrm{P}(Y = 1|X) = F(X^\top \theta_0)$$

and ensuring covariate balance in the sense that

$$\mathrm{E}\left[\left\{\frac{Y}{F(X^\top \theta_0)} - \frac{1 - Y}{1 - F(X^\top \theta_0)}\right\} X\right] = \mathbf{0}.$$

Balancing here amounts to enforcing a collection of moment conditions and is therefore naturally studied in a generalized method of moments (GMM) framework. However, specifying

F to be the logistic CDF Λ, covariate balancing can be achieved via M-estimation of θ_0 based on the loss function

$$m(t, y) = (1 - y)\,\mathrm{e}^t + y\mathrm{e}^{-t} + (1 - 2y)\,t,$$

which is also convex in t. (See Tan (2017) for details.) □

Example 5 (Expectile Model). Newey and Powell (1987) study the conditional τ^{th} *expectile model* $\mu_\tau(Y|\, X) = X^\top \theta_0$, where $\tau \in (0, 1)$ is a known number, and propose the *asymmetric least squares* estimator of θ_0 in this model. This estimator can be understood as an M-estimator with the loss function

$$m(t, y) = \rho_\tau(y - t)\,, \tag{2.7}$$

where $\rho_\tau : \mathbf{R} \to \mathbf{R}$ is the piecewise quadratic and continuously differentiable function defined by

$$\rho_\tau(u) = |\tau - \mathbf{1}(u < 0)|\, u^2 = \begin{cases} (1-\tau)\,u^2, & u < 0, \\ \tau u^2, & u \geqslant 0, \end{cases}$$

a smooth analogue of the 'check' function known from the quantile regression literature. This estimator can also be interpreted as a maximum likelihood estimator when model disturbances arise from a normal distribution with unequal weights placed on positive and negative disturbances (Aigner et al., 1976). Note that $m(\cdot, y)$ in (2.7) is convex but not twice differentiable (at zero) unless $\tau = 1/2$. □

Example 6 (Panel Logit Model). Consider the *panel logit model*

$$\mathrm{P}(Y_\tau = 1|X, \gamma, Y_0, \ldots, Y_{\tau-1}) = \Lambda(\gamma + X_\tau^\top \theta_0), \quad \tau = 1, 2,$$

where $Y = (Y_1, Y_2)^\top$ is a pair of outcome variables, $X = (X_1^\top, X_2^\top)^\top$ is a vector of regressors, and γ is a unit-specific unobserved fixed effect. Rasch (1960) shows that θ_0 in this model can be identified by $\theta_0 = \operatorname{argmin}_{\theta \in \mathbf{R}^p} \mathrm{E}[m((X_1 - X_2)^\top \theta, Y)]$, where

$$m(t, y) = \mathbf{1}(y_1 \neq y_2)\left[\ln\left(1 + \mathrm{e}^t\right) - y_1 t\right], \tag{2.8}$$

which is convex in t.[5] □

Example 7 (Panel Censored Model). Consider the *panel censored model*

$$Y_\tau = \max\left(0, \gamma + X_\tau^\top \theta_0 + \varepsilon_\tau\right), \quad \tau = 1, 2,$$

[5] See also Chamberlain (1984, Section 3.2) and Wooldridge (2010, Section 15.8.3).

where $Y = (Y_1, Y_2)^\top \in \mathbf{R}_+^2$ is a pair of outcome variables, $(X_1^\top, X_2^\top)^\top$ is a vector of regressors, γ is a unit-specific unobserved fixed effect, and ε_1 and ε_2 are unobserved error terms, which may or may not be centered. Honoré (1992) shows that under certain conditions, including exchangeability of ε_1 and ε_2 conditional on (X_1, X_2, γ), θ_0 in this model can be identified by $\theta_0 = \operatorname{argmin}_{\theta \in \mathbf{R}^p} \mathrm{E}[m(X^\top\theta, Y)]$, with $X = X_1 - X_2$ and m being the *trimmed loss* function

$$m(t, y) = \begin{cases} \Xi(y_1) - (y_2 + t)\,\xi(y_1), & t \leqslant -y_2, \\ \Xi(y_1 - y_2 - t), & -y_2 < t < y_1, \\ \Xi(-y_2) - (t - y_1)\,\xi(-y_2), & y_1 \leqslant t, \end{cases} \tag{2.9}$$

and either $\Xi = |\cdot|$ or $\Xi = (\cdot)^2$ and ξ its derivative (when defined).[6] These choices lead to *trimmed least absolute deviations* (LAD) and *trimmed least squares* (LS) estimators, respectively, both of which are based on loss functions convex in t. Here, $\Xi = |\cdot|$ leads to a nondifferentiable loss. □

Example 8 (**Panel Duration Model**)**.** Consider the *panel duration model* with a log-linear specification:

$$\ln h_\tau(y) = X_\tau^\top \theta_0 + h_0(y), \quad \tau = 1, 2,$$

where h_τ denotes the hazard for spell τ and both h_0 and h_τ are allowed to be unit-specific. This model is a special case of the duration models studied in Chamberlain (1985, Section 3.1). Chamberlain presumes that the spells Y_1 and Y_2 are (conditionally) independent of each other and shows that the partial log-likehood contribution is[7]

$$\theta \mapsto \mathbf{1}(Y_1 < Y_2) \ln \Lambda((X_1 - X_2)^\top \theta) + \mathbf{1}(Y_1 \geqslant Y_2) \ln\left(1 - \Lambda((X_1 - X_2)^\top \theta)\right).$$

The implied loss function

$$m(t, y) = \ln\left(1 + \mathrm{e}^t\right) - \mathbf{1}(y_1 < y_2)\, t \tag{2.10}$$

is of the logit form (see Example 1), hence convex in t. With more than two completed spells, the partial log-likelihood takes a conditional-logit form (*ibid.*), and the resulting loss is therefore still a convex function (albeit involving multiple indices). □

[6] When $\Xi = |\cdot|$, we set $\xi(0) := 0$ to make (2.9) consistent with formulas in Honoré (1992).

[7] See also Lancaster (1992, Chapter 9, Section 2.10.2).

3 Nonasymptotic Bounds on Estimation Error

In this section, we derive bounds on the error of the ℓ^1-penalized M-estimator (1.2) in the ℓ^1 and ℓ^2 norms. The argument reveals which quantities one needs to control in order to ensure good behavior of the estimator, motivating the choice of the penalty parameter λ in the following sections. We split the section into two subsections. In Section 3.1 we derive bounds via an empirical error function. In Section 3.2 we derive a bound on the empirical error function itself.

3.1 Bounds via Empirical Error Function

Denote

$$M(\theta) := \mathrm{E}[m(X^\top\theta, Y)] \quad \text{and} \quad \widehat{M}(\theta) := \mathbb{E}_n[m(X_i^\top\theta, Y_i)], \quad \theta \in \Theta,$$

Also, let

$$T := \{j \in \{1, \dots, p\}; \theta_{0j} \neq 0\}$$

and for any $\widetilde{c} > 1$, let $\mathcal{R}(\widetilde{c})$ denote the *restricted set*

$$\mathcal{R}(\widetilde{c}) := \{\delta \in \mathbf{R}^p; \|\delta_{T^c}\|_1 \leqslant \widetilde{c}\|\delta_T\|_1 \text{ and } \theta_0 + \delta \in \Theta\}. \tag{3.1}$$

In addition, fix $c_0 > 1$, and define the (random) *empirical error function* $\epsilon : \mathbf{R}_+ \to \mathbf{R}_+$ by

$$\epsilon\left(u\right) := \sup_{\substack{\delta \in \mathcal{R}(\bar{c}_0), \\ \|\delta\|_2 \leqslant u}} \left|(\mathbb{E}_n - \mathrm{E})\left[m\left(X_i^\top(\theta_0 + \delta), Y_i\right) - m\left(X_i^\top\theta_0, Y_i\right)\right]\right|, \quad u \in \mathbf{R}_+,$$

where $\bar{c}_0 := (c_0 + 1) / (c_0 - 1)$. Moreover, define the *excess risk function* $\mathcal{E} : \Theta \to \mathbf{R}_+$ by

$$\mathcal{E}\left(\theta\right) := M\left(\theta\right) - M\left(\theta_0\right) = \mathrm{E}\left[m\left(X^\top\theta, Y\right) - m\left(X^\top\theta_0, Y\right)\right], \quad \theta \in \Theta.$$

In this subsection, we derive bounds on $\|\widehat{\theta}(\lambda) - \theta_0\|_1$ and $\|\widehat{\theta}(\lambda) - \theta_0\|_2$ via the empirical error function. Our bounds will be based on the following four assumptions.

Assumption 1 (**Parameter Space**)**.** *The parameter space Θ is a convex subset of $\mathbf{R}^p$ for which θ_0 is interior.*

Assumption 2 (**Convexity**)**.** *The function $t \mapsto m\left(t, y\right)$ is convex for all $y \in \mathcal{Y}$.*

Assumption 3 (**Differentiability and Integrability**)**.** *The derivative $m_1'(X^\top\theta, Y)$ exists almost surely and $\mathrm{E}[|m(X^\top\theta, Y)|] < \infty$ for all $\theta \in \Theta$.*

Assumption 4 (**Margin**)**.** *There exist constants c_M and c_M' in $\mathbf{R}_{++}$ such that*

$$\theta \in \Theta \textit{ and } \|\theta - \theta_0\|_1 \leqslant c_M' \quad \textit{imply} \quad \mathcal{E}(\theta) \geqslant c_M \|\theta - \theta_0\|_2^2.$$

Assumption 1 is a minor regularity condition. Assumption 2 is satisfied in all examples from the previous section. These assumptions imply that the (random) function $\widehat{M}$ is convex, hence subdifferentiable on the domain interior (Rockafellar, 1970, Theorem 23.4).[8] The first part of Assumption 3 strengthens this conclusion to (full) differentiability, except possibly on a set of zero probability. This assumption is satisfied in all examples from the previous section except for Example 7 with the trimmed LAD loss function, where it is satisfied if the conditional distribution of $(\varepsilon_1, \varepsilon_2)$ given (X_1, X_2, γ) is continuous. In fact, for all our results except for those in Section 5.2, it would be sufficient to assume that the derivative $m_1'(X^\top \theta, Y)$ exists almost surely for $\theta = \theta_0$ only. The second part of Assumption 3 is a minor regularity condition. Assumption 4 is expected to be satisfied in most applications where the parameters are well identified—see Appendix B for low-level sufficient conditions and verification in the examples from Section 2.

Define $S \in \mathbf{R}^p$ as the derivative of the objective function $\widehat{M}$ at θ_0,

$$S := \mathbb{E}_n[(\partial/\partial\theta) m(X_i^\top \theta, Y_i)|_{\theta=\theta_0}] = \mathbb{E}_n[m_1'(X_i^\top \theta_0, Y_i) X_i], \tag{3.2}$$

which is almost surely well defined by Assumption 3. In this paper we refer to S as the *score*. Theorem 1 below establishes error guarantees for $\widehat{\theta}(\lambda)$ in terms of bounds on the score, penalty and empirical error. To arrive at these bounds, we introduce the following three events: For some constants λ_ϵ and $\overline{\lambda}$ in $\mathbf{R}_{++}$, let

$$\begin{aligned} \mathscr{S} &:= \{\lambda \geqslant c_0 \|S\|_\infty\}, && \text{(score domination)} \\ \mathscr{L} &:= \{\lambda \leqslant \overline{\lambda}\}, && \text{(penalty majorization)} \\ \mathscr{E} &:= \{\epsilon(u_0) \leqslant \lambda_\epsilon u_0\}, && \text{(empirical error control)} \end{aligned}$$

where

$$u_0 := \frac{2}{c_M} \left(\lambda_\epsilon + (1 + \bar{c}_0) \overline{\lambda} \sqrt{s}\right). \tag{3.3}$$

On $\mathscr{S}$ the penalty is large enough to provide a sufficient level of regularization, and on $\mathscr{L}$ the penalty is not "too large."[9] For the purpose of the deterministic calculation of this section,

[8] A function f defined on $\mathbf{R}^m$ is subdifferentiable at x if its subdifferential $\partial f(x) := \{y \in \mathbf{R}^m; f(z) \geqslant f(x) + y^\top (z - x) \text{ for all } z \in \mathbf{R}^m\}$ is nonempty. A convex function f is differentiable at x if and only if $\partial f(x)$ is singleton (with its gradient then given by the single point).

[9] Since S may be well defined only almost surely, $\mathscr{S}$ should technically include "S exists." We omit this

the event $\mathscr{L}$ plays little to no role, and one may enforce it by simply setting $\overline{\lambda} = \lambda$. However, in later sections, the penalty level will be a random quantity, $\overline{\lambda}$ will play the role of a (high-probability) bound on λ, and $\mathscr{L}$ facilitates easy reference. The constant λ_ϵ appearing in the event $\mathscr{E}$ represents a (high-probability) deterministic modulus of continuity of the empirical error function ϵ in a neighborhood of zero of size u_0.

We next present a theorem that provides guarantees for the ℓ^1- and ℓ^2-estimation error for the ℓ^1-penalized M-estimator. The proof, given in Appendix C, builds on an argument of Belloni and Chernozhukov (2011b). Similar statements appear also in van de Geer (2008), Bickel et al. (2009) and Negahban et al. (2012), among others. We make no claims of originality for these deterministic bounds and include the theorem solely for expositional purposes.

Theorem 1 (**Nonasymptotic Bounds**). *Let Assumptions 1, 2, 3, and 4 hold and suppose that* $(1+\bar{c}_0)u_0\sqrt{s} \leqslant c'_M$. *Then on the event* $\mathscr{S} \cap \mathscr{L} \cap \mathscr{E}$, *we have both*

$$\|\widehat{\theta}\,(\lambda) - \theta_0\|_2 \leqslant \frac{2}{c_M}\left(\lambda_\epsilon + (1+\bar{c}_0)\,\overline{\lambda}\sqrt{s}\right) \quad \textit{and} \tag{3.4}$$

$$\|\widehat{\theta}\,(\lambda) - \theta_0\|_1 \leqslant \frac{2\,(1+\bar{c}_0)}{c_M}\left(\lambda_\epsilon\sqrt{s} + (1+\bar{c}_0)\,\overline{\lambda}s\right). \tag{3.5}$$

This theorem motivates our choices of the penalty parameter λ. In particular, we will in Section 3.2 show that empirical error control ($\mathscr{E}$) holds with probability approaching one if $\lambda_\epsilon = C_\epsilon\sqrt{s\ln(pn)/n}$ for a sufficiently large constant $C_\epsilon \in \mathbf{R}_{++}$. (See Lemma 1.) Therefore, setting $\overline{\lambda} = \lambda$, such that penalty majorization ($\mathscr{L}$) holds trivially, Theorem 1 yields bounds of the form (1.4) given in the Introduction, and we arrive at the following principle: choose λ as small as possible subject to the constraint that the score domination event $\mathscr{S} = \{\lambda \geqslant c_0\|S\|_\infty\}$ occurs with high probability. It is exactly this principle that guides our choices of λ in the following sections.

Remark 1 (**Uniqueness**). Like similar statements appearing in the literature, Theorem 1 actually concerns the set of optimizers to the convex minimization problem (1.2) for a fixed value of λ. While the objective function $\widehat{M}$ is convex, it need not be strictly convex, such that the global minimum may be attained at more than one point $\widehat{\theta}(\lambda)$. The bounds stated here (and in what follows) hold for any of these optimizers. □

Remark 2 (**Loss Structure**). The proof of Theorem 1 requires neither the index structure placed on the loss function nor the separation of a datum W into regressors X and outcome(s) Y. The deterministic bounds in Theorem 1 continue to hold if $(w,\theta) \mapsto m(x^\top\theta, y)$ is replaced

qualifier throughout.

by a general loss $(w, \theta) \mapsto m_\theta(w)$, which is convex in θ and P-integrable in w. The theorem therefore also allows settings with multiple indices provided the loss is jointly convex in the index arguments. For example, the theorem accommodates loss functions from multinomial and conditional logit models (see Section 6.2), which are prevalent in economics. □

Remark 3 (**Margin**)**.** Our convexity, interiority, and differentiability assumptions suffice to show that the risk M is differentiable at θ_0 (Bertsekas, 1973, Proposition 2.3). Consequently, our estimand θ_0 must satisfy the population first-order condition $\nabla M(\theta_0) = \mathbf{0}$. Assumption 4 therefore amounts to assuming that the population criterion M admits a quadratic *margin* near θ_0. The name *margin condition* appears to originate from Tsybakov (2004, Assumption A1), who invokes a similar assumption in a classification context. van de Geer (2008, Assumption B) contains a more general formulation of margin behavior for estimation purposes. We consider the (focal) quadratic case for the sake of simplicity. □

Remark 4 (**Sparsity**)**.** For the Theorem 1 bounds to be interesting, our notion of sparsity must be *exact* (or *strong*), i.e. the number $s = \sum_{j=1}^{p} \mathbf{1}(\theta_{0j} \neq 0)\}$ of nonzero coefficients must be small (relative to n). However, one may entertain weaker notions such as *approximate* (or *weak*) *sparsity*. For example, one could instead assume that θ_0 belongs to an ℓ^q "ball" $\mathbb{B}_q(R_q) := \{\theta \in \mathbf{R}^p; \sum_{j=1}^{p} |\theta|^q \leqslant R_q\}$ of "radius" R_q for some fixed $q \in (0, 1]$, which would allow many nonzero but small cofficients. Step 1 in the proof of Theorem 1 shows that, under our exact sparsity assumption (corresponding to $q = 0$) and on $\mathscr{S}$, the error vector $\widehat{\theta}(\lambda) - \theta_0$ belongs to the restricted set $\mathcal{R}(\bar{c}_0)$ defined in (3.1). The structure of the restricted set allows us to swap the ℓ^1-norm of any of its elements for the corresponding ℓ^2-norm at the cost of $\sqrt{s}$ (up to a constant depending on c_0). In the terminology of Negahban et al. (2012), $\sqrt{s}$ is the *subspace compatibility constant* linking the ℓ^1-regularizer and ℓ^2-error norm on the s-dimensional coordinate subspace $\{\delta \in \mathbf{R}^p; \delta_{T^c} = \mathbf{0}\}$. With only approximate sparsity ($q > 0$), the relevant restricted set takes a more complicated form, but a careful thresholding argument still implies useful error bounds. Although these bounds involve more terms, they are akin to the ones in Theorem 1 with $\overline{\lambda}\sqrt{s}$ replaced by $\overline{\lambda}^{1-q/2}\sqrt{R_q}$. As Theorem 1 is not our contribution, we do not include the extension in this paper. See Negahban et al. (2012) for error guarantees which accommodate approximate sparsity as well as other norm-based regularizers. Simulation evidence suggests that our methods are relevant also in settings where only approximate sparsity is satisfied. (See Section 7.) □

Remark 5 (**Free Parameter**)**.** The free parameter $c_0 > 1$ in Theorem 1 serves as a trade-off between the degree (or likelihood) of score domination on the one hand and the sample size threshold and bound quality on the other. A smaller $c_0 > 1$ makes $\mathscr{S}$ more probable, but it also increases the sample size threshold and bounds. Note that a free parameter appears either explicitly or implicitly in existing bounds.[10] Given that the penalty methods proposed here seek to probabilistically control the score domination event $\mathscr{S}$, a free parameter will be present in our penalty expressions below. Our finite-sample experiments (Section 7) indicate that increasing c_0 away from one worsens performance, but also that setting c_0 to any value near one, including one itself, does not impact the results by much (cf. Figures 7.5 and 7.6). Our current recommendation is therefore to set $c_0 = 1.1$, a value slightly above unity. Similar observations (and the same recommendation) were made by Belloni et al. (2012, Footnote 7) in the context of the LASSO and linear model. □

3.2 Empirical Error Function Control

In this subsection, we consider the problem of gaining control over the empirical error event $\mathscr{E} = \{\epsilon(u_0) \leqslant \lambda_\epsilon u_0\}$. More precisely, we present conditions under which one may ensure a linear modulus of continuity of the function ϵ in a neighborhood of zero with high probability. To do so, we will use the following two assumptions.

Assumption 5 (**Covariates**). *There exist sequences ζ_n and B_n of constants in $\mathbf{R}_{++}$ and $[1,\infty)$, respectively, and a constant C_X in $\mathbf{R}_{++}$ such that (1) $\zeta_n \to 0$, (2) $\max_{1\leqslant i\leqslant n}\|X_i\|_\infty \leqslant B_n$ with probability at least $1-\zeta_n$, and (3) $\max_{1\leqslant j\leqslant p}\sum_{i=1}^n X_{ij}^4 \leqslant nC_X^4$ with probability at least $1-\zeta_n$.*

Assumption 6 (**Local Loss Behavior**). *There exist constants c_L and C_L in $\mathbf{R}_{++}$ and a function $L : \mathcal{W} \to \mathbf{R}_+$ such that*

1. *for all $w = (x, y) \in \mathcal{W}$ and all $(t_1, t_2) \in \mathbf{R}^2$ satisfying $|t_1| \vee |t_2| \leqslant c_L$,*

$$\left|m\left(x^\top\theta_0 + t_1, y\right) - m\left(x^\top\theta_0 + t_2, y\right)\right| \leqslant L(w)\,|t_1 - t_2| \tag{3.6}$$

with $\mathrm{E}[L(W)^8] \leqslant (C_L/2)^8$;

2. *for all $t \in [0, c_L]$ and $\theta \in \Theta$ satisfying $\|\theta - \theta_0\|_2 \leqslant t$, we have*

$$\mathrm{E}\left[\left\{m\left(X^\top\theta, Y\right) - m\left(X^\top\theta_0, Y\right)\right\}^2\right] \leqslant C_L^2 t^2.$$

Assumption 5 is satisfied if the regressors have sufficiently light tails. For example, with X_{ij} distributed $\mathrm{N}(0,\sigma_j^2)$, by the union bound, Assumption 5(ii) is satisfied with $B_n = \sqrt{2\ln(2pn/\zeta_n)}\max_{1\leqslant j\leqslant p}\sigma_j$ for any vanishing sequence ζ_n in $(0,1)$, and by Lemmas E.3

[10] A free parameter is explicit in both Belloni and Chernozhukov (2011b) and van de Geer (2008). In deriving their bounds both Bickel et al. (2009) (for the LASSO) and Negahban et al. (2012) set $c_0 = 2$.

and E.4 in Chernozhukov et al. (2017), Assumption 5(iii) is satisfied in this case with a sufficiently large constant C_X depending only on $\max_{1\leqslant j\leqslant p}\sigma_j$ and $\zeta_n = n^{-1}$ as long as $\ln^3(pn)\max_{1\leqslant j\leqslant p}\sigma_j^4 \leqslant n$. In the setup of our simulations (Section 7), the regressors are standard Gaussian, so taking $\zeta_n = n^{-1}$, we get that B_n grows like $\sqrt{\log(pn)}$. Assumption 6.1 requires that the function $t \mapsto m(x^\top\theta_0 + t, y)$ is locally Lipschitz continuous for all $w = (x, y) \in \mathcal{W}$ with Lipschitz constant $L(w)$, and that the eighth moment of $L(W)$ is finite. Assumption 6.2 essentially amounts to the loss being mean-square continuous at θ_0. Given that every convex function $f : C \to \mathbf{R}$ is Lipschitz relative to any compact subset of the interior of its domain C (Rockafellar, 1970, Theorem 10.4), it follows that local Lipschitz continuity (3.6) is actually implied by Assumption 2, and so Assumption 6.1 should be regarded as a mild regularity condition restricting the moments of the implied random variable $L(W)$. In Appendix B we provided low-level conditions leading to verification of this assumption in the examples from Section 2. Note that in the special case where the loss $m(\cdot, y)$ is (globally) Lipschitz with Lipschitz constant not depending on y, Assumption 6.1 is trivially satisfied. Moreover, in this case Assumption 6.2 reduces to the assumption that the eigenvalues of the matrix $\mathrm{E}[XX^\top]$ are bounded from above, which is often imposed in the literature on high-dimensional models.

We now present a result showing that one may take the high-probability local modulus of continuity λ_ϵ appearing in the empirical error event $\mathscr{E} = \{\epsilon(u_0) \leqslant \lambda_\epsilon u_0\}$ proportional to $\sqrt{s\ln(pn)/n}$:

Lemma 1 (**Empirical Error Bound**). *Let Assumptions 5 and 6 hold, and define the constant $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X \in \mathbf{R}_{++}$. Then provided $s\ln(pn) \geqslant 16C_L^2/C_\epsilon^2$ and $0 < u \leqslant c_L/\left[(1+\bar{c}_0)B_n\sqrt{s}\right]$, we have*

$$\epsilon(u) \leqslant C_\epsilon u\sqrt{s\ln(pn)/n}$$

with probability at least $1 - 5n^{-1} - 8\zeta_n$.

Remark 6 (**Alternative Nonasymptotic Bounds**). If the loss function m is (globally) Lipschitz in its first argument with Lipschitz constant not depending on y, and the regressors are bounded, then symmetrization, contraction, and concentration arguments may be used to bound the modified empirical error

$$\widetilde{\epsilon}(u) := \sup_{\substack{\delta\in\mathbf{R}^p;\\ \|\delta\|_1\leqslant u}} \left|(\mathbb{E}_n - \mathrm{E})\left[m\left(X_i^\top(\theta_0+\delta), Y_i\right) - m\left(X_i^\top\theta_0, Y_i\right)\right]\right|, \quad u \in \mathbf{R}_+,$$

now defined with respect to the ℓ^1 norm and without the restricted set $\mathcal{R}(\bar{c}_0)$. This is the approach taken by van de Geer (2008), who shows that, under the above assumptions, there exists a constant $\widetilde{C} \in \mathbf{R}_{++}$ such that with probability approaching one,

$$\widetilde{\epsilon}(u)/u \leqslant \widetilde{C}\left(\sqrt{\frac{\ln p}{n}} + \frac{\ln p}{n}\right), \quad u \in \mathbf{R}_{++}.$$

van de Geer (2008) demonstrates that bounds on the estimation error of $\widehat{\theta}(\lambda)$ can be derived if λ is chosen to exceed the right-hand side of this inequality, which motivates alternative methods to choose λ. Unfortunately, $\widetilde{C}$ typically relies on design constants unknown to the researcher. Moreover, even if these constants were known, the resulting values of $\widetilde{C}$ would typically be prohibitively large, yielding choices of λ leading to trivial estimates of the vector θ_0 in moderate samples; see Section 7 for simulation results based on the choices in van de Geer (2008). Our bounds therefore seem more suitable for devising methods to choose λ. □

4 Analytic Method

In this section we develop our analytic method to choose the penalty parameter λ. To do so, recall that we would like to choose the penalty parameter as small as possible while making score domination, $\mathscr{S} = \{\lambda \geqslant c_0\|S\|_\infty\}$, a high-probability event. Recall also that $S = \mathbb{E}_n[m_1'(X_i^\top\theta_0, Y_i)X_i]$ is the derivative of $\widehat{M}$ at $\theta = \theta_0$. By analogy with the linear mean regression, we refer to $m_1'(X^\top\theta_0, Y)$ as the *residual*. Our analytic method controls the residual through the following two assumptions.

Assumption 7 (**Conditional Mean Zero**). *The residual $m_1'(X^\top\theta_0, Y)$ is such that with probability one,* $\mathrm{E}[m_1'(X^\top\theta_0, Y)\,|X] = 0$.

Assumption 8 (**Residual: Analytic Method**). *There exist functions $\underline{r}, \overline{r} : \mathcal{X} \to \mathbf{R}$ and a known constant $d \in \mathbf{R}_{++}$ such that both $m_1'(X^\top\theta_0, Y) \in [\underline{r}(X), \overline{r}(X)]$ and $\overline{r}(X) - \underline{r}(X) \leqslant d$ almost surely.*

Assumptions 7 and 8 presume that the residual $m_1'(X^\top\theta_0, Y)$ is centered conditional on the regressors and resides in a bounded interval of known width (*d*iameter), respectively. The former assumption is satified in all of the examples from Section 2. As we explain at the end of this section, the latter assumption is satisfied in several, but not all, of the same examples.

Using these assumptions and appealing to Hoeffding's inequality (Vershynin, 2018, Theorem 2.2.6) conditional on the X_i's, we see that for any coordinate j and any $t \in \mathbf{R}_{++}$,

$$\mathrm{P}\left(|S_j| > t \,|\{X_i\}_{i=1}^n\right) \leqslant 2\exp\left(-\frac{2nt^2}{d^2\mathbb{E}_n\left[X_{ij}^2\right]}\right) \text{ a.s.}$$

The union bound then implies that for any $t \in \mathbf{R}_{++}$,

$$\mathrm{P}\left(\|S\|_\infty > t \,|\{X_i\}_{i=1}^n\right) \leqslant 2p\exp\left(-\frac{2nt^2}{d^2\max_{1\leqslant j\leqslant p}\mathbb{E}_n\left[X_{ij}^2\right]}\right) \text{ a.s.}$$

Equating the right-hand side with $\alpha \in (0,1)$ and solving for the resulting t, we arrive at the data-dependent penalty level

$$\widehat{\lambda}_\alpha^{\mathtt{am}} := c_0 d\sqrt{\frac{\ln(2p/\alpha)}{2n}\max_{1\leqslant j\leqslant p}\mathbb{E}_n\left[X_{ij}^2\right]}. \tag{4.1}$$

By construction, $\widehat{\lambda}_\alpha^{\mathtt{am}} \geqslant c_0\|S\|_\infty$ with conditional probability at least $1-\alpha$ for almost every realization of the X_i's, and so $\widehat{\lambda}_\alpha^{\mathtt{am}} \geqslant c_0\|S\|_\infty$ with probability at least $1-\alpha$ also unconditionally. Given that this penalty level is available in closed form, we refer to this method for obtaining a penalty level as the *analytic method* (AM). Note that, under Assumption 5, the analytic penalty level admits the bound

$$\widehat{\lambda}_\alpha^{\mathtt{am}} \leqslant c_0 C_X d\sqrt{\frac{\ln(p/\alpha)}{n}} =: \overline{\lambda}_\alpha^{\mathtt{am}}, \tag{4.2}$$

with probability at least $1-\zeta_n$ as long as $p \geqslant 2$. Use of the analytic method leads to the following result:

Theorem 2 (Nonasymptotic High-Probability Bounds: Analytic Method). *Let Assumptions 1–8 hold and let $\widehat{\theta} := \widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}})$ be a solution to the ℓ^1-penalized M-estimation problem (1.2) with penalty level $\lambda = \widehat{\lambda}_\alpha^{\mathtt{am}}$ given in (4.1). Define the constants $C_\epsilon := 16\sqrt{2}(1+\overline{c}_0)C_L C_X \in \mathbf{R}_{++}$, $C_\lambda^{\mathtt{am}} := c_0 C_X d \in \mathbf{R}_{++}$, and*

$$u_0 := \frac{2}{c_M}\left(C_\epsilon\sqrt{\frac{s\ln(pn)}{n}} + (1+\overline{c}_0)\,C_\lambda^{\mathtt{am}}\sqrt{\frac{s\ln(p/\alpha)}{n}}\right) \in \mathbf{R}_{++}.$$

In addition, suppose that

$$s\ln(pn) \geqslant \frac{16C_L^2}{C_\epsilon^2} \text{ and } (1+\overline{c}_0)\,u_0\sqrt{s} \leqslant \frac{c_L}{B_n} \wedge c_M'. \tag{4.3}$$

Then both

$$\|\widehat{\theta}-\theta_0\|_2 \leqslant \frac{2}{c_M}\left(C_\epsilon\sqrt{\frac{s\ln(pn)}{n}}+(1+\bar{c}_0)\,C_\lambda^{\mathtt{am}}\sqrt{\frac{s\ln(p/\alpha)}{n}}\right) \quad \textit{and}$$
$$\|\widehat{\theta}-\theta_0\|_1 \leqslant \frac{2(1+\bar{c}_0)}{c_M}\left(C_\epsilon\sqrt{\frac{s^2\ln(pn)}{n}}+(1+\bar{c}_0)\,C_\lambda^{\mathtt{am}}\sqrt{\frac{s^2\ln(p/\alpha)}{n}}\right)$$

with probability at least $1-\alpha-5n^{-1}-9\zeta_n$.

Theorem 2 gives nonasymptotic bounds on the estimation error of the ℓ^1-penalized M-estimator based on the penalty parameter λ chosen according to the analytic method. From this theorem, we immediately obtain the corresponding convergence rates:

Corollary 1 (**Convergence Rate Based on Analytic Method**). *Let Assumptions 1–8 hold and let* $\widehat{\theta}:=\widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}})$ *be a solution to the* ℓ^1*-penalized M-estimation problem (1.2) with penalty level* $\lambda=\widehat{\lambda}_\alpha^{\mathtt{am}}$ *given in (4.1). In addition, suppose that*

$$\alpha\to 0 \quad \textit{and} \quad \frac{B_n^2 s^2\ln(pn/\alpha)}{n}\to 0. \tag{4.4}$$

Then there exists a constant $C\in\mathbf{R}_{++}$ *depending only on the constants appearing in the aforementioned assumptions such that both*

$$\|\widehat{\theta}-\theta_0\|_2\leqslant C\sqrt{\frac{s\ln(pn/\alpha)}{n}} \quad \textit{and} \quad \|\widehat{\theta}-\theta_0\|_1\leqslant C\sqrt{\frac{s^2\ln(pn/\alpha)}{n}}$$

with probability approaching one.

The proof of Corollary 1 actually reveals that even if we drop the $\alpha\to 0$ requirement, we get

$$\mathrm{P}\left(\|\widehat{\theta}-\theta_0\|_2\leqslant C\sqrt{\frac{s\ln(pn/\alpha)}{n}}\text{ and }\|\widehat{\theta}-\theta_0\|_1\leqslant C\sqrt{\frac{s^2\ln(pn/\alpha)}{n}}\right)\geqslant 1-\alpha+o(1).$$

For example, with a fixed probability tolerance α (i.e. not depending on n), the previous display implies

$$\liminf_{n\to\infty}\mathrm{P}\left(\|\widehat{\theta}-\theta_0\|_2\leqslant C\sqrt{\frac{s\ln(pn/\alpha)}{n}}\text{ and }\|\widehat{\theta}-\theta_0\|_1\leqslant C\sqrt{\frac{s^2\ln(pn/\alpha)}{n}}\right)\geqslant 1-\alpha.$$

We conclude this section by pointing out examples from Section 2 where Assumption 8 is satisfied, and so our analytic method can be applied.

Example 1 (Binary Response Model, Continued). The logit loss function (2.2) is differentiable in t with $m_1'(t,y) = \Lambda(t) - y$. The logit residual $m_1'(X^\top\theta_0, Y)$ thus resides in the interval $[\Lambda(X^\top\theta_0) - 1, \Lambda(X^\top\theta_0)]$, and so satisfies Assumption 8 with $d = 1$.

More generally, let F admit an everywhere positive log-concave PDF $f = F'$. Then the binary-response loss (2.1) is differentiable with partial derivative

$$m_1'(t,y) = \frac{f(t)}{F(t)[1 - F(t)]}[F(t) - y]. \tag{4.5}$$

The binary nature of the outcome implies that

$$\min_{y\in\{0,1\}} m_1'\left(X^\top\theta_0, y\right) \leqslant m_1'\left(X^\top\theta_0, Y\right) \leqslant \max_{y\in\{0,1\}} m_1'\left(X^\top\theta_0, y\right).$$

From (4.5) we may deduce $m_1'(t,1) = -f(t)/F(t) < 0 < f(t)/[1-F(t)] = m_1'(t,0)$. Inserting and simplifying, we therefore arrive at

$$\max_{y\in\{0,1\}} m_1'(t,y) - \min_{y\in\{0,1\}} m_1'(t,y) = \frac{f(t)}{F(t)[1 - F(t)]}.$$

Hence, for any distribution such that $f/[F(1-F)]$ is also bounded from above, Assumption 8 is satisfied with

$$d = \sup_{t\in\mathbf{R}} \frac{f(t)}{F(t)[1-F(t)]},$$

which only requires solving an unconstrained, univariate maximization problem. For example, as discussed earlier, $f/[F(1-F)]$ is bounded from above if F is a t-distribution F_ν with $0 < \nu < \infty$ degrees of freedom. For $0 < \nu \leqslant 5$, the (unique) mode of $f_\nu/[F_\nu(1-F_\nu)]$ is zero, such that $d = d_\nu = 4f_\nu(0) = 4\Gamma((\nu+1)/2)/[\sqrt{\nu\pi}\Gamma(\nu/2)]$. For example, $d_1 = 4/\pi \approx 1.41$ for the standard Cauchy distribution. For higher degrees of freedom, the solution is more complicated, the exact d_ν being somewhat larger than the value $\frac{1}{2}\sqrt{\nu}$ of the (asymptotic) program $\sup_{t\in\mathbf{R}} |t|/(1+t^2/\nu)$. For example, $\nu = 9$ produces $d_9 \approx 1.68 > \frac{3}{2}$.

As a side note, observe also that in contrast to the logit loss function, the probit loss function (2.3) does not satisfy Assumption 8. Indeed, this loss function is differentiable in t with

$$m_1'(t,y) = \frac{\varphi(t)}{\Phi(t)[1-\Phi(t)]}[\Phi(t) - y]$$

but here $m_1'(t,0) - m_1'(t,1) = \varphi(t)/[1-\Phi(t)] + \varphi(t)/\Phi(t) \sim t$ as $t \to \infty$. The probit residual is thus not confined to any bounded interval, violating Assumption 8. We could in principle reconcile the probit loss function with Assumption 8 by assuming that we know a constant $C_d \in \mathbf{R}_{++}$ such that $\|\theta_0\|_1 \leqslant C_d$ and setting

$$d = \sup_{t \in [-\overline{X} C_d, \overline{X} C_d]} \frac{\varphi(t)}{\Phi(t)\left[1 - \Phi(t)\right]},$$

where $\overline{X} = \max_{1 \leqslant i \leqslant n} \|X_i\|_\infty$. While the resulting d is a known function of the X_i's, this procedure would likely lead to very large values of the penalty parameter λ, thus making the analytic method impractical. □

Example 2 (Ordered Response Model, Continued). Provided the distribution F admits an everywhere positive log-concave PDF $f = F'$, the ordered-response loss (2.4) is differentiable in t with partial derivative

$$m_1'(t, y) = \sum_{j=0}^{J} \mathbf{1}(y = j) \frac{f(\alpha_{j+1} - t) - f(\alpha_j - t)}{F(\alpha_{j+1} - t) - F(\alpha_j - t)}, \tag{4.6}$$

where we interpret $f(\pm\infty)$ and $F(-\infty)$ as zero and $F(+\infty)$ as one. The discrete nature of the outcome and (4.6) imply that

$$\min_{0 \leqslant j \leqslant J} \frac{f(\alpha_{j+1} - t) - f(\alpha_j - t)}{F(\alpha_{j+1} - t) - F(\alpha_j - t)} \leqslant m_1'(t, y) \leqslant \max_{0 \leqslant j \leqslant J} \frac{f(\alpha_{j+1} - t) - f(\alpha_j - t)}{F(\alpha_{j+1} - t) - F(\alpha_j - t)},$$

Hence, for a distribution F and cut-off points $\{\alpha_j\}$ such that the difference between the upper and lower bounds is bounded from above in t, Assumption 8 is satisfied with

$$d = \sup_{t \in \mathbf{R}} \left\{ \max_{0 \leqslant j \leqslant J-1} \frac{f(\alpha_{j+1} - t) - f(\alpha_j - t)}{F(\alpha_{j+1} - t) - F(\alpha_j - t)} - \min_{1 \leqslant j \leqslant J} \frac{f(\alpha_{j+1} - t) - f(\alpha_j - t)}{F(\alpha_{j+1} - t) - F(\alpha_j - t)} \right\},$$

where we have used our knowledge of the signs of the first and last elements to reduce the candidates for a minimum and maximum, respectively. With knowledge of F and the α_j's, this quantity may at least in principle be computed. For the logistic distribution $F = \Lambda$ we have $f = \Lambda(1 - \Lambda)$, such that d simplifies to

$$\begin{aligned} d &= \sup_{t \in \mathbf{R}} \left\{ \max_{0 \leqslant j \leqslant J-1} \left\{1 - \Lambda(\alpha_{j+1} - t) - \Lambda(\alpha_j - t)\right\} - \min_{1 \leqslant j \leqslant J} \left\{1 - \Lambda(\alpha_{j+1} - t) - \Lambda(\alpha_j - t)\right\} \right\} \\ &= \sup_{t \in \mathbf{R}} \left\{ \max_{1 \leqslant j \leqslant J} \left\{\Lambda(\alpha_{j+1} - t) + \Lambda(\alpha_j - t)\right\} - \min_{0 \leqslant j \leqslant J-1} \left\{\Lambda(\alpha_{j+1} - t) + \Lambda(\alpha_j - t)\right\} \right\}. \end{aligned}$$

The maxi-/minimands are here ascending in j for each $t \in \mathbf{R}$, so the pointwise maximum and minimum equal $1 + \Lambda(\alpha_J - t)$ and $\Lambda(\alpha_1 - t)$, respectively. The resulting d is

$$d = \sup_{t \in \mathbf{R}} \{1 + \Lambda(\alpha_J - t) - \Lambda(\alpha_1 - t)\} = 2\Lambda\left(\frac{\alpha_J - \alpha_1}{2}\right),$$

the supremum being attained at the midpoint $t = (\alpha_1 + \alpha_J)/2$. The previous display shows that (i) d depends only on the difference $\alpha_J - \alpha_1$ between the largest and smallest threshold, and that (ii) $d \in (1, 2)$ for all such threshold values. For example, a difference of $\alpha_J - \alpha_1 = 2$ produces $d = 2\mathrm{e}/(1 + \mathrm{e}) \approx 1.46$. In the limiting case $\alpha_J - \alpha_1 \to 0$, we recover the value $d = 1$ for the binary logit, as expected. □

Example 6 (Panel Logit Model, Continued). The panel logit loss function (2.8) is differentiable in t with $m_1'(t, y) = \mathbf{1}(y_1 \neq y_2)[\Lambda(t) - y_1]$. Thus, the residual $m_1'((X_1 - X_2)^\top \theta_0, Y)$ resides in the interval $[\Lambda((X_1 - X_2)^\top \theta_0) - 1, \Lambda((X_1 - X_2)^\top \theta_0)]$, and so satisfies Assumption 8 with $d = 1$. □

Example 7 (Panel Censored Model, Continued). The trimmed LAD loss function (2.9) is differentiable in t and satisfies $|m_1'(t, y)| \leqslant 1$ if $t \neq y_1 - y_2$ or $y_1 = y_2 = 0$. Thus, as long as the conditional distribution of $(\varepsilon_1, \varepsilon_2)$ given (X_1, X_2, γ) is absolutely continuous (as implied by Honoré (1992, Assumption E.1)), this loss function satisfies Assumption 8 with $d = 2$. Note, however, that the trimmed LS loss function does not satisfy Assumption 8. □

Example 8 (Panel Duration Model, Continued). Since the loss function (2.10) here is of the logit form, Assumption 8 is satisfied with $d = 1$ (cf. Example 1). □

5 Bootstrap-after-Cross-Validation Method

The analytic method of the previous section relies on Assumption 8. As explained there, this assumption is satisfied in quite a few applications. However, there are also many other applications where this assumption is not satisfied. Examples include the probit model, the logit model with estimation based on the logistic calibration loss function, and the panel censored model with estimation based on the trimmed LS loss function. (See Examples 1, 3 and 7, respectively.) Moreover, even if Assumption 8 is satisfied, the analytic penalty level $\widehat{\lambda}_\alpha^{\mathsf{am}}$ in (4.1) follows from a union-bound argument and may thus be quite conservative. In this section we therefore seek to provide a method to choose the penalty parameter which is not conservative and broadly available, yet amenable to theoretical analysis.

We split the section into two subsections. In Section 5.1, we develop a generic bootstrap method that allows for choosing the penalty parameter λ assuming availability of some generic estimators $\widehat{U}_i$ of the residuals $U_i = m_1'(X_i^\top \theta_0, Y_i)$. We then briefly discuss a plug-in strategy for residual estimation (Remark 7). In Section 5.2, we explain how to obtain suitable estimators $\widehat{U}_i$ via cross-validation, which is broadly available. For convenience of the reader, we have gathered the implementation details for our two main methods in Appendix A. See also Section 6 for extensions of these methods to richer modelling frameworks.

5.1 Bootstrapping the Penalty Level

To develop some intuition, suppose for the moment that residuals $U_i = m_1'\left(X_i^\top \theta_0, Y_i\right)$ are observable. In this case, we can estimate the $(1-\alpha)$ quantile of the score $S = \mathbb{E}_n[U_i X_i]$,

$$q\,(1-\alpha) := (1-\alpha)\text{-quantile of } \max_{1\leqslant j\leqslant p} \left|\mathbb{E}_n\left[U_i X_{ij}\right]\right|,$$

via the Gaussian multiplier bootstrap.[11] To this end, let $\{e_i\}_{i=1}^n$ be independent standard normal random variables that are independent of the data $\{W_i\}_{i=1}^n$. We then estimate $q(1-\alpha)$ by

$$\widetilde{q}\,(1-\alpha) := (1-\alpha)\text{-quantile of } \max_{1\leqslant j\leqslant p} \left|\mathbb{E}_n\left[e_i U_i X_{ij}\right]\right| \text{ given } \{W_i\}_{i=1}^n.$$

It is rather standard to show that, under certain regularity conditions, $\widetilde{q}(1-\alpha)$ delivers a good approximation to $q(1-\alpha)$, even if the dimension p of the X_i's is much larger than the sample size n. To see why this is the case, let Z be a centered random vector in $\mathbf{R}^p$ and let $Z_1, \ldots, Z_n$ be independent copies of Z. As established in Chernozhukov et al. (2013, 2017), the random vectors $Z_1, \ldots, Z_n$ satisfy the following high-dimensional versions of the central limit and Gaussian multiplier bootstrap theorems: If for some constant $b \in \mathbf{R}_{++}$ and a sequence $\widetilde{B}_n$ of constants in $[1, \infty)$, possibly growing to infinity, one has

$$\min_{1\leqslant j\leqslant p} \mathrm{E}[Z_{ij}^2] \geqslant b, \quad \max_{k\in\{1,2\}} \max_{1\leqslant j\leqslant p} \mathrm{E}\left[|Z_{ij}|^{2+k}\right] / \widetilde{B}_n^k \leqslant 1 \quad \text{and} \quad \mathrm{E}\Big[\max_{1\leqslant j\leqslant p} Z_{ij}^4\Big] \leqslant \widetilde{B}_n^4,$$

then there exist a constant $C_b \in \mathbf{R}_{++}$, depending only on b, such

$$\sup_{A\in\mathcal{A}_p} \left| \mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i \in A\right) - \mathrm{P}\big(\mathrm{N}(\mathbf{0}, \mathrm{E}[ZZ^\top]) \in A\big)\right| \leqslant C_b \left(\frac{\widetilde{B}_n^4 \ln^7(pn)}{n}\right)^{1/6}, \tag{5.1}$$

and, with probability approaching one,

$$\sup_{A\in\mathcal{A}_p} \left| \mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i Z_i \in A \middle| \{Z_i\}_{i=1}^n\right) - \mathrm{P}\left(\mathrm{N}(\mathbf{0}, \mathrm{E}[ZZ^\top]) \in A\right)\right| \leqslant C_b \left(\frac{\widetilde{B}_n^4 \ln^6(pn)}{n}\right)^{1/6}. \tag{5.2}$$

[11] Recall that S is well-defined and unique a.s. We omit the qualifier throughout this section.

Here $\mathcal{A}_p$ denotes the collection of all (hyper)rectangles in $\mathbf{R}^p$. Provided $\widetilde{B}_n^4 \ln^7(pn)/n \to 0$, combination of these two results suggests that the Gaussian multiplier bootstrap yields a good approximation to the law of the potentially high-dimensional vector $n^{-1/2}\sum_{i=1}^n Z_i$ when restricted to (hyper)rectangles.

Consider now the family of rectangles defined by

$$A_t := \Big\{u \in \mathbf{R}^p;\ \max_{1\leqslant j\leqslant p} |u_j| \leqslant t\Big\}, \quad t \geqslant 0.$$

We can then write

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |\mathbb{E}_n\left[U_i X_{ij}\right]| \leqslant t\right) = \mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i X_i \in A_{t\sqrt{n}}\right).$$

The $U_i X_i$'s are centered under Assumption 7, and so the aforementioned results can be applied in our context of ℓ^1-penalized M-estimation.

Of course, we typically do not observe the residuals $U_i = m_1'(X_i^\top\theta_0, Y_i)$, and so the method described above is infeasible. Fortunately, the result (5.2) continues to hold upon replacing $\{Z_i\}_{i=1}^n$ with estimators $\{\widehat{Z}_i\}_{i=1}^n$, provided these estimators are "sufficiently good." Suppose therefore that residual estimators $\{\widehat{U}_i\}_{i=1}^n$ are available. We then compute

$$\widehat{q}\,(1-\alpha) := (1-\alpha)\text{-quantile of } \max_{1\leqslant j\leqslant p}\left|\mathbb{E}_n\big[e_i \widehat{U}_i X_{ij}\big]\right| \text{ given } \{(W_i, \widehat{U}_i)\}_{i=1}^n, \tag{5.3}$$

and a feasible penalty level follows as

$$\widehat{\lambda}_\alpha^{\mathsf{bm}} := c_0 \widehat{q}\,(1-\alpha)\,. \tag{5.4}$$

We refer to this method for obtaining a penalty level as the *bootstrap method* (BM) and to $\widehat{\lambda}_\alpha^{\mathsf{bm}}$ itself as the *bootstrapped penalty level.*

To ensure that $\widehat{q}\,(1-\alpha)$ delivers a good approximation to $q\,(1-\alpha)$, we invoke the following assumptions, where we denote $U = m_1'(X^\top\theta_0, Y)$ and $Z = (Z_1, \ldots, Z_p)^\top = UX$.

Assumption 9 (**Residual: Bootstrap Method**). *There exist constants c_U and C_U in $\mathbf{R}_{++}$ and a sequence $\widetilde{B}_n$ of constants in $[1,\infty)$ such that (1) $\mathrm{E}[U^8] \leqslant (C_U/2)^8$, (2) $\mathrm{E}[Z_j^2] \geqslant c_U$ for all $j \in \{1, \ldots, p\}$, (3) $\mathrm{E}[|Z_j|^{2+k}] \leqslant \widetilde{B}_n^k$ for all $k \in \{1,2\}$ and all $j \in \{1,\ldots,p\}$, and (4) $\mathrm{E}\left[\|Z\|_\infty^4\right] \leqslant \widetilde{B}_n^4$.*

Assumption 10 (**Residual Estimation**). *There exist sequences β_n and δ_n of constants in $\mathbf{R}_{++}$ both converging to zero such that $\mathbb{E}_n[(\widehat{U}_i - U_i)^2] \leqslant \delta_n^2/\ln^2(pn)$ with probability at least $1-\beta_n$.*

Assumption 9 contains a set of moment conditions tailored to a high-dimensional multiplier bootstrap theorem (see Appendix E.4). Other moment conditions are certainly possible (cf. Chernozhukov et al., 2017). In the binary logit setup used in our simulations (see Section 7 as well as Example 1), the residual is bounded in absolutely value by one, thus verifying Assumption 9.1 with $C_U = 2$. Since the regressors are there taken to be standard Gaussian, the third and fourth absolute moments of the Z_j's are bounded by absolute constants. Moreover, properties of L^p and exponential Orlisz norms (see, e.g., van der Vaart and Wellner, 1996, Section 2.2) show that $\mathrm{E}\left[\|Z\|_\infty^4\right]$ grows no faster than $\ln^2(p)$. In this case, Assumptions 9.3 and 9.4 are satisfied with $\widetilde{B}_n$ proportional to $\sqrt{\ln p}$.

Use of the bootstrap method leads to following result:

Theorem 3 (**Nonasymptotic High-Probability Bounds: Bootstrap Method**). *Let Assumptions 1–7, 9 and 10 hold with $B_n\delta_n \to 0$, and let $\widehat{\theta} := \widehat{\theta}(\widehat{\lambda}_\alpha^{\mathsf{bm}})$ be a solution to the ℓ^1-penalized M-estimation problem (1.2) with penalty level $\lambda = \widehat{\lambda}_\alpha^{\mathsf{bm}}$ given in (5.4). Define the constants $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X \in \mathbf{R}_{++}$, $C_\lambda^{\mathsf{bm}} := 4(2+\sqrt{2})c_0 C_U C_X \in \mathbf{R}_{++}$, and*

$$u_0 := \frac{2}{c_M}\left(C_\epsilon\sqrt{\frac{s\ln(pn)}{n}} + (1+\bar{c}_0)\,C_\lambda^{\mathsf{bm}}\sqrt{\frac{s\ln(p/\alpha)}{n}}\right) \in \mathbf{R}_{++}.$$

In addition, suppose that

$$s\ln(pn) \geqslant \frac{16C_L^2}{C_\epsilon^2}, \quad (1+\bar{c}_0)\,u_0\sqrt{s} \leqslant \frac{c_L}{B_n} \wedge c_M' \quad \text{and} \quad B_n\delta_n \leqslant C_U C_X \ln(pn). \tag{5.5}$$

Then there exists a constant $C \in \mathbf{R}_{++}$, depending only on c_U, such that for

$$\rho_n := C\max\left\{\beta_n + \zeta_n, B_n\delta_n, \left(\frac{\widetilde{B}_n^4\ln^7(pn)}{n}\right)^{1/6}, \frac{1}{\ln^2(pn)}\right\},$$

we have both

$$\|\widehat{\theta}-\theta_0\|_2 \leqslant \frac{2}{c_M}\left(C_\epsilon\sqrt{\frac{s\ln(pn)}{n}} + (1+\bar{c}_0)\,C_\lambda^{\mathsf{bm}}\sqrt{\frac{s\ln(p/\alpha)}{n}}\right) \quad \text{and}$$
$$\|\widehat{\theta}-\theta_0\|_1 \leqslant \frac{2(1+\bar{c}_0)}{c_M}\left(C_\epsilon\sqrt{\frac{s^2\ln(pn)}{n}} + (1+\bar{c}_0)\,C_\lambda^{\mathsf{bm}}\sqrt{\frac{s^2\ln(p/\alpha)}{n}}\right)$$

with probability at least $1-\alpha-6n^{-1}-\beta_n-10\zeta_n-\rho_n$.

The idea of using a bootstrap procedure to select the penalty level in high-dimensional estimation is itself not new. Chernozhukov et al. (2013) use a Gaussian multiplier bootstrap to tune the Dantzig selector for the high-dimensional linear model allowing both non-Gaussian and heteroskedastic errors. Note, however, that Theorem 4.2 of the same paper presumes access to a preliminary Dantzig selector which is used to estimate residuals. Assumption 10 is similarly 'high-level' in the sense that it does not specify how one performs residual estimation in practice. Our primary contribution lies in providing methods for coming up with good residual estimators, which we turn in the next subsection.

Remark 7 (Plug-In Residual Estimation). If the residuals are estimated using the plug-in estimator $\widehat{U}_i = m_1'(X_i^\top \widehat{\theta}, Y_i)$ based on some preliminary estimator $\widehat{\theta}$ of θ_0, and the function $m_1'(\cdot, y)$ is (globally) $L_1(y)$-Lipschitz for some function $L_1 : \mathcal{Y} \to \mathbf{R}_+$, then $\mathbb{E}_n[(\widehat{U}_i - U_i)^2] \leqslant \mathbb{E}_n[L_1(Y_i)^2 |X_i^\top(\widehat{\theta} - \theta_0)|^2]$. Hence, under Assumption 5 and provided $\mathbb{E}_n[L_1(Y_i)^2]$ is bounded in probability, the right-hand side is bounded in probability by the product of B_n^2 and the squared ℓ^1-error of the preliminary estimator. In the case where L_1 is itself bounded, up to a constant multiple, $\mathbb{E}_n[(\widehat{U}_i - U_i)^2]$ is bounded by the squared prediction error $\mathbb{E}_n[|X_i^\top(\widehat{\theta} - \theta_0)|^2]$, which may lead to faster convergence than for the (squared) ℓ^1 estimation error. (See, e.g., Bühlmann and van de Geer (2011, Theorem 6.1) and Belloni and Chernozhukov (2013, Theorem 1) for the LASSO and linear model.) Thus, in some cases, up to a logarithmic factor, the residual estimator inherits the probability and error terms β_n and δ_n appearing in Assumption 10 from the preliminary estimator $\widehat{\theta}$. □

Example 1 (Binary Response Model, Continued). Consider the binary logit model and loss $m(t, y) = \ln(1 + \mathrm{e}^t) - yt$. Then the residual $U = m_1'(X^\top \theta_0, Y) = \Lambda(X^\top \theta_0) - Y$ resides in an interval of width one, and we may consider an ℓ^1-penalized logit estimator $\widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}})$ based on the analytic method (4.1) with $d = 1$. Let $\widehat{U}_i^{\mathtt{am}} := \Lambda(X_i^\top \widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}})) - Y_i$ be the resulting plug-in estimators. Under the assumptions of Theorem 2, the same theorem establishes the existence of a constant $C \in \mathbf{R}_{++}$ such that $\|\widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}}) - \theta_0\|_1 \leqslant C\sqrt{s \ln(pn/\alpha)/n}$ with probability at least $1 - \alpha - 5n^{-1} - 9\zeta_n$. The logit loss is twice differentiable in its first argument with $m_{11}''(t, y) = \Lambda'(t) = \mathrm{e}^t/(1 + \mathrm{e}^t)^2$, which is bounded by one in absolute value. Hence, by Remark 7 and further bounding the squared prediction error $\mathbb{E}_n[|X_i^\top(\widehat{\theta} - \theta_0)|^2]$ by $\max_{1 \leqslant i \leqslant n} \|X_i\|_\infty^2 \|\widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}}) - \theta_0\|_1^2$, the residual estimates implied by the analytic method then satisfy Assumption 10 with $\beta_n = \alpha + 5n^{-1} + 10\zeta_n$ and δ_n^2 proportional to $B_n^2 s \ln^2(pn) \ln(pn/\alpha)/n$ (provided both sequences vanish as $n \to \infty$). We consider the ℓ^1-penalized logit estimator arising from bootstrapping after the analytic method in our simulations—see Section 7. □

The previous remark and example illustrate that a plug-in estimation strategy, perhaps based on a preliminary estimator using the analytic method from Section 4, can yield suitable residual estimators under additional (smoothness) assumptions. In Section 5.2, we show how to obtain residual estimators $\{\widehat{U}_i\}_{i=1}^n$ via cross-validation, thus obtaining the bootstrap-after-cross-validation method.

5.2 Cross-Validating Residuals

In this subsection, we explain how residual estimation can be performed via cross-validation (CV). To describe our CV residual estimator, fix any integer $K \geqslant 2$, and let $I_1, \ldots, I_K$ partition the sample indices $\{1, \ldots, n\}$. Provided n is divisible by K, the even partition

$$I_k = \{(k-1)\, n/K + 1, \ldots, kn/K\}, \quad k \in \{1, \ldots, K\}, \tag{5.6}$$

is natural, but not necessary. For the formal results below, we only require that each I_k specifies a "substantial" subsample (see Assumption 11 below).

Define the *subsample criterion* $\widehat{M}_I$ to be the sample criterion

$$\widehat{M}_I(\theta) := \mathbb{E}_I\left[m\left(X_i^\top \theta, Y_i\right)\right], \quad \theta \in \Theta, \quad \emptyset \neq I \subsetneq \{1, \ldots, n\}, \tag{5.7}$$

based only on observations $i \in I$, and let Λ_n denote a finite subset of $\mathbf{R}_{++}$ composed by candidate penalty levels. We require Λ_n to be "sufficiently rich" (see Assumption 12 below). Our CV procedure then goes as follows. First, estimate parameters θ_0 by

$$\widehat{\theta}_{I_k^c}(\lambda) \in \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{\widehat{M}_{I_k^c}(\theta) + \lambda \|\theta\|_1\right\}, \tag{5.8}$$

for each candidate penalty level $\lambda \in \Lambda_n$ and holding out each subsample $k \in \{1, \ldots, K\}$ in turn. Second, determine the penalty level

$$\widehat{\lambda}^{\mathsf{cv}} \in \underset{\lambda \in \Lambda_n}{\operatorname{argmin}} \sum_{k=1}^K \sum_{i \in I_k} m\big(X_i^\top \widehat{\theta}_{I_k^c}(\lambda), Y_i\big) \tag{5.9}$$

by minimizing the out-of-sample loss over the set of candidate penalties. Third, estimate residuals $U_i = m_1\left(X_i^\top \theta_0, Y_i\right), i \in \{1, \ldots, n\}$, by predicting out of each estimation sample, i.e.,

$$\widehat{U}_i^{\mathsf{cv}} := m_1'\big(X_i^\top \widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}), Y_i\big), \quad i \in I_k, \quad k \in \{1, \ldots, K\}. \tag{5.10}$$

Combining the bootstrap estimate $\widehat{\lambda}_\alpha^{\mathsf{bm}} = c_0 \widehat{q}(1-\alpha)$ from the previous subsection with the CV residual estimates $\widehat{U}_i = \widehat{U}_i^{\mathsf{cv}}$ from this subsection, we obtain the *bootstrap-after-cross-validation* (BCV) *method* for choosing the penalty parameter $\lambda = \widehat{\lambda}_\alpha^{\mathsf{bcv}}$.

To ensure good performance of the estimator resulting from initiating the bootstrap method with the CV residual estimates (5.10), we invoke the following assumptions.

Assumption 11 (**Data Partition**). *The number $K \in \{2,3,\dots\}$ of subsamples is constant and does not depend on n. There exists a constant $c_D \in (0,1)$ such that $\min_{1\leqslant k\leqslant K} |I_k| \geqslant c_D n$.*

Assumption 12 (**Candidate Penalties**). *There exists constants c_Λ and C_Λ in $\mathbf{R}_{++}$ and $a \in (0,1)$ such that*

$$\Lambda_n = \left\{C_\Lambda a^\ell; a^\ell \geqslant c_\Lambda/n, \ell \in \{0,1,2,\dots\}\right\}.$$

Assumption 13 (**Residual: Cross-Validation Method**). *There exist constants σ and C_{ms1} in $\mathbf{R}_{++}$ such that:*

1. *For all $t \in \mathbf{R}$,*

$$\ln \mathrm{E}\left[\exp\left(t m_1'\left(X^\top\theta_0, Y\right)\right)\middle| X\right] \leqslant \frac{\sigma^2 t^2}{2} \text{ a.s.}$$

2. *For all $\theta \in \Theta$,*

$$\mathrm{E}\left[\left\{m_1'\left(X^\top\theta, Y\right) - m_1'\left(X^\top\theta_0, Y\right)\right\}^2\right] \leqslant C_{ms1}^2\left(\sqrt{\mathcal{E}(\theta)} \vee \mathcal{E}(\theta)\right).$$

Assumption 14 (**Global Loss Behaviour**). *There exists a constant $C_{ms} \in \mathbf{R}_{++}$ such that for all $\theta \in \Theta$,*

$$\mathrm{E}\left[\left\{m\left(X^\top\theta, Y\right) - m\left(X^\top\theta_0, Y\right)\right\}^2\right] \leqslant C_{ms}^2\left(\mathcal{E}\left(\theta\right) \vee \mathcal{E}\left(\theta\right)^2\right).$$

Assumption 11 means that we rely upon the classical K-fold CV with fixed K. This assumption does rule out leave-one-out CV, since $K = n$ and $I_k = \{k\}$ imply $|I_k|/n \to 0$. Assumption 12 allows for a rather large candidate set Λ_n of penalty values. Note that the largest penalty value (C_Λ) can be set arbitrarily large and the smallest value (c_Λ/n) converges rapidly to zero. In Lemma C.1 we show that these properties ensure that the set Λ_n eventually contains a "good" penalty candidate, say λ_*, in the sense of leading to a uniform bound on the excess risk of subsample estimators $\widehat{\theta}_{I_k^c}\left(\lambda_*\right), k \in \{1,\dots,K\}$. Other candidate penalty sets leading to a bound on the subsample estimator excess risk are certainly possible. Assumptions 13 and 14 are high-level but rather mild. We provide a set of low-level conditions suitable for each of the examples from Section 2 in Appendix B.

Use of CV residual estimators leads to the following result:

Theorem 4 (**High-Probability CV-Residual Error Bound**). *Let Assumptions 1–6 and 11–14 hold. Define the constants $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X$, $C_S := 2C_X\sigma/((K-1)c_D)$,*

$$C_{\mathcal{E}} := \sqrt{\frac{2}{c_M}}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\bar{c}_0)\,c_0 C_S}{a}\right) \tag{5.11}$$

and

$$\widetilde{u}_0 := \frac{2}{c_M}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\bar{c}_0)\,c_0 C_S}{a}\right)\sqrt{\frac{s\ln(pn)}{n}}, \tag{5.12}$$

which are all in $\mathbf{R}_{++}$*. In addition, suppose that*

$$s\ln(pn) \geqslant \frac{16\,(K-1)\,c_D C_L^2}{C_\epsilon^2}, \quad (1+\bar{c}_0)\,\widetilde{u}_0\sqrt{s} \leqslant \frac{c_L}{B_n}\wedge c_M', \tag{5.13}$$

$$\frac{\ln(pn)}{n} \leqslant \left(\frac{C_\Lambda a}{c_0 C_S}\right)^2, \quad n\ln(pn) \geqslant \left(\frac{c_\Lambda}{c_0 C_S}\right)^2 \quad \textit{and} \quad n \geqslant \frac{1}{c_\Lambda \wedge a}. \tag{5.14}$$

Then for any $t \in \mathbf{R}_{++}$ *satisfying*

$$\frac{48C_{ms}^2}{c_D^2\ln(1/a)}\frac{t\ln n}{n} + \frac{8C_{\mathcal{E}}^2}{c_D}\frac{s\ln(pn)}{n} \leqslant 1, \tag{5.15}$$

we have

$$\mathbb{E}_n\big[(\widehat{U}_i^{\mathsf{cv}} - U_i)^2\big] \leqslant \frac{12C_{ms1}^2 t\ln n}{\ln(1/a)}\left(\frac{3C_{ms}^2}{c_D^2\ln(1/a)}\frac{t\ln n}{n} + \frac{C_{\mathcal{E}}^2}{2c_D}\frac{s\ln(pn)}{n}\right)^{1/2} \tag{5.16}$$

with probability at least $1 - K(5n^{-1} + 9\zeta_n + [(K-1)c_D n]^{-1} + 2t^{-1})$*.*

This theorem provides an avenue for verification of Assumption 10. Specifically, it implies that for any sequence t_n of constants in $\mathbf{R}_{++}$ satisfying (5.15), we can take β_n to be $K(5n^{-1} + 9\zeta_n + [(K-1)c_D n]^{-1} + 2t_n^{-1})$ and specify δ_n^2 as the right-hand side of (5.16) multiplied by $\ln^2(pn)$.

Theorem 4 indicates that the CV residual estimators are reasonable inputs for the bootstrap method. Combining Theorems 3 and 4, we obtain convergence rates for the ℓ^1-penalized M-estimator based on the penalty parameter λ chosen according to the BCV method:

Corollary 2 (**Convergence Rate Based on Bootstrap after CV Method**)**.** *Let Assumptions 1–7, 9, and 11–14 hold and let* $\widehat{\theta} := \widehat{\theta}(\widehat{\lambda}_\alpha^{\mathsf{bcv}})$ *be a solution to the* ℓ^1*-penalized M-estimation problem* (1.2) *with penalty level* $\lambda = \widehat{\lambda}_\alpha^{\mathsf{bcv}}$*. In addition, suppose that*

$$\alpha \to 0, \quad \frac{B_n^2 s^2\ln(pn/\alpha)}{n} \to 0, \quad \frac{B_n^4 s\ln^5(pn)(\ln n)^2}{n} \to 0 \quad \textit{and} \quad \frac{\widetilde{B}_n^4\ln^7(pn)}{n} \to 0. \tag{5.17}$$

Then there exists a constant C *in* $\mathbf{R}_{++}$ *depending only on the constants appearing in the aforementioned assumptions such that both*

$$\|\widehat{\theta} - \theta_0\|_2 \leqslant C\sqrt{\frac{s\ln(pn/\alpha)}{n}} \quad \textit{and} \quad \|\widehat{\theta} - \theta_0\|_1 \leqslant C\sqrt{\frac{s^2\ln(pn/\alpha)}{n}}$$

with probability approaching one.

Corollaries 1 and 2 demonstrate that both analytic and bootstrap-after-cross-validation methods with α, for example, proportional to $1/n$ yield ℓ^1-penalized M-estimators whose convergence rates in the ℓ^2 and ℓ^1 norms are $\sqrt{s\ln(pn)/n}$ and $\sqrt{s^2\ln(pn)/n}$, respectively. These are typical rates that one expects in the high-dimensional settings under sparsity. For example, it is well-known that these rates are minimax optimal in the case of the high-dimensional linear mean regression model; see Rigollet and Tsybakov (2011) and Chetverikov et al. (2016).

Remark 8 (**Derivatives**)**.** The CV residual estimator $\widehat{U}_i^{\mathsf{cv}}$ in (5.10) implicitly assumes the loss derivative m_1' tractable. If not, one may replace it with a numerical derivative, in which case the CV residual estimator is of the form $\widehat{U}_i^{\mathsf{cv}} = m_1'\big(X_i^\top \widehat{\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})}, Y_i\big)$. It may be possible to extend Theorem 4 to accommodate the additional error resulting from numerical differentiation. Of course, any numerical approach to differentiation introduces additional tuning parameters in the form of the difference method and step size. □

6 Extensions

The single-index structure in (1.1) allows simple notation but also rules out some settings of interest. For example, this structure does not accommodate multiple (unordered) response models, which are popular in applied work in economics and related fields. Extending the generic bootstrap and bootstrap-after-cross-validation (BCV) methods of Section 5 to richer modeling frameworks is, at least conceptually, straightforward. We describe the BCV method with a general loss function in Section 6.1. Extensions of the analytic method of Section 4 are more delicate and depend on the loss structure. We discuss analytic methods with multiple indices in Section 6.2 and illustrate the calculations involved through examples.

6.1 Bootstrapping-after-Cross-Validation with General Loss

Conceptually, the generic bootstrap method introduced in Section 5.1 has nothing to do with the index structure, and the implementation of our bootstrap-after-cross-validation (BCV) method (see Section 5.2) immediately extends beyond the single-index case. To see this, let $(w,\theta)\mapsto m_\theta(w)$ denote a general loss (as in Remark 2). The score then takes the form $S=\mathbb{E}_n[S_i]$, where $S_i := (\partial/\partial\theta)m_\theta(W_i)|_{\theta=\theta_0}$ denotes the ith score contribution (when it exists). The general BCV method is then summarized by the following pseudo code:

Pseudo Code: Bootstrap-after-Cross-Validation Method with General Loss

(1) **Cross-Validation:** Obtain $\widehat{\lambda}^{\mathsf{cv}}$ via K-fold cross-validation with folds $\{I_k\}_{k=1}^K$.

(2) **Estimate Score Contributions** S_i using the hold-out estimators:

$$\widehat{S}_i = \frac{\partial}{\partial\theta} m_\theta\left(W_i\right)\Big|_{\theta=\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})}, \quad i \in I_k, \quad k \in \{1, \ldots, K\}. \tag{6.1}$$

(3) **Estimate the** $(1-\alpha)$ **Score Quantile**: Holding the data $\{W_i\}_{i=1}^n$ fixed, calculate

$$\widehat{q}^{\mathsf{bcv}}\left(1-\alpha\right) = \left(1-\alpha\right)\text{-quantile of } \max_{1\leqslant j\leqslant p}\left|\mathbb{E}_n\left[e_i\widehat{S}_{ij}\right]\right|$$

via simulation of independent standard Gaussian multipliers $\{e_i\}_{i=1}^n$.

(4) **Declare Penalty:** $\widehat{\lambda}_\alpha^{\mathsf{bcv}} = c_0\widehat{q}^{\mathsf{bcv}}(1-\alpha)$.

The score-contribution estimates $\widehat{S}_i$ in (6.1) described in the above pseudo code take the place of $\widehat{U}_iX_i$ in (5.3).

6.2 Analytic Method with Multiple Indices

In this section we discuss extensions of the analytic method to losses with multiple indices. While additional work is needed to extend the formal result in Theorem 2, we here focus on modifying Assumptions 7 and 8 to arrive at an explicit penalty formula akin to (4.1).

In this section we consider a model where the true parameter value $\theta_0 := (\mathrm{vec}(\mathbf{\Delta}_0)^\top, \gamma_0^\top)^\top \in \mathbf{R}^{L_1p_1+p_2}$ follows from the (assumed unique) solution

$$(\mathbf{\Delta}_0, \gamma_0) = \operatorname*{argmin}_{\substack{\mathbf{\Delta}\in\mathbf{R}^{p_1\times L_1}\\ \gamma\in\mathbf{R}^{p_2}}} \mathrm{E}\left[m\left(Z^\top\mathbf{\Delta}, \gamma^\top\mathbf{V}, Y\right)\right],$$

where $\mathrm{vec}(\mathbf{A})$ denotes the column vector arising from stacking the columns of a matrix $\mathbf{A}$, and $m : \mathbf{R}^{L_1+L_2}\times\mathcal{Y} \to \mathbf{R}$ is now a known loss function that is convex in its first L_1+L_2 arguments. For simplicity, we also take m to be everywhere differentiable in its first $L_1 + L_2$ arguments. The first L_1 indices share the candidate regressors $Z := (Z_{i1}, \ldots, Z_{ip_1})^\top \in \mathbf{R}^{p_1}$, but each of these indices have unrelated parameter vectors $\delta_\ell = (\delta_{\ell 1}, \ldots, \delta_{\ell p_1})^\top, \ell \in \{1, \ldots, L_1\}$, here gathered in the matrix

$$\mathbf{\Delta} = [\delta_1 : \cdots : \delta_{L_1}] \in \mathbf{R}^{p_1\times L_1}.$$

The last L_2 indices share the parameters $\gamma \in \mathbf{R}^{p_2}$, but the candidate regressors $V_\ell := (V_{\ell 1}, \ldots, V_{\ell p_2})^\top$, here gathered in the matrix

$$\mathbf{V} = [V_1 : \cdots : V_{L_2}] \in \mathbf{R}^{p_2\times L_2},$$

may vary with the index $\ell \in \{1, \ldots, L_2\}$. As before, $Y \in \mathcal{Y}$ denotes one or more outcome variables.[12] Our extended modelling framework is motivated by the following multinomial response models.

Example 9 (**Multinomial Logit**). The *multinomial logit* models a discrete outcome $Y \in \{0, 1, \ldots, J\}$ as

$$\mathrm{P}\left(Y = k \middle| X\right) = \frac{\mathrm{e}^{X^\top \delta_{k0}}}{\sum_{h=0}^{J} \mathrm{e}^{X^\top \delta_{h0}}}, \quad k \in \{0, 1, \ldots, J\},$$

where $\delta_{00} = \mathbf{0}$ to allow model identification. The log-likelihood yields the loss function

$$m\left(t_1, \ldots, t_J, y\right) = \ln\left(1 + \sum_{h=1}^{J} \mathrm{e}^{t_h}\right) - \sum_{k=1}^{J} \mathbf{1}\left(y = k\right) t_k, \tag{6.2}$$

which is convex in $(t_1, \ldots, t_J)$. In this example, $L_1 = J$ and $L_2 = 0$. □

Example 10 (**Conditional Logit**). The *conditional logit* models a discrete outcome $Y \in \{0, 1, \ldots, J\}$ as

$$\mathrm{P}\left(Y = k \middle| \mathbf{X}\right) = \frac{\mathrm{e}^{X_k^\top \theta_0}}{\sum_{h=0}^{J} \mathrm{e}^{X_h^\top \theta_0}}, \quad k \in \{0, 1, \ldots, J\},$$

where we define the X_k's to be deviations of regressors relative to those of alternative zero (i.e., $X_0 \equiv \mathbf{0}$) and gather these in $\mathbf{X} \in \mathbf{R}^{p \times J}$. The log-likelihood yields the convex loss (6.2). In this example, $L_1 = 0$ and $L_2 = J$. □

Example 11 (**Mixed Logit**). The *mixed logit* models a discrete outcome $Y \in \{0, 1, \ldots, J\}$ as

$$\mathrm{P}\left(Y = k \middle| \mathbf{V}, Z\right) = \frac{\mathrm{e}^{V_k^\top \gamma_0 + Z^\top \delta_{k0}}}{\sum_{h=0}^{J} \mathrm{e}^{V_h^\top \gamma_0 + Z^\top \delta_{h0}}}, \quad k \in \{0, 1, \ldots, J\},$$

where, as for the multinomial logit (Example 9), we set $\delta_{00} = \mathbf{0}$ to allow model identification. The log-likelihood yields the loss function

$$m\left(t_1, \ldots, t_J, t_0', t_1', \ldots, t_J', y\right) = \widetilde{m}(t_0', t_1 + t_1', \ldots, t_J + t_J', y)$$

where $\widetilde{m}$ is given by

$$\widetilde{m}\left(\widetilde{t}_0, \widetilde{t}_1 \ldots, \widetilde{t}_J, y\right) = \ln\left(\sum_{h=0}^{J} \mathrm{e}^{\widetilde{t}_h}\right) - \sum_{k=0}^{J} \mathbf{1}\left(y = k\right) \widetilde{t}_k,$$

[12] For simplicity, we here employ the full parameter space $\mathbf{R}^{L_1 p_1 + p_2}$, a convex set for which θ_0 is interior.

a function convex in its first $1+J$ arguments. Since the mapping $(t_1,\dots,t_J,t_0',t_1',\dots,t_J')\mapsto(t_0',t_1+t_1',\dots,t_J+t_J')$ is linear, m is convex in its first $2J+1$ arguments. In this example, $L_1=J$ and $L_2=1+J$. □

The score $S\in\mathbf{R}^{L_1p_1+p_2}$ is now of the form

$$S=\mathbb{E}_n\big[\,(\partial/\partial\theta)m(Z_i^\top\mathbf{\Delta},\gamma^\top\mathbf{V}_i,Y_i)\big|_{\theta=\theta_0}\,\big]=\mathbb{E}_n\left[\begin{pmatrix}U_{i,1:L_1}\otimes Z_i\\ \mathbf{V}_iU_{i,L_1+1:L_1+L_2}\end{pmatrix}\right],$$

where

$$U_{i\ell}=m_\ell'\left(Z_i^\top\mathbf{\Delta}_0,\gamma_0^\top\mathbf{V}_i,Y_i\right),\quad \ell\in\{1,\dots,L_1+L_2\},$$

and $m_\ell'(t_1,\dots,t_{L_1+t_2},y)=(\partial/\partial t_\ell)m(t_1,\dots,t_{L_1+t_2},y)$ denotes the derivative of m with respect to the ℓ^{th} argument. Abbreviate $\mathbf{X}=(Z,\mathbf{V})$ and write $\mathcal{X}$ for its support. Our analytic method for models with multiple indices employs the following two assumptions.

Assumption 7′ (Conditional Mean Zero with Multiple Indices). *The residual vector* $U=(U_1,\dots,U_{L_1+L_2})^\top$ *with* $U_\ell=m_\ell'\left(Z^\top\mathbf{\Delta}_0,\gamma_0^\top\mathbf{V},Y\right),\ell\in\{1,\dots,L_1+L_2\}$, *is such that with probability one,* $\mathrm{E}\left[U|\,\mathbf{X}\right]=\mathbf{0}$.

Assumption 8′ (Residual: Analytic Method with Multiple Indices). *(1) For each* $\ell\in\{1,\dots,L_1\}$, *there exists* $\underline{r}_\ell,\overline{r}_\ell:\mathcal{X}\to\mathbf{R}$ *nonrandom and a known constant* d_ℓ *in* $\mathbf{R}_{++}$ *such that both* $U_\ell=m_\ell'\left(Z_i^\top\mathbf{\Delta}_0,\gamma_0^\top\mathbf{V}_i,Y_i\right)\in\left[\underline{r}_\ell\left(\mathbf{X}\right),\overline{r}_\ell\left(\mathbf{X}\right)\right]$ *and* $\overline{r}_\ell\left(\mathbf{X}\right)-\underline{r}_\ell\left(\mathbf{X}\right)\leqslant d_\ell$ *almost surely. (2) There exist known constants* $q\in[1,\infty]$ *and* $\widetilde{d}_{L_2}\in(0,\infty)$ *such that* $\|U_{L_1+1:L_1+L_2}\|_q\leqslant\widetilde{d}_{L_2}$ *almost surely.*

Under Assumptions 7′ and 8′.1, appealing to Hoeffding's inequality conditional on the $\mathbf{X}_i$'s and the union bound, we may reuse the argument leading to (4.1) to see that for any t in $\mathbf{R}_{++}$,

$$\mathrm{P}\Big(\max_{1\leqslant\ell\leqslant L_1}\max_{1\leqslant j\leqslant p_1}\left|S_{(\ell-1)p_1+j}\right|>t\Big|\,\{\mathbf{X}_i\}_{i=1}^n\Big)\leqslant 2L_1p_1\exp\left(-\frac{2nt^2}{d_{(L_1)}^2\max\limits_{1\leqslant j\leqslant p_1}\mathbb{E}_n\left[Z_{ij}^2\right]}\right)\text{ a.s.},$$

where we have introduced $d_{(L_1)}:=\max_{1\leqslant\ell\leqslant L_1}d_\ell$. The previous expression allows control over the score elements pertaining to indices with common regressors. To deal with the remaining indices we use Assumption 8′.2 and Hölder's inequality to see that the score contributions $S_{i,L_1p_1+j}=\sum_{\ell=1}^{L_2}U_{i,L_1+\ell}V_{i\ell j}$, $j\in\{1,\dots,p_2\}$, satisfy

$$|S_{i,L_1p_1+j}|\leqslant\|U_{i,L_1+1:L_1+L_2}\|_q\,\|V_{i\cdot j}\|_{q^*}\leqslant\widetilde{d}_{L_2}\,\|V_{i\cdot j}\|_{q^*}\text{ a.s.},\tag{6.3}$$

with $V_{i\cdot j}:=(V_{i1j},\dots,V_{iL_2j})^\top$ and $q^*\in[1,\infty]$ the exponent conjugate to q (i.e. $1/q+1/q^*=$

1). We may now appeal to Hoeffding's inequality and the union bound once more to see that, for any t in $\mathbf{R}_{++}$, also

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p_2} \left|S_{L_1p_1+j}\right| > t \middle| \{\mathbf{X}_i\}_{i=1}^n \right) \leqslant 2p_2 \exp\left(-\frac{nt^2}{2\widetilde{d}^2_{L_2} \max\limits_{1\leqslant j\leqslant p_2} \mathbb{E}_n\left[\|V_{i\cdot j}\|^2_{q^*} \right]} \right) \text{ a.s.} \tag{6.4}$$

Combining (6.3) and (6.4), by the union bound and inverting the resulting upper bound for a given probability tolerance α in $(0,1)$, we arrive at the data-dependent penalty level

$$\widehat{\lambda}^{\texttt{am}}_{\alpha} = c_0 \sqrt{\frac{\ln\left(2\left(L_1p_1+p_2\right)/\alpha\right)}{n} \max\left\{ \frac{d^2_{(L_1)}}{2} \max_{1\leqslant j\leqslant p_1} \mathbb{E}_n\left[Z^2_{ij}\right], 2\widetilde{d}^2_{L_2} \max_{1\leqslant j\leqslant p_2} \mathbb{E}_n\left[\|V_{i\cdot j}\|^2_{q^*} \right] \right\}}. \tag{6.5}$$

With only a single index, $L_1 = 1$ and $L_2 = 0$, and we recover the analytic penalty level in (4.1) upon relabeling. In this sense, (6.5) strictly generalizes the analytic method of Section 4 to the case of multiple indices. If several candidates for q and $\widetilde{d}_{L_2}$ are available, then we may choose the pair leading to smallest penalty level. This minimization idea is illustrated in Examples 10 and 11 below.

Example 9 (Multinomial Logit, Continued) The loss m in (6.2) has partial derivatives

$$m'_\ell\left(t_1,\ldots,t_J,y\right) = \frac{\mathrm{e}^{t_\ell}}{1+\sum_{h=1}^J \mathrm{e}^{t_h}} - \mathbf{1}\left(y=\ell\right), \quad \ell\in\{1,\ldots,J\}, \tag{6.6}$$

so each residual element

$$U_\ell = m'_\ell\left(X^\top \mathbf{\Delta}_0, Y\right) = \mathrm{P}\left(Y=\ell \middle| Z\right) - \mathbf{1}\left(Y=\ell\right), \quad \ell\in\{1,\ldots,J\},$$

lies in an interval of width $d_\ell = 1$. It follows that $d_{(J)} = 1$, and

$$\widehat{\lambda}^{\texttt{MNL}}_{\alpha} := c_0 \sqrt{\frac{\ln\left(2Jp/\alpha\right)}{2n} \max_{1\leqslant j\leqslant p} \mathbb{E}_n\left[X^2_{ij}\right]} \tag{6.7}$$

constitutes an analytic penalty for the multinomial logit. As p and J enter (6.7) only through a logarithmic term and a maximum, the analytic penalty $\widehat{\lambda}^{\texttt{MNL}}_{\alpha}$ allows both many common regressors and many alternatives. In the binary response case ($J = 1$), the expression (6.7) reduces to our previous expression for the binary logit (Example 1). □

Example 10 (Conditional Logit, Continued) The partial derivatives (6.6) yield residual elements

$$U_\ell = m'_\ell\left(\gamma_0^\top \mathbf{V}, Y\right) = \mathrm{P}\left(Y=\ell \middle| \mathbf{V}\right) - \mathbf{1}\left(Y=\ell\right), \quad \ell\in\{1,\ldots,J\},$$

which are probability differences. It follows that the ℓ^q-norm of the residual vector $U = (U_1, \dots, U_J)$ is bounded by the ℓ^q-diameter of the J-dimensional probability simplex, which is given by $2^{1/q}$. We may therefore use any $(q, \widetilde{d}_J)$ pair $(q, 2^{1/q}), q \in [1, \infty]$, and

$$\widehat{\lambda}_\alpha^{\texttt{CL}} = c_0 \sqrt{\frac{2\ln(2p/\alpha)}{n}} \inf_{q \in [1,\infty]} 2^{1/q} \sqrt{\max_{1 \leqslant j \leqslant p} \mathbb{E}_n \left[\|X_{i \cdot j}\|_{q^*}^2 \right]} \tag{6.8}$$

constitutes an analytic penalty for the conditional logit. The calculation of (6.8) requires optimization over $q \in [1, \infty]$. Since $q = 1$ is feasible, but not necessarily optimal, one may alternatively employ

$$\widetilde{\lambda}_\alpha^{\texttt{CL}} = 2c_0 \sqrt{\frac{2\ln(2p/\alpha)}{n}} \sqrt{\max_{1 \leqslant j \leqslant p} \mathbb{E}_n \left[\|X_{i \cdot j}\|_{\infty}^2 \right]}, \tag{6.9}$$

which is free of optimization, but may be larger. As p and J enter (6.9) only through a logarithmic term and maxima, respectively, both the analytic penalty $\widetilde{\lambda}_\alpha^{\texttt{CL}}$ and the possibly smaller $\widehat{\lambda}_\alpha^{\texttt{CL}}$ allow many alternatives and many alternative-varying regressors. In the binary response case ($J = 1$), $\|X_{i \cdot j}\|_{q^*} = |X_{i1j}|$ does not depend on q, and the infimum is attained at $q = \infty$. We then recover the analytic penalty for the binary logit (see Example 1) up to a factor of two.[13] □

Example 11 (Mixed Logit, Continued) Arguments parallel to the ones for the multinomial and conditional logit models (Examples 9 and 10) show that for the mixed logit each residual element $U_\ell = m'_\ell \left(Z^\top \mathbf{\Delta}_0, \gamma_0^\top \mathbf{V}, Y \right), \ell \in \{1, \dots, J\}$, lies in an interval of width $d_\ell = 1$, and the residual vector $U_{J+1:2J+1} = (U_{J+1}, \dots, U_{2J+1})$ is the difference between two probability vectors. It follows that we may take $d_{(J)} = 1$ alongside any $(q, \widetilde{d}_J)$ pair $(q, 2^{1/q}), q \in [1, \infty]$. Hence, an analytic penalty for the mixed logit is given by

$$\widehat{\lambda}_\alpha^{\texttt{ML}} = c_0 \sqrt{\frac{\ln(2(Jp_1 + p_2)/\alpha)}{n} \max \left\{ \frac{1}{2} \max_{1 \leqslant j \leqslant p_1} \mathbb{E}_n \left[Z_{ij}^2 \right], 2 \inf_{q \in [1,\infty]} 2^{2/q} \max_{1 \leqslant j \leqslant p_2} \mathbb{E}_n \left[\|V_{i \cdot j}\|_{q^*}^2 \right] \right\}}. \tag{6.10}$$

As for the conditional logit (Example 10), an alternative penalty free of optimization is given by

$$\widetilde{\lambda}_\alpha^{\texttt{ML}} = c_0 \sqrt{\frac{\ln(2(Jp_1 + p_2)/\alpha)}{n} \max \left\{ \frac{1}{2} \max_{1 \leqslant j \leqslant p_1} \mathbb{E}_n \left[Z_{ij}^2 \right], 8 \max_{1 \leqslant j \leqslant p_2} \mathbb{E}_n \left[\|V_{i \cdot j}\|_{\infty}^2 \right] \right\}}. \tag{6.11}$$

[13] The factor of two arises from using an absolute value bound instead of lower and upper bounds on the (then scalar) residual.

As p_1, p_2 and J all enter (6.11) through only a logarithmic term and maxima, both $\widehat{\lambda}_\alpha^{\mathtt{ML}}$ and $\widetilde{\lambda}_\alpha^{\mathtt{ML}}$ allow many common regressors, many alternative-varying regressors, and many alternatives.

□

7 Simulations

In this section we investigate the finite-sample behavior of estimators based on the analytic and bootstrap-based methods for obtaining penalty levels proposed in Sections 4 and 5, respectively and compare our penalty selection methods to already existing methods.

7.1 Simulation Design

For concreteness, we consider a data-generating process (DGP) of the form

$$Y_i = \mathbf{1}\left(\sum_{j=1}^{p} \theta_{0j} X_{ij} + \varepsilon_i > 0\right), \quad \varepsilon_i | \, X_{i1}, \ldots, X_{ip} \sim \text{Logistic}\,(0,1)\,, \quad i \in \{1, \ldots, n\}\,,$$

thus leading to a binary logit model. The regressors $X = (X_1, \ldots, X_p)$ are jointly centered Gaussian with a covariance matrix of the Toeplitz form

$$\text{cov}\,(X_{ij}, X_{ik}) = \text{E}\,[X_{ij} X_{ik}] = \rho^{|j-k|}, \quad j,k \in \{1, \ldots, p\}\,,$$

such that ρ determines the overall correlation level. We allow $\rho \in \{0, .1, \ldots, .9\}$, thus running the gamut of (positive) correlation levels. Since the ε_i's are standard Logistic, the "noise" in our DGP is fixed at $\text{var}\,(\varepsilon_i) = \pi^2/3 \approx 3.3$. However, the "signal" $\text{var}(\sum_{j=1}^{p} \theta_{0j} X_{ij}) = \theta_0^\top \text{E}\left[X_i X_i^\top\right] \theta_0$ depends on both the correlation level and coefficient pattern. We consider both sparse and dense coefficient patterns.

The *sparse* coefficient *pattern* has only nonzero coefficients for the first couple of regressors,

$$\text{Pattern 1 (Sparse):} \quad \theta_0 = (1, 1, 0, \ldots, 0)^\top\,,$$

thus yielding $s = 2$ relevant regressors among the p candidates. The implied signals are here given by

$$\text{var}\left(\sum_{j=1}^{p} \theta_{0j} X_{ij}\right) = 2\,(1+\rho) \in \{2, 2.2, \ldots, 3.8\}\,,$$

further implying a signal-to-noise ratio (SNR) range of about .6 to about 1.2. Compared to existing simulations studies for the high-dimensional logit, the signals considered here are rather low.[14]

The *dense* coefficient *pattern* have all nonzero coefficients,

$$\text{Pattern 2 (Dense):} \quad \theta_{0j} = \left(1/\sqrt{2}\right)^{j-1}, \quad j \in \{1, \ldots, p\},$$

thus implying $s = p$. The base $(1/\sqrt{2})$ was here chosen to (approximately) equate the signals arising from the dense and sparse coefficient patterns in the baseline case of uncorrelated regressors ($\rho = 0$), which, in turn, amounts to $\|\theta_0\|_2^2$. We consider sample sizes $n \in \{100, 200, 400\}$ and limit attention to the high-dimensional regime by fixing $p = n$ throughout.

With a sparse coefficient pattern, the nonzero coefficients are well separated from zero and should be relatively easy to detect—at least with larger sample sizes. With a dense coefficient pattern, every regressor is in principle relevant, and our implicit assumption of exact sparsity fails ($s = p = n$). Note, however, that the relevance of the regressors, as measured by their coefficient, is rapidly decaying in the regressor index, such that the vast majority of the signal is still captured by a small fraction of the regressors. For example, in the baseline case of uncorrelated regressors ($\rho = 0$), the first 10 regressors account for 99.9 pct. of the total signal, and the model may be interpreted as effectively sparse.

Specifically, with the dense coefficient pattern

$$\sum_{j=1}^{p} |\theta_{0j}|^q \leqslant \frac{1}{2^{q/2} - 1} \text{ for all } q \in (0, 1],$$

so θ_0 is approximately sparse in that it lies in every ℓ^q "ball" $\mathbb{B}_q(R_q), q \in (0,1]$, each of which has fixed "radius" $R_q := 1/(1 - 2^{-q/2})$. A suitable extension of Theorem 1 therefore ought to apply. (See Remark 4 for discussion and definitions.) However, as R_q is quite large for small $q > 0$, one might also expect the estimation error to be relatively large.

7.2 Estimators

Both analytic and bootstrap-after-cross-validation methods require specifying the constant c_0. Unless otherwise noted, we here set $c_0 = 1.1$, which reflects one of the standard recommendations in the LASSO literature (see, e.g., Belloni and Chernozhukov, 2011a). We also specify the probability tolerance as $\alpha = \alpha_n = 10/n$, thus leading to an α of $10, 5$ and 2.5 percent for $n = 100, 200$ and 400, respectively. We consider three feasible estimators based on the analytic and bootstrap methods. With our binary logit design, Assumption 8

[14] For example, the design in Friedman, Hastie, and Tibshirani (2010, Section 5.2) implies a SNR of three. In Ng (2004, Section 5), the SNR is over 30.

is satisfied with $d = 1$. (See Example 1.) We specify the penalty using the *analytic method* in (4.1), and $\widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}})$ constitutes our *first* estimator. Our *second* estimator is based on the *bootstrap method* (5.3) initiated with residual estimates resulting from the analytic method,

$$\widehat{U}_i^{\mathtt{am}} = \Lambda\big(X_i^\top \widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}})\big) - Y_i, \quad i \in \{1, \dots, n\}.$$

Since the logit loss is globally Lipschitz, the justification for this plug-in residual estimator essentially follows from the consistency of $\widehat{\theta}(\widehat{\lambda}_\alpha^{\mathtt{am}})$ and the light tails of the normal distribution. See Remark 7 and the discussion following Assumption 5 for details. Our *third* estimator follows similarly, except that we initiate the bootstrap method with *cross-validation residual estimates* (5.10).[15] To introduce a positive benchmark, we consider the infeasible estimator arising from the bootstrap method using the true residuals. We refer to the latter three bootstrap-based estimators as *bootstrapping after the analytic method* (BAM), *bootstrapping after cross validation* (BCV), and the *oracle bootstrap* (Oracle).[16]

All simulations are carried out in Matlab® with optimization and cross validation done using the user-contributed `glmnet` package.[17] For each sample size $n(= p)$, each correlation level ρ, and each coefficient pattern (sparse or dense), we use 2,000 simulation draws and 1,000 standard Gaussian bootstrap draws per simulation draw (when applicable). In constructing the candidate penalty set Λ_n, we use the `glmnet` default setting, which constructs a log-scale equi-distant grid of a 100 candidate penalties from the threshold penalty level to essentially zero. The threshold is the (approximately) smallest level of penalization needed to set every coefficient to zero, thus resulting in a trivial (null) model.[18]

7.3 Simulation Results

Figure 7.1 shows the mean ℓ^2 estimation error (over the 2,000 simulation draws) as a function of the correlation level ρ for each of the three bootstrap-based estimators and each sample size $n(= p)$ obtained with a sparse coefficient pattern. We also include a line (Zeros) at $\|\theta_0\|_2$, which allows for comparison with the trivial "estimator" $\widehat{\theta} = \mathbf{0}$. For each of the

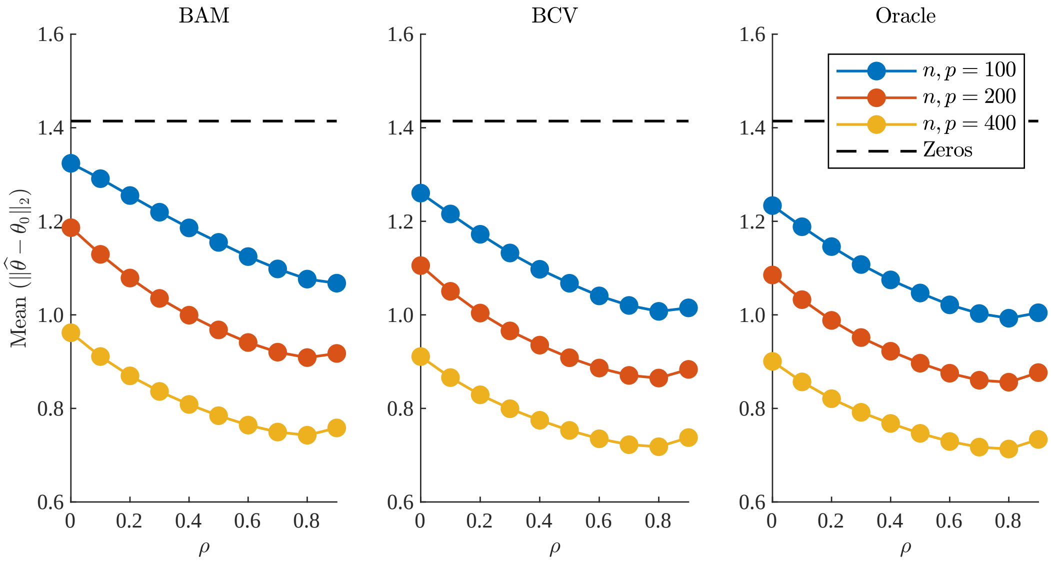
[15] We use 10-fold cross validation, splitting the data evenly, and assign folds according to (5.6) to ensure replicability. As a result, $K = 10$ and $c_D = \frac{1}{10}$.

[16] For implementation details for the BAM and BCV methods more broadly, see Appendix A.

[17] We use the August 30, 2013 version of `glmnet` for Matlab®, available for download at https://web.stanford.edu/~hastie/glmnet_matlab/. Cross validation is done using `cvglmnet`, which automatically stores the out-of-fold predictions $X_i^\top \widehat{\theta}_{I_k^c}(\lambda), i \in I_k$, for each candidate penalty.

[18] Log-scale equi-distance from a "large" candidate value to essentially zero fits well with the form of Λ_n in our Assumption 12 (interpreting $c_\Lambda/n \approx 0$). However, the threshold penalty is a function of the data and, thus, random. The resulting candidate penalty set used in our simulations is therefore also random, and thus, strictly speaking, not allowed by Assumption 12. Moreover, the number of candidate values is here held fixed. We believe these deviations from our theory to be only a minor issue.

Figure 7.1: Consistency of Bootstrap-Based Estimators with Exact Sparsity



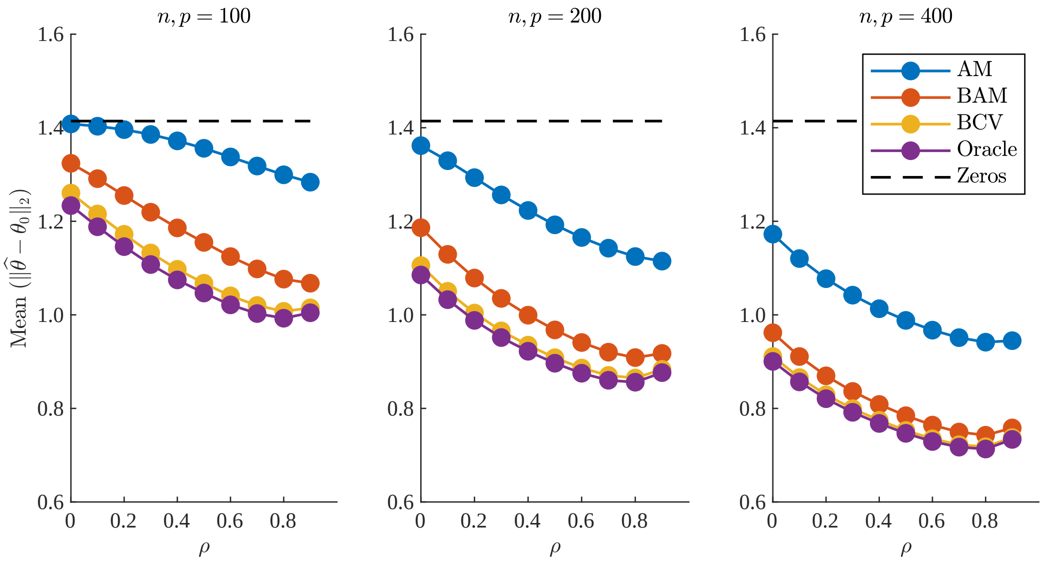
bootstrap-based estimators, we see that the mean estimation error decreases with sample size. Convergence appears to take place even though the number of candidate regressors matches the sample size and no matter the level of regressor correlation.[19] This finding indicates that our bootstrap method is useful for high-dimensional estimation, not only in the best-case scenario where residuals are observed, but also when residuals are estimated by a pilot method—whether it be analytic or computational.

Figure 7.2 rearranges the plots in Figure 7.1 in order to facilitate comparison of the various estimators, now including the estimator based on the analytic method (AM). For each of the three sample sizes/numbers of candidate regressors, we see that the oracle performs better than the other two bootstrap-based estimators. Bootstrapping after cross validation appears to outperform bootstrapping after the analytic method, which, in turn, improves greatly upon the analytic method itself. While residual estimation comes at a price, bootstrapping after cross-validation achieves near-oracle performance even with our smallest sample size—and is essentially indistinguishable from the oracle at $n = 400$. Bootstrapping after the analytic method here comes in close second place among the feasible estimators, which indicates that BAM provides a computationally inexpensive way of obtaining quality results.

Figures 7.3 and 7.4 reproduce Figures 7.1 and 7.2, respectively, with results stemming from the dense coefficient pattern (approximate sparsity). The plots in Figure 7.3 are also indicative of consistency, although convergence is slowed down by the weaker form of sparsity

[19] That mean error is downward sloping for moderate ρ levels is due to a higher signal (7.1) and need not translate to other simulation designs. The observed increase in mean error as ρ approaches one is likely due to the margin condition (Assumption 4) not being appropriate for highly correlated designs.

Figure 7.2: Comparing Estimators with Exact Sparsity



(compare with Figure 7.1). The ranking of estimators in Figure 7.2 is preserved in Figure 7.4. These findings suggest that our methods remain relevant under a less stringent assumption than exact sparsity.

We next investigate the impact of the choice of c_0. Figures 7.5 and 7.6 plot the mean ℓ^2 estimation error for $c_0 = 1, 1.05$, and (the previously used) 1.1, each sample size and coefficient pattern and for the BAM and BCV estimators, respectively. Our finite-sample experiments suggest that increasing c_0 away from one worsens (mean) performance. However, while our theory takes $c_0 > 1$, any value near one—including the limit case of one itself—appears to lead to near identical results. Similar findings were mentioned in Belloni et al. (2012) for the case of the LASSO and linear model.

We also compare our analytic and bootstrap methods to existing penalty methods formally justifiable in our binary logit model. Specifically, we here compare with the analytic penalty levels provided in Bunea (2008b, Theorem 2.4), van de Geer (2008, Theorem 2.1) and van de Geer (2016, Theorem 12.1). Across all of our designs and simulation draws, the *smallest* Bunea (2008a) penalty is larger than the *largest* van de Geer (2008) penalty, which, in turn, is similar in size to her (2016) penalty level. We therefore restrict attention to the latter. In our notation, the van de Geer (2016, Theorem 12.1) penalty takes the form

$$\widehat{\lambda}_{\alpha}^{\texttt{vdG16}} = 8c_0\sqrt{\frac{2\ln(2p/\alpha)}{n}\max_{1\leqslant j\leqslant p}\mathbb{E}_n\left[X_{ij}^2\right]}, \tag{7.1}$$

which is nothing more than 16 times our analytic penalty level (4.1). (Recall that $d = 1$.)

Figure 7.3: Consistency of Bootstrap-Based Estimators with Approximate Sparsity

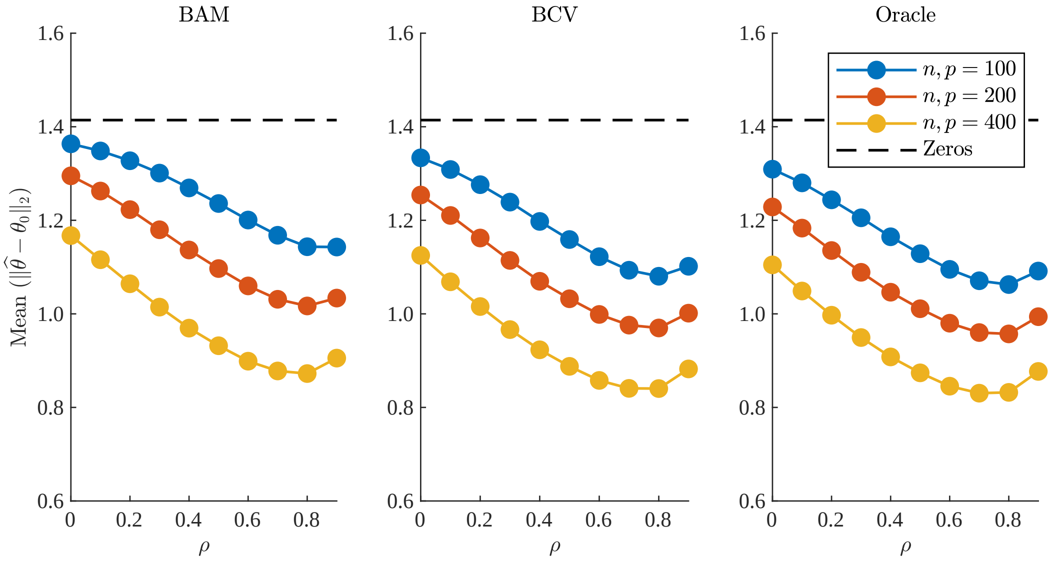


Figure 7.4: Comparing Estimators with Approximate Sparsity

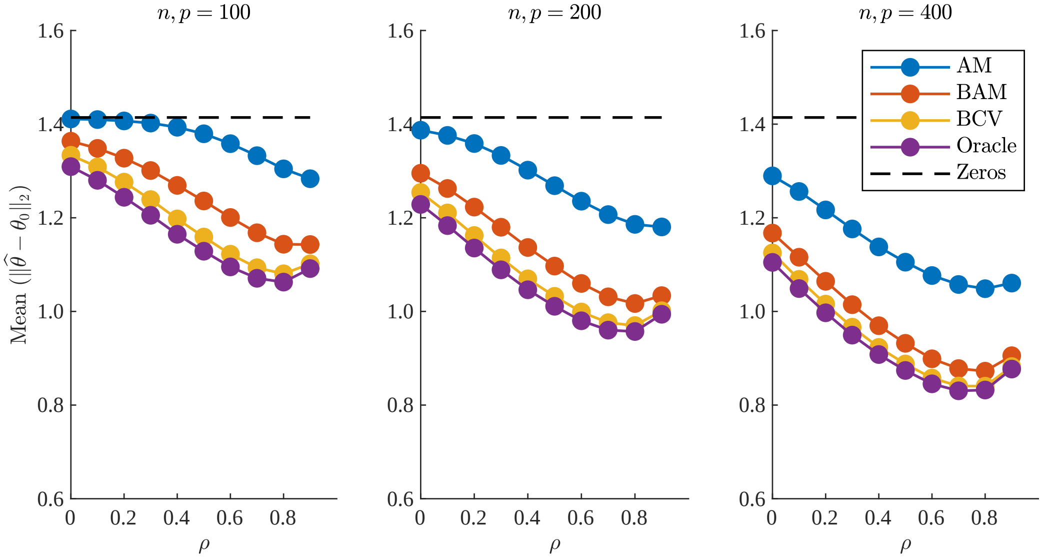


Figure 7.5: Bootstrapping after the Analytic Method for Different c_0

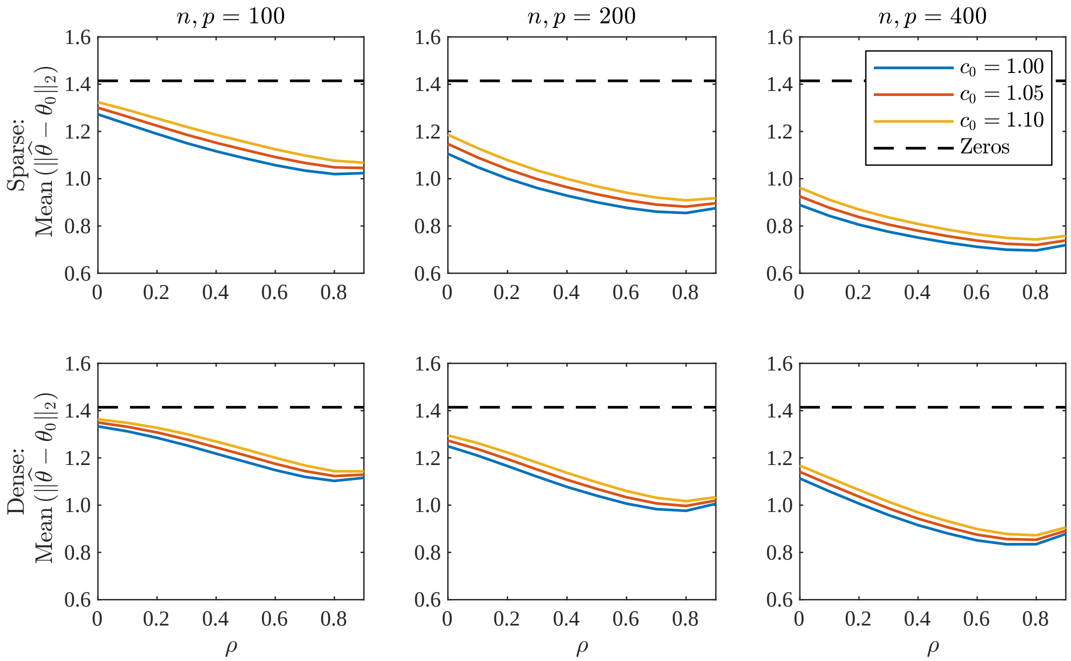


Figure 7.6: Bootstrapping after Cross-Validation for Different c_0

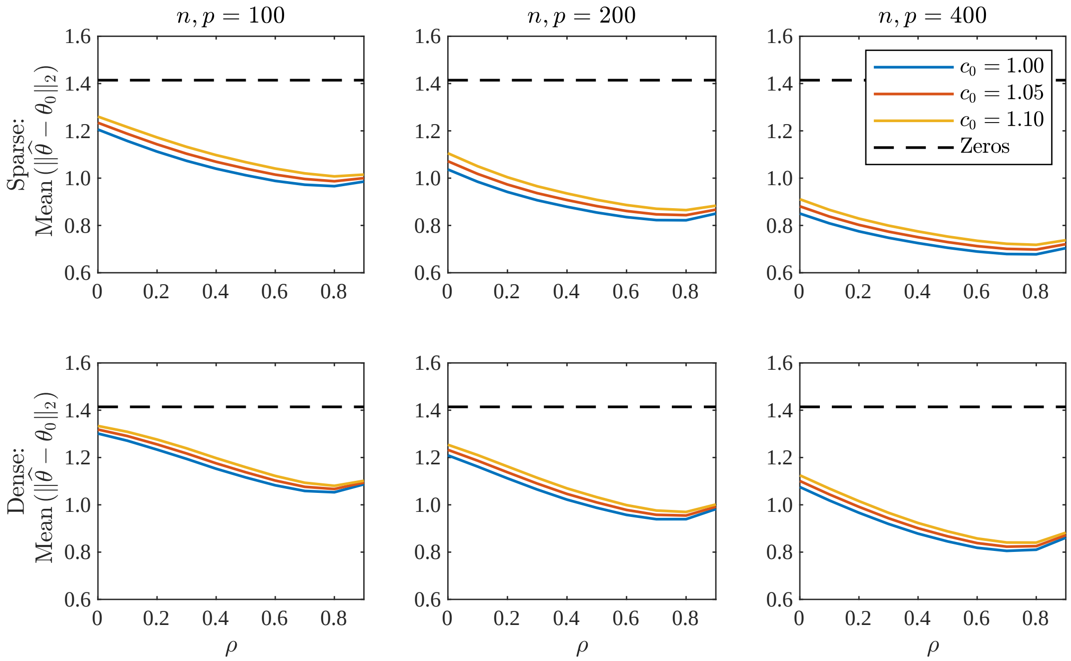


Figure 7.7: Kernel Density Estimates of Penalty Distributions

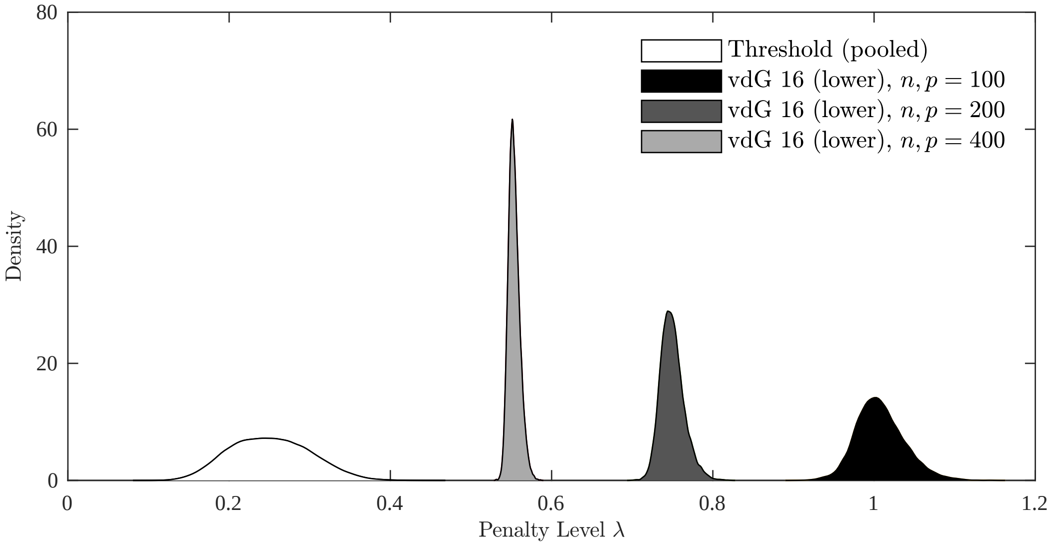
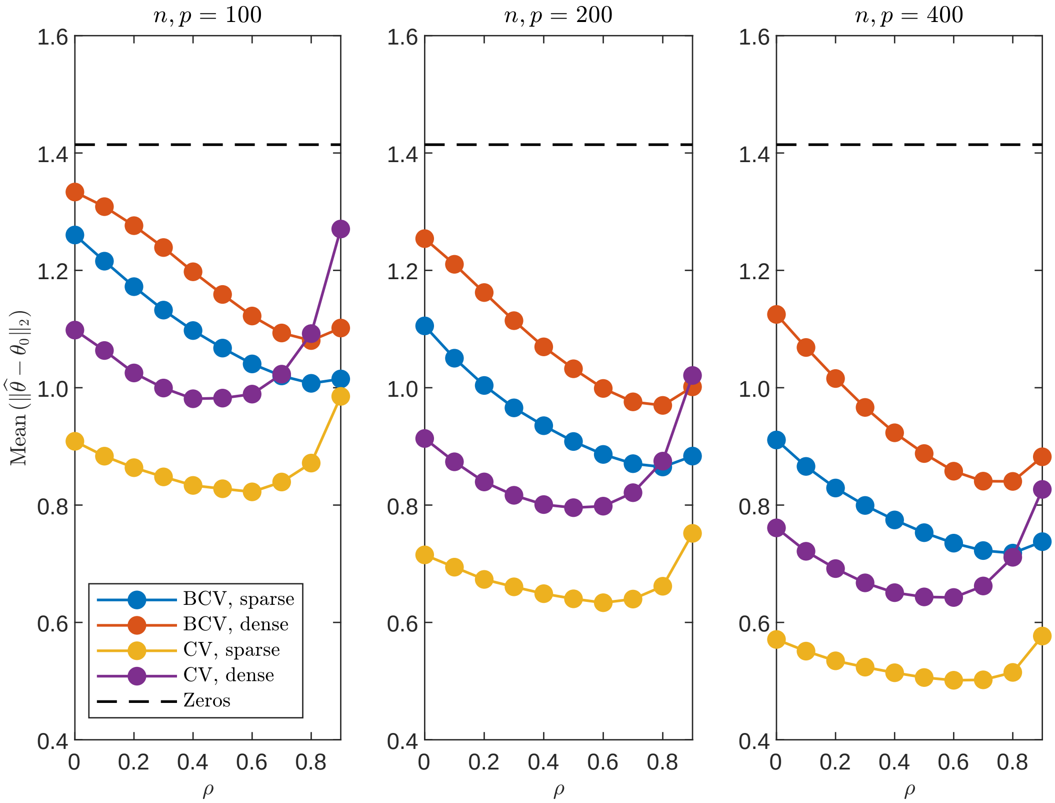


Figure 1.1 in the Introduction displays the distribution of the van de Geer penalty level as a function of the sample size $n(= p)$, pooling over both correlation levels and coefficient patterns. For comparison, we include the distribution of the threshold penalty pooled over all designs.[20] The latter threshold is the (approximately) smallest level of penalization needed to set every coefficient to zero, thus resulting in a trivial (null) model. The figure shows that the distribution of the threshold penalty is an order of magnitude closer to the origin than the van de Geer penalty. As a consequence, the latter penalty results in a trivial model estimate across *all* of our designs and simulation draws. The estimators resulting from the Bunea and van de Geer penalties (with $c_0 = 1.1$) are therefore all represented by the "Zeros" lines in Figures 7.1–7.6 and 7.8–7.9. Inspection of the proof underlying van de Geer (2016, Theorem 12.1) suggests that the factor of 8 in (7.1) may be reduced to a 2, when restricting attention to our framework. However, even with this lower bound on the multiplier, the supports of these penalty distributions remain separated (Figure 7.7).

Our findings should not be interpreted as a critique of these authors, whose work were intended as primarily of theoretical interest. For example, van de Geer (2008, p. 621) explicitly states that other penalty choices should be used in practice. It is, however, not immediately clear how one should modify the penalty choices of these authors without disconnecting theory from practice. In contrast, the simulation results of this section demonstrate that our analytic and bootstrap methods are not only theoretically justifiable, but also practically useful.

[20] All density estimates in Figures 1.1 and 7.7 were created using the Matlab® package `ksdensity` with default settings.

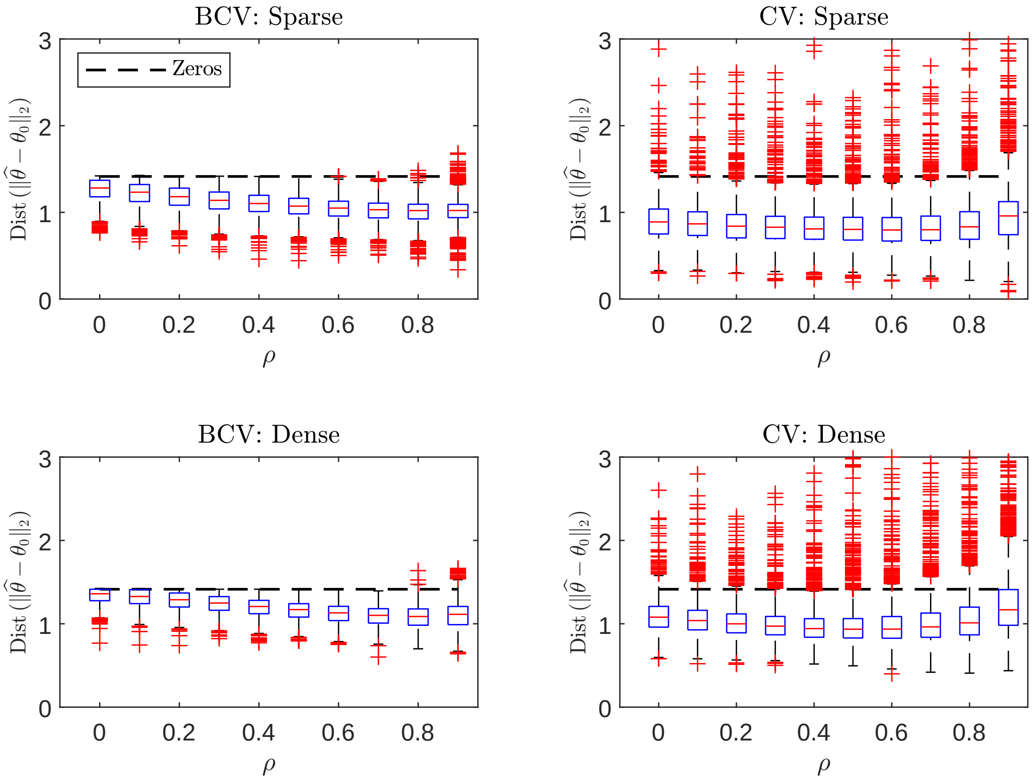
Figure 7.8: Cross-Validation and Bootstrapping after Cross-Validation ($c_0 = 1.1$)



As a final exercise, in Figures 7.8 and 7.9 we compare the estimators proposed in this paper to the CV estimator $\widehat{\theta}(\widehat{\lambda}^{\texttt{cv}})$. The latter estimator lacks formal justification but is popular in practice. Since BCV outperforms BAM in our simulation setting (see Figures 7.2 and 7.4), we limit attention to the former. Figure 7.8 shows that CV tends to do somewhat better than BCV under sparsity. However, in the case of a dense coefficient pattern, the two estimators cannot be ranked by their mean errors. Moreover, the two estimators differ in terms of more than their mean error, as illustrated in Figure 7.9, summarizing the empirical distributions of ℓ^2-error for the case $n = p = 100$ using box-and-whisker plots.[21] Both in the case of sparse and dense coefficient patterns, CV may take on quite extreme values, thus resulting in high variance. Taken together, these figures show that neither method dominates the other. While we do not formally allow for a dense coefficient pattern, being designed to dominate the score, our methods appear to translate well to this more challenging modelling framework, both formulaically and in simulations.

[21] All box-and-whisker plots were created using the Matlab® package `boxplot` with default settings.

Figure 7.9: Cross-Validation and Bootstrapping after Cross-Validation ($c_0 = 1.1$), $n, p = 100$.



Notes: (i) In the box-and-whisker plots, the central mark indicates the median, and the bottom and top edges of the box correspond to the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' marker symbol. (ii) To improve visibility, the upper limit on the range of the secondary axis has been set to three. As a result, 23 and 59 data points do not appear in the figures for CV in the case of the sparse and dense coefficient pattern, respectively. No data points were dropped for BCV.

References

Adler, R. J. and J. E. Taylor (2007): *Random fields and geometry*, Springer Science & Business Media.

Aigner, D. J., T. Amemiya, and D. J. Poirier (1976): "On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function," *International Economic Review*, 377–396.

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

Belloni, A. and V. Chernozhukov (2011a): *High dimensional sparse econometric models: An introduction*, Springer.

——— (2011b): "l1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, 39, 82–130.

——— (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018a): "High-dimensional econometrics and regularized GMM," *Working Paper*.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018b): "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in Z-Estimation Framework," *The Annals of Statistics*, 3643–3675.

Bertsekas, D. P. (1973): "Stochastic optimization problems with nondifferentiable cost functionals," *Journal of Optimization Theory and Applications*, 12, 218–231.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 1705–1732.

Boucheron, S., G. Lugosi, and P. Massart (2012): *Concentration inequalities: A nonasymptotic theory of independence*, Clarendon press, Oxford.

Buchinsky, M. and J. Hahn (1998): "An alternative estimator for the censored quantile regression model," *Econometrica*, 653–671.

Bühlmann, P. and S. van de Geer (2011): *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Bunea, F. (2008a): "Consistent selection via the Lasso for high dimensional approximating regression models," in *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, Institute of Mathematical Statistics, 122–137.

——— (2008b): "Honest variable selection in linear and logistic regression models via l1 and l1+ l2 penalization," *Electronic Journal of Statistics*, 2, 1153–1194, publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Chamberlain, G. (1984): "Panel data," *Handbook of econometrics*, 2, 1247–1318.

——— (1985): "Heterogeneity, omitted variable bias, and duration dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer, Cambridge University Press, 3–38.

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

——— (2017): "Central limit theorems and bootstrap in high dimensions," *The Annals of Probability*, 45, 2309–2352.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2016): "On cross-validated Lasso," *arXiv preprint arXiv:1605.02214*.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, 33, 1.

Hastie, T., R. Tibshirani, and M. Wainwright (2015): *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press.

Honoré, B. E. (1992): "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects," *Econometrica: journal of the Econometric Society*, 533–565.

Imai, K. and M. Ratkovic (2014): "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 243–263.

Lancaster, T. (1992): *The econometric analysis of transition data*, 17, Cambridge university press.

LECUE, G. AND G. MITCHELL (2012): "Oracle inequalities for cross-validation type procedures," *Electronic Journal of Statistics*, 1803–1837.

MIOLANE, L. AND A. MONTANARI (2018): "The distribution of the Lasso: uniform control over sparse balls and adaptive parameter tuning," *arXiv:1811.01212*.

NEGAHBAN, S., P. RAVIKUMAR, M. WAINWRIGHT, AND B. YU (2012): "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*.

NEWEY, W. K. AND J. L. POWELL (1987): "Asymmetric least squares estimation and testing," *Econometrica: Journal of the Econometric Society*, 819–847.

NG, A. Y. (2004): "Feature selection, L 1 vs. L 2 regularization, and rotational invariance," in *Proceedings of the twenty-first international conference on Machine learning*, ACM, 78.

NINOMIYA, Y. AND S. KAWANO (2016): "AIC for the Lasso in generalized linear models," *Electronic Journal of Statistics*.

PRATT, J. W. (1981): "Concavity of the log likelihood," *Journal of the American Statistical Association*, 76, 103–106, publisher: Taylor & Francis.

RASCH, G. (1960): "Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests." .

RIGOLLET, P. AND A. TSYBAKOV (2011): "Exponential screening and optimal rates of sparse estimation," *Annals of Statistics*, 731–771.

ROCKAFELLAR, R. T. (1970): "Convex Analysis (Princeton Mathematical Series)," *Princeton University Press*, 46, 49.

TALAGRAND, M. (2010): *Mean field models for spin glasses: Volume I: Basic examples*, vol. 54, Springer Science & Business Media.

TAN, Z. (2017): "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data," *arXiv:1710.08074v1*.

TSYBAKOV, A. B. (2004): "Optimal aggregation of classifiers in statistical learning," *The Annals of Statistics*, 32, 135–166, publisher: Institute of Mathematical Statistics.

VAN DE GEER, S. (2016): "Estimation and testing under sparsity," *Lecture notes in mathematics*, 2159, publisher: Springer.

van de Geer, S. A. (2008): “High-Dimensional Generalized Linear Models and the Lasso,” *The Annals of Statistics*, 36, 614–645.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Vershynin, R. (2018): *High-dimensional probabilty: An introduction with applications in data science*, Cambridge university press.

Wainwright, M. (2019): *High-dimensional statistics: a non-asymptotic viewpoint*, Cambridge university press.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

Appendix

We split the appendix into five parts. Appendix A contains implementation details. In Appendix B, we provide low-level conditions that are sufficient for Assumptions 4, 6, 13, and 14 in each of the examples considered in the main text. In Appendix C, we provide proofs of the results stated in the main text. In Appendix D, we provide auxiliary proofs. In Appendix E, we provide a collection of technical tools used to prove the main results.

A Implementation Details

We organize the implementation details concerning our bootstrap penalty methods into algorithms. Both the bootstrap-after-analytic (BAM) and bootstrap-after-cross-validation (BCV) methods require a choice of the markup $c_0 > 1$ and probability tolerance $\alpha \in (0,1)$. The BCV method additionally requires on a choice $K \geqslant 2$ of folds, partition $I_1, \dots, I_K$ of $\{1,\dots,n\}$, and candidate penalty set $\Lambda_n \subset \mathbf{R}_{++}$. In our simulations (Section 7) we used $c_0 = 1.1$, $\alpha = 10/n$, $K = 10$, the I_k in (5.6), and set Λ_n equal to a log-scale equi-distant grid of a 100 candidate penalties from essentially zero to (approximately) the smallest level of penalization needed to set every coefficient to zero. In the binary logit model the (approximate) threshold is $\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[(.5 - Y_i)X_{ij}]|$.

Algorithm 1 (Bootstrap after Analytic Method). (a) Analytically derive a $d \in \mathbf{R}_{++}$ (preferably the smallest possible) such that Assumption 8 holds. (b) Calculate the analytic penalty $\widehat{\lambda}^{\mathtt{am}}_{\alpha}$ using (4.1), the estimates $\widehat{\theta}(\widehat{\lambda}^{\mathtt{am}}_{\alpha})$ resulting from the analytic method using (1.2), and the implied residual estimates $\widehat{U}^{\mathtt{am}}_i = m_1'\big(X_i^\top \widehat{\theta}(\widehat{\lambda}^{\mathtt{am}}_{\alpha}), Y_i\big)$. (c) Calculate the quantile $\widehat{q}^{\mathtt{bam}}(1-\alpha)$ via simulation using the Gaussian multiplier bootstrap (5.3) with $\widehat{U}_i = \widehat{U}^{\mathtt{am}}_i$, and set $\widehat{\lambda}^{\mathtt{bam}}_{\alpha} = c_0 \widehat{q}^{\mathtt{bam}}(1-\alpha)$.

Algorithm 2 (Bootstrap-after-Cross-Validation Method). (a) Calculate the cross-validated penalty level $\widehat{\lambda}^{\mathtt{cv}}$ in (5.9), and produce residual estimates $\widehat{U}^{\mathtt{cv}}_i$ as in (5.10) using subsample estimates $\widehat{\theta}_{I_k^c}(\lambda), \lambda \in \Lambda_n$, in (5.8), $k = 1,\dots,K$. (b) Calculate the quantile $\widehat{q}^{\mathtt{bcv}}(1-\alpha)$ via simulation using the Gaussian multiplier bootstrap (5.3) and $\widehat{U}_i = \widehat{U}^{\mathtt{cv}}_i$, and set $\widehat{\lambda}^{\mathtt{bcv}}_{\alpha} = c_0 \widehat{q}^{\mathtt{bcv}}(1-\alpha)$.

B Verification of High-Level Assumptions

In this section, we verify Assumptions 4, 6, 13 and 14 in each of the examples from Section 2. Throughout this section, we use $m_1'(t,y)$ and $m_{11}''(t,y)$ to denote the first and the second derivatives of the function $t \mapsto m(t,y)$, whenever they exist. We suppose that Assumptions 1, 2 and 7 hold and that there exists a constant C_d in $\mathbf{R}_{++}$ such that

$$\|\theta\|_1 \leqslant C_d \quad \text{for all } \theta \in \Theta. \tag{B.1}$$

In addition, we suppose that there exist constants c_{ev} and C_{ev} in $\mathbf{R}_{++}$ such that

$$c_{ev} \leqslant \lambda_{\min}(\mathrm{E}[XX^\top]) \leqslant \lambda_{\max}(\mathrm{E}[XX^\top]) \leqslant C_{ev}, \tag{B.2}$$

where $\lambda_{\min}(\mathrm{E}[XX^\top])$ and $\lambda_{\max}(\mathrm{E}[XX^\top])$ are the smallest and the largest eigenvalues of the matrix $\mathrm{E}[XX^\top]$. Moreover, we suppose that there exists a finite constant C_X in $\mathbf{R}_{++}$ such that $\|X\|_\infty \leqslant C_X$ with probability one, which amounts to a special case of Assumption 5. This last assumption can be avoided but helps make other assumptions more transparent.

Before we proceed, we make two observations that are helpful to verify the assumptions of interest. First, to verify Assumption 4, it suffices to show that the function $f(t,x) := \mathrm{E}[m(t,Y)|X=x]$, defined for all $(t,x) \in \mathbf{R} \times \mathcal{X}$ has the following two properties: (i) its derivative $f_1'(t,x)$ with respect to t exists and is bounded on $[-C_XC_d, C_XC_d] \times \mathcal{X}$, and in addition (ii) for all $x \in \mathcal{X}$, the function $t \mapsto f_1'(t,x)$ is absolutely continuous on $[-C_XC_d, C_XC_d]$ with (a version of) the derivative $t \mapsto f_{11}''(t,x)$ being bounded below from zero by some constant c_1 in $\mathbf{R}_{++}$, which is independent of x. Indeed, since $M(\theta_0) \leqslant M(\theta)$ for all $\theta \in \Theta$, it is possible to show that under the aforementioned assumptions, $\mathrm{E}[f_1'(X^\top\theta_0, X)X^\top(\theta-\theta_0)] = 0$ for all $\theta \in \Theta$.[22] Therefore, by the first-order Taylor expansion with remainder in the integral form, for all $\theta \in \Theta$,

$$\begin{aligned} M(\theta) - M(\theta_0) = \mathrm{E}[f(X^\top\theta, X) - f(X^\top\theta_0, X)] &= \mathrm{E}\left[\int_{X^\top\theta_0}^{X^\top\theta} f_{11}''(t,X)(X^\top\theta - t)dt\right] \\ &= \mathrm{E}\left[\int_0^1 f_{11}''(X^\top\theta_0 + sX^\top(\theta-\theta_0), X)\{X^\top(\theta-\theta_0)\}^2(1-s)ds\right] \\ &\geqslant c_1\mathrm{E}\left[\{X^\top(\theta-\theta_0)\}^2\right]\int_0^1(1-t)dt \geqslant c_1c_{ev}\|\theta-\theta_0\|_2^2/2, \end{aligned}$$

which yields Assumption 4 with $c_M = c_1c_{ev}/2$ and arbitrary c_M'.

[22] To show this claim, fix any $\theta \in \Theta$ and observe that $\theta_0 + s(\theta-\theta_0) \in \Theta$ for all $s \in [0,1]$ by convexity (Assumption 1). Therefore, $\mathrm{E}[f(X^\top\theta_0 + sX^\top(\theta-\theta_0), X)] = M(\theta_0 + s(\theta-\theta_0)) \geqslant M(\theta_0) = \mathrm{E}[f(X^\top\theta_0, X)]$. Also, by the mean-value theorem, $f(x^\top\theta_0 + sx^\top(\theta-\theta_0)) - f(x^\top\theta_0, x) = sf_1'(x^\top\tilde\theta_x, x)x^\top(\theta-\theta_0)$ for some $\tilde\theta_x$ on the line connecting θ_0 and $\theta_0 + s(\theta-\theta_0)$ for all $x \in \mathcal{X}$. Hence, $s\mathrm{E}[f_1'(X^\top\tilde\theta_X, X)X^\top(\theta-\theta_0)] \geqslant 0$. Thus, given that $f_1' : [-C_XC_d, C_XC_d] \times \mathcal{X} \to \mathbf{R}$ is continuous in t and bounded, sending $s \to 0$ and applying the dominated convergence theorem gives $\mathrm{E}[f_1'(X^\top\theta_0, X)X^\top(\theta-\theta_0)] \geqslant 0$, which is the lower bound. To obtain the upper bound, observe that by interiority (Assumption 1), there exists $s_0 < 0$ such that $\theta_0 + s_0(\theta-\theta_0) \in \Theta$. The upper bound then follows from the same argument as that used above with $s \in [s_0, 0]$.

Second, to verify Assumptions 13 and 14, it is helpful to note that under Assumption 4 and (B.1), there exists a constant c_e in $\mathbf{R}_{++}$ such that

$$\mathcal{E}(\theta) \geqslant c_e \|\theta - \theta_0\|_2^2 \quad \text{for all } \theta \in \Theta. \tag{B.3}$$

To see why this bound holds, fix any $\theta \in \Theta$ and observe that if $\|\theta - \theta_0\|_1 \leqslant c'_M$, then $\mathcal{E}(\theta) \geqslant c_M \|\theta - \theta_0\|_2^2$ by Assumption 4. Therefore, we only need to consider the case $\|\theta - \theta_0\|_1 > c'_M$. In this case, for $t := c'_M / \|\theta - \theta_0\|_1$, by convexity (Assumption 2) we have that

$$t\mathcal{E}(\theta) + (1-t)\mathcal{E}(\theta_0) \geqslant \mathcal{E}(t\theta + (1-t)\theta_0),$$

and so, given that $\mathcal{E}(\theta_0) = 0$, it follows from (B.1) that

$$\mathcal{E}(\theta) \geqslant \frac{\mathcal{E}(\theta_0 + t(\theta - \theta_0))}{t} \geqslant \frac{c_M t^2 \|\theta - \theta_0\|_2^2}{t} = \frac{c_M c'_M \|\theta - \theta_0\|_2^2}{\|\theta - \theta_0\|_1} \geqslant \frac{c_M c'_M}{2C_d} \|\theta - \theta_0\|_2^2.$$

This gives (B.3) with $c_e = c_M \wedge (c_M c'_M / 2C_d)$.

We are now in the position to verify Assumptions 4, 6, 13, and 14 in each of the examples from Section 2. We proceed example by example.

Example 1 (Binary Response Model, Continued). For simplicity, we only consider logit and probit models. In the case of the logit loss function (2.2), we have

$$f(t,x) = \mathrm{E}[m(t,Y) | X = x] = \ln(1 + e^t) - \Lambda(x^\top \theta_0) t$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Here $f'_1(t,x) = \Lambda(t) - \Lambda(x^\top\theta_0)$ is bounded and absolutely continuous in t with derivative $f''_{11}(t,x) = \Lambda(t)(1 - \Lambda(t))$ not depending on x. Since $f''_{11}(t,x)$ is bounded away from zero on any bounded set, including $[-C_X C_d, C_X C_d] \times \mathcal{X}$, Assumption 4 is satisfied by the discussion in the beginning of the section. Further, we have $m'_1(t,y) = \Lambda(t) - y$ and $m''_{11}(t,y) = \Lambda(t)(1 - \Lambda(t))$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$. Therefore, $|m'_1(t,y)| \leqslant 1$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$, and so Assumption 6.1 is satisfied with $c_L \in \mathbf{R}_{++}$ arbitrary, $L(w) = 1$ for all $w = (x,y) \in \mathcal{W}$, and $C_L = 2$. Also, for all $\theta \in \Theta$, we have

$$\begin{aligned} \mathrm{E}\left[\left\{m\left(X^\top\theta, Y\right) - m\left(X^\top\theta_0, Y\right)\right\}^2\right] &\leqslant \mathrm{E}\left[m'_1(X^\top \widetilde{\theta}_{X,Y}, Y)^2 |X^\top(\theta - \theta_0)|^2\right] \\ &\leqslant \mathrm{E}\left[|X^\top(\theta - \theta_0)|^2\right] \leqslant C_{ev} \|\theta - \theta_0\|_2^2 \leqslant (C_{ev}/c_e)\mathcal{E}(\theta), \end{aligned}$$

where the first inequality follows from the mean-value theorem with $\widetilde{\theta}_{X,Y}$ being a value on the line connecting θ_0 and θ, the third from (B.2), and the fourth from (B.3). Increasing C_L if necessary, these bounds yield both Assumptions 6.2 and 14. Moreover, since $|m'_1(t,y)| \leqslant 1$

for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$, Assumption 13.1 is satisfied by Hoeffding's lemma (Boucheron et al., 2012, Lemma 2.2). Finally, since $|m''_{11}(t,y)| \leqslant 1$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$, Assumption 13.2 is satisfied since for all $\theta \in \Theta$, we have

$$\begin{aligned}\mathrm{E}\left[\left\{m'_1(X^\top\theta, Y) - m'_1(X^\top\theta_0, Y)\right\}^2\right] &= \mathrm{E}\left[m''_{11}(X^\top\widetilde{\theta}_{X,Y}, Y)^2 |X^\top(\theta-\theta_0)|^2\right] \\ &\leqslant \mathrm{E}\left[|X^\top(\theta-\theta_0)|^2\right] \leqslant C_{ev}\|\theta-\theta_0\|_2^2 \leqslant (C_{ev}/c_e)\mathcal{E}(\theta),\end{aligned}$$

where the equality follows from the mean-value theorem with $\widetilde{\theta}_{X,Y}$ being a value on the line connecting θ_0 and θ, the second inequality from (B.2), and the third from (B.3).

In the case of the probit loss function (2.3), we have

$$f(t,x) = \mathrm{E}[m(t,Y)|X=x] = -\Phi(x^\top\theta_0)\ln\Phi(t) - (1-\Phi(x^\top\theta_0))\ln(1-\Phi(t))$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Here, letting φ denote the PDF of the standard normal distribution,

$$f'_1(t,x) = -\frac{\Phi(x^\top\theta_0)\varphi(t)}{\Phi(t)} + \frac{(1-\Phi(x^\top\theta_0))\varphi(t)}{1-\Phi(t)}$$

is bounded on $[-C_XC_d, C_XC_d] \times \mathcal{X}$ and absolutely continuous in t on $[-C_XC_d, C_XC_d]$ with derivative

$$f''_{11}(t,x) = \frac{\Phi(x^\top\theta_0)t\varphi(t)}{\Phi(t)} + \frac{\Phi(x^\top\theta_0)\varphi(t)^2}{\Phi(t)^2} - \frac{(1-\Phi(x^\top\theta_0))t\varphi(t)}{1-\Phi(t)} + \frac{(1-\Phi(x^\top\theta_0))\varphi(t)^2}{[1-\Phi(t)]^2}.$$

By (1.2.2) in Adler and Taylor (2007), $f''_{11}(t,x)$ is strictly positive and continuous for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Hence, $f''_{11}(t,x)$ is bounded away from zero on $[-C_XC_d, C_XC_d] \times \mathcal{X}$, and so Assumption 4 is satisfied by the argument given in the beginning of the section. Further, we have

$$m'_1(t,y) = -\frac{y\varphi(t)}{\Phi(t)} + \frac{(1-y)\varphi(t)}{1-\Phi(t)}$$

and

$$m''_{11}(t,y) = \frac{yt\varphi(t)}{\Phi(t)} + \frac{y\varphi(t)^2}{\Phi(t)^2} - \frac{(1-y)t\varphi(t)}{1-\Phi(t)} + \frac{(1-y)\varphi(t)^2}{[1-\Phi(t)]^2}$$

for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$. Since $\mathcal{Y} = \{0,1\}$ and the functions $t \mapsto m'_1(t,0)$, $t \mapsto m'_1(t,1)$, $t \mapsto m''_{11}(t,0)$, and $t \mapsto m''_{11}(t,1)$ are all continuous, it follows that both $m'_1(t,y)$ and $m''_{11}(t,y)$ are bounded in absolute value on the compact $[-C_XC_d, C_XC_d] \times \mathcal{Y}$. Therefore, Assumptions 6, 13, and 14 are satisfied by the same arguments as those used in the logit case. □

Example 2 (Ordered Response Model, Continued). For simplicity, we only consider the ordered logit model. Since $m(t,y) = -\sum_{j=0}^{J} \mathbf{1}(y=j)\ln(\Lambda(\alpha_{j+1}-t) - \Lambda(a_j - t))$ in this case,

$f(t,x) = \mathrm{E}[m(t,Y)|X=x]$ is given by

$$f(t,x) = -\sum_{j=0}^{J} \left[\Lambda(\alpha_{j+1} - x^\top\theta_0) - \Lambda(\alpha_j - x^\top\theta_0)\right] \ln\left(\Lambda(\alpha_{j+1} - t) - \Lambda(\alpha_j - t)\right)$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. (Recall that we interpret $\Lambda(-\infty)$ as zero and $\Lambda(+\infty)$ as one.) Here,

$$f_1'(t,x) = \sum_{j=0}^{J} \left[\Lambda(\alpha_{j+1} - x^\top\theta_0) - \Lambda(\alpha_j - x^\top\theta_0)\right] \left[1 - \Lambda(\alpha_{j+1} - t) - \Lambda(\alpha_j - t)\right]$$

is bounded and absolutely continuous in t with derivative

$$f_{11}''(t,x) = \sum_{j=0}^{J} \left[\Lambda(\alpha_{j+1} - x^\top\theta_0) - \Lambda(\alpha_j - x^\top\theta_0)\right] \times \big(\Lambda(\alpha_{j+1} - t)\left[1 - \Lambda(\alpha_{j+1} - t)\right] + \Lambda(\alpha_j - t)\left[1 - \Lambda(\alpha_j - t)\right]\big).$$

Since $f_{11}''(t,x)$ is bounded away from zero on $[-C_X C_d, C_X C_d] \times \mathcal{X}$, Assumption 4 is satisfied by the argument given in the beginning of the section.

Further, we have

$$m_1'(t,y) = \sum_{j=0}^{J} \mathbf{1}(y=j)\left[1 - \Lambda(\alpha_{j+1} - t) - \Lambda(\alpha_j - t)\right]$$

and

$$m_{11}''(t,y) = \sum_{j=0}^{J} \mathbf{1}(y=j)\big(\Lambda(\alpha_{j+1} - t)(1 - \Lambda(\alpha_{j+1} - t)) + \Lambda(\alpha_j - t)(1 - \Lambda(\alpha_j - t))\big)$$

for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$. Since both $m_1'(t,y)$ and $m_{11}''(t,y)$ are bounded in absolute value on $\mathbf{R} \times \mathcal{Y}$, Assumptions 6, 13, and 14 are satisfied by the same arguments as those used in the logit case of Example 1. □

Example 3 (Logistic Calibration, Continued). Since $m(t,y) = y\mathrm{e}^{-t} + (1-y)t$ in this example,

$$f(t,x) = \mathrm{E}[m(t,Y)|X=x] = \Lambda(x^\top\theta_0)\mathrm{e}^{-t} + (1 - \Lambda(x^\top\theta_0))t$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Here, $f_1'(t,x) = -\Lambda(x^\top\theta_0)\mathrm{e}^{-t} + 1 - \Lambda(x^\top\theta_0)$ is bounded and absolutely continuous in t on $[-C_X C_d, C_X C_d] \times \mathcal{X}$ with derivative $f_{11}''(t,x) = \Lambda(x^\top\theta_0)\mathrm{e}^{-t}$. Since $f_{11}''(t,x)$ is bounded away from zero on $[-C_X C_d, C_X C_d] \times \mathcal{X}$, Assumption 4 is satisfied by the argument given in the beginning of the section.

Further, we have $m_1'(t,y) = -y\mathrm{e}^{-t} + 1 - y$ and $m_{11}''(t,y) = y\mathrm{e}^{-t}$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$. Since $\mathcal{Y} = \{0,1\}$ and the functions $t \mapsto m_1'(t,0)$, $t \mapsto m_1'(t,1)$, $t \mapsto m_{11}''(t,0)$, and $t \mapsto m_{11}''(t,1)$ are all continuous, it follows that both $m_1'(t,y)$ and $m_{11}''(t,y)$ are bounded in absolute value on $[-C_X C_d, C_X C_d] \times \mathcal{Y}$. Therefore, it follows that Assumptions 6, 13, and 14 are satisfied by the same arguments as those used in the logit case of Example 1. □

Example 4 (Logistic Balancing, Continued). Since $m(t,y) = (1-y)\mathrm{e}^{t} + y\mathrm{e}^{-t} + (1-2y)t$ in this example,

$$f(t,x) = \mathrm{E}[m(t,Y)|X=x] = \left[1 - \Lambda(x^\top\theta_0)\right]\mathrm{e}^{t} + \Lambda(x^\top\theta_0)\mathrm{e}^{-t} + \left[1 - 2\Lambda(x^\top\theta_0)\right]t$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Here, $f_1'(t,x) = [1 - \Lambda(x^\top\theta_0)]\mathrm{e}^{t} - \Lambda(x^\top\theta_0)\mathrm{e}^{-t} + 1 - 2\Lambda(x^\top\theta_0)$ is bounded and absolutely continuous in t on $[-C_X C_d, C_X C_d] \times \mathcal{X}$ with derivative $f_{11}''(t,x) = [1 - \Lambda(x^\top\theta_0)]\mathrm{e}^{t} + \Lambda(x^\top\theta_0)\mathrm{e}^{-t}$. Since $f_{11}''(t,x)$ is bounded away from zero on $[-C_X C_d, C_X C_d] \times \mathcal{X}$, Assumption 4 is satisfied by the argument given in the beginning of the section.

Further, we have $m_1'(t,y) = (1-y)\mathrm{e}^{t} - y\mathrm{e}^{-t} + 1 - 2y$ and $m_{11}''(t,y) = (1-y)\mathrm{e}^{t} + y\mathrm{e}^{-t}$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$. Since $\mathcal{Y} = \{0,1\}$ and the functions $t \mapsto m_1'(t,0)$, $t \mapsto m_1'(t,1)$, $t \mapsto m_{11}''(t,0)$, and $t \mapsto m_{11}''(t,1)$ are all continuous, it follows that both $m_1'(t,y)$ and $m_{11}''(t,y)$ are bounded in absolute value on $[-C_X C_d, C_X C_d] \times \mathcal{Y}$. Therefore, it follows that Assumptions 6, 13, and 14 are satisfied by the same arguments as those used in the logit case of Example 1. □

Example 5 (Expectile Model, Continued). Throughout this example, we assume that for all $x \in \mathcal{X}$, the conditional distribution of Y given $X = x$ is absolutely continuous with PDF $g_{Y|X=x}$ and that there exists a constant C in $\mathbf{R}_{++}$ such that

$$\mathrm{P}\left(|Y| \geqslant t \middle| X = x\right) \leqslant 2\exp\left(-t^2/C\right), \quad \text{for all } t \in \mathbf{R}_{++} \text{ and } x \in \mathcal{X}. \tag{B.4}$$

Since $m(t,y) = |\tau - \mathbf{1}(y - t < 0)|(y-t)^2$ in this example,

$$f\left(t,x\right) = (1-\tau)\int_{-\infty}^{t}(y-t)^2 g_{Y|X=x}(y)\mathrm{d}y + \tau\int_{t}^{+\infty}(y-t)^2 g_{Y|X=x}(y)\mathrm{d}y$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Here,

$$f_1'(t,x) = 2(1-\tau)\int_{-\infty}^{t}(t-y)g_{Y|X=x}(y)\mathrm{d}y + 2\tau\int_{t}^{+\infty}(t-y)g_{Y|X=x}(y)\mathrm{d}y$$

is bounded on $[-C_X C_d, C_X C_d] \times \mathcal{X}$ since

$$|f_1'(t,x)| \leqslant 2\int_{-\infty}^{+\infty} (|t|+|y|) g_{Y|X=x}(y)\mathrm{d}y \leqslant 2C_X C_d + 2\overline{C}$$

for all $(t,x) \in [-C_X C_d, C_X C_d] \times \mathcal{X}$ and some constant $\overline{C}$ in $\mathbf{R}_{++}$, where we have used (B.4). In addition, $f_1'(t,x)$ is absolutely continuous in t on $[-C_X C_d, C_X C_d] \times \mathcal{X}$ with derivative

$$f_{11}''(t,x) = 2(1-\tau)\int_{-\infty}^{t} g_{Y|X=x}(y)\mathrm{d}y + 2\tau\int_{t}^{+\infty} g_{Y|X=x}(y)\mathrm{d}y$$

Since $f_{11}''(t,x) \geqslant 2[\tau \wedge (1-\tau)]$ for all $(t,x) \in [-C_X C_d, C_X C_d] \times \mathcal{X}$, Assumption 4 is satisfied by the argument given in the beginning of the section.

Further, we have $m_1'(t,y) = 2|\tau - \mathbf{1}(y-t<0)|(t-y)$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$ and $m_{11}''(t,y) = 2|\tau - \mathbf{1}(y-t<0)|$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$ such that $y \neq t$. Hence, Assumption 6.1 is satisfied for any c_L in $\mathbf{R}_{++}$ and $L(w) = 2(c_L + |y|)$ for all $w = (x,y) \in \mathcal{W}$, where the inequality $\mathrm{E}[|L(W)|^8] \leqslant (C_L/2)^8$ holds for some constant C_L in $\mathbf{R}_{++}$ again using (B.4). Also, there exists a constant $\widetilde{C}$ in $\mathbf{R}_{++}$ such that for all θ in Θ, we have

$$\begin{aligned}
\mathrm{E}\left[\left\{m\left(X^\top\theta, Y\right) - m\left(X^\top\theta_0, Y\right)\right\}^2\right] &\leqslant \mathrm{E}\left[m_1'(X^\top\tilde{\theta}_{X,Y}, Y)^2 |X^\top(\theta-\theta_0)|^2\right] \\
&\leqslant 4\mathrm{E}\left[(Y - X^\top\tilde{\theta}_{X,Y})^2 |X^\top(\theta-\theta_0)|^2\right] \\
&\leqslant 8\mathrm{E}\left[(Y^2 + (X^\top\tilde{\theta}_{X,Y})^2) |X^\top(\theta-\theta_0)|^2\right] \\
&\leqslant \widetilde{C}\mathrm{E}\left[|X^\top(\theta-\theta_0)|^2\right] \leqslant (\widetilde{C}C_{ev}/c_e)\mathcal{E}(\theta),
\end{aligned}$$

where the first line follows from the mean-value theorem with $\tilde{\theta}_{X,Y}$ being a value on the line connecting θ_0 and θ, and the fourth line follows from (B.1), (B.2), (B.3), and (B.4). This gives Assumptions 6.2 and 14. Moreover, under (B.4), Assumption 13.1 follows from Proposition 2.5.2 in Vershynin (2018). Finally,

$$\begin{aligned}
\mathrm{E}\left[\left\{m_1'\left(X^\top\theta, Y\right) - m_1'\left(X^\top\theta_0, Y\right)\right\}^2\right] &= \mathrm{E}\left[\left(\int_{X^\top\theta_0}^{X^\top\theta} m_{11}''(t,Y)dt\right)^2\right] \\
&\leqslant 4\mathrm{E}\left[|X^\top(\theta-\theta_0)|^2\right] \leqslant (4C_{ev}/c_e)\mathcal{E}(\theta),
\end{aligned}$$

where the second line follows from (B.2) and (B.3). This gives Assumption 13.2. □

Example 6 (Panel Logit Model, Continued). Since $m(t,y) = \mathbf{1}(y_1 \neq y_2)[\ln(1+\mathrm{e}^t) - y_1 t]$ in this example,

$$f(t,x)=\mathrm{E}[m(t,Y)|X=x]=\mathrm{P}(Y_1\neq Y_2|X=x)\ln(1+\mathrm{e}^t)-\mathrm{P}(Y_1=1,Y_2=0|X=x)t$$

for all $(t,x)\in\mathbf{R}\times\mathcal{X}$. Here,

$$f_1'(t,x)=\mathrm{P}(Y_1\neq Y_2|X=x)\Lambda(t)-\mathrm{P}(Y_1=1,Y_2=0|X=x)$$

is bounded and absolutely continuous in t on $[-C_XC_d,C_XC_d]\times\mathcal{X}$ with derivative $f_{11}''(t,x)=\mathrm{P}(Y_1\neq Y_2|X=x)\Lambda(t)[1-\Lambda(t)]$. Therefore, assuming that $\mathrm{P}(Y_1\neq Y_2|X=x)$ is bounded away from zero on $\mathcal{X}$, it follows that $f_{11}''(t,x)$ is bounded away from zero on $[-C_XC_d,C_XC_d]\times\mathcal{X}$, and so Assumption 4 is satisfied by the argument given in the beginning of the section.

Further, we have $m_1'(t,y)=\mathbf{1}(y_1\neq y_2)[\Lambda(t)-y_1]$ and $m_{11}''(t,y)=\mathbf{1}(y_1\neq y_2)\Lambda(t)[1-\Lambda(t)]$ for all $(t,y)\in\mathbf{R}\times\mathcal{Y}$. Therefore, $|m_1'(t,y)|\leqslant 1$ and $|m_{11}''(t,y)|\leqslant 1$ for all $(t,y)\in\mathbf{R}\times\mathcal{Y}$, and so Assumptions 6, 13, and 14 are satisfied by the same arguments as those used in the logit case of Example 1. □

Example 7 (Panel Censored Model, Continued). Denote latent outcomes $Y_\tau^*:=\gamma+X_\tau^\top\theta_0+\varepsilon_\tau$ for $\tau=1,2$. Recall that the loss function m given in (2.9) depends on Ξ. We will consider the cases $\Xi=|\cdot|$ and $(\cdot)^2$ separately, starting with $\Xi=|\cdot|$. In this case, we assume that for all x in $\mathcal{X}$, the conditional distribution of (Y_1^*,Y_2^*) given $X=x$ is absolutely continuous with PDF $g_{(Y_1^*,Y_2^*)|X=x}$, and that there exists a constant c in $\mathbf{R}_{++}$ such that $\int_0^\infty g_{(Y_1^*,Y_2^*)|X=x}(s+t,s)\mathrm{d}s\geqslant c$ for all $(t,x)\in[0,C_XC_d]\times\mathcal{X}$ and $\int_0^\infty g_{(Y_1^*,Y_2^*)|X=x}(s,s-t)\mathrm{d}s\geqslant c$ for all $(t,x)\in[-C_XC_d,0]\times\mathcal{X}$. In addition, we assume that the PDF of the conditional distribution of Y_2^* given (Y_1^*,X_1,X_2) and the PDF of the conditional distribution of Y_1^* given (Y_2^*,X_1,X_2) are both bounded from above by some constant C in $\mathbf{R}_{++}$.

Then it follows from Lemma A.1 in Honoré (1992) that for the function $f(t,x)=\mathrm{E}[m(t,Y)|X=x]$ defined on $\mathbf{R}\times\mathcal{X}$, we have

$$f_1'(t,x)=\mathrm{E}[\mathbf{1}(Y_2>0,Y_2>Y_1-t)|X=x]-\mathrm{E}[\mathbf{1}(Y_1>0,Y_1>Y_2+t)|X=x]$$

for all $(t,x)\in\mathbf{R}\times\mathcal{X}$. Thus, $|f_1'(t,x)|\leqslant 1$ for all $(t,x)\in\mathbf{R}\times\mathcal{X}$. In addition, it follows from the proof of Lemma A.3 in Honoré (1992) that $f_1'(t,x)$ is absolutely continuous in t on $\mathbf{R}\times\mathcal{X}$ with derivative $f_{11}''(t,x)$ satisfying

$$f_{11}''(t,x)\geqslant\begin{cases}2\int_0^\infty g_{(Y_1,Y_2)|X=x}(s+t,s)\mathrm{d}s, & t\geqslant 0,\\ 2\int_0^\infty g_{(Y_1,Y_2)|X=x}(s,s-t)\mathrm{d}s, & t<0.\end{cases}$$

Thus, $f''_{11}(t,x) \geqslant c > 0$ for all $(t,x) \in [-C_X C_d, C_X C_d] \times \mathcal{X}$, and so Assumption 4 is satisfied by the discussion in the beginning of the section. A calculation shows that $m(\cdot, y)$ is (globally) Lipschitz with Lipschitz constant equal to one, and that

$$m'_1(t,y) = \begin{cases} 0, & y_1 = y_2 = 0, \\ \mathbf{1}(t > -y_2), & y_1 = 0, y_2 > 0, \\ -\mathbf{1}(t < y_1), & y_1 > 0, y_2 = 0, \\ -1 + 2\mathbf{1}(t > y_1 - y_2), & y_1 > 0, y_2 > 0, \end{cases}$$

for (Lebesgue) almost every $(t,y) \in \mathbf{R} \times \mathcal{Y}$. Assumptions 6, 13.1 and 14 are then satisfied by the same arguments as those used in the logit case of Example 1. Moreover, for all $\theta \in \Theta$, denoting $t_1 := \min((X_1 - X_2)^\top \theta, (X_1 - X_2)^\top \theta_0)$ and $t_2 := \max((X_1 - X_2)^\top \theta, (X_1 - X_2)^\top \theta_0)$, we have

$$\begin{aligned} &\mathrm{E}\left[\left\{m'_1\left((X_1 - X_2)^\top \theta, Y\right) - m'_1\left((X_1 - X_2)^\top \theta_0, Y\right)\right\}^2\right] \\ &\quad \leqslant \mathrm{P}(Y_1 = 0, Y_2 > 0, -t_2 < Y_2 \leqslant -t_1) \\ &\qquad + \mathrm{P}(Y_1 > 0, Y_2 = 0, t_1 < Y_1 \leqslant t_2) \\ &\qquad + 4\mathrm{P}(Y_1 > 0, Y_2 > 0, t_1 \leqslant Y_1 - Y_2 < t_2) \\ &\quad \leqslant \mathrm{P}(-t_2 < Y_2^* \leqslant -t_1) + \mathrm{P}(t_1 < Y_1^* \leqslant t_2) + 4\mathrm{P}(t_1 \leqslant Y_1^* - Y_2^* < t_2) \\ &\quad \leqslant 6C\mathrm{E}\left[|(X_1 - X_2)^\top(\theta - \theta_0)|\right] \leqslant 6C\left(\mathrm{E}\left[|(X_1 - X_2)^\top(\theta - \theta_0)|^2\right]\right)^{1/2} \leqslant \widetilde{C}\sqrt{\mathcal{E}(\theta)}, \end{aligned}$$

with $\widetilde{C} = 6C\sqrt{C_{ev}/c_e}$, where the first inequality in the last line follows from our assumption on the conditional distributions of Y_1^* given (Y_2^*, X_1, X_2) and of Y_2^* given (Y_1^*, X_1, X_2), the second from Jensen's inequality, and the third from (B.2) and (B.3). This gives Assumption 13.2.

Next, consider the case $\Xi = (\cdot)^2$. In this case, we assume that there exist constants c and C in $\mathbf{R}_{++}$ such that $\mathrm{P}(-Y_2 < t < Y_1 | X = x) \geqslant c$ and $\mathrm{P}(Y_1 \vee Y_2 \leqslant t | X = x) \leqslant 2\exp(-t^2/C)$ for all $(t,x) \in [-C_X C_d, C_X C_d] \times \mathcal{X}$. Then it follows from Lemma A.1 in Honoré (1992) that for the function $f(t,x) = \mathrm{E}[m(t,Y)|X = x]$ defined on $\mathbf{R} \times \mathcal{X}$, we have

$$f'_1(t,x) = 2\mathrm{E}[Y_2 \mathbf{1}(t \geqslant Y_1) - Y_1 \mathbf{1}(t \leqslant -Y_2) + (Y_1 - Y_2 + t)\mathbf{1}(-Y_2 < t < Y_1)|X = x]$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Thus, under our assumptions, $|f'_1(t,x)| \leqslant \widetilde{C}$ for all $(t,x) \in [-C_X C_d, C_X C_d] \times \mathcal{X}$ and some constant $\widetilde{C}$ in $\mathbf{R}_{++}$. In addition, it follows from Lemma A.3 in Honoré (1992) that $f'_1(t,x)$ is absolutely continuous in t on $\mathbf{R} \times \mathcal{X}$ with derivative

$f''_{11}(t,x) = 2\mathrm{P}(-Y_2 < t < Y_1 | X = x)$. Thus, under our assumptions, $f''_{11}(t,x) \geqslant c/2 > 0$ for all $(t,x) \in [-C_X C_d, C_X C_d] \times \mathcal{X}$, and so Assumption 4 is satisfied by the argument given in the beginning of this section. Further, we have

$$m'_1(t,y) = \begin{cases} -2y_1, & t \leqslant -y_2, \\ 2(t + y_2 - y_1), & -y_2 < t < y_1, \\ 2y_2, & y_1 \leqslant t, \end{cases}$$

and

$$m''_{11}(t,y) = \begin{cases} 0, & t < -y_2, \\ 2, & -y_2 < t < y_1, \\ 0, & y_1 < t. \end{cases}$$

Hence, Assumption 6.1 is satisfied with any constant c_L in $\mathbf{R}_{++}$ and $L(w) = 2(|y_1| + |y_2|)$ for all $w = (x,y) \in \mathcal{W}$, where the inequality $\mathrm{E}[|L(W)|^8] \leqslant (C_L/2)^8$ holds for some constant C_L in $\mathbf{R}_{++}$ by our assumptions. Also, there exists a constant $\widetilde{C}$ in $\mathbf{R}_{++}$ such that for all $\theta \in \Theta$, we have

$$\begin{aligned} &\mathrm{E}\left[\left\{m\left(X^\top\theta, Y\right) - m\left(X^\top\theta_0, Y\right)\right\}^2\right] \\ &\quad \leqslant \mathrm{E}\left[m'_1(X^\top\tilde{\theta}_{X,Y}, Y)^2 |X^\top(\theta - \theta_0)|^2\right] \\ &\quad \leqslant 4\mathrm{E}\left[(Y_1 \vee Y_2)^2 |X^\top(\theta - \theta_0)|^2\right] \\ &\quad \leqslant \widetilde{C}\mathrm{E}\left[|X^\top(\theta - \theta_0)|^2\right] \leqslant (\widetilde{C}C_{ev})\|\theta - \theta_0\|_2^2 \leqslant (\widetilde{C}C_{ev}/c_e)\mathcal{E}(\theta), \end{aligned}$$

where the first inequality follows from the mean-value theorem with $\tilde{\theta}_{X,Y}$ being a value on the line connecting θ_0 and θ, and the last from (B.2) and (B.3). This gives Assumptions 6.2 and 14. Moreover, under our assumptions, Assumption 13.1 follows from Proposition 2.5.2 in Vershynin (2018). Finally,

$$\begin{aligned} \mathrm{E}\left[\left\{m'_1\left(X^\top\theta, Y\right) - m'_1\left(X^\top\theta_0, Y\right)\right\}^2\right] &= \mathrm{E}\left[\left(\int_{X^\top\theta_0}^{X^\top\theta} m''_{11}(t,Y)dt\right)^2\right] \\ &\leqslant 4\mathrm{E}\left[|X^\top(\theta - \theta_0)|^2\right] \leqslant (4C_{ev}/c_e)\mathcal{E}(\theta), \end{aligned}$$

where the second line follows from (B.2) and (B.3). This gives Assumption 13.2. □

Example 8 (Panel Duration Model, Continued). Since $m(t,y) = \ln(1 + \mathrm{e}^t) - \mathbf{1}(y_1 < y_2)t$ in this example,

$$f(t,x) := \mathrm{E}[m(t,Y)|X=x] = \ln(1+\mathrm{e}^t) - \mathrm{P}(Y_1 < Y_2|X=x)t$$

for all $(t,x) \in \mathbf{R} \times \mathcal{X}$. Here,

$$f_1'(t,x) = \Lambda(t) - \mathrm{P}(Y_1 < Y_2|X=x)$$

is bounded and absolutely continuous in t on $[-C_X C_d, C_X C_d] \times \mathcal{X}$ with derivative $f_{11}''(t,x) = \Lambda(t)[1-\Lambda(t)]$. Since $f_{11}''(t,x)$ is bounded away from zero on $[-C_X C_d, C_X C_d] \times \mathcal{X}$, Assumption 4 is satisfied by the argument given in the beginning of the section. Further, we have $m_1'(t,y) = \Lambda(t) - \mathbf{1}(y_1 < y_2)$ and $m_{11}''(t,y) = \Lambda(t)[1-\Lambda(t)]$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$. Therefore, $|m_1'(t,y)| \leqslant 1$ and $|m_{11}''(t,y)| \leqslant 1$ for all $(t,y) \in \mathbf{R} \times \mathcal{Y}$, and so Assumptions 6, 13, and 14 are satisfied by the same arguments as those used in the logit case of Example 1. □

C Proofs for Statements in Main Text

In this section, we provide proofs of all results stated in the main text.

C.1 Proofs for Section 3

Proof of Theorem 1. We proceed in two steps.

Step 1: Abbreviate $\widehat{\theta} := \widehat{\theta}(\lambda)$. By minimization and the triangle inequality,

$$\mathbb{E}_n[m(X_i^\top \widehat{\theta}, Y_i) - m(X_i^\top \theta_0, Y_i)] \leqslant \lambda(\|\theta_0\|_1 - \|\widehat{\theta}\|_1) \leqslant \lambda(\|\widehat{\delta}_T\|_1 - \|\widehat{\delta}_{T^c}\|_1),$$

where $\widehat{\delta} := \widehat{\theta} - \theta_0$. By convexity followed by Hölder's inequality and score domination ($\mathscr{S}$),

$$\mathbb{E}_n[m(X_i^\top \widehat{\theta}, Y_i) - m(X_i^\top \theta_0, Y_i)] \geqslant S^\top(\widehat{\theta} - \theta_0) \geqslant -\|S\|_\infty \|\widehat{\delta}\|_1 \geqslant -\frac{\lambda}{c_0}(\|\widehat{\delta}_T\|_1 + \|\widehat{\delta}_{T^c}\|_1).$$

Combining the two previous displays, we get

$$\|\widehat{\delta}_{T^c}\|_1 \leqslant \frac{c_0+1}{c_0-1}\|\widehat{\delta}_T\|_1 = \bar{c}_0 \|\widehat{\delta}_T\|_1.$$

Therefore, on the event $\mathscr{S}$, we have $\widehat{\delta} \in \mathcal{R}(\bar{c}_0)$.

Step 2: Seeking a contradiction, suppose that we are on the event $\mathscr{S} \cap \mathscr{L} \cap \mathscr{E}$ but

$\|\widehat{\delta}\|_2 > u_0$. Since Step 1 implies that $\widehat{\delta} \in \mathcal{R}(\overline{c}_0)$, it then follows by minimization that

$$0 \geqslant \inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 > u_0}} \left\{ \mathbb{E}_n[m(X_i^\top(\theta_0 + \delta), Y_i) - m(X_i^\top \theta_0, Y_i)] + \lambda \left(\|\theta_0 + \delta\|_1 - \|\theta_0\|_1 \right) \right\}.$$

Now, observe that $\delta \mapsto \{\mathbb{E}_n[m(X_i^\top(\theta_0 + \delta), Y_i) - m(X_i^\top \theta_0, Y_i)] + \lambda(\|\theta_0 + \delta\|_1 - \|\theta_0\|_1)\}$ is a (random) convex function taking the value 0 when $\delta = \mathbf{0} \in \mathbf{R}^p$. In addition, since Θ is convex (Assumption 1), it follows that $\delta \in \mathcal{R}(\overline{c}_0)$ implies $t\delta \in \mathcal{R}(\overline{c}_0)$ for any $t \in (0, 1)$. Therefore,

$$0 \geqslant \inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \left\{ \mathbb{E}_n[m(X_i^\top(\theta_0 + \delta), Y_i) - m(X_i^\top \theta_0, Y_i)] + \lambda \left(\|\theta_0 + \delta\|_1 - \|\theta_0\|_1 \right) \right\}.$$

By superadditivity of infima and definition of the empirical error function, on the event $\mathscr{L}$, the right-hand side here is bounded from below by

$$\begin{aligned} &\inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \mathrm{E}[m(X^\top(\theta_0 + \delta), Y_i) - m(X^\top \theta_0, Y_i)] \\ &\quad + \inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} (\mathbb{E}_n - \mathrm{E}) \left[m(X_i^\top(\theta_0 + \delta), Y_i) - m(X_i^\top \theta_0, Y_i) \right] + \lambda \inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \{ \|\theta_0 + \delta\|_1 - \|\theta_0\|_1 \} \\ &\quad \geqslant \inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \mathrm{E}[m(X^\top(\theta_0 + \delta), Y_i) - m(X^\top \theta_0, Y_i)] - \epsilon(u_0) - \overline{\lambda} \sup_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \big| \|\theta_0 + \delta\|_1 - \|\theta_0\|_1 \big|. \end{aligned}$$

Next, since we assume that $(1 + \overline{c}_0) u_0 \sqrt{s} \leqslant c'_M$, any $\delta \in \mathcal{R}(\overline{c}_0)$ such that $\|\delta\|_2 = u_0$ must satisfy

$$\|\delta\|_1 \leqslant (1 + \overline{c}_0)\|\delta_T\|_1 \leqslant (1 + \overline{c}_0)\sqrt{s}\|\delta_T\|_2 \leqslant (1 + \overline{c}_0)\sqrt{s}\|\delta\|_2 \leqslant c'_M. \tag{C.1}$$

Therefore, by Assumption 4,

$$\inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \mathrm{E}[m(X_i^\top(\theta_0 + \delta), Y_i) - m(X_i^\top \theta_0, Y_i)] \geqslant c_M \inf_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \|\delta\|_2^2 = c_M u_0^2.$$

Also, by the triangle inequality,

$$\sup_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \big| \|\theta_0 + \delta\|_1 - \|\theta_0\|_1 \big| \leqslant \sup_{\substack{\delta \in \mathcal{R}(\overline{c}_0), \\ \|\delta\|_2 = u_0}} \|\delta\|_1 \leqslant (1 + \overline{c}_0) u_0 \sqrt{s}.$$

In addition, $\epsilon(u_0) \leqslant \lambda_\epsilon u_0$ on the event $\mathscr{E}$. Therefore, it follows that

$$
\begin{aligned}
0 &\geqslant \inf_{\substack{\delta\in\mathcal{R}(\bar{c}_0),\\ \|\delta\|_2=u_0}} \mathrm{E}[m(X^\top(\theta_0+\delta),Y)-m(X^\top\theta_0,Y)] - \epsilon(u_0) - \overline{\lambda}\sup_{\substack{\delta\in\mathcal{R}(\bar{c}_0),\\ \|\delta\|_2=u_0}}\{\|\theta_0\|_1-\|\theta_0+\delta\|_1\} \\
&\geqslant c_M u_0^2 - \lambda_\epsilon u_0 - (1+\bar{c}_0)\overline{\lambda}u_0\sqrt{s}.
\end{aligned}
$$

However, by definition of u_0,

$$
c_M u_0^2 - \lambda_\epsilon u_0 - (1+\bar{c}_0)\overline{\lambda}u_0\sqrt{s} = c_M u_0^2 - \left(\lambda_\epsilon + (1+\bar{c}_0)\,\overline{\lambda}\sqrt{s}\right)u_0 = \frac{c_M}{2}u_0^2 > 0,
$$

yielding the desired contradiction. We therefore conclude that on the event $\mathscr{S}\cap\mathscr{L}\cap\mathscr{E}$ we have $\|\widehat{\delta}\|_2 \leqslant u_0$, which establishes the ℓ^2 bound (3.4). The ℓ^1 bound (3.5) then follows from the ℓ^2 bound and (C.1). □

C.2 Proofs for Section 3.2

Proof of Lemma 1. The claim will follow from an application of the maximal inequality in Lemma E.1. Setting up for such an application, fix $0 < u \leqslant c_L/\left[B_n(1+\bar{c}_0)\sqrt{s}\right]$ and define $\Delta(u) := \mathcal{R}(\bar{c}_0) \cap \{\delta \in \mathbf{R}^p; \|\delta\|_2 \leqslant u\}$. The zero vector in $\mathbf{R}^p$ belongs to both $\mathcal{R}(\bar{c}_0)$ and $\{\delta\in\mathbf{R}^p; \|\delta\|_2\leqslant u\}$, so $\Delta(u)$ is a nonempty subset of $\mathbf{R}^p$. By definition of $\Delta(u)$, any $\delta\in\Delta(u)$ must satisfy

$$
\|\delta\|_1 \leqslant (1+\bar{c}_0)\|\delta_T\|_1 \leqslant (1+\bar{c}_0)\sqrt{s}\|\delta_T\|_2 \leqslant (1+\bar{c}_0)\sqrt{s}\|\delta\|_2 \leqslant (1+\bar{c}_0)u\sqrt{s},
$$

thus implying

$$
\|\Delta(u)\|_1 := \sup_{\delta\in\Delta(u)}\|\delta\|_1 \leqslant (1+\bar{c}_0)u\sqrt{s}. \tag{C.2}
$$

Next, define $h : \mathbf{R}\times\mathcal{W}\to\mathbf{R}$ by $h(t,w) := m\left(x^\top\theta_0+t,y\right) - m\left(x^\top\theta_0,y\right)$ for all $t\in\mathbf{R}$ and $w=(x,y)\in\mathcal{W}$. By Assumption 6.1, the function $h:[-c_L,c_L]\times\mathcal{W}\to\mathbf{R}$ is Lipschitz in its first argument and satisfies $h(0,\cdot)\equiv 0$, thus verifying Condition 1 of Lemma E.1. Condition 2 of the same lemma follows from Hölder's inequality, Assumption 5.2, (C.2), and the upper bound on u:

$$
\max_{1\leqslant i\leqslant n}\sup_{\delta\in\Delta(u)}|X_i^\top\delta| \leqslant \max_{1\leqslant i\leqslant n}\|X_i\|_\infty\|\Delta(u)\|_1 \leqslant B_n(1+\bar{c}_0)u\sqrt{s}\leqslant c_L
$$

with probability at least $1-\zeta_n$. Assumption 6.2 implies that

$$\sup_{\delta\in\Delta(u)} \mathrm{E}[h(X^\top\delta, W)^2] \leqslant C_L^2 u^2,$$

and so Condition 3 of Lemma E.1 holds for $B_{1n} = C_L u$. Finally, given that $\mathbb{E}_n[L(W_i)^4] \leqslant C_L^4$ with probability at least $1-n^{-1}$ by Assumption 6.1 and Chebyshev's inequality, it follows from Assumption 5.3 that

$$\max_{1\leqslant j\leqslant p} \mathbb{E}_n[L\left(W_i\right)^2 X_{ij}^2] \leqslant \sqrt{\mathbb{E}_n[L(W_i)^4]} \max_{1\leqslant j\leqslant p} \sqrt{\mathbb{E}_n[X_{ij}^4]} \leqslant C_L^2 C_X^2$$

with probability at least $1-n^{-1}-\zeta_n$. Condition 4 of Lemma E.1 therefore holds with $B_{2n} = C_L C_X$ and $\gamma_n = n^{-1}+\zeta_n$. Lemma E.1 combined with the bounds on $\left\|\Delta\left(u\right)\right\|_1$ from (C.2) and $\ln\left(8pn\right) \leqslant 4\ln\left(pn\right)$ (which follows from $p \geqslant 2$) therefore shows that for all $n \in \mathbf{N}$,

$$\begin{aligned}
&\mathrm{P}\left(\epsilon\left(u\right) > \left(\{4C_L\} \vee \{C_\epsilon\sqrt{s\ln\left(pn\right)}\}\right)u/\sqrt{n}\right) \\
&\quad = \mathrm{P}\left(\sup_{\delta\in\Delta(u)} \left|\mathbb{G}_n[h(X_i^\top\delta, W_i)]\right| > \left(\{4C_L\} \vee \{C_\epsilon\sqrt{s\ln\left(pn\right)}\}\right)u\right) \leqslant 5n^{-1} + 8\zeta_n.
\end{aligned}$$

The claim now follows since we assume that $s\ln\left(pn\right) \geqslant 16C_L^2/C_\epsilon^2$. □

C.3 Proofs for Section 4

Proof of Theorem 2. We set up for an application of Theorem 1. To this end, define $\lambda_\epsilon := C_\epsilon\sqrt{s\ln(pn)/n}$ and $\overline{\lambda} := \overline{\lambda}_\alpha^{\texttt{am}}$ [see (4.2)], which are positive and finite under our assumptions. Then it follows from Lemma 1, whose application is justified by the inequalities in (4.3), that $\epsilon\left(u_0\right) \leqslant \lambda_\epsilon u_0$ with probability at least $1-5n^{-1}-8\zeta_n$, meaning that $\mathrm{P}(\mathscr{E}) \geqslant 1-5n^{-1}-8\zeta_n$. Also, observe that by the choice of penalty (4.1) and Assumptions 7 and 8, the event $\widehat{\lambda}_\alpha^{\texttt{am}} \geqslant c_0\left\|S\right\|_\infty$ occurs with probability at least $1-\alpha$, as discussed in the main text, meaning that $\mathrm{P}(\mathscr{S}) \geqslant 1-\alpha$. In addition, $\mathrm{P}(\mathscr{L}) \geqslant 1-\zeta_n$ by (4.2). Therefore, the asserted claims follow from the union bound and Theorem 1, whose application is again justified by inequalities in (4.3). □

Proof of Corollary 1. The assumption (4.4) ensures that (4.3) holds for all n large enough. Therefore, the asserted claim follows immediately from Theorem 2. □

C.4 Proofs for Section 5.1

Proof of Theorem 3. We set up for an application of Theorem 1. To this end, define $\lambda_\epsilon := C_\epsilon\sqrt{s\ln(pn)/n}$. Then it follows from Lemma 1, whose application is justified by inequalities in (5.5), that $\epsilon(u_0) \leqslant \lambda_\epsilon u_0$ with probability at least $1 - 5n^{-1} - 8\zeta_n$, meaning that $\mathrm{P}(\mathscr{E}) \geqslant 1 - 5n^{-1} - 8\zeta_n$.

Further, observe that conditional on $\{(W_i, \widehat{U}_i)\}_{i=1}^n$, the random vector $\mathbb{E}_n[e_i\widehat{U}_iX_i]$ is centered Gaussian in $\mathbf{R}^p$ with jth coordinate variance $n^{-1}\mathbb{E}_n[\widehat{U}_i^2X_{ij}^2]$. Lemma E.2 therefore shows that

$$\widehat{q}\,(1-\alpha) \leqslant (2+\sqrt{2})\sqrt{\frac{\ln{(p/\alpha)}}{n}\max_{1\leqslant j\leqslant p}\mathbb{E}_n[\widehat{U}_i^2X_{ij}^2]}.$$

In addition, with probability at least $1 - n^{-1} - \beta_n - 2\zeta_n$,

$$\begin{aligned}\max_{1\leqslant j\leqslant p}\mathbb{E}_n[\widehat{U}_i^2X_{ij}^2] &\leqslant 2\max_{1\leqslant j\leqslant p}\Big(\mathbb{E}_n[U_i^2X_{ij}^2] + \mathbb{E}_n[(\widehat{U}_i - U_i)^2X_{ij}^2]\Big)\\ &\leqslant 2\sqrt{\mathbb{E}_n[U_i^4]}\sqrt{\max_{1\leqslant j\leqslant p}\mathbb{E}_n[X_{ij}^4]} + 2B_n^2\mathbb{E}_n[(\widehat{U}_i - U_i)^2]\\ &\leqslant 2C_U^2C_X^2 + 2B_n^2\delta_n^2/\ln^2(pn) \leqslant 4C_U^2C_X^2,\end{aligned}$$

where the first inequality follows from the elementary inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$, the second inequality from Hölder's inequality and Assumption 5.2, the third from Assumptions 5.3, 9.1 and 10 and Chebyshev's inequality, and the fourth from (5.5). Hence, with the same probability,

$$\widehat{\lambda}_\alpha^{\mathsf{bm}} \leqslant \overline{\lambda}_\alpha^{\mathsf{bm}} := 4(2+\sqrt{2})c_0C_UC_X\sqrt{\frac{\ln{(p/\alpha)}}{n}},$$

meaning that $\mathrm{P}(\mathscr{L}) \geqslant 1 - n^{-1} - \beta_n - 2\zeta_n$ when taking $\overline{\lambda} = \overline{\lambda}_\alpha^{\mathsf{bm}}$.

Next, Assumptions 9.2, 9.3, and 9.4 imply that the moment conditions (E.1) for $Z_{ij} = U_iX_{ij}$ hold with b and B_n there replaced by c_U and $\widetilde{B}_n$, respectively. Further, Assumptions 5.2 and 10 imply that the estimation error condition (E.3) for $\widehat{Z}_{ij} = \widehat{U}_iX_{ij}$ hold with δ_n and β_n there replaced by $B_n\delta_n$ and $\beta_n + \zeta_n$, respectively. Since the Z_i's are centered (by Assumption 7), Theorem E.4 therefore shows that there exists a finite constant C depending only on c_U such that[23]

$$\sup_{\alpha\in(0,1)} \left|\mathrm{P}\big(\|S\|_\infty > \widehat{q}\,(1-\alpha)\big) - \alpha\right| \\ \leqslant C \max\left\{\beta_n + \zeta_n, B_n\delta_n, \left(\frac{\widetilde{B}_n^4 \ln^7(pn)}{n}\right)^{1/6}, \frac{1}{\ln^2(pn)}\right\}.$$

Taking ρ_n to be this upper bound, it thus follows by construction of the bootstrap penalty level $\widehat{\lambda}_\alpha^{\mathtt{bm}} = c_0\widehat{q}\,(1-\alpha)$ that the event $\widehat{\lambda}_\alpha^{\mathtt{bm}} \geqslant c_0 \|S\|_\infty$ occurs with probability at least $1-\alpha-\rho_n$, meaning that $\mathrm{P}(\mathscr{S}) \geqslant 1-\alpha-\rho_n$.

Therefore, the asserted claims follow from Theorem 1, whose application is again justified by the inequalities in (5.5). □

C.5 Proofs for Section 5.2

For the arguments in this section, we introduce some additional notation. For any nonempty $I \subsetneq \{1,\dots,n\}$, define the *subsample score*

$$S_I := \mathbb{E}_I\left[m_1'\left(X_i^\top\theta_0, Y_i\right)X_i\right]$$

and *subsample empirical error*

$$\epsilon_I(u) := \sup_{\substack{\delta\in\mathcal{R}(\bar{c}_0),\\ \|\delta\|_2\leqslant u}} \left|(\mathbb{E}_I - \mathrm{E})\left[m\left(X_i^\top(\theta_0+\delta), Y_i\right) - m\left(X_i^\top\theta_0, Y_i\right)\right]\right|, \quad u\in\mathbf{R}_+.$$

In proving Theorem 4, we rely on the following lemmas.

Lemma C.1. *Let Assumption 12 hold. Then for any constant $C\in\mathbf{R}_{++}$ satisfying $n^{-1}\ln(pn) \leqslant (C_\Lambda a/C)^2$ and $n\ln(pn) \geqslant (c_\Lambda/C)^2$, the candidate penalty set Λ_n and the interval*

$$\left[C\sqrt{n^{-1}\ln(pn)}, (C/a)\sqrt{n^{-1}\ln(pn)}\right]$$

have an element in common.

Lemma C.2. *Let Assumptions 5, 6, and 11 hold, and define the constant $C_\epsilon := 16\sqrt{2}(1+$*

[23] We here invoke the scaling property that $q_{tV}(\alpha) = tq_V(\alpha)$ for $t\in\mathbf{R}_{++}$ and $\alpha\in(0,1)$ and $q_V(\alpha)$ denoting the α quantile of the random variable V.

$\bar{c}_0)C_LC_X \in \mathbf{R}_{++}$. *Then provided*

$$s\ln(pn) \geqslant 16(K-1)c_D\left(\frac{C_L}{C_\epsilon}\right)^2 \quad \textit{and} \quad 0 < u \leqslant \frac{c_L}{B_n(1+\bar{c}_0)\sqrt{s}},$$

we have

$$\max_{1\leqslant k\leqslant K} \epsilon_{I_k^c}(u) \leqslant \frac{C_\epsilon u}{(K-1)c_D}\sqrt{\frac{s\ln(pn)}{n}}$$

with probability at least $1 - K\left(4n^{-1} + 8\zeta_n + [(K-1)c_Dn]^{-1}\right)$.

Lemma C.3. *Let Assumptions 5, 11 and 13 hold and define the constant* $C_S := 2C_X\sigma/[(K-1)c_D] \in \mathbf{R}_{++}$. *Then*

$$\max_{1\leqslant k\leqslant K} \|S_{I_k^c}\|_\infty \leqslant C_S\sqrt{\ln(pn)/n}$$

with probability at least $1 - K(n^{-1} + \zeta_n)$.

Lemma C.4. *Let Assumptions 1–4 hold. Fix some constants* λ_ϵ *and* $\bar{\lambda}$ *in* $\mathbf{R}_{++}$ *and* $k \in \{1,\ldots,K\}$ *and define* $u_0 := (2/c_M)(\lambda_\epsilon + (1+\bar{c}_0)\bar{\lambda}\sqrt{s})$. *In addition, suppose that* $(1+\bar{c}_0)u_0\sqrt{s} \leqslant c'_M$. *Then for any (possibly random)* $\lambda \in \Lambda_n$, *on the event* $\{\lambda \geqslant c_0\|S_{I_k^c}\|_\infty\} \cap \{\lambda \leqslant \bar{\lambda}\} \cap \{\epsilon_{I_k^c}(u_0) \leqslant \lambda_\epsilon u_0\}$, *we have*

$$\mathcal{E}\left(\widehat{\theta}_{I_k^c}(\lambda)\right) \leqslant \frac{2}{c_M}\left(\lambda_\epsilon + (1+\bar{c}_0)\bar{\lambda}\sqrt{s}\right)^2.$$

Lemma C.5. *Let Assumptions 1–6 and 11–13 hold and define the constants* $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_LC_X, C_S := 2C_X\sigma/((K-1)c_D)$ *and*

$$\widetilde{u}_0 := \frac{2}{c_M}\left(\frac{C_\epsilon}{(K-1)c_D} + \frac{(1+\bar{c}_0)c_0C_S}{a}\right)\sqrt{\frac{s\ln(pn)}{n}}, \tag{C.3}$$

all in $\mathbf{R}_{++}$. *In addition, suppose that the following inequalities hold:*

$$\left\{\begin{array}{rcl} s\ln(pn) & \geqslant & 16(K-1)c_DC_L^2/C_\epsilon^2, \\ (1+\bar{c}_0)\widetilde{u}_0\sqrt{s} & \leqslant & (c_L/B_n) \wedge c'_M, \\ n^{-1}\ln(pn) & \leqslant & (C_\Lambda a/c_0C_S)^2, \\ \textit{and} \quad n\ln(pn) & \geqslant & (c_\Lambda/c_0C_S)^2, \end{array}\right\}. \tag{C.4}$$

Then there exists a candidate penalty level $\lambda_* \in \Lambda_n$ *(possibly depending on* n*), such that*

$$\max_{1\leqslant k\leqslant K} \mathcal{E}\left(\widehat{\theta}_{I_k^c}(\lambda_*)\right) \leqslant \frac{2}{c_M}\left(\frac{C_\epsilon}{(K-1)c_D} + \frac{(1+\bar{c}_0)c_0C_S}{a}\right)^2\frac{s\ln(pn)}{n}$$

with probability at least $1 - K\left(5n^{-1} + 9\zeta_n + [(K-1)\,c_D n]^{-1}\right)$,

Lemma C.6. *Let Assumptions 1–6 and 11–14 hold and define the constants* $C_\epsilon := 16\sqrt{2}(1+\overline{c}_0)C_L C_X$, $C_S := 2C_X\sigma/((K-1)c_D)$ *and*

$$C_{\mathcal{E}} := \sqrt{\frac{2}{c_M}}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\overline{c}_0)\,c_0 C_S}{a}\right),$$

all in $\mathbf{R}_{++}$. *In addition, suppose that the inequalities (C.4) hold with* $\widetilde{u}_0$ *appearing in (C.3). Then for any* $t \in \mathbf{R}_{++}$ *such that*

$$\left\{n \geqslant \frac{1}{c_\Lambda \wedge a}, \quad \frac{C_{\mathcal{E}}^2 s \ln(pn)}{n} \leqslant 1, \quad \textit{and} \quad C_{ms}\sqrt{\frac{6t\ln n}{c_D \ln(1/a)\, n}} \leqslant \frac{1}{2}\right\}, \tag{C.5}$$

we have

$$\max_{1\leqslant k\leqslant K} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) \leqslant \frac{48C_{ms}^2}{c_D^2\ln(1/a)}\frac{t\ln n}{n} + \frac{8C_{\mathcal{E}}^2}{c_D}\frac{s\ln(pn)}{n} \tag{C.6}$$

with probability at least $1 - K\left(5n^{-1} + 9\zeta_n + [(K-1)\,c_D n]^{-1} + t^{-1}\right)$.

Proof of Theorem 4. Fix any $t \in \mathbf{R}_{++}$ satisfying (5.15) and $\lambda \in \Lambda_n$. For all $k \in \{1,\ldots,K\}$, by Assumption 13 and Markov's inequality applied conditional on $\{W_i\}_{i\in I_k^c}$, we have

$$\mathrm{P}\Big(\mathbb{E}_{I_k}\Big[\big\{m_1'\big(X_i^\top\widehat{\theta}_{I_k^c}(\lambda),Y_i\big) - m_1'\big(X_i^\top\theta_0,Y_i\big)\big\}^2\Big] > C_{ms1}^2 t\left[\sqrt{\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda)\big)} \vee \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda)\big)\right]\Big) \leqslant t^{-1}.$$

In addition, since $n \geqslant 1/(c_\Lambda \wedge a)$ by (5.14), Assumption 12 implies that $|\Lambda_n| \leqslant 3(\ln n)/\ln(1/a)$. (See the proof of Lemma C.6 for more details.) Therefore, by the union bound, for all $k \in \{1,\ldots,K\}$,

$$\begin{aligned}\mathrm{P}\Big(&\exists\lambda\in\Lambda_n \text{ s.t. } \mathbb{E}_{I_k}\Big[\big\{m_1'\big(X_i^\top\widehat{\theta}_{I_k^c}(\lambda),Y_i\big) - m_1'\big(X_i^\top\theta_0,Y_i\big)\big\}^2\Big]\\ &> \frac{3C_{ms1}^2 t\ln n}{\ln(1/a)}\left[\sqrt{\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda)\big)} \vee \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda)\big)\right]\Big) \leqslant \frac{1}{t}.\end{aligned}$$

Next, introduce events $\mathscr{C} := \cap_{k=1}^K \mathscr{C}_k$, where

$$\begin{aligned}\mathscr{C}_k := \Big\{&\mathbb{E}_{I_k}\Big[\big\{m_1'\big(X_i^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}),Y_i\big) - m_1'\big(X_i^\top\theta_0,Y_i\big)\big\}^2\Big]\\ &\leqslant \frac{3C_{ms1}^2 t\ln n}{\ln(1/a)}\left[\sqrt{\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda^{\mathsf{cv}})\big)} \vee \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda^{\mathsf{cv}})\big)\right]\Big\},\end{aligned}$$

and

$$\mathscr{R} := \Big\{ \max_{1\leqslant k\leqslant K} \mathcal{E}\big(\widehat{\theta}_{I_k^c}\big(\widehat{\lambda}^{\mathsf{cv}}\big)\big) \leqslant \frac{48C_{ms}^2}{c_D^2 \ln(1/a)} \frac{t\ln n}{n} + \frac{8C_{\mathcal{E}}^2}{c_D}\frac{s\ln(pn)}{n} \Big\}.$$

Given that the cross-validated penalty $\widehat{\lambda}^{\mathsf{cv}}$ is a random element of Λ_n, it follows that $\max_{1\leqslant k\leqslant K} \mathrm{P}\left(\mathscr{C}_k^c\right) \leqslant 1/t$, and so, by the union bound, $\mathrm{P}\left(\mathscr{C}^c\right) \leqslant K/t$. Moreover, by Lemma C.6, whose application is justified by the inequalities in (5.13), (5.14), and (5.15), we have $\mathrm{P}\left(\mathscr{R}^c\right) \leqslant K(5n^{-1} + 9\zeta_n + [(K-1)\, c_D n]^{-1} + t^{-1})$. Therefore, again by the union bound, $\mathrm{P}(\mathscr{C}\cap\mathscr{R}) \geqslant 1 - K(5n^{-1} + 9\zeta_n + [(K-1)\, c_D n]^{-1} + 2t^{-1})$. But on $\mathscr{C}\cap\mathscr{R}$, we have

$$\begin{aligned}
\mathbb{E}_n\big[(\widehat{U}_i^{\mathsf{cv}} - U_i)^2\big] &= \sum_{k=1}^K \frac{|I_k|}{n}\mathbb{E}_{I_k}\Big[\big\{m_1'\big(X_i^\top\widehat{\theta}_{I_k^c}\big(\widehat{\lambda}^{\mathsf{cv}}\big), Y_i\big) - m_1'\big(X_i^\top\theta_0, Y_i\big)\big\}^2\Big] \\
&\leqslant \frac{3C_{ms1}^2 t\ln n}{\ln(1/a)} \sum_{k=1}^K \frac{|I_k|}{n}\left[\sqrt{\mathcal{E}\big(\widehat{\theta}_{I_k^c}\,(\lambda^{\mathsf{cv}})\,\big)} \vee \mathcal{E}\big(\widehat{\theta}_{I_k^c}\,(\lambda^{\mathsf{cv}})\,\big)\right] \\
&\leqslant \frac{3C_{ms1}^2 t\ln n}{\ln(1/a)} \left(\frac{48C_{ms}^2}{c_D^2\ln(1/a)}\frac{t\ln n}{n} + \frac{8C_{\mathcal{E}}^2}{c_D}\frac{s\ln(pn)}{n}\right)^{1/2},
\end{aligned}$$

where the first inequality follows from $\mathscr{C}$ and the second from $\mathscr{R}$ and (5.15). This gives the asserted claim and completes the proof. □

Proof of Corollary 2. The assumption (5.17) ensures that there exists a sequence t_n of constants in $\mathbf{R}_{++}$ such that both

$$t_n \to \infty \quad \text{and} \quad \frac{t_n^3 B_n^4 s\ln^5(pn)(\ln n)^2}{n} \to 0. \tag{C.7}$$

Therefore, for $C_{\mathcal{E}}$ appearing in the statement of Theorem 4, setting

$$\delta_n^2 := \frac{12C_{ms1}^2 t_n\ln n}{\ln(1/a)}\left(\frac{3C_{ms}^2}{c_D^2\ln(1/a)}\frac{t_n\ln n}{n} + \frac{C_{\mathcal{E}}^2}{2c_D}\frac{s\ln(pn)}{n}\right)^{1/2} \ln^2(pn) \in \mathbf{R}_{++}$$

and

$$\beta_n := K\left(5n^{-1} + 9\zeta_n + [(K-1)c_D n]^{-1} + 2t_n^{-1}\right) \in \mathbf{R}_{++},$$

we have both $\delta_n \to 0$ and $\beta_n \to 0$. Using (5.17), (5.13) and (5.14) must hold for all n large enough. In addition, (5.17) and (C.7) imply that (5.15) with $t = t_n$ holds for all n large enough. Hence, Theorem 4 implies that Assumption 10 holds with δ_n and β_n thus chosen and all n large enough. Given that (5.17) and (C.7) also ensure $B_n\delta_n \to 0$ for δ_n thus chosen, (5.5) must hold for all n large enough. The asserted claim now follows from Theorem 3. □

D Proofs for Supporting Lemmas

In this section, we prove Lemmas C.1–C.6 used in the proof of Theorem 4.

Proof of Lemma C.1. Fix $C \in \mathbf{R}_{++}$ satisfying both $n^{-1}\ln(pn) \leqslant (C_\Lambda a/C)^2$ and $n\ln(pn) \geqslant (c_\Lambda/C)^2$, and denote $b_n := C\sqrt{n^{-1}\ln(pn)}$. We will show that there exists an integer $\ell_0 \in \{0,1,2,\dots\}$ such that

$$c_\Lambda/n \leqslant b_n \leqslant C_\Lambda a^{\ell_0} \leqslant b_n/a \leqslant C_\Lambda. \tag{D.1}$$

By Assumption 12, this will imply that $C_\Lambda a^{\ell_0}$ belongs to both the candidate penalty set Λ_n and the interval $[C\sqrt{n^{-1}\ln(pn)}, (C/a)\sqrt{n^{-1}\ln(pn)}]$.

To prove (D.1), note that the condition $n^{-1}\ln(pn) \leqslant (C_\Lambda a/C)^2$ implies that

$$0 \leqslant \frac{\ln(b_n/C_\Lambda)}{\ln a} - 1. \tag{D.2}$$

In addition, there exists an integer ℓ_0 such that

$$\frac{\ln(b_n/C_\Lambda)}{\ln a} - 1 \leqslant \ell_0 \leqslant \frac{\ln(b_n/C_\Lambda)}{\ln a}. \tag{D.3}$$

Combining (D.2) and (D.3), we obtain $b_n \leqslant C_\Lambda a^{\ell_0} \leqslant b_n/a \leqslant C_\Lambda$. Moreover, the condition $n\ln(pn) \geqslant (c_\Lambda/C)^2$ implies that $c_\Lambda/n \leqslant b_n$. Combining these inequalities gives (D.1) and completes the proof of the lemma. □

Proof of Lemma C.2. The claim will follow from an application of the maximal inequality in Lemma E.1 in a manner very similar to the proof of Lemma 1. Verification of Conditions 1–3 of Lemma E.1 follow exactly as in the proof of Lemma 1. It thus remains to verify Condition 4. To do so, fix a (hold-out) subsample $k \in \{1,\dots,K\}$. We then have

$$\max_{1\leqslant j\leqslant p} \mathbb{E}_{I_k^c}[L(W_i)^2 X_{ij}^2] \leqslant \frac{1}{(K-1)\,c_D} \max_{1\leqslant j\leqslant p} \mathbb{E}_n[L(W_i)^2 X_{ij}^2] \leqslant \frac{C_L^2 C_X^2}{(K-1)\,c_D}.$$

with probability at least $1 - n^{-1} - \zeta_n$, where the first inequality follows from Assumption 11 and the second from the same argument as in the proof of Lemma E.1. Condition 4 of Lemma E.1 thus holds with $\gamma_n = n^{-1} + \zeta_n$ and the now (K, c_D)-dependent $B_{2n} =$

$C_L C_X/\sqrt{(K-1)\,c_D}$. Lemma E.1 then shows that for any $0 < u \leqslant c_L/\left[B_n\left(1+\bar{c}_0\right)\sqrt{s}\right]$,

$$\mathrm{P}\left(\sqrt{|I_k^c|}\epsilon_{I_k^c}(u) > \left(\{4C_L\} \vee \left\{C_\epsilon\sqrt{\frac{s\ln(pn)}{(K-1)c_D}}\right\}\right) u\right)$$
$$\leqslant 4n^{-1} + 8\zeta_n + |I_k^c|^{-1} \leqslant 4n^{-1} + 8\zeta_n + \left[(K-1)\,c_D n\right]^{-1},$$

where the second inequality again uses Assumption 11. Now, given that we assume $s\ln(pn) \geqslant 16\,(K-1)\,c_D C_L^2/C_\epsilon^2$, it follows that

$$\epsilon_{I_k^c}(u) \leqslant \frac{C_\epsilon u}{(K-1)\,c_D}\sqrt{\frac{s\ln(pn)}{n}}$$

with probability at least $1-(4n^{-1}+8\zeta_n+[(K-1)c_D n]^{-1})$, where we also used Assumption 11. The asserted claim now follows from combining this inequality and the union bound. □

Proof of Lemma C.3. Fix a (hold-out) subsample $k \in \{1,\ldots,K\}$. It follows from Assumption 13.1 that for each $t \in \mathbf{R}$ and each $j \in \{1,\ldots,p\}$, the random variable $S_{I_k^c,j} = |I_k^c|^{-1}\sum_{i\in I_k^c} m_1'(X_i^\top\theta_0, Y_i)X_{ij}$ satisfies

$$\ln\mathrm{E}\left[\mathrm{e}^{tS_{I_k^c,j}}\middle|\,\{X_i\}_{i=1}^n\right] \leqslant \frac{\sigma^2 t^2}{2|I_k^c|^2}\sum_{i=1}^n X_{ij}^2 \text{ a.s.}$$

Hence, by Chernoff's inequality, for any $t > 0$,

$$\mathrm{P}\left(\left|S_{I_k^c,j}\right| > t\middle|\,\{X_i\}_{i=1}^n\right) \leqslant 2\exp\left(-\frac{|I_k^c|^2 t^2}{2\sigma^2\sum_{i=1}^n X_{ij}^2}\right) \text{ a.s.}$$

Therefore, by the union bound,

$$\mathrm{P}\left(\left\|S_{I_k^c}\right\|_\infty > t\middle|\,\{X_i\}_{i=1}^n\right) \leqslant 2p\exp\left(-\frac{|I_k^c|^2 t^2}{2\sigma^2\max_{1\leqslant j\leqslant p}\sum_{i=1}^n X_{ij}^2}\right) \text{ a.s.}$$

Hence, by Assumption 5.3, the elementary inequality $\mathrm{P}(A) \leqslant \mathrm{P}(A\cap B)+\mathrm{P}(B^c)$, and iterating expectations, we arrive at

$$\mathrm{P}\left(\left\|S_{I_k^c}\right\|_\infty > t\right) \leqslant 2p\exp\left(-\frac{|I_k^c|^2 t^2}{2n\sigma^2 C_X^2}\right) + \zeta_n.$$

Thus,

$$\mathrm{P}\left(\left\|S_{I_k^c}\right\|_\infty > C_X\sigma\sqrt{\frac{2n\ln(2pn)}{|I_k^c|^2}}\right) \leqslant n^{-1} + \zeta_n.$$

In addition,

$$\frac{2n\ln(2pn)}{|I_k^c|^2} \leqslant \frac{4n\ln(pn)}{[(K-1)c_D n]^2} \leqslant \frac{4\ln(pn)}{(K-1)^2 c_D^2 n},$$

by Assumption 11 and the fact that $p \geqslant 2$. Combining these inequalities and applying the union bound, we obtain the asserted claim. □

Proof of Lemma C.4. Denote $\widehat{\theta} := \widehat{\theta}_{I_k^c}(\lambda)$ and $\widehat{\delta} := \widehat{\theta} - \theta_0$. By Theorem 1, we then have $\|\widehat{\delta}\|_1 \leqslant (1+\overline{c}_0)u_0\sqrt{s}$ and $\|\widehat{\delta}\|_2 \leqslant u_0$. An argument parallel to Step 1 of the proof of Theorem 1 also shows that the assumed $\lambda \geqslant c_0\|S_{I_k^c}\|_\infty$ implies $\widehat{\delta} \in \mathcal{R}(\overline{c}_0)$. Therefore,

$$\begin{aligned}
\mathcal{E}(\widehat{\theta}) &= \widehat{M}_{I_k^c}(\widehat{\theta}) - \widehat{M}_{I_k^c}(\theta_0) - \left[\widehat{M}_{I_k^c}(\theta_0+\widehat{\delta}) - \widehat{M}_{I_k^c}(\theta_0) - M(\theta_0+\widehat{\delta}) + M(\theta_0)\right] \\
&\leqslant \lambda\left(\|\theta_0\|_1 - \|\widehat{\theta}\|_1\right) + \left|\widehat{M}_{I_k^c}(\theta_0+\widehat{\delta}) - \widehat{M}_{I_k^c}(\theta_0) - M(\theta_0+\widehat{\delta}) + M(\theta_0)\right| \\
&\leqslant \overline{\lambda}\|\widehat{\delta}\|_1 + \epsilon_{I_k^c}(u_0) \leqslant \left(\lambda_\epsilon + (1+\overline{c}_0)\overline{\lambda}\sqrt{s}\right)u_0 = \frac{2}{c_M}\left(\lambda_\epsilon + (1+\overline{c}_0)\overline{\lambda}\sqrt{s}\right)^2,
\end{aligned}$$

where the second line follows from the definition of $\widehat{\theta}$ and the third from $\widehat{\delta} \in \mathcal{R}(\overline{c}_0)$, the definition of $\epsilon_{I_k^c}(u_0)$, imposed conditions, and the triangle inequality. This gives the asserted claim. □

Proof of Lemma C.5. By (C.4) and Lemma C.1,

$$\left[c_0 C_S\sqrt{\frac{\ln(pn)}{n}}, \frac{c_0 C_S}{a}\sqrt{\frac{\ln(pn)}{n}}\right] \cap \Lambda_n \neq \emptyset,$$

so we can fix a penalty $\lambda_* \in \Lambda_n$ satisfying

$$c_0 C_S\sqrt{\frac{\ln(pn)}{n}} \leqslant \lambda_* \leqslant \frac{c_0 C_S}{a}\sqrt{\frac{\ln(pn)}{n}} =: \overline{\lambda}.$$

Further, denote

$$\lambda_\epsilon := \frac{C_\epsilon}{(K-1)c_D}\sqrt{\frac{s\ln(pn)}{n}}$$

and for all $k \in \{1,\ldots,K\}$, define events

$$\mathscr{L}_k := \left\{\|S_{I_k^c}\|_\infty \leqslant C_S\sqrt{n^{-1}\ln(pn)}\right\} \quad \text{and} \quad \mathscr{E}_k := \left\{\epsilon_{I_k^c}(\tilde{u}_0) \leqslant \lambda_\epsilon \tilde{u}_0\right\}.$$

Also, note that using λ_ϵ and $\overline{\lambda}$, $\widetilde{u}_0$ can be written as

$$\widetilde{u}_0 = \frac{2}{c_M}\left(\lambda_\epsilon + (1+\overline{c}_0)\overline{\lambda}\sqrt{s}\right).$$

Lemma C.4 and (C.4) therefore imply that on $\mathscr{L}_k \cap \mathscr{E}_k$,

$$\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big) \leqslant \frac{2}{c_M}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\bar{c}_0)\,c_0 C_S}{a}\right)^2 \frac{s\ln(pn)}{n}. \tag{D.4}$$

In turn, Lemma C.2 and (C.4) show that

$$\mathrm{P}\big((\cap_{k=1}^K \mathscr{E}_k)^c\big) \leqslant K\left(4n^{-1} + 8\zeta_n + [(K-1)\,c_D n]^{-1}\right).$$

Also, Lemma C.3 shows that

$$\mathrm{P}\big((\cap_{k=1}^K \mathscr{L}_k)^c\big) \leqslant K(n^{-1} + \zeta_n).$$

It thus follows from the union bound that (D.4) holds simultaneously for all $k \in \{1,\ldots,K\}$ with probability at least $1 - K\left(5n^{-1} + 9\zeta_n + [(K-1)\,c_D n]^{-1}\right)$. □

Proof of Lemma C.6. For any $\theta_1, \theta_2 \in \Theta$ and $k \in \{1,\ldots,K\}$, let

$$f_k(\theta_1,\theta_2) := (\mathbb{E}_{I_k} - \mathrm{E})[m(X_i^\top\theta_1, Y_i) - m(X_i^\top\theta_2, Y_i)].$$

Also, let $\lambda_* \in \Lambda_n$ be a value of λ satisfying the bound of Lemma C.5 and define events

$$\mathscr{R} := \left\{\max_{1\leqslant k\leqslant K} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big) \leqslant C_{\mathcal{E}}^2 \frac{s\ln(pn)}{n}\right\} \quad \text{and} \quad \mathscr{C}(t) := \bigcap_{k=1}^K \mathscr{C}_k(t),$$

where for each $k \in \{1,\ldots,K\}$,

$$\begin{aligned}\mathscr{C}_k(t) := \Bigg\{&\left|f_k(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}), \widehat{\theta}_{I_k^c}(\lambda_*))\right| \\ &\leqslant \sqrt{\frac{3t\ln n}{c_D\ln(1/a)\,n}}\sqrt{\mathrm{E}_{X,Y}\left[\left\{m\big(X^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}), Y\big) - m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda_*), Y\big)\right\}^2\right]}\Bigg\}.\end{aligned}$$

Now, fix a subsample $k \in \{1,\ldots,K\}$ and observe that for any $\lambda \in \Lambda_n$, the variance of the conditional distribution of

$$\mathbb{E}_{I_k}\left[m\big(X_i^\top\widehat{\theta}_{I_k^c}(\lambda), Y_i\big) - m\big(X_i^\top\widehat{\theta}_{I_k^c}(\lambda_*), Y_i\big)\right]$$

given $\{(X_i,Y_i)\}_{i\in I_k^c}$ is bounded from above by

$$|I_k|^{-1}\,\mathrm{E}_{X,Y}\left[\left\{m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda), Y\big) - m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda_*), Y\big)\right\}^2\right].$$

In addition, by (C.5) and Assumption 12, we have

$$|\Lambda_n| \leqslant 2\,(\ln n)\,/\,\ln(1/a) + 1 \leqslant 3(\ln n)/\ln(1/a).$$

Therefore, by the union bound and Chebyshev's inequality applied conditional on $\{(X_i, Y_i)\}_{i\in I_k^c}$, we have

$$\begin{aligned}
&\mathrm{P}\Bigg(\exists\lambda \in \Lambda_n \text{ s.t. } \left|f_k(\widehat{\theta}_{I_k^c}(\lambda), \widehat{\theta}_{I_k^c}(\lambda_*))\right| \\
&\qquad > \sqrt{\frac{3t\ln n}{c_D n\ln(1/a)}}\sqrt{\mathrm{E}_{X,Y}\left[\left\{m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda), Y\big) - m\big(X^\top\widehat{\theta}_{I_k}(\lambda_*), Y\big)\right\}^2\right]}\Bigg) \\
&\leqslant \sum_{\lambda\in\Lambda_n}\frac{c_D n\ln(1/a)}{3t|I_k|\ln n} \leqslant \frac{1}{t},
\end{aligned}$$

where the second inequality follows from Assumption 11. Hence, by the union bound and Lemma C.5,

$$\mathrm{P}\left((\mathscr{R}\cap\mathscr{C}(t))^c\right) \leqslant K\left(5n^{-1} + 9\zeta_n + \left[(K-1)\,c_D n\right]^{-1} + t^{-1}\right).$$

We will now prove that (C.6) holds on $\mathscr{R}\cap\mathscr{C}(t)$. For the rest of proof, we therefore remain on this event.

Given that

$$\widehat{\lambda}^{\mathsf{cv}} \in \operatorname*{argmin}_{\lambda\in\Lambda_n}\sum_{k=1}^{K}\sum_{i\in I_k} m\big(X_i^\top\widehat{\theta}_{I_k^c}(\lambda), Y_i\big),$$

a problem for which λ_* is feasible, we must have

$$\sum_{k=1}^{K}|I_k|\,\mathbb{E}_{I_k}\left[m\big(X_i^\top\widehat{\theta}_{I_k^c}(\lambda_*), Y_i\big)\right] \geqslant \sum_{k=1}^{K}|I_k|\,\mathbb{E}_{I_k}\left[m\big(X_i^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}), Y_i\big)\right]$$

It therefore follows from the triangle inequality and $\mathscr{C}(t)$ that

$$\begin{aligned}
&\sum_{k=1}^{K}\frac{|I_k|}{n}\Big[\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)-\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)\Big] \\
&\quad=\sum_{k=1}^{K}\frac{|I_k|}{n}\mathrm{E}_{X,Y}\Big[m\big(X^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}),Y\big)-m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda_*),Y\big)\Big] \\
&\quad\leqslant\sum_{k=1}^{K}\frac{|I_k|}{n}\big|f_k(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}),\widehat{\theta}_{I_k^c}(\lambda_*))\big| \\
&\quad\leqslant\sum_{k=1}^{K}\frac{|I_k|}{n}\sqrt{\frac{3t\ln n}{c_D\ln(1/a)\,n}}\sqrt{\mathrm{E}_{X,Y}\Big[\big\{m\big(X^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}),Y\big)-m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda_*),Y\big)\big\}^2\Big]}. \qquad \text{(D.5)}
\end{aligned}$$

In addition, on $\mathscr{R}$, we have

$$\max_{1\leqslant k\leqslant K}\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)\leqslant\frac{C_{\mathcal{E}}^2 s\ln(pn)}{n}\leqslant 1, \qquad \text{(D.6)}$$

where the second inequality follows from (C.5). Assumption 14 therefore yields

$$\mathrm{E}_{X,Y}\Big[\big\{m\big(X^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}),Y\big)-m\big(X^\top\theta_0,Y\big)\big\}^2\Big]\leqslant C_{ms}^2\big[\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)\vee\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)^2\big],$$

and

$$\mathrm{E}_{X,Y}\Big[\big\{m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda_*),Y\big)-m\big(X^\top\theta_0,Y\big)\big\}^2\Big]\leqslant C_{ms}^2\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)$$

for all $k\in\{1,\dots,K\}$. Thus, using the well-known inequality $(a+b)^2\leqslant 2a^2+2b^2$, we get

$$\begin{aligned}
&\mathrm{E}_{X,Y}\Big[\big\{m\big(X^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}),Y\big)-m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda_*),Y\big)\big\}^2\Big] \\
&\leqslant 2\mathrm{E}_{X,Y}\Big[\big\{m\big(X^\top\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}}),Y\big)-m\big(X^\top\theta_0,Y\big)\big\}^2\Big] \\
&\quad+2\mathrm{E}_{X,Y}\Big[\big\{m\big(X^\top\widehat{\theta}_{I_k^c}(\lambda_*),Y\big)-m\big(X^\top\theta_0,Y\big)\big\}^2\Big] \\
&\leqslant 2C_{ms}^2\big[\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)+\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)^2+\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)\big].
\end{aligned}$$

Substituting this bound into (D.5), we obtain

$$
\begin{aligned}
&\sqrt{\frac{c_D \ln(1/a) n}{3t \ln n}} \sum_{k=1}^{K} \frac{|I_k|}{n} \Big[\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) - \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big) \Big] \\
&\quad \leqslant \sum_{k=1}^{K} \frac{|I_k|}{n} \sqrt{2C_{ms}^2 \big[\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) + \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)^2 + \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big) \big]} \\
&\quad \leqslant \sqrt{2} C_{ms} \sum_{k=1}^{K} \frac{|I_k|}{n} \left(\sqrt{\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)} + \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) + \sqrt{\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)} \right) \\
&\quad \leqslant \sqrt{2} C_{ms} \left(\sqrt{\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)} + \sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) + \sqrt{\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)} \right),
\end{aligned}
$$

where the last line follows from Jensen's inequality. Rearranging this bound and using the last inequality in (C.5), we now obtain

$$
\begin{aligned}
\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) \leqslant 2C_{ms} \sqrt{\frac{6t \ln n}{c_D \ln(1/a) n}} &\left(\sqrt{\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)} + \sqrt{\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)} \right) \\
&+ 2 \sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big).
\end{aligned}
$$

Thus, given that the inequality $x \leqslant 2a(\sqrt{x} + \sqrt{y}) + 2y$ for $x, y \geqslant 0$ implies that $\sqrt{x} \leqslant a + [(a + \sqrt{y})^2 + y]^{1/2} \leqslant 2a + 2\sqrt{y}$, so that $x \leqslant 8a^2 + 8y$, it follows that

$$
\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) \leqslant \frac{48 C_{ms}^2}{c_D \ln(1/a)} \frac{t \ln n}{n} + 8 \sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big).
$$

Combining this bound with Assumption 11 and using (D.6), we obtain

$$
\max_{1 \leqslant k \leqslant K} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) \leqslant \frac{48 C_{ms}^2}{c_D^2 \ln(1/a)} \frac{t \ln n}{n} + \frac{8 C_{\mathcal{E}}^2}{c_D} \frac{s \ln(pn)}{n},
$$

which completes the proof of the lemma. □

E Fundamental Tools

E.1 Maximal Inequality

Let $\mathbb{G}_n\left[f\left(W_i\right)\right] := \sqrt{n}\left\{\mathbb{E}_n\left[f\left(W_i\right)\right] - \mathrm{E}\left[f\left(W\right)\right]\right\}$ abbreviate the centered and scaled empirical average.

Lemma E.1 (**Maximal Inequality Based on Contraction Principle**). *Let $\{W_i\}_{i=1}^n$ be independent copies of a random vector W, with support $\mathcal{W}$, of which X is a p-dimensional subvector, let Δ be a nonempty subset of $\mathbf{R}^p$, and let $h : \mathbf{R} \times \mathcal{W} \to \mathbf{R}$ be a measurable map satisfying $h\left(0, \cdot\right) \equiv 0$. Suppose that there exist constants $C_h, B_{1n}, B_{2n} \in \mathbf{R}_+, \zeta_n, \gamma_n \in (0,1)$ and a measurable function $L : \mathcal{W} \to \mathbf{R}_+$ such that*

1. *for all $w \in \mathcal{W}$ and all $t_1, t_2 \in \mathbf{R}$ satisfying $|t_1| \vee |t_2| \leqslant C_h$,*
$$\left|h\left(t_1, w\right) - h\left(t_2, w\right)\right| \leqslant L\left(w\right)\left|t_1 - t_2\right|;$$
2. *$\max_{1\leqslant i\leqslant n}\sup_{\delta\in\Delta}\left|X_i^\top\delta\right| \leqslant C_h$ with probability at least $1-\zeta_n$;*
3. *$\sup_{\delta\in\Delta}\mathrm{E}[h\left(X^\top\delta, W\right)^2] \leqslant B_{1n}^2$; and,*
4. *$\max_{1\leqslant j\leqslant p}\mathbb{E}_n[L\left(W_i\right)^2 X_{ij}^2] \leqslant B_{2n}^2$ with probability at least $1-\gamma_n$.*

Then, denoting $\left\|\Delta\right\|_1 := \sup_{\delta\in\Delta}\left\|\delta\right\|_1$, we have

$$\mathrm{P}\left(\sup_{\delta\in\Delta}\left|\mathbb{G}_n[h(X_i^\top\delta, W_i)]\right| > u\right) \leqslant 4\zeta_n + 4\gamma_n + n^{-1},$$

provided $u \geqslant \{4B_{1n}\} \vee \{8\sqrt{2}B_{2n}\left\|\Delta\right\|_1\sqrt{\ln\left(8pn\right)}\}$.

Proof. The claim follows from the proof of Belloni et al. (2018a, Lemma D.3), where in Step 1, we replace the set Ω by the intersection of Ω and $\{\max_{1\leqslant i\leqslant n}\sup_{\delta\in\Delta}\left|X_i^\top\delta\right| \leqslant C_h\}$. □

E.2 Gaussian Inequality

Lemma E.2 (**Gaussian Quantile Bound**). *Let $(Y_1, \ldots, Y_p)$ be centered Gaussian in $\mathbf{R}^p$ with $\sigma^2 := \max_{1\leqslant j\leqslant p}\mathrm{E}\left[Y_j^2\right]$ and $p \geqslant 2$. Let $q^Y\left(1-\alpha\right)$ denote the $(1-\alpha)$-quantile of $\max_{1\leqslant j\leqslant p}|Y_j|$ for $\alpha \in (0,1)$. Then $q^Y\left(1-\alpha\right) \leqslant (2+\sqrt{2})\sigma\sqrt{\ln\left(p/\alpha\right)}$.*

Proof. By the Borell-TIS (Tsirelson-Ibragimov-Sudakov) inequality (Adler and Taylor, 2007, Theorem 2.1.1), for any $t > 0$ we have

$$\mathrm{P}\Big(\max_{1\leqslant j\leqslant p}|Y_j| > \mathrm{E}\big[\max_{1\leqslant j\leqslant p}|Y_j|\big] + \sigma t\Big) \leqslant \mathrm{e}^{-t^2/2}.$$

This inequality translates to the quantile bound

$$q^Y(1-\alpha) \leqslant \mathrm{E}\big[\max_{1\leqslant j\leqslant p}|Y_j|\big] + \sigma\sqrt{2\ln(1/\alpha)}.$$

Talagrand (2010, Proposition A.3.1) shows that

$$\mathrm{E}\big[\max_{1\leqslant j\leqslant p}|Y_j|\big] \leqslant \sigma\sqrt{2\ln(2p)},$$

thus implying

$$q^Y(1-\alpha) \leqslant \sigma\Big(\sqrt{2\ln(2p)} + \sqrt{2\ln(1/\alpha)}\Big).$$

The claim now follows from $p \geqslant 2$. □

E.3 CLT and Bootstrap in High Dimensions

Throughout this section we let $Z_1, \ldots, Z_n$ be independent centered $\mathbf{R}^p$-valued random variables and denote their scaled average and variance by

$$S_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i \quad \text{and} \quad \Sigma := \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[Z_i Z_i^\top\right],$$

respectively. (The existence of Σ is guaranteed by our assumptions below.) For $\mathbf{R}^p$-valued random variables U and V, define the distributional measure of distance

$$\rho(U,V) := \sup_{A\in\mathcal{A}_p}\left|\mathrm{P}(U\in A) - \mathrm{P}(V\in A)\right|,$$

where $\mathcal{A}_p$ denotes the collection of hyperrectangles in $\mathbf{R}^p$. Also, for $M \in \mathbf{R}^{p\times p}$ symmetric positive definite, write N_M for a centered Gaussian random vector $\mathrm{N}(\mathbf{0}, M)$ with variance M.

Theorem E.1 (**High-Dimensional CLT**). *If for some constants $b \in \mathbf{R}_{++}$ and $B_n \in [1,\infty)$,*

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[Z_{ij}^2\right] \geqslant b, \quad \frac{1}{n}\sum_{i=1}^n \mathrm{E}\big[|Z_{ij}|^{2+k}\big] \leqslant B_n^k \quad \text{and} \quad \mathrm{E}\Big[\max_{1\leqslant j\leqslant p} Z_{ij}^4\Big] \leqslant B_n^4, \tag{E.1}$$

for all $i \in \{1,\ldots,n\}, j\in\{1,\ldots,p\}$ and $k\in\{1,2\}$, then there exists a constant $C_b \in \mathbf{R}_{++}$, depending only on b, such that

$$\rho(S_n, N_\Sigma) \leqslant C_b\left(\frac{B_n^4\ln^7(pn)}{n}\right)^{1/6}. \tag{E.2}$$

Proof. The claim follows from Chernozhukov et al. (2017, Proposition 2.1). □

Let $\widehat{Z}_i$ be an estimator of Z_i, and let $e_1, \ldots, e_n$ be i.i.d. standard Gaussians independent of both the Z_i's and the $\widehat{Z}_i$'s. Define $\widehat{S}_n^e := n^{-1/2} \sum_{i=1}^n e_i \widehat{Z}_i$ and let P_e denote the (conditional) probability measure computed with respect to the e_i's for fixed Z_i's and $\widehat{Z}_i$'s. Also, abbreviate

$$\widetilde{\rho}(\widehat{S}_n^e, N_\Sigma) := \sup_{A \in \mathcal{A}_p} \left| \mathrm{P}_e \big(\widehat{S}_n^e \in A \big) - \mathrm{P}\left(N_\Sigma \in A \right) \right|,$$

with the tilde stressing that $\widetilde{\rho}(\widehat{S}_n^e, N_\Sigma)$ is a random quantity.

Theorem E.2 (**Multiplier Bootstrap for Many Approximate Means**). *Let (E.1) hold for some constants $b \in \mathbf{R}_{++}$ and $B_n \in [1, \infty)$, and let β_n and δ_n be sequences of constants in $\mathbf{R}_{++}$ both converging to zero such that*

$$\mathrm{P}\left(\max_{1 \leqslant j \leqslant p} \frac{1}{n} \sum_{i=1}^n (\widehat{Z}_{ij} - Z_{ij})^2 > \frac{\delta_n^2}{\ln^2 (pn)} \right) \leqslant \beta_n. \tag{E.3}$$

Then there exists a constant $C_b \in \mathbf{R}_{++}$, depending only on b, such that with probability at least $1 - \beta_n - 1/\ln^2 (pn)$,

$$\widetilde{\rho}(\widehat{S}_n^e, N_\Sigma) \leqslant C_b \left(\delta_n \vee \left(\frac{B_n^4 \ln^6 (pn)}{n} \right)^{1/6} \right). \tag{E.4}$$

Proof. The claim follows from the proof of Belloni et al. (2018a, Theorem 2.2), which is here rephrased in order to highlight the dependence on the sequences β_n and δ_n. (Note that their Theorem 2.2 does not actually require their Condition A(i).) □

For any M symmetric positive definite, define $q_M^N : \mathbf{R} \to \mathbf{R} \cup \{\pm\infty\}$ as the (extended) quantile function of $\|N_M\|_\infty$,

$$q_M^N (\alpha) := \inf \left\{ t \in \mathbf{R}; \mathrm{P}\left(\|N_M\|_\infty \leqslant t \right) \geqslant \alpha \right\}, \quad \alpha \in \mathbf{R}.$$

Here we interpret $q_M^N (\alpha)$ as $+\infty (= \inf \emptyset)$ if $\alpha \geqslant 1$, and $-\infty (= \inf \mathbf{R})$ if $\alpha \leqslant 0$, such that q_M^N is monotone increasing.

Lemma E.3. *Let $M \in \mathbf{R}^{p \times p}$ be symmetric positive definite, let U be an $\mathbf{R}^p$-valued random variable, and let q denote the quantile function of $\|U\|_\infty$. Then*

$$q_M^N \left(\alpha - 2\rho \left(U, N_M \right) \right) \leqslant q \left(\alpha \right) \leqslant q_M^N \left(\alpha + \rho \left(U, N_M \right) \right) \; \textit{for all } \alpha \in (0, 1) \,.$$

Proof. Given positive definiteness of M, by the union bound, for any $t \in \mathbf{R}$ we have

$$\mathrm{P}\left(\|N_M\|_\infty = t\right) \leqslant \sum_{j=1}^{p} \mathrm{P}\left(|\mathrm{N}\left(0, M_{jj}\right)| = t\right) = 0.$$

It follows that for each $\alpha \in (0,1)$, $q_M^N(\alpha)$ is uniquely defined by

$$\mathrm{P}\left(\|N_M\|_\infty \leqslant q_M^N(\alpha)\right) = \alpha.$$

In establishing the *lower* bound we may take $\rho(U, N_M) < \alpha$. (Otherwise $q_M^N(\alpha - \rho(U, N_M)) = -\infty$ and there is nothing to show.) Then $\left[-q_M^N(\alpha - \rho(U, N_M)), q_M^N(\alpha - \rho(U, N_M))\right]^p$ is a rectangle and

$$\mathrm{P}\left(\|U\|_\infty \leqslant q_M^N(\alpha - 2\rho(U, N_M))\right) \leqslant \mathrm{P}\left(\|N_M\|_\infty \leqslant q_M^N(\alpha - 2\rho(U, N_M))\right) + \rho(U, N_M) < \alpha,$$

which implies the lower bound. In establishing the *upper* bound we may assume $\rho(U, N_M) < 1 - \alpha$. (Otherwise $q_M^N(\alpha + \rho(U, N_M)) = +\infty$ and there is nothing to show.) Then from the rectangle $\left[-q_M^N(\alpha + \rho(U, N_M)), q_M^N(\alpha + \rho(U, N_M))\right]^p$, a parallel calculation shows

$$\mathrm{P}\left(\|U\|_\infty \leqslant q_M^N(\alpha + \rho(U, N_M))\right) \geqslant \alpha,$$

which by definition of quantiles implies the upper bound. □

Now, define $q_n(\alpha)$ as the α-quantile of $\|S_n\|_\infty$

$$q_n(\alpha) := \inf\left\{t \in \mathbf{R}; \mathrm{P}(\|S_n\|_\infty \leqslant t) \geqslant \alpha\right\}, \quad \alpha \in (0,1),$$

and let $\widehat{q}_n(\alpha)$ be the α-quantile of $\|\widehat{S}_n^e\|_\infty$ computed conditional on X_i's and $\widehat{X}_i$'s,

$$\widehat{q}_n(\alpha) := \inf\left\{t \in \mathbf{R}; \mathrm{P}_e(\|\widehat{S}_n^e\|_\infty \leqslant t) \geqslant \alpha\right\}, \quad \alpha \in (0,1).$$

Theorem E.3 (**Quantile Comparison**). *If (E.1) holds for some constants $b \in \mathbf{R}_{++}$ and $B_n \in [1, \infty)$, and*

$$\rho_n := 2C_b\left(\frac{B_n^4 \ln^7(pn)}{n}\right)^{1/6}$$

denotes two times the upper bound (E.2) in Theorem E.1, then

$$q_\Sigma^N(1 - \alpha - \rho_n) \leqslant q_n(1 - \alpha) \leqslant q_\Sigma^N(1 - \alpha + \rho_n) \text{ for all } \alpha \in (0,1).$$

(ii) If, in addition, (E.3) holds for some sequences β_n and δ_n of constants in $\mathbf{R}_{++}$ both converging to zero, and

$$\rho_n' := 2C_b' \left(\delta_n \vee \left(\frac{B_n^4 \ln^6 (pn)}{n} \right)^{1/6} \right)$$

denotes two times the upper bound (E.4) in Theorem E.2, then with probability at least $1 - \beta_n - 1/\ln^2(pn)$,

$$q_\Sigma^N (1 - \alpha - \rho_n') \leqslant \widehat{q}_n (1 - \alpha) \leqslant q_\Sigma^N (1 - \alpha + \rho_n') \text{ for all } \alpha \in (0, 1) .$$

Proof. Apply Lemma E.3 with $U = S_n$ to obtain

$$q_\Sigma^N (1 - \alpha - 2\rho(S_n, N_\Sigma)) \leqslant q_n (1 - \alpha) \leqslant q_\Sigma^N (1 - \alpha + \rho(S_n, N_\Sigma)) \text{ for all } \alpha \in (0, 1) .$$

The first pair of inequalities then follows from $2\rho(S_n, N_\Sigma) \leqslant \rho_n$ (Theorem E.1). To establish the second claim, apply Lemma E.3 with $U = \widehat{S}_n^e$ and conditional on the X_i's and $\widehat{X}_i$'s to obtain

$$q_\Sigma^N (1 - \alpha - 2\widetilde{\rho}(\widehat{S}_n^e, N_\Sigma)) \leqslant \widehat{q}_n (1 - \alpha) \leqslant q_\Sigma^N (1 - \alpha + \widetilde{\rho}(\widehat{S}_n^e, N_\Sigma)) \text{ for all } \alpha \in (0, 1) .$$

The second pair of inequalities then follows on the event $2\widetilde{\rho}(\widehat{S}_n^e, N_\Sigma) \leqslant \rho_n'$, which by Theorem E.2 occurs with probability at least $1 - \beta_n - 1/\ln^2(pn)$. □

Theorem E.4 (**Multiplier Bootstrap Consistency**)**.** *Let (E.1) and (E.3) hold for some constants $b \in \mathbf{R}_{++}$ and $B_n \in [1, \infty)$ and some sequences δ_n and β_n of constants in $\mathbf{R}_{++}$ both converging to zero. Then there exists a constant $C_b \in \mathbf{R}_{++}$, depending only on b, such that*

$$\sup_{\alpha \in (0,1)} \left| \mathrm{P}\big(\|S_n\|_\infty > \widehat{q}_n (1 - \alpha) \big) - \alpha \right| \leqslant C_b \max \left\{ \beta_n, \delta_n, \left(\frac{B_n^4 \ln^7 (pn)}{n} \right)^{1/6}, \frac{1}{\ln^2 (pn)} \right\} .$$

Thus, if in addition $B_n^4 \ln^7 (pn) / n \to 0$, then

$$\sup_{\alpha \in (0,1)} \left| \mathrm{P}\big(\|S_n\|_\infty > \widehat{q}_n (1 - \alpha) \big) - \alpha \right| \to 0.$$

Proof. By Theorems E.1 and E.3,

$$\begin{aligned}\mathrm{P}\big(\|S_n\|_\infty \leqslant \widehat{q}_n\,(1-\alpha)\,\big) &\leqslant \mathrm{P}\big(\|S_n\|_\infty \leqslant q^N_\Sigma\,(1-\alpha+\rho'_n)\,\big) + \beta_n + \frac{1}{\ln^2(pn)}\\ &\leqslant \mathrm{P}\left(\|N_\Sigma\|_\infty \leqslant q^N_\Sigma\,(1-\alpha+\rho'_n)\right) + \rho_n + \beta_n + \frac{1}{\ln^2(pn)}\\ &\leqslant 1-\alpha+\rho'_n+\rho_n+\beta_n+\frac{1}{\ln^2(pn)}.\end{aligned}$$

A parallel argument shows

$$\mathrm{P}\big(\|S_n\|_\infty \leqslant \widehat{q}_n\,(1-\alpha)\,\big) \geqslant 1-\alpha-\left(\rho'_n+\rho_n+\beta_n+\frac{1}{\ln^2(pn)}\right).$$

The claim now follows from combining and rearranging the previous two displays. □