.

Biases Of Z-Estimators Of Parameters In General Estimating Equations For Both Fixed And Growing Dimension

Hanxiang Peng[1]

[1]*Department of Mathematical Science, Indiana University Purdue University Indianapolis, e-mail:* hanxpeng@iu.edu

Abstract: In this article, we give the analytic formulas of the first order biases of the Z-estimators of parameters in GEE, and prove the rate $O(p^{9/2}n^{-3/2})$ for the remainder as the dimension p and sample size n tend to infinity but $p = o(n^{4/13})$ under suitable conditions. We show that the rates can be improved to $O(p^{5/2}n^{-3/2} + p^{7/2}n^{-2})$ and $p = o(n^{4/7})$ under additional conditions. The biases in regularized regression models are obtained, and the assumptions are verified in generalized linear models.

AMS 2000 subject classifications: Primary 62F10; 62F12.
Keywords and phrases: Bias reduction; General estimating equations; Infinite dimension; Regularized Regression; Single Index Models.

1. Introduction

Let $\{\mathbf{z}_{ni} : 1 \le i \le n,\, n \ge 1\}$ be a sequence of independent random variables (rv) defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$. Consider a triangular array of smooth functions $\{\boldsymbol{\psi}_{ni}(\mathbf{z}_{ni}; \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{B} \subset \mathbb{R}^p,\, 1 \le i \le n,\, n \ge 1\}$ taking values in $\mathbb{R}^p$ and satisfying $\mathbb{E}(\boldsymbol{\psi}_{ni}(\mathbf{z}_{ni}; \boldsymbol{\beta}_0)) = 0$ for some unique unknown $\boldsymbol{\beta}_0 \in \mathbb{B}$ for all i and $n \ge 1$. We estimate $\boldsymbol{\beta}_0$ by the solution $\hat{\boldsymbol{\beta}}_n$ to the general estimating equations (GEE),

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\psi}_{ni}(\mathbf{z}_{ni}; \boldsymbol{\beta}) = 0. \tag{1.1}$$

Note that $\hat{\boldsymbol{\beta}}_n$ is not well defined on the whole space Ω but only on a subspace of it for any finite n. Typically, one extends the definition to the whole Ω by defining it to be an arbitrary constant on the complement of the subspace (denoted it by the *s-space*). For instance, consider estimating the probability $\beta \in (0,\, 1)$ in the Bernoulli distribution,

$$\bar{z}_n/\beta - (1 - \bar{z}_n)/(1 - \beta) = 0, \tag{1.2}$$

where $\bar{z}_n$ denotes the average of 1's. The maximum likelihood estimator (MLE) $\hat{\beta}_n = \bar{z}_n$ exists only on the subspace $\{0 < z_1 + \cdots + z_n < n\}$ of Ω.

Let $B_{n,0}$ be the event on which $\hat{\boldsymbol{\beta}}_n$ does not satisfy (1.1). Naturally, the s-space can be taken to be $B_{n,0}$. Without regard to its detailed structure, the consistency and asymptotic distribution can be rigorously established under

suitable conditions. As a result, the s-space $B_{n,0}$ is negligible. Often, however, there is no statement with regard to the bias of the estimator as it generally does *not* exist. Consider estimating the parameter β in the normal $\mathcal{N}(1/\beta, 1)$ by the MLE $\hat{\beta}_n = 1/\bar{z}_n$. While $\hat{\beta}_n$ is not defined at $\bar{z}_n = 0$, the consistency and asymptotic normality (ASN) of $\hat{\beta}_n$ hold at $\beta_0 \neq 0$. The bias, however, does *not* exist because $\mathbb{E}(\hat{\beta}_n)$ diverges.

The preceding $B_{n,0}$ is not suitable for the ongoing analysis of bias. Bias calculation necessitates specification of the s-space, although practical calculation may be carried out without it provided that the s-space is adequately negligible. The probability of the s-space must, certainly, tend to zero faster than that of $B_{n,0}$. But this is not enough, and we need to know that the gradient matrix $\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta})$ is non-singular at the values of certain random vectors (i.e. $\widetilde{\mathbf{B}}_1$ in (1.3)) as we shall rely on the generalized vector multivariate mean value theorem (MVT) to solve for the bias. In contrast, the non-singularity of the expected gradient at the true value $\boldsymbol{\beta}_0$, $\mathbb{E}(\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta}_0))$, is sufficient for proving the consistency and asymptotic distribution.

Consider the case that $\hat{\boldsymbol{\beta}}_n$ satisfies (1.1) (i.e. on $B^c_{n,0}$). By the MVT in (6.5), there is a matrix point $\tilde{\mathbf{B}}_1 \in \mathbb{B}^{p\times p}$ lying in $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$ such that

$$0 = \boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{\Psi}_n(\boldsymbol{\beta}_0) + \dot{\boldsymbol{\Psi}}_n(\tilde{\mathbf{B}}_1)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0), \tag{1.3}$$

where $\dot{\boldsymbol{\Psi}}_n(\mathbf{B}), \mathbf{B} \in \mathbb{B}^{p\times p}$ denotes the *generalized* gradient matrix. Eqt (1.3) reveals that the event that $\dot{\boldsymbol{\Psi}}_n(\tilde{\mathbf{B}}_1)$ is singular must be included in the s-space. In other words, the singularity event $C_{n,0}$ must be included in the s-space, where

$$C_{n,0} = \left\{ \omega \in \Omega : \begin{array}{c} \text{There exists } \mathbf{B} \in \mathbb{B}^{p\times p} \text{ s.t. } \sigma_{\min}(\dot{\boldsymbol{\Psi}}_n(\mathbf{B}))(\omega) = 0 \\ \text{and that (1.3) holds at } \mathbf{B} = \widetilde{\mathbf{B}}_1(\omega). \end{array} \right\} \tag{1.4}$$

Noting that the number N_n of points $\widetilde{\mathbf{B}}_1$ can't diverge infinity too fast. For notational brevity, we shall assume that $N_n = 1$ almost surely.

The s-space is now taken to be the union $A_{n,0} = B_{n,0} \cup C_{n,0}$, on which define $\hat{\boldsymbol{\beta}}_n$ to be an arbitrary but fixed constant $\mathbf{b}$. As a consequence, any moment of $\hat{\boldsymbol{\beta}}_n$ can be "calculated" and, in particular, the variance-covariance and mean squared error (MSE) of $\hat{\boldsymbol{\beta}}_n$.

The concept of bias, dating back to the early years, is fundamental in statistical science. The analysis of bias has recently gained momentum as the complexity of models used in practice increases. For instance, regularization is commonly employed to reduce the model complexity, which, however, leads to *biased* estimators such as LASSO estimators in high dimensional linear models. Estimators which are asymptotically unbiased have finite sample biases that can lead to significant loss of performance of standard inferential procedures, see a comprehensive review by Kosmidis (2014)[13]. Biases are *not* negligible in high dimensional parameter estimation. The squared bias in the decomposition of MSE, for instance, can have higher order of magnitude than the variance as the dimension grows to infinity, while it is dominated by the variance and negligible in the case of fixed dimension, see (2.7).

There is an extensive amount of literature on the analysis of bias, in which the bootstrap, the jackknife, and the approximation are perhaps three most popular methods. The tremendous success of the first two methods rests on the idea of *resampling*, which is computationally intensive. In the Era of Big Data, the two methods are confronted with the challenge that data are of massive size and often accompanied with an estimation of a myriad of unknown parameters. The method of approximation provides a remedy, with which we shall be concerned in this article. A significant achievement in this method is the formulas of the first order biases for MLE in a closed form. Cox and Snell (1968)[5] and McCullagh (1987)[15] gave the formulas for i.i.d. observations achieved through the systematic use of index notation and tensors; Cook, *et al.* (1986)[4] provided the formula in their Eqt (3) for normal nonlinear regression; Cordeiro and McCullagh (1991)[6] obtained the formula in their Eqt (4.2) for GLM; Kosmidis and Firth (2010)[12] presented the formula in their Eqt (2.4) in matrix form.

Efron (1975)[7] investigated the biases of an important class of estimators, noting that the estimators were, apparently, assumed to be well defined on the whole space. He proved the biases and rigorously established the rate $o(1/n)$ for the remainder based on some large deviation results.

While the analytic formulas are convenient in constructing bias-corrected estimators, the analysis of bias in literature is often conducted in a framework of some sort of specific bias formulas or of conditional biases. This appearing to be permissible from a practitioner's viewpoint, there would lead to a lack of a rigorous analysis of bias with generality to the best of our knowledge, and this article is an attempt to fill the gap. Briefly, we provide a bias formula with generality and prove two rates. To be specific, we rely on the generalized MVT to carry out the analysis of bias for Z-estimators in GEE for independent observations, give the bias formulas in matrices in (2.2), and rigorously establish the rates for the remainders with tedious calculation for both fixed and growing dimension. Our spirit here is the same as Efron's (1975)[7], and our approach is elementary and there is no difficulty in extending our results to other cases such as correlated data and incomplete data.

As an application, we derive the bias formulas for Lasso estimators in linear regression model, and the penalized estimators in generalized linear models (GLM) and single index models (SIM). Our analysis of bias indicates that the Lasso estimators vanishes at the fastest rate as the dimension tends to infinity among all Bridge estimators, see Section 4 for more details.

The bias formulas for MLE are well documented in the literature, and agree with ours. Moreover, we exhibit that the bias of a Z-estimator admits an expansion of the form,

$$\text{Bias}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_0 = \mathbf{b}_1 n^{-1} + O(p^{9/2} n^{-3/2}), \tag{1.5}$$

where $\mathbf{b}_1 n^{-1} = O(p^{3/2} n^{-1})$ is the first-order bias, which dominates the remainder for $p = o(n^{1/6})$ among other conditions in Remark 2.1. The rates can be improved to $p = o(n^{1/2})$ and $O(p^{7/2} n^{-2} + p^{5/2} n^{-3/2})$ for the remainder under the conditions in Remark 2.4. For p-dimensional parametric exponential families, Portnoy (1988)[18] showed that MLE are ASN as p, n tend to infinity but

$p = o(n^{1/2})$. Note that the bias expansion found in literature is often of the form Bias $= \mathbf{b}_1 n^{-1} + \mathbf{b}_2 n^{-2} + \cdots$, in decreasing powers of n faster than $n^{-1/2}$ in (1.5). See Wu (1986)[22] for LSE in linear regression, page 203 in the monograph by McCullagh (1987)[15] and Eqt (2.2) of Firth (1993)[8]. In fact, the rate $O(n^{-2})$ was given on page 210 in the monograph. Our result here is that the biases have a slower rate $n^{-1/2}$ of negligibility than MLE.

For bias correction by the bootstrap and jackknife methods, see Wu (1986)[22] and the monograph by Shao and Tu (1995)[20] among others; For indirect inference appeared in the Econometrics literature, see the comprehensive review by Jiang and Turnbull (2004)[10]; Other references include Firt (1993)[8], Gart, *et al.* (1985)[9], Lin and Breslow (1996)[14], Mehrabi and Matthews (1995)[16], Pettitt, *et al.* (1998)[17], Schaefer (1983)[19], and Simas, *et al.* (2010)[21].

The article is structured as follows: The main results are presented in Section 2. The biases in regularized regression models are given in Section 4. Section 5 contains the assumptions, which are verified in GLM in Section 3. Proofs are collected in Section 6.

2. The Bias Formulas

In this section, we present the analytical formulas for the biases and rates for the remainders for both fixed and growing dimension, with the proof delayed.

Write $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\psi}_{ni}(\boldsymbol{\beta}) = \boldsymbol{\psi}_{ni}(\mathbf{z}_{ni}; \boldsymbol{\beta})$, $\boldsymbol{\psi}_{ni} = \boldsymbol{\psi}_{ni}(\boldsymbol{\beta}_0)$, $\dot{\boldsymbol{\Psi}}_n = \dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta}_0)$, etc. Let $\Psi_{n,d}(\boldsymbol{\beta})$ be the d-th component of $\boldsymbol{\Psi}_n(\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathbb{R}^p$ with $\dot{\Psi}_{n,d}(\mathbf{B}), \mathbf{B} \in \mathbb{R}^{p \times p}$ denoting the second order generalized partial derivative, which is the d-th block of the stacking matrix $\ddot{\boldsymbol{\Psi}}_n(\mathbf{B}), \mathbf{B} \in \mathbb{R}^{p^2 \times p}$. Let

$$\mathbf{J}_n = \mathbb{E}(\boldsymbol{\Psi}_n^{\otimes 2}) = \sum_{i=1}^{n} \mathbb{E}(\boldsymbol{\psi}_{ni}^{\otimes 2}), \quad \mathbf{H}_n = \mathbb{E}(\dot{\boldsymbol{\Psi}}_n), \quad \mathbf{G}_{n,d} = \mathbb{E}(\ddot{\Psi}_{n,d}). \tag{2.1}$$

Assume throughout that $\mathbf{H}_n$ is invertible. Let $\vartheta_n = O(\|\mathbf{H}_n\|_o)$. Typically $\vartheta_n = O(pn)$, but $\vartheta_n = O(n)$ is possible when $\mathbf{H}_n$ is a matrix of certain structure. For a random matrix $\mathbf{X}$, define the "under tilde" and "bar" operations as $\underset{\sim}{\mathbf{X}} = \mathbf{H}_n^{-1}\mathbf{X}$ and $\bar{\mathbf{X}} = \mathbf{X} - \mathbb{E}(\mathbf{X})$ unless otherwise specified.

Let $\mathbf{q}_n$ be a vector with components $q_{n,d} = \mathrm{Tr}(\mathbf{H}_n^{-\top}\mathbf{G}_{n,d}\mathbf{H}_n^{-1}\mathbf{J}_n)$, and let

$$\underset{\sim}{\mathbf{b}}_n = \mathbf{H}_n^{-1}(\mathbf{b}_{n1} - 2^{-1}\mathbf{q}_n), \quad \text{where} \quad \mathbf{b}_{n1} = \mathbb{E}(\dot{\boldsymbol{\Psi}}_n \mathbf{H}_n^{-1} \boldsymbol{\Psi}_n). \tag{2.2}$$

As justified by the main theorem below, $\underset{\sim}{\mathbf{b}}_n$ is the *(first-order) bias* under the assumptions specified in Section 5.

Theorem 2.1. *Let $\hat{\boldsymbol{\beta}}_n$ be a solution of (1.1).*

(i) (Unbiasedness) Assume (B0)(a), (B2)(a) and (B5). Then Bias$(\hat{\boldsymbol{\beta}}) = o(1)$.

(ii) (Bias) Assume (B0)(c), (B2), (B3) with $|\tilde{\xi}_n|_4 = O(p^2)$, (B6) with $(\nu_0, \nu_1, \nu_2, \nu_3) = (8, 4, 2, 8)$, and $\kappa(\mathbf{H}_n) = O(1)$. If $p = o(\vartheta_n^{4/13})$, then

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}}) = \underset{\sim}{\mathbf{b}}_n + O(p^{9/2}\vartheta_n^{-3/2}). \tag{2.3}$$

Remark 2.1. *For the case of $\vartheta_n = O(n)$, $\underset{\sim}{\mathbf{b}}_n = O(p^{3/2}n^{-1})$ with the rate $O(p^{9/2}n^{-3/2})$ for the remainder, provided that p and n satisfy $p = o(n^{4/13})$ among other conditions. Moreover, for the remainder to be negligible, p must be much smaller, $p = o(n^{1/6})$. These relationships between p and n reveal that higher dimension has severer adverse effect on the bias.*

Consider a classical case of $\nu_0' = \nu_1' = 2$ and $n^{-1}\mathbf{J}_n = O(p) = n^{-1}\mathbf{H}_n$. By Theorem 2.1 (i), $\hat{\boldsymbol{\beta}}$ is *asymptotically unbiased* if (B0)(a), (B1) with $\nu = 2$ and (B2)(a) hold and $p = o(n)$ as $\mathbb{E}(\|\underset{\sim}{\boldsymbol{\Psi}}_n\|^2) = O(pn^{-1}) = o(1)$. In the case of a constant gradient $\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta})$, (B5) holds with $\nu_0' = \infty$ and $\nu_1' = 1$. Hence, the unbiasedness result holds if $\mathbb{E}(\|n^{-1}\boldsymbol{\Psi}_n\|) = o(\sqrt{p})$.

If only $\mathbf{b}_{n1}$ is involved, then the assumptions required are weaker.

Proposition 2.1. *If $p = o(\vartheta_n^{2/5})$, then*

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}}) = \mathbf{H}_n^{-1}\mathbf{b}_{n1} + O(p^{5/2}\vartheta_n^{-1}). \tag{2.4}$$

MLE. Suppose that $\boldsymbol{\psi}_{ni}(\boldsymbol{\beta})$ are the gradients of a likelihood function. The Z-estimator $\hat{\boldsymbol{\beta}}$ of (1.1) is then the MLE. In this case, $\mathbf{G}_{n,d} = -2\mathbb{E}(\boldsymbol{\Psi}_n\dot{\Psi}_{n,d}) - \mathbb{E}(\dot{\boldsymbol{\Psi}}_n\Psi_{n,d}) - \mathbb{E}(\boldsymbol{\Psi}_n^{\otimes 2}\Psi_{n,d}), d = 1,\ldots,p$, we thus have

$$q_{n,d} = -2\mathrm{Tr}\big(\mathbb{E}(\dot{\Psi}_{n,d}\mathbf{J}_n^{-1}\boldsymbol{\Psi}_n)\big) - \mathrm{Tr}\big(\mathbb{E}(\mathbf{J}_n^{-1}\dot{\Psi}_n\Psi_{n,d})\big) - \mathrm{Tr}\big(\mathbb{E}(\mathbf{J}_n^{-1}\boldsymbol{\Psi}_n^{\otimes 2}\Psi_{n,d})\big).$$

For $\vartheta_n = O(n)$, the d-th component of the bias simplifies to

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}})_d = -2^{-1}\mathbf{J}_n^{-1}\mathrm{Tr}(\mathbf{J}_n^{-1}\mathbb{E}((\dot{\boldsymbol{\Psi}}_n + \boldsymbol{\Psi}_n^{\otimes 2})\Psi_{n,d})) + O(p^{9/2}n^{-3/2}). \tag{2.5}$$

Remark 2.2. *For i.i.d. rv's, the first order bias for MLE in (2.5) is identical to the bias given by McCullagh (1987)[15], Kosmidis and Firth (2010)[12], the formula (20) of Cox and Snell (1968)[5] and (10.21) of Efron (1975)[7].*

Bias correction. Let $\tilde{\boldsymbol{\beta}}$ be a pilot estimator of $\boldsymbol{\beta}_0$, and let $\tilde{\mathbf{J}}, \tilde{\mathbf{H}}, \tilde{\mathbf{G}}$ be estimates of $\mathbf{J}_n, \mathbf{H}_n, \mathbf{G}_n$, respectively. Typically, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$, $\tilde{\mathbf{J}} = \boldsymbol{\Psi}_n^{\otimes 2}(\tilde{\boldsymbol{\beta}})$, $\tilde{\mathbf{H}} = \dot{\boldsymbol{\Psi}}_n(\tilde{\boldsymbol{\beta}})$ and $\tilde{\mathbf{G}} = \ddot{\boldsymbol{\Psi}}_n(\tilde{\boldsymbol{\beta}})$. Other choices are possible or even necessary such as in the Analysis of Big Data in which a subsampling estimator must be used. The bias $\underset{\sim}{\mathbf{b}}_n$ can be estimated by

$$\tilde{\mathbf{b}} = \tilde{\mathbf{H}}^{-1}\big(\tilde{\mathbf{b}}_1 - 2^{-1}\mathbb{T}\mathrm{r}\big(\tilde{\mathbf{J}}(\tilde{\mathbf{H}}^{-\top}\circ\tilde{\mathbf{G}})\tilde{\mathbf{H}}^{-1}\big)\big), \tag{2.6}$$

where $\tilde{\mathbf{b}}_1 = \sum_{i=1}^n \dot{\boldsymbol{\psi}}_{ni}(\tilde{\boldsymbol{\beta}})\tilde{\mathbf{H}}^{-1}\boldsymbol{\psi}_{ni}(\tilde{\boldsymbol{\beta}})$. Consider a function $\mathbf{g}(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$. A bias-corrected estimator of $\hat{\mathbf{g}} = \mathbf{g}(\hat{\boldsymbol{\beta}})$ must be corrected the bias of $\hat{\mathbf{g}}$ itself.

Remark 2.3. *Let $\mathbf{g}(\boldsymbol{\beta})$ be differentiable at $\boldsymbol{\beta}_0$. A bias-corrected estimator of $\mathbf{g}(\boldsymbol{\beta}_0)$ is given by*

$$\hat{\mathbf{g}}_{bc}(\hat{\boldsymbol{\beta}}) = \mathbf{g}(\hat{\boldsymbol{\beta}}) - \dot{\mathbf{g}}(\tilde{\boldsymbol{\beta}})\tilde{\mathbf{b}}.$$

By Theorem 2.1, $\hat{\mathbf{g}}_{bc}(\hat{\mathbf{b}}) - \mathbf{g}(\boldsymbol{\beta}_0) = O(p^{9/2}n^{-3/2})$ under suitable assumptions.

The block matrices $\ddot{\mathbf{\Psi}}_{n,d}(\mathbf{B})$ in the stacking partial derivative $\ddot{\mathbf{\Psi}}_n(\mathbf{B})$ contain structural information. Using the quasinorm $\|\cdot\|_{oe}$ introduced in Section 6, we formulate Assumptions (B2$^+$) – (B3$^+$) and (B6$^+$) in Section 5. The global properties of the block matrices lead to the reduction of the adverse effect of high dimensionality, and yield faster rates as stated below.

Theorem 2.2. *If (B0)(b), (B2) and (B6) in Proposition 2.1 are strengthened to (B0)(c), (B2$^+$) and (B6$^+$), then p and the rate for the remainder in Proposition 2.1 are improved to $p = o(\vartheta_n^{2/3})$ and $O(p^{3/2}/\vartheta_n)$, respectively.*

Furthermore, if (B0)(c) and (B3) in Theorem 2.1 (ii) are strengthened to (B0)(d) and (B3$^+$), then the results in Theorem 2.1 are improved to $p = o(\vartheta_n^{4/7})$ and $O(p^{7/2}/\vartheta_n^2 + p^{5/2}/\vartheta_n^{3/2})$ for the remainder.

Remark 2.4. *Consider the case in Remark 2.1. By Theorem 3.1, for the remainder to be negligible, $p = o(n^{1/2})$ among other conditions. Using the structural information, p is much larger than $p = o(n^{1/6})$ in Remark 2.1.*

MSE. Although MSE is commonly used to study estimation bias and efficiency, its existence has yet been rigorously investigated. Similar to the bias expansion, $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ and $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$ can be expanded. Here we are satisfied with a heuristic expansion for dimension asymptotics. Recall

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2) = \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) + \|\mathrm{Bias}(\hat{\boldsymbol{\beta}})\|^2. \tag{2.7}$$

Consider a typical case of $\vartheta_n = O(n) = \lambda_n$, and $\mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) = O(pn^{-1})$ and $\|\mathrm{Bias}(\hat{\boldsymbol{\beta}})\|^2 = O(p^3 n^{-2})$ by (2.2) provided that $p = o(n^{1/6})$ or $p = o(n^{1/2})$ for the case in Remarks 2.1 or 2.4 among other conditions, respectively. As a consquence, the trace of the covariance matrix in (2.7) dominates the squared bias. The dominance relationship, however, does not hold anymore if p grows to infinity faster than the preceding rates.

3. Biases in GLM

In this section, we verify the assumptions. Let Y_i be independent rv with density in the canonical exponential class,

$$f(y;\theta_i) = c(y)\exp(y\theta_i - b(\theta_i)), \quad \theta \in \Theta \subset \mathbb{R}, \tag{3.1}$$

where $b(\theta)$ has continuous third derivative and $c(y)$ is a normalizing constant. In GLM, the response Y_i and covariate $\mathbf{x}_i$ satisfy

$$\mathbb{E}(Y_i|\mathbf{x}_i) = \mu(\theta_i) = h(\eta_i), \quad \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i = 1, \ldots, n, \tag{3.2}$$

where $\boldsymbol{\beta} \in \mathbb{B}$ is an unknown regression parameter and h is an inverse link. A canonical link yields $\theta_i = \eta_i$, so that $h(\theta) = \mu(\theta), \theta \in \Theta$. In this case, $h'(\theta) = \mu'(\theta) = b''(\theta) = V(\theta)$ and $h''(\theta) = \mu''(\theta) = b'''(\theta) = V'(\theta)$.

The MLE $\hat{\boldsymbol{\beta}}_n$ solves the estimating equations,

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Y_i - \mu_i(\boldsymbol{\beta})}{V_i(\boldsymbol{\beta})} h_i'(\boldsymbol{\beta})\mathbf{x}_i = 0, \tag{3.3}$$

where, as is customary, $\mu_i(\boldsymbol{\beta}) =: \mu(\theta_i) = b'(\theta_i)$, $V_i(\boldsymbol{\beta}) =: V(\theta_i) = b''(\theta_i)$, $h_i(\boldsymbol{\beta}) = h(\eta_i) = h(\mathbf{x}_i^\top\boldsymbol{\beta})$, $h_i'(\boldsymbol{\beta}) = h'(\eta_i)$, $\varepsilon_i(\boldsymbol{\beta}) = y_i - \mu_i(\boldsymbol{\beta})$, etc. and $\mu_i = \mu_i(\boldsymbol{\beta}_0)$, $V_i = V_i(\boldsymbol{\beta}_0)$, $h_i = h(\mathbf{x}_i^\top\boldsymbol{\beta}_0)$, $h_i' = h_i'(\boldsymbol{\beta}_0)$, etc..

Verification of (B0)–(B1). Let $s_i(\boldsymbol{\beta}) = (Y_i - \mu_i(\boldsymbol{\beta}))/V_i(\boldsymbol{\beta})$. Noticing $\boldsymbol{\psi}_{ni}(\boldsymbol{\beta}) = s_i(\boldsymbol{\beta})h_i'(\boldsymbol{\beta})\mathbf{x}_i = [(Y_i - \mu(\theta_i))/V(\theta_i)]h'(\eta_i)\mathbf{x}_i$, we let

$$u_i(\boldsymbol{\beta}) = -\frac{d}{d\eta_i}(s_i h_i')(\boldsymbol{\beta}) = \Big(\frac{h_i'^2}{V_i} + \frac{V_i' h_i'^2 - V_i^2 h_i''}{V_i^3}\varepsilon_i\Big)(\boldsymbol{\beta}). \tag{3.4}$$

Then $\dot{\boldsymbol{\psi}}_{ni}(\boldsymbol{\beta}) = -u_i(\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i^\top$, whose d-th row $\dot{\boldsymbol{\psi}}_{ni,d}(\boldsymbol{\beta}) = -u_i(\boldsymbol{\beta})x_{i,d}\mathbf{x}_i^\top$ has the partial derivative matrix $\ddot{\boldsymbol{\psi}}_{ni,d}(\boldsymbol{\beta}) = -u_i'(\boldsymbol{\beta})x_{i,d}\mathbf{x}_i\mathbf{x}_i^\top$. Likewise, the e-th row of the latter, $\ddot{\boldsymbol{\psi}}_{ni,d,e}(\boldsymbol{\beta}) = -u_i'(\boldsymbol{\beta})x_{i,d}x_{i,e}\mathbf{x}_i^\top$, has the derivative $\dddot{\boldsymbol{\psi}}_{ni,d,e}(\boldsymbol{\beta}) = -u_i''(\boldsymbol{\beta})x_{i,d}x_{i,e}\mathbf{x}_i\mathbf{x}_i^\top$, where $u_i'(\boldsymbol{\beta}) = (d/d\eta_i)u_i(\boldsymbol{\beta})$ and $u_i''(\boldsymbol{\beta}) = (d^2/d^2\eta_i)u_i(\boldsymbol{\beta})$. Stacking them up respectively, we have $\ddot{\boldsymbol{\psi}}_{ni}(\boldsymbol{\beta}) = -u_i'(\boldsymbol{\beta})\mathbf{M}_i$ and $\dddot{\boldsymbol{\psi}}_{ni}(\boldsymbol{\beta}) = -u_i''(\boldsymbol{\beta})\mathbf{x}_i \otimes \mathbf{M}_i$. Hence, with $\mathbf{M}_i = \mathbf{x}_i \otimes (\mathbf{x}_i\mathbf{x}_i^\top)$,

$$\begin{aligned} &\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta}) = -\mathbf{X}^\top\mathbf{U}(\boldsymbol{\beta})\mathbf{X}, \qquad \mathbf{U}(\boldsymbol{\beta}) = \mathrm{Diag}(u_i(\boldsymbol{\beta})), \\ &\ddot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta}) = -\sum_{i=1}^{n} u_i'(\boldsymbol{\beta})\mathbf{M}_i, \qquad \dddot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta}) = -\sum_{i=1}^{n} u_i''(\boldsymbol{\beta})\mathbf{x}_i \otimes \mathbf{M}_i. \end{aligned} \tag{3.5}$$

Assume $\upsilon_n = \inf\{\min_i u_i(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^p\} > 0$ (it still holds for $\upsilon_n < 0$). Then

$$-\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{U}(\boldsymbol{\beta})\mathbf{X} \succeq \upsilon_n\mathbf{X}^\top\mathbf{X}, \quad \boldsymbol{\beta} \in \mathbb{B},$$

where $\mathbf{X}$ is the matrix consisting of rows $\mathbf{x}_i^\top$. As $\mathbb{E}(u_i|\mathbf{x}_i) = h_i'^2/V_i$, one has $\mathbf{H}_n = -\mathbb{E}(\mathbf{X}^\top\Sigma\mathbf{X}) = -\mathbf{J}_n$ with $\Sigma = \mathbb{E}(\mathbf{U}|\mathbf{X}) = \mathrm{Diag}(h_i'^2/V_i)$, and

$$\sigma_{\min}^{-1}(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\boldsymbol{\beta})) \le \sigma_{\min}^{-1}(\mathbf{H}_n)\upsilon_n^{-1}\sigma_{\min}^{-1}(\mathbf{X}^\top\mathbf{X}), \quad \boldsymbol{\beta} \in \mathbb{B},\ n \ge 1.$$

By Hölder's inequality, for $1/u_1 + 1/u_2 = 1$ with $1 \le u_1, u_2 \le \infty$ and $k \ge 1$,

$$\mathbb{E}(\sigma_{\min}^{-k}(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\boldsymbol{\beta}))) \le \sigma_{\min}^{-k}(\mathbf{H}_n) \cdot |\upsilon_n^{-1}|_{ku_1}^k \cdot |\sigma_{\min}^{-1}(\mathbf{X}^\top\mathbf{X})|_{ku_2}^k, \quad \boldsymbol{\beta} \in \mathbb{B}.$$

Thus (B1) can be established using the results in Peng (2024). If $\upsilon_n \ge b > 0$ for some constant b, then one can take $u_1 = \infty$ and $u_2 = 1$. By Lemma **??** in Peng (2024), if the rows of $\mathbf{X}$ are sub-Gaussian and isotropic, then (B0) holds.

Verification of (Bk) and (Bk^+) ($k = 2, 3, 6$). Assume $|x_{i,d}| + |u_i'(\boldsymbol{\beta})| \le m, \forall\boldsymbol{\beta}, i, d$ for some constant m. Then all $\|\ddot{\boldsymbol{\psi}}_{ni,d}(\boldsymbol{\beta})\| \le m^2\|\mathbf{x}_i\|^2$. Assume also $\mathbf{H}_n = \mathbb{E}(\dot{\boldsymbol{\Psi}}_n) = O(n)$ (so that $\vartheta_n = O(n)$). Then (B2) is met with $\mathbb{E}(\tilde{\eta}_n) = O(p^{3/2})$. Assume, furthermore, all $|u_i''(\boldsymbol{\beta})| \le m$. Then $\|\dddot{\boldsymbol{\psi}}_{ni,d,e}(\boldsymbol{\beta})\| \le m^3\|\mathbf{x}_i\|^2$, so that (B3) is met with $\mathbb{E}(\underset{\sim}{\xi}_n) = O(p^2)$.

Note that all $\ddot{\boldsymbol{\psi}}_{ni,d}(\boldsymbol{\beta})$ and $\dddot{\boldsymbol{\psi}}_{ni,d,e}(\boldsymbol{\beta})$ are symmetric matrices. While (B2) is an entry-wise condition, (B2$^+$) uses the global block information. Clearly, $\ddot{\boldsymbol{\Psi}}_{n,d}(\boldsymbol{\beta}) \preceq m^2\mathbf{X}^\top\mathbf{X} = \mathbf{H}_{n,d}, \forall d$ as

$$\ddot{\boldsymbol{\psi}}_{ni,d}(\boldsymbol{\beta}) \preceq |\ddot{\boldsymbol{\psi}}_{ni,d}(\boldsymbol{\beta})| = |u_i'(\boldsymbol{\beta})x_{i,d}|\mathbf{x}_i\mathbf{x}_i^\top \preceq m^2\mathbf{x}_i\mathbf{x}_i^\top =: \mathbf{H}_{ni,d}, \quad \boldsymbol{\beta} \in \mathbb{B}, \forall i, d.$$

By (2) and (4) of Lemma 6.1, we thus have

$$\|\ddot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta})\|_o \le \|\ddot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta})\|_{oe} \le \|\mathbf{H}_n\|_{oe}.$$

Assume, furthermore, the condition number of $\mathbf{X}$ satisfies $\kappa_n(\mathbf{X}) = O(1)$. Then $\mathbb{E}(\|\underset{\sim}{\mathbf{H}}_n\|_{oe}) = O(\sqrt{p})$, so that (B2$^+$) is established.

Analogously, (B3$^+$) is satisfied since, for $\forall\boldsymbol{\beta} \in \mathbb{B}$ and i, d, e,

$$\dddot{\boldsymbol{\psi}}_{ni,d,e}(\boldsymbol{\beta}) \preceq |\dddot{\boldsymbol{\psi}}_{ni,d}(\boldsymbol{\beta})| = |u_i''(\boldsymbol{\beta})x_{i,d}x_{i,e}|\mathbf{x}_i\mathbf{x}_i^\top \preceq m^3\mathbf{x}_i\mathbf{x}_i^\top =: \boldsymbol{\Xi}_{ni,d,e},$$

yielding all $\dddot{\boldsymbol{\Psi}}_{n,d,e}(\boldsymbol{\beta}) \preceq m^3\mathbf{X}^\top\mathbf{X} = \boldsymbol{\Xi}_{n,d,e}$, hence $\mathbb{E}(\|\underset{\sim}{\boldsymbol{\Xi}}_n\|_{oe}) = O(p)$. For the rest of the verification, see the discussions behind the assumptions.

The Bias Formulas. One caculates $\mathbf{J}_n = \mathbb{E}(\mathbf{X}^\top\Sigma\mathbf{X}) = -\mathbf{H}_n$ and

$$\mathbf{G}_n = \sum_{i=1}^n \mathbb{E}\big((2h_i'^3V_i'/V_i^3 - 3h_i'h_i''/V_i)\mathbf{M}_i\big) =: \sum_{i=1}^n \mathbb{E}(g_i\mathbf{M}_i), \quad \text{say.}$$

The d-th block of $\mathbf{G}_n$ is $\mathbf{G}_{n,d} = \sum_i \mathbb{E}(g_ix_{i,d}\mathbf{x}_i\mathbf{x}_i^\top)$, so that

$$q_{n,d} = \mathrm{Tr}(\mathbf{J}_n^{-1} \circ \mathbf{G}_{n,d}) = \sum_{i=1}^n \mathbb{E}((2h_i'V_i'/V_i^2 - 3h_i''/h_i')\mathbb{H}_{i,i}x_{i,d}), \tag{3.6}$$

where $\mathbb{H}_{i,i} = \mathbf{x}_i^\top\mathbf{J}_n^{-1}\mathbf{x}_i \cdot h_i'^2/V_i$ are the diagonal entries of

$$\mathbb{H} = \Sigma^{1/2}\mathbf{X}\big(\mathbf{X}^\top\mathbf{J}_n^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^\top\Sigma^{1/2}. \tag{3.7}$$

(When $\mathbf{x}_i$ are nonrandom, $\mathbb{H}$ is referred to as the *generalized hat matrix* in the literature). Thus $\mathbf{q}_n = \sum_i \mathbb{E}((2h_i'V_i'/V_i^2 - 3h_i''/h_i')\mathbb{H}_{i,i}\mathbf{x}_i)$. One has

$$\underset{\sim}{\mathbf{b}}_{n1} = \mathbf{J}_n^{-1}\mathbb{E}((h_i''/h_i' - V_i'h_i'/V_i^2)\mathbb{H}_{i,i}\mathbf{x}_i).$$

By Theorem 3.1 and (2.2), if $p = O(n^{2/3})$, then the first order bias is $\underset{\sim}{\mathbf{b}}_{n1}$ with the rate $O(p^{3/2}/n)$ for the remainder. Further, if $p = o(n^{4/7})$, then the bias with a sharper rate for the remainder is

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}}) = -\frac{1}{2}\mathbf{J}_n^{-1}\sum_{i=1}^n \mathbb{E}\left(\mathbb{H}_{i,i}\mathbf{x}_ih_i''/h_i'\right) + O(p^{7/2}/n^2 + p^{5/2}/n^{3/2}). \tag{3.8}$$

This can also be obtained using the simplified bias formula (2.5). For a canonical link, $h_i' = V_i$ and $h_i'' = V_i'$, so that $\Sigma = \mathrm{Diag}(V_i)$, $\mathbf{J}_n = \mathbb{E}(\mathbf{X}^\top\mathrm{Diag}(V_i)\mathbf{X})$. The bias then simplifies to

$$\mathbf{b}_{n1} = 0, \quad \underset{\sim}{\mathbf{b}}_n = -\frac{1}{2}\mathbf{J}_n^{-1}\sum_{i=1}^n \mathbb{E}\big(V_i'(\mathbf{x}_i^\top\mathbf{J}_n^{-1}\mathbf{x}_i)\mathbf{x}_i\big). \tag{3.9}$$

Theorem 3.1. *If (B0)(b), (B2) and (B6) in Proposition 2.1 are strengthened to (B0)(c), (B2$^+$) and (B6$^+$), then p and the rate for the remainder in Proposition 2.1 are improved to $p = o(\vartheta_n^{2/3})$ and $O(p^{3/2}/\vartheta_n)$, respectively.*

Furthermore, if (B0)(c) and (B3) in Theorem 2.1 (ii) are strengthened to (B0)(d) and (B3$^+$), then the results in Theorem 2.1 are improved to $p = o(\vartheta_n^{4/7})$ and $O(p^{7/2}\vartheta_n^{-2} + p^{5/2}\vartheta_n^{-3/2})$ for the remainder.

4. Biases in Regularized Regression Models

Consider the observations of the form $\mathbf{z}_{ni} = (\mathbf{x}_i, Y_i), i = 1, \ldots, n$, where each Y_i is a scalar esponse and $\mathbf{x}_i$ is a p-dimensional covariate. For large p possibly $p > n$, a form of complexity regularization is often used. The penalized estimator of $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(\mathbf{x}_i, Y_i; \boldsymbol{\beta}) + \mathrm{P}(\boldsymbol{\beta}; \boldsymbol{\lambda}) \right\},$$

where $\rho(\mathbf{x}, Y; \boldsymbol{\beta})$ is a loss function, $\mathrm{P}(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is a penality function, and $\boldsymbol{\lambda}$ is a penality parameter vector. If the penality function $\mathrm{P}(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is differentiable w.r.t. $\boldsymbol{\beta}$, then one readily caculates the partial derivative to obtain the GEE for the penalized estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. If $\mathrm{P}(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is not differentiable w.r.t. $\boldsymbol{\beta}$, then one caculates the subgrient to obtain the GEE for $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. As the ℓ_1-norm has a linear gradient, the second order gradients are zero. This fact suggests to applying the bias formula obtained in the article to the ℓ_1-penalty (or heuristically, and we conjecture that the result still holds).

The LASSO Bias in LM. Consider a high dimensional linear model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. For the quadratic loss $\rho(\mathbf{x}, Y; \boldsymbol{\beta}) = (Y - \mathbf{x}^\top\boldsymbol{\beta})^2$ and the ℓ_1-penality $\mathrm{P}(\boldsymbol{\beta}; \lambda) = \lambda\|\boldsymbol{\beta}\|_1$ with $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$, the LASSO estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ satisfies the estimating equations,

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) = n^{-1}\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \lambda\mathbf{u}(\boldsymbol{\beta}) = 0,$$

where $\mathbf{u}(\boldsymbol{\beta})$ is a subgradient of the ℓ_1 norm at $\boldsymbol{\beta}$, specifically, the jth component satisfies $u_j(\boldsymbol{\beta}) = u_j(\beta_j) \in [-1, 1]$ if $\beta_j = 0$ and $u_j(\boldsymbol{\beta}) = u_j(\beta_j) = \mathrm{sgn}(\beta_j)$ otherwise if $\beta_j \neq 0$. One calculatees $\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta}) = -n^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{H}_n$, and $\ddot{\boldsymbol{\Psi}}_{n,d}(\boldsymbol{\beta}) = 0$ for $d = 1, \ldots, p$. The LASSO estimator thus has the bias given by

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) = \lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{u}, \tag{4.1}$$

where $\mathbf{u} = \mathbf{u}(\boldsymbol{\beta}_0)$ is a subgradient at the true value $\boldsymbol{\beta}_0$ of parameter.

The LASSO Bias in GLM. The Lasso estimator is now defined by penalizing the negative log-likelihood with the ℓ_1-norm. By (3.1)–(3.2), $\rho(\mathbf{x}, y; m, \boldsymbol{\beta}) = -\theta y - b(\theta)$, where θ satisfies $\mu(\theta) = m + \mathbf{x}_1^\top\boldsymbol{\beta}_1$ with $\boldsymbol{\beta} = (m, \boldsymbol{\beta}_1^\top)^\top$ and

$\mathbf{x} = (1, \mathbf{x}_1^\top)^\top$. As the intercept m is often not penalized, the penality takes the form $\mathrm{P}(\boldsymbol{\beta}; \lambda) = \lambda \|\boldsymbol{\beta}_1\|_1$. By (3.3), the GEE for $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = (\hat{m}(\lambda), \hat{\boldsymbol{\beta}}_1(\lambda)^\top)^\top$ is

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i - \mu_i(\boldsymbol{\beta})}{V_i(\boldsymbol{\beta})} h_i'(\boldsymbol{\beta})\mathbf{x}_i - \lambda \mathbf{v}(\boldsymbol{\beta}_1) = 0, \tag{4.2}$$

where $\mathbf{v}(\boldsymbol{\beta}_1) = (0, \mathbf{u}(\boldsymbol{\beta}_1)^\top)^\top$. Similar to the bias calculation for the GLM in Section 3, both $\mathbf{H}_n$ and $\mathbf{G}_n$ are the same as those there since $\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta})$, $\ddot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta})$, $\dddot{\boldsymbol{\Psi}}_n(\boldsymbol{\beta})$ are the same as those in (3.5). Meanwhile, $\mathbf{J}_n = \mathbf{J}_{\mathrm{glm},n} + \lambda^2 \mathbf{v}^{\otimes 2}$, where $\mathbf{J}_{\mathrm{glm},n} = \mathbb{E}(\mathbf{X}^\top \Sigma \mathbf{X}) = \mathbf{H}_n$ is the $\mathbf{J}_n$ given in the GLM. One has $\mathbf{b}_{n1} = \mathbf{b}_{\mathrm{glm},n1} - \lambda \mathbf{v}$ and $q_{n,d} = q_{\mathrm{glm},n,d} + \lambda^2 q_{n,d}(\mathbf{v})$, where $q_{n,d}(\mathbf{v}) = \mathbf{v}^\top \mathbf{H}_n^{-\top} \mathbf{G}_{n,d} \mathbf{H}_n^{-1} \mathbf{v}$. By Theorem 3.1, if $p = o(n^{4/7})$, then the bias of the LASSO estimator satisfies

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) = \underset{\sim}{\mathbf{b}}_{\mathrm{glm},n} + \lambda \mathbf{J}_{\mathrm{glm},n}^{-1}(\mathbf{v} + 2^{-1}\lambda \mathbf{q}_n(\mathbf{v})) + O(p^{7/2}n^{-2} + p^{5/2}n^{-3/2}), \tag{4.3}$$

where $\underset{\sim}{\mathbf{b}}_{\mathrm{glm},n}$ is the bias for the GLM given in (3.8) or (3.9).

Biases in Penalized SIM. The response y_i and the covriate $\mathbf{x}_i$ satisfy

$$y_i = m_0(\boldsymbol{\beta}_0^\top \mathbf{x}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{4.4}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the *index parameter* which satisfies $\|\boldsymbol{\beta}_0\| = 1$ with its first component $\beta_1 > 0$ for identifiability, $m_0(x)$ is a smooth nonparametric function on $\mathbb{R}$, and $\epsilon_1 \ldots, \epsilon_n$ are i.i.d. random errors with zero mean and constant variance $\sigma_0^2 = \mathrm{Var}(\epsilon_i) > 0$. The mean function $m_0(x)$ is expressed by

$$m(x) = \boldsymbol{\delta}^\top \mathbf{B}(x), \quad x \in \mathbb{R}, \tag{4.5}$$

where $\boldsymbol{\delta} \in \mathbb{R}^d$ is an unknown vector of coefficients, and $\mathbf{B}(x)$ is a vector of basis functions. Apparently, $m(x)$ approximates $m_0(x)$ and converges to it as $d \to \infty$. From a practical viewpoint, one takes $m_0 = m$ in (4.4). Our study of the bias concerns with the case of $p + d \to \infty$. The SIM is a hybrid of parametric and nonparametric models. It generalizes LM by introducing a nonparametric link function, and extends GLM by using an unknown link.

The parameters $\boldsymbol{\beta}, \boldsymbol{\delta}$ can be estimated by using the quadratic loss $\rho(\mathbf{x}, y; \boldsymbol{\beta}, \boldsymbol{\delta}) = (y - \boldsymbol{\delta}^\top \mathbf{B}(\boldsymbol{\beta}^\top \mathbf{x}))^2$ and the penality $\mathrm{P}(\boldsymbol{\beta}, \boldsymbol{\delta}; \lambda) = \lambda P(\boldsymbol{\beta}, \boldsymbol{\delta})$ subject to the constraints $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$. As is customary, the constraints on $\boldsymbol{\beta}$ can be handled by reparametrization,

$$\boldsymbol{\beta}(\boldsymbol{\phi}) = (1, \boldsymbol{\phi}^\top)^\top / \sqrt{1 + \|\boldsymbol{\phi}\|^2}, \quad \boldsymbol{\phi} \in \mathbb{R}^{p-1}. \tag{4.6}$$

The parameters to be estimated become $\boldsymbol{\theta} = (\boldsymbol{\phi}^\top, \boldsymbol{\delta}^\top)^\top \in \mathbb{R}^{p+d-1}$ using the transformed objective,

$$Q_n(\boldsymbol{\theta}) =: \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \boldsymbol{\delta}^\top \mathbf{B}(\boldsymbol{\beta}(\boldsymbol{\phi})^\top \mathbf{x}_i)\right)^2 + \lambda P(\boldsymbol{\theta}). \tag{4.7}$$

Let $f_i(\boldsymbol{\theta}) = \boldsymbol{\delta}^\top \mathbf{B}(\mathbf{x}_i^\top \boldsymbol{\beta}(\boldsymbol{\phi}))$ and $\boldsymbol{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_n(\boldsymbol{\theta}))^\top$, and let $e_i(\boldsymbol{\theta}) = y_i - f_i(\boldsymbol{\theta})$. Then $\mathbb{E}(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$. The GEE for the penalized

estimator $\hat{\boldsymbol{\theta}}$ reads

$$\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} 2e_i(\boldsymbol{\theta})\dot{f}_i^\top(\boldsymbol{\theta}) - n\lambda \dot{P}^\top(\boldsymbol{\theta}) = 0. \tag{4.8}$$

As a result, $\mathbf{J}_n = 4\sigma^2 \dot{\mathbf{f}}^{\otimes 2} + n^2\lambda^2 \dot{P}^\top \dot{P}$.

Drop the indices for now and write $\mathbf{x} = \mathbf{x}_i, f(\boldsymbol{\theta}) = f_i(\boldsymbol{\theta})$, $\eta = \mathbf{x}^\top\boldsymbol{\beta}$ with $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\phi})$, etc. Recall that $\dot{f}(\boldsymbol{\theta}) = \partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^\top$ is a row vector and $\ddot{f}(\boldsymbol{\theta}) = \partial \dot{f}(\boldsymbol{\theta})^\top/\boldsymbol{\theta}^\top$. One has $\dot{\eta} = \mathbf{x}^\top\dot{\boldsymbol{\beta}}$, $\dot{e}(\boldsymbol{\theta}) = -\dot{f}(\boldsymbol{\theta})$, $\ddot{e}(\boldsymbol{\theta}) = -\ddot{f}(\boldsymbol{\theta})$, and

$$\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} 2(e_i\ddot{f}_i - \dot{f}_i^\top \dot{f}_i)(\boldsymbol{\theta}) - n\lambda \ddot{P}(\boldsymbol{\theta}).$$

As a result, $\mathbf{H}_n = -2\dot{\mathbf{f}}^{\otimes 2} - n\lambda\ddot{P}$. For the kth component $\Psi_{n,k}(\boldsymbol{\theta})$ of $\boldsymbol{\Psi}_n(\boldsymbol{\theta})$,

$$\ddot{\Psi}_{n,k}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} 2(\dot{f}_{i,k}\ddot{f}_i + \ddot{f}_{i,k}^\top \dot{f}_i + \dot{f}_i^\top \ddot{f}_{i,k} - e_i \dddot{f}_{i,k})(\boldsymbol{\theta}) - n\lambda \dddot{P}_k(\boldsymbol{\theta}),$$

where $\dot{f}_{i,k}$ ($\ddot{f}_{i,k}$) denotes the kth component (row) of $\dot{f}_i$ ($\ddot{f}_i$) and $\dddot{P}_k = \ddot{D}_{P,k}$ with $D_{P,k} = \dot{P}_k$. Hence

$$\mathbf{G}_{n,k} = -\sum_{i=1}^{n} 2\mathbb{E}(\dot{f}_{i,k}\ddot{f}_i + \ddot{f}_{i,k}^\top \dot{f}_i + \dot{f}_i^\top \ddot{f}_{i,k}) - n\lambda \dddot{P}_k, \quad k = 1, \ldots, d+p-1.$$

Let $\mathbf{A} = \boldsymbol{\delta}^\top \ddot{\mathbf{B}}(\eta)(\dot{\boldsymbol{\beta}}^\top \mathbf{x})^{\otimes 2} + \boldsymbol{\delta}^\top \dot{\mathbf{B}}(\eta)(\ddot{\beta}_{(1)}^\top \mathbf{x}, \cdots, \ddot{\beta}_{(p-1)}^\top \mathbf{x})^\top$, where $\ddot{\beta}_{(j)} = \dot{\mathbf{c}}_j$ with $\mathbf{c}_j$ denoting the jth column of $\dot{\boldsymbol{\beta}}$. Then

$$\dot{f}(\boldsymbol{\theta}) = (\boldsymbol{\delta}^\top \dot{\mathbf{B}}(\eta)(\mathbf{x}^\top\dot{\boldsymbol{\beta}}),\, \mathbf{B}(\eta)^\top), \quad \ddot{f}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{A} & (\dot{\boldsymbol{\beta}}^\top \mathbf{x})\dot{\mathbf{B}}(\eta)^\top \\ \dot{\mathbf{B}}(\eta)(\mathbf{x}^\top\dot{\boldsymbol{\beta}}) & \mathbf{0}_{d\times d} \end{pmatrix}.$$

Therefore, the first-order bias for $\hat{\boldsymbol{\theta}}$ is $\mathrm{Bias}_1(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\delta}}) = \mathbf{H}_n^{-1}(\mathbf{b}_{n1} - 2^{-1}\mathbf{q}_n)$, where $q_{n,k} = \mathrm{Tr}(\mathbf{H}_n^{-\top}\mathbf{G}_{n,k}\mathbf{H}_n^{-1}\mathbf{J}_n)$, and

$$\mathbf{b}_{n1} = 4\sigma^2 \sum_i \ddot{f}_i \mathbf{H}_n^{-1} \dot{f}_i^\top - 2n\lambda \sum_i \dot{f}_i^\top \dot{f}_i \mathbf{H}_n^{-1}\dot{P}^\top + 2n^2\lambda^2 \ddot{P}\mathbf{H}_n^{-1}\dot{P}^\top. \tag{4.9}$$

By Remark 2.3, the first-order bias for $\boldsymbol{\beta}_0$ is now given by $\mathrm{Bias}_1(\hat{\boldsymbol{\beta}}) = \mathbf{J}(\hat{\boldsymbol{\phi}})\mathrm{Bias}_1(\hat{\boldsymbol{\phi}})$, where $\mathbf{J}^\top(\boldsymbol{\phi}) = (-\boldsymbol{\phi},\, (1+\|\boldsymbol{\phi}\|^2)\mathbf{I} - \boldsymbol{\phi}^{\otimes 2})(1+\|\boldsymbol{\phi}\|^2)^{-3/2}$.

Let $\mathbf{J}^0, \mathbf{H}^0, \mathbf{G}^0$ denote the respective values of $n^{-1}\mathbf{J}_n, n^{-1}\mathbf{H}_n, n^{-1}\mathbf{G}_n$ when $\lambda = 0$ (unpenalized). Let $\mathbf{J}^1 = \dot{P}^\top\dot{P}$, $\mathbf{H}^2 = -\ddot{P}$ and $\mathbf{G}_k^1 = -\dddot{P}_k$. Then

$$n^{-1}\mathbf{J}_n = \mathbf{J}^0 + n\lambda^2\mathbf{J}^1, \quad n^{-1}\mathbf{H}_n = \mathbf{H}^0 + \lambda\mathbf{H}^1, \quad n^{-1}\mathbf{G}_{n,k} = \mathbf{G}_k^0 + \lambda\mathbf{G}_k^1.$$

Typically, $(n^{-1}\mathbf{H}_n)^{-1} = (\mathbf{H}^0)^{-1} + \lambda\mathbf{H}_-^1$ for some matrix $\mathbf{H}_-^1$. It then follows from the trace expression of $q_{n,k}$ that

$$q_{n,k} = q_k^0 + (n\lambda + 1)q_k^1(\lambda), \quad k = 1, \ldots, p+d-1, \tag{4.10}$$

where q_k^0 is the value of $q_{n,k}$ when $\lambda = 0$, and $q_k^1(\lambda) = \mathrm{Tr}(\lambda \mathbf{C}_1 + \lambda^2 \mathbf{C}_2 + \lambda^3 \mathbf{C}_3)$ with $\mathbf{C}_k, k = 1, 2, 3$ being matrices independen of λ. Similarly,

$$\mathbf{b}_{n1} = \mathbf{b}_1^0 + n\lambda(1 + \lambda + \lambda^2)\mathbf{b}_1^1, \tag{4.11}$$

where $\mathbf{b}_1^0$ is the value of $\mathbf{b}_{n1}$ when $\lambda = 0$, and $\mathbf{b}_1^1$ is some averaged vector independent of λ. Since each $\mathbf{C}_k$ is a product of four square matrices of dimension $p + d - 1$, it follows from (6.1) that $\mathrm{Tr}(\mathbf{C}_k) = O((p + d)^4)$, so that $q_k^1(\lambda) = (\lambda + \lambda^2 + \lambda^3)O((p + d)^4)$ for $k = 1, \ldots, d + p - 1$. Also from the expression in (4.9), one obtains $\mathbf{b}_1^1 = O((p+d)^{5/2})$. By (4.10)–(4.11), we thus derive that the first-order bias satisfies

$$\underset{\sim}{\mathbf{b}}_n = n^{-1}\mathbf{H}_0^{-1}(\mathbf{b}_1^0 - 2^{-1}\mathbf{q}_0) + \lambda(1 + \lambda + \cdots + \lambda^4)O((p + d)^{11/2}). \tag{4.12}$$

Note that the first term on the left-hand side is the first-order bias for the unpenalized estimate $\hat{\boldsymbol{\theta}}$, which is of order $O(n^{-1}(p + d)^{11/2})$. It is celebrated that the optimal rate for the penality is $\lambda = O(\sqrt{\log(p + d)/n})$ under suitable conditions, see e.g. Bickel, *et al.*(2009). Thus for the bias to be negligible, the dimension $p + d$ must grow at an extremely slow rate $(p + d)\sqrt[11]{\log(p + d)} = o(n^{1/11})$ as $p + d$ tends to infinity.

For the ridge regression $P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2$, $\dddot{P}_k(\boldsymbol{\theta}_0) = 0$ for all k. Then $\mathbf{q}_n \equiv 0$, and $\mathbf{b}_{n1} = \mathbf{b}_1^0 + n\lambda(1 + \lambda)\mathbf{b}_1^1$ from (4.11). As $\mathbf{b}_1^1 = O((p + d)^{5/2})$, we have

$$\underset{\sim}{\mathbf{b}}_n = n^{-1}\mathbf{H}_0^{-1}\mathbf{b}_1^0 + \lambda(1 + \lambda + \lambda^2)O((p + d)^{5/2}). \tag{4.13}$$

Consider $\ddot{P}(\boldsymbol{\theta}_0) = 0$. The ℓ_1-penality is in this case (loosely speaking). Clearly, $\mathbf{H}_n = -2\dot{\mathbf{f}}^{\otimes 2}$, $\mathbf{q}_n \equiv 0$, and $\mathbf{b}_{n1} = \mathbf{b}_1^0 + n\lambda\mathbf{b}_1^1$ from (4.11). Similarly, the bias satisfies

$$\underset{\sim}{\mathbf{b}}_n = n^{-1}\mathbf{H}_0^{-1}\mathbf{b}_1^0 + \lambda O((p + d)^{5/2}). \tag{4.14}$$

The Penality Function $P(\boldsymbol{\theta})$. For Bridge estimators, $P(\boldsymbol{\theta}) = \sum_{j=1}^D |\theta_j|^\gamma$, $\gamma > 0$. The cases of $\gamma = 1, 2$ are the ℓ_1/ℓ_2-penalities. Clearly, $\dot{P}(\boldsymbol{\theta})$ is a D-dimensional row vector with components $\dot{P}_k(\boldsymbol{\theta}) = \gamma|\theta_k|^{\gamma-1}\mathrm{sgn}(\theta_k)$, $k = 1, \ldots, D$; the kth row $\ddot{P}_k(\boldsymbol{\theta})$ of $\ddot{P}(\boldsymbol{\theta})$ consists of zero components except the kth component equals $\gamma(\gamma - 1)|\theta_k|^{\gamma-2}\mathrm{sgn}(\theta_k)$.

For the ridge regression $P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2$, one has $\dot{P}(\boldsymbol{\theta}) = 2\boldsymbol{\theta}^\top$, $\ddot{P}(\boldsymbol{\theta}) = 2\mathbf{I}$, and $\dddot{P}_k(\boldsymbol{\theta}) = 0$ for all k, so that $\mathbf{H}_n = -2\dot{\mathbf{f}}^{\otimes 2} - 2n\lambda\mathbf{I}$ and $\mathbf{q}_n = 0$. Hence $\mathrm{Bias}_1(\hat{\boldsymbol{\theta}}) = \mathbf{H}_n^{-1}\mathbf{b}_{n1}$, where by (4.9)

$$\mathbf{b}_{n1} = 4\sigma^2 \sum_i \ddot{f}_i \mathbf{H}_n^{-1} \dot{f}_i^\top - 4n\lambda \sum_i \dot{f}_i^\top \dot{f}_i \mathbf{H}_n^{-1}\boldsymbol{\theta}_0.$$

Examining (4.12)–(4.13), we see that the bias in this case vanishes at a much faster rate $(p+d)\sqrt[5]{\log(p + d)} = o(n^{1/5})$ than that for the biases when $\dddot{P}_k(\boldsymbol{\theta}) \neq 0$ for some k. Likewise, the bias for the Lasso estimator in (4.1) vanishes faster than that for the ℓ_2-penality (the ridge regression).

In SIM, one popular model is the P-spline in which the spline basis is the truncated powers given by $\mathbf{B}(u) = (1, u, \ldots, u^q, (u - \kappa_1)_+^q, \ldots, (u - \kappa_K)_+^q)^\top$.

The penality function is of the form of the ridge regression, $P(\boldsymbol{\theta}) = \boldsymbol{\delta}^\top \mathbf{D}\boldsymbol{\delta}$, where $\mathbf{D}$ is a diagonal matrix with the last K diagonal entries equal to 1 and the rest equal to 0. This implies that the penalty λ works to avoid overfitting by penalizing the last K components of $\boldsymbol{\delta}$. The cubic spline ($q = 3$) is often used.

5. Assumptions

In this section, we introduce the assumptions. Recall $\vartheta_n = O(\|\mathbf{H}_n\|_o)$ and $A_{n,0}$ defined in Section 1.

(B0) (a) $\mathbb{P}(A_{n,0}) = o(1)$; Furthermore, (b) $\mathbb{P}(A_{n,0}) = O(\vartheta_n^{-1})$; (c) $\mathbb{P}(A_{n,0}) = O(\vartheta_n^{-2})$; (d) $\mathbb{P}(A_{n,0}) = O(p^{-2}\vartheta_n^{-3} + \vartheta_n^{-4} + p^4\vartheta_n^{-6})$.

(B1) For $\nu > 0$, there is a constant B such that

$$\mathbb{E}\big(\sup\{\sigma_{\min}^{-\nu}\big(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\boldsymbol{\beta})\mathbf{1}[\Omega_{n,0}^c(\boldsymbol{\beta})]\big) : \boldsymbol{\beta} \in \mathbb{B}\}\big) \le B, \quad p \ge 1,\, n \ge 1,$$

where $\Omega_{n,0}(\boldsymbol{\beta}) = \{\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\boldsymbol{\beta}) \text{ is singular}\}$.

(B2) (a) Each $\boldsymbol{\psi}_{ni,d}(\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathbb{B}$ is twice continuously differentiable, and (b) there exists a sequence of rv's $\tilde{\eta}_n$ with $\mathbb{E}(\tilde{\eta}_n) = O(p^{3/2})$ such that

$$\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\boldsymbol{\beta})\|_o \le \tilde{\eta}_n, \quad \boldsymbol{\beta} \in \mathbb{B},\, n \ge 1. \tag{5.1}$$

(B3) Each $\boldsymbol{\psi}_{ni,d}(\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathbb{B}$ is thrice continuously differentiable, and there exists a sequence of rv's $\underset{\sim}{\xi}_n$ with $\mathbb{E}(\underset{\sim}{\xi}_n) = O(p^2)$ such that

$$\|\dddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\boldsymbol{\beta})\|_o \le \underset{\sim}{\xi}_n, \quad \boldsymbol{\beta} \in \mathbb{B},\, n \ge 1. \tag{5.2}$$

(B4) $\kappa(\mathbf{J}_n) =: \lambda_{\max}(\mathbf{J}_n)/\lambda_{\min}(\mathbf{J}_n) = O(1)$.

(B5) There exist constants $1 \le \nu_0', \nu_1' \le \infty$ with $1/\nu_0' + 1/\nu_1' = 1$ such that (B1) holds with $\nu = \nu_0'$ and $|\underset{\sim}{\boldsymbol{\Psi}}_n|_{\nu_1'} = o(1)$ for $p \ge 1$ and $n \ge 1$.

(B6) There exist constants $1 \le \nu_0, \nu_1 \le \infty$, $2 \le \nu_2, \nu_3 \le \infty$ with $1/\nu_0 + \cdots + 1/\nu_3 = 1$ such that (B1) holds with $\nu = 2\nu_0$, and

$$|\underset{\sim}{\boldsymbol{\Psi}}_n|_{2\nu_1} = O(p^{1/2}\vartheta_n^{-1/2}), \quad |\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n|_{2\nu_2} = O(p\vartheta_n^{-1/2}), \quad |\bar{\tilde{\eta}}_n|_{\nu_3} = O(p^{3/2}\vartheta_n^{-1/2}).$$

(B0)–(B1) and (B5)–(B6) involve conditions that concern with the singularity of the gradient $\dot{\boldsymbol{\Psi}}_n(\mathbf{B}), \mathbf{B} \in \mathbb{B}^{p\times p}$. While (B1) is a uniform integrability condition for the inverse of the random matrix $\dot{\boldsymbol{\Psi}}_n(\mathbf{B})$ over $\mathbf{B} \in \mathbb{R}^{p\times p}$ and is investigated in Peng (2024), (B0) is a mild condition as the probability that a random matrix of interest is singular often decays exponentially with n, and the rates from (a) to (d) get vanishing faster. (Bk) and (Bk$^+$), $k = 2, 3$ below are typical assumptions. All are verified for GLM in Section 3. By Lemma 6.2, (B6) implies (B5).

(B2$^+$) Same as (B2) except replacing the existence of a sequence of $\tilde{\eta}_n$ with that of $p^2 \times p$ random matrices $\underset{\sim}{\mathbf{H}}_n$ such that $\mathbb{E}(\|\underset{\sim}{\mathbf{H}}_n\|_{oe}) = O(\sqrt{p})$ and

$$\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\boldsymbol{\beta})\|_o \le \|\underset{\sim}{\mathbf{H}}_n\|_{oe}, \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

(B3$^+$) Same as (B3) except replacing the existence of a sequence of $\underset{\sim}{\xi}_n$ with that of $p^3 \times p$ random matrices $\mathbf{\Xi}_n$ such that $\mathbb{E}(\|\mathbf{\Xi}_n\|_{oe}) = O(p)$ and

$$\|\ddot{\underset{\sim}{\mathbf{\Psi}}}_n(\boldsymbol{\beta})\|_o \leq \|\underset{\sim}{\mathbf{\Xi}}_n\|_{oe}, \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

(B6$^+$) Same as (B6) except replacing $\tilde{\eta}_n$ with $\underset{\sim}{\mathbf{H}}_n$ such that $|\dot{\underset{\sim}{\mathbf{\Psi}}}_n|_{2\nu_2} = O(1/\sqrt{\vartheta_n})$ and $|\bar{\underset{\sim}{\mathbf{H}}}_n|_{\nu_3} = O(\sqrt{p/\vartheta_n})$.

Remark 5.1. *It is possible to relax (B1)–(B3) to hold a neighborhood of the true value $\boldsymbol{\beta}_0$ of parameter, although we won't pursue it in this article.*

6. Proofs

In this section, we introduce the toolkit and prove the theorems.

6.1. Notation and the generalizated MVT

Write $\mathbf{A} \otimes \mathbf{B}$ for the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^\top$, $\mathbf{A}^{-\top} = (\mathbf{A}^{-1})^\top$, and $\mathrm{Vec}(\mathrm{Diag}(\mathbf{A})) = (A_{1,1}, \ldots, A_{n,n})^\top$. Let $\mathbf{A}^{1/2}$ ($\mathbf{A}^{\top/2}$) be the left (right) square root of the positive definite matrix $\mathbf{A}$. Write $\lambda_{\max}(\mathbf{A})$ for the maximum eigenvalue of $\mathbf{A}$, etc. Write $\|\mathbf{A}\|$ for the euclidean norm and $\|\mathbf{A}\|_o$ for the operator (spectral) norm, defined by $\|\mathbf{A}\|^2 = \mathrm{Tr}(\mathbf{A}^\top\mathbf{A})$ and $\|\mathbf{A}\|_o = \lambda_{\max}^{1/2}(\mathbf{A}^\top\mathbf{A})$. The singular values $\sigma(\mathbf{A})$ of $\mathbf{A}$ are the positive square roots of the eigenvalues $\lambda(\mathbf{A}^\top\mathbf{A})$ of $\mathbf{A}^\top\mathbf{A}$. The maximum (minimum) singular value is $\sigma_{\max}(\mathbf{A}) = \lambda_{\max}^{1/2}(\mathbf{A}^\top\mathbf{A}) = \|\mathbf{A}\|_o$ ($\sigma_{\min}(\mathbf{A}) = \lambda_{\min}^{1/2}(\mathbf{A}^\top\mathbf{A})$).

For $p \times q$ and $q \times r$ matrices $\mathbf{A}, \mathbf{B}$, write $\mathbf{A} = O(\sqrt{pq})$ and $\mathbf{A} = O(\sqrt{qr})$ if $\|\mathbf{A}\| = O(\sqrt{pq})$ and $\|\mathbf{B}\| = O(\sqrt{qr})$. It then follows

$$\mathbf{A}\mathbf{B} = O(\sqrt{pq^2r}), \quad \mathrm{Tr}(\mathbf{A}\mathbf{B}) = p^2(\text{ if } p = q = r). \tag{6.1}$$

For matrices $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{p\times p}$ and $\mathbf{B}^\top = (\mathbf{B}_1, \ldots, \mathbf{B}_q)$ with $\mathbf{B}_d \in \mathbb{R}^{p\times p}$ (if $\mathbf{B}_q$ has less rows, then add zero rows), define the Hadamard-type product $\mathbf{V}\circ\mathbf{B}\circ\mathbf{W}$ to be the $q \times p$ block matrix consisting of blocks $\mathbf{V}\mathbf{B}_d\mathbf{W}, d = 1, \ldots, q$. Clearly, $\mathbf{V}\circ\mathbf{B}\circ\mathbf{W} = (\mathbf{V}\circ\mathbf{B})\mathbf{W}$, and it is not associative. Throughout it is understood that it precedes the usual multiplication. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ and a compatible block $qp \times p$ matrix $\mathbf{B}$, we have

$$\|\mathbf{u}^\top \circ \mathbf{B}\|_o \leq \|\mathbf{B}\|_o\|\mathbf{u}\|, \quad \|\mathbf{u}^\top \circ \mathbf{B} \circ \mathbf{v}\| \leq \|\mathbf{B}\|_o\|\mathbf{u}\|\,\|\mathbf{v}\|. \tag{6.2}$$

Define the *quasinorm* by

$$\|\mathbf{B}\|_{oe} = \sqrt{\|\mathbf{B}_1\|_o^2 + \cdots + \|\mathbf{B}_q\|_o^2}. \tag{6.3}$$

Note that different block partitions of $\mathbf{B}$ have different values. Define $\mathbb{T}\mathrm{r}(\mathbf{B})$ to be the block-wise trace vector of $\mathbf{B}$, that is, $\mathbb{T}\mathrm{r}(\mathbf{B}) = (\mathrm{Tr}(\mathbf{B}_1), \ldots, \mathrm{Tr}(\mathbf{B}_q))^\top$.

Let $|\mathbf{A}|$ stand for the absolute value of $\mathbf{A}$, that is, $|\mathbf{A}|$ is the unique symmetric semipositive definite matrix $(\mathbf{A}^\top\mathbf{A})^{1/2}$. For a matrix $\mathbf{A}$ consisting of blocks $\mathbf{A}_k$, define the (blockwise) absolute value $|\mathbf{A}|$ to be the matrix consisting of blocks $|\mathbf{A}_k|$. We summarize some useful facts from textbooks below.

Lemma 6.1. *(1)* $\|\mathbf{A}_1+\cdots+\mathbf{A}_q\|_{oe}\le\sqrt{q}(\|\mathbf{A}_1\|_{oe}+\cdots+\|\mathbf{A}_q\|_{oe})$.
(2) $\|\mathbf{B}\|_o\le\|\mathbf{B}\|_{oe}\le\|\mathbf{B}\|$.
(3)(Absolute) $\||\mathbf{A}|\|_{oe}=\|\mathbf{A}\|_{oe}$ *and* $\||\mathbf{A}|\|_o=\|\mathbf{A}\|_o$.
(4) (Monotone) If $|\mathbf{A}_k|\preceq|\mathbf{B}_k|,\forall k$ *(blockwise), then* $\||\mathbf{A}|\|_{oe}\le\||\mathbf{B}|\|_{oe}$. *Thus, if all the blocks* $\mathbf{A}_k,\mathbf{B}_k$ *are symmetric, then* $\|\mathbf{A}\|_{oe}\le\|\mathbf{B}\|_{oe}$.
(5) If $|\mathbf{A}_k|^2\preceq|\mathbf{B}_k|^2,\forall k$, *then* $|\mathbf{A}_k|\preceq|\mathbf{B}_k|$ *and* $\||\mathbf{A}|\|_o\le\|\mathbf{B}\|_o$.

For a random matrix $\mathbf{X}$, define $|\mathbf{X}|_p=(\mathbb{E}(\|\mathbf{X}\|_o^p))^{1/p}$, $p>0$. Note that if $|\mathbf{X}|_t=O(a_n)$ then $|\mathbf{X}|_s=O(a_n)$ for $0\le s\le t$ and $a_n\ge 0$ by the moment inequality. This, Lemma 6.2 and Hölder's inequality for a few rv's shall be frequently used in our proofs. Here we shall state the inequality for three rv's X,Y,Z: for $1\le a,b,c\le\infty$ with $1/a+1/b+1/c=1$,

$$\mathbb{E}(|XYZ|)\le|X|_a\cdot|Y|_b\cdot|Z|_c. \tag{6.4}$$

Lemma 6.2. *If* $|X|_a+|Y|_b+|Z|_c<\infty$ *for* $1\le a,b,c\le\infty$ *with* $1/a+1/b+1/c=1$, *then there exist positive* a',b' *with* $a'<a,b'<b$ *and* $1/a'+1/b'=1$ *such that* $E(|XY|)\le|X|_{a'}|X|_{b'}\le|X|_a|X|_b$. *As a result,* $E(|XY|)+E(|XZ|)+E(|YZ|)+E(|XYZ|)<\infty$.

MVT. For $\boldsymbol{\psi}:\mathbb{R}^p\to\mathbb{R}^q$, define $\dot{\boldsymbol{\psi}}:\mathbb{R}^{q\times p}\to\mathbb{R}^{q\times p}$ to be the $q\times p$ matrix $\dot{\boldsymbol{\psi}}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_q)$ with the d-th row equal to the partial derivative $\dot{\psi}_d(\boldsymbol{\beta}_d),\boldsymbol{\beta}_d\in\mathbb{R}^p$. Define $\ddot{\boldsymbol{\psi}}:\mathbb{R}^{qp\times p}\to\mathbb{R}^{qp\times p}$ to be the matrix consisting of stacking q $p\times p$ partial derivative matrices $\ddot{\psi}_d(\boldsymbol{\beta}_{d,1},\ldots,\boldsymbol{\beta}_{d,p})$. Similarly, define $\boldsymbol{\psi}^{(k)}:\mathbb{R}^{(qp^{k-2})p\times p}\to\mathbb{R}^{(qp^{k-2})p\times p}$ for $k\ge 2$. Let $\mathbf{f}:\mathbb{R}^p\to\mathbb{R}^q$ be k-th continuously differentiable. For $\mathbf{x}_0,\mathbf{t}\in\mathbb{R}^p$, there exists $\widetilde{\mathbf{X}}_k\in\mathbb{R}^{qp^{k-1}\times p}$ (whose row vectors) lying in $\mathbf{x}_0$ and $\mathbf{x}_0+\mathbf{t}$ such that

$$\mathbf{f}^{(k)}(\mathbf{x}_0+\mathbf{t})=\mathbf{f}^{(k)}(\mathbf{x}_0)+\mathbf{t}^\top\circ\mathbf{f}^{(k+1)}(\widetilde{\mathbf{X}}_k),\quad k\ge 0. \tag{6.5}$$

The multivariate MVT holds in the Laplacian form. Stacking up the partial derivative matrices leads to the integral MVT,

$$\mathbf{f}^{(k)}(\mathbf{x}_0+\mathbf{t})=\mathbf{f}^{(k)}(\mathbf{x}_0)+\mathbf{t}^\top\circ\int_0^1\mathbf{f}^{(k+1)}(\mathbf{x}_0+u\mathbf{t})\,du,\quad k\ge 0. \tag{6.6}$$

Clearly, (6.5)–(6.6) are equivalent, showing that the boundedness of $\mathbf{f}^{(k+1)}(\mathbf{x})$ implies that of $\mathbf{f}^{(k+1)}(\widetilde{\mathbf{X}}_k)$. This fact is repeatedly used without reference to it.

6.2. Proof of Proposition 2.1 and Theorem 2.1

The proof of Theorem 3.1 is similar to Theorem 2.1 and is omitted.

The main idea of the proof is as follows: take the expected values across Eqt (1.3) or (6.28) and solve for the bias formula; solve the same equation for $\hat{\boldsymbol{\beta}}$ and substitute it in the bias; identify the dominate terms and calculate tediously the rates of the remainders. Note we shall repeatedly use the moment-type inequalities in Lemma 6.2 without explicit reference.

PROOF (of Proposition 2.1). Let $A_{n,1} = \{\dot{\boldsymbol{\Psi}}_n(\tilde{\mathbf{B}}_1) \text{ is singular}\}$. Then $A_{n,1} \subset A_{n,0}$. Recall the definition of $\hat{\boldsymbol{\beta}}$ in (1.3) – (1.4) associated with $\Omega_{n,0}$. Setting $\mathbf{l} = -\underset{\sim}{\boldsymbol{\Psi}}_n$, we obtain from (1.3) an important decomposition,

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\tilde{\mathbf{B}}_1)^{-1}\mathbf{l} \cdot \mathbf{1}[A_{n,1}^c] + (\mathbf{b} - \boldsymbol{\beta}_0)\mathbf{1}[A_{n,1}] \\ &=: (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_+ + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_0, \quad \text{say.} \end{aligned} \tag{6.7}$$

Taking expectation across (1.3) and using $\mathbb{E}(\boldsymbol{\Psi}_n) = 0$, we obtain

$$\text{Bias}(\hat{\boldsymbol{\beta}}_n) = \mathbb{E}\big((\mathbf{I}_p - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big) + \underset{\sim}{\boldsymbol{\Delta}}, \tag{6.8}$$

$$\underset{\sim}{\boldsymbol{\Delta}} = \mathbb{E}\big((\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\tilde{\mathbf{B}}_1))(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big). \tag{6.9}$$

Again from (1.3) we derive the representation,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = -\underset{\sim}{\boldsymbol{\Psi}}_n + \boldsymbol{\alpha}_1 =: \mathbf{l} + \underset{\sim}{\boldsymbol{\alpha}}_1, \tag{6.10}$$

where $\underset{\sim}{\boldsymbol{\alpha}}_1 = (\mathbf{I}_p - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\tilde{\mathbf{B}}_1))(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. Substitution of (6.10) in the expectation of (6.8) yields

$$\mathbb{E}\big((\mathbf{I}_p - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big) = -\mathbb{E}\big(\bar{\dot{\underset{\sim}{\boldsymbol{\Psi}}}}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big) = \mathbb{E}\big(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n \underset{\sim}{\boldsymbol{\Psi}}_n\big) + \underset{\sim}{\boldsymbol{\delta}}_1, \tag{6.11}$$

where $\underset{\sim}{\boldsymbol{\delta}}_1 = -\mathbb{E}(\bar{\dot{\underset{\sim}{\boldsymbol{\Psi}}}}_n \underset{\sim}{\boldsymbol{\alpha}}_1)$. Noting $\mathbb{E}(\boldsymbol{\psi}_{ni}) = 0$ for all i and by independence, we substitute (6.11) in (6.8) and get

$$\text{Bias}(\hat{\boldsymbol{\beta}}) =: \sum_{i=1}^{n} \mathbb{E}\big(\dot{\underset{\sim}{\boldsymbol{\psi}}}_{ni} \underset{\sim}{\boldsymbol{\psi}}_{ni}\big) + \mathbf{r}_n, \tag{6.12}$$

where $\mathbf{r}_n = \underset{\sim}{\boldsymbol{\delta}}_1 + \underset{\sim}{\boldsymbol{\Delta}}$. Corresponding to the decomposition (6.7), we write $\underset{\sim}{\boldsymbol{\alpha}}_1 = \underset{\sim}{\boldsymbol{\alpha}}_{1+} + \underset{\sim}{\boldsymbol{\alpha}}_{10}$, resulting in $\underset{\sim}{\boldsymbol{\delta}}_1 = \underset{\sim}{\boldsymbol{\delta}}_{1+} + \underset{\sim}{\boldsymbol{\delta}}_{10}$. Analogously,

$$\underset{\sim}{\boldsymbol{\Delta}} = \underset{\sim}{\boldsymbol{\Delta}}_+ + \underset{\sim}{\boldsymbol{\Delta}}_0. \tag{6.13}$$

It is shown below

$$\|\underset{\sim}{\boldsymbol{\Delta}}_+\| = O\big(p^{5/2}/\vartheta_n\big), \tag{6.14}$$

$$\|\underset{\sim}{\boldsymbol{\delta}}_{1+}\| = O(p^{7/2}/\vartheta_n^{3/2}), \tag{6.15}$$

$$\|\underset{\sim}{\boldsymbol{\delta}}_{10}\| = \mathbb{P}^{1/2}(A_{n,1})O(p^{7/2}/\vartheta_n^{1/2}), \tag{6.16}$$

$$\|\underset{\sim}{\boldsymbol{\Delta}}_0\| = \mathbb{P}^{1/2}(A_{n,1})O\big(p^{5/2}/\vartheta_n^{1/2}\big) + \mathbb{P}(A_{n,1})O\big(p^{5/2}\big). \tag{6.17}$$

To determine the magnitude of p relative to n (via ϑ_n), we first find the maximum magnitude of p such that (6.14) – (6.15) converge to zero as p and n tend

to infinity; then determine the rate r_n (the slowest among all the rates); finally, determine the rate of the singularity probability so that $\mathbb{P}(A_{n,1}) = O(r_n)$. In this case, $p = o(\vartheta_n^{2/5})$, $r_n = O(p^{5/2}/\vartheta_n)$. Setting $\|\underset{\sim}{\boldsymbol{\delta}}_{10}\| \asymp r_n$, $\|\underset{\sim}{\boldsymbol{\Delta}}_0\| \asymp r_n$, we find $\mathbb{P}(A_{n,1}) \asymp \vartheta_n^{-1}$. By (B0)(b), we thus prove $\mathbf{r}_n = O(p^{5/2}/\vartheta_n)$ by (6.12).

To prove (6.15), we write $\underset{\sim}{\boldsymbol{\alpha}}_{1+} = \underset{\sim}{\boldsymbol{\alpha}}_{1+a} + \underset{\sim}{\boldsymbol{\alpha}}_{1+b}$, where

$$\underset{\sim}{\boldsymbol{\alpha}}_{1+a} = (\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\tilde{\mathbf{B}}_1))(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_+, \quad \underset{\sim}{\boldsymbol{\alpha}}_{1+b} = (\mathbf{I}_p - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_+. \tag{6.18}$$

Thus $\underset{\sim}{\boldsymbol{\delta}}_{1+} = \underset{\sim}{\boldsymbol{\delta}}_{1+a} + \underset{\sim}{\boldsymbol{\delta}}_{1+b}$. Let $s_n = \sigma_{\min}(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\tilde{\mathbf{B}}_1)\mathbf{1}[\Omega^c_{n,0}(\tilde{\mathbf{B}}_1)])$. Then

$$\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_+\| \le s_n^{-1}\|\mathbf{l}\|\mathbf{1}[A^c_{n,1}]. \tag{6.19}$$

By Hölder's inequality for three rv's in (6.4), we now get

$$\|\underset{\sim}{\boldsymbol{\delta}}_{1+b}\| \le \mathbb{E}\big(s_n^{-1}\|\mathbf{l}\|\,\big\|\bar{\dot{\underset{\sim}{\boldsymbol{\Psi}}}}_n\big\|_o^2\big) = O(p^{5/2}/\vartheta_n^{3/2}), \quad \text{by (B6).}$$

Thus (6.15) follows from

$$\|\underset{\sim}{\boldsymbol{\delta}}_{1+a}\| = O(p^{7/2}/\vartheta_n^{3/2}), \tag{6.20}$$

which is shown next. To this end, (B2)(a) allows us to apply (6.5) with $k = 2$, there being $\widetilde{\mathbf{B}}_2$ in between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and get on $A^c_{n,1}$ by (B2)(b) that

$$\begin{aligned}&\|\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_1) - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n\|_o \le \|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_2)\|_o\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \le \tilde{\eta}_n\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \\ &= \big(\bar{\tilde{\eta}}_n + \mathbb{E}(\tilde{\eta}_n)\big)s_n^{-1}\|\mathbf{l}\| =: a_1 + a_2, \quad \text{say.}\end{aligned} \tag{6.21}$$

Corresponding to the last sum, we bound $\underset{\sim}{\boldsymbol{\delta}}_{1+a} \le \mathbf{d}_1 + \mathbf{d}_2$, and obtain

$$\begin{aligned}\|\mathbf{d}_1\| &\le \mathbb{E}\big(s_n^{-2}\|\mathbf{l}\|^2\big\|\bar{\dot{\underset{\sim}{\boldsymbol{\Psi}}}}_n\big\|_o|\bar{\tilde{\eta}}_n|\big) \\ &\le |s_n^{-1}|^2_{2\nu_0}|\mathbf{l}|^2_{2\nu_1}\big|\bar{\dot{\underset{\sim}{\boldsymbol{\Psi}}}}_n\big|_{\nu_2}|\bar{\tilde{\eta}}_n|_{\nu_3} = O(p^{7/2}/\vartheta_n^2), \text{ by (B6),} \qquad (6.22)\\ \|\mathbf{d}_2\| &\le \mathbb{E}\big(s_n^{-2}\|\mathbf{l}\|^2\big\|\bar{\dot{\underset{\sim}{\boldsymbol{\Psi}}}}_n\big\|_o\big)\mathbb{E}(\tilde{\eta}_n) = O(p^{7/2}/\vartheta_n^{3/2}), \text{ by (B2)\&(B6).} \qquad (6.23)\end{aligned}$$

These prove (6.20). Using (6.21), we also have

$$\|\underset{\sim}{\boldsymbol{\Delta}}_+\| \le \mathbb{E}\big((|a_1| + |a_2|)s_n^{-1}\|\mathbf{l}\|\big) =: b_1 + b_2.$$

Thus (6.14) follows from

$$\begin{aligned}b_1 &\le \mathbb{E}(s_n^{-2}\|\mathbf{l}\|^2|\bar{\tilde{\eta}}|) \le |s_n^{-1}|^2_{2\nu_0}|\mathbf{l}|^2_{2\nu_1}|\bar{\tilde{\eta}}|_{\nu_3} = O(p^{5/2}/\vartheta_n^{3/2}), \text{ by (B6),} \\ b_2 &\le \mathbb{E}(s_n^{-2}\|\mathbf{l}\|^2)\mathbb{E}(\tilde{\eta}_n) = O(p^{5/2}/\vartheta_n), \quad \text{by (B2) \& (B6).}\end{aligned}$$

To show (6.16) – (6.17), analogous to (6.18) we write $\underset{\sim}{\boldsymbol{\alpha}}_{10} = \mathbf{a}_a + \mathbf{a}_b$, where

$$\mathbf{a}_a = (\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\tilde{\mathbf{B}}_1))\mathbf{1}[A_{n,1}](\mathbf{b} - \boldsymbol{\beta}_0), \ \mathbf{a}_b = (\mathbf{I}_p - \dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\mathbf{1}[A_{n,1}](\mathbf{b} - \boldsymbol{\beta}_0).$$

Thereby $\underset{\sim}{\boldsymbol{\delta}}_{10} = \mathbf{d}_a + \mathbf{d}_b$. Similar to (6.20), we have by (R9),

$$\|\mathbf{d}_b\| \le \mathbb{P}^{1/2}(A_{n,1})\mathbb{E}^{1/2}\big(\big\|\bar{\underset{\sim}{\dot{\Psi}}}_n\big\|_o^4\big)\|\mathbf{b}-\boldsymbol{\beta}_0\| = \mathbb{P}^{1/2}(A_{n,1})O(p^{5/2}/\vartheta_n). \quad (6.24)$$

Thus (6.16) follows from

$$\|\mathbf{d}_a\| = \mathbb{P}^{1/2}(A_{n,1})O(p^{7/2}/\sqrt{\vartheta_n}), \quad (6.25)$$

which is shown next. Similar to (6.21), we have on $A_{n,1}^c$,

$$\|\dot{\underset{\sim}{\Psi}}_n(\widetilde{\mathbf{B}}_1) - \dot{\underset{\sim}{\Psi}}_n\|_o \le \big(\bar{\tilde{\eta}}_n + \mathbb{E}(\tilde{\eta}_n)\big)\|\mathbf{b}-\boldsymbol{\beta}_0\| =: A_1 + A_2. \quad (6.26)$$

Correspondingly, $\|\mathbf{d}_a\| \le \|\mathbf{d}_{a1}\| + \|\mathbf{d}_{a2}\|$. Similar to (6.22),

$$\begin{aligned}
\|\mathbf{d}_{a1}\| &\le \mathbb{P}^{1/2}(A_{n,1})\mathbb{E}^{1/2}\big(\big\|\bar{\underset{\sim}{\dot{\Psi}}}_n\big\|_o^2|\bar{\tilde{\eta}}|^2\big)\|\mathbf{b}-\boldsymbol{\beta}_0\|^2 = \mathbb{P}^{1/2}(A_{n,1})O(p^{7/2}/\vartheta_n),\\
\|\mathbf{d}_{a2}\| &\le \mathbb{P}^{1/2}(A_{n,1})\mathbb{E}(\tilde{\eta}_n)\mathbb{E}^{1/2}\big(\big\|\bar{\underset{\sim}{\dot{\Psi}}}_n\big\|_o^2\big)\|\mathbf{b}-\boldsymbol{\beta}_0\|^2\\
&+ \mathbb{P}^{1/2}(A_{n,1})O(p^{7/2}/\sqrt{\vartheta_n}), \quad \text{by (B2) \& (B6).}
\end{aligned}$$

These prove (6.25). Recalling $\boldsymbol{\Delta}_0$ in (6.13) and using (6.26), we get

$$\|\underset{\sim}{\boldsymbol{\Delta}}_0\| \le \mathbb{E}\big((A_1+A_2)\big)\|\mathbf{b}-\boldsymbol{\beta}_0\| =: D_{01} + D_{02}.$$

Thus the desired (6.17) follows from

$$\begin{aligned}
D_{01} &\le \mathbb{P}^{1/2}(A_{n,1})\mathbb{E}^{1/2}(|\bar{\tilde{\eta}}|^2)\|\mathbf{b}-\boldsymbol{\beta}_0\|^2 = \mathbb{P}^{1/2}(A_{n,1})O(p^{5/2}/\sqrt{\vartheta_n}),\\
D_{02} &\le \mathbb{P}(A_{n,1})\mathbb{E}(\tilde{\eta}_n)\|\mathbf{b}-\boldsymbol{\beta}_0\|^2 = \mathbb{P}(A_{n,1})O(p^{5/2}). \qquad \square
\end{aligned}$$

Proof (of Theorem 2.1). By Hölder's, the main term in (6.7) satisfies

$$\mathbb{E}(\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)_+\|) \le |\dot{\underset{\sim}{\Psi}}_n(\tilde{\mathbf{B}}_1)^{-1}\mathbf{1}[A_{n,1}^c]|_{\nu_0'} \cdot |\mathbf{l}|_{\nu_1'},$$

where ν_0', ν_1' are given in (B5), by which the above product is $O(1)$ uniformly in p, n. Since the other term in (6.7) is obviously bounded, it follows $\mathrm{Bias}(\hat{\boldsymbol{\beta}}) = O(1)$ uniformly in p, n. As (B0)(a) and (B5) imply that the product is $o(1)$, we prove part (i) of Theorem 2.1.

The rest of the proof is similar to Proposition 2.1. By (B2), we expand $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}_n) = 0$ at $\boldsymbol{\beta}_0$,

$$0 = \boldsymbol{\Psi}_n + \dot{\underset{\sim}{\Psi}}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + 1/2\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \circ \ddot{\underset{\sim}{\Psi}}_n(\widetilde{\mathbf{B}}_2) \circ (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0), \quad (6.27)$$

where $\widetilde{\mathbf{B}}_2$ lies in $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. Taking expectation across the equality yields

$$\begin{aligned}
0 &= \mathbb{E}\big(\dot{\underset{\sim}{\Psi}}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\big) + 1/2\,\mathbb{E}\big((\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \circ \ddot{\underset{\sim}{\Psi}}_n(\widetilde{\mathbf{B}}_2) \circ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big)\\
&= \mathbb{E}\big(\bar{\underset{\sim}{\dot{\Psi}}}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big) + \mathbb{E}(\dot{\underset{\sim}{\Psi}}_n)\mathrm{Bias}(\hat{\boldsymbol{\beta}}_n) + \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2, \quad \text{where} \quad (6.28)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Delta}_1 &= 1/2\,\mathbb{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \circ \ddot{\underset{\sim}{\Psi}}_n \circ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big),\\
\boldsymbol{\Delta}_2 &= 1/2\,\mathbb{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \circ \big(\ddot{\underset{\sim}{\Psi}}_n(\widetilde{\mathbf{B}}_2) - \ddot{\underset{\sim}{\Psi}}_n\big) \circ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big).
\end{aligned} \quad (6.29)$$

By (6.27) again and analogous to (6.10) but with a higher order expansion,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathbf{l} + \underset{\sim}{\boldsymbol{\alpha}}_2, \quad \text{say}, \tag{6.30}$$

where $\underset{\sim}{\boldsymbol{\alpha}}_2 = \underset{\sim}{\boldsymbol{\alpha}}_{21} + \underset{\sim}{\boldsymbol{\alpha}}_{22}$ with

$$\underset{\sim}{\boldsymbol{\alpha}}_{21} = \bar{\underset{\sim}{\dot{\boldsymbol{\Psi}}}}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{6.31}$$

$$\underset{\sim}{\boldsymbol{\alpha}}_{22} = -1/2\,\mathbf{H}_n^{-1}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \circ \ddot{\boldsymbol{\Psi}}_n(\widetilde{\mathbf{B}}_2) \circ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big). \tag{6.32}$$

Write $\Delta_1 = \Delta_{11} + \Delta_{12}$, where

$$\boldsymbol{\Delta}_{11} = 1/2\,\mathbb{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \circ \bar{\ddot{\boldsymbol{\Psi}}}_n \circ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big), \tag{6.33}$$

$$\boldsymbol{\Delta}_{12} = 1/2\,\mathbb{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \circ \mathbb{E}\big(\ddot{\boldsymbol{\Psi}}_n\big) \circ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big). \tag{6.34}$$

Plugging (6.30) in Δ_{12}, we get

$$\boldsymbol{\Delta}_{12} = 1/2\,\mathbb{E}\big(\mathbf{l}^\top \circ \mathbb{E}\big(\ddot{\boldsymbol{\Psi}}_n\big) \circ \mathbf{l}\big) + \boldsymbol{\delta}_2, \tag{6.35}$$

where $\boldsymbol{\delta}_2 = \mathbb{E}\big(\underset{\sim}{\boldsymbol{\alpha}}_2^\top \circ \mathbb{E}(\ddot{\boldsymbol{\Psi}}_n) \circ \mathbf{l} + 1/2\,\underset{\sim}{\boldsymbol{\alpha}}_2^\top \circ \mathbb{E}(\ddot{\boldsymbol{\Psi}}_n) \circ \underset{\sim}{\boldsymbol{\alpha}}_2\big)$. Using independence and $\mathbb{E}(\underset{\sim}{\boldsymbol{\psi}}_{ni}) = 0$ for all i, we simplify

$$\boldsymbol{\Delta}_{12} = \frac{1}{2}\sum_{i,j=1}^{n} \mathbb{E}\big(\underset{\sim}{\boldsymbol{\psi}}_{ni}^\top \circ \mathbb{E}\big(\ddot{\boldsymbol{\Psi}}_n\big) \circ \underset{\sim}{\boldsymbol{\psi}}_{nj}\big) + \boldsymbol{\delta}_2 =: \frac{1}{2}\sum_{i=1}^{n} \mathbf{b}_{nii} + \boldsymbol{\delta}_2, \tag{6.36}$$

Substituting (6.11) and (6.36) in (6.28), we arrive at the expansion,

$$\text{Bias}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\big(\mathbb{E}(\dot{\underset{\sim}{\boldsymbol{\psi}}}_{ni}\underset{\sim}{\boldsymbol{\psi}}_{ni}) - \frac{1}{2}\underset{\sim}{\mathbf{b}}_{nii}\big) + \underset{\sim}{\boldsymbol{\delta}}_1 - \underset{\sim}{\boldsymbol{\delta}}_2 - \underset{\sim}{\boldsymbol{\Delta}}_{11} - \underset{\sim}{\boldsymbol{\Delta}}_2. \tag{6.37}$$

We now derive the rates for the remainders. Using the decomposition in (6.7), Hölder's inequality with indices $(2, 6, 3)$ and Cauchy,

$$\begin{aligned}
\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{21}\|^2) &= \mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{21+}\|^2) + \mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{210}\|^2) \\
&\leq \mathbb{E}(\|\bar{\underset{\sim}{\dot{\boldsymbol{\Psi}}}}_n s_n^{-1}\mathbf{l}\|^2) + \mathbb{E}(\|\bar{\underset{\sim}{\dot{\boldsymbol{\Psi}}}}_n\|_o^2 \mathbf{1}[A_{n,1}])\|\mathbf{b} - \boldsymbol{\beta}_0\|^2 \\
&\leq |\bar{\underset{\sim}{\dot{\boldsymbol{\Psi}}}}_n|_4^2 \cdot |s_n^{-1}|_{12}^2 \cdot |\mathbf{l}|_6^2 + |\bar{\underset{\sim}{\dot{\boldsymbol{\Psi}}}}_n|_4^2 \cdot \mathbb{P}^{1/2}(A_{n,1})\|\mathbf{b} - \boldsymbol{\beta}_0\|^2 \\
&= |s_n^{-1}|_{12}^2 O(p^3/\vartheta_n^2) + \mathbb{P}^{1/2}(A_{n,1})O(p^3/\vartheta_n).
\end{aligned} \tag{6.38}$$

The decomposition also yields $\underset{\sim}{\boldsymbol{\alpha}}_{22} = \underset{\sim}{\boldsymbol{\alpha}}_{22+} + \underset{\sim}{\boldsymbol{\alpha}}_{220}$, $\underset{\sim}{\boldsymbol{\Delta}}_{11} = \underset{\sim}{\boldsymbol{\Delta}}_{11+} + \underset{\sim}{\boldsymbol{\Delta}}_{110}$ and $\underset{\sim}{\boldsymbol{\Delta}}_2 = \underset{\sim}{\boldsymbol{\Delta}}_{2+} + \underset{\sim}{\boldsymbol{\Delta}}_{20}$. Recalling $\kappa_n = \kappa(\mathbf{H}_n)$, we show bellow

$$\|\underset{\sim}{\boldsymbol{\Delta}}_{11+}\| = \kappa_n |s_n^{-1}|_8^2 O(p^{5/2}/\vartheta_n^{3/2}), \tag{6.39}$$

$$\|\underset{\sim}{\boldsymbol{\Delta}}_{110}\| = \kappa_n \mathbb{P}^{1/2}(A_{n,1})O(p^{5/2}/\sqrt{\vartheta_n}), \tag{6.40}$$

$$\|\underset{\sim}{\boldsymbol{\Delta}}_{2+}\| = \kappa_n |s_n^{-1}|_8^3 O(p^{7/2}/\vartheta_n^{3/2}), \tag{6.41}$$

$$\|\underset{\sim}{\boldsymbol{\Delta}}_{20}\| = \kappa_n\big(\mathbb{P}^{1/2}(A_{n,1})O(p^{7/2}/\sqrt{\vartheta_n}) + \mathbb{P}(A_{n,1})O(p^{7/2})\big), \tag{6.42}$$

$$\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{22+}\|^2) = \kappa_n^2|s_n^{-1}|_{16}^4 O(p^5/\vartheta_n^2), \tag{6.43}$$

$$\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{220}\|^2) = \kappa_n^2\big(\mathbb{P}^{1/2}(A_{n,1})O(p^5/\vartheta_n) + \mathbb{P}(A_{n,1})O(p^5)\big). \tag{6.44}$$

By (6.38) and (6.43) and using $\|\underset{\sim}{\boldsymbol{\alpha}}_{2+}\|^2 \le 2\|\underset{\sim}{\boldsymbol{\alpha}}_{21+}\|^2 + 2\|\underset{\sim}{\boldsymbol{\alpha}}_{22+}\|^2$, we get

$$\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{2+}\|^2) \le |s_n^{-1}|_{12}^2 O(p^3/\vartheta_n^2) + \kappa_n^2|s_n^{-1}|_{16}^4 O(p^5/\vartheta_n^2) = \kappa_n^2|s_n^{-1}|_{16}^4 O(p^5/\vartheta_n^2).$$

Similarly, by (6.38) and (6.44),

$$\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{20}\|^2) \le \kappa_n^2\big(\mathbb{P}^{1/2}(A_{n,1})O(p^5/\vartheta_n) + \mathbb{P}(A_{n,1})O(p^5)\big).$$

One has $\boldsymbol{\delta}_2 = \boldsymbol{\delta}_{2+} + \boldsymbol{\delta}_{20}$, where

$$\boldsymbol{\delta}_{2s} = \mathbb{E}\big(\underset{\sim}{\boldsymbol{\alpha}}_{2s}^\top \circ \mathbb{E}(\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n) \circ \mathbf{l}_s + 1/2\,\underset{\sim}{\boldsymbol{\alpha}}_{2s}^\top \circ \mathbb{E}(\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n) \circ \underset{\sim}{\boldsymbol{\alpha}}_{2s}\big), \quad s = +,\, 0.$$

By (6.2) and Cauchy,

$$\begin{aligned} \|\boldsymbol{\delta}_{2s}\| &\le \|\mathbb{E}(\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\|_o\big(\mathbb{E}\big(\|\underset{\sim}{\boldsymbol{\alpha}}_{2s}\| \cdot \|\mathbf{l}\|\big) + 1/2\,\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{2s}\|^2)\big) \\ &\le \|\mathbb{E}(\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\|_o\big(|\underset{\sim}{\boldsymbol{\alpha}}_{2s}|_2 \cdot |\mathbf{l}|_2 + 1/2\,|\underset{\sim}{\boldsymbol{\alpha}}_{2s}|_2^2\big), \quad s = +, 0. \end{aligned} \tag{6.45}$$

Noting $\|\mathbb{E}(\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\|_o = O(p^{3/2})$, it thus follows

$$\begin{aligned} &\|\underset{\sim}{\boldsymbol{\delta}}_{2+}\| = \kappa_n^2|s_n^{-1}|_{16}^4 O\big(p^{13/2}/\vartheta_n^2 + p^{9/2}/\vartheta_n^{3/2}\big) =: R_+, \\ &\|\underset{\sim}{\boldsymbol{\delta}}_{20}\| = \mathbb{P}^{1/4}(A_{n,1})\kappa_n O(p^{9/2}/\vartheta_n) + \mathbb{P}(A_{n,1})\kappa_n^2 O(p^{13/2}) \\ &\quad + \mathbb{P}^{1/2}(A_{n,1})\big(\kappa_n O(p^{9/2}/\sqrt{\vartheta_n}) + \kappa_n^2 O(p^{13/2}/\vartheta_n)\big) =: R_0. \end{aligned} \tag{6.46}$$

By (6.15), (6.39), (6.41) and (6.16), (6.40), (6.42), we derive

$$\begin{aligned} &\|\underset{\sim}{\boldsymbol{\delta}}_{1+}\| + \|\underset{\sim}{\boldsymbol{\delta}}_{2+}\| + \|\underset{\sim}{\boldsymbol{\Delta}}_{11+}\| + \|\underset{\sim}{\boldsymbol{\Delta}}_{2+}\| =: R_+, \\ &\|\underset{\sim}{\boldsymbol{\delta}}_{10}\| + \|\underset{\sim}{\boldsymbol{\delta}}_{20}\| + \|\underset{\sim}{\boldsymbol{\Delta}}_{110}\| + \|\underset{\sim}{\boldsymbol{\Delta}}_{20}\| =: R_0. \end{aligned}$$

We shall calculate the rate r_n using the method described in Proposition 2.1. Note first that the remainder in (6.37) is equal to $R_+ + R_0$. As $|s_n^{-1}|_{16} + \kappa_n = O(1)$, we see that $R_+ = o(1)$ leads to $p = \vartheta_n^{4/13}$, which results in the rate $R_+ \asymp p^{9/2}/\vartheta_n^{3/2} = r_n$, while $R_0 = O(r_n)$ leads to $\mathbb{P}(A_{n,1}) \asymp \vartheta_n^{-2}$. By (B0)(c), we thus prove the desired rate in (2.3).

To show (6.41), we use (B3) and (6.5) with $k = 3$, there being $\widetilde{\mathbf{B}}_3$ in between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and get

$$\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_1) - \ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n\|_o \le \|\dddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_3)\|_o\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \le \underset{\sim}{\xi}_n\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|. \tag{6.47}$$

Recalling (6.7), we thus have

$$\begin{aligned} 2\|\underset{\sim}{\boldsymbol{\Delta}}_{2+}\| &\le \kappa_n\mathbb{E}\big(\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_2) - \ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n\|_o s_n^{-2}\|\mathbf{l}\|^2\big) \\ &\le \kappa_n\mathbb{E}\big((\bar{\underset{\sim}{\xi}}_n + \mathbb{E}(\underset{\sim}{\xi}_n))s_n^{-3}\|\mathbf{l}\|^3\big) =: \kappa_n(D_1 + D_2). \end{aligned} \tag{6.48}$$

By (6.19), (B3) and Hölder's inequality with indices $(8/3, 8/3, 4)$ and $(1/2, 1/2)$ for (6.49) and (6.50), respectively, we get

$$\begin{aligned} D_1 &\leq |s_n^{-1}|_8^3 \cdot |\mathbf{l}|_8^3 \cdot |\bar{\underset{\sim}{\xi}}_n|_4 = |s_n^{-1}|_8^3 \, O(p^{7/2}/\vartheta_n^2), && (6.49)\\ D_2 &\leq |s_n^{-1}|_6^3 \cdot |\mathbf{l}|_6^3 \cdot \mathbb{E}(\underset{\sim}{\xi}_n) = |s_n^{-1}|_6^3 \, O(p^{7/2}/\vartheta_n^{3/2}), && (6.50) \end{aligned}$$

yielding (6.41). Similarly, it follows (6.42) from

$$\begin{aligned} 2\|\underset{\sim}{\boldsymbol{\Delta}}_{20}\| &\leq \kappa_n \mathbb{E}\big(\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_2) - \ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n\|_o \mathbf{1}[A_{n,1}]\big)\|\mathbf{b} - \boldsymbol{\beta}_0\|^2 \\ &\leq \kappa_n \mathbb{E}\big((\bar{\underset{\sim}{\xi}}_n + \mathbb{E}(\underset{\sim}{\xi}_n))\mathbf{1}[A_{n,1}]\big)\|\mathbf{b} - \boldsymbol{\beta}_0\|^3 \\ &=: \kappa_n (D_{10} + D_{20}) O(p^{3/2}), \quad \text{where} \\ D_{10} &\leq |\bar{\underset{\sim}{\xi}}_n|_2 \mathbb{P}^{1/2}(A_{n,1}) = \mathbb{P}^{1/2}(A_{n,1}) O(p^2/\sqrt{\vartheta_n}), && (6.51)\\ D_{20} &\leq \mathbb{E}(\underset{\sim}{\xi}_n)\mathbb{P}(A_{n,1}) = \mathbb{P}(A_{n,1}) O(p^2). && (6.52) \end{aligned}$$

By (B2), $\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_2)\|_o \leq \tilde{\eta}_n$, so that

$$\begin{aligned} 4\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{22+}\|^2) &\leq \|\mathbb{E}^{-1}(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\|_o^2 \mathbb{E}(\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_2)\|_o^2 s_n^{-4}\|\mathbf{l}\|^4) \leq \kappa_n^2 \mathbb{E}(\tilde{\eta}_n^2 s_n^{-4}\|\mathbf{l}\|^4) \\ &\leq 2\kappa_n^2 \mathbb{E}(\bar{\tilde{\eta}}_n^2 + \mathbb{E}^2(\tilde{\eta}_n)) s_n^{-4}\|\mathbf{l}\|^4) =: 2\kappa_n^2(E_1 + E_2). && (6.53) \end{aligned}$$

By Hölder's inequality with power indices $(4, 4, 2)$, we obtain

$$\begin{aligned} E_1 &\leq |s_n^{-1}|_{16}^4 \cdot |\bar{\tilde{\eta}}_n|_8^2 \cdot |\mathbf{l}|_8^4 \leq |s_n^{-1}|_{16}^4 O(p^5/\vartheta_n^3), && (6.54)\\ E_2 &\leq |s_n^{-1}|_8^4 \cdot \mathbb{E}^2(\tilde{\eta}_n) \cdot |\mathbf{l}|_8^4 = |s_n^{-1}|_8^4 O(p^5/\vartheta_n^2), && (6.55) \end{aligned}$$

yielding (6.43), whereas (6.44) follows from

$$\begin{aligned} 4\mathbb{E}(\|\underset{\sim}{\boldsymbol{\alpha}}_{220}\|^2) &\leq \|\mathbb{E}^{-1}(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\|_o^2 \mathbb{E}\big(\|\ddot{\underset{\sim}{\boldsymbol{\Psi}}}_n(\widetilde{\mathbf{B}}_2)\|_o^2 \mathbf{1}[A_{n,1}]\big)\|\mathbf{b} - \boldsymbol{\beta}_0\|^4 \\ &\leq \kappa_n^2 \mathbb{E}\big(\tilde{\eta}_n^2 \mathbf{1}[A_{n,1}]\big)\|\mathbf{b} - \boldsymbol{\beta}_0\|^4 \\ &\leq 2\kappa_n^2 \mathbb{E}\big((\bar{\tilde{\eta}}_n^2 + \mathbb{E}^2(\tilde{\eta}_n))\mathbf{1}[A_{n,1}]\big)\|\mathbf{b} - \boldsymbol{\beta}_0\|^4 \\ &\leq 2\kappa_n^2 \big(|\bar{\eta}_n|_4^2 \mathbb{P}^{1/2}(A_{n,1}) + \mathbb{E}^2(\tilde{\eta}_n)\mathbb{P}(A_{n,1})\big)\|\mathbf{b} - \boldsymbol{\beta}_0\|^4 \\ &= \kappa_n^2 \big(\mathbb{P}^{1/2}(A_{n,1}) O(p^5/\vartheta_n) + \mathbb{P}(A_{n,1}) O(p^5)\big). && (6.56) \end{aligned}$$

Using the last Hölder's inequality, (6.39) – (6.40) follow from

$$\begin{aligned} 2\|\underset{\sim}{\boldsymbol{\Delta}}_{11+}\| &\leq \|\mathbb{E}^{-1}(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\|_o \mathbb{E}(\|\ddot{\bar{\underset{\sim}{\boldsymbol{\Psi}}}}_n\|_o s_n^{-2}\|\mathbf{l}\|^2) \\ &\leq \kappa_n \mathbb{E}(\|\ddot{\bar{\underset{\sim}{\boldsymbol{\Psi}}}}_n\|_o s_n^{-2}\|\mathbf{l}\|^2) \leq \kappa_n |\ddot{\bar{\underset{\sim}{\boldsymbol{\Psi}}}}_n|_2 \cdot |s_n^{-1}|_8^2 \cdot |\mathbf{l}|_8^2 = \kappa_n |s_n^{-1}|_8^2 \cdot O(p^{5/2}/\vartheta_n^{3/2}), \\ 2\|\underset{\sim}{\boldsymbol{\Delta}}_{110}\| &\leq \|\mathbb{E}^{-1}(\dot{\underset{\sim}{\boldsymbol{\Psi}}}_n)\|_o \mathbb{E}(\|\ddot{\bar{\underset{\sim}{\boldsymbol{\Psi}}}}_n\|_o \mathbf{1}[A_{n,1}])\|\mathbf{b} - \boldsymbol{\beta}_0\|^2 \\ &\leq \kappa_n |\ddot{\bar{\underset{\sim}{\boldsymbol{\Psi}}}}_n|_2 \cdot \mathbb{P}^{1/2}(A_{n,1}) \cdot \|\mathbf{b} - \boldsymbol{\beta}_0\|^2 = \kappa_n \mathbb{P}^{1/2}(A_{n,1}) \cdot O(p^{5/2}/\sqrt{\vartheta_n}). \end{aligned}$$

□

References

[1] Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37**: 1705–1732.

[2] Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**: 81–91.

[3] Chatterjee, S. and Bose, A. (2002). Dimension asymptotics for generalized bootstrap in linear regression. *Ann. Statist.* **54** (2): 367–381.

[4] Cook, R. D., Tsai, C.-L. and Wei, B. C. (1986). Bias in nonlinear regression. *Biometrika* **73**: 615–623.

[5] Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *J. R. Statist. Soc. B*, **30**: 248–275.

[6] Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *J. R. Statist. Soc. B*, **53**(3): 629–643.

[7] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann. Statist.* **3**: 1189–1217.

[8] Firt, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** (1): 27–38.

[9] Gart, J. J., Pettigrew, H.M., and Thomas, D.G. (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika* **72**: 179–190.

[10] Jiang W. and Turnbull B. (2004). The indirect method: inference based on intermediate statistics – synthesis and examples. *Statist. Sci.* **19**: 239–263.

[11] Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96** (4): 793–804.

[12] Kosmidis, I. and D. Firth (2010). A generic algorithm for reducing bias in parametric estimation. *Electron. J. Stat.* **4** (1): 1097–1112. DOI: 10.1214/10-EJS579

[13] Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *WIREs Comput. Stat.* **6**: 185–196. doi: 10.1002/wics.1296

[14] Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Stat. Assoc.* **91**: 1007–1016

[15] McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

[16] Mehrabi, Y. and Matthews, J. N. S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* **51**: 1543–1549.

[17] Pettitt, A. N., Kelly, J.M., and Gao, J.T. (1998). Bias correction for censored data with exponential lifetimes. *Statistica Sinica* **8**: 941–964.

[18] Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**: 356–366.

[19] Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine* **2**: 71–78.

[20] Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics Series.

[21] Simas, A. B., Barreto-Souza, W., and Rocha, A.V. (2010). Improved estimators for a general class of beta regression models. *Comput. Stat. Data Anal.* **54** (2): 348–366.

[22] Wu, C.F.J.(1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *Ann. Statist.* **14** (4): 1261-1295.