éCetverikov, Denis N.; Sørensen, Jesper R-V

Working Paper

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

cemmap working paper, No. CWP20/21

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: éCetverikov, Denis N.; Sørensen, Jesper R-V (2021) : Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of highdimensional M-estimators, cemmap working paper, No. CWP20/21, Centre for Microdata Methods and Practice (cemmap), London,
https://doi.org/10.47004/wp.cem.2021.2021

This Version is available at:
https://hdl.handle.net/10419/241956

cemmap
centre for microdata methods and practice

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 CWP20/21

UKRI Economic and Social Research Council

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

Denis Chetverikov† Jesper R.-V. Sørensen‡

April 20, 2021

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

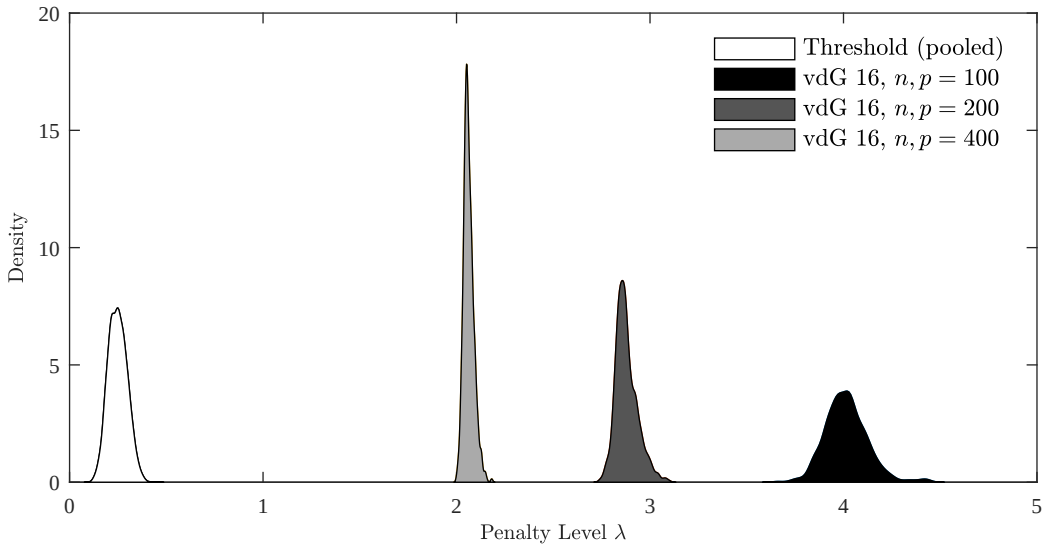
High-dimensional models have attracted substantial attention both in the econometrics and in the statistics/machine learning literature, e.g. see Belloni et al. (2018a) and Hastie et al. (2015), and ℓ^1-penalized estimators have emerged among the most useful methods for learning parameters of such models. 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: April 2021. 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 6 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, $X = (X_1, \ldots, X_p)^\top \in \mathcal{X} \subseteq \mathbf{R}^p$ 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 random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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 if $c_0 > 1$ is some constant, then there exists a constant C, depending on the distribution of the pair (X, Y) and c_0, such that under mild regularity conditions, with probability approaching one, the event

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

[1] We consider the single-index setup solely for notational convenience. A more general setup entails L indices $\{X_{(\ell)}^\top \theta_{0(\ell)}\}_{\ell=1}^L$ and a loss function $m : \mathbf{R}^L \times \mathcal{Y} \to \mathbf{R}$. Multiple indices occur naturally in, e.g., multinomial models such as the multinomial and conditional logit models. Our treatment readily extends to L greater than one provided this constant does not depend on n.

[2] We take $s \geqslant 1$ throughout. This assumption is innocuous as we may always redefine s as $\max\{1, s\}$. Also, our notion of sparsity is *exact*: the number of nonzero coefficients is small. One may entertain weaker notions such as *approximate sparsity*, which allows for many nonzero but small cofficients. Simulation evidence suggests that our methods are relevant also in settings where only approximate sparsity is satisfied (see Section 6).

implies both

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

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 nondifferentiable). These bounds suggest the following principle: choose λ as small as possible subject to the event (1.3) occurring with high probability. We therefore wish to set $\lambda = c_0 q(1-\alpha)$, where

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

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 $\lambda > c_0 q\,(1-\alpha)$.

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 $\lambda \approx c_0 q(1-\alpha)$.[3]

[3]Both analytic and bootstrap-after-cross-validation methods require specifying the constant c_0. While our

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. Moreover, both methods substantially outperform the choices of λ discussed in van de Geer (2016).

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 $\|\widehat{\theta}(\lambda) - \theta_0\|_1$. 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, the 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). In the same fashion, when λ is selected using cross-validation, neither ℓ^1 nor ℓ^2 estimation error bounds are typically known. The only exception we are aware of is the linear mean regression model. The bounds for this model have been derived in Chetverikov et al. (2016) and Miolane and Montanari (2018), but even in this special case the bounds derived are not as sharp as 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 recommendations 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 results in a very general setting, and when specialized to our setting (1.2) their results become quite similar to our result 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

theory only requires that $c_0 > 1$, simulations suggest that both ℓ^1 and ℓ^2 estimation errors of the ℓ^1-penalized M-estimator are increasing in c_0 for $c_0 > 1$. We therefore recommend setting $c_0 = 1.1$, which reflects one of the standard recommendations in the LASSO literature (see, e.g., Belloni and Chernozhukov, 2011a).

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 focus on fixed-p asymptotics, thus excluding 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 discuss the analytic method in Section 4 and the bootstrap-after-cross-validation method in Section 5. In Section 6 we illustrate our methods via a simulation study and compare them with existing methods. We give all the proofs in the Appendix, which also provides low-level conditions sufficient for some of the assumptions made in the main text.

Notation

Throughout $W_i := (X_i, Y_i), i \in \{1, \dots, n\}$, 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 $I \subsetneq \{1, \dots, n\}$ 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 $I \subseteq \{1, \dots, n\}$, I^c denotes the elements of $\{1, \dots, n\}$ not in I. Given a vector $\delta \in \mathbf{R}^p$ and a nonempty set of indices $J \subseteq \{1, \dots, p\}$, 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^1, ℓ^2 and ℓ^∞ norms by $\|\delta\|_1$, $\|\delta\|_2$, and $\|\delta\|_\infty$, respectively. We abbreviate $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$, and take $n \geqslant 3$ and $p \geqslant 2$ throughout. We introduce more notation as needed in the appendices.

[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.

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. 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 $f(t) \propto \mathrm{e}^{-|t|^\alpha}$ for some $\alpha \geqslant 1$ or $f(t) \propto t^{a-1}\mathrm{e}^{-t}$ for $t \geqslant 0$ and some $a \geqslant 1$, 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

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

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 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. In particular, it leads to average treatment effect estimators that enjoy particularly nice robustness properties. Specifically, under some conditions, these estimators remain root-n consistent and asymptotically normal even if the model (2.5) is misspecified. □

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

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

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

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

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

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

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

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

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

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 $m(t,y)$ in (2.7) is convex but not twice differentiable unless $\tau=1/2$. □

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

$$\mathrm{P}(Y_t=1|X,\alpha,Y_0,\ldots,Y_{t-1})=\Lambda(\alpha+X_t^{\top}\theta_0),\quad t=1,2,$$

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 α is a unit-specific unobserved fixed effect. Rasch (1960) shows[5] that θ_0 in this model can be identified by $\theta_0=\mathrm{argmin}_{\theta\in\mathbf{R}^p}\,\mathrm{E}[m((X_1-X_2)^{\top}\theta,Y)]$, where

$$m\left(t,y\right)=\mathbf{1}\left(y_1\neq y_2\right)\left[\ln\left(1+\mathrm{e}^{t}\right)-y_1t\right], \tag{2.8}$$

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

which is convex in t. □

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

$$Y_t = \max\left(0, \alpha + X_t^\top \theta_0 + \varepsilon_t\right), \quad t = 1, 2,$$

where $Y = (Y_1, Y_2)^\top \in \mathbf{R}_+^2$ is a pair of outcome variables, $X = (X_1^\top, X_2^\top)^\top$ is a vector of regressors, α is a unit-specific unobserved fixed effect, and ε_1 and ε_2 are unobserved error terms. Honoré (1992) shows that under certain conditions—including exchangeability of ε_1 and ε_2 conditional on X_1, X_2, α—θ_0 in this model can be identified by $\theta_0 = \operatorname{argmin}_{\theta \in \mathbf{R}^p} \mathrm{E}[m((X_1 - X_2)^\top \theta, Y)]$, with 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 have 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:

$$\ln h_t(y) = X_t^\top \theta_0 + h_0(y), \quad t = 1, 2,$$

where h_t denotes the hazard for spell t and both h_0 and h_t 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

[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).

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). □

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

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

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

$$\mathcal{R}(c) := \{\delta \in \mathbf{R}^p; \|\delta_{T^c}\|_1 \leqslant c\|\delta_T\|_1\}.$$

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

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

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

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

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 loss function $t \mapsto m(t, y)$ 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 finite constants* $c_M, c_M' > 0$ *such that*

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

Assumption 1 is a minor regularity condition. Assumption 2 is satisfied in all examples from the previous section. This assumption implies that the (random) function $\widehat{M}$ is convex, hence subdifferentiable at the interior point θ_0 (Rockafellar, 1970, Theorem 23.4). The first part of Assumption 3 is satisfied in the sure sense 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 (α, X_1, X_2) 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 as well. We provide a set of conditions that are sufficient for Assumption 4 at the end of this subsection, which are then straightforward to verify in all of the examples from the previous section.

Next, for some score selection S from the subdifferential $\partial\widehat{M}(\theta_0)$, which always exist by Assumption 2 and is almost surely singleton by Assumption 3, and for some constants $\lambda_\epsilon, \overline{\lambda} > 0$, define the events

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

where

$$u_0 := \frac{2}{c_M}\left(\lambda_\epsilon + (1 + \bar{c}_0)\,\overline{\lambda}\sqrt{s}\right). \tag{3.1}$$

Here, the event $\mathscr{S}$ ensures that the penalty is large enough to provide a sufficient level of regularization and the event $\mathscr{L}$ ensures that the penalty is not too large. For the purpose of the deterministic calculation of this section, 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, and $\mathscr{L}$ facilitates easy reference. The constant λ_ϵ appearing in the event $\mathscr{E}$ represents a deterministic modulus of continuity of the empirical error function ϵ in a neighborhood of zero of size u_0.

We now have the following result on the error bounds for the ℓ^1-penalized M-estimator:

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

$$\|\widehat{\theta}(\lambda) - \theta_0\|_2 \leqslant \frac{2}{c_M}\left(\lambda_\epsilon + (1+\overline{c}_0)\,\overline{\lambda}\sqrt{s}\right) \quad \textit{and} \tag{3.2}$$

$$\|\widehat{\theta}(\lambda) - \theta_0\|_1 \leqslant \frac{2\,(1+\overline{c}_0)}{c_M}\left(\lambda_\epsilon\sqrt{s} + (1+\overline{c}_0)\,\overline{\lambda}s\right). \tag{3.3}$$

This theorem motivates our choices of the penalty parameter λ. In particular, we will show in the next subsection that empirical error control ($\mathscr{E}$) holds with probability approaching one if $\lambda_\epsilon = C_\epsilon\sqrt{s\log(pn)/n}$ for a sufficiently large constant $C_\epsilon > 0$. Therefore, setting $\overline{\lambda} = \lambda$, such that penalty majorization ($\mathscr{L}$) holds trivially, Theorem 1 gives the same bounds as those appearing 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 (On Uniqueness). Theorem 1 actually concerns the set of optimizers to the convex minimization problem (1.2) for a fixed value of λ. While 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 (On 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 $(w,\theta) \mapsto m\left(x^\top\theta, y\right)$ is replaced by a general loss $(w,\theta) \mapsto m_\theta(w)$, which is convex in θ and P-integrable in w. □

Remark 3 (On Quadratic 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* at θ_0. The proof of Theorem 1 readily extends to more-than-quadratic margin behavior (thus leading to a flatter population criterion at θ_0 and, hence, worse identification). 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 concentrate on the (focal) quadratic case for the sake of simplicity. □

We end this subsection with a proposition providing somewhat more primitive conditions

for ensuring a quadratic margin, i.e. Assumption 4.[8]

Proposition 1 (**Quadratic Margin**). *Let Assumptions 1 and 2 hold. In addition, suppose that* $\|X\|_\infty \leqslant C_X$ *with probability one, the smallest eigenvalue of* $\mathrm{E}[XX^\top]$ *is at least* $c_e \in \mathbf{R}_{++}$*, and* $t \mapsto \mathrm{E}[m(t,Y)|X]$ *is twice differentiable in a neighborhood* $\{t \in \mathbf{R}; |t - X^\top\theta_0| < c_1\}$ *of* $X^\top\theta_0$ *with radius* $c_1 \in \mathbf{R}_{++}$ *and with second derivative at least* $c_2 \in \mathbf{R}_{++}$ *with probability one. Then Assumption 4 holds with* $c'_M \in (0, c_1/C_X)$ *and* $c_M \in (0, c_2 c_e/2]$.

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, let $\mathcal{S}_{p-1}$ denote the unit sphere in $\mathbf{R}^p$,

$$\mathcal{S}_{p-1} := \{\delta \in \mathbf{R}^p; \|\delta\|_2 = 1\}.$$

We establish empirical error control using the following three assumptions.

Assumption 5 (**Boundedness**). *There exists a finite constant* $C_X > 0$ *such that* $\|X\|_\infty \leqslant C_X$ *almost surely.*

Assumption 6 (**Locally Lipschitz Loss**). *There exist finite constants* $c_L, C_L > 0$ *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* $|t_1| \vee |t_2| \leqslant c_L$,

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

2. $\mathrm{E}[L(W)^4] \leqslant (C_L/2)^4$.

Assumption 7 (**Weighted Population Design Matrix**). *There exists a finite constant* $C_{L,e} > 0$ *such that all eigenvalues of the matrix* $\mathrm{E}[L(W)^2 XX^\top]$ *are bounded from above by* $C_{L,e}^2$.

Assumption 5 is a regularity condition that some researchers may find rather strong but we emphasize that it can be relaxed. We have chosen to impose it in order to abstract from unnecessary technicalities. Note also that the same assumption was used, for example, in van de Geer (2008). Assumption 6.1 requires that the function $t \mapsto m(x^\top\theta_0 + t, y)$ is locally

[8] We state the result as a proposition, since we do not provide the regularity conditions necessary for interchanging the order of differentiation and integration.

Lipschitz continuous for all $w = (x, y) \in \mathcal{W}$ with Lipschitz constant $L(w)$, and Assumption 6.2 requires that the fourth moment of $L(W)$ is finite. Given that every convex function $f : C \to \mathbf{R}$ is Lipschitz relative to any compact subset S of the interior of its domain C (Rockafellar, 1970, Theorem 10.4), it follows that Assumption 6.1 is actually implied by Assumption 2, and so Assumption 6 should be regarded as a mild regularity condition restricting the moments of the random variable $L(W)$. At the end of this subsection, we illustrate the calculation of this random variable and implied restrictions on the data-generating process via the examples on binary and ordered choice models from Section 2. Finally, Assumption 7 restricts eigenvalues of the weighted population design matrix. This assumption is similar to conditions often imposed in the literature on high-dimensional models, where it is assumed that the eigenvalues of the matrix $\mathrm{E}[XX^\top]$ are bounded from above.

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

Lemma 1 (**Empirical Error Bound**). *Let Assumptions 5, 6, and 7 hold, and define the finite constant* $C_\epsilon := 16\sqrt{2}(1 + \bar{c}_0)C_L C_X > 0$. *Then provided* $s \ln(pn) \geqslant 16 C_{L,e}^2 / C_\epsilon^2$ *and* $0 < u \leqslant c_L / [(1 + \bar{c}_0) C_X \sqrt{s}]$, *we have*

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

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

Remark 4 (Alternative Nonasymptotic Bounds). If the loss function m is *globally* Lipschitz in its first argument, then Assumption 6 holds with the function L being a constant. In this case, symmetrization, contraction, and concentration arguments may be used to bound the modified empirical error

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

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 there exists a constant $\widetilde{C}$ such that with probability approaching one,

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

She then demonstrates that useful 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 in moderate samples; see Section 6 for simulation results based on the choices in van de Geer (2008). Our bounds therefore seem more suitable for devising methods to choose λ. □

We conclude this section by illustrating Assumption 6 via the binary and ordered response model examples from Section 2.

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 loss function is therefore 1-Lipschitz in t, and so in this case one can take $L(w) = 1$ for all $w \in \mathcal{W}$, making Assumption 6 trivial. The probit loss function (2.3) is differentiable in t with

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

and $\varphi(t) = (2\pi)^{-1/2}\,\mathrm{e}^{-t^2/2}$ being the standard normal PDF. One can show that $m_1'(t,1) = -\varphi(t)/\Phi(t) \sim t$ as $t \to -\infty$ and $m_1'(t,0) = \varphi(t)/\left[1-\Phi(t)\right] \sim t$ as $t \to \infty$, so the probit loss function is *not* globally Lipschitz in t. However, a mean-value argument shows that for any finite $c > 0$, $|t_1| \vee |t_2| \leqslant c$, and $w = (x,y) \in \mathcal{W}$,

$$\left|m\left(x^\top\theta_0 + t_1, y\right) - m\left(x^\top\theta_0 + t_2, y\right)\right| \leqslant \sup_{|t|\leqslant c} \frac{\varphi(x^\top\theta_0 + t)}{\Phi(x^\top\theta_0 + t)\left[1-\Phi(x^\top\theta_0 + t)\right]}\left|t_1 - t_2\right|.$$

The function $\varphi/[\Phi(1-\Phi)]$ is convex and even, so the right-hand side supremum is attained at both boundary points $\pm c$. It follows that the probit loss function is locally Lipschitz with

$$L(w) = \sup_{|t|\leqslant c} \frac{\varphi(x^\top\theta_0 + t)}{\Phi(x^\top\theta_0 + t)\left[1-\Phi(x^\top\theta_0 + t)\right]} \leqslant 2(|x^\top\theta_0| + c),$$

see (1.2.2) in Adler and Taylor (2007). In this case, Assumption 6 therefore reduces to the requirement that $\mathrm{E}[|X^\top\theta_0|^4]$ is bounded from above, which is a very mild regularity condition.

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

$$m_1'(t,y) = \frac{f(t)}{F(t)\left[1-F(t)\right]}\left[F(t) - y\right]. \tag{3.4}$$

Given the binary outcome, we have $|F(t) - y| \leqslant 1$, and so, for any distribution such that $f/\left[F(1-F)\right]$ is also bounded from above, the binary-response loss function is L-Lipschitz

in t with

$$L = \sup_{t \in \mathbf{R}} \frac{f(t)}{F(t)\left[1 - F(t)\right]}.$$

For example, one can show that the t-distribution with $0 < \nu < \infty$ degrees of freedom satisfies $f_\nu(t)/[1 - F_\nu(t)] \sim t/(1 + t^2/\nu)$ as $t \to \infty$ and $-f_\nu(t)/F_\nu(t) \sim t/(1 + t^2/\nu)$ as $t \to -\infty$, demonstrating that the resulting loss function is globally Lipschitz and that Assumption 6 holds trivially in this case. □

Example 2 (Ordered Response Model, Continued). With F admitting an everywhere positive log-concave PDF $f = F'$, the loss (2.4) is differentiable with partial derivative

$$m_1'(t, y) = \sum_{j=0}^{J} \mathbf{1}(y = j) \frac{f(\alpha_{j+1} - t) - f(\alpha_j - t)}{F(\alpha_{j+1} - t) - F(\alpha_j - t)}. \tag{3.5}$$

(We here interpret $f(\pm\infty)$ and $F(-\infty)$ as zero and $F(+\infty)$ as one.) It follows from the mean-value theorem that the ordered-response loss is L-Lipschitz in t for any distribution F and cut-off points $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_J < \alpha_{J+1} = +\infty$ such that

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

is finite. For the logistic distribution $F = \Lambda$ and $f = \Lambda(1 - \Lambda)$, such that L simplifies to

$$L = \max_{0 \leqslant j \leqslant J} \sup_{t \in \mathbf{R}} |1 - \Lambda(\alpha_{j+1} - t) - \Lambda(\alpha_j - t)| = 1.$$

The ordered-logit loss is therefore globally Lipschitz, and Assumption 6 holds trivially. □

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 denotes a selection from the subdifferential $\partial \widehat{M}(\theta_0)$, which is nonempty under Assumption 2 and almost-surely singleton under Assumption 3. Therefore,

$$S = \mathbb{E}_n[(\partial/\partial\theta) m(X_i^\top \theta, Y_i)|_{\theta=\theta_0}] = \mathbb{E}_n[m_1'(X_i^\top \theta_0, Y_i) X_i] \quad \text{a.s.}$$

By analogy with the linear mean regression, we refer to $m_1'(X^\top \theta_0, Y)$ as the residual. Our analytic method makes use of the following two assumptions.

Assumption 8 (**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 9 (**Residual: Analytic Method**). *There exist functions $a, b : \mathcal{X} \to \mathbf{R}$ and a known finite constant $d > 0$ such that both $m_1'(X^\top\theta_0, Y) \subseteq [a(X), b(X)]$ and $b(X) - a(X) \leqslant d$ almost surely.*

Assumptions 8 and 9 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 $t > 0$,

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

A union bound then implies that for any $t > 0$,

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

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

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

By construction, $\widehat{\lambda}_\alpha^{\texttt{am}} \geqslant c_0\|S\|_\infty$ with conditional probability at least $1-\alpha$ for almost every realization of the X_i's, and so $\widehat{\lambda}_\alpha^{\texttt{am}} \geqslant c_0\|S\|_\infty$ 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 almost-sure bound

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

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 1–9 hold and let $\widehat{\theta} := \widehat{\theta}(\widehat{\lambda}_\alpha^{\texttt{am}})$ be a solution to the ℓ^1-penalized M-estimation problem (1.2) with*

penalty level $\lambda = \widehat{\lambda}^{\mathtt{am}}_{\alpha}$ *given in (4.1). Define the finite constants* $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X > 0$, $C^{\mathtt{am}}_\lambda := c_0 C_X d > 0$, *and*

$$u_0 := \frac{2}{c_M}\left(C_\epsilon\sqrt{\frac{s\ln(pn)}{n}} + (1+\bar{c}_0)\, C^{\mathtt{am}}_\lambda\sqrt{\frac{s\ln(p/\alpha)}{n}}\right) > 0.$$

In addition, suppose that

$$s\ln(pn) \geqslant \frac{16C^2_{L,e}}{C^2_\epsilon} \text{ and } (1+\bar{c}_0)\, u_0\sqrt{s} \leqslant \frac{c_L}{C_X} \wedge c'_M. \tag{4.3}$$

Then both

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

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

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 1–9 hold and let* $\widehat{\theta} := \widehat{\theta}(\widehat{\lambda}^{\mathtt{am}}_{\alpha})$ *be a solution to the* ℓ^1*-penalized M-estimation problem (1.2) with penalty level* $\lambda = \widehat{\lambda}^{\mathtt{am}}_{\alpha}$ *given in (4.1). In addition, suppose that*

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

Then there exists a constant C *depending only on the constants appearing in the aforementioned assumptions such that both*

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

with probability $1-\alpha-o(1)$.

We conclude this section by pointing out examples from Section 2 where Assumption 9 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 9 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 (3.4). 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 (3.4) 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

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

Hence, for any distribution such that $f/\left[F\left(1-F\right)\right]$ is also bounded from above, Assumption 9 is satisfied with

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

which only requires solving an unconstrained, univariate maximization problem. For example, as discussed earlier, $f/\left[F\left(1-F\right)\right]$ 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 9. Indeed, this loss function is differentiable in t with

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

but here $m_1'(t,0) - m_1'(t,1) = \varphi(t)/\left[1-\Phi(t)\right] + \varphi(t)/\Phi(t) \sim t$ as $t \to \infty$. The probit residual is thus not confined to any bounded interval, violating Assumption 9. We could in principle reconcile the probit loss function with Assumption 9 by assuming that we know a constant $C_d > 0$ 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 (3.5). The discrete nature of the outcome and (3.5) imply that

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

where we interpret $f(\pm\infty)$ and $F(-\infty)$ as zero and $F(+\infty)$ as one. 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 9 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

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

The second line shows that for *any* ordered logit we may use the possibly crude upper bound of $d = 2$. In the case of *trinary* ordered logit, the relevant pointwise maximum and minimum equal $1 + \Lambda(\alpha_2 - t)$ and $\Lambda(\alpha_1 - t)$, respectively. The resulting d is

$$d = \sup_{t \in \mathbf{R}} \{1 + \Lambda(\alpha_2 - t) - \Lambda(\alpha_1 - t)\} = 1 + \Lambda\left(\frac{\alpha_2 - \alpha_1}{2}\right) - \Lambda\left(\frac{\alpha_1 - \alpha_2}{2}\right),$$

the supremum being attained at $t = (\alpha_1 + \alpha_2)/2$. Further specializing to symmetric cut-offs $\alpha_2 = \alpha > 0$ and $\alpha_1 = -\alpha$ yields $d = 2\Lambda(\alpha) = 2/(1 + \mathrm{e}^{-\alpha})$, which is in $(1, 2)$ for all $\alpha > 0$ and asymptotes to two as $\alpha \to \infty$. For example, a value of $\alpha = 1$ produces $d = 2\mathrm{e}/(1 + \mathrm{e}) \approx 1.4621$. In the limiting case of $\alpha = 0$, 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

9 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 (α, X_1, X_2) is continuous (as implied by Honoré (1992, Assumption E.1)), this loss function satisfies Assumption 9 with $d = 2$. Note, however, that the trimmed LS loss function does not satisfy Assumption 9. □

Example 8 (Panel Duration Model, Continued). Since the loss function (2.10) here is of the logit form, it satisfies Assumption 9 with $d = 1$. □

5 Bootstrap-after-Cross-Validation Method

The analytic method of the previous section relies on Assumption 9. 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. Moreover, even if Assumption 9 is satisfied, the analytic penalty level $\widehat{\lambda}_\alpha^{\texttt{am}}$ 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 $U_i = m_1'(X_i^\top \theta_0, Y_i)$. In Section 5.2, we explain how to obtain suitable estimators $\widehat{U}_i$ via cross-validation.

5.1 Bootstrapping 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. To this end, let $e_1, \ldots, e_n$ be independent standard normal random variables that are independent of the data $W_1, \ldots, W_n$. 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 any 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 finite constants $b, B > 0$, one has

$$\min_{1 \leqslant j \leqslant p} \mathrm{E}[Z_{ij}^2] \geqslant b \text{ and } \mathrm{E}\Big[\max_{1 \leqslant j \leqslant p} Z_{ij}^4\Big] \leqslant B,$$

then

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

and, with probability approaching one,

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

where $\mathcal{A}_p$ denotes the collection of all (hyper)rectangles in $\mathbf{R}^p$, and the constants $C_{b,B}$ and $C'_{b,B}$ depend only on b and B. Provided $\ln^7(pn)/n \to 0$, 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

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

The $U_i X_i$'s are centered under Assumption 8, 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

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

A feasible penalty level is then given by

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

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

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 10 (**Residual: Bootstrap Method**). *There exist finite constants $c_U, C_U > 0$ such that (1) $\mathrm{E}[U^4] \leqslant C_U^4$ and (2) $\mathrm{E}[Z_j^2] \geqslant c_U$ for all $j \in \{1, \ldots, p\}$.*

Assumption 11 (**Residual Estimation**). *There exist sequences β_n and δ_n of constants in $\mathbf{R}_{++}$ 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 $1-\beta_n$.*

Use of the bootstrap method leads to following result:

Theorem 3 (**Nonasymptotic High-Probability Bounds: Bootstrap Method**). *Let Assumptions 1–8, 10, and 11 hold and let $\widehat{\theta} := \widehat{\theta}(\widehat{\lambda}_\alpha^{\texttt{bm}})$ be a solution to the ℓ^1-penalized M-estimation problem (1.2) with penalty level $\lambda = \widehat{\lambda}_\alpha^{\texttt{bm}}$ given in (5.4). Define the finite constants $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X > 0$, $C_\lambda^{\texttt{bm}} := 2(2+\sqrt{2})c_0 C_X(C_U + \delta_n/\ln(pn)) > 0$, and*

$$u_0 := \frac{2}{c_M}\left(C_\epsilon\sqrt{\frac{s\ln(pn)}{n}} + (1+\bar{c}_0)\,C_\lambda^{\texttt{bm}}\sqrt{\frac{s\ln(p/\alpha)}{n}}\right) > 0.$$

In addition, suppose that

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

Then there exists a finite constant C, depending only on c_U, such that for

$$\rho_n := C\max\left\{\beta_n, C_X\delta_n, \left(\frac{B^4\ln^7(pn)}{n}\right)^{1/6}, \frac{1}{\ln^2(pn)}\right\} \text{ and } B := (C_X C_U)^3 \vee 1,$$

we have both

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

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

5.2 Cross-Validating Residuals

Assumption 11 is 'high-level' in the sense that it does not specify how one performs residual estimation in practice. 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 12 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 13 below). Our CV procedure then goes as follows. First, estimate parameters θ_0 by

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

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

$$\widehat{\lambda}^{\texttt{cv}}\in\operatorname*{argmin}_{\lambda\in\Lambda_n}\sum_{k=1}^{K}\sum_{i\in I_k}m\big(X_i^\top\widehat{\theta}_{I_k^c}(\lambda),Y_i\big) \tag{5.8}$$

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.,

$$\widehat{U}_i^{\mathtt{cv}} := m_1'\big(X_i^\top \widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathtt{cv}}), Y_i\big), \quad i \in I_k, \quad k \in \{1, \dots, K\}\,. \tag{5.9}$$

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

To ensure good performance of the estimator resulting from initiating the bootstrap method with the CV residual estimates (5.9), we invoke the following assumptions.

Assumption 12 (**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 13 (**Candidate Penalties**). *There exists finite constants $c_\Lambda, C_\Lambda > 0$ and $a \in (0,1)$ such that*

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

Assumption 14 (**Residual**). *There exist finite constants $\sigma, C_{ms1} > 0$ such that:*

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

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

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 15 (**Mean-Square Loss Continuity**). *There exists a finite constant $C_{ms} > 0$ 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 12 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 $I_k = \{k\}$ implies $|I_k|/n \to 0$. Assumption 13 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 B.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 $\widehat{\theta}_{I_k^c}(\lambda_*), k \in \{1, \dots, K\}$. Other candidate penalty sets leading to a bound on the subsample estimator excess risk are certainly possible. Assumptions 14 and 15 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 A.

Use of CV residual estimators leads to the following result:

Theorem 4 (**High-Probability CV-Residual Error Bound**). *Let Assumptions 1–8 and 12–15 hold. Define the finite constants* $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X > 0$, $C_S := 2C_X\sigma/\sqrt{(K-1)\,c_D} > 0$,

$$C_{\mathcal{E}} := \sqrt{\frac{2}{c_M}}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\bar{c}_0)\,c_0 C_S}{a}\right) > 0 \tag{5.10}$$

and

$$\widetilde{u}_0 := \frac{2}{c_M}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\bar{c}_0)\,c_0 C_S}{a}\right)\sqrt{\frac{s\ln(pn)}{n}} > 0. \tag{5.11}$$

In addition, suppose that

$$s\ln(pn) \geqslant 16\,(K-1)\,c_D C_{L,e}^2/C_\epsilon^2, \quad (1+\bar{c}_0)\,\widetilde{u}_0\sqrt{s} \leqslant (c_L/C_X)\wedge c_M', \tag{5.12}$$

$$\ln(pn)/n \leqslant (C_\Lambda a/c_0 C_S)^2\,,\ n\ln(pn) \geqslant (c_\Lambda/c_0 C_S)^2\,,\ n \geqslant 1/c_\Lambda. \tag{5.13}$$

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

$$\frac{32C_{ms}^2}{c_D^2\ln(1/a)}\frac{t\ln n}{n} + \frac{8C_{\mathcal{E}}^2}{c_D}\frac{s\ln(pn)}{n} \leqslant 1, \tag{5.14}$$

we have

$$\mathbb{E}_n\big[(\widehat{U}_i^{\mathsf{cv}} - U_i)^2\big] \leqslant \frac{8C_{ms1}^2 t\ln n}{\ln(1/a)}\left(\frac{2C_{ms}^2}{c_D^2\ln(1/a)}\frac{t\ln n}{n} + \frac{C_{\mathcal{E}}^2}{2c_D}\frac{s\ln(pn)}{n}\right)^{1/2} \tag{5.15}$$

with probability at least $1 - K(4n^{-1} + 2[(K-1)c_D n]^{-1} + 2t^{-1})$.

This theorem provides an avenue for verification of Assumption 11. Specifically, it implies that for any sequence t_n of constants in $\mathbf{R}_{++}$ satisfying (5.14), we can take δ_n^2 and β_n in Assumption 11 to be the right-hand side of (5.15) multiplied by $\ln^2(pn)$ and $K(4n^{-1} + 2[(K-1)c_D n]^{-1} + 2t_n^{-1})$, respectively. 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 1–8, 10, and 12–15 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*

$$\frac{s^2\ln(pn/\alpha)}{n} \to 0, \quad \frac{s\ln^5(pn)(\ln n)^2}{n} \to 0, \quad \textit{and} \quad \frac{\ln^7(pn/\alpha)}{n} \to 0. \tag{5.16}$$

Then there exists a constant C 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 $1 - \alpha - o(1)$.

Corollaries 1 and 2 demonstrate that both analytic and bootstrap-after-cross-validation methods with α, for example, equal 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).

6 Simulations

In this section we investigate the finite-sample behavior of estimators based on the analytic and bootstrap-after-cross-validation methods for obtaining penalty levels proposed in Sections 4 and 5, respectively.

6.1 Simulation Design

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

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

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

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

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.[9]

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

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

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 attempt sample sizes $n \in \{100, 200, 400\}$ and limit attention to the high-dimensional regime "$p \geqslant n$" 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 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. One may therefore hope that our methods also apply (to some extent) under this *approximate sparsity* alone.

6.2 Estimators and Implementation

For estimation purposes, we mark up the score by $c_0 = 1.1$ and specify the 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 let α decrease with n, such that the error bounds in Theorems 2 and 3

[9] 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.

may be interpreted as holding with probability approaching one. We consider three feasible estimators based on the analytic and bootstrap methods. With our binary-logit design, the *analytic method* (4.1) for specifying the penalty level is justified with $d = 1$, and $\widehat{\theta}(\widehat{\lambda}^{\texttt{am}}_{\alpha})$ constitutes our *first* estimator. Our *second* estimator is based on the *bootstrap method* (5.3) initiated with residual estimates

$$\widehat{U}^{\texttt{am}}_i := m_1'\big(X_i^\top \widehat{\theta}(\widehat{\lambda}^{\texttt{am}}_{\alpha}), Y_i\big), \quad i \in \{1, \ldots, n\},$$

resulting from the analytic method. Our *third* estimator follows similarly, except that we initiate the bootstrap method with *cross-validation residual estimates* (5.9).[10] To introduce a benchmark, we also 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). All simulations are carried out in Matlab® with optimization and cross validation done using the user-contributed `glmnet` package.[11] 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.[12]

6.3 Simulation Results

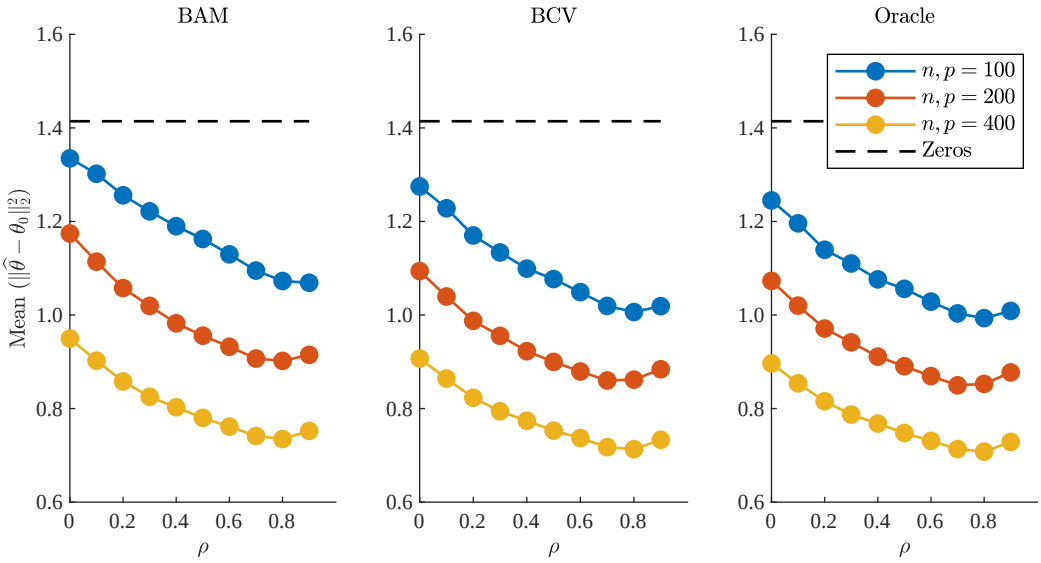
Figure 6.1 shows the mean-square ℓ^2 estimation error (over the 2,000 simulation draws) as a function of the correlation level ρ for each of the three bootstrap-based estimators and each sample size $n(= p)$, obtained with a sparse coefficient pattern. For each of these three estimators, we see that 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. This finding indicates that our bootstrap

[10] We use 10-fold cross validation, splitting the data evenly. As a result, $K = 10$ and $c_D = \frac{1}{10}$.

[11] 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 conveniently 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.

[12] Log-scale equi-distance from a "large" candidate value to essentially zero fits well with the form of Λ_n in our Assumption 13 (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 13. We believe this deviation from our theory to be only a minor issue.

Figure 6.1: Consistency of Bootstrap-Based Estimators with Exact Sparsity



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 6.2 rearranges the plots in Figure 6.1 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/number 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 6.3 and 6.4 reproduce Figures 6.1 and 6.2, respectively, with results stemming from the dense coefficient pattern (approximate sparsity). The plots in Figure 6.3 are also indicative of consistency, although convergence is slowed down by the lack of exact sparsity (compare with Figure 6.1). The ranking of estimators in Figure 6.2 is preserved in Figure 6.4. These findings suggest that our methods remain relevant under a less stringent assumption than exact sparsity.

As a final exercise, we 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

Figure 6.2: Comparing Estimators with Exact Sparsity

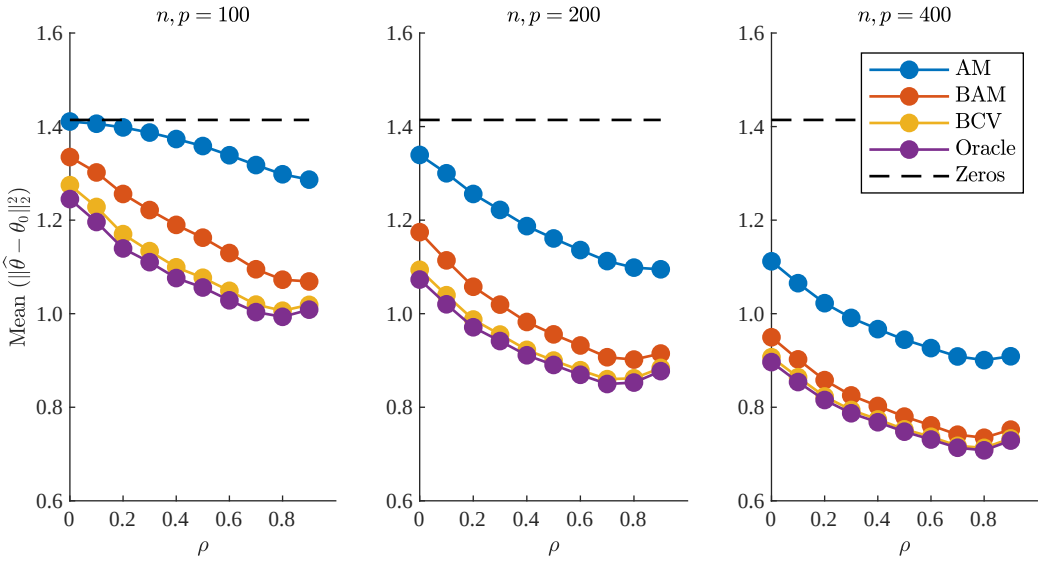


Figure 6.3: Consistency of Bootstrap-Based Estimators with Approximate Sparsity

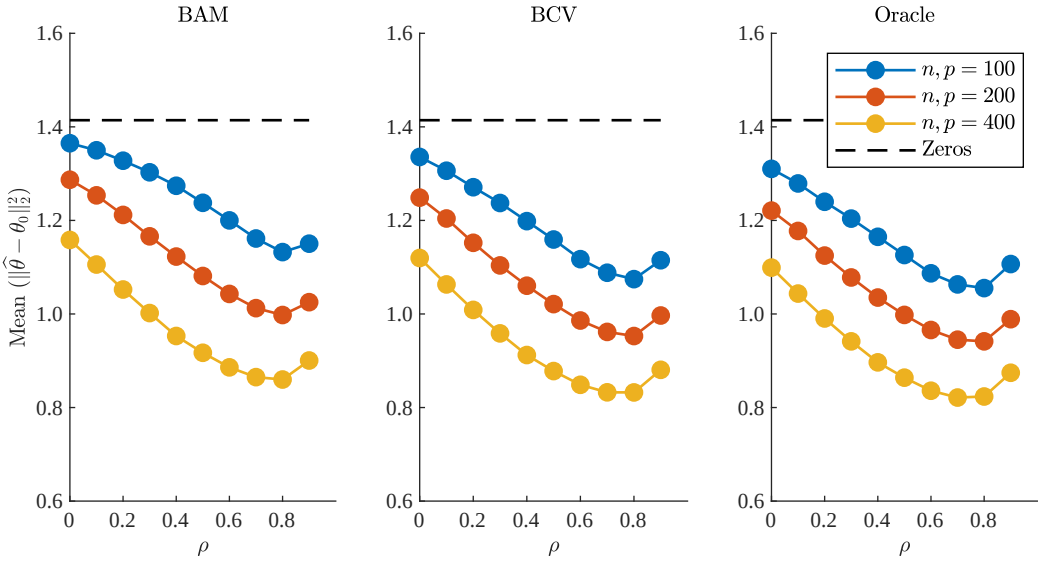
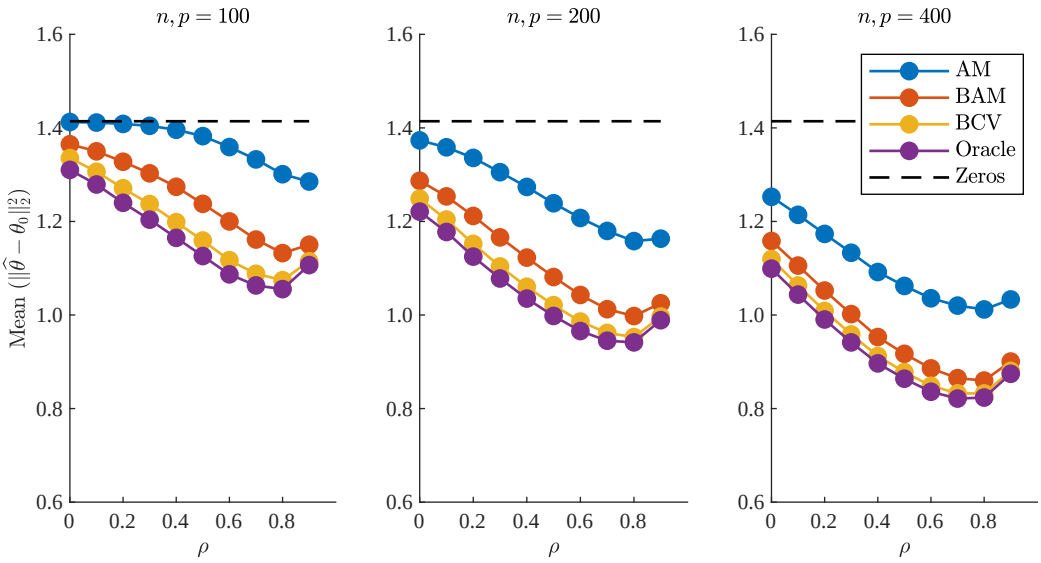


Figure 6.4: Comparing Estimators with Approximate Sparsity



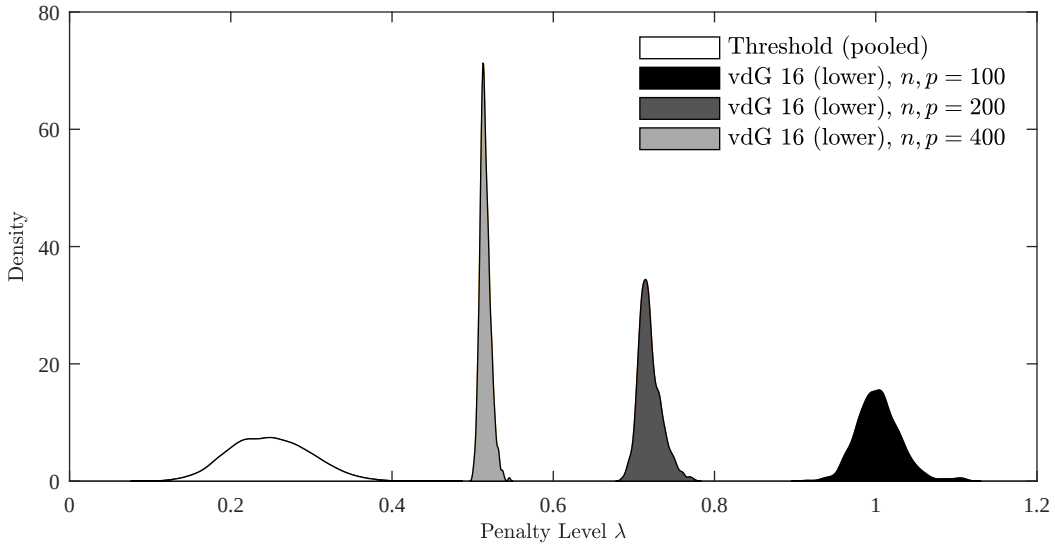
2.1) and van de Geer (2016, Theorem 12.1). Across all of our designs and simulation draws, the *smallest* Bunea (2008a) penalty is larger than the *largest* van de Geer (2008) penalty, which, in turn, is similar in size to her (2016) penalty level. We therefore restrict attention to the latter. In our notation, the van de Geer (2016, Theorem 12.1) penalty takes the form

$$\widehat{\lambda}_{\alpha}^{\texttt{vdG16}} = 8c_0\sqrt{\frac{2\ln(2p/\alpha)}{n}\max_{1\leqslant j\leqslant p}\mathbb{E}_n\left[X_{ij}^2\right]}, \tag{6.1}$$

which is nothing more than 16 times our analytic penalty level (4.1). (Recall that $d = 1$.) 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.[13] 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 are therefore all represented by the "Zeros" lines in Figures 6.1–6.4. Inspection of the proof underlying van de Geer (2016, Theorem 12.1) suggests that the factor of 8 in (6.1) may be reduced to a 2, when restricting attention to our framework. However,

[13] All density estimates in Figures 1.1 and 6.5 are created using the Matlab® package `ksdensity` with default settings.

Figure 6.5: Kernel Density Estimates of Penalty Distributions



even with this lower bound on the multiplier, the supports of these penalty distributions remain separated (Figure 6.5).

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.

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. 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.

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.

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.

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.

Ledoux, M. and M. Talagrand (1991): *Probability in Banach Spaces: Isoperimetry and Processes*, vol. 23, Springer Science & Business Media.

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.

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.

Online Appendix

We split the online appendix into four parts. In Appendix A, we provide low-level conditions that are sufficient for Assumptions 14 and 15 in each of the examples considered in the main text. In Appendix B, we provide proofs of the results stated in the main text. In Appendix C, we provide auxiliary proofs. In Appendix D, we provide a collection of technical tools used to prove the main results.

A Verification of Assumptions 14 and 15

In this section, we verify Assumptions 14 and 15 in each of the examples from Section 2. Throughout this section, we use $m''_{11}(t,y)$ to denote the second derivative of the function $t \mapsto m(t,y)$, whenever it exists. Also, we suppose that Assumptions 1, 2, 4, 5 and 8 hold and that there exists a finite constant $C_d > 0$ such that

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

Under these assumptions, there exists a finite constant $c_e > 0$ such that

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

which we use extensively throughout this section. 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 $t := c'_M / \|\theta - \theta_0\|_1$, we have by Assumption 2 that

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

and so, given that $\mathcal{E}(\theta_0) = 0$, we have

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

Hence, (A.2) holds with $c_e = c_M \wedge (c_M c'_M / 2C_d)$.

Finally, again throughout the section, we assume that there exists a finite constant $C_{ev} > 0$ such that

$$\lambda_{\max}(\mathrm{E}[XX^\top]) \leqslant C_{ev}, \tag{A.3}$$

where $\lambda_{\max}(\mathrm{E}[XX^\top])$ denotes the largest eigenvalue of the matrix $\mathrm{E}[XX^\top]$.

Example 1 (Binary Response Model, Continued). In the case of the logit loss function (2.2), we have $m_1'(t,y) = \Lambda(t) - y$, so that $|m_1'(t,y)| \leqslant 1$ for all $t \in \mathbf{R}$ and $y \in \mathcal{Y}$. Hence, Assumption 14.1 holds by Hoeffding's Lemma (Boucheron et al., 2012, Lemma 2.2) and Assumption 15 holds by noting that for all $\theta \in \Theta$, we have

$$\begin{aligned}\mathrm{E}\left[\left\{m\left(X^\top\theta, Y\right) - m\left(X^\top\theta_0, Y\right)\right\}^2\right] &\leqslant \mathrm{E}\left[m_1'(X^\top\widetilde{\theta}, Y)^2 |X^\top(\theta - \theta_0)|^2\right] \\ &\leqslant \mathrm{E}\left[|X^\top(\theta - \theta_0)|^2\right] \leqslant C_{ev}\|\theta - \theta_0\|_2^2 \leqslant (C_{ev}/c_e)\mathcal{E}(\theta),\end{aligned}$$

where the first inequality follows from the mean-value theorem with $\widetilde{\theta}$ being a value on the line connecting θ_0 and θ, the third from (A.3), and the fourth from (A.2). In addition, $m_{11}''(t,y) = \Lambda'(t) = \mathrm{e}^t/(1+\mathrm{e}^t)^2$, so that $|m_{11}''(t,y)| \leqslant 1$ for all $t \in \mathbf{R}$ and $y \in \mathcal{Y}$. Hence, Assumption 14.2 holds since for all $\theta \in \Theta$, we have

$$\begin{aligned}\mathrm{E}\left[\left\{m_1'(X^\top\theta, Y) - m_1'(X^\top\theta_0, Y)\right\}^2\right] &\leqslant \mathrm{E}\left[m_{11}''(X^\top\widetilde{\theta}, Y)^2 |X^\top(\theta - \theta_0)|^2\right] \\ &\leqslant \mathrm{E}\left[|X^\top(\theta - \theta_0)|^2\right] \leqslant C_{ev}\|\theta - \theta_0\|_2^2 \leqslant (C_{ev}/c_e)\mathcal{E}(\theta),\end{aligned}$$

where the first inequality follows from the mean-value theorem with $\widetilde{\theta}$ being a value on the line connecting θ_0 and θ, the third from (A.3), and the fourth from (A.2).

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

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

and

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

Therefore, given that the functions

$$t \mapsto \frac{\varphi(t)}{\Phi(t)} + \frac{\varphi(t)}{1-\Phi(t)}$$

and

$$t \mapsto \frac{|t|\varphi(t)}{\Phi(t)} + \frac{\varphi(t)^2}{\Phi(t)^2} + \frac{|t|\varphi(t)}{1-\Phi(t)} + \frac{\varphi(t)^2}{[1-\Phi(t)]^2}$$

are continuous, it follows that both $|m_1'(t,y)|$ and $m_{11}''(t,y)$ are bounded from above uniformly over $y \in \mathcal{Y}$ and $t \in [-C_X C_d, C_X C_d]$. Hence, Assumptions 14 and 15 hold by the same argument as that in the logit case since $|X^\top\theta_0| \leqslant C_X C_d$ almost surely by Assumption 5 and (A.1). □

Example 2 (Ordered Response Model, Continued). In the ordered logit model

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

thus implying

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

where $\Lambda'(\pm\infty)$ and $\Lambda(-\infty)$ are interpreted as zero and $\Lambda(+\infty)$ as one. Both Λ and Λ' are bounded (by one). Hence, verification of Assumptions 14 and 15 follows from an argument parallel to the one given for the binary logit. The argument for the ordered probit similarly runs in parallel to the one given in the binary case (see Example 1). □

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

$$m_1'(t,y) = -y\mathrm{e}^{-t}+1-y \quad \text{and} \quad m_{11}''(t,y) = y\mathrm{e}^{-t}.$$

Therefore, given that the function $t \mapsto \mathrm{e}^{-t}$ is continuous, it follows that both $|m_1'(t,y)|$ and $m_{11}''(t,y)$ are bounded from above uniformly over $y \in \mathcal{Y}$ and $t \in [-C_X C_d, C_X C_d]$. Hence, Assumptions 14 and 15 hold by noting that $|X^\top\theta_0| \leqslant C_X C_d$ almost surely by Assumption 5 and (A.1) like in 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, we have

$$m_1'(t,y) = (1-y)\mathrm{e}^{t}-y\mathrm{e}^{-t}+1-2y \quad \text{and} \quad m_{11}''(t,y) = (1-y)\mathrm{e}^{t}+y\mathrm{e}^{-t}.$$

Therefore, given that the functions $t \mapsto \mathrm{e}^{t}$ and $t \mapsto \mathrm{e}^{-t}$ are continuous, it follows that both $|m_1'(t,y)|$ and $|m_{11}''(t,y)|$ are bounded from above uniformly over $y \in \mathcal{Y}$ and $t \in [-C_X C_d, C_X C_d]$. Hence, Assumptions 14 and 15 hold by noting that $|X^\top\theta_0| \leqslant C_X C_d$ almost surely by Assumption 5 and (A.1) like in Example 1. □

Example 5 (Expectile Model, Continued). Since $m(t,y) = |\tau - \mathbf{1}(y-t<0)|(y-t)^2$ in this example, we have

$$m_1'(t,y) = 2|\tau - \mathbf{1}(y-t<0)|(t-y)$$

and

$$m_{11}''(t,y) = 2|\tau - \mathbf{1}(y-t<0)| \quad \text{if } y \neq t.$$

Therefore, given that the function $t \mapsto |\tau - \mathbf{1}(y-t<0)|$ is bounded (by one), Assumption

14.1 holds by noting that $|X^\top\theta_0| \leqslant C_X C_d$ almost surely by Assumption 5 and (A.1) and assuming that there exist a finite constant $C > 0$ such that

$$\mathrm{P}(|Y| \geqslant t|X) \leqslant 2\exp(-t^2/C) \quad \text{for all } t > 0 \text{ a.s.;} \tag{A.4}$$

see Proposition 2.5.2 in Vershynin (2018). If we further assume that Y is continuously distributed conditional on X, then $m_{11}''(X^\top\theta, Y)$ exists with probability one for each $\theta \in \Theta$, and the mean-value theorem yields

$$\begin{aligned}\mathrm{E}\left[\left\{m_1'\left(X^\top\theta, Y\right) - m_1'\left(X^\top\theta_0, Y\right)\right\}^2\right] &\leqslant \mathrm{E}\left[m_{11}''(X^\top\widetilde{\theta}, Y)^2|X^\top(\theta-\theta_0)|^2\right] \\ &\leqslant 4\mathrm{E}\left[|X^\top(\theta-\theta_0)|^2\right] \leqslant (4C_{ev}/c_e)\mathcal{E}(\theta),\end{aligned}$$

where $\widetilde{\theta}$ is some convex combination of θ and θ_0 and the rest is similar to the logit case in the binary response model above (Example 1). Assumption 14.2 follows.

To verify Assumption 15, note that there exists a finite constant $\widetilde{C} > 0$ such that for all $\theta \in \Theta$, we have

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

where the fourth inequality follows from (A.4) and (A.1) and Assumption 5 and the rest is similar to the logit case in the binary response model above (Example 1). □

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, the verification of Assumptions 14 and 15 is analogous to that in the logit case of the binary response model above (Example 1). □

Example 7 (Panel Censored Model, Continued). In the case $\Xi = (\cdot)^2$, 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 $m_1'(\cdot, y)$ is 2-Lipschitz for all $y \in \mathbf{R}_+^2$. Therefore, Assumptions 14 and 15 hold by noting that $|X^\top\theta_0| \leqslant C_X C_d$ almost surely by Assumption 5 and (A.1) and assuming that there

exist a finite constant $C > 0$ such that $\mathrm{P}(|Y_1| \geqslant t|X) \leqslant 2\exp(-t^2/C)$ and $\mathrm{P}(|Y_2| \geqslant t|X) \leqslant 2\exp(-t^2/C)$ for all $t > 0$ almost surely; see Proposition 2.5.2 in Vershynin (2018).

In the case $\Xi = |\cdot|$, the subdifferential $\partial_1 m(t, y)$ of the loss with respect to its first argument is contained in $[-1, 1]$ for all $t \in \mathbf{R}$ and all $y \in \mathbf{R}^2_+$. Since the subgradients of $m(\cdot, y)$ are bounded in absolute value by one, Assumptions 14.1 and 15 hold as in the logit case of the binary response model above (Example 1). To verify Assumption 14.2, note that for all $\theta \in \Theta$, $m((X_1 - X_2)^\top \theta, Y)$ is differentiable at θ except when $Y_1 - Y_2 = (X_1 - X_2)^\top \theta$ and at least one of the Y_t's is positive. Moreover, the piecewise linearity of the loss implies that the subgradients $g(\theta) \in \partial_1 m((X_1 - X_2)^\top \theta, Y)$ and $g(\theta_0) \in \partial_1 m((X_1 - X_2)^\top \theta_0, Y)$ may differ only if $(X_1 - X_2)^\top \theta$ and $(X_1 - X_2)^\top \theta_0$ are located at or on opposite sides of $Y_1 - Y_2$ and at least one of the Y_t's is nonzero. Since these subgradients can be at most 2 apart, assuming that $(\varepsilon_1, \varepsilon_2)$ is continuously distributed conditional on (X_1, X_2, α) and that both the conditional PDF of ε_2 given $(\varepsilon_1, X_1, X_2, \alpha)$ and the conditional PDF of ε_1 given $(\varepsilon_2, X_1, X_2, \alpha)$ are bounded from above by some finite constant $C > 0$, we get

$$
\begin{aligned}
&\mathrm{E}\left[\left\{m_1'\left((X_1 - X_2)^\top \theta, Y\right) - m_1'\left((X_1 - X_2)^\top \theta_0, Y\right)\right\}^2\right] \\
&\leqslant 4\mathrm{E}\left[\mathbf{1}\left(Y_1 > 0 \,\text{or}\, Y_2 > 0\right) \mathbf{1}\left(\operatorname{sgn}\left(Y_1 - Y_2 - (X_1 - X_2)^\top \theta\right) \neq \operatorname{sgn}\left(Y_1 - Y_2 - (X_1 - X_2)^\top \theta_0\right)\right)\right] \\
&\leqslant 4C\mathrm{E}\left[\left|(X_1 - X_2)^\top (\theta - \theta_0)\right|\right] \\
&\leqslant 4C\left(\mathrm{E}\left[\left|(X_1 - X_2)^\top (\theta - \theta_0)\right|^2\right]\right)^{1/2} \\
&\leqslant 4C\sqrt{C_{ev}}\|\theta - \theta_0\|_2 \leqslant (4C\sqrt{C_{ev}}/c_e)\sqrt{\mathcal{E}(\theta)},
\end{aligned}
$$

where the second inequality follows from Jensen's inequality, the third from (A.3) (with C_{ev} now denoting the upper bound on the eigenvalues of $E[(X_1 - X_2)(X_1 - X_2)^\top]$), and the fourth from (A.2). This gives Assumption 14.2, as desired. □

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, the verification of Assumptions 14 and 15 is analogous to that in the logit case of the binary response model above (Example 1). □

B Proofs for Statements in Main Text

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

B.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

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

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

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

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

Now, since $\delta \mapsto \{\mathbb{E}_n[m(X_i^\top(\theta_0 + \delta), Y_i) - m(X_i^\top \theta_0, Y_i)] + \lambda(\|\theta_0 + \delta\|_1 - \|\theta_0\|_1)\}$ is a (random) convex function taking the value 0 when $\delta = \mathbf{0} \in \mathbf{R}^p$ and $\mathcal{R}(\overline{c}_0)$ a cone (i.e., $\delta \in \mathcal{R}(\overline{c}_0)$ implies $t\delta \in \mathcal{R}(\overline{c}_0)$ for any $t \in \mathbf{R}_{++}$), the previous display implies that

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

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

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

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

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

Therefore, by Assumption 4,

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

Also, by the triangle inequality,

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

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. The ℓ^1 bound then follows from the ℓ^2 bound and (B.1). □

Proof of Proposition 1. The interiority, convexity and differentiability assumptions imply that $\nabla M(\theta_0) = \mathbf{0}$. Let $\theta \in \Theta$ satisfy $\|\theta-\theta_0\|_1 < c_1/C_X$. Then for any convex combination $\overline{\theta}$ of θ and θ_0

$$|X^\top\overline{\theta} - X^\top\theta_0| \leqslant \|X\|_\infty \|\overline{\theta}-\theta_0\|_1 \leqslant \|X\|_\infty \|\theta-\theta_0\|_1 < c_1$$

almost surely. Provided that we may interchange the order of differentiation and integration,

a second-order mean-value expansion yields

$$\begin{aligned} M(\theta) - M(\theta_0) &= \frac{1}{2}(\theta-\theta_0)^\top \nabla^2 M(\bar{\theta})(\theta-\theta_0) \\ &= \frac{1}{2}\mathrm{E}\left[\frac{\partial^2}{\partial^2 t}\mathrm{E}[m(t,Y)|X]\Big|_{t=X^\top\bar{\theta}}\{X^\top(\theta-\theta_0)\}^2\right] \\ &\geqslant \frac{1}{2}c_2\mathrm{E}\left[\{X^\top(\theta-\theta_0)\}^2\right] \\ &\geqslant \frac{1}{2}c_2c_e\|\theta-\theta_0\|_2^2, \end{aligned}$$

hence yielding the asserted claim and completing the proof of the proposition. □

B.2 Proofs for Section 3.2

Proof of Lemma 1. The claim will follow from an application of the maximal inequality in Lemma D.1. Setting up for such an application, fix $0 < u \leqslant c_L/[C_X(1+\bar{c})\sqrt{s}]$ 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

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

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, $h : [-c_L, c_L]\times\mathcal{W} \to \mathbf{R}$ is Lipschitz in its first argument and satisfies $h(0,\cdot) \equiv 0$, thus verifying Condition 1 of Lemma D.1. Condition 2 of the same lemma follows from Hölder's inequality, Assumption 5, (B.2), the upper bound on u and the calculation

$$\sup_{\delta\in\Delta(u)}|X^\top\delta| \leqslant \|X\|_\infty\|\Delta(u)\|_1 \leqslant C_X(1+\bar{c}_0)u\sqrt{s} \leqslant c_L \text{ a.s.}$$

Given that $h : [-c_L, c_L] \times \mathcal{W} \to \mathbf{R}$ is both Lipschitz in its first argument and satisfies $h(0, \cdot) \equiv 0$ as well as $\sup_{\delta \in \Delta(u)} |X^\top \delta| \leqslant c_L$ almost surely,

$$\begin{aligned}\sup_{\delta \in \Delta(u)} \mathrm{E}[h(X^\top \delta, W)^2] &\leqslant \sup_{\delta \in \Delta(u)} \mathrm{E}[L(W)^2 (X^\top \delta)^2] \\ &\leqslant u^2 \sup_{\delta \in \mathcal{S}_{p-1}} \mathrm{E}[L(W)^2 (X^\top \delta)^2] \leqslant C_{L,e}^2 u^2\end{aligned}$$

by Assumption 7. Condition 3 of Lemma D.1 therefore holds for $B_{1n} = C_{L,e} u$. Given that $\|X\|_\infty \leqslant C_X$ almost surely and $\mathbb{E}_n[L(W_i)^2] \leqslant C_L^2$ with probability at least $1 - n^{-1}$ by Assumption 6.2 and Chebyshev's inequality, it follows that

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

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

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

The claim now follows since we assume that $s \ln(pn) \geqslant 16 C_{L,e}^2 / C_\epsilon^2$. □

B.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 $\overline{\lambda} := \overline{\lambda}_\alpha^{\mathtt{am}}$ [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 $\epsilon(u_0) \leqslant \lambda_\epsilon u_0$ with probability at least $1 - 5n^{-1}$, meaning that $\mathrm{P}(\mathscr{E}) \geqslant 1 - 5n^{-1}$. Also, observe that by the choice of penalty (4.1) and Assumptions 8 and 9, the event $\widehat{\lambda}_\alpha^{\mathtt{am}} \geqslant c_0 \|S\|_\infty$ occurs with probability at least $1 - \alpha$, as discussed in the main text, meaning that $\mathrm{P}(\mathscr{S}) \geqslant 1 - \alpha$. In addition, $\mathrm{P}(\mathscr{L}) = 1$ by (4.2). Therefore, the asserted claims follow from 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. □

B.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 $1-5n^{-1}$, meaning that $\mathrm{P}(\mathscr{E}) \geqslant 1-5n^{-1}$.

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

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

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

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

where the first inequality follows from Assumption 5, the second from the elementary inequality $(a+b)^2 \leqslant 2a^2+2b^2$, and the third from Assumptions 10 and 11 and Chebyshev's inequality. Hence, with the same probability,

$$\widehat{\lambda}_\alpha^{\mathtt{bm}} \leqslant \overline{\lambda} := 2(2+\sqrt{2})c_0C_X\big(C_U+\delta_n/\ln(pn)\big)\sqrt{\frac{\ln(p/\alpha)}{n}},$$

meaning that $\mathrm{P}(\mathscr{L}) \geqslant 1-\beta_n-n^{-1}$.

Next, Assumptions 5 and 10 imply that the moment conditions (D.1) for $Z_{ij}=U_iX_{ij}$ hold with b replaced by c_U and B_n replaced by $B=(C_XC_U)^3\vee 1$. Further, the same assumptions also imply the estimation error condition (D.2) for $\widehat{Z}_{ij}=\widehat{U}_iX_{ij}$ hold with δ_n replaced by $C_X\delta_n$. Since the Z_i's are centered (by Assumption 8), Theorem D.4 therefore shows that there exists a finite constant C depending only on c_U such that[14]

$$\begin{aligned}&\sup_{\alpha\in(0,1)}\big|\mathrm{P}\big(\|S\|_\infty > \widehat{q}\,(1-\alpha)\big)-\alpha\big| \\ &\qquad\leqslant C\max\Big\{\beta_n, C_X\delta_n, \Big(\frac{B^4\ln^7(pn)}{n}\Big)^{1/6}, \frac{1}{\ln^2(pn)}\Big\} = \rho_n.\end{aligned}$$

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

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

by the inequalities in (5.5). □

B.5 Proofs for Section 5.2

For the arguments in this section, we introduce some additional notation. For any nonempty $I \subsetneq \{1, \ldots, n\}$, 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 B.1. *Let Assumption 13 hold. Then for any finite constant $C > 0$ 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 $[C\sqrt{n^{-1}\ln(pn)}, (C/a)\sqrt{n^{-1}\ln(pn)}]$ have an element in common.*

Lemma B.2. *Let Assumptions 5, 6, 7, and 12 hold, and define the finite constant $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0) C_L C_X > 0$. Then provided*

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

we have

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

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

Lemma B.3. *Let Assumptions 5, 12 and 14 hold and define the finite constant $C_S := 2C_X\sigma / \sqrt{(K-1)c_D}$. Then*

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

with probability at least $1 - K[(K-1)c_D n]^{-1}$.

Lemma B.4. *Let Assumptions 1, 2, 3, and 4 hold. Fix some finite constants $\lambda_\epsilon, \overline{\lambda} > 0$ and $k \in \{1, \ldots, K\}$ and define $u_0 := (2/c_M)(\lambda_\epsilon + (1+\bar{c}_0)\overline{\lambda}\sqrt{s})$. 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* $\{\lambda \geqslant c_0\|S_{I_k^c}\|_\infty\} \cap \{\lambda \leqslant \overline{\lambda}\} \cap \{\epsilon_{I_k^c}(u_0) \leqslant \lambda_\epsilon u_0\}$, *we have*

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

Lemma B.5. *Let Assumptions 1–7 and 12–14 hold and define the finite constants* $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X > 0$. $C_S := 2C_X\sigma/\sqrt{(K-1)\,c_D} > 0$, *and*

$$\widetilde{u}_0 := \frac{2}{c_M}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\bar{c}_0)\,c_0 C_S}{a}\right)\sqrt{\frac{s\ln(pn)}{n}} > 0. \tag{B.3}$$

In addition, suppose that the following inequalities hold:

$$\left\{\begin{array}{rcl} s\ln(pn) & \geqslant & 16\,(K-1)\,c_D C_{L,e}^2/C_\epsilon^2, \\ (1+\bar{c}_0)\,\widetilde{u}_0\sqrt{s} & \leqslant & (c_L/C_X)\wedge c'_M, \\ n^{-1}\ln(pn) & \leqslant & \left(C_\Lambda a/c_0 C_S\right)^2, \\ \text{and} \quad n\ln(pn) & \geqslant & \left(c_\Lambda/c_0 C_S\right)^2, \end{array}\right\}. \tag{B.4}$$

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

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

with probability at least $1 - K\left(4n^{-1} + 2\left[(K-1)\,c_D n\right]^{-1}\right)$,

Lemma B.6. *Let Assumptions 1–7 and 12–15 hold and define the finite constants* $C_\epsilon := 16\sqrt{2}(1+\bar{c}_0)C_L C_X > 0$, $C_S := 2C_X\sigma/\sqrt{(K-1)\,c_D} > 0$ *and*

$$C_{\mathcal{E}} := \sqrt{\frac{2}{c_M}}\left(\frac{C_\epsilon}{(K-1)\,c_D} + \frac{(1+\bar{c}_0)\,c_0 C_S}{a}\right) > 0.$$

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

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

we have

$$\max_{1\leqslant k\leqslant K}\mathcal{E}\left(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\right) \leqslant \frac{32C_{ms}^2}{c_D^2\ln(1/a)}\frac{t\ln n}{n} + \frac{8C_{\mathcal{E}}^2}{c_D}\frac{s\ln(pn)}{n} \tag{B.6}$$

with probability at least $1 - K\left(4n^{-1} + 2\left[(K-1)\,c_D n\right]^{-1} + t^{-1}\right)$.

PROOF OF THEOREM 4. Fix any $t \in \mathbf{R}_{++}$ satisfying (5.14) and $\lambda \in \Lambda_n$. For all $k \in \{1, \dots, K\}$, by Assumption 14 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 $n \geqslant 1/c_\Lambda$ by (5.13), Assumption 13 implies that $|\Lambda_n| \leqslant 2\,(\ln n)\,/\ln(1/a)$. Therefore, by the union bound, for all $k \in \{1, \dots, K\}$,

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

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

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

and

$$\mathscr{R} := \Big\{ \max_{1 \leqslant k \leqslant K} \mathcal{E}\big(\widehat{\theta}_{I_k^c}\big(\widehat{\lambda}^{\mathsf{cv}}\big)\big) \leqslant \frac{32 C_{ms}^2}{c_D^2 \ln(1/a)} \frac{t \ln n}{n} + \frac{8 C_{\mathcal{E}}^2}{c_D} \frac{s \ln(pn)}{n} \Big\}.$$

Given that the cross-validated penalty $\widehat{\lambda}^{\mathsf{cv}}$ is a random element of Λ_n, it follows that $\max_{1 \leqslant k \leqslant K} \mathrm{P}(\mathscr{C}_k^c) \leqslant 1/t$, and so, by the union bound, $\mathrm{P}(\mathscr{C}^c) \leqslant K/t$. Moreover, by Lemma B.6, whose application is justified by the inequalities in (5.12), (5.13), and (5.14), we have $\mathrm{P}(\mathscr{R}^c) \leqslant K(4n^{-1} + 2\,[(K-1)\,c_D n]^{-1} + t^{-1})$. Therefore, again by the union bound, $\mathrm{P}(\mathscr{C} \cap \mathscr{R}) \geqslant 1 - K(4n^{-1} + 2\,[(K-1)\,c_D n]^{-1} + 2t^{-1})$. But on $\mathscr{C} \cap \mathscr{R}$, we have

$$\begin{aligned} \mathbb{E}_n\big[(\widehat{U}_i^{\mathsf{cv}} - U_i)^2\big] &= \sum_{k=1}^K \frac{|I_k|}{n} \mathbb{E}_{I_k}\Big[\big\{ m_1'\big(X_i^\top \widehat{\theta}_{I_k^c}\big(\widehat{\lambda}^{\mathsf{cv}}\big), Y_i\big) - m_1'\big(X_i^\top \theta_0, Y_i\big) \big\}^2 \Big] \\ &\leqslant \frac{2C_{ms1}^2 t \ln n}{\ln(1/a)} \sum_{k=1}^K \frac{|I_k|}{n} \left[\sqrt{\mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda)\big)} \vee \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda)\big) \right] \\ &\leqslant \frac{2C_{ms1}^2 t \ln n}{\ln(1/a)} \Big(\frac{32 C_{ms}^2}{c_D^2 \ln(1/a)} \frac{t \ln n}{n} + \frac{8 C_{\mathcal{E}}^2}{c_D} \frac{s \ln(pn)}{n} \Big)^{1/2}, \end{aligned}$$

where the first inequality follows from $\mathscr{C}$ and the second from $\mathscr{R}$ and (5.14). This gives the asserted claim and completes the proof. □

Proof of Corollary 2. The assumption (5.16) ensures that there exists a sequence t_n of constants $\mathbf{R}_+$ such that both

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

Therefore, setting $\delta_n > 0$ such that

$$\delta_n^2 := \frac{8C_{ms1}^2 t_n \ln n}{\ln(1/a)} \left(\frac{2C_{ms}^2}{c_D^2 \ln(1/a)} \frac{t_n \ln n}{n} + \frac{C_{\mathcal{E}}^2}{2c_D} \frac{s \ln(pn)}{n} \right)^{1/2} \ln^2(pn)$$

and

$$\beta_n := K \left(4n^{-1} + 2[(K-1)c_D n]^{-1} + 2t_n^{-1} \right),$$

we have both $\delta_n \to 0$ and $\beta_n \to 0$. In addition, (B.7) implies that (5.14) with $t = t_n$ holds for all n large enough. Also, (5.16) ensures that (5.12) and (5.13) hold for all n large enough as well. Hence, Theorem 4 implies that Assumption 11 with δ_n and β_n thus chosen holds for all n large enough. Thus, given that (5.16) also ensures that (5.5) holds for all n large enough, the asserted claim follows from Theorem 3. □

C Proofs for Supporting Lemmas

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

Proof of Lemma B.1. Denote $b_n := C\sqrt{n^{-1} \ln(pn)}$. We will show that there exists an integer ℓ_0 such that

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

By Assumption 13, 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 (C.1), note that the condition $n^{-1}\ln(pn) \leqslant (C_\Lambda a / C)^2$ implies that

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

In addition, there exists an integer ℓ_0 such that

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

Combining (C.2) and (C.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 $c_\Lambda / n \leqslant b$. Combining these inequalities gives (C.1) and

completes the proof of the lemma. □

Proof of Lemma B.2. The claim will follow from an application of the maximal inequality in Lemma D.1 in a manner very similar to the proof of Lemma 1. Verification of Conditions 1–3 of Lemma D.1 are unrelated to sampling. It thus remains to verify Condition 4. To do so, fix a (hold-out) subsample $k \in \{1, \ldots, K\}$. Given that $\|X\|_\infty \leqslant C_X$ almost surely (Assumption 5), $|I_k| \geqslant c_D n$ (Assumption 12) and $\mathbb{E}_n[L(W_i)^2] \leqslant C_L^2$ with probability at least $1 - n^{-1}$ (Assumption 6 and Chebyshev's inequality), it follows that

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

with the same probability. Condition 4 of the lemma thus holds with $\gamma_n = n^{-1}$ and the now (K, c_D)-dependent $B_{2n} = C_L C_X / \sqrt{\left(K-1\right) c_D}$. Lemma D.1 then shows that for any $0 < u \leqslant c_L / \left[\left(1 + \bar{c}_0\right) C_X \sqrt{s}\right]$,

$$\begin{aligned} &\mathrm{P}\left(\sqrt{|I_k^c|}\epsilon_{I_k^c}(u) > \left(\{4C_{L,e}\} \vee \left\{C_\epsilon \sqrt{\frac{s \ln(pn)}{(K-1)c_D}}\right\}\right) u\right) \\ &\qquad \leqslant 4n^{-1} + |I_k^c|^{-1} \leqslant 4n^{-1} + \left[\left(K-1\right) c_D n\right]^{-1}, \end{aligned}$$

where the second inequality follows from $|I_k| \geqslant c_D n$. Now $s \ln\left(pn\right) \geqslant 16\left(K-1\right) c_D C_{L,e}^2 / C_\epsilon^2$ is equivalent to $4C_{L,e} \leqslant C_\epsilon \{s \ln\left(pn\right) / [\left(K-1\right) c_D]\}^{1/2}$, and so with probability at least $1 - (4n^{-1} + [(K-1)c_D n]^{-1})$,

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

where we used the bound $|I_k^c| \geqslant (K-1)c_D n$. The asserted claim now follows from combining this inequality and the union bound. □

Proof of Lemma B.3. Fix a (hold-out) subsample $k \in \{1, \ldots, K\}$. Assumptions 5 and 14 imply that for each $t \in \mathbf{R}$ and each $j \in \{1, \ldots, p\}$, the random variables $S_{I_k^c, j} = |I_k^c|^{-1} \sum_{i \in I_k^c} m_1'(X_i^\top \theta_0, Y_i) X_{ij}$ satisfy

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

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

$$\mathrm{P}\left(\left|S_{I_k^c, j}\right| > t\right) \leqslant 2 \exp\left(-\frac{|I_k^c| t^2}{2 C_X^2 \sigma^2}\right).$$

The union bound then implies that

$$\mathrm{P}\left(\left\|S_{I_k^c}\right\|_\infty > t\right) \leqslant 2p \exp\left(-\frac{|I_k^c| t^2}{2C_X^2 \sigma^2}\right), \quad t > 0,$$

from which we obtain

$$\mathrm{P}\left(\left\|S_{I_k^c}\right\|_\infty > C_X \sigma \sqrt{\frac{2\ln(2p|I_k^c|)}{|I_k^c|}}\right) \leqslant |I_k^c|^{-1} \leqslant \left[(K-1)\, c_D n\right]^{-1},$$

where the second inequality follows from Assumption 12. In addition,

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

where we again used Assumption 12 (recall also that we take $p \geqslant 2$). Combining these inequalities and applying the union bound, we obtain the asserted claim. □

Proof of Lemma B.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,

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

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 B.5. By (B.4) and Lemma B.1,

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

and 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 $k \in \{1,\dots,K\}$, consider events

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

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 B.4 and (B.4) therefore imply that on $\mathscr{Z}_k \cap \mathscr{E}_k$,

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

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

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

Also, Lemma B.3 shows that

$$\mathrm{P}\big((\cap_{k=1}^K \mathscr{Z}_k)^c\big) \leqslant K\,[(K-1)\,c_D n]^{-1}.$$

It thus follows from the union bound that (C.4) holds simultaneously for all $k \in \{1,\dots,K\}$ with probability at least $1 - K\left(4n^{-1} + 2\,[(K-1)\,c_D n]^{-1}\right)$. □

Proof of Lemma B.6. For any $\theta_1, \theta_2 \in \Theta$ and $k \in \{1,\dots,K\}$, 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 B.5 and consider 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 $k \in \{1, \ldots, K\}$,

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

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 (B.5) and Assumption 13, we have $|\Lambda_n| \leqslant 2(\ln n)/\ln(1/a)$. Therefore, the union bound followed by Chebyshev's inequality applied conditional on $\{(X_i, Y_i)\}_{i \in I_k^c}$ gives

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

where the second inequality follows from Assumption 12. Hence, by the union bound and Lemma B.5,

$$\mathrm{P}\left((\mathscr{R} \cap \mathscr{C}(t))^c \right) \leqslant K\left(4n^{-1} + 2\left[(K-1) c_D n \right]^{-1} + t^{-1} \right).$$

We will now prove that (B.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

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

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

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

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

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

where the second inequality follows from (B.5). Assumption 15 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 $k \in \{1, \ldots, K\}$. 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 (C.5), we obtain

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

or rearranging and using the last inequality in (B.5),

$$\begin{aligned}
\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) \leqslant 4C_{ms} \sqrt{\frac{t \ln n}{c_D \ln(1/a)\, n}} &\left(\sqrt{\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big)} + \sqrt{\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big)} \right) \\
&+ 2 \sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big).
\end{aligned}$$

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

$$\sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) \leqslant \frac{32C_{ms}^2}{c_D \ln(1/a)} \frac{t \ln n}{n} + 8 \sum_{k=1}^{K} \frac{|I_k|}{n} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\lambda_*)\big).$$

Combining this bound with Assumption 12 and using (C.6), we obtain

$$\max_{1 \leqslant k \leqslant K} \mathcal{E}\big(\widehat{\theta}_{I_k^c}(\widehat{\lambda}^{\mathsf{cv}})\big) \leqslant \frac{32C_{ms}^2}{c_D^2 \ln(1/a)} \frac{t \ln n}{n} + \frac{8C_{\mathcal{E}}^2}{c_D} \frac{s \ln(pn)}{n},$$

which completes the proof of the lemma. □

D Fundamental Tools

D.1 Maximal Inequality

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

Lemma D.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* $h(0, \cdot) \equiv 0$*. Suppose that there exists constants* $C_h, B_{1n}, B_{2n} \in \mathbf{R}_+, \gamma_n \in (0,1)$ *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* $|t_1| \vee |t_2| \leqslant C_h$,

$$|h(t_1, w) - h(t_2, w)| \leqslant L(w)\,|t_1 - t_2|\,;$$

2. $\sup_{\delta \in \Delta} \left|X^\top \delta\right| \leqslant C_h$ *a.s.;*

3. $\sup_{\delta \in \Delta} \mathrm{E}[h\left(X^\top \delta, W\right)^2] \leqslant B_{1n}^2$*; and,*

4. $\max_{1 \leqslant j \leqslant p} \mathbb{E}_n[L(W_i)^2 X_{ij}^2] \leqslant B_{2n}^2$ *with probability at least* $1 - \gamma_n$*.*

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

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

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

Proof. The claim follows from Belloni et al. (2018a, Lemma D.3), which, in turn, follows from a variant of an argument given in Ledoux and Talagrand (1991). □

D.2 Gaussian Inequality

Lemma D.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* $q^Y(1-\alpha)$ *denote the* $(1-\alpha)$*-quantile of* $\max_{1 \leqslant j \leqslant p} |Y_j|$ *for* $\alpha \in (0,1)$*. Then* $q^Y(1-\alpha) \leqslant (2+\sqrt{2})\sigma\sqrt{\ln(p/\alpha)}$.

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 $p\geqslant 2$. □

D.3 CLT and Bootstrap in High Dimensions

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

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

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 $N_M:=\mathrm{N}(\mathbf{0},M)$.

Theorem D.1 (**High-Dimensional CLT**). *If for some finite constants $b>0$ and $B_n\geqslant 1$,*

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

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

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

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

Let $\widehat{Z}_i$ be an estimator of Z_i, and let $e_1,\dots,e_n$ be i.i.d. $\mathrm{N}(0,1)$ and 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 D.2 (**Multiplier Bootstrap for Many Approximate Means**). *Let (D.1) hold for some finite constants $b > 0$ and $B_n \geqslant 1$, and let $\{\beta_n\}_1^\infty$ and $\{\delta_n\}_1^\infty$ be sequences in $\mathbf{R}_{++}$ both converging to zero such that*

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

Then there exists a finite constant C_b, depending only on b, such that with probability at least $1 - \beta_n - 1/\ln^2(pn)$,

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

Proof. The claim follows from Belloni et al. (2018a, Theorem 2.2), which is here restated 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 D.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*

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

Proof. Given positive definiteness of M, for any $t \in \mathbf{R}$,

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

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 D.3 (**Quantile Comparison**). *If (D.1) holds for some finite constants $b > 0$ and $B_n \geqslant 1$, and*

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

denotes the upper bound in Theorem D.1 multiplied by two, 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, (D.2) holds for some sequences $\{\delta_n\}_1^\infty$ and $\{\beta_n\}_1^\infty$ in $\mathbf{R}_{++}$ both converging to zero, and

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

denotes the upper bound in Theorem D.2 multiplied by two, then with probability at least

$1 - \beta_n - 1/\ln^2(pn)$,

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

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

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

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

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

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 D.2 occurs with probability at least $1 - \beta_n - 1/\ln^2(pn)$. □

Theorem D.4 (**Multiplier Bootstrap Consistency**). *Let (D.1) and (D.2) hold for some finite constants $b > 0$ and $B_n \geqslant 1$ and some sequences $\{\delta_n\}_1^\infty$ and $\{\beta_n\}_1^\infty$ in $\mathbf{R}_{++}$ both converging to zero. Then there exists a finite constant C_b, depending only on b, such that*

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

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

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

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

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

A parallel argument shows

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

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