Asymptotics for estimating a diverging number of parameters — with and without sparsity

Jana Gauss and Thomas Nagler

Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)

November 26, 2024

Abstract

We consider high-dimensional estimation problems where the number of parameters diverges with the sample size. General conditions are established for consistency, uniqueness, and asymptotic normality in both unpenalized and penalized estimation settings. The conditions are weak and accommodate a broad class of estimation problems, including ones with non-convex and group structured penalties. The wide applicability of the results is illustrated through diverse examples, including generalized linear models, multi-sample inference, and stepwise estimation procedures.

1 Introduction

In modern applications, statisticians are facing increasingly complex and high-dimensional problems. Many data sets have a huge number of variables, calling for similarly many parameters p. In other scenarios, the number of variables is moderate, but adequately modeling the data requires highly complex, non-linear models with many parameters. The traditional fixed-p-large-n paradigm is inadequate in such situations.

This article adopts an asymptotic perspective, allowing both the sample size n and the number of parameters p_n to diverge. We consider general parametric problems where the estimator $\hat{\boldsymbol{\theta}}$ solves an estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \mathbf{0} \in \mathbb{R}^{p_n}, \tag{1}$$

with some function $\phi \colon \mathbb{R}^{p_n} \to \mathbb{R}^{p_n}$. A classical example are risk minimization problems, where ϕ is the gradient of a loss function. The estimating equation framework is also suited for more complex methods, such as stepwise estimation procedures, where an optimization-based formulation is less convenient. The main question we address is: under what conditions on the data-generating process, the function ϕ, and the growth of p_n is an estimator solving (1) consistent and asymptotically normal? Both penalized and unpenalized estimation problems are considered in this article.

Unpenalized estimation The study of this problem dates back at least to Huber (1973), who focused on M-estimators in linear models. Following his seminal work, Yohai and Maronna (1979), Portnoy (1984, 1985), Welsh (1989), and Mammen (1989) established consistency and asymptotic normality under various conditions; see Li et al. (2011) for a comprehensive overview. The sharpest known conditions are $p_n \ln p_n/n \to 0$ for consistency (Portnoy, 1984), and $(p_n \ln n)^{3/2}/n \to 0$ (Portnoy, 1985) or $p_n^{3/2} \ln n/n \to 0$ (Mammen, 1989) for asymptotic normality. Fan and Peng (2004) extend these results from (generalized) linear models to general maximum likelihood problems under more restrictive conditions, requiring $p_n^4/n \to 0$ for consistency and $p_n^5/n \to 0$ for asymptotic normality. He and Shao (2000) derived asymptotics for M-estimators with convex loss under a generic stochastic equicontinuity condition, though this approach provides limited insight into the settings where it applies.

The *first contribution* of this work is to extend existing results to much broader classes of estimation problems under verifiable but weak conditions. We provide general conditions for the existence and consistency (Theorem 1), uniqueness (Theorem 2), and asymptotic normality (Theorem 3) of the estimator $\hat{\boldsymbol{\theta}}$ in (1). When specialized, these results closely align with the sharpest known conditions for (generalized) linear models and substantially improve upon those for general maximum-likelihood problems. Importantly, our conditions accommodate more complex estimation settings, including stepwise or multi-sample procedures, as we demonstrate by several corollaries.

Penalized estimation Even the condition $p_n/n \to 0$ is overly restrictive for many modern problems. When p_n exceeds n, consistent estimation is still possible if the true parameter $\boldsymbol{\theta}^*$ is sparse, containing only a few nonzero entries. In such settings, penalized estimation offers a viable solution. Penalized estimation was initially studied in the context of linear models and its variants. Various sparsity-inducing penalties have been proposed, including the Lasso (Tibshirani, 1996), group-structured penalties like the Group Lasso (Yuan and Lin, 2006), and bias-reducing, non-convex penalties such as ℓ_q-penalties (Knight and Fu, 2000), SCAD (Fan, 1997, Fan and Li, 2001), and MCP (Zhang, 2010).

A key question is when such penalized estimators are consistent, both in terms of estimation error and the identification of nonzero parameters. The developments in this area are comprehensively summarized in the recent monograph by Wainwright (2019). State-of-the-art results address general M-estimation problems and broad classes of penalties (Negahban et al., 2012, Lee et al., 2013, Loh and Wainwright, 2015, 2017, Loh, 2017). A unifying theme is the focus on deterministic bounds, where conditions are based on the realized sample rather than population-level quantities. One may then show that these conditions hold with high probability under appropriate assumptions, often at the expense of constants too large for practical inference. Although an asymptotic perspective may sometimes be misleading for finite samples, it offers two distinct advantages. First, regularity conditions are naturally formulated at the population level, providing clarity on the settings where the results are applicable. Second, it enables a precise characterization of estimation uncertainty through distributional limits.

Our *second contribution* is to complement the existing literature with such asymptotic results while broadening their applicability. Penalized inference within the general estimating equation framework (1) can be appropriately formulated as

$$\frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) \in \partial p_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}), \tag{2}$$

where $\partial p_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}})$ denotes the subdifferential of a penalty function $p_{\boldsymbol{\lambda}}$ at $\hat{\boldsymbol{\theta}}$, and $\boldsymbol{\lambda}$ is a vector of tuning parameters. We derive general conditions for consistency (Theorem 4), existence (Theorem 5), and uniqueness (Theorem 6) of such penalized estimators. The solution constructed in Theorem 5 is particularly appealing, because it is selection consistent. Its existence relies on a generalization of a condition known as *mutual incoherence* in the literature. For this solution, we also establish asymptotic normality (Theorem 7). For suitable non-convex penalties, the estimator satisfies the oracle property: it behaves as if the zero entries of $\boldsymbol{\theta}^*$ were known in advance and no penalization is needed. The assumptions on the penalty are mild. Notably, it does not need to be convex or coordinate-separable and can involve multiple tuning parameters with varying strength. We are unaware of other results that apply to such a broad class of penalties. The population-level regularity conditions are weak, permitting up to $p_n = e^{O(n)}$, and relatively straightforward to verify in applications, as we shall illustrate in several corollaries.

Outline The remainder of this article is structured as follows. Section 2 presents the main results for unpenalized estimation. Section 3.1 explains formulation (2) for penalized estimation within general estimating equations. Our conditions on the penalty, along with several examples, are discussed in Section 3.2. Section 3.3 contains the main consistency results and a detailed discussion of the mutual incoherence condition. The asymptotic normality result is provided in Section 3.4. In Section 4, we demonstrate the wide applicability of our results through examples, including M-estimation for generalized linear models (Section 4.1), multi-sample estimation problems in distributed inference and quality control (Section 4.2), and stepwise procedures in causal inference and stochastic optimization (Section 4.3). All proofs are provided in the supplementary material.

2 Unpenalized estimation

2.1 Setup and notation

Suppose we observe independent random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathcal{X}$ and want to estimate a parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p_n})^\top \in \mathbb{R}^{p_n}$ with $p_n \to \infty$ as $n \to \infty$. We do not assume identical distributions to also cover multi-sample estimation problems. The target value $\boldsymbol{\theta}^*$ is the solution to the system of equations $\sum_{i=1}^n \mathbb{E}[\phi_i(\boldsymbol{\theta}^*)] = \mathbf{0}$, with continuous $\phi_i(\boldsymbol{\theta}) = \phi(\boldsymbol{X}_i; \boldsymbol{\theta}) \in \mathbb{R}^{p_n}$; for example, the gradient of a log-likelihood or loss function. The k-th entry of $\phi_i(\boldsymbol{\theta})$ is denoted by $\phi_i(\boldsymbol{\theta})_k, k = 1, \ldots, p_n$. The estimator $\hat{\boldsymbol{\theta}}$ is the solution of

$$\Phi_n(\hat{\boldsymbol{\theta}}) := \frac{1}{n} \sum_{i=1}^n \phi_i(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{3}$$

Define the two $p_n \times p_n$ matrices

$$I(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathrm{Cov}[\phi_i(\boldsymbol{\theta})], \quad J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \mathbb{E}[\phi_i(\boldsymbol{\theta})] = \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_l} \mathbb{E}\left[\phi_i(\boldsymbol{\theta})_k\right] \right)_{k,l=1,\ldots,p_n}.$$

The parameter $\boldsymbol{\theta}^*$, functions $\phi(\boldsymbol{\theta}), J(\boldsymbol{\theta}), I(\boldsymbol{\theta})$, and the support and distribution of the $\boldsymbol{X}_i$ all depend on n, but we suppress this in the notation to avoid clutter.

Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm $\|A\| = \sup_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|$ for matrices. Define $r_n = \sqrt{\mathrm{tr}(I(\boldsymbol{\theta}^*))/n}$ and let $\Theta_n \subset \mathbb{R}^{p_n}$ be a sequence of sets with $\Theta_n \supset \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant r_n C\}$ for all $C < \infty$ and n large.

2.2 Consistency and uniqueness

Our main assumption for consistency of the estimator $\hat{\boldsymbol{\theta}}$ is the following:

(A1) There exists a sequence of symmetric, matrix-valued functions $H_n(\boldsymbol{x})$ such that:

(i) For all $\boldsymbol{u}$ such that $\boldsymbol{\theta}^* + \boldsymbol{u} \in \Theta_n$ and $\boldsymbol{x} \in \mathcal{X}$, it holds

$$\boldsymbol{u}^\top [\phi(\boldsymbol{x}; \boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{x}; \boldsymbol{\theta}^*)] \leqslant \boldsymbol{u}^\top H_n(\boldsymbol{x})\boldsymbol{u};$$

(ii) $\limsup_{n\to\infty} \lambda_{\max}(n^{-1}\sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)]) \leqslant -c < 0$;

(iii) For some sequence $B_n = o(n/\ln p_n)$, it holds

$$\frac{1}{n}\left\|\sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)^2 \mathbb{1}_{\|H_n(\boldsymbol{X}_i)\| \leqslant B_n}]\right\| = o(n/\ln p_n),$$
$$\frac{1}{n}\sum_{i=1}^n \int_{B_n}^\infty \mathbb{P}(\|H_n(\boldsymbol{X}_i)\| > t)dt = o(1).$$

Before discussing the assumptions in detail, we state our main results.

Theorem 1. *Under assumption (A1), the following holds with probability tending to 1:*

(i) The sets Θ_n contain at least one solution of the estimating equation (3).

(ii) Every solution $\hat{\boldsymbol{\theta}} \in \Theta_n$ satisfies

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\sqrt{\frac{\operatorname{tr}(I(\boldsymbol{\theta}^*))}{n}}\right).$$

Theorem 2. *Suppose that (A1) holds, and for any $\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{u} \in \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant a_n r_n\}$ with $a_n \to \infty$ arbitrarily slowly and all $\boldsymbol{x} \in \mathcal{X}$, it holds*

$$\boldsymbol{u}^\top [\phi(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{u}) - \phi(\boldsymbol{x}; \boldsymbol{\theta})] \leqslant \boldsymbol{u}^\top H_n(\boldsymbol{x})\boldsymbol{u}. \tag{4}$$

Then the solution $\hat{\boldsymbol{\theta}}$ is unique on Θ_n with probability tending to 1.

The convergence rate $r_n = \sqrt{\operatorname{tr}(I(\boldsymbol{\theta}^*))/n}$ is typically $\sqrt{p_n/n}$, for example when $\max_{i,k} \mathbb{E}\left[\phi(\boldsymbol{X}_i; \boldsymbol{\theta}^*)_k^2\right] = O(1)$. However, the above formulation allows the entries of ϕ to diverge. This is useful in settings with a slower rate of convergence, e.g., the multi-sample estimation problems discussed in Section 4.2.

To understand condition (A1), it is instructive to outline the core idea of the proof. A sufficient condition for the estimating equation (3) to have solution on a ball $\{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant r_n C\}$ is (Fierro et al., 2004, Theorem 2.3):

$$\sup_{\|\boldsymbol{u}\| = r_n C} \boldsymbol{u}^\top \Phi_n(\boldsymbol{\theta}^* + \boldsymbol{u}) \leqslant 0.$$

Condition (i) allows us to majorize the left-hand side via

$$\boldsymbol{u}^\top \Phi_n(\boldsymbol{\theta}^* + \boldsymbol{u}) \leqslant \boldsymbol{u}^\top \Phi_n(\boldsymbol{\theta}^*) + \boldsymbol{u}^\top \left(\frac{1}{n}\sum_{i=1}^n H_n(\boldsymbol{X}_i)\right)\boldsymbol{u}.$$

Standard arguments give $\|\Phi_n(\boldsymbol{\theta}^*)\| = O_p(r_n)$, and the sample average in the quadratic form can be approximated by its expectation, using condition (iii) and concentration inequalities for random matrices. This expectation is negative definite by (ii), so the second term on the right is strictly negative. For $\|\boldsymbol{u}\| = r_n C$ with sufficiently large C, the second term in the upper bound dominates the first. This shows that the estimating equation has a solution on $\{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant r_n C\}$ with high probability. Consistency of other solutions and uniqueness follow from variations of the same argument.

The majorization approach provides uniform control over the curvature around $\boldsymbol{\theta}^*$ without relying on covering arguments. This appears essential for establishing non-restrictive conditions on the growth rate of p_n. Examples of constructing H_n in various settings are provided in Section 4. The matrix-valued function H_n may be interpreted as an envelope for the family of symmetrized Jacobians

$$\left\{ H_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{2}\, \nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{x}; \boldsymbol{\theta}) + \frac{1}{2}\, \nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{x}; \boldsymbol{\theta})^\top : \boldsymbol{\theta} \in \Theta_n \right\}$$

in the positive semi-definite ordering $\preceq$ of matrices[1], but condition (i) is slightly weaker. Condition (ii) is an on-average concavity constraint, but $\boldsymbol{u} \mapsto \boldsymbol{u}^\top H_n(\boldsymbol{x})\boldsymbol{u}$ does not have to be concave for every $\boldsymbol{x}$. Rate conditions as in (iii) and similar form will be used throughout this article. Lemma 1 and Lemma 2 in the appendix are useful tools for converting them into more interpretable moment conditions. For example, the second part of (iii) is satisfied if $p_n \ln p_n / n \to 0$, $\max_i \mathbb{E}[\|H_n(\boldsymbol{X}_i)\|] = O(p_n)$, and $\max_i \mathbb{V}\mathrm{ar}\,[\|H_n(\boldsymbol{X}_i)\|] = O(p_n)$. Further, the first part of (iii) follows automatically from the second if all $H_n(\boldsymbol{x})$ are negative semi-definite and $n^{-1}\sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)] = O(1)$. The weakest conditions are achieved by choosing $\Theta_n = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant r_n a_n\}$ for $a_n \to \infty$ arbitrarily slow. In this case, only the local behavior around a shrinking neighborhood of $\boldsymbol{\theta}^*$ is relevant. The conditions become stronger when considering larger sets Θ_n because they potentially require larger envelopes H_n in (A1)(i) and (4).

The assumptions of the theorems implicitly restrict the rate of growth of the number of parameters p_n. Results of this form have been studied primarily in the context of M-estimation of (generalized) linear regression models, where $\phi_i(\boldsymbol{\theta}) = \psi(Y_i, \boldsymbol{X}_i^\top\boldsymbol{\theta})\boldsymbol{X}_i$ for some ψ. In this setting, we obtain the condition $p_n \ln p_n / n \to 0$ under standard assumptions. This corresponds to the best known results obtained by Portnoy (1984); see Section 4.1 for more details. For maximum-likelihood estimation in general non-linear models, Fan and Peng (2004) required the much stronger condition $p_n^4/n \to 0$. Technically, we do not even require $p_n/n \to 0$ unless dictated by condition (A1)(iii). This is sensible, because the vector equation (3) may be composed of entirely unrelated estimating equations for the individual parameters θ_k, $k = 1, \dots, p_n$. In such cases, we may choose H_n as a diagonal matrix. If, for example, all entries of H_n are uniformly bounded, (A1)(iii) holds with $B_n = O(1)$ as long as $p_n \leqslant e^{o(n)}$.

2.3 Asymptotic normality

We now turn to the asymptotic normality of the estimator. Because the dimension of the parameter space grows with the sample size, we state this in terms of convergence of finite-dimensional projections. Let $A_n \in \mathbb{R}^{q \times p_n}$ be some matrix.

[1] $A \preceq B$ if $B - A$ is positive semi-definite or, equivalently, $\boldsymbol{u}^\top A\boldsymbol{u} - \boldsymbol{u}^\top B\boldsymbol{u} \leqslant 0$ for all $\boldsymbol{u}$.

(A2) For every $C \in (0, \infty)$ and some sequence $B_n = o(\sqrt{n}/(r_n p_n))$, it holds

$$\sup_{\|\boldsymbol{u}\|, \|\boldsymbol{u}'\| \leqslant r_n C} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{E}\left[\|A_n[\phi_i(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi_i(\boldsymbol{\theta}^* + \boldsymbol{u}')]\|^2\right]}{\|\boldsymbol{u} - \boldsymbol{u}'\|^2} = o\left(\frac{1}{r_n^2 p_n}\right),$$

$$\sum_{i=1}^{n} \mathbb{P}\left(\sup_{\|\boldsymbol{u}\|, \|\boldsymbol{u}'\| \leqslant r_n C} \frac{\|A_n[\phi_i(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi_i(\boldsymbol{\theta}^* + \boldsymbol{u}')]\|}{\|\boldsymbol{u} - \boldsymbol{u}'\|} > B_n\right) = o(1),$$

$$\sup_{\|\boldsymbol{u}\| \leqslant r_n C} \|A_n[J(\boldsymbol{\theta}^* + \boldsymbol{u}) - J(\boldsymbol{\theta}^*)]\| = o\left(\frac{1}{\sqrt{n}\, r_n}\right).$$

(A3) It holds $\max_{1 \leqslant i \leqslant n} \mathbb{E}\left[\|A_n \phi_i(\boldsymbol{\theta}^*)\|^4\right] = o(n)$.

Assumption (A2) is a stochastic smoothness condition required to control fluctuations of the estimating equation. The moment condition (A3) typically requires $p_n^2/n \to 0$, e.g., if $\|A_n\| = O(1)$ and $\max_{i,k} \mathbb{E}[\phi_i(\boldsymbol{\theta}^*)_k^4] = O(1)$.

Theorem 3. *If conditions (A1)–(A3) hold for some matrix $A_n \in \mathbb{R}^{q \times p_n}$ for which $\Sigma = \lim_{n\to\infty} A_n I(\boldsymbol{\theta}^*) A_n^\top$ exists, it holds*

$$\sqrt{n} A_n J(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to_d \mathcal{N}(\mathbf{0}, \Sigma).$$

In the finite-dimensional setting ($p_n = p < \infty$), we can choose $A_n = J(\boldsymbol{\theta}^*)^{-1}$. Then, Theorem 3 yields the typical result

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to_d \mathcal{N}(\mathbf{0}, J(\boldsymbol{\theta}^*)^{-1} I(\boldsymbol{\theta}^*) J(\boldsymbol{\theta}^*)^{-\top}).$$

The proof of Theorem 3 uses mixed-entropy inequalities (Van der Vaart and Wellner, 2023, Section 2.14) building on Talagrand's work on generic chaining (Talagrand, 2005) to obtain weak conditions. For M-estimators in the linear regression model, we obtain asymptotic normality if $p_n^2 \ln n/n \to 0$ under standard assumptions, see Corollary 1 in Section 4.1. Portnoy (1985) and Mammen (1989) obtain the slightly weaker condition $p_n^{3/2} \ln n/n \to 0$. Both authors exploit the linear model structure through fourth-order expansions and a design-dependent standardization, effectively alleviating the need for the sample Hessian to converge. Portnoy (1986) showed that a general CLT cannot hold unless $p_n^2/n \to 0$, so our result appears to be close to optimal. Fan and Peng (2004) require the much stronger condition $p_n^5/n \to 0$ for maximum-likelihood estimation in general non-linear models.

3 Penalized estimation

3.1 Penalization of estimating equations

To motivate our formulation, consider a penalized risk minimization problem. Let L be some differentiable loss function and p_{λ_n} be a function penalizing the magnitude of $\boldsymbol{\theta}$, e.g., the ℓ_1-penalty $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \|\boldsymbol{\theta}\|_1$. The estimator is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left(\frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{X}_i; \boldsymbol{\theta}) + p_{\lambda_n}(\boldsymbol{\theta})\right).$$

If both penalty and loss are differentiable, the first-order condition for $\hat{\boldsymbol{\theta}}$ is

$$\frac{1}{n}\sum_{i=1}^{n} -\nabla_{\boldsymbol{\theta}} L(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) - p'_{\lambda_n}(\hat{\boldsymbol{\theta}}) = \mathbf{0},$$

where $p'_{\lambda_n}(\hat{\boldsymbol{\theta}})$ is the gradient of p_{λ_n} at $\hat{\boldsymbol{\theta}}$. Sparsity inducing penalties are discontinuous at points where some coordinates are zero. The vector $\hat{\boldsymbol{\theta}}$ is a solution to the minimization problem when the gradients cross $\mathbf{0}$ at $\hat{\boldsymbol{\theta}}$, either continuously or with a jump. This is illustrated in the left panel of Fig. 1.

To appropriately deal with the non-differentiability of the penalty, we reformulate the estimation problem. Recall from Clarke (1990, Chapter 2) that $\boldsymbol{z} \in \mathbb{R}^p$ is a *generalized gradient* or *subgradient* of a Lipschitz function $f : \mathbb{R}^p \to \mathbb{R}$ at $\boldsymbol{\theta}$, if $f(\boldsymbol{\theta} + \boldsymbol{\Delta}) - f(\boldsymbol{\theta}) \geqslant \langle \boldsymbol{z}, \boldsymbol{\Delta} \rangle$, for all $\boldsymbol{\Delta}$ in a neighborhood of $\mathbf{0}$. It coincides with the usual derivative if the function is differentiable. The collection of all subgradients of f at $\boldsymbol{\theta}$ is called subdifferential and denoted by $\partial f(\boldsymbol{\theta})$. For example, $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \|\boldsymbol{\theta}\|_1$ has subdifferentials $\partial p_{\lambda_n}(\boldsymbol{\theta})_k = \{\lambda_n \operatorname{sign}(\theta_k)\}$, where $\operatorname{sign}(0)$ is allowed to be any number in $[-1, 1]$. With this notation, the estimator $\hat{\boldsymbol{\theta}}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} -\nabla_{\boldsymbol{\theta}} L(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) \in \partial p_{\lambda_n}(\hat{\boldsymbol{\theta}}).$$

More generally, we define a penalized estimator $\hat{\boldsymbol{\theta}}$ as a solution to

$$\Phi_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \phi_i(\boldsymbol{\theta}) \in \partial p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}). \tag{5}$$

We allow for a vector of tuning parameters $\boldsymbol{\lambda}_n = (\lambda_{n,1}, \dots, \lambda_{n,p_n})$. This encompasses, for example, penalties of the form $p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}) = \sum_{k=1}^{p_n} \lambda_{n,k}\, p(\theta_k)$, which can be useful when the estimator $\hat{\boldsymbol{\theta}}$ is composed of solutions to sub-problems of different dimensionality or sample size. The target parameter $\boldsymbol{\theta}^*$ still solves the unpenalized population equation $\sum_{i=1}^{n} \mathbb{E}[\phi_i(\boldsymbol{\theta}^*)] = \mathbf{0}$. In what follows, we assume without loss of generality that the true $\boldsymbol{\theta}^*$ can be written as $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)})$ with $\boldsymbol{\theta}^*_{(1)} \in \mathbb{R}^{s_n}$ and $\boldsymbol{\theta}^*_{(2)} = \mathbf{0} \in \mathbb{R}^{p_n - s_n}$. Similarly, write $\boldsymbol{v}_{(1)} = (v_1, \dots, v_{s_n})$, $\boldsymbol{v}_{(2)} = (v_{s_n+1}, \dots, v_{p_n})$ for any vector $\boldsymbol{v} \in \mathbb{R}^{p_n}$.

A penalty that is not differentiable at $\mathbf{0}$ can induce sparsity. To illustrate this, consider the Lasso penalty $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \|\boldsymbol{\theta}\|_1$ with scalar λ_n. Recall that the subdifferential of the Lasso penalty $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \|\boldsymbol{\theta}\|$ at $\mathbf{0}$ is the set $\partial p_{\lambda_n}(\boldsymbol{\theta}) = [-\lambda_n, \lambda_n]^{p_n}$. More generally, sparsity-inducing penalties typically satisfy

$$\partial p_{\lambda_n}(\boldsymbol{\theta})_{(2)} \supseteq [-\lambda_n, \lambda_n]^{p_n - s_n},$$

for all $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_{(2)} = \mathbf{0}$. The subdifferential collapses to the gradient if it exists, i.e., $\partial_{\theta_k} p_{\lambda_n}(\boldsymbol{\theta}) = \{\lambda_n \operatorname{sign}(\theta_k)\}$ is a singleton for $\theta_k \neq 0$ for the Lasso. This means that a solution with $\hat{\theta}_k > 0$ must satisfy $\Phi_n(\hat{\boldsymbol{\theta}})_k = \lambda_n$, and one with $\hat{\theta}_k < 0$ must satisfy $\Phi_n(\hat{\boldsymbol{\theta}})_k = -\lambda_n$. This biases the estimating equation to favor parameters with smaller magnitude. For $\hat{\boldsymbol{\theta}}$ with $\hat{\theta}_k = 0$ to be a solution, we only need $\Phi_n(\hat{\boldsymbol{\theta}})_k \in [-\lambda_n, \lambda_n]$. Here, we give the criterion $\Phi_n(\hat{\boldsymbol{\theta}})_k$ extra wiggle room to make it more likely that the penalized criterion has solutions with $\hat{\theta}_k = 0$.

This principle is illustrated in Fig. 1 for a one-dimensional θ. In the left column, the dashed lines show two possible realizations of $\Phi_n(\theta)$. Its root, the unpenalized estimator, is denoted by $\hat{\theta}_u$. Now we add a Lasso penalty with $\lambda = 1$. The solid lines show $\Phi_n(\theta) - p'_\lambda(\theta)$, whose root is the penalized estimator $\hat{\theta}$. The Lasso is not differentiable at zero, so $\Phi_n(\theta) - p'_\lambda(\theta)$ is not continuous and has a jump of size 2λ at $\theta = 0$. If this jump crosses 0, then $\hat{\theta} = 0$, which is the

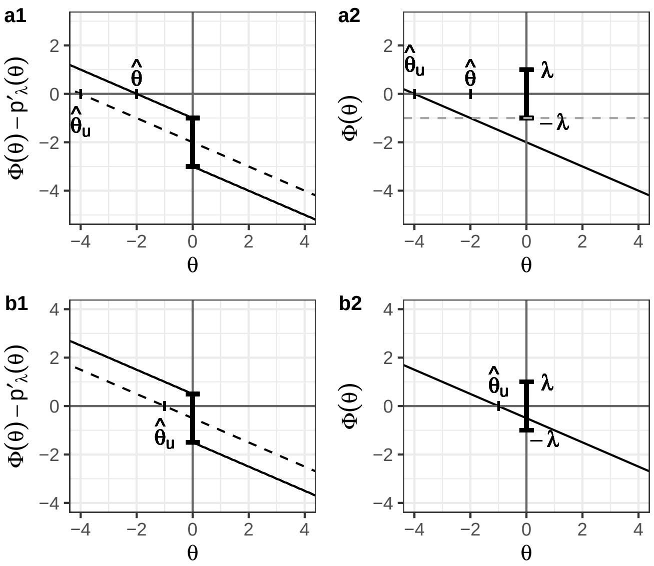

Figure 1: Two views on sparsity-inducing penalties: Left column: realization of $\Phi_n(\theta)$ (dashed line) and Lasso-penalized version (solid line). Right column: Penalized estimator as solution to $\Phi_n(\theta) \in \partial p_\lambda(\theta)$.

case in the lower left corner. The right column illustrates the view described above. Here, the solid line denotes $\Phi_n(\theta)$. A negative $\hat{\theta}$ must satisfy $\Phi_n(\hat{\theta}) = -\lambda$, which is the case in the upper right corner. The penalty shrinks the coefficient, i.e., $|\hat{\theta}| < |\hat{\theta}_u|$. The penalized estimator is zero if $\Phi_n(0) \in [-\lambda, \lambda]$, which is shown in the lower right corner.

More generally, we may define the penalized estimator $\hat{\boldsymbol{\theta}}$ as a solution to $\Phi_n(\hat{\boldsymbol{\theta}}) \in \mathcal{I}_n(\hat{\boldsymbol{\theta}})$ with some function $\mathcal{I}_n$ that maps $\boldsymbol{\theta}$ to p_n-dimensional rectangles. We shall continue with the penalty formulation for better interpretability. Adapting the following results to more general interval-valued functions $\mathcal{I}_n(\boldsymbol{\theta})$ is straightforward.

3.2 Examples and conditions for the penalty

Some common penalty functions covered by our framework are given in the following examples.

Example 3.1 (Lasso)**.** The ℓ_1 penalty (Tibshirani, 1996) $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \|\boldsymbol{\theta}\|_1$ is not differentiable at 0 and the subgradient is given by $p'_{\lambda_n}(\boldsymbol{\theta})_k = \lambda_n \operatorname{sign}(\theta_k)$, where $\operatorname{sign}(0)$ is allowed to be any number in $[-1, 1]$.

Example 3.2 (Elastic Net)**.** The elastic net penalty (Zou and Hastie, 2005) is given by $p_{\lambda_{n,1},\lambda_{n,2}}(\boldsymbol{\theta}) = \lambda_{n,1}\|\boldsymbol{\theta}\|_1 + \lambda_{n,2}\|\boldsymbol{\theta}\|_2^2$. It holds $p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})_k = \lambda_{n,1} \operatorname{sign}(\theta_k) + 2\lambda_{n,2}\theta_k$. For $\lambda_{n,1} > 0$, the elastic net can induce sparsity.

Example 3.3 (Group Lasso)**.** The Group Lasso penalty (Yuan and Lin, 2006) is given by $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \sum_{i=1}^K \|\boldsymbol{\theta}_{G_i}\|_2$ with groups $G_1, \dots, G_K \subset \{1, \dots, p_n\}$. In this case, $\boldsymbol{\theta}_{(1)}$ contains all θ_k belonging to groups G_i with at least one non-zero θ_k^* and $\boldsymbol{\theta}_{(2)}$ contains all θ_k belonging to at least one group G_i with $\boldsymbol{\theta}^*_{G_i} = \mathbf{0}$. The subgradient is given by

$$p'_{\lambda_n}(\boldsymbol{\theta})_k = \lambda_n \sum_{i=1}^K \mathbb{1}\{k \in G_i\} \left[\mathbb{1}\{\|\boldsymbol{\theta}_{G_i}\|_2 \neq 0\} \frac{\theta_k}{\|\boldsymbol{\theta}_{G_i}\|_2} + \mathbb{1}\{\|\boldsymbol{\theta}_{G_i}\|_2 = 0\} \operatorname{sign}(\theta_k) \right].$$

The Group Lasso is non-differentiable at points where $\|\boldsymbol{\theta}_{G_i}\|_2 = 0$ for some of the groups.

Example 3.4 (ℓ_q penalty)**.** The ℓ_q penalty (Frank and Friedman, 1993) $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \|\boldsymbol{\theta}\|_q^q$ with $q \in (0, 1]$ is not differentiable at 0 and can therefore induce sparsity. The ℓ_q penalty is also not Lipschitz around 0 for $q < 1$, but the definition of a subdifferential can be extended to non-smooth functions as in Clarke (1990, Section 2.4). This leads to subdifferentials of the form $\partial_{\theta_k} p_{\lambda_n}(\boldsymbol{\theta}) = \{q \operatorname{sign}(\theta_k)|\theta_k|^{q-1}\}$ for $\theta_k \neq 0$ and $\partial_{\theta_k} p_{\lambda_n}(\boldsymbol{\theta}) = \mathbb{R}$ for $\theta_k = 0$. The ℓ_q penalty with $q < 1$ is special in the sense that $\boldsymbol{\theta} = \mathbf{0}$ is always a solution to the penalized estimating equation.

Example 3.5 (SCAD)**.** The SCAD penalty (*smoothly clipped absolute deviation*) (Fan, 1997, Fan and Li, 2001) was developed with the aim of obtaining a penalized estimator that is unbiased for large parameters. For a single parameter, the SCAD penalty and its derivative are defined as

$$p_\lambda(\theta) = \begin{cases} \lambda|\theta| & \text{if } |\theta| \leqslant \lambda, \\ \frac{2a\lambda|\theta| - \theta^2 - \lambda^2}{2(a-1)} & \text{if } \lambda < |\theta| \leqslant a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\theta| > a\lambda, \end{cases} \qquad p'_\lambda(\theta) = \begin{cases} \lambda \operatorname{sign}(\theta) & \text{if } |\theta| \leqslant \lambda, \\ \frac{\operatorname{sign}(\theta)(a\lambda - |\theta|)}{(a-1)} & \text{if } \lambda < |\theta| \leqslant a\lambda, \\ 0 & \text{if } |\theta| > a\lambda, \end{cases}$$

for some $a > 2$. For multiple parameters, the SCAD penalty is used componentwise as $p_{\lambda_n}(\boldsymbol{\theta}) = \sum_{k=1}^{p_n} p_{\lambda_n}(\theta_k)$. Around the origin, SCAD coincides with the Lasso penalty, so it can induce sparsity. Since the derivative $p'_\lambda(\theta)$ is zero for all $|\theta| \geqslant a\lambda$, it leads to an unbiased estimating equation for large parameters.

Example 3.6 (MCP)**.** A penalty that is unbiased for large coefficients (i.e., $p'_\lambda(|\theta|) = 0$ for $|\theta| \geqslant a\lambda$ with some $a > 0$) and induces sparsity (i.e., $\lim_{\theta \downarrow 0} p'_\lambda(\theta) = \lambda$) must be nonconvex. The *minimax concave penalty* (MCP) (Zhang, 2010) is the "most convex" penalty among the penalties satisfying unbiasedness and sparsity, i.e., it minimizes the maximum concavity. For a single parameter, it is given by

$$p_\lambda(\theta) = \mathbb{1}\{|\theta| \leqslant a\lambda\} \left(\lambda|\theta| - \frac{\theta^2}{2a} \right) + \mathbb{1}\{|\theta| > a\lambda\} \frac{a\lambda^2}{2},$$

with derivative

$$p'_\lambda(\theta) = \mathbb{1}\{|\theta| \leqslant a\lambda\} \operatorname{sign}(\theta) \frac{(a\lambda - |\theta|)}{a}$$

for some $a > 0$ (Fan and Lv, 2010). The penalty shares a similar behavior to SCAD in that it is equivalent to Lasso around $\theta = 0$ and leads to unbiased estimating equations for $|\theta|$ large.

Example 3.7 (Fusion penalty)**.** The aim of a *fusion penalty* is to reduce the number of different coefficients, which is of particular interest for categorical data (Tibshirani et al., 2005). This is achieved by penalizing differences between coefficients, e.g., $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \sum_{k=1}^{p_n - 1} |\theta_{k+1} - \theta_k|$. This penalty fits in the presented framework using a reparametrization, i.e., by defining $\beta_1 = \theta_1$ and $\beta_k = \theta_k - \theta_{k-1}$ and adapting the estimation function ϕ accordingly. Then, zero-entries of $\boldsymbol{\beta}_n$ correspond to parameters being "fused".

All these penalties can be used with a vector of tuning parameters $\boldsymbol{\lambda}_n$, i.e., specific penalty parameters $\lambda_{n,k}$ for each θ_k. Most penalties can then be expressed as $p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}) = \sum_{k=1}^{p_n} p_{\lambda_{n,k}}(\theta_k)$. For Group Lasso, a group-specific λ_{n,G_i} can be introduced for each group G_i. Different $\lambda_{n,k}$ are especially desirable in settings of grouped or stepwise estimation of parameters. The examples show a wide range of different behaviors. In particular, we do not assume that the penalties are convex or decomposable into a sum of component penalties. To state our conditions, let r_n by the target rate of convergence defined in Section 3.3.1 ahead. Define $\Theta'_n = \{(\boldsymbol{\theta}_{(1)}, \mathbf{0}) \colon \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant r_n a_n\}$ with $a_n \to \infty$ arbitrarily slowly, and let $\Theta_n \supset \Theta'_n$ be sets specified further in the theorems to follow.

(P1) The penalty $p_{\boldsymbol{\lambda}_n}$ is twice continuously differentiable with respect to $\boldsymbol{\theta}_{(1)}$ at $\boldsymbol{\theta} \in \Theta_n'$ with

$$\sup_{\boldsymbol{\theta} \in \Theta_n'} \|\nabla^2_{\boldsymbol{\theta}_{(1)}} p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})\| = o(1).$$

(P2) The subdifferential satisfies

$$\partial p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})_{(2)} \supseteq [-\lambda_{n,s_n+1}, \lambda_{n,s_n+1}] \times \ldots \times [-\lambda_{n,p_n}, \lambda_{n,p_n}] \quad \text{for all } \boldsymbol{\theta} \in \Theta_n'.$$

(P3) There is $\mu_n \geqslant 0$ such that for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_n$ and valid subgradients $p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}), p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}')$,

$$\langle \boldsymbol{\theta}' - \boldsymbol{\theta}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}') - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}) \rangle \geqslant -\frac{1}{2}\mu_n \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2.$$

(P4) The Hessian matrix of $p_{\boldsymbol{\lambda}_n}$ with respect to $\boldsymbol{\theta}_{(1)}$ satisfies

$$\sup_{\boldsymbol{\theta} \in \Theta_n'} \|\nabla^2_{\boldsymbol{\theta}_{(1)}} p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})\| = o\left(\frac{1}{\sqrt{n}\, r_n}\right).$$

For the Lasso and SCAD, differentiability in (P1) requires $\min_{1 \leqslant k \leqslant s_n} |\theta_k^*|/r_n \to \infty$: to remain identifiable under penalization, the nonzero parameters $\theta_1^*, \ldots, \theta_{s_n}^*$ are not allowed to vanish too fast. For the Group Lasso, the maximal coefficient of each group with at least one $\theta_k^* \neq 0$ must not decay too fast. The Hessian condition is mild and typically implied by $\max_k \lambda_{n,k} \to 0$. For Lasso, Group Lasso, SCAD, and MCP penalties, condition (P2) holds with equality instead of $\supseteq$, but using $\supseteq$ also allows for ℓ_q penalties with $q < 1$ or asymmetric penalties of the form $p_{\lambda_n}(\boldsymbol{\theta})_k = \lambda_n(c_1|\theta_k| \cdot \mathbb{1}\{\theta_k \geqslant 0\} + c_2|\theta_k| \cdot \mathbb{1}\{\theta_k < 0\})$. Assumption (P2) ensures that the penalty induces sparsity. Assumption (P3) limits the degree of non-convexity of the penalty and is required to guarantee uniqueness. For convex penalties, the assumption is always true with $\mu_n = 0$. It also holds, e.g., for the non-convex SCAD and MLP penalties with $\mu_n = (a-1)^{-1}$ and $\mu_n = a^{-1}$, respectively. It fails for ℓ_q-penalties with $q < 1$, which cannot lead to unique solutions unless $\boldsymbol{\theta}^* = \mathbf{0}$. The condition is implied by, but weaker than the μ-amenability condition in Loh and Wainwright (2017) and Loh (2017). Assumption (P4) is a refinement of (P1) and only required for asymptotic normality. For most penalties (including SCAD and Lasso), (P4) is trivial since $\|\nabla^2_{\boldsymbol{\theta}_{(1)}} p_{\lambda_n}(\boldsymbol{\theta})\| = \mathbf{0}$ for small enough λ_n.

3.3 Estimation and selection consistency

3.3.1 Notation and assumptions

To simplify some conditions, we shall assume from now on that $p_n \geqslant n^a$ for some $a > 0$. This only excludes uninteresting edge cases where p_n is effectively constant. To state the required assumptions, let

$$b_n^* = \|p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)}\|_\infty, \qquad r_n = \sqrt{\frac{\operatorname{tr}(I(\boldsymbol{\theta}^*)_{(1)})}{n}} + \sqrt{s_n} b_n^*,$$

and recall $\Theta_n' = \{(\boldsymbol{\theta}_{(1)}, \mathbf{0}) \colon \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant r_n a_n\}$ with $a_n \to \infty$ arbitrarily slowly. Further, define the cones

$$\Theta^*(\nu_n) = \{\boldsymbol{\theta} \colon \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_1 \leqslant \sqrt{\nu_n} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2\},$$

and suppose that Θ_n are sets such that $\Theta' \subseteq \Theta_n \subseteq \Theta^*(\nu_n)$. Define

$$
\begin{aligned}
J(\boldsymbol{\theta})_{(1)} &= \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}_{(1)}} \mathbb{E}[\phi(\boldsymbol{X}_i;\boldsymbol{\theta})_{(1)}] \in \mathbb{R}^{s_n \times s_n},\\
J(\boldsymbol{\theta})_{(2,1)} &= \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}_{(1)}} \mathbb{E}[\phi(\boldsymbol{X}_i;\boldsymbol{\theta})_{(2)}] \in \mathbb{R}^{(p_n - s_n) \times s_n},\\
J(\boldsymbol{\theta})_{k,(1)} &= \frac{1}{n}\sum_{i=1}^n \nabla^\top_{\boldsymbol{\theta}_{(1)}} \mathbb{E}[\phi(\boldsymbol{X}_i;\boldsymbol{\theta})_{k}] \in \mathbb{R}^{1 \times s_n},\\
I(\boldsymbol{\theta})_{(1)} &= \frac{1}{n}\sum_{i=1}^n \mathrm{Cov}[\phi(\boldsymbol{X}_i;\boldsymbol{\theta})_{(1)}] \in \mathbb{R}^{s_n \times s_n},\\
\bar{b}_n &= \sup_{\boldsymbol{v} \in \partial p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_n} \|\boldsymbol{v}\|_\infty,
\end{aligned}
$$

and let η_n be any sequence such that

$$
\eta_n \geqslant 2\sigma_n \sqrt{\frac{\ln p_n}{n}}, \qquad \text{where} \qquad \sigma_n = \max_{1 \leqslant k \leqslant p_n} \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi_i(\boldsymbol{\theta}^*)_k^2]}. \tag{6}
$$

Our main regularity conditions for penalized estimation are as follows:

(A4) There exists a sequence of symmetric, matrix-valued functions $H_n(\boldsymbol{x})$ such that:

- (i) For all $\boldsymbol{\theta}^* + \boldsymbol{u} \in \Theta_n$ and $\boldsymbol{x} \in \mathcal{X}$, it holds
$$\boldsymbol{u}^\top[\phi(\boldsymbol{x};\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{x};\boldsymbol{\theta}^*)] \leqslant \boldsymbol{u}^\top H_n(\boldsymbol{x})\boldsymbol{u};$$
- (ii) $\limsup_{n\to\infty} \lambda_{\max}(n^{-1}\sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)]) \leqslant -c < 0$;
- (iii) For some sequence $B_n = o(n/(\nu_n \ln p_n))$, it holds
$$\max_{1 \leqslant j,k \leqslant p_n} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)_{j,k}^2] = o\left(\frac{n}{\nu_n^2 \ln p_n}\right)$$
$$\sum_{i=1}^n \mathbb{P}\left(\max_{1 \leqslant j,k \leqslant p_n} |H_n(\boldsymbol{X}_i)_{j,k}| > B_n\right) = o(1).$$

(A5) It holds $\displaystyle\sum_{i=1}^n \mathbb{P}\left(\|\phi_i(\boldsymbol{\theta}^*)\|_\infty > \sigma_n \sqrt{\frac{n}{4\ln p_n}}\right) = o(1)$.

(A6) There is $\alpha \in [0,1)$ such that

$$
\left\| \mathrm{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} J(\boldsymbol{\theta})_{(2,1)}\ J^{-1}(\boldsymbol{\theta})_{(1)}\ p'_{\boldsymbol{\lambda}_n}(\tilde{\boldsymbol{\theta}})_{(1)} \right\|_\infty \leqslant \alpha \tag{7}
$$

for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta'$.

(A7) The subvector $\boldsymbol{\lambda}_{n(2)}$ must fulfill

$$
\lambda_{n,k} \geqslant \frac{4}{1-\alpha} J_{n,k}\eta_n \qquad \text{for all } k = s_n + 1, \dots, p_n,
$$

where α is defined in (A6), η_n is defined in (6) and

$$
J_{n,k} = \max\left\{1, \sup_{\boldsymbol{\theta} \in \Theta'} \left\|(J(\boldsymbol{\theta})_{k,(1)}\ J^{-1}(\boldsymbol{\theta})_{(1)})^\top\right\|_1\right\}.
$$

(A8) It holds

$$\max_{1 \leqslant k \leqslant p_n} \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta'} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{E}[|\phi_i(\boldsymbol{\theta})_k - \phi_i(\boldsymbol{\theta}')_k|^2]}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2} = o\left(\frac{n\eta_n^2}{r_n^2(s_n + \ln p_n)}\right),$$

$$\sum_{i=1}^{n} \mathbb{P}\left(\sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta'} \frac{\|\phi_i(\boldsymbol{\theta}) - \phi_i(\boldsymbol{\theta}')\|_\infty}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} > \tilde{B}_n\right) = o(1),$$

with η_n as defined in (A7) and some sequence $\tilde{B}_n = o(n\eta_n/(r_n s_n + r_n \ln p_n))$.

Assumption (A4) is a variant of (A1) that guarantees a restricted form of concavity with high probability; see the following section for further comments. Assumption (A5) is a tail condition trading off the moments of the estimating equation with the number of parameters p_n. For example, if $\sigma_n = O(1)$ and all $\phi(\boldsymbol{X}_i; \boldsymbol{\theta}^*)_k$ have sub-Gaussian tail, we may take p_n as large as $p_n \sim e^{n^a}$ for some $a \in (0, 1/2)$. The remaining conditions are only required for constructing an explicit solution that is selection consistent. Assumption (A6) generalizes the *mutual incoherence* or *irrepresentable* conditions (Wainwright, 2019, Zhao and Yu, 2006, Bühlmann and van de Geer, 2011) and is discussed in more detail in Section 3.3.3. Assumption (A7) requires the penalty parameters $\lambda_{n,k}$ to be large enough for obtaining sparse solutions. In the recent literature (Negahban et al., 2012, Loh and Wainwright, 2015), such conditions are often stated with reference to the realization of $n^{-1}\|\sum_{i=1}^{n} \phi_i(\boldsymbol{\theta}^*)\|_\infty$, which we replace by a population bound η_n. The final condition (A8) guarantees sufficiently fast convergence of the estimating equation to its population counterpart. It leads to growth restrictions on p_n analogous to our discussion of (A5). The condition becomes weaker if we make η_n in (6) large, but this requires us to use a larger penalty parameter $\boldsymbol{\lambda}_n$ in (A7). This is less problematic for non-convex penalties, see the discussion following Theorem 5.

3.3.2 Main results

We will provide three main results for penalized estimators. The first guarantees that any sparse solution to the penalized estimating equation is close to the true parameter vector.

Theorem 4. *Suppose that (A4) and (A5) holds. Then any solution $\hat{\boldsymbol{\theta}} \in \Theta_n \subseteq \Theta_n^*(\nu_n)$ satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\sqrt{\nu_n}(\eta_n + \bar{b}_n)\right).$$

The optimal rate of convergence $\sqrt{s_n}\eta_n$ is attained if $\nu_n = O(s_n)$ and $\bar{b}_n = O(\eta_n)$ which typically holds if $\|\boldsymbol{\lambda}_n\|_\infty = O(\eta_n)$. In most classical applications, this rate simplifies to the usual convergence rate $\sqrt{s_n \ln p_n / n}$. A similar result (with the same rate) was obtained by Loh and Wainwright (2015) under a *restricted strong convexity* (RSC) condition. The RSC condition makes assumptions about the realization of a sample, while Theorem 4 gives conditions on the population level. In the proof, we see that (A4) implies that an RSC-type condition holds with high probability. Specifically, in most standard cases we have

$$\langle \boldsymbol{u}, \Phi_n(\boldsymbol{\theta}^*) - \Phi_n(\boldsymbol{\theta}^* + \boldsymbol{u}) \rangle \geqslant c\|\boldsymbol{u}\|^2 - c_1\|\boldsymbol{u}\|_1^2 \eta_n,$$

for some constant $c_1 \geqslant 0$. A second difference is that we restrict the statement to cones $\Theta^*(\nu_n)$, while Loh and Wainwright (2015) allow for larger sets of the form $\{\|\boldsymbol{\theta}\|_1 \leqslant k_n\}$. They facilitate this by a stronger RSC condition of the form

$$\langle \boldsymbol{u}, \Phi_n(\boldsymbol{\theta}^*) - \Phi_n(\boldsymbol{\theta}^* + \boldsymbol{u}) \rangle \geqslant c\|\boldsymbol{u}\|^2 - c_1\|\boldsymbol{u}\|_1^2 \eta_n^2. \tag{8}$$

Under mild assumptions on the penalty and tuning parameter, this stronger condition implies that $\hat{\boldsymbol{\theta}} \in \Theta_n^*(\nu_n)$ with $\nu_n = O(s_n)$, as shown in Lemma 3 in the appendix. To the best of our knowledge, the stronger condition (8) has only been verified for variations of the linear model, where the fact that H_n is negative semi-definite and rank-1 can be exploited. Establishing this for general nonlinear problems appears much harder, if not entirely infeasible. The preemptive restriction to cones is a way to circumvent this issue with little practical consequence. Whenever a ν_n-sparse solution has been found, we also know that it belongs to $\Theta^*(\nu_n)$ and Theorem 4 applies.

Theorem 4 does not say anything about existence and selection consistency, which also explains why it does not require a lower bound on the tuning parameters. The next theorem shows that there is an estimation and selection consistent solution to the penalized equation with high probability. Specifically, we show that, with high probability, the reduced problem

$$\frac{1}{n}\sum_{i=1}^{n} \phi_i((\boldsymbol{\theta}_{(1)}, \mathbf{0}))_{(1)} \in \partial p_{\boldsymbol{\lambda}_n}((\boldsymbol{\theta}_{(1)}, \mathbf{0}))_{(1)}. \tag{9}$$

has a solution $\hat{\boldsymbol{\theta}}_{(1)}$ close to $\boldsymbol{\theta}^*_{(1)}$, and that $(\hat{\boldsymbol{\theta}}_{(1)}, \mathbf{0})$ is also a solution to the full problem (5).

Theorem 5. *Suppose the reduced problem* (9) *satisfies (A1) on sets* $\{\boldsymbol{\theta}_{(1)} : \boldsymbol{\theta} \in \Theta'\}$. *Suppose further that (P1), (P2) and (A5)–(A8) hold. Then, with probability tending to* 1, *the sets* Θ'_n *contain a solution* $\hat{\boldsymbol{\theta}}$ *of the penalized estimating equation (5) such that*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\sqrt{\frac{\operatorname{tr}(I(\boldsymbol{\theta}^*)_{(1)})}{n}} + \sqrt{s_n} b_n^*\right), \quad \hat{\boldsymbol{\theta}}_{(2)} = \mathbf{0}.$$

To simplify the discussion, suppose $\operatorname{tr}(I(\boldsymbol{\theta}^*)_{(1)}) = O(s_n)$, which is the most common situation. The rate of convergence depends on the number of non-zero coefficients s_n of $\boldsymbol{\theta}^*$ and the bias b_n^* induced by the penalty. For example, we have $b_n^* = \lambda_n$ for the Lasso, and $b_n^* = 0$ for SCAD for n large enough if $\lambda_n \to 0$. For a Lasso-type penalty with different $\lambda_{n,k}$, i.e., $p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}) = \sum_{k=1}^{p_n} \lambda_{n,k}|\theta_k|$, we obtain $b_n^* = \max_{k=1,\dots,s_n} \lambda_{n,k}$. The rate of convergence does not explicitly involve the overall number of parameters p_n, although it may enter through the bias b_n^* (typically logarithmically). Similarly, the regularity conditions above usually depend on p_n only logarithmically. This shows that estimation of $\boldsymbol{\theta}^*$ is possible even if the total number of parameters p_n is much larger than the sample size n. The non-asymptotic results of Negahban et al. (2012) and Loh and Wainwright (2015) imply a rate of $O_p(\sqrt{s_n \ln p_n / n})$. For the SCAD and similar penalties, this is still suboptimal as Theorem 5 gives the rate $O_p(\sqrt{s_n/n})$.

In practice, we do not know which parameters θ_k are 0, so it is unclear whether the solution from Theorem 5 can be found. Our final theorem gives conditions under which the solution is unique.

Theorem 6. *Suppose that the conditions of Theorem 4 are satisfied and (P3) holds with* $\limsup_{n\to\infty} \mu_n < 2c$, *where* c *is defined in (A4). Suppose further that for all* $\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{u} \in \Theta_n \cap \{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant a_n\sqrt{\nu_n}(\eta_n + \bar{b}_n)\}$ *with* $a_n \to \infty$ *arbitrarily slowly and all* $\boldsymbol{x} \in \mathcal{X}$, *it holds*

$$\boldsymbol{u}^\top[\phi(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{u}) - \phi(\boldsymbol{x}; \boldsymbol{\theta})] \leqslant \boldsymbol{u}^\top H_n(\boldsymbol{x})\boldsymbol{u}. \tag{10}$$

Then, with probability tending to 1, there is at most one solution in Θ_n.

Theorem 6 implies the uniqueness of any solution in Θ_n. If the conditions of Theorem 5 hold, the selection consistent estimator $\hat{\boldsymbol{\theta}}$ is unique on Θ_n with probability tending to 1.

3.3.3 Mutual incoherence & irrepresentable condition

The condition on $\boldsymbol{\lambda}_n$ in (A6) requires the existence of some $\alpha \in [0, 1)$ such that

$$\left\| \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} J(\boldsymbol{\theta})_{(2,1)}\ J^{-1}(\boldsymbol{\theta})_{(1)}\ p'_{\boldsymbol{\lambda}_n}(\tilde{\boldsymbol{\theta}})_{(1)} \right\|_\infty \leqslant \alpha$$

for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta'$. For penalties such as SCAD or MCP that are unbiased for large coefficients if $\lambda_n \to 0$, this condition is always satisfied with $\alpha = 0$ since $p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})_{(1)} = \mathbf{0}$. For the Lasso and a scalar λ_n, the terms involving λ_n cancel and the condition can be simplified to

$$\left\| J(\boldsymbol{\theta})_{(2,1)}\ J^{-1}(\boldsymbol{\theta})_{(1)} \right\|_\infty \leqslant \alpha. \tag{11}$$

This condition is independent of λ_n and therefore a characteristic of the population equation. The $(p_n - s_n) \times s_n$ matrix $J(\boldsymbol{\theta})_{(2,1)}$ is the expected gradient of $\phi(\boldsymbol{\theta})_{(2)}$ with respect to $\boldsymbol{\theta}_{(1)}$ and describes the average effect of errors in $\boldsymbol{\theta}_{(1)}$ on $\phi(\boldsymbol{\theta})_{(2)}$. The matrix $J^{-1}(\boldsymbol{\theta})_{(1)}$ can be seen a normalization term. Condition (A6) requires that this effect must not be too big.

For the linear model, condition (11) has an even nicer interpretation: In this case, one obtains $J(\boldsymbol{\theta})_{(2,1)} = \mathbb{E}[\boldsymbol{X}_{(2)}\boldsymbol{X}_{(1)}^\top]$ and $J(\boldsymbol{\theta})_{(1)} = \mathbb{E}[\boldsymbol{X}_{(1)}\boldsymbol{X}_{(1)}^\top]$. Assuming without loss of generality that $\mathbb{E}[\boldsymbol{X}] = 0$, these matrices correspond to covariances between the covariates. If the covariates are correlated too strongly, the Lasso cannot select the correct variables. This is a population version of the well-known *mutual incoherence* condition Wainwright (2019, Section 7.5.1), which requires some sort of approximate orthogonality. Zhao and Yu (2006) and Bühlmann and van de Geer (2011) obtain a similar condition under the name *irrepresentable condition*, as this means that the $X_j, j = s_n + 1, \dots, p_n$ are not represented “too well” by the covariates $\boldsymbol{X}_{(1)}$ of the true model. Zhao and Yu (2006) and Zou (2006) also show that the condition in Eq. (11) is necessary for variable selection consistency of the Lasso in the linear model. Similar to our results, Loh and Wainwright (2017) show that such incoherence conditions are not necessary for the SCAD, MCP, and similar non-convex penalties.

To better understand the condition, consider the Lasso. Assume an *iid* problem with two-dimensional parameter (θ_1, θ_2) with $\theta_1^* \neq 0$ and $\theta_2^* = 0$. The zero $\theta_2^{(0)}$ of $\theta_2 \mapsto \mathbb{E}[\phi((\theta_1^*, \theta_2))]$ is $\theta_2^* = 0$. However, if some $\hat{\theta}_1$ is plugged in instead of θ_1^*, then $\theta_2^{(0)}$ is not necessarily equal to $\theta_2^* = 0$. The more θ_1 affects $\phi(\boldsymbol{\theta})_2$, the further away $\theta_2^{(0)}$ is from 0. In general, biased estimation of $\boldsymbol{\theta}_{(1)}$ implies biased estimation of $\boldsymbol{\theta}_{(2)}$. If the influence of $\boldsymbol{\theta}_{(1)}$ on the estimation of $\boldsymbol{\theta}_{(2)}$ is too big, $\hat{\boldsymbol{\theta}}_{(2)}$ is too large to be shrunk to $\mathbf{0}$. Choosing a larger λ_n does not help, since this also increases the bias of $\hat{\boldsymbol{\theta}}_{(1)}$. If the estimation of $\boldsymbol{\theta}_{(1)}$ does not affect the estimation of $\boldsymbol{\theta}_{(2)}$, e.g., if the k-th entry of $\phi(\boldsymbol{\theta})$ only depends on θ_k for all k, then $J(\boldsymbol{\theta})_{(2,1)} = 0$ and $\alpha = 0$.

The definition of the penalty in Section 3.1 also allows for vectors $\boldsymbol{\lambda}_n = (\lambda_{n,1}, \dots, \lambda_{n,p_n})$, e.g., a weighted Lasso penalty $p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}) = \sum_{k=1}^{p_n} \lambda_{n,k}|\theta_k|$. Here, the condition characterizes a trade-off between the magnitudes of the entries of $\boldsymbol{\lambda}_n$, depending on how sensitive the equation is to errors in the respective parameters. Assuming $J(\boldsymbol{\theta})_{(1)} = I_{s_n}$ for simplicity, condition (A6) becomes

$$\frac{1}{\lambda_{n,j}} \sum_{k=1}^{s_n} \lambda_{n,k} \left| \mathbb{E}\left[\frac{\partial}{\partial \theta_k} \phi(\boldsymbol{\theta})_j \right] \right| \leqslant \alpha < 1 \quad \text{for all } j = s_n + 1, \dots, p_n.$$

Recall from above that we need this assumption because a biased estimation of $\boldsymbol{\theta}_{(1)}$ leads to a bias in $\hat{\boldsymbol{\theta}}_{(2)}$. If the effect of errors in $\theta_k, k \in \{1, \dots, s_n\}$ on the estimating equation for $\theta_j, j \in \{s_n + 1, \dots, p_n\}$ are large, we must either choose $\lambda_{n,k}$ small or $\lambda_{n,j}$ large. Since we do not know the true support of $\boldsymbol{\theta}^*$ in advance, tuning $\lambda_{n,k}$ differently depending on whether it belongs to $\boldsymbol{\lambda}_{n,(1)}$ or $\boldsymbol{\lambda}_{n(2)}$ is not practical. However, it is sometimes possible to gauge the magnitudes of $\mathbb{E}[\nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta})_j]$ and choose $\lambda_{n,j}$ large if this magnitude is high. One example is if $\phi_i(\boldsymbol{\theta})_j$ only depends on a subset of $\boldsymbol{\theta}$, as is often the case in step-wise estimation procedures.

3.4 Asymptotic normality and the oracle property

Finally, we establish asymptotic normality of the estimator $\hat{\boldsymbol{\theta}}_{(1)}$.

Theorem 7. *Suppose that (P4) holds and that for some matrix $A_n \in \mathbb{R}^{q \times s_n}$ with $\|A_n\| = O(1)$, the reduced problem (9) satisfies the conditions of Theorem 3 with $r_n = \sqrt{\operatorname{tr}(I(\boldsymbol{\theta}^*)_{(1)})/n} + \sqrt{s_n} b_n^*$. Then, the estimator $\hat{\boldsymbol{\theta}}_{(1)}$ in Theorem 5 is asymptotically normal with*

$$\sqrt{n} A_n \left[J(\boldsymbol{\theta}^*)_{(1)} (\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}) - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} \right] \to_d \mathcal{N}(0, \Sigma),$$

where $\Sigma = \lim_{n\to\infty} A_n I(\boldsymbol{\theta}^*)_{(1)} A_n^\top$.

Recall that $\hat{\boldsymbol{\theta}}$ has the oracle property if it behaves like the hypothetical *oracle estimator* that knows $\boldsymbol{\theta}^*_{(2)} = \mathbf{0}$ in advance. This is the case if $\hat{\boldsymbol{\theta}}_{(2)} = \mathbf{0}$ and $p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} = o(1/\sqrt{n})$. For SCAD and MCP with $\lambda_n \to 0$, one obtains $b_n^* = 0$, so the rate of convergence is the same as if $\boldsymbol{\theta}^*_{(2)} = \mathbf{0}$ is known in advance. Since $p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} = \mathbf{0}$, the SCAD and MCP-penalized estimators have the same efficiency as the oracle estimator. As long as $\boldsymbol{\lambda}_n$ is small enough that $p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} = \mathbf{0}$ asymptotically, it can be chosen large enough such that (A8) is fulfilled.

For the Lasso, we have $b_n^* = \lambda_n$ and $p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} = O(\sqrt{s_n}\lambda_n)$. For the oracle property to hold, we would need $\lambda_n = o(1/\sqrt{s_n n})$. This condition cannot be satisfied, because (A6) requires $\lambda_n \geqslant \eta_n \approx \sqrt{\ln p_n/n}$. This is in line with the results from Zou (2006), who shows that in the linear model, the Lasso is only variable selection consistent at the cost of a slower rate of convergence.

4 Applications

Our general results are stated under rather abstract regularity conditions to keep them widely applicable. In the following, we provide several examples of how these conditions simplify in specific applications.

4.1 M-Estimation of generalized linear models

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be *iid* and consider an estimator $\hat{\boldsymbol{\theta}}$ that satisfies

$$\sum_{i=1}^n \phi_i(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \psi(Y_i, \boldsymbol{X}_i^\top \hat{\boldsymbol{\theta}}) \boldsymbol{X}_i = \mathbf{0} \tag{12}$$

with some function ψ. This includes the least squares estimator and other M-estimators in the linear model studied by Huber (1973), Portnoy (1984, 1985) and Mammen (1989) as well as likelihood inference in generalized linear models as special cases. If the function ψ is smooth with nonpositive derivative, the matrix H_n in (A1) can be constructed as

$$H_n(\boldsymbol{x}) = -\inf_{\boldsymbol{\theta}\in\Theta_n} |\psi'(Y_i, \boldsymbol{x}^\top\boldsymbol{\theta})| \boldsymbol{x}\boldsymbol{x}^\top,$$

where $\psi'(Y_i, \eta) = \frac{\partial}{\partial\eta}\psi(Y_i, \eta)$.

Corollary 1. *Let $\hat{\boldsymbol{\theta}}$ solve (12) and let ψ' be nonpositive, Lipschitz in η and uniformly bounded. Suppose that $\max_k \mathbb{E}[\phi_i(\boldsymbol{\theta}^*)_k^4] = O(1)$ and for all $\boldsymbol{a} \in \mathbb{R}^{p_n}$ with $\|\boldsymbol{a}\| = O(1)$,*

$$\begin{aligned} &\|\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^\top]\| = O(1), \quad \mathbb{E}[|\boldsymbol{a}^\top\boldsymbol{X}_i|^4] = O(\|\boldsymbol{a}\|^4), \quad \mathbb{E}[\rho(|\boldsymbol{a}^\top\boldsymbol{X}_i|^2)] = O(1) \\ &\operatorname{Var}[\|\boldsymbol{X}_i\|^2] = O(p_n), \quad \mathbb{E}[\rho(|\|\boldsymbol{X}_i\| - \mathbb{E}[\|\boldsymbol{X}_i\|]|^2)] = O(1) \end{aligned} \tag{13}$$

with some increasing and strictly convex function $\rho\colon (0,\infty) \to (0,\infty)$. Suppose further that

$$\lambda_{\min}\left(\mathbb{E}\left[\inf_{\boldsymbol{\theta}\in\Theta_n} |\psi'(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\theta})| \boldsymbol{X}_i \boldsymbol{X}_i^\top\right]\right) \geqslant c \tag{14}$$

holds with some $c > 0$ and sets $\Theta_n \supset \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant r_n C\}$ for all $C < \infty$ and n large.

(i) If $p_n \ln p_n / n \to 0$, $\hat{\boldsymbol{\theta}}$ is a $\sqrt{n/p_n}$-consistent estimator of $\boldsymbol{\theta}^*$ and unique on Θ_n with probability tending to 1.

(ii) If the matrix A_n in Theorem 3 satisfies $\|A_n\| = O(1)$, and $p_n^2 \rho^{-1}(n)/n \to 0$, $\hat{\boldsymbol{\theta}}$ is asymptotically normal.

The design conditions (13) are relatively mild and only the first and fourth are required for consistency. The second requires a weak form of isotropy. The third is a tail condition. The fourth and fifth require some form of concentration. All conditions are easily satisfied for, e.g., independent sub-Gaussian variables, for which $\rho(x) = \exp(x)$ works, and the growth bound for asymptotic normality becomes $p_n^2 \ln n/n \to 0$. In contrast to the results of Portnoy (1985), the corollary also applies to covariates with heavier tails. For example, taking $\rho(x) = x^2$ (which gives a fourth moment condition), the estimator is asymptotically normal as long as $p_n^4/n \to 0$. For the least squares estimator, ψ' is constant, so the eigenvalue condition (14) simplifies to $\lambda_{\min}\left(\mathbb{E}\left[\boldsymbol{X}_i \boldsymbol{X}_i^\top\right]\right) \geqslant c$. This simplification applies more generally if ψ' is uniformly bounded away from zero.

Now we turn to penalized estimation. We consider the Group Lasso penalty from Example 3.3, noting that this includes the usual Lasso a special case. Now, $\boldsymbol{\theta}_{(2)}$ consists of all groups G_k with $\boldsymbol{\theta}^*_{G_k} = \mathbf{0}$.

Corollary 2. *Suppose the reduced problem (9) satisfies the conditions from Corollary 1 with some function ρ with $\rho(x) = O(\exp(x))$, and*

$$\max_{1\leqslant k\leqslant p_n} \mathbb{E}[\rho(X_{i,k}^2 \psi(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\theta}^*)^2)] = O(1). \tag{15}$$

Let

$$\sup_{\boldsymbol{\theta}\in\Theta'} \left\| \mathbb{E}[\psi'(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\theta}) \boldsymbol{X}_{(2)} \boldsymbol{X}_{(1)}^\top]\, \mathbb{E}[\psi'(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\theta}) \boldsymbol{X}_{(1)} \boldsymbol{X}_{(1)}^\top]^{-1} \right\|_\infty \leqslant \alpha \tag{16}$$

for some $\alpha \in [0,1)$, suppose that $\sigma^2 = \max_{1\leqslant k\leqslant p_n} \mathbb{E}[\phi_i(\boldsymbol{\theta}^*)_k^2]$ is bounded away from zero and infinity, either $\boldsymbol{\theta}^*_{G_i} = \mathbf{0}$ or $\|\boldsymbol{\theta}^*_{G_i}\|/r_n \to \infty$ with $r_n = \sqrt{s_n \ln p_n / n}$ and

$$\lambda_n \geqslant \frac{8}{1-\alpha}\sqrt{\frac{\sigma^2 \ln p_n}{n}}.$$

Then, if

$$s_n \ln p_n = o\left(\sqrt{n}\right), \qquad p_n = o\left(\frac{\rho(n/(s_n \ln p_n)^2)}{n}\right), \tag{17}$$

the Group-Lasso-penalized equation has a solution $\hat{\boldsymbol{\theta}}$ with $\hat{\boldsymbol{\theta}}_{(2)} = \mathbf{0}$ with probability tending to 1, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p(\sqrt{s_n \ln p_n / n})$, and $\hat{\boldsymbol{\theta}}_{(1)}$ is asymptotically normal (with $\|A_n\| = O(1)$). If $|\psi'|$ is uniformly bounded away from 0 and $\lambda_{\min}\left(\mathbb{E}\left[\boldsymbol{X}_i \boldsymbol{X}_i^\top\right]\right) \geqslant c > 0$, this solution is unique with probability tending to 1.

The moment conditions in (13) and (15) constrain the growth of p_n through the function ρ. Assuming $s_n = O(1)$ for simplicity, the choice $\rho(x) = \exp(x)$ allows $p_n \sim \exp(n^{1/2-\varepsilon})$ for any $\varepsilon > 0$. Similarly, polynomial moment bounds translate into polynomial growth conditions on p_n. For example, the choice $\rho(x) = x^3$ requires sixth moments and allows $p_n \sim n^{2-\varepsilon}$.

The two corollaries also apply to non-parametric regression problems, in which $\boldsymbol{X}_i$ consists of appropriate basis functions. For example, Corollary 2 implies consistency of Group-Lasso-assisted variable selection in high-dimensional nonparametric additive models (Huang et al., 2010) under appropriate conditions.

4.2 Multi-sample estimation

Consider a multi-sample estimation problem: The data are given by $(k_1, \boldsymbol{X}_1), \dots, (k_n, \boldsymbol{X}_n)$, where $k_i \in \{1, \dots, K_n\}$ indicates to which of the K_n samples $\boldsymbol{X}_i$ belongs. Assume for simplicity that $p_n = K_n$ and that each θ_k is estimated using only the k-th sample. The estimation function can be written as

$$\phi(k_i, \boldsymbol{X}_i; \boldsymbol{\theta}) = \begin{pmatrix} \mathbb{1}\{k_i = 1\} \, \frac{n}{n_1} \phi_1(\boldsymbol{X}_i; \theta_1) \\ \vdots \\ \mathbb{1}\{k_i = K_n\} \, \frac{n}{n_{K_n}} \phi_{K_n}(\boldsymbol{X}_i; \theta_{K_n}) \end{pmatrix},$$

where $n_k = \sum_{i=1}^n \mathbb{1}\{k_i = k\}$ denotes the sample size of the k-th sample. The standardization n/n_k is necessary to ensure that the eigenvalues of $n^{-1} \sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)]$ are bounded away from 0. The Jacobian of ϕ is a diagonal matrix given by

$$\nabla_{\boldsymbol{\theta}} \phi(k_i, \boldsymbol{X}_i; \boldsymbol{\theta}) = \operatorname{diag}\left(\mathbb{1}\{k_i = k\} \, \frac{n}{n_k} \phi_k'(\boldsymbol{X}_i; \theta_k) \right)_{k=1,\dots,K_n},$$

where $\phi_k'(\boldsymbol{X}; \theta_k) = \frac{\partial}{\partial \theta_k} \phi_k(\boldsymbol{X}; \theta_k)$. Straightforward computations give

$$I(\boldsymbol{\theta}^*) = \operatorname{diag}\left(\frac{n}{n_k} \mathbb{E}[\phi_k(\boldsymbol{X}_i; \theta_k^*)^2] \right)_{k=1,\dots,K_n},$$

which yields $r_n = O((\sum_{k=1}^{K_n} n_k^{-1})^{1/2})$ assuming that $\mathbb{E}[\phi_k(\boldsymbol{X}_i; \theta_k^*)^2]$ is bounded for each n and k. This framework can easily be extended to multiple parameters $\boldsymbol{\theta}_k \in \mathbb{R}^{p_k}$, potentially shared across subsamples. In the following examples, we stick to the single-parameter case for simplicity.

Remark 8. *There is an alternative framework for modelling multi-sample problems. One could model the data as* iid *draws from a mixture distribution with K_n components, where the k-th component has weight $\pi_k = n_k/n$. The main difference in this approach is that the sizes of the sub-samples are random, but it otherwise leads to essentially the same results.*

4.2.1 Example: Distributed inference

An interesting application arises in distributed inference. Here, *iid* data is distributed over K_n different locations, and the goal is to estimate a parameter $\theta^* \in \mathbb{R}$ from the distributed data. This is a common setup in federated learning, where data is distributed over different devices, and the goal is to estimate a common model. Our setup further allows for differing sample sizes and population characteristics between locations. This may happen if, for example, the data is collected over hospitals who may share their estimate but not the data for privacy reasons. The distributed estimates $\hat{\theta}_k$ can be reconciled into a global estimate through averaging

$\hat{\theta}_{K_n+1} = K_n^{-1} \sum_{k=1}^{K_n} \hat{\theta}_k$. To put this in our framework, we stack the individual estimating equations $\phi_k(\boldsymbol{X}_i; \theta_k) = \psi(\boldsymbol{X}_i; \theta_k)$ as above, and append the reconciliation function

$$\phi_{K_n+1}(\boldsymbol{X}_i; \theta) = \frac{1}{K_n} \sum_{k=1}^{K_n} \theta_k - \theta_{K_n+1}.$$

Corollary 3. *Let $\Theta_0 \subseteq \mathbb{R}$ and $n_1 = \cdots = n_{K_n} = n/K_n$. Suppose that, for $\theta \in \Theta_0$, $\psi(\boldsymbol{x}; \theta)$ is uniformly bounded and $\psi'(\boldsymbol{x}; \theta) = \partial_\theta \psi(\boldsymbol{x}; \theta)$ is negative, uniformly bounded away from 0 and $-\infty$ and Lipschitz in θ. Then, if $K_n^3/n \to 0$, it holds*

$$\begin{aligned} &\sqrt{n/K_n}(\hat{\theta}_k - \theta_k^*) \to N(0, \sigma_k^2), \quad k = 1, \dots, K_n, \\ \textit{and} \quad &\sqrt{n}(\hat{\theta}_{K_n+1} - \theta^*) \to N(0, \sigma^2), \end{aligned}$$

with

$$\sigma_k^2 = \frac{\mathbb{E}[\psi(\boldsymbol{X}_i; \theta_k^*)^2]}{\mathbb{E}[\psi'(\boldsymbol{X}_i; \theta_k^*)]^2}, \quad \sigma^2 = \lim_{n \to \infty} \frac{1}{K_n} \sum_{k=1}^{K_n} \sigma_k^2.$$

If the samples have identical distributions, $\sigma_1 = \cdots = \sigma_{K_n}$, the averaged estimator is as efficient as one computed from the pooled sample. If the samples have different distributions, however, there might be a loss in efficiency.

4.2.2 Example: Quality control

To give an example where our penalized results are useful, consider the following statistical quality control problem. We have K_n machines producing items, and we want to ensure that the items maintain a prescribed quality level. The quality of an item i produced by machine k_i is given by $q_{k_i}(\boldsymbol{X}_i)$. A machine is considered OK if

$$\mathbb{E}[q_k(\boldsymbol{X}_i) \mid k_i = k] = a,$$

where a is the targeted quality. To detect potentially faulty machines, define parameters θ_k^* such that

$$\theta_k^* = \mathbb{E}[q_k(\boldsymbol{X}_i) \mid k_i = k] - a,$$

which can be estimated using $\phi_k(\boldsymbol{X}_i; \theta_k) = q_k(\boldsymbol{X}_i) - a - \theta_k$ in the above setup. To only detect faulty machines, we can penalize the equation using, e.g., the Lasso penalty $p_{\lambda_n}(\boldsymbol{\theta}) = \lambda_n \sum_{k=1}^{K_n} |\theta_k|$. The following corollary guarantees that the penalized estimator $\hat{\boldsymbol{\theta}}$ detects all faulty machines with probability tending to 1.

Corollary 4. *In the above quality control example, suppose that (A5) holds and*

$$\lambda_n \geqslant 8 \max_{1 \leqslant k \leqslant K_n} \max_{1 \leqslant i \leqslant n} \sqrt{\frac{\mathbb{E}[(q_k(\boldsymbol{X}_i) - a - \theta_k^*)^2] \ln K_n}{n_k}}.$$

Suppose further that $s_n^2 \ln K_n / \min_k n_k \to 0$ and either $\theta_k^* = 0$ or $\theta_k^*/r_n \to \infty$ with $r_n = \sqrt{s_n} \lambda_n$ for all k. Then

$$\mathbb{P}\left(\{k \colon \hat{\theta}_k \neq 0\} = \{k \colon \theta_k^* \neq 0\}\right) \to 1.$$

Note that we do not require the samples from the machines to be identically distributed, so the result also applies to situations where different types of machines are used or the machine only fails after some time.

4.3 Stepwise estimation

Another setting that shows the flexibility of our results is stepwise estimation. Assume the parameter vector can be grouped as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K_n})$. The parameters are estimated sequentially using the estimates $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{k-1}$ from previous iterations:

$$\hat{\boldsymbol{\theta}}_k = \arg\max_{\theta_k} \sum_{i=1}^n f_k(\boldsymbol{X}_i; \boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{k-1}), \quad \boldsymbol{\theta}_k^* = \arg\max_{\boldsymbol{\theta}_k} \mathbb{E}\left[f_k(\boldsymbol{X}; \boldsymbol{\theta}_k, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k-1}^*)\right],$$

for some functions f_k. Denote $\phi_k(\boldsymbol{X}_i; \boldsymbol{\theta}_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}) = \nabla_{\boldsymbol{\theta}_k} f_k(\boldsymbol{X}_i; \boldsymbol{\theta}_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1})$. Then, the sequential estimator $\hat{\boldsymbol{\theta}}$ can be expressed as the solution of $\sum_{i=1}^n \phi(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \mathbf{0}$ with

$$\phi(\boldsymbol{X}_i; \boldsymbol{\theta}) = \begin{pmatrix} \phi_1(\boldsymbol{X}_i; \boldsymbol{\theta}_1) \\ \phi_2(\boldsymbol{X}_i; \boldsymbol{\theta}_2, \boldsymbol{\theta}_1) \\ \vdots \\ \phi_{K_n}(\boldsymbol{X}_i; \boldsymbol{\theta}_{K_n}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K_n - 1}) \end{pmatrix},$$

and $\boldsymbol{\theta}^*$ is the solution of $\mathbb{E}[\phi(\boldsymbol{X}; \boldsymbol{\theta}^*)] = \mathbf{0}$.

4.3.1 Example: Causal inference

As a concrete example, suppose we want to estimate the causal effect of some covariates $\boldsymbol{Z}$ on an outcome Y in the presence of confounders $\boldsymbol{C}$ from *iid* observational data. Part of the population has received a treatment, which we indicate by the binary treatment indicator T. Under the usual conditions for no unmeasured confounding, the *conditional average treatment effect (CATE)* can be defined as

$$\text{CATE}(\boldsymbol{z}) = \mathbb{E}\left[\frac{YT}{\mathbb{P}(T = 1 \mid \boldsymbol{W})} - \frac{Y(1-T)}{\mathbb{P}(T = 0 \mid \boldsymbol{W})} \mid \boldsymbol{Z} = \boldsymbol{z}\right],$$

where $\boldsymbol{W} = (\boldsymbol{Z}, \boldsymbol{C})$; see for example Huber (2023). Now we model the treatment probabilities and CATE by

$$\mathbb{P}(T = 1 \mid \boldsymbol{w}) = \sigma(\boldsymbol{w}^\top \boldsymbol{\theta}_1), \quad \text{CATE}(\boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{\theta}_2,$$

where σ is an appropriate link function. The parameters can be estimated by first estimating $\boldsymbol{\theta}_1$ using maximum-likelihood, and then estimating $\boldsymbol{\theta}_2$ by the plug-in least squares estimator

$$\hat{\boldsymbol{\theta}}_2 = \arg\min_{\boldsymbol{\theta}_2} \sum_{i=1}^n \left[\frac{Y_i T_i}{\sigma(\boldsymbol{W}_i^\top \hat{\boldsymbol{\theta}}_1)} - \frac{Y_i(1-T_i)}{1 - \sigma(\boldsymbol{W}_i^\top \hat{\boldsymbol{\theta}}_1)} - \boldsymbol{Z}_i^\top \boldsymbol{\theta}_2\right]^2.$$

This step-wise procedure can be reformulated as solving the estimating equation

$$\sum_{i=1}^n \phi(Y_i, T_i, \boldsymbol{Z}_i, \boldsymbol{W}_i; \hat{\boldsymbol{\theta}}) = \mathbf{0}$$

with

$$\phi(Y_i, T_i, \boldsymbol{Z}_i, \boldsymbol{W}_i; \boldsymbol{\theta}) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}_1}\left[T_i(\ln \sigma(\boldsymbol{W}_i^\top \boldsymbol{\theta}_1) + (1-T_i)\ln[1 - \sigma(\boldsymbol{W}_i^\top \boldsymbol{\theta}_1)]\right] \\ -\left[\frac{Y_i T_i}{\sigma(\boldsymbol{W}_i^\top \boldsymbol{\theta}_1)} - \frac{Y_i(1-T_i)}{1-\sigma(\boldsymbol{W}_i^\top \boldsymbol{\theta}_1)} - \boldsymbol{Z}_i^\top \boldsymbol{\theta}_2\right] \boldsymbol{Z}_i \end{pmatrix}.$$

Corollary 5. *Suppose that σ is bounded away from zero and 1 and twice continuously differentiable with uniformly bounded derivatives. Suppose further that $|Y| \leqslant 1$ and $\boldsymbol{W} \in \mathbb{R}^{p_n}$ satisfies the design conditions from* (13)*. Let $\overline{\sigma} = 1 - \sigma$ and define*

$$\begin{aligned}
\alpha_1(T, \boldsymbol{W}) &= \sup_{\boldsymbol{\theta} \in \Theta_n} \left[T(\ln \sigma)''(\boldsymbol{W}^\top \boldsymbol{\theta}_1) + (1-T)(\ln \overline{\sigma})''(\boldsymbol{W}^\top \boldsymbol{\theta}_1) \right], \\
\alpha_2(T, Y, \boldsymbol{W}) &= \sup_{\boldsymbol{\theta} \in \Theta_n} \left| \frac{TY\sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{2\sigma(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2} \right| + \left| \frac{(1-T)Y\sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{2\overline{\sigma}(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2} \right| - 1,
\end{aligned}$$

and suppose there is $c > 0$ such that

$$\lambda_{\min}\left(\mathbb{E}\left[\alpha_1(T, \boldsymbol{W})\boldsymbol{W}\boldsymbol{W}^\top\right]\right) \leqslant -c, \qquad \lambda_{\min}\left(\mathbb{E}\left[\alpha_2(T, Y, \boldsymbol{W})\boldsymbol{Z}\boldsymbol{Z}^\top\right]\right) \leqslant -c.$$

Then, if $p_n \ln p_n / n \to 0$, the estimating equation has a unique solution $\hat{\boldsymbol{\theta}}$ on Θ_n with

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\sqrt{\frac{p_n}{n}}\right).$$

If $p_n^2 \rho^{-1}(n)/n \to 0$ (with ρ as defined in (13)*), $\hat{\boldsymbol{\theta}}$ is also asymptotically normal.*

The matrices inside the expectations of the eigenvalue condition are the blocks of a block-diagonal matrix H_n constructed in the proof. The condition is easiest to verify if $\ln \sigma$ is concave, which is the case for the logistic and probit link functions.

If the number of covariates or confounders is large, we may want to add a sparsity penalty $p_{\lambda_n}(\boldsymbol{\theta})$. For simplicity, suppose that the parameters are reordered such that $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_{(1)}, \boldsymbol{0})$. The following corollary guarantees that the SCAD-penalized estimator $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal.

Corollary 6. *Suppose that the regularity conditions from Corollary 5 hold, $\boldsymbol{W}$ satisfies the design conditions* (15) *(with $|\psi|$ uniformly bounded), $\sqrt{n/s_n} \min_{1 \leqslant k \leqslant s_n} |\theta_k^*| \to \infty$, and that the SCAD penalty is used with $a > 1 + \frac{1}{2c}$ and*

$$\lambda_n \geqslant 8\sigma_n \sqrt{\frac{\ln p_n}{n}}.$$

Suppose that $s_n \ln p_n = o(\sqrt{n})$ and $p_n = o(\rho(n/(s_n \ln p_n)^2)/n)$. Then, with probability tending to 1, the penalized equation has a unique solution $\hat{\boldsymbol{\theta}}$ on Θ_n with $\hat{\boldsymbol{\theta}}_{(2)} = \boldsymbol{0}$ and $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p(\sqrt{s_n/n})$. Additionally, $\hat{\boldsymbol{\theta}}_{(1)}$ is asymptotically normal and as efficient as the oracle solution.

4.3.2 Example: Stochastic optimization

Our general results also apply to cases where $K_n \to \infty$. Such situations arise frequently in the analysis of iterative procedures, such as gradient descent or boosting algorithms. To illustrate this, suppose we want to learn a parameter $\theta^*_\infty \in \mathbb{R}$ solving

$$\mathbb{E}[f(\boldsymbol{X}; \theta^*_\infty)] = 0.$$

Let $\theta^*_0 \in \mathbb{R}$ be an initial value and define the iterative solutions

$$\theta^*_k = \theta^*_{k-1} - \alpha \mathbb{E}[f(\boldsymbol{X}; \theta^*_{k-1})].$$

Under appropriate conditions on the learning rate α and smoothness of f, the sequence θ_k^* can be shown to converge geometrically fast to θ_∞^* as $k \to \infty$. Now define $\hat{\theta}_0 = \theta_0^*$ and $\hat{\theta}_k, 1 \leqslant k \leqslant K_n$, as the solutions of the batched sample equation

$$\hat{\theta}_k = \hat{\theta}_{k-1} - \alpha \frac{K_n}{n} \sum_{i \in \mathcal{B}_k} f(\boldsymbol{X}_i; \hat{\theta}_{k-1}),$$

where $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are *iid* samples from the distribution of $\boldsymbol{X}$, and $\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_{K_n} = \{1, \dots, n\}$ with $|\mathcal{B}_k| = n/K_n$ is a partition of the sample indices. We can define the entire iteration path $\hat{\boldsymbol{\theta}}$ as solution of a single estimating equation $n^{-1} \sum_{i=1}^n \phi(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \mathbf{0}$ with

$$\phi(\boldsymbol{X}; \boldsymbol{\theta}) = \begin{pmatrix} K_n \mathbb{1}_{i \in \mathcal{B}_1} [\theta_0^* - \theta_1 - \alpha f(\boldsymbol{X}_i; \theta_0^*)] \\ \vdots \\ K_n \mathbb{1}_{i \in \mathcal{B}_{K_n}} [\theta_{K_n - 1} - \theta_{K_n} - \alpha f(\boldsymbol{X}_i; \theta_{K_n - 1})] \end{pmatrix}, \tag{18}$$

and similarly for the population version. The following is a possible result under simple conditions.

Corollary 7. *Let $f'(\boldsymbol{x}; \theta) = \partial_\theta f(\boldsymbol{x}; \theta)$. Suppose that $K_n^3/n \to 0$, $\|\boldsymbol{\theta}^*\|_\infty = O(1)$, $\sup_{\theta \in \Theta_n} \mathbb{E}[f(\boldsymbol{X}; \theta)^4] = O(1)$, $f' \in [\kappa, L]$ and $|f''| \leqslant L$ for some $\kappa, L \in (0, \infty)$, and that $0 < \alpha \leqslant 1/L$. Then*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\frac{K_n}{\sqrt{n}}\right),$$

and $\hat{\boldsymbol{\theta}}$ is unique with probability tending to 1, and for any $A_n \in \mathbb{R}^{q \times K_n}$ with $\|A_n\| = O(1)$, we have

$$\sqrt{n/K_n} A_n (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to_d \mathcal{N}\left(0, \lim_{n \to \infty} A_n \Sigma_n A_n^\top\right),$$

where Σ_n is symmetric and, with the convention $\prod_{j=i}^{i-1} a_j = 1$, it holds for $i \leqslant j$,

$$\Sigma_{i,j} = \alpha^2 \sum_{k=1}^{i} \mathbb{V}\mathrm{ar}[f(\boldsymbol{X}; \theta_{k-1}^*)] \left[\prod_{m=k}^{i-1} (1 - \alpha \mathbb{E}[f'(\boldsymbol{X}; \theta_m^*)])\right]^2 \left[\prod_{m=i}^{j-1} (1 - \alpha \mathbb{E}[f'(\boldsymbol{X}; \theta_m^*)])\right].$$

The corollary couples the iterates on the sample equation to the iterates on the population equation. The first part shows that the iteration paths are globally close to one another. The second part shows that finite-dimensional linear summaries of the solution path converge to a Gaussian limit. Several interesting special cases arise from particular choices of the sequence A_n. For example, $A_n = (0, \dots, 0, 1)$ implies $\sqrt{n/K_n}$-convergence of the final iterate and $A_n = (1/\sqrt{K_n}, \dots, 1/\sqrt{K_n})$ gives $\sqrt{n}$-convergence of the averaged iterate. Another interesting choice is $A_n = (\boldsymbol{e}_{\lceil K_n/q \rceil}, \boldsymbol{e}_{\lceil 2K_n/q \rceil}, \dots, \boldsymbol{e}_{\lceil qK_n/q \rceil})^\top$, where $\boldsymbol{e}_k$ is the kth standard unit vector. This corresponds to a discretized approximation of the process $\hat{\theta}(t) = \hat{\theta}_{\lceil tK_n \rceil}$, $t \in [0, 1]$. One can verify that for $n, K_n \to \infty$,

$$\sqrt{n/K_n} A_n (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to_d \mathcal{N}(0, V),$$

with diagonal matrix V. This suggests that the solution path $\hat{\theta}(t)$ behaves like a white noise process around the population path $\theta^*(t)$. The assumptions of the corollary can be relaxed in various ways using more sophisticated assumptions and arguments in the proof. For example, we may let the learning rate α decay slowly or impose only probabilistic bounds on f'. Other iterative algorithms and multivariate versions can be handled similarly.

References

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

Clarke, F. H. (1990). *Optimization and nonsmooth analysis*. SIAM.

Fan, J. (1997). Comments on "Wavelets in statistics: A review" by A. Antoniadis. *Journal of the Italian Statistical Society*, 6(2):131.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148. PMC3092303.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928 – 961.

Fierro, R., Yañez, C., and Morales, C. (2004). The aftermath of the intermediate value theorem. *Fixed Point Theory and Applications*, 2004.

Frank, l. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282 – 2313.

Huber, M. (2023). *Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R*. MIT Press.

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799 – 821.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, 28(5):1356 – 1378.

Lee, J. D., Sun, Y., and Taylor, J. E. (2013). On model selection consistency of penalized M-estimators: a geometric theory. *Advances in Neural Information Processing Systems*, 26.

Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized m-estimation with a diverging number of parameters. *Statistica Sinica*, 21(1):391–419.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45(2):866–896.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19):559–616.

Loh, P.-L. and Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455 – 2482.

Mammen, E. (1989). Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap. *The Annals of Statistics*, 17(1):382 – 400.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4).

Portnoy, S. (1984). Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *The Annals of Statistics*, 12(4):1298 – 1309.

Portnoy, S. (1985). Asymptotic Behavior of M Estimators of p Regression Parameters when p^2/n is Large; II. Normal Approximation. *The Annals of Statistics*, 13(4):1403 – 1417.

Portnoy, S. (1986). On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$. *Probability theory and related fields*, 73(4):571–583.

Talagrand, M. (2005). *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Van der Vaart, A. and Wellner, J. A. (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Nature.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Welsh, A. H. (1989). On M-Processes and M-Estimation. *The Annals of Statistics*, 17(1):337 – 361.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic Behavior of M-Estimators for the Linear Model. *The Annals of Statistics*, 7(2):258 – 268.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90):2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

A Proofs of Theorems

To simplify the notation in the following proofs and results, we shall use the following notation:

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i), \qquad Pf = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(\boldsymbol{X}_i)].$$

A.1 Proof of Theorem 1

We first show that the sets Θ_n contain a solution of the estimating equation (3) with probability tending to 1. From an extension of the intermediate value theorem in Fierro et al. (2004, Theorem 2.3), it follows that if

$$\sup_{\|\boldsymbol{u}\|=1} \langle r_n C\boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})\rangle \leqslant 0$$

holds, there is a solution $\hat{\boldsymbol{\theta}}$ of $\mathbb{P}_n\phi(\boldsymbol{\theta}) = \boldsymbol{0}$ that satisfies $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leqslant r_n C$. We show that by choosing C large enough, the probability that

$$(r_nC)^{-1} \sup_{\|\boldsymbol{u}\|=1} \langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})\rangle \leqslant -c < 0 \tag{19}$$

holds for some $c > 0$ becomes arbitrarily close to 1. We have

$$\begin{aligned} &(r_nC)^{-1} \sup_{\|\boldsymbol{u}\|=1} \langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})\rangle \\ \leqslant\ &(r_nC)^{-1} \sup_{\|\boldsymbol{u}\|=1} \langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^*)\rangle + (r_nC)^{-1} \sup_{\|\boldsymbol{u}\|=1} \langle \boldsymbol{u}, \mathbb{P}_n[\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]\rangle. \end{aligned}$$

The first term is of order $O_p(1/C)$, since

$$\sup_{\|\boldsymbol{u}\|=1} |\langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^*)\rangle| = \|\mathbb{P}_n\phi(\boldsymbol{\theta}^*)\| = O_p(n^{-1/2}\sqrt{\operatorname{tr}(I(\boldsymbol{\theta}^*))}) = O_p(r_n)$$

by Lemma 5. Choosing C large enough, it suffices that the second term remains below some $-c$ with probability going to 1. It holds

$$(r_nC)^{-1}\langle \boldsymbol{u}, \mathbb{P}_n[\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]\rangle \leqslant \mathbb{P}_n[\boldsymbol{u}^\top H_n \boldsymbol{u}] = P[\boldsymbol{u}^\top H_n \boldsymbol{u}] + (\mathbb{P}_n - P)[\boldsymbol{u}^\top H_n \boldsymbol{u}],$$

with H_n as defined in assumption (A1)(i). Hence,

$$(r_nC)^{-1} \sup_{\|\boldsymbol{u}\|=1} \langle \boldsymbol{u}, \mathbb{P}_n[\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]\rangle \leqslant \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[H_n(\boldsymbol{X}_i)]\right) + \|(\mathbb{P}_n - P)H_n\|.$$

By assumption (A1)(ii), we have $\lambda_{\max}(n^{-1}\sum_{i=1}^{n}\mathbb{E}[H_n(\boldsymbol{X}_i)]) \leqslant -c$ for some $c > 0$ and large enough n, and Lemma 7 gives $\|(\mathbb{P}_n - P)H_n\| = o_p(1)$. This proves (19).

We now show that every solution in Θ_n must satisfy $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leqslant Cr_n$ for some $C < \infty$, with probability tending to 1. Suppose this is not the case and define $C_n^{-1} = r_n/\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = o_p(1)$. Then, we can write $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + r_n C_n \hat{\boldsymbol{u}}$ with $\|\hat{\boldsymbol{u}}\| = 1$. It holds

$$\begin{aligned} 0 &= (r_nC_n)^{-1}\langle \hat{\boldsymbol{u}}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + r_n C_n\hat{\boldsymbol{u}})\rangle \leqslant \sup_{\|\boldsymbol{u}\|=1} (r_nC_n)^{-1}\langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + r_n C_n\boldsymbol{u})\rangle \\ &\leqslant O_p(C_n^{-1}) - c + o_p(1) = -c + o_p(1), \end{aligned}$$

where the second inequality follows from the above arguments. Hence, $\hat{\boldsymbol{\theta}}$ cannot be a solution with probability tending to 1.

A.2 Proof of Theorem 2

The claim is trivial when no solution exists. Otherwise, let $\hat{\boldsymbol{\theta}}$ be any two solution to the estimating equation (1). By Theorem 1, we may assume that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leqslant r_n C$ for some $C \in (0, \infty)$ and n large enough. Suppose there is another solution $\hat{\boldsymbol{\theta}} + \boldsymbol{u}$. By the strengthened assumption (4),

$$\begin{aligned}\langle \boldsymbol{u}, \mathbb{P}_n \phi(\hat{\boldsymbol{\theta}} + \boldsymbol{u})\rangle = \langle \boldsymbol{u}, \mathbb{P}_n \phi(\hat{\boldsymbol{\theta}} + \boldsymbol{u})\rangle - \langle \boldsymbol{u}, \mathbb{P}_n \phi(\hat{\boldsymbol{\theta}})\rangle &\leqslant \mathbb{P}_n[\boldsymbol{u}^\top H_n \boldsymbol{u}] \\ &= P[\boldsymbol{u}^\top H_n \boldsymbol{u}] + (\mathbb{P}_n - P)[\boldsymbol{u}^\top H_n \boldsymbol{u}] \\ &\leqslant \|\boldsymbol{u}\|^2(-c + o_p(1)),\end{aligned}$$

uniformly on the set $\{\boldsymbol{u} \colon \|\hat{\boldsymbol{\theta}} + \boldsymbol{u} - \boldsymbol{\theta}^*\| \leqslant r_n C\}$ using (A1) and Lemma 7. The right-hand side is strictly negative on the subset where $\boldsymbol{u} \neq \mathbf{0}$ with probability tending to 1, so it must hold $\boldsymbol{u} = \mathbf{0}$.

A.3 Proof of Theorem 3

We have

$$\begin{aligned}\mathbf{0} = \mathbb{P}_n \phi(\hat{\boldsymbol{\theta}}) &= \mathbb{P}_n \phi(\boldsymbol{\theta}^*) + \mathbb{P}_n[\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)] \\ &= \mathbb{P}_n \phi(\boldsymbol{\theta}^*) + P[\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)] + (\mathbb{P}_n - P)[\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)] \\ &= \mathbb{P}_n \phi(\boldsymbol{\theta}^*) + \nabla_{\boldsymbol{\theta}} P\phi(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + (\mathbb{P}_n - P)[\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)],\end{aligned}$$

with some $\tilde{\boldsymbol{\theta}}$ between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. We have $\nabla_{\boldsymbol{\theta}} P\phi(\tilde{\boldsymbol{\theta}}) = J(\tilde{\boldsymbol{\theta}})$, so

$$-J(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \mathbb{P}_n \phi(\boldsymbol{\theta}^*) + [J(\tilde{\boldsymbol{\theta}}) - J(\boldsymbol{\theta}^*)](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + (\mathbb{P}_n - P)[\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)]$$

and

$$\begin{aligned}-\sqrt{n} A_n J(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = &\sqrt{n} A_n \mathbb{P}_n \phi(\boldsymbol{\theta}^*) \\ &+ \sqrt{n} A_n [J(\tilde{\boldsymbol{\theta}}) - J(\boldsymbol{\theta}^*)](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &+ \sqrt{n}(\mathbb{P}_n - P) A_n [\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)].\end{aligned}$$

The second and the third term are negligible, since

$$\sqrt{n} A_n [J(\tilde{\boldsymbol{\theta}}) - J(\boldsymbol{\theta}^*)](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = o\left(\sqrt{n} \frac{1}{\sqrt{n} r_n} r_n\right) = o(1)$$

by assumption (A2), and Lemma 10 yields

$$\sqrt{n}(\mathbb{P}_n - P) A_n [\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)] = o_p(1).$$

It remains to show a central limit theorem for

$$\sqrt{n} A_n \mathbb{P}_n \phi(\boldsymbol{\theta}^*) = \sum_{i=1}^n \frac{1}{\sqrt{n}} A_n \phi_i(\boldsymbol{\theta}^*) \coloneqq \sum_{i=1}^n \mathbf{Y}_i.$$

Since

$$\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{Y}_i\|^2 \mathbb{1}\{\|\mathbf{Y}_i\| > \varepsilon\}\right] \leqslant \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{Y}_i\|^2 \mathbb{1}\{\|\mathbf{Y}_i\| > \varepsilon\} \|\mathbf{Y}_i\|^2/\varepsilon^2\right] \leqslant \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{Y}_i\|^4\right]/\varepsilon^2,$$

and $\mathbb{E}[\|\boldsymbol{Y}_i\|^4] = n^{-2}\mathbb{E}[\|A_n\phi_i(\boldsymbol{\theta}^*)\|^4] = o(n^{-1})$ for all $i = 1, \ldots, n$, by (A3), we have

$$\sum_{i=1}^{n} \mathbb{E}\left[\|\boldsymbol{Y}_i\|^2 \mathbb{1}\{\|\boldsymbol{Y}_i\| > \varepsilon\}\right] \to 0 \text{ for every } \varepsilon > 0.$$

Since $\mathbb{E}[\boldsymbol{Y}_i] = \mathbf{0}$ for all $i = 1, \ldots, n$ and

$$\sum_{i=1}^{n} \mathrm{Cov}(\boldsymbol{Y}_i) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}[A_n\phi_i(\boldsymbol{\theta}^*)] = \frac{1}{n}A_n\sum_{i=1}^{n} \mathrm{Cov}[\phi_i(\boldsymbol{\theta}^*)]A_n^\top = A_n I(\boldsymbol{\theta}^*)A_n^\top \to \Sigma,$$

the conditions of the Lindeberg-Feller central limit theorem (van der Vaart, 1998, Section 2.8) are satisfied, and we obtain

$$\sqrt{n}A_n J(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to_d \mathcal{N}(\mathbf{0}, \Sigma).$$

A.4 Proof of Theorem 4

Let $\hat{\boldsymbol{\theta}} \in \Theta_n$ be any solution of (5), which we write as $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \boldsymbol{u}$. It holds

$$\begin{aligned} 0 &= \langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u})\rangle \\ &\leqslant \langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^*)\rangle + \langle \boldsymbol{u}, \mathbb{P}_n H_n\boldsymbol{u}\rangle - \langle \boldsymbol{u}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u})\rangle. \end{aligned} \tag{20}$$

Using Hölder's inequality and Lemma 13, the first term in (20) can be bounded by

$$\langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^*)\rangle \leqslant \|\boldsymbol{u}\|_1\eta_n \leqslant \sqrt{\nu_n}\|\boldsymbol{u}\|_2\eta_n$$

with probability tending to 1. For the second term in (20), (A4) and Hölder's inequality yield

$$\begin{aligned} \langle \boldsymbol{u}, \mathbb{P}_n H_n\boldsymbol{u}\rangle &= \langle \boldsymbol{u}, PH_n\boldsymbol{u}\rangle + \langle \boldsymbol{u}, (\mathbb{P}_n - P)H_n\boldsymbol{u}\rangle \\ &\leqslant -c\|\boldsymbol{u}\|^2 + \nu_n\|\boldsymbol{u}\|^2 \max_{1\leqslant j,k\leqslant p_n} |(\mathbb{P}_n - P)H_{n,j,k}| \\ &= -c\|\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 o_p(1), \end{aligned}$$

where the last step follows from Lemma 8. For the third term in (20), Hölder's inequality gives

$$-\langle \boldsymbol{u}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u})\rangle \leqslant \sqrt{\nu_n}\|\boldsymbol{u}\|_2\bar{b}_n.$$

Altogether, we have shown

$$0 \leqslant \|\boldsymbol{u}\|_2\sqrt{\nu_n}(\eta_n + \bar{b}_n) - \|\boldsymbol{u}\|_2^2[c + o_p(1)].$$

Rearranging terms gives

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = \|\boldsymbol{u}\|_2 \leqslant \frac{\sqrt{\nu_n}(\eta_n + \bar{b}_n)}{c + o_p(1)} = O_p\left(\sqrt{\nu_n}(\eta_n + \bar{b}_n)\right),$$

as claimed.

A.5 Proof of Theorem 5

The proof is split in two steps:

1. Show that there is a solution $\hat{\boldsymbol{\theta}}_{(1)}$ to
$$\Phi_n((\boldsymbol{\theta}_{(1)}, \mathbf{0}))_{(1)} \in \partial_{\boldsymbol{\theta}_{(1)}} p_{\boldsymbol{\lambda}_n}((\boldsymbol{\theta}_{(1)}, \mathbf{0})) \in \mathbb{R}^{s_n}$$
with $\|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}\| = O_p(r_n)$.

2. Show that $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{(1)}, \mathbf{0})$ is also a valid solution to
$$\Phi_n(\hat{\boldsymbol{\theta}})_{(2)} \in \partial_{\boldsymbol{\theta}_{(2)}} p_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}}).$$
Since
$$\partial p_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}})_{(2)} \supseteq [-\lambda_{n,s_n+1}, \lambda_{n,s_n+1}] \times \ldots \times [-\lambda_{n,p_n}, \lambda_{n,p_n}]$$
by (P2), it suffices to verify $\|\operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}\Phi_n(\hat{\boldsymbol{\theta}})_{(2)}\|_\infty \leqslant 1$.

Together this implies that $\hat{\boldsymbol{\theta}}$ is a valid solution to the full problem (5).

Step 1: Observe that (P1) implies that $p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})$ is differentiable on Θ_n'. Similar as in the proof of Theorem 1, it therefore suffices to show that

$$(r_n C)^{-1} \sup_{\|\boldsymbol{u}\|=1, \boldsymbol{u}_{(2)}=\mathbf{0}} \langle \boldsymbol{u}_{(1)}, \Phi_n(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})_{(1)} - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})_{(1)} \rangle \leqslant -c < 0, \tag{21}$$

for large enough C and high probability. The left-hand side of (21) is upper bounded by

$$(r_n C)^{-1} \sup_{\|\boldsymbol{u}\|=1, \boldsymbol{u}_{(2)}=\mathbf{0}} \langle \boldsymbol{u}_{(1)}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})_{(1)} \rangle - (r_n C)^{-1} \inf_{\|\boldsymbol{u}\|=1, \boldsymbol{u}_{(2)}=\mathbf{0}} \langle \boldsymbol{u}_{(1)}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})_{(1)} \rangle \tag{22}$$

For the first term, one can proceed similarly to the proof of Theorem 1 and show that, using (A1) and choosing C large enough,

$$(r_n C)^{-1} \sup_{\|\boldsymbol{u}\|=1, \boldsymbol{u}_{(2)}=\mathbf{0}} \langle \boldsymbol{u}_{(1)}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})_{(1)} \rangle \leqslant -c < 0$$

holds with arbitrarily high probability. It remains to show that the second term in (22) is sufficiently small. A Taylor expansion yields

$$\begin{aligned}
&(r_n C)^{-1}\langle \boldsymbol{u}_{(1)}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + r_n C\boldsymbol{u})_{(1)} \rangle \\
&= (r_n C)^{-1}\langle \boldsymbol{u}_{(1)}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} \rangle + \langle \boldsymbol{u}_{(1)}, \nabla^2_{\boldsymbol{\theta}_{(1)}} p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + r_n C\boldsymbol{u}')\boldsymbol{u}_{(1)} \rangle
\end{aligned}$$

with some $\boldsymbol{u}'$ with $\|\boldsymbol{u}'\| \leqslant 1, \boldsymbol{u}'_{(2)} = \mathbf{0}$. Using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
(r_n C)^{-1} \sup_{\|\boldsymbol{u}\|=1, \boldsymbol{u}_{(2)}=\mathbf{0}} \left|\langle \boldsymbol{u}_{(1)}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} \rangle\right| &\leqslant (r_n C)^{-1} \sup_{\|\boldsymbol{u}\|=1, \boldsymbol{u}_{(2)}=\mathbf{0}} \|p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)}\|\|\boldsymbol{u}\| \\
&\leqslant (r_n C)^{-1}\sqrt{s_n} \max_{k=1,\ldots,s_n} \{|p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_k|\} \\
&= (r_n C)^{-1}\sqrt{s_n} b_n^* = O(1/C),
\end{aligned}$$

which becomes negligible by choosing C large enough. For the second term, (P1) gives

$$\sup_{\|\boldsymbol{u}\|=1, \boldsymbol{u}_{(2)}=\mathbf{0}} \langle \boldsymbol{u}_{(1)}, \nabla^2_{\boldsymbol{\theta}_{(1)}} p_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + r_n C\boldsymbol{u}')\boldsymbol{u}_{(1)} \rangle = o(1),$$

which proves (21).

Step 2: Lemma 11 yields

$$
\begin{aligned}
&\| \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \Phi_n(\hat{\boldsymbol{\theta}})_{(2)} \|_\infty \\
&\leqslant \quad \| \underbrace{\operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} J(\tilde{\boldsymbol{\theta}})_{(2,1)}\ J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)}\ p'_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}})_{(1)}}_{=:\boldsymbol{v}_1} \|_\infty \\
&\quad + \| \underbrace{\operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \left(\mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(2)} - J(\tilde{\boldsymbol{\theta}})_{(2,1)}\ J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)}\ \mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(1)} \right)}_{=:\boldsymbol{v}_2} \|_\infty \\
&\quad + o_p(1)
\end{aligned}
$$

with some $\tilde{\boldsymbol{\theta}}$ between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$. It holds $\|\boldsymbol{v}_1\|_\infty \leqslant \alpha \in [0,1)$ by the definition of α in (A6). Further, Lemma 13 implies $\|\mathbb{P}_n \phi(\boldsymbol{\theta}^*)\|_\infty \leqslant \eta_n$ with probability tending to 1. On this event, the definitions of η_n in (6) and of $\boldsymbol{\lambda}_{n(2)}$ in (A7) give

$$
\begin{aligned}
\|\boldsymbol{v}_2\|_\infty &\leqslant \frac{1-\alpha}{4} \frac{\|\mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(2)}\|_\infty}{\eta_n} + \| \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} J(\tilde{\boldsymbol{\theta}})_{(2,1)}\ J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)} \|_\infty\ \|\mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(1)}\|_\infty \\
&\leqslant \frac{1-\alpha}{4} + \frac{1-\alpha}{4} \frac{\max_{k=s_n+1,\ldots,p_n} \frac{1}{J_{n,k}} \|(J(\tilde{\boldsymbol{\theta}})_{k,(1)} J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)})^\top\|_1\ \|\mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(1)}\|_\infty}{\eta_n} \\
&\leqslant \frac{1}{2}(1-\alpha).
\end{aligned}
$$

Together, we have shown

$$
\| \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \Phi_n(\hat{\boldsymbol{\theta}})_{(2)} \|_\infty \leqslant \alpha + \frac{1}{2}(1-\alpha) + o_p(1) = \frac{1}{2}(1+\alpha) + o_p(1) < 1
$$

with probability going to 1.

A.6 Proof of Theorem 6

Suppose there is a solution $\hat{\boldsymbol{\theta}} \in \Theta_n$ and a further solution $\widetilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \widetilde{\boldsymbol{u}} \in \Theta_n$. From Theorem 4 we know that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\sqrt{\nu_n}(\eta_n + \|\boldsymbol{\lambda}_n\|_\infty)\right)$ and $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\sqrt{\nu_n}(\eta_n + \|\boldsymbol{\lambda}_n\|_\infty)\right)$. Similar to the proofs of Theorem 2 and Theorem 4, it must hold

$$
\begin{aligned}
0 &= \langle \widetilde{\boldsymbol{u}}, \mathbb{P}_n \phi(\widetilde{\boldsymbol{\theta}}) - p'_{\boldsymbol{\lambda}_n}(\widetilde{\boldsymbol{\theta}}) \rangle \\
&= \langle \widetilde{\boldsymbol{u}}, \mathbb{P}_n[\phi(\widetilde{\boldsymbol{\theta}}) - \phi(\hat{\boldsymbol{\theta}})] \rangle - \langle \widetilde{\boldsymbol{u}}, p'_{\boldsymbol{\lambda}_n}(\widetilde{\boldsymbol{\theta}}) - p'_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}}) \rangle \\
&\leqslant -c\|\widetilde{\boldsymbol{u}}\|_2^2 + \nu_n \|\widetilde{\boldsymbol{u}}\|_2^2 \max_{1 \leqslant j,k \leqslant p_n} |(\mathbb{P}_n - P)H_{j,k}| + \frac{1}{2}\mu_n \|\widetilde{\boldsymbol{u}}\|_2^2 \\
&\leqslant \|\widetilde{\boldsymbol{u}}\|_2^2 [-c + \frac{1}{2}\mu_n + o_p(1)],
\end{aligned}
$$

where we used (A4) and (P3) in the first inequality, and Lemma 8 in the second. Because $\mu_n < 2c$ asymptotically, it must hold $\|\widetilde{\boldsymbol{u}}\|_2 = 0$ or, equivalently, $\widetilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ with probability tending to 1.

A.7 Proof of Theorem 7

Similar to the proof of Theorem 3, we obtain

$$
\begin{aligned}
\mathbf{0} &= \mathbb{P}_n \phi(\hat{\boldsymbol{\theta}})_{(1)} - p'_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}})_{(1)} \\
&= \mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(1)} + J(\tilde{\boldsymbol{\theta}})_{(1)}(\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}) + (\mathbb{P}_n - P)[\phi(\hat{\boldsymbol{\theta}})_{(1)} - \phi(\boldsymbol{\theta}^*)_{(1)}] \\
&\quad - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)} - \nabla^2_{\boldsymbol{\theta}_{(1)}} p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}')(\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)})
\end{aligned}
$$

for and some $\tilde{\boldsymbol{\theta}}$ on the line segment from $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}^*$. It then holds

$$\begin{aligned}\sqrt{n}A_n\left[-J(\boldsymbol{\theta}^*)_{(1)}(\hat{\boldsymbol{\theta}}_{(1)}-\boldsymbol{\theta}^*_{(1)})+p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^*)_{(1)}\right] = &\sqrt{n}A_n\mathbb{P}_n\phi(\boldsymbol{\theta}^*)_{(1)}\\ &+\sqrt{n}A_n[J(\tilde{\boldsymbol{\theta}})_{(1)}-J(\boldsymbol{\theta}^*)_{(1)}](\hat{\boldsymbol{\theta}}_{(1)}-\boldsymbol{\theta}^*_{(1)})\\ &+\sqrt{n}A_n(\mathbb{P}_n-P)[\phi(\hat{\boldsymbol{\theta}})_{(1)}-\phi(\boldsymbol{\theta}^*)_{(1)}]\\ &-\sqrt{n}A_n\nabla^2_{\boldsymbol{\theta}_{(1)}}p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}')(\hat{\boldsymbol{\theta}}_{(1)}-\boldsymbol{\theta}^*_{(1)}).\end{aligned}$$

Adapting the proof of Theorem 3, one can show a central limit theorem for the first term and that the second and third term are of order $o_p(1)$ by (A2) and Lemma 10. For the last term, we have

$$\sqrt{n}A_n\nabla^2_{\boldsymbol{\theta}_{(1)}}p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}')(\hat{\boldsymbol{\theta}}_{(1)}-\boldsymbol{\theta}^*_{(1)}) = o_p(1)$$

by assumption (P4) and $\|A_n\| = O(1)$, which concludes the proof.

B Proofs of Corollaries

B.1 Proof of Corollary 1

Consistency For (A1), we can choose

$$H_n(\boldsymbol{X}) = -\inf_{\boldsymbol{\theta}\in\Theta_n}|\psi'(Y_i,\boldsymbol{X}_i^\top\boldsymbol{\theta})|\boldsymbol{X}_i\boldsymbol{X}_i^\top.$$

Then, (A1)(i) is fulfilled by the definition of H_n and (ii) follows from (14). For (iii), $\inf_{\boldsymbol{\theta}\in\Theta_n}|\psi'(Y_i,\boldsymbol{X}_i^\top\boldsymbol{\theta})|$ is negligible as this term is bounded, so it remains to verify

$$\frac{1}{n}\left\|\sum_{i=1}^n\mathbb{E}[(\boldsymbol{X}_i\boldsymbol{X}_i^\top)^2\mathbb{1}_{\|\boldsymbol{X}_i\|^2\leqslant B_n}]\right\| = o(n/\ln p_n)\quad\text{and}\quad\int_{B_n}^\infty\mathbb{P}\left(\|\boldsymbol{X}_i\boldsymbol{X}_i^\top\|>t\right) = o(1).$$

The second condition holds for $B_n = p_na_n$ with $a_n\to\infty$ arbitrarily slowly by Lemma 1 with $\beta(x) = x$ and $F_n(\boldsymbol{X}) = \|\boldsymbol{X}\boldsymbol{X}^\top\|$. This a valid B_n since $p_n\ln p_n/n\to 0$ and the second condition follows from Lemma 2. Now the consistency result follows from Theorem 1. Since our choice of H_n does not rely on $\boldsymbol{\theta}^*$, (A1)(i) holds with $\boldsymbol{\theta}^*$ replaced by any $\boldsymbol{\theta}\in\Theta_n$, and the resulting estimator is unique by Theorem 2. The rate of convergence is $\sqrt{p_n/n}$, asinces $\mathrm{tr}(I(\boldsymbol{\theta}^*)) = O(p_n)$ follows from $\mathbb{E}[\phi(\boldsymbol{X}_i;\boldsymbol{\theta}^*)_k^4] = O(1)$.

Asymptotic normality It suffices to verify (A2) for each row $\boldsymbol{a}_n$ of A_n, as $A_n\phi_i(\boldsymbol{\theta})$ is a finite dimensional vector. Boundedness of A_n implies $\|\boldsymbol{a}_n\| = O(1)$. Using a Taylor expansion and boundedness of $\psi'(Y_i,\boldsymbol{X}_i^\top\boldsymbol{\theta})$, it suffices to use $\boldsymbol{a}_n^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\tilde{\boldsymbol{u}}$ with $\tilde{\boldsymbol{u}} := \boldsymbol{u}-\boldsymbol{u}'$ instead of $\boldsymbol{a}_n^\top[\phi_i(\boldsymbol{\theta}^*+\boldsymbol{u})-\phi_i(\boldsymbol{\theta}^*+\boldsymbol{u}')]$. We obtain

$$\mathbb{E}[|\boldsymbol{a}_n^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\tilde{\boldsymbol{u}}|^2] = \mathbb{E}[|\boldsymbol{a}_n^\top\boldsymbol{X}_i|^2|\boldsymbol{X}_i^\top\tilde{\boldsymbol{u}}|^2]\leqslant\sqrt{\mathbb{E}[|\boldsymbol{a}_n^\top\boldsymbol{X}_i|^4]\mathbb{E}[|\tilde{\boldsymbol{u}}^\top\boldsymbol{X}_i|^4]} = O(\|\tilde{\boldsymbol{u}}\|^2),$$

verifying the first condition in (A2) (with $p_n^2/n = o(1)$). For the second, note that the first condition in (13) implies $\mathbb{E}[\|\boldsymbol{X}_i\|] = O(\sqrt{p_n})$ since $\mathbb{E}[\|\boldsymbol{X}_i\|]\leqslant\sqrt{\mathbb{E}[\|\boldsymbol{X}_i\|^2]} = \sqrt{\sum_{k=1}^{p_n}\mathbb{E}[X_{i,k}^2]} = O(\sqrt{p_n})$. Set $B_n = \rho^{-1}(n\omega_n)\sqrt{p_n}$ with $\omega_n\to\infty$ arbitrarily slowly. This is a valid choice since

$$\frac{B_nr_np_n}{\sqrt{n}} = \frac{\rho^{-1}(n\omega_n)p_n^2}{n} = o(1),$$

by assumption. We have

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathbb{P}\left(\sup_{\|\boldsymbol{u}\|\leqslant r_n C} \frac{|\boldsymbol{a}_n^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{u}|}{\|\boldsymbol{u}\|} > B_n\right) \\
&\leqslant n\mathbb{P}\left(|\boldsymbol{a}_n^\top \boldsymbol{X}_i|\|\boldsymbol{X}_i\| > B_n\right) \\
&\leqslant n\mathbb{P}\left(|\boldsymbol{a}_n^\top \boldsymbol{X}_i|\mathbb{E}[\|\boldsymbol{X}_i\|] > B_n\right) + n\mathbb{P}\left(|\boldsymbol{a}_n^\top \boldsymbol{X}_i|(\|\boldsymbol{X}_i\| - \mathbb{E}[\|\boldsymbol{X}_i\|]) > B_n\right) \\
&\leqslant n\mathbb{P}\left(|\boldsymbol{a}_n^\top \boldsymbol{X}_i|\mathbb{E}[\|\boldsymbol{X}_i\|] > B_n\right) + n\mathbb{P}\left(|\boldsymbol{a}_n^\top \boldsymbol{X}_i|^2 > B_n\right) + n\mathbb{P}\left(|\|\boldsymbol{X}_i\| - \mathbb{E}[\|\boldsymbol{X}_i\|]|^2 > B_n\right) \\
&\leqslant \frac{n\mathbb{E}[\rho(|\boldsymbol{a}_n^\top \boldsymbol{X}_i|)]}{\rho(B_n/\mathbb{E}[\|\boldsymbol{X}_i\|])} + \frac{n\mathbb{E}[\rho(|\boldsymbol{a}_n^\top \boldsymbol{X}_i|^2)]}{\rho(B_n)} + \frac{n\mathbb{E}[\rho(|\|\boldsymbol{X}_i\| - \mathbb{E}[\|\boldsymbol{X}_i\|]|^2)]}{\rho(B_n)} \\
&= o(1),
\end{aligned}
$$

since all expectations are bounded by assumption, $\rho(x)$ increasing and $B_n/\mathbb{E}[\|\boldsymbol{X}_i\|] \geqslant \rho^{-1}(n\omega_n)$ for n large enough.

Because ψ' is Lipschitz, we further have

$$
\begin{aligned}
\|\boldsymbol{a}_n^\top[J(\boldsymbol{\theta}^* + \boldsymbol{u}) - J(\boldsymbol{\theta}^*)]\| &\lesssim \|\mathbb{E}\left[\boldsymbol{a}_n^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{X}_i^\top \boldsymbol{u}\right]\| \\
&= \sup_{\|\boldsymbol{u}'\|=1} |\mathbb{E}\left[\boldsymbol{a}_n^\top \boldsymbol{X}_i \boldsymbol{u}'^\top \boldsymbol{X}_i \boldsymbol{u}^\top \boldsymbol{X}_i\right]| \\
&\leqslant |\mathbb{E}[|\boldsymbol{a}_n^\top \boldsymbol{X}_i|^3]^{1/3} \mathbb{E}[|\boldsymbol{u}'^\top \boldsymbol{X}_i|^3]^{1/3} \mathbb{E}[|\boldsymbol{u}^\top \boldsymbol{X}_i|^3]^{1/3}| \\
&= O(\|\boldsymbol{a}_n\|\|\boldsymbol{u}\|\|\tilde{\boldsymbol{u}}\|) = O(\sqrt{p_n/n}) = o(1/\sqrt{p_n}),
\end{aligned}
$$

since $p_n^2/n = o_(1)$, verifying the third condition in (A2).

Finally, since $\|A_n \phi_i(\boldsymbol{\theta})\|^4 \leqslant \|A_n\|^4 \|\phi_i(\boldsymbol{\theta})\|^4$,

$$
\|\phi_i(\boldsymbol{\theta})\|^4 = \left(\sum_{k=1}^{p_n} \phi_i(\boldsymbol{\theta})_k^2\right)^2 = \sum_{k=1}^{p_n}\sum_{k'=1}^{p_n} \left(\phi_i(\boldsymbol{\theta})_k \phi_i(\boldsymbol{\theta})_{k'}\right)^2,
$$

and $\max_k \mathbb{E}[\phi(\boldsymbol{X}_i; \boldsymbol{\theta}^*)_k^4] = O(1)$, we have $\mathbb{E}\left[\|A_n \phi_i(\boldsymbol{\theta}^*)\|^4\right] = O(p_n^2) = o(n)$, since $p_n^2/n = o(1)$, verifying (A3). We have checked all conditions of Theorem 3, which yields the claim.

B.2 Proof of Corollary 2

We apply Theorems 5–7. The rate of convergence in Theorem 5 is $r_n = \sqrt{s_n \ln p_n / n}$, since $a_n = \lambda_n = O(\eta_n J_{n,k})$, $\eta_n = O(\sqrt{\ln p_n / n})$ and $J_{n,k} \leqslant \alpha$ by (A7), see also Section 3.3.3.

- The conditions on the penalty (P1)–(P3) are satisfied for the Group Lasso by the assumptions on $\|\boldsymbol{\theta}_{G_i}^*\|$. (P4) follows from the assumptions on s_n and p_n.
- That the reduced problem (9) satisfies (A1) follows from the proof of Corollary 1.
- For (A5), our assumptions give $\sigma_n = \sigma$ which is bounded away from zero and infinity. Further the union bound, Markov's inequality, (15), and (17) imply

$$
\begin{aligned}
\sum_{i=1}^{n} \mathbb{P}\left(\|\phi_i(\boldsymbol{\theta}^*)\|_\infty > \sqrt{n\sigma^2/4\ln p_n}\right) &\leqslant np_n \max_k \mathbb{P}\left(\phi_i(\boldsymbol{\theta}^*)_k^2 > n\sigma^2/4\ln p_n\right) \\
&\leqslant \frac{np_n \max_k \mathbb{E}[\rho(\phi_i(\boldsymbol{\theta}^*)_k^2)]}{\rho(n\sigma^2/4\ln p_n)} \\
&= O\left(\frac{np_n}{\rho(n/\ln p_n)}\right) \\
&= o(1).
\end{aligned}
$$

- Eq. (16) implies (A6), and (A7) holds with the proposed choice of λ_n.
- Observe that for $|\psi'| \leqslant K$ and $\boldsymbol{e}_k$ the kth unit vector, we have

$$|\phi_i(\boldsymbol{\theta})_k - \phi_i(\boldsymbol{\theta}')_k| \leqslant K|\boldsymbol{e}_k^\top \boldsymbol{X}_i||\boldsymbol{X}_{i(1)}^\top(\boldsymbol{\theta}_{(1)} - \boldsymbol{\theta}_{'(1)})|,$$

  using that $\boldsymbol{\theta}_{(2)} = \mathbf{0}$ for $\boldsymbol{\theta} \in \Theta_n'$. By our design conditions, we get

$$\max_{1 \leqslant k \leqslant p_n} \mathbb{E}[|\phi_i(\boldsymbol{\theta})_k - \phi_i(\boldsymbol{\theta}')_k|^2] = O(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2),$$

  as in the proof of Corollary 1, so the first condition of (A8) holds because $(s_n^2 + s_n \ln p_n)/n = o(1)$. For the second condition, choose $\tilde{B}_n = K\sqrt{s_n}\rho^{-1}(np_n\omega_n)$ with $\omega_n \to \infty$ arbitrarily slowly. It holds

$$\begin{aligned}
\sum_{i=1}^n \mathbb{P}\left(\sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta'} \frac{\|\phi_i(\boldsymbol{\theta}) - \phi_i(\boldsymbol{\theta}')\|_\infty}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} > \tilde{B}_n\right) &\leqslant n\mathbb{P}\left(K\|\boldsymbol{X}_i\|_\infty \|\boldsymbol{X}_{i(1)}\| > \tilde{B}_n\right) \\
&\leqslant n\mathbb{P}\left(K\sqrt{s_n}\|\boldsymbol{X}_i\|_\infty^2 > \tilde{B}_n\right) \\
&\leqslant np_n \max_{1 \leqslant k \leqslant p_n} \mathbb{P}\left(|X_{i,k}|^2 > \tilde{B}_n/(K\sqrt{s_n})\right) \\
&\leqslant O\left(\frac{np_n}{\rho(\tilde{B}_n/(K\sqrt{s_n}))}\right) \\
&= O\left(\frac{np_n}{\rho(\rho^{-1}(np_n\omega_n))}\right) = o(1).
\end{aligned}$$

  This choice satisfies

$$\frac{\tilde{B}_n r_n(s_n + \ln p_n)}{n\eta_n} \leqslant \frac{\tilde{B}_n\sqrt{s_n}(s_n + \ln p_n)}{2\sigma n} = \frac{K}{2\sigma}\frac{(s_n^2 + s_n \ln p_n)\rho^{-1}(np_n\omega_n)}{n} = o(1),$$

  by (17) as required.
- The first two conditions in (A2) can be verified as in the proof of Corollary 1 by the choice $B_n = \rho^{-1}(n\omega_n)\sqrt{s_n}$ and $\omega_n \to \infty$ arbitrarily slowly. This is a valid choice because

$$\frac{\rho^{-1}(n\omega_n)\sqrt{s_n}r_n s_n}{\sqrt{n}} = \frac{\rho^{-1}(n\omega_n)s_n^2\sqrt{\ln p_n}}{n} \leqslant \frac{\rho^{-1}(np_n)s_n^2\sqrt{\ln p_n}}{n} = o(1)$$

  by (17). Similarly, it follows

$$\sup_{\|\boldsymbol{u}\| \leqslant r_n C} \|A_n[J(\boldsymbol{\theta}^* + \boldsymbol{u}) - J(\boldsymbol{\theta}^*)]\| = O(r_n) = o\left(\frac{1}{\sqrt{n}r_n}\right),$$

  because

$$r_n^2\sqrt{n} = \frac{s_n \ln p_n}{\sqrt{n}} = o(1),$$

  by (17). Assumptions (A3) is verified exactly as in Corollary 1, so asymptotic normality of $\hat{\boldsymbol{\theta}}_{(1)}$ follows from Theorem 7.

- To verify the conditions of Theorem 6, we can choose H_n as in Corollary 1. This construction is independent of $\boldsymbol{\theta}^*$. Since $|\psi'|$ is bounded away from zero, the eigenvalue condition implies that (A4)(i) and (ii) hold for all $\boldsymbol{\theta} \in \Theta_n$. It remains to verify (A4)(iii) with $\nu_n = s_n$. As $\max_{1 \leqslant k \leqslant p_n} \mathbb{E}[X_{i,k}^4] = O(1)$ by (13), the first condition holds since $s_n^2 \ln p_n / n = o(1)$. For the second condition, choose $B_n = \rho^{-1}(n p_n \omega_n)$ with $\omega_n \to \infty$ arbitrarily slowly. Then

$$\begin{aligned}
\sum_{i=1}^n \mathbb{P}\left(\max_{1 \leqslant j,k \leqslant p_n} |X_{i,j} X_{i,k}| > B_n\right) &\leqslant \sum_{i=1}^n \mathbb{P}\left(\max_{1 \leqslant k \leqslant p_n} X_{i,k}^2 > B_n\right) \\
&\leqslant n p_n \max_{1 \leqslant j,k \leqslant p_n} \mathbb{P}\left(X_{i,k}^2 > B_n\right) \\
&\leqslant \frac{n p_n \max_{1 \leqslant j,k \leqslant p_n} \mathbb{E}[\rho(X_{i,k}^2)]}{\rho(B_n)} \\
&= O\left(\frac{n p_n}{n p_n \omega_n}\right) = o(1),
\end{aligned}$$

using (15). This choice satisfies $B_n = o(n/(s_n \ln p_n))$ since $\rho^{-1}(n p_n) s_n \ln p_n / n = o(1)$ by (17).

B.3 Proof of Corollary 3

We first verify (A1). Since

$$\nabla_{\boldsymbol{\theta}} \phi_i(\boldsymbol{\theta}) = \begin{pmatrix} \operatorname{diag}\left(K_n \mathbb{1}\{k_i = k\} \psi'(\boldsymbol{X}_i; \theta_k)\right)_{k=1,\ldots,K_n} & \mathbf{0} \\ \frac{1}{K_n} \mathbf{1}^\top & -1 \end{pmatrix},$$

we can choose

$$H_n(\boldsymbol{x}_i, k_i) = \begin{pmatrix} \operatorname{diag}\left(K_n \mathbb{1}\{k_i = k\}(-\inf_{\theta_k \in \Theta^0} |\psi'(\boldsymbol{x}_i; \theta_k)|)\right)_{k=1,\ldots,K_n} + I_{K_n}/\sqrt{4K_n} & \mathbf{0} \\ \mathbf{0} & -1 + 1/\sqrt{4K_n} \end{pmatrix}$$

by Lemma 4. Then

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[H_n(\boldsymbol{X}_i, k_i)\right] = \begin{pmatrix} \operatorname{diag}\left(\mathbb{E}[-\inf_{\theta_k \in \Theta^0} |\psi'(\boldsymbol{X}_i; \theta_k)|]\right)_{k=1,\ldots,K_n} + I_{K_n}/\sqrt{4K_n} & \mathbf{0} \\ \mathbf{0} & -1 + 1/\sqrt{4K_n}. \end{pmatrix},$$

so $\limsup_{n \to \infty} \lambda_{\max}(n^{-1} \sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)]) < 0$ because ψ' is negative and bounded away from 0. Further,

$$\mathbb{E}[H_n(\boldsymbol{X}_i, k_i)^2] = \begin{pmatrix} \operatorname{diag}\left(K_n^2 \mathbb{1}\{k_i = k\} \mathbb{E}[\inf_{\theta_k \in \Theta^0} |\psi'(\boldsymbol{X}_i; \theta_k)|^2]\right)_{k=1,\ldots,K_n} & \mathbf{0} \\ \mathbf{0} & (1 + 1/\sqrt{4K_n})^2 \end{pmatrix},$$

so

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i, k_i)^2] = \begin{pmatrix} \operatorname{diag}\left(K_n \mathbb{E}[\inf_{\theta_k \in \Theta^0} |\psi'(\boldsymbol{X}_i; \theta_k)|^2]\right)_{k=1,\ldots,K_n} & \mathbf{0} \\ \mathbf{0} & (1 - 1/\sqrt{4K_n})^2 \end{pmatrix},$$

using $n_k/n = 1/K_n$. This gives $n^{-1} \| \sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)^2] \| = O(K_n)$ because $|\psi'(\boldsymbol{x}_i; \theta)|$ is bounded. Since $\max_{1 \leqslant i \leqslant n} \|H_n(\boldsymbol{X}_i)\| = O(K_n)$, we can choose $B_n = K_n C$ in (A1) with some large enough C. Then, (A1) is satisfied since $K_n \ln K_n / n \to 0$. Because the given $H_n(\boldsymbol{x}_i, k_i)$ is valid for all $\boldsymbol{\theta} \in \Theta_0^{K_n + 1}$, the solution is also unique by Theorem 2. Next, we have

$$I(\boldsymbol{\theta}^*) = K_n \operatorname{diag}\left(\mathbb{E}[\psi(\boldsymbol{X}_i; \theta_1^*)^2], \ldots, \mathbb{E}[\psi(\boldsymbol{X}_i; \theta_{K_n}^*)^2], 0\right),$$

which implies that the convergence rate of the stacked parameter vector $\hat{\boldsymbol{\theta}}$ is $\sqrt{\operatorname{tr}(I(\boldsymbol{\theta}^*))/n} = \sqrt{K_n^2/n} = \sqrt{K_n/n_1}$.

We now verify (A2) and (A3). It holds

$$J(\boldsymbol{\theta}^*) = \begin{pmatrix} \operatorname{diag}\left(\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_1^*)],\dots,\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_{K_n}^*)]\right) & \mathbf{0} \\ \frac{1}{K_n}\mathbf{1}^\top & -1 \end{pmatrix},$$

for which the block inversion formula yields

$$J(\boldsymbol{\theta}^*)^{-1} = \begin{pmatrix} \operatorname{diag}\left(\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_1^*)]^{-1},\dots,\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_{K_n}^*)]^{-1}\right) & \mathbf{0} \\ -\frac{1}{K_n}\operatorname{vec}\left(\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_1^*)]^{-1},\dots,\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_{K_n}^*)]^{-1}\right) & -1 \end{pmatrix}.$$

Choosing $A_n = \boldsymbol{a}_n^\top J(\boldsymbol{\theta}^*)^{-1} \in \mathbb{R}^{1\times p_n}$ with $\boldsymbol{a}_n^\top = (0,\dots,0,1)$ gives the statement for $\hat{\theta}_{K_n+1}$, as

$$A_n = \boldsymbol{a}_n^\top J(\boldsymbol{\theta}^*)^{-1} = \operatorname{vec}\left(-\frac{1}{K_n}\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_1^*)]^{-1},\dots,-\frac{1}{K_n}\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_{K_n}^*)]^{-1},-1\right),$$

and, therefore,

$$\boldsymbol{a}_n^\top J(\boldsymbol{\theta}^*)^{-1} I(\boldsymbol{\theta}^*) J(\boldsymbol{\theta}^*)^{-\top}\boldsymbol{a}_n = \frac{1}{K_n}\sum_{k=1}^{K_n}\frac{\mathbb{E}[\psi(\boldsymbol{X}_i;\theta_k^*)^2]}{\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_k^*)]^2}.$$

We have

$$\begin{aligned}&A_n[\phi_i(\boldsymbol{\theta}^*+\boldsymbol{u}) - \phi_i(\boldsymbol{\theta}^*+\boldsymbol{u}')] \\ &= u_{K_n+1} - u'_{K_n+1} - \frac{1}{K_n}\sum_{k=1}^{K_n}(u_k - u'_k) - \sum_{k=1}^{K_n}\frac{\mathbb{1}\{k_i = k\}\psi'(\boldsymbol{X}_i;\theta_k^* + \tilde{u}_k)(u_k - u'_k)}{\mathbb{E}[\psi'(\theta_k^*)]} = O(\|\boldsymbol{u}-\boldsymbol{u}'\|)\end{aligned}$$

since ψ' is bounded away from 0 and $-\infty$. Choosing $B_n = C$ with some large enough C, the first two conditions in (A2) are satisfied since $r_n^2 p_n = K_n^3/n \to 0$. Further

$$A_n[J(\boldsymbol{\theta}^*+\boldsymbol{u}) - J(\boldsymbol{\theta}^*)] = -\frac{1}{K_n}\operatorname{vec}\left(\frac{\mathbb{E}[\psi'(\theta_1^*+u_1) - \psi'(\theta_1^*)]}{\mathbb{E}[\psi'(\theta_1^*)]},\dots,\frac{\mathbb{E}[\psi'(\theta_{K_n}^*+u_{K_n}) - \psi'(\theta_{K_n}^*)]}{\mathbb{E}[\psi'(\theta_{K_n}^*)]},0\right).$$

Since ψ' is Lipschitz and $\|\boldsymbol{a}^\top\| \leqslant \sqrt{K_n}\max_i|a_i|$, we have $\|A_n[J(\boldsymbol{\theta}^*+\boldsymbol{u}) - J(\boldsymbol{\theta}^*)]\|/\|\boldsymbol{u}\| = O(K_n^{-1/2})$, so the third condition in (A2) holds since $K_n^3/n \to 0$ We further have

$$A_n\phi_i(\boldsymbol{\theta}^*) = -\sum_{k=1}^{K_n}\frac{\mathbb{1}\{k_i = k\}\,\psi(\boldsymbol{X}_i;\theta_k^*)}{\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_k^*)]} - \frac{1}{K_n}\sum_{k=1}^{K_n}\theta_k^* - \theta_{K_n+1}^* = -\sum_{k=1}^{K_n}\frac{\mathbb{1}\{k_i = k\}\,\psi(\boldsymbol{X}_i;\theta_k^*)}{\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_k^*)]}.$$

This impliues that $\max_{1\leqslant i\leqslant n}\mathbb{E}\left[\|A_n\phi_i(\boldsymbol{\theta}^*)\|^4\right] = O(1)$ using that $\max_{i,k}\mathbb{E}[\psi(\boldsymbol{X}_i;\theta_k^*)^4] = O(1)$ since ψ is bounded and $\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_k^*)]$ is bounded away from 0. This verifies (A3).

To obtain the asymptotic distribution of $\sqrt{n/K_n}(\hat{\theta}_k - \theta_k^*)$, choose $\boldsymbol{a}_n^\top = K_n^{-1/2}\boldsymbol{e}_k^\top$ and

$$A_n = \boldsymbol{a}_n^\top J(\boldsymbol{\theta}^*)^{-1} = K_n^{-1/2}\mathbb{E}[\psi'(\boldsymbol{X}_i;\theta_k^*)]^{-1}\boldsymbol{e}_k^\top.$$

Simple calculations yield $O(1)$, $B_n = \sqrt{K_n}C$ and $O(K_n^{-1/2})$ for the three quantities in (A2), so the conditions of Theorem 3 are satisfied since $K_n^3/n \to 0$.

B.4 Proof of Corollary 4

Due to simplicity of the estimating equation, most regularity conditions are automatic. In particular, $J(\boldsymbol{\theta}) = -I_{p_n}$ so that conditions (A6) becomes void and (A7) holds with $J_{n,k} = 1$ for all k and

$$\eta_n = 2 \max_{1 \leqslant k \leqslant K_n} \sqrt{\frac{\frac{1}{n}\sum_{i=1}^n \frac{n^2}{n_k^2}\mathbb{E}[\phi_k(\boldsymbol{X}_i, \theta_k^*)^2] \ln K_n}{n}} = 2 \max_{1 \leqslant k \leqslant K_n} \sqrt{\frac{\mathbb{E}[(q_k(\boldsymbol{X}_i) - a - \theta_k^*)^2] \ln K_n}{n_k}}.$$

Since $\phi_i(\boldsymbol{\theta})_k - \phi_i(\boldsymbol{\theta}')_k = \mathbb{1}\{k_i = k\}\frac{n}{n_k}(\theta_{k'} - \theta_k)$, (A8) can be verified using and $s_n^2 \ln K_n / \min_k n_k \to 0$. Further, (P1) is satisfied because either $\theta_k^* = 0$ or $\theta_k^*/r_n \to \infty$, (P2) always holds for the Lasso and (P3) is easily verified with $\mu_n = 0$. As

$$\langle \boldsymbol{u}, \phi_i(\boldsymbol{\theta} + \boldsymbol{u}) - \phi_i(\boldsymbol{\theta}) \rangle = -\sum_{k=1}^{K_n} \mathbb{1}\{k_i = k\}\frac{n}{n_k}u_k^2$$

the s_n-dimensional subproblem satisfies (A1) with $H_n(k_i, \boldsymbol{X}_i) = -\operatorname{diag}(\mathbb{1}\{k_i = k\}\frac{n}{n_k})_{k=1,\dots,s_n}$ and $\ln K_n / \min_k n_k \to 0$. Existence of a solution with the claimed detection property now follows from Theorem 5. The additional conditions in Theorem 6 also hold with $H_n = -\frac{n}{\min_k n_k} I_{p_n}$ since $s_n^2 \ln K_n / \min_k n_k \to 0$, which guarantees uniqueness of the solution.

B.5 Proof of Corollary 5

First note that we can multiply the first block of ϕ with any $\kappa > 0$ without changing the solution. Denote $\boldsymbol{X} = (Y, T, \boldsymbol{Z}, \boldsymbol{W})$. By the mean value theorem, for any $\boldsymbol{u}$, there exists $s \in (0,1)$ such that

$$\langle \boldsymbol{u}, \phi(\boldsymbol{X}; \boldsymbol{\theta} + \boldsymbol{u}) - \phi(\boldsymbol{X}; \boldsymbol{\theta}) \rangle = \boldsymbol{u}^\top \nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{X}; \boldsymbol{\theta} + s\boldsymbol{u})\boldsymbol{u},$$

where

$$\nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{X}; \boldsymbol{\theta}) = \begin{pmatrix} \kappa[T(\ln \sigma)''(\boldsymbol{W}^\top \boldsymbol{\theta}_1) + (1-T)(\ln \bar{\sigma})''(\boldsymbol{W}^\top \boldsymbol{\theta}_1)]\boldsymbol{W}\boldsymbol{W}^\top & 0 \\ \left[\frac{YT}{\sigma(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2} + \frac{Y(1-T)}{\bar{\sigma}(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2}\right]\sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)\boldsymbol{Z}\boldsymbol{W}^\top & -\boldsymbol{Z}\boldsymbol{Z}^\top \end{pmatrix}.$$

To obtain suitable matrices $H_n(\boldsymbol{X})$, we distinguish the two cases $T = 0, 1$. Consider first the case $T = 1$, in which we can simplify

$$\begin{aligned}
&\boldsymbol{u}^\top \nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{X}; \boldsymbol{\theta})\boldsymbol{u} \\
&= \boldsymbol{u}^\top \begin{pmatrix} \kappa(\ln \sigma)''(\boldsymbol{W}^\top \boldsymbol{\theta}_1)\boldsymbol{W}\boldsymbol{W}^\top & 0 \\ \frac{Y\sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{\sigma(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2}\boldsymbol{Z}\boldsymbol{W}^\top & -\boldsymbol{Z}\boldsymbol{Z}^\top \end{pmatrix} \boldsymbol{u} \\
&= \begin{pmatrix} \boldsymbol{u}_1^\top \boldsymbol{W} \\ \boldsymbol{u}_2^\top \boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \kappa(\ln \sigma)''(\boldsymbol{W}^\top \boldsymbol{\theta}_1) & 0 \\ \frac{Y\sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{\sigma(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2} & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{W}^\top \boldsymbol{u}_1 \\ \boldsymbol{Z}^\top \boldsymbol{u}_2 \end{pmatrix} \\
&\leqslant \begin{pmatrix} \boldsymbol{u}_1^\top \boldsymbol{W} \\ \boldsymbol{u}_2^\top \boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \kappa(\ln \sigma)''(\boldsymbol{W}^\top \boldsymbol{\theta}_1) + \left|\frac{Y\sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{2\sigma(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2}\right| & 0 \\ 0 & -1 + \left|\frac{Y\sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{2\sigma(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2}\right| \end{pmatrix} \begin{pmatrix} \boldsymbol{W}^\top \boldsymbol{u}_1 \\ \boldsymbol{Z}^\top \boldsymbol{u}_2 \end{pmatrix}
\end{aligned}$$

using Lemma 4. By the same arguments, we get a similar result for $T = 0$. Denoting $\overline{\sigma} = 1 - \sigma$, we obtain

$$\begin{aligned}
&\boldsymbol{u}^\top \nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{X}; \boldsymbol{\theta}) \boldsymbol{u} \\
&\leqslant \begin{pmatrix} \boldsymbol{u}_1^\top \boldsymbol{W} \\ \boldsymbol{u}_2^\top \boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \kappa (\ln \overline{\sigma})''(\boldsymbol{W}^\top \boldsymbol{\theta}_1) + \left| \frac{Y \sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{2\overline{\sigma}(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2} \right| & 0 \\ 0 & -1 + \left| \frac{Y \sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)}{2\overline{\sigma}(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2} \right| \end{pmatrix} \begin{pmatrix} \boldsymbol{W}^\top \boldsymbol{u}_1 \\ \boldsymbol{Z}^\top \boldsymbol{u}_2 \end{pmatrix}.
\end{aligned}$$

By the assumptions on $\boldsymbol{W}$, σ and Y, there is $K \in (0, \infty)$ such that

$$|Y \sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)/(2\sigma(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2)| \leqslant K, \quad |Y \sigma'(\boldsymbol{W}^\top \boldsymbol{\theta}_1)/(2\overline{\sigma}(\boldsymbol{W}^\top \boldsymbol{\theta}_1)^2)| \leqslant K, \quad \|\mathbb{E}[\boldsymbol{W}\boldsymbol{W}^\top]\| \leqslant K.$$

Now the matrix

$$H_n(T, Y, \boldsymbol{W}) = \begin{pmatrix} [\kappa \alpha_1(T, \boldsymbol{W}) + K] \boldsymbol{W}\boldsymbol{W}^\top & 0 \\ 0 & \alpha_2(T, Y, \boldsymbol{W}) \boldsymbol{Z}\boldsymbol{Z}^\top, \end{pmatrix}$$

satisfies (A1)(i)–(ii) with $\kappa \geqslant 2K^2/c$. Finally, because α_1 and α_2 are uniformly bounded, (A1)(iii), (A2) and (A3) can be verified exactly as for the generalized linear model (Corollary 1).

B.6 Proof of Corollary 6

Conditions (P1), (P2) (since $\sqrt{n/s_n} \min_{1 \leqslant k \leqslant s_n} \theta_k^* \to \infty$) and (P4) are satisfied by the SCAD penalty. Consistency with $r_n = \sqrt{s_n/n}$ and asymptotic normality of $\hat{\boldsymbol{\theta}}_{(1)}$ as well as its oracle property (efficiency) follow from Corollary 5, $\rho^{-1}(n) s_n^2/n \to 0$ and the properties of the SCAD penalty with $\lambda_n \to 0$.

(A5) follows from the assumptions on p_n, see the proof of Corollary 2. (A6) always holds for SCAD. The proposed choice of λ_n satisfies (A7) and (A8) can be verified as in the proof of Corollary 2, so $\hat{\boldsymbol{\theta}}_{(2)} = \boldsymbol{0}$ with probability tending to 1. (A4) follows from the assumptions on p_n. The assumption on a, which determines the non-convexity of SCAD, implies that (P3) holds with $\mu < 2c$, so $\hat{\boldsymbol{\theta}}$ is unique with probability tending to 1 by Theorem 6.

B.7 Proof of Corollary 7

The Jacobian of ϕ is a bidiagonal matrix with

$$\nabla \phi(\boldsymbol{X}; \boldsymbol{\theta}) = K_n \begin{pmatrix} -\mathbb{1}_{i \in \mathcal{B}_1} & & & \\ [1 - \alpha f'(\boldsymbol{X}_i; \theta_1)] \mathbb{1}_{i \in \mathcal{B}_2} & -\mathbb{1}_{i \in \mathcal{B}_2} & & \\ & \ddots & & \\ & & [1 - \alpha f'(\boldsymbol{X}_i; \theta_{K_n - 1})] \mathbb{1}_{i \in \mathcal{B}_{K_n}} & -\mathbb{1}_{i \in \mathcal{B}_{K_n}} \end{pmatrix}.$$

We have

$$
\begin{aligned}
&K_n^{-1}\boldsymbol{u}^\top \nabla\phi(\boldsymbol{X};\boldsymbol{\theta})\boldsymbol{u} \\
&= -\sum_{k=1}^{K_n} u_k^2 \mathbb{1}_{i\in\mathcal{B}_k} + \sum_{k=1}^{K_n-1} u_k u_{k+1}[1-\alpha f'(\boldsymbol{X}_i;\theta_k)]\mathbb{1}_{i\in\mathcal{B}_{k+1}} \\
&\leqslant -\sum_{k=1}^{K_n} u_k^2 \mathbb{1}_{i\in\mathcal{B}_k} + (1-\alpha\kappa)\sum_{k=1}^{K_n-1}\sqrt{u_k^2 u_{k+1}^2}\mathbb{1}_{i\in\mathcal{B}_{k+1}} && (\kappa \leqslant f' \leqslant 1/\alpha) \\
&\leqslant -\sum_{k=1}^{K_n} u_k^2 \mathbb{1}_{i\in\mathcal{B}_k} + (1-\alpha\kappa)\frac{1}{2}\sum_{k=1}^{K_n-1}(u_k^2 + u_{k+1}^2)\mathbb{1}_{i\in\mathcal{B}_{k+1}} && \text{(AM-GM inequality)} \\
&\leqslant -\sum_{k=1}^{K_n} u_k^2 \mathbb{1}_{i\in\mathcal{B}_k} + (1-\alpha\kappa)\sum_{k=1}^{K_n} u_k^2 \mathbb{1}_{i\in\mathcal{B}_k} \\
&\leqslant -\alpha\kappa\sum_{k=1}^{K_n} u_k^2 \mathbb{1}_{i\in\mathcal{B}_k}.
\end{aligned}
$$

Hence, (A1) is satisfied with

$$
H_n(\boldsymbol{x}) = -\alpha\kappa K_n \operatorname{diag}(\mathbb{1}_{i\in\mathcal{B}_k})_{k=1,\dots,K_n},
$$

$c = \alpha\kappa$, $B_n = K_n \ln n$ and $K_n^3/n \to 0$. Further, (A1)(i) is valid for all $\boldsymbol{\theta} \in \mathbb{R}^{K_n}$. Finally,

$$
I(\boldsymbol{\theta}^*) = \mathbb{C}\mathrm{ov}[\phi(\boldsymbol{\theta}^*)] = \alpha^2 K_n \operatorname{diag}\left(\mathbb{V}\mathrm{ar}[f'(\boldsymbol{X};\theta_{k-1}^*)]\right)_{k=1,\dots,K_n} =: \alpha^2 K_n \Gamma,
$$

so $\operatorname{tr}(I(\boldsymbol{\theta}^*)) = O(K_n^2)$ and $r_n = O(\sqrt{K_n^2/n})$. Now Theorem 1 and Theorem 2 show that, with probability tending to 1, a unique solution path $\hat{\boldsymbol{\theta}}$ exists and satisfies $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p(\sqrt{K_n^2/n})$.

We now verify the conditions of Theorem 3 for a matrix $\tilde{A}_n$ such that $\|\tilde{A}_n\| = 1/\sqrt{K_n}$ to be chosen later. It holds

$$
\|\phi_i(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi_i(\boldsymbol{\theta}^* + \boldsymbol{u}')\| \leqslant (2+\alpha L)K_n\|\boldsymbol{u} - \boldsymbol{u}'\| \leqslant 3K_n\|\boldsymbol{u} - \boldsymbol{u}'\|,
$$

so the first two conditions of (A2) are satisfied for any matrix $\tilde{A}_n$ with $\|\tilde{A}_n\| = O(1/\sqrt{K_n})$ and $B_n = K_n$ since $K_n^3/n \to 0$. Next,

$$
J(\boldsymbol{\theta}) = \mathbb{E}[\nabla\phi(\boldsymbol{X};\boldsymbol{\theta})] = \begin{pmatrix}
-1 & & & \\
1-\alpha\mathbb{E}[f'(\boldsymbol{X};\theta_1)] & -1 & & \\
 & \ddots & & \\
 & & 1-\alpha\mathbb{E}[f'(\boldsymbol{X};\theta_{K_n-1})] & -1
\end{pmatrix}.
$$

We have

$$
\begin{aligned}
\|\tilde{A}_n J(\boldsymbol{\theta}^* + \boldsymbol{u}) - \tilde{A}_n J(\boldsymbol{\theta}^* + \boldsymbol{u}')\| &\leqslant \alpha\|\tilde{A}_n\|\sqrt{\sum_{j=1}^{K_n-1}\left(\mathbb{E}[f'(\boldsymbol{X};\theta_k^* + u_k)] - \mathbb{E}[f'(\boldsymbol{X};\theta_k^* + u_k')]\right)^2} \\
&\leqslant \alpha\|\tilde{A}_n\| L\|\boldsymbol{u} - \boldsymbol{u}'\| = O(\|\boldsymbol{u} - \boldsymbol{u}'\|/\sqrt{K_n}) \\
&= o(\|\boldsymbol{u} - \boldsymbol{u}'\|/(\sqrt{n}r_n^2)),
\end{aligned}
$$

where we used $r_n^2 = K_n^2/n$ and $K_n^3/n \to 0$. This verifies the third condition in (A2). Assumption (A3) also holds because only one entry of $\phi_i(\boldsymbol{\theta}^*)$ can be non-zero at a time and, thus,

$$
\begin{aligned}
&\mathbb{E}[\|\tilde{A}_n \phi_i(\boldsymbol{\theta}^*)\|^4] \\
&= O(1/K_n^2) \times \mathbb{E}[\|\phi_i(\boldsymbol{\theta}^*)\|^4] \\
&= O(1/K_n^2) \times K_n^4 \sum_{k=1}^{K_n} \sum_{j=1}^{K_n} \mathbb{E}[(\theta_{k-1}^* - \theta_k^* - f'(\boldsymbol{X}_i; \theta_k^*))^2 (\theta_{j-1}^* - \theta_j^* - f'(\boldsymbol{X}_i; \theta_j^*))^2 \mathbb{1}_{i \in \mathcal{B}_k \cap \mathcal{B}_j}] \\
&= O(K_n^2) \times \sum_{k=1}^{K_n} \mathbb{E}[(\theta_{k-1}^* - \theta_k^* - f'(\boldsymbol{X}_i; \theta_k^*))^4 \mathbb{1}_{i \in \mathcal{B}_k}] \\
&= O(K_n^2) = o(n),
\end{aligned}
$$

again using $K_n^3/n \to 0$.

We can now apply Theorem 3 and it remains to verify the asymptotic covariance structure. By the inversion formula for bidiagonal matrices,

$$
(J(\boldsymbol{\theta})^{-1})_{i,j} = \begin{cases} -1, & i = j \\ -\prod_{\ell=j}^{i-1} (1 - \alpha \mathbb{E}[f'(\boldsymbol{X}; \theta_\ell)]), & i > j \\ 0, & i < j. \end{cases}
$$

Let $\tilde{A}_n = A_n J(\boldsymbol{\theta}^*)^{-1}/\sqrt{K_n} = O(1/\sqrt{K_n})$, so that

$$
\tilde{A}_n I(\boldsymbol{\theta}^*) \tilde{A}_n^\top = \frac{1}{K_n} A_n J(\boldsymbol{\theta}^*)^{-1} I(\boldsymbol{\theta}^*) J(\boldsymbol{\theta}^*)^{-\top} A_n^\top.
$$

Since $I(\boldsymbol{\theta}^*) = \alpha^2 K_n \Gamma$ with Γ diagonal, we have for $i \leqslant j$,

$$
\begin{aligned}
\Sigma_n &\coloneqq \frac{1}{K_n} (J(\boldsymbol{\theta}^*)^{-1} I(\boldsymbol{\theta}^*) J(\boldsymbol{\theta}^*)^{-\top})_{i,j} \\
&= \alpha^2 \sum_{\ell=1}^{K_n} \sum_{k=1}^{K_n} (J(\boldsymbol{\theta})^{-1})_{i,\ell} \Gamma_{\ell,k} (J(\boldsymbol{\theta})^{-1})_{j,k} \\
&= \alpha^2 \sum_{\ell=1}^{i} \sum_{k=1}^{j} (J(\boldsymbol{\theta})^{-1})_{i,\ell} \Gamma_{\ell,k} (J(\boldsymbol{\theta})^{-1})_{j,k} \\
&= \alpha^2 \sum_{k=1}^{i} \Gamma_{k,k} (J(\boldsymbol{\theta})^{-1})_{i,k} (J(\boldsymbol{\theta})^{-1})_{j,k} \\
&= \alpha^2 \sum_{k=1}^{i} \Gamma_{k,k} \left[\prod_{m=k}^{i-1} (1 - \alpha \mathbb{E}[f'(\boldsymbol{X}; \theta_m^*)]) \right] \left[\prod_{m=k}^{j-1} (1 - \alpha \mathbb{E}[f'(\boldsymbol{X}; \theta_m^*)]) \right] \\
&= \alpha^2 \sum_{k=1}^{i} \Gamma_{k,k} \left[\prod_{m=k}^{i-1} (1 - \alpha \mathbb{E}[f'(\boldsymbol{X}; \theta_m^*)]) \right]^2 \left[\prod_{m=i}^{j-1} (1 - \alpha \mathbb{E}[f'(\boldsymbol{X}; \theta_m^*)]) \right].
\end{aligned}
$$

C Lemmas

For any strictly increasing, convex function $\beta \colon (0, \infty) \to (0, \infty)$ and any positive function g, define the norm-like quantity

$$
\|g\|_\beta = \max_{1 \leqslant i \leqslant n} \mathbb{E}[g(\boldsymbol{X}_i)] + \frac{1}{n} \sum_{i=1}^{n} \beta^{-1} \left(\mathbb{E} \left[\tilde{\beta} \left(|g(\boldsymbol{X}_i) - \mathbb{E}[g(\boldsymbol{X}_i)]| \right) \right] \right),
$$

where $\tilde{\beta}(x) = x\beta(x)$. This 'norm' measures both the absolute size of $g(\boldsymbol{X})$ and its concentration. For example, the choice $\beta(x) = x$ corresponds to

$$\|g\|_\beta = \max_{1 \leqslant i \leqslant n} \mathbb{E}[g(\boldsymbol{X}_i)] + \frac{1}{n}\sum_{i=1}^n \mathbb{V}\mathrm{ar}\left[g(\boldsymbol{X}_i)\right].$$

Another important example is $\beta(x) = \exp(x)$ for exponential concentration inequalities.

Lemma 1. *Let F_n be a positive function and $\beta\colon (0,\infty) \to (0,\infty)$ strictly increasing and convex. For any B_n such that $B_n/\|F_n\|_\beta \to \infty$, it holds*

$$\frac{1}{n}\sum_{i=1}^n \int_{B_n}^\infty \mathbb{P}\left(F_n(\boldsymbol{X}_i) > t\right) dt = o(1).$$

Proof. Since $B_n/\|F_n\|_\beta \to \infty$ and the convexity of β imply $B_n/\max_{1\leqslant i \leqslant n}\mathbb{E}[F_n(\boldsymbol{X}_i)] \to \infty$, it holds $\max_{1\leqslant i\leqslant n}\mathbb{E}[F_n(\boldsymbol{X}_i)] \leqslant B_n/2$ for n large enough. We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \int_{B_n}^\infty \mathbb{P}\left(F_n(\boldsymbol{X}_i) > t\right) dt &\leqslant \frac{1}{n}\sum_{i=1}^n \int_{B_n}^\infty \mathbb{P}\left(|F_n(\boldsymbol{X}_i) - \mathbb{E}[F_n(\boldsymbol{X}_i)]| > t - B_n/2\right) dt \\
&= \frac{1}{n}\sum_{i=1}^n \int_{B_n/2}^\infty \mathbb{P}\left(|F_n(\boldsymbol{X}_i) - \mathbb{E}[F_n(\boldsymbol{X}_i)]| > t\right) dt \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|F_n(\boldsymbol{X}_i) - \mathbb{E}[F_n(\boldsymbol{X}_i)]|\mathbb{1}_{F_n(\boldsymbol{X}_i) > B_n/2}] \\
&\overset{(*)}{\leqslant} \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}[\tilde{\beta}(|F_n(\boldsymbol{X}_i) - \mathbb{E}[F_n(\boldsymbol{X}_i)]|)]}{\beta(B_n/2)} \\
&\leqslant \frac{\beta(\|F_n\|_\beta)}{\beta(B_n/2)} = o(1),
\end{aligned}$$

since $B_n/\|F_n\|_\beta \to \infty$. In $(*)$ we use that for a positive random variable Y it holds

$$\mathbb{E}\left[Y\,\mathbb{1}_{Y > B}\right] \leqslant \mathbb{E}\left[\frac{Y\beta(Y)}{\beta(B)}\right]. \qquad \square$$

Lemma 2. *Suppose that $H_n(\boldsymbol{x})$ is negative semi-definite for all $\boldsymbol{x} \in \mathcal{X}$, $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)] = O(1)$, and $B_n = o(n/\ln p_n)$. Then*

$$\frac{1}{n}\left\|\sum_{i=1}^n \mathbb{E}[H_n(\boldsymbol{X}_i)^2 \mathbb{1}_{\|H_n(\boldsymbol{X}_i)\| \leqslant B_n}]\right\| = o(n/\ln p_n).$$

Proof. Using $\operatorname{tr}(AB) \leqslant |\operatorname{tr}(A)|\|B\|$, we have

$$\begin{aligned}
\boldsymbol{u}^\top H_n(\boldsymbol{X}_i)^2 \boldsymbol{u}\mathbb{1}_{\|H_n(\boldsymbol{X}_i)\| \leqslant B_n} &= \operatorname{tr}(\boldsymbol{u}^\top H_n(\boldsymbol{X}_i)^2 \boldsymbol{u})\mathbb{1}_{\|H_n(\boldsymbol{X}_i)\| \leqslant B_n} \\
&= \operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^\top H_n(\boldsymbol{X}_i)^2)\mathbb{1}_{\|H_n(\boldsymbol{X}_i)\| \leqslant B_n} \\
&\leqslant |\operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^\top H_n(\boldsymbol{X}_i))|\|H_n(\boldsymbol{X}_i)\|\mathbb{1}_{\|H_n(\boldsymbol{X}_i)\| \leqslant B_n} \\
&\leqslant |\boldsymbol{u}^\top H_n(\boldsymbol{X}_i)\boldsymbol{u}| B_n \\
&= -\boldsymbol{u}^\top H_n(\boldsymbol{X}_i)\boldsymbol{u} B_n.
\end{aligned}$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^\top\mathbb{E}[H_n(\boldsymbol{X}_i)^2\mathbb{1}_{\|H_n(\boldsymbol{X}_i)\|\leqslant B_n}]\boldsymbol{u} \leqslant -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^\top\mathbb{E}[H_n(\boldsymbol{X}_i)]\boldsymbol{u}B_n,$$

which implies the claim. □

Lemma 3. *Let $\boldsymbol{\lambda}_n = \lambda_n\mathbf{1}$ with $\lambda_n \to 0$ and $\lambda_n \geqslant 2\eta_n$. Denote $\tilde{\Theta}_n = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_1 \leqslant k_n\}$ with some $k_n = o(\eta_n^{-1})$ such that $\boldsymbol{\theta}^* \in \tilde{\Theta}_n$. Suppose that (P3) and (A5) hold,*

$$\sup_{\boldsymbol{\theta}\in\tilde{\Theta}_n}\|p'_{\lambda_n}(\boldsymbol{\theta})\|_\infty = O(\lambda_n), \tag{23}$$

and for all $\|\boldsymbol{u}\| = o(\lambda_n)$ and large enough n,

$$\langle \boldsymbol{u}_{(2)}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u})_{(2)}\rangle \geqslant \lambda_n\|\boldsymbol{u}_{(2)}\|_1. \tag{24}$$

Then if

$$\left\langle \boldsymbol{u}, \frac{1}{n}\sum_{i=1}^{n}[\phi_i(\boldsymbol{\theta}^*) - \phi_i(\boldsymbol{\theta}^* + \boldsymbol{u})]\right\rangle \geqslant c\|\boldsymbol{u}\|^2 - c_1\|\boldsymbol{u}\|_1^2\eta_n^2 \tag{25}$$

holds with $\limsup_{n\to\infty}\mu_n < 2c$, any solution $\hat{\boldsymbol{\theta}} \in \tilde{\Theta}_n$ satisfies

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leqslant \sqrt{\nu_n}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|,$$

for some $\nu_n = O(s_n)$.

Proof. Write $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \boldsymbol{u}$ with $\|\boldsymbol{u}\|_1 \leqslant 2k_n$. We will show that $\|\boldsymbol{u}\|_1 \leqslant O(\sqrt{s_n})\|\boldsymbol{u}\|$. Let $t_n = o(1)$ fast enough such that $\|t_n\boldsymbol{u}\| = o(\lambda_n)$. It holds,

$$\begin{aligned}
\langle \boldsymbol{u}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u})\rangle &= \langle \boldsymbol{u}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u}) - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + t_n\boldsymbol{u})\rangle + \langle \boldsymbol{u}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + t_n\boldsymbol{u})\rangle \\
&= \frac{1}{1-t_n}\langle \boldsymbol{u}(1-t_n), p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u}) - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + t_n\boldsymbol{u})\rangle + \langle \boldsymbol{u}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + t_n\boldsymbol{u})\rangle \\
&\geqslant -\frac{\mu_n}{2}\|\boldsymbol{u}\|^2 + \langle \boldsymbol{u}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + t_n\boldsymbol{u})\rangle \\
&= -\frac{\mu_n}{2}\|\boldsymbol{u}\|^2 + \langle \boldsymbol{u}_{(1)}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + t_n\boldsymbol{u})_{(1)}\rangle + \frac{1}{t_n}\langle t_n\boldsymbol{u}_{(2)}, p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + t_n\boldsymbol{u})_{(2)}\rangle \\
&\geqslant -\frac{\mu_n}{2}\|\boldsymbol{u}\|^2 - O(\lambda_n)\|\boldsymbol{u}_{(1)}\|_1 + \lambda_n\|\boldsymbol{u}_{(2)}\|_1.
\end{aligned}$$

using (P3) and $|1-t_n| \leqslant 1$ in the first, and (23)–(24) in the second inequality. Furthermore, by (25), (A5) and Lemma 13,

$$\begin{aligned}
\langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^* + \boldsymbol{u})\rangle &= \langle \boldsymbol{u}, \mathbb{P}_n[\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]\rangle + \langle \boldsymbol{u}, \mathbb{P}_n\phi(\boldsymbol{\theta}^*)\rangle \\
&\leqslant -c\|\boldsymbol{u}\|^2 + c_1\eta_n^2\|\boldsymbol{u}\|_1^2 + \eta_n\|\boldsymbol{u}\|_1 \\
&\leqslant -c\|\boldsymbol{u}\|^2 + [1 + o(1)]\eta_n\|\boldsymbol{u}\|_1 \\
&= -c\|\boldsymbol{u}\|^2 + O(\lambda_n)\|\boldsymbol{u}_{(1)}\|_1 + [1 + o(1)]\eta_n\|\boldsymbol{u}_{(2)}\|_1,
\end{aligned}$$

where we have used $\eta_n \|\boldsymbol{u}\|_1 \leqslant \eta_n k_n = o(1)$ in the third and $\lambda_n \geqslant 2\eta_n$ in the last step. Together this yields

$$\begin{aligned}
0 &= \langle \boldsymbol{u}, \mathbb{P}_n \phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - p'_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta}^* + \boldsymbol{u}) \rangle \\
&\leqslant -(c - \mu_n/2)\|\boldsymbol{u}\|^2 + O(\lambda_n)\|\boldsymbol{u}_{(1)}\|_1 - \frac{\lambda_n}{2}[1 + o(1)]\|\boldsymbol{u}_{(2)}\|_1,
\end{aligned}$$

as $\eta_n - \lambda_n \leqslant -\lambda_n/2$ by assumption. Since $c - \mu_n/2$ is strictly positive asymptotically, it must hold

$$\|\boldsymbol{u}_{(2)}\|_1 \leqslant O(1)\|\boldsymbol{u}_{(1)}\|_1,$$

Now the claim follows from

$$\|\boldsymbol{u}\|_1 = \|\boldsymbol{u}_{(1)}\|_1 + \|\boldsymbol{u}_{(2)}\|_1 \leqslant O(1)\|\boldsymbol{u}_{(1)}\|_1 \leqslant O(\sqrt{s_n})\|\boldsymbol{u}_{(1)}\| \leqslant O(\sqrt{s_n})\|\boldsymbol{u}\|. \qquad \square$$

Lemma 4. *For any $A \in \mathbb{R}^{q_1 \times q_1}$, $B \in \mathbb{R}^{q_1 \times q_2}$, $C \in \mathbb{R}^{q_2 \times q_1}$, $D \in \mathbb{R}^{q_2 \times q_2}$, $q_1, q_2 \in \mathbb{N}$, it holds*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \preceq \begin{pmatrix} A + I_{q_1}(\|B\| + \|C\|)/2 & 0 \\ 0 & D + I_{q_2}(\|B\| + \|C\|)/2. \end{pmatrix}$$

Proof. Let $\boldsymbol{y}_1 \in \mathbb{R}^{q_1}$ and $\boldsymbol{y}_2 \in \mathbb{R}^{q_2}$ arbitrary. Then using sub-multiplicativity of the norm (first step) and the AM-GM inequality (second step), we get

$$\begin{aligned}
&\begin{pmatrix} \boldsymbol{y}_1^\top & \boldsymbol{y}_2^\top \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \\
&= \boldsymbol{y}_1^\top A \boldsymbol{y}_1 + \boldsymbol{y}_1^\top B \boldsymbol{y}_2 + \boldsymbol{y}_2^\top C \boldsymbol{y}_1 + \boldsymbol{y}_2^\top D \boldsymbol{y}_2 \\
&\leqslant \boldsymbol{y}_1^\top A \boldsymbol{y}_1 + \|\boldsymbol{y}_1\|\|\boldsymbol{y}_2\|(\|B\| + \|C\|) + \boldsymbol{y}_2^\top D \boldsymbol{y}_2 \\
&\leqslant \boldsymbol{y}_1^\top A \boldsymbol{y}_1 + \frac{1}{2}(\|\boldsymbol{y}_1\|^2 + \|\boldsymbol{y}_2\|^2)(\|B\| + \|C\|) + \boldsymbol{y}_2^\top D \boldsymbol{y}_2 \\
&= \begin{pmatrix} \boldsymbol{y}_1^\top & \boldsymbol{y}_2^\top \end{pmatrix} \begin{pmatrix} A + I_{q_1}(\|B\| + \|C\|)/2 & 0 \\ 0 & D + I_{q_2}(\|B\| + \|C\|)/2 \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}. \qquad \square
\end{aligned}$$

Lemma 5. *It holds*

$$\|\mathbb{P}_n \phi(\boldsymbol{\theta}^*)\| = O_p\left(\sqrt{\frac{\operatorname{tr}(I(\boldsymbol{\theta}^*))}{n}}\right).$$

Proof. Using $Y = O_p(\mathbb{E}(Y^2)^{1/2})$ and $\|\boldsymbol{u}\| = \sqrt{\operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^\top)}$, we obtain

$$\|\mathbb{P}_n \phi(\boldsymbol{\theta}^*)\| = O_p(\sqrt{\mathbb{E}[\operatorname{tr}(\mathbb{P}_n \phi(\boldsymbol{\theta}^*)\, \mathbb{P}_n \phi(\boldsymbol{\theta}^*)^\top)]}) = O_p(\sqrt{\operatorname{tr}(\mathbb{E}[\mathbb{P}_n \phi(\boldsymbol{\theta}^*)\, \mathbb{P}_n \phi(\boldsymbol{\theta}^*)^\top])})$$

and

$$\mathbb{E}[\mathbb{P}_n \phi(\boldsymbol{\theta}^*)\, \mathbb{P}_n \phi(\boldsymbol{\theta}^*)^\top] = \operatorname{Cov}[\mathbb{P}_n \phi(\boldsymbol{\theta}^*)] = \frac{1}{n^2}\sum_{i=1}^n \operatorname{Cov}[\phi_i(\boldsymbol{\theta}^*)] = \frac{1}{n} I(\boldsymbol{\theta}^*). \qquad \square$$

Lemma 6. *Let $\mathcal{F}_n$ be classes of real-valued functions from $\mathcal{X}$ to $\mathbb{R}$, F_n be any function with $\sup_{f \in \mathcal{F}_n} |f| \leqslant F_n$, and B_n be any sequence.*

(i) It holds

$$\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f|\leqslant\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{F_n\leqslant B_n}|+O_p\left(\int_{B_n}^{\infty}P\mathbb{1}_{F_n>t}\,dt,\right).$$

(ii) If $P\mathbb{1}_{F_n>B_n}=o(1/n)$, we have

$$\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f|\leqslant\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{F_n\leqslant B_n}|+o_p\left(\sqrt{\frac{\sup_{f\in\mathcal{F}_n}Pf^2}{n}}\right).$$

Proof. It holds

$$\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f|\leqslant\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{F_n\leqslant B_n}|+\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{F_n>B_n}|.$$

We prove two different bounds for the second term on the right.

(i) We have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|(\mathbb{P}_n-P)f\mathbb{1}_{F_n>B_n}|\right]\leqslant\mathbb{E}\left[\mathbb{P}_nF_n\mathbb{1}_{F_n>B_n}\right]+PF_n\mathbb{1}_{F_n>B_n}\leqslant 2PF_n\mathbb{1}_{F_n>B_n}.$$

Noting that $PF_n\mathbb{1}_{F_n>B_n}=\int_{B_n}^{\infty}P\mathbb{1}_{F_n>t}dt$ and Markov's inequality give

$$\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{F_n>B_n}|=O_p\left(\int_{B_n}^{\infty}P\mathbb{1}_{F_n>t}\,dt\right).$$

(ii) Decompose

$$\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{F_n>B_n}|\leqslant|\mathbb{P}_nF_n\mathbb{1}_{F_n>B_n}|+\sup_{f\in\mathcal{F}}|Pf\mathbb{1}_{F_n>B_n}|.$$

Since $P\mathbb{1}_{F_n>B_n}=o(1/n)$ it holds

$$\mathbb{P}(\exists i\colon F_n(\boldsymbol{X}_i)>B_n)\leqslant nP\mathbb{1}_{F_n>B_n}=o(1),$$

so $\mathbb{P}_nF_n\mathbb{1}_{F_n>B_n}=0$ with probability tending to 1. Now the claim follows from the Cauchy Schwarz inequality:

$$\sup_{f\in\mathcal{F}}|Pf\mathbb{1}_{F_n>B_n}|\leqslant\sqrt{\sup_{f\in\mathcal{F}}Pf^2}\sqrt{P\mathbb{1}_{F_n>B_n}}=o_p\left(\sqrt{\frac{\sup_{f\in\mathcal{F}_n}Pf^2}{n}}\right).\qquad\square$$

Lemma 7. *Under assumption (A1)(iii), it holds*

$$\|(\mathbb{P}_n-P)H_n\|=o_p(1).$$

Proof. Let $S_n = \{\boldsymbol{x}\colon \|H_n(\boldsymbol{x})\| \leqslant B_n\}$ with B_n as in (A1)(iii). Applying Lemma 6 (i), we obtain

$$\|(\mathbb{P}_n - P)H_n\| \leqslant \|(\mathbb{P}_n - P)H_n\mathbb{1}_{S_n}\| + o_p(1)$$

by (A1)(iii). Defining $M_n^2 = PH_n^2\mathbb{1}_{S_n}$, the Bernstein inequality for random matrices (Wainwright, 2019, Theorem 6.17) yields

$$\|(\mathbb{P}_n - P)H_n\mathbb{1}_{S_n}\| = O_p\left(\sqrt{\frac{M_n^2 \ln p_n}{n}} + \frac{B_n \ln p_n}{n}\right) = o_p(1). \qquad \square$$

Lemma 8. *Under assumption (A4), it holds*

$$\max_{1\leqslant j,k\leqslant p_n} |(\mathbb{P}_n - P)H_{n,j,k}| = o_p(1/\nu_n).$$

Proof. Define

$$M_n^2 = \max_{1\leqslant j,k\leqslant p_n} PH_{n,j,k}^2, \qquad S_n = \left\{\boldsymbol{x}\colon \max_{1\leqslant j,k\leqslant p_n} |H_n(\boldsymbol{x})_{n,j,k}| \leqslant B_n\right\}.$$

Lemma 6 (ii) and (A4) give

$$\begin{aligned}
\max_{1\leqslant j,k\leqslant p_n} |(\mathbb{P}_n - P)H_{n,j,k}| &\leqslant \max_{1\leqslant j,k\leqslant p_n} |(\mathbb{P}_n - P)H_{n,j,k}\mathbb{1}_{S_n}| + o_p\left(\sqrt{M_n^2/n}\right) \\
&= \max_{1\leqslant j,k\leqslant p_n} |(\mathbb{P}_n - P)H_{n,j,k}\mathbb{1}_{S_n}| + o_p(1/\nu_n),
\end{aligned}$$

with probability tending to 1. The union bound and Bernstein's inequality give

$$\mathbb{P}\left(\max_{1\leqslant j,k\leqslant p_n} |(\mathbb{P}_n - P)H_{n,j,k}\mathbb{1}_{S_n}| > \varepsilon\right) \leqslant 2p_n^2 \exp\left(-\frac{n\varepsilon^2}{2M_n^2 + B_n\varepsilon}\right).$$

The claim follows upon choosing

$$\varepsilon = C\sqrt{\frac{M_n^2 \ln p_n}{n}} + C\frac{B_n \ln p_n}{n},$$

with some large enough constant C and the assumptions on the growth of M_n and B_n from (A4). $\square$

Lemma 9. *For some $c_n \in (0,\infty)$, $d_n, K_n \in \mathbb{N}$, let*

$$\mathcal{F}_n = \{f_{\boldsymbol{u},k}\colon \boldsymbol{u} \in \mathbb{R}^{d_n}, \|\boldsymbol{u}\| \leqslant c_n, f_{\boldsymbol{0},k} \equiv 0, k = 1,\dots,K_n\},$$

be classes of functions such that

$$\begin{aligned}
&\max_{1\leqslant k\leqslant K_n} \sup_{\|\boldsymbol{u}\|,\|\boldsymbol{u}'\|\leqslant c_n} \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}[|f_{\boldsymbol{u},k}(\boldsymbol{X}_i) - f_{\boldsymbol{u}',k}(\boldsymbol{X}_i)|^2]}{\|\boldsymbol{u}-\boldsymbol{u}'\|^2} \leqslant M_n^2, \\
&\sum_{i=1}^n \mathbb{P}\left(\max_{1\leqslant k\leqslant K_n} \sup_{\|\boldsymbol{u}\|,\|\boldsymbol{u}'\|\leqslant c_n} \frac{|f_{\boldsymbol{u},k}(\boldsymbol{X}_i) - f_{\boldsymbol{u}',k}(\boldsymbol{X}_i)|}{\|\boldsymbol{u}-\boldsymbol{u}'\|} > B_n\right) = o(1).
\end{aligned}$$

Then

$$\sup_{f\in\mathcal{F}_n} |(\mathbb{P}_n - P)f| = O_p\left(\sqrt{\frac{M_n^2 c_n^2(d_n + \ln K_n)}{n}} + \frac{B_n c_n(d_n + \ln K_n)}{n}\right),$$

Proof. We proceed in three steps.

Step 1: Truncation We start with a truncation argument. Let

$$F_n(\boldsymbol{x}) = \max_{1\leqslant k\leqslant K_n} \sup_{\|\boldsymbol{u}\|,\|\boldsymbol{u}'\|\leqslant c_n} \frac{|f_{\boldsymbol{u},k}(\boldsymbol{x}) - f_{\boldsymbol{u}',k}(\boldsymbol{x})|}{\|\boldsymbol{u}-\boldsymbol{u}'\|}.$$

Since $f_{k,\mathbf{0}} \equiv 0$ by assumption, F_n is an envelope for the functions in $c_n^{-1}\mathcal{F}_n$:

$$\sup_{f\in\mathcal{F}_n} c_n^{-1}|f| \leqslant \max_{1\leqslant k\leqslant K_n} \sup_{\|\boldsymbol{u}\|\leqslant c_n} c_n^{-1}\|\boldsymbol{u}\| \frac{|f_{\boldsymbol{u},k}(\boldsymbol{x}) - 0|}{\|\boldsymbol{u}-\mathbf{0}\|} \leqslant F_n(\boldsymbol{x}).$$

Now Lemma 6 (ii) and our assumptions give

$$c_n^{-1} \sup_{f\in\mathcal{F}_n} |(\mathbb{P}_n - P)f| \leqslant c_n^{-1} \sup_{f\in\mathcal{F}_n} |(\mathbb{P}_n - P)f\mathbb{1}_{F_n\leqslant B_n}| + o_p(M_n/\sqrt{n}),$$

so

$$\sup_{f\in\mathcal{F}_n} |(\mathbb{P}_n - P)f| \leqslant \sup_{f\in\mathcal{F}_n} |(\mathbb{P}_n - P)f\mathbb{1}_{F_n\leqslant B_n}| + o_p(\sqrt{M_n^2c_n^2/n}).$$

Step 2: Bounding covering numbers Let $\mathcal{A}$ be some set equipped with a norm $\|\cdot\|$. A collection of N balls $B(a_k,\epsilon) = \{a\in\mathcal{A}\colon \|a-a_k\|\leqslant\epsilon\}$ is called an ϵ-covering of $\mathcal{A}$ with respect to the norm $\|\cdot\|$ if $\mathcal{A}\subset\bigcup_{k=1}^N B(a_k,\epsilon)$. The minimal number of balls of radius ϵ needed to cover $\mathcal{A}$ is the covering number $N(\epsilon,\mathcal{A},\|\cdot\|)$.

Fix k and consider $\mathcal{F}_n^{(k)} = \{f_{\boldsymbol{u},k}\mathbb{1}_{S_n}\colon \boldsymbol{u}\in\mathbb{R}^{p_n}, \|\boldsymbol{u}\|\leqslant c_n\}$. Recall that by our definition of Pf, the $L_2(P)$-norm is defined as $\|f-g\|_{L_2(P)}^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|f(X_i)-g(X_i)|^2]$. We will show that

$$\ln N(\varepsilon,\mathcal{F}_n^{(k)},L_2(P)) \lesssim d_n\ln(3M_nc_n/\varepsilon), \quad \ln N(\varepsilon,\mathcal{F}_n^{(k)},\|\cdot\|_\infty) \lesssim d_n\ln(3B_nc_n/\varepsilon), \tag{26}$$

where $\lesssim$ means "bounded up to a universal constant". Let $\boldsymbol{u}_1,\dots,\boldsymbol{u}_N$ be the centers of an η-covering of $\{\boldsymbol{u}\in\mathbb{R}^{d_n}\colon \|\boldsymbol{u}\|\leqslant c_n\}$, which we can find with $N=(3c_n/\eta)^{d_n}$. Then, by the definitions of M_n and B_n, the functions $f_{\boldsymbol{u}_1,k},\dots,f_{\boldsymbol{u}_N,k}$ are centers of a $M_n\eta$-covering of $\mathcal{F}_n^{(k)}$ in $L_2(P)$, and a $B_n\eta$ covering for $\|\cdot\|_\infty$, respectively. Choosing $\eta=\varepsilon/M_n$ and $\eta=\varepsilon/B_n$, respectively, gives (26). Now we can take a union over the coverings of all $\mathcal{F}_n^{(k)}$ to find a covering of $\mathcal{F}_n$, which gives

$$\begin{aligned}\ln N(\varepsilon,\mathcal{F}_n,L_2(P)) &\lesssim \ln K_n + d_n\ln(3M_nc_n/\varepsilon),\\ \ln N(\varepsilon,\mathcal{F}_n,\|\cdot\|_\infty) &\lesssim \ln K_n + d_n\ln(3B_nc_n/\varepsilon).\end{aligned} \tag{27}$$

Step 3: Bounding the truncated process Denote $S_n = \{\boldsymbol{x} : F_n(\boldsymbol{x})\leqslant B_n\}$. Theorem 2.14.21 of Van der Vaart and Wellner (2023) gives

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{S_n}|\right] \lesssim \frac{\int_0^{M_nc_n}\sqrt{1+\ln N(\epsilon,\mathcal{F}_n,L_2(P))}d\epsilon}{\sqrt{n}} + \frac{\int_0^{B_nc_n}[1+\ln N(\epsilon,\mathcal{F}_n,\|\cdot\|_\infty)]d\epsilon}{n}.$$

Substituting the covering number bounds (27) and the changes of variables $t=\epsilon/M_nc_n$ and $t=\epsilon/B_nc_n$, respectively, gives

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_n}|(\mathbb{P}_n-P)f\mathbb{1}_{S_n}|\right] \lesssim \frac{M_nc_n\sqrt{d_n+\ln K_n}}{\sqrt{n}} + \frac{B_nc_n(d_n+\ln K_n)}{n}.$$

Now the result follows from Markov's inequality. □

Lemma 10. *Under assumption (A2), it holds for all $C < \infty$,*

$$\sup_{\|\boldsymbol{u}\| \leqslant r_n C} \|(\mathbb{P}_n - P) A_n [\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]\| = o_p(1/\sqrt{n}).$$

Proof. We show that for each row $\boldsymbol{a}_n$ from $A_n \in \mathbb{R}^{q \times p_n}$, it holds

$$\sup_{\|\boldsymbol{u}\| \leqslant r_n C} |(\mathbb{P}_n - P) \boldsymbol{a}_n^\top [\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]| = o_p(1/\sqrt{n}).$$

Since $A_n[\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]$ is a finite dimensional vector, this implies the claim. Let $\boldsymbol{a}_n$ be some row of A_n. We have

$$\sup_{\|\boldsymbol{u}\| \leqslant r_n C} |(\mathbb{P}_n - P) \boldsymbol{a}_n^\top [\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]| = \sup_{f_{\boldsymbol{u}} \in \mathcal{F}_n} |(\mathbb{P}_n - P) f_{\boldsymbol{u}}|$$

with $\mathcal{F}_n = \{\boldsymbol{a}_n^\top [\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)] : \|\boldsymbol{u}\| \leqslant r_n C\}$. Now apply Lemma 9 with $d_n = p_n$, $c_n = r_n C$, $K_n = 1$. This gives

$$\sup_{f_{\boldsymbol{u}} \in \mathcal{F}_n} |(\mathbb{P}_n - P) f_{\boldsymbol{u}}| = o_p\left(\frac{1}{\sqrt{n}} M_n r_n \sqrt{p_n} + \frac{1}{\sqrt{n}} \frac{B_n r_n p_n}{\sqrt{n}} \right) = o_p(1/\sqrt{n})$$

since $M_n = o(1/(r_n \sqrt{p_n}))$ and $B_n = o(\sqrt{n}/(r_n p_n))$ by (A2). □

Lemma 11. *Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{(1)}, \mathbf{0})$ be a solution to $\Phi_n((\hat{\boldsymbol{\theta}}_{(1)}, \mathbf{0}))_{(1)} \in \partial_{\boldsymbol{\theta}_{(1)}} p_{\boldsymbol{\lambda}_n}((\hat{\boldsymbol{\theta}}_{(1)}, \mathbf{0}))$. Under assumptions (A7) and (A8), it holds*

$$\begin{aligned}
&\| \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \mathbb{P}_n \phi(\hat{\boldsymbol{\theta}})_{(2)} \|_\infty \\
\leqslant \quad & \| \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} J(\tilde{\boldsymbol{\theta}})_{(2,1)} \, J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)} \, p'_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}})_{(1)} \|_\infty \\
& + \| \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \left(\mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(2)} - J(\tilde{\boldsymbol{\theta}})_{(2,1)} \, J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)} \, \mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(1)} \right) \|_\infty + o_p(1)
\end{aligned}$$

with some $\tilde{\boldsymbol{\theta}}$ on the line segment from $\boldsymbol{\theta}^*$ to $\hat{\boldsymbol{\theta}}$.

Proof. We have

$$\operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \mathbb{P}_n \phi(\hat{\boldsymbol{\theta}})_{(2)} = \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(2)} + \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1} \mathbb{P}_n \left[\phi(\hat{\boldsymbol{\theta}})_{(2)} - \phi(\boldsymbol{\theta}^*)_{(2)} \right]$$

and

$$\begin{aligned}
\mathbb{P}_n \left[\phi(\hat{\boldsymbol{\theta}})_{(2)} - \phi(\boldsymbol{\theta}^*)_{(2)} \right] &= P \left[\phi(\hat{\boldsymbol{\theta}})_{(2)} - \phi(\boldsymbol{\theta}^*)_{(2)} \right] + (\mathbb{P}_n - P) \left[\phi(\hat{\boldsymbol{\theta}})_{(2)} - \phi(\boldsymbol{\theta}^*)_{(2)} \right] \\
&= J(\tilde{\boldsymbol{\theta}})_{(2,1)} (\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}) + (\mathbb{P}_n - P) \left[\phi(\hat{\boldsymbol{\theta}})_{(2)} - \phi(\boldsymbol{\theta}^*)_{(2)} \right]
\end{aligned}$$

with some $\tilde{\boldsymbol{\theta}}$ between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ and $J(\boldsymbol{\theta})_{(2,1)}$ as defined in Section 3.3.1. Similar to the proof of Theorem 3, one obtains

$$\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)} = J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)} \left[-\mathbb{P}_n \phi(\boldsymbol{\theta}^*)_{(1)} - (\mathbb{P}_n - P)[\phi(\hat{\boldsymbol{\theta}})_{(1)} - \phi(\boldsymbol{\theta}^*)_{(1)}] + p'_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}})_{(1)} \right],$$

with $J(\boldsymbol{\theta})_{(1)}$ as defined in Section 3.3.1. We may take the same $\tilde{\boldsymbol{\theta}}$ as above here since we are using the same expansion for the term $P\phi(\hat{\boldsymbol{\theta}}) - P\phi(\boldsymbol{\theta}^*)$ in both arguments. The three last displays together yield

$$
\begin{aligned}
& \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}\mathbb{P}_n\phi(\hat{\boldsymbol{\theta}})_{(2)} \\
= \quad & \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}\mathbb{P}_n\phi(\boldsymbol{\theta}^*)_{(2)} \\
& + \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}(\mathbb{P}_n - P)\left[\phi(\hat{\boldsymbol{\theta}})_{(2)} - \phi(\boldsymbol{\theta}^*)_{(2)}\right] \\
& + \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}J(\tilde{\boldsymbol{\theta}})_{(2,1)}\; J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)}\left[-\mathbb{P}_n\phi(\boldsymbol{\theta}^*)_{(1)} + p'_{\boldsymbol{\lambda}_n}(\hat{\boldsymbol{\theta}})_{(1)}\right] \\
& + \operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}J(\tilde{\boldsymbol{\theta}})_{(2,1)}\; J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)}\left[-(\mathbb{P}_n - P)[\phi(\hat{\boldsymbol{\theta}})_{(1)} - \phi(\boldsymbol{\theta}^*)_{(1)}]\right].
\end{aligned}
$$

The claim follows if we show that the the second and fourth terms are negligible (with respect to the $\|\cdot\|_\infty$-norm). With η_n as defined in (A7), it holds

$$
\|\operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}\|_\infty \leqslant \frac{1}{\min_{k=s_n+1,\dots,p_n}\lambda_{n,k}} \leqslant \frac{1}{\eta_n},
$$

and

$$
\begin{aligned}
\|\operatorname{diag}(\boldsymbol{\lambda}_{n(2)})^{-1}J(\tilde{\boldsymbol{\theta}})_{(2,1)}\; J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)}\|_\infty &\leqslant \frac{1}{\eta_n}\max_{k=s_n+1,\dots,p_n}\frac{1}{J_{n,k}}\|(J(\tilde{\boldsymbol{\theta}})_{k,(1)}\; J^{-1}(\tilde{\boldsymbol{\theta}})_{(1)})^\top\|_1 \\
&\leqslant \frac{1}{\eta_n},
\end{aligned}
$$

Further, Lemma 12 gives

$$
\left\|(\mathbb{P}_n - P)[\phi(\hat{\boldsymbol{\theta}}) - \phi(\boldsymbol{\theta}^*)]\right\|_\infty = o_p(\eta_n).
$$

Now the claim follows from combining the previous displays. ☐

Lemma 12. *Under (A8), it holds*

$$
\sup_{\substack{\boldsymbol{u}\in\mathbb{R}^{p_n},\boldsymbol{u}_{(2)}=\boldsymbol{0}\\ \|\boldsymbol{u}\|\leqslant r_nC}} \|(\mathbb{P}_n - P)[\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]\|_\infty = o_p(\eta_n)
$$

for every $C < \infty$.

Proof. We have

$$
\sup_{\substack{\boldsymbol{u}\in\mathbb{R}^{p_n},\boldsymbol{u}_{(2)}=\boldsymbol{0}\\ \|\boldsymbol{u}\|\leqslant r_nC}} \|(\mathbb{P}_n - P)[\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \phi(\boldsymbol{\theta}^*)]\|_\infty = \sup_{f_{\boldsymbol{u},k}\in\mathcal{F}_n}|(\mathbb{P}_n - P)f_{\boldsymbol{u},k}|
$$

with

$$
\mathcal{F}_n = \{\phi(\boldsymbol{\theta}^* + (\boldsymbol{u}_{(1)},\boldsymbol{0}))_k - \phi(\boldsymbol{\theta}^*)_k : \boldsymbol{u}_{(1)}\in\mathbb{R}^{s_n}, \boldsymbol{u}_{(2)} = \boldsymbol{0}, \|\boldsymbol{u}\| \leqslant r_nC, k = 1,\dots p_n\}.
$$

Define

$$
\tilde{M}_n^2 = \max_{1\leqslant k\leqslant p_n}\sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta'}\frac{P|\phi_i(\boldsymbol{\theta})_k - \phi_i(\boldsymbol{\theta}')_k|^2}{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2}
$$

By assumption (A8), we can apply Lemma 9 with $d_n = s_n, K_n = p_n$, $c_n = r_n C$, $M_n = \tilde{M}_n$ and $B_n = \tilde{B}_n$. This gives

$$\mathbb{E}\left(\sup_{f_{\boldsymbol{u},k}\in\mathcal{F}_n} |(\mathbb{P}_n - P)f_{\boldsymbol{u},k}|\right) = O\left(\sqrt{\frac{\tilde{M}_n^2 r_n^2 C^2(s_n + \ln p_n)}{n}} + \frac{\tilde{B}_n r_n C(s_n + \ln p_n)}{n}\right) = o(\eta_n),$$

where we used the growth bounds from (A8) in the last step. Now the claim follows from Markov's inequality. □

Lemma 13. *Under assumption (A5) and $p_n \to \infty$, it holds*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \phi_i(\boldsymbol{\theta}^*)\right\|_\infty \leqslant 2\sigma_n\sqrt{\frac{\ln p_n}{n}}\right) \to 1.$$

Proof. Let

$$\eta_n = 2\sigma_n\sqrt{\frac{\ln p_n}{n}}, \quad B_n = \sqrt{\frac{n}{4\ln p_n}}.$$

Using Lemma 6 (ii), (A5), and $\sqrt{\max_k P\phi(\boldsymbol{\theta}^*)_k^2} \leqslant \sqrt{n}\eta_n$, we get

$$\|\mathbb{P}_n\phi_i(\boldsymbol{\theta}^*)\|_\infty \leqslant \left\|(\mathbb{P}_n - P)\phi_i(\boldsymbol{\theta}^*)\mathbb{1}_{\|\phi_i(\boldsymbol{\theta}^*)\|_\infty \leqslant B_n}\right\|_\infty + o_p(\eta_n),$$

Further, the union bound and Bernstein's inequality give

$$\begin{aligned}
\mathbb{P}\left(\left\|(\mathbb{P}_n - P)\phi_i(\boldsymbol{\theta}^*)\mathbb{1}_{\|\phi_i(\boldsymbol{\theta}^*)\|_\infty \leqslant B_n}\right\|_\infty > \eta_n\right) &\leqslant 2\,p_n \max_{1\leqslant k\leqslant p_n} \exp\left(-\frac{\frac{1}{2}\eta_n^2}{\frac{1}{n}\sigma_n^2 + \frac{1}{3}\eta_n B_n/n}\right) \\
&\leqslant 2\exp\left(\ln p_n - \frac{\eta_n^2 n}{2\sigma_n^2 + \eta_n B_n}\right) \\
&\leqslant 2\exp\left(\ln p_n - \frac{\eta_n^2 n}{3\sigma_n^2}\right) \\
&= 2\exp\left(\ln p_n - \frac{4}{3}\ln p_n\right) \\
&= o(1).
\end{aligned}$$

□