Adaptive Huber Regression on Markov-dependent Data

Jianqing Fan, Yongyi Guo, and Bai Jiang*

Abstract

High-dimensional linear regression has been intensively studied in the community of statistics in the last two decades. For the convenience of theoretical analyses, classical methods usually assume independent observations and sub-Gaussian-tailed errors. However, neither of them hold in many real high-dimensional time-series data. Recently [Sun, Zhou, Fan, 2019, J. Amer. Stat. Assoc., in press] proposed Adaptive Huber Regression (AHR) to address the issue of heavy-tailed errors. They discover that the robustification parameter of the Huber loss should adapt to the sample size, the dimensionality, and the moments of the heavy-tailed errors. We progress in a vertical direction and justify AHR on dependent observations. Specifically, we consider an important dependence structure – Markov dependence. Our results show that the Markov dependence impacts on the adaption of the robustification parameter and the estimation of regression coefficients in the way that the sample size should be discounted by a factor depending on the spectral gap of the underlying Markov chain.

Keywords: Adaptive Huber Regression, dependent observations, Markov chain, high-dimensional regression, heavy-tailed errors.

1   Introduction

In the Big Data era, massive and high-dimensional data characterize many modern statistical problems, arising from biomedical sciences, econometrics, finance, engineering and social sciences. Examples include the gene expression data, the functional magnetic resonance imaging (fMRI) data, the macroeconomic data, the high-frequency financial data, the high-resolution image data, the e-commerce data, among others. The data abundance, the high dimensionality and other complex structures have given rise to a few statistical and computational challenges (**?**).

An important and fundamental problem is high-dimensional linear regression, in which the dimensionality (the number of covariates) is much larger than the sample size so that the ordinary least squares estimation of the regression coefficients

is highly unstable. To solve this problem, statisticians have made extensive progresses in the development of high-dimensional statistical inference (**?????**). They commonly make the sparsity assumption that only a small number of covariates contribute to the response variable, and enforce the sparsity of regression coefficients by regularization techniques. For a comprehensive and systematic overview of this high-dimensional regression problem, we refer to **?** and **?**.

For the convenience of theoretical analyses, many high-dimensional regression methods assume Gaussian or sub-Gaussian-tailed errors in the model. However, this assumption is unsatisfied in a broad range of real datasets. See **?** for asset return data, **??** for macroeconomic data, **?** for RNA-seq gene expression data, **?** for microarray gene expression data, **?** for fMRI data, to name a few. **?** argue that heavy-tailed errors are stylized features of high-dimensional data. These heavy-tailed errors impair the consistency of many high-dimensional regression methods, as the ordinary least squares loss function is non-robust to outliers.

Robust loss functions, which are less sensitive to outliers, have been considered to address the issue of heavy-tailed errors. Huber's seminar work (**??**) introduced robust M-estimators ("M" for "maximum likelihood-type") and provided the initial theory on robust regression methods. Asymptotic properties of robust M-estimators in the low-dimensional setting have been well studied by **??????**. Recently **?** investigated theoretical properties of a general class of regularized robust M-estimators for the high-dimensional linear model with heavy-tailed errors.

More recently, **?** studied a specific regularized robust M-estimator, which minimizes the ℓ_1-regularized Huber loss, for the high-dimensional linear model with heavy-tailed errors, which admit finite $(1 + \delta)$-moments for some $\delta > 0$. Remarkably, they observe that the robustification parameter of the Huber loss should adapt to the sample size, the dimensionality and $(1 + \delta)$-moments of heavy-tailed errors for an optimal tradeoff between bias and robustness. The adaption of the robustification parameter exhibits a smooth phase transition between regimes of $0 < \delta < 1$ and $\delta > 1$. Thereafter, they name their method *Adaptive Huber Regression* (AHR), to highlight its difference from others' Huber regression methods with fixed robustification parameter.

Apart from heavily-tailed errors, another common feature of high-dimensional data, especially those collected in a temporal order, is dependent observations. For example, fMRI time series data are usually collected from a few different regions over a time period (**??**); the monthly data of macroeconomic variables spanning the time period of decades are now benchmark datasets used by many econometric studies (**???**).

High-dimensional regression has been applied to fMRI time-series data (**???**) and economic time-series data (**???**), although its theoretical framework takes little consideration of the dependence structure of observations. For fMRI time-series

data, the lasso or elastic-net regression method is one of the main tools to find a small number of functionally-connected regions of a specific region in human brains from temporally-ordered observations across all the regions (**?**). On a macroeconomic dataset, (**?**) regressed the U.S. bond premia on macroeconomic variables over the time period spanning from January, 1964 to December, 2003. Nevertheless, the applicability of high-dimensional regression methods to these time-series data is not fully understood, because these methods assume either unconditional independence of data samples directly (as in the random design setup), or conditional independence of data samples given covariates satisfying some conditions (as in the fixed design setup) (**??**). In the latter case, high probabilities of these conditions are usually verified with independent observations of covariates.

This paper aims to close the gap between theories of high-dimensional regression methods and practical needs in addressing both heavy-tailed errors and dependent observations of real data. Inspired by the optimality of AHR dealing with heavy-tailed errors in the independent setup, we consider extending AHR to cope with the dependence structure of observations. Albeit of the theoretical results on AHR by **?**, it is still unclear whether AHR works for the dependent data. This unclearness puts the applicability of AHR to many real high-dimensional datasets in doubt. Even if AHR is justifiable under some type of dependence structure, we are still curious about the degree to which the data dependence influences the error rate of the AHR estimator. As an initial step towards full answers for these questions, we narrow down to the Markov dependence, an important and widely-used dependence structure, and analyze AHR on Markov-dependent data. Specially, we assume that covariates are functions of an underlying Markov chain and heteroskedastic heavy-tailed errors are dependent on the Markov chain.

In this Markov-dependent setup, we show under moderate conditions that AHR exhibits a similar phase transition of the adaptation of the robustification parameter and the estimation of regression coefficients between regimes of $0 < \delta < 1$ and $\delta > 1$, compared to that in the usual independent setup. The only difference is that the sample size should be discounted by a factor depending on the spectral gap of the underlying Markov chain.

The core of the proof is to bound in ℓ_∞-norm the gradient of the Huber loss at the true sparse vector of coefficients, denoted by $\boldsymbol{\beta}^\star$, and to establish the restricted eigenvalue condition of the Hessian of the Huber loss over a neighborhood of $\boldsymbol{\beta}^\star$. In the usual independent setting, these tasks are accomplished by applying Bernstein's inequality for independent random variables. The Bernstein-type mixture of the sub-Gaussian and sub-exponential tails is the key to derive the trade-off of bias and robustness in AHR. However, the analogous tasks in the Markov-dependent setup are non-trivial, due to the lack of Bernstein's inequalities for (possibly non-identical) functions of Markov chains. A very recent work (**?**) establishes the ex-

act counterpart of Bernstein's inequality for **bounded**, non-identical functions of Markov chains. But it still does not fully meet the requirements of the theoretical analyses in this paper, because covariates and errors involved in the theoretical analyses are **unbounded**. We develop a truncation argument for the extension of the Berstein-type inequality in (**?**) to unbounded functions.

The rest of this paper is organized as follows. Section **??** introduces the high-dimensional linear model and the methodology of AHR in the Markov-dependent setup. Section **??** presents the assumptions and the main theorem. Section **??** sketches the proof of the main theorem. Other technical proofs are collected in Section **??**. Section **??** concludes the paper with a brief discussion.

2   Model and Methodology

The goal is to estimate the high-dimensional linear model with Markov-dependent covariates and heavy-tailed errors. Start with the linear model as follows.

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where y_i is the response, $\boldsymbol{x}_i \in \mathbb{R}^d$ is the vector of d covariates, ε_i is the error, $\boldsymbol{\beta} \in \mathbb{R}^d$ is the vector of d regression coefficients. In the high-dimensional regime, d is much larger than the sample size n. To make the model identifiable, assume only a small number s of covariates contribute to the response, i.e., the vector of true regression coefficients $\boldsymbol{\beta}^\star$ contains at most s non-zero elements.

Suppose covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are functions of a stationary Markov chain $\{Z_i\}_{i=1}^n$ on a general state space, i.e., for a collection of d-dimensional vectorial functions $\{\boldsymbol{f}_i\}_{i=1}^n$, where $\boldsymbol{f}_i = (f_{i1}, \ldots, f_{id})^{\mathrm{T}}$,

$$x_{ij} = f_{ij}(Z_i).$$

The errors are conditionally independent given the underlying Markov chain and possibly heteroskedastic. There exists a conditional distribution $g(\cdot|z)$ such that

$$\varepsilon_i | Z_i \sim g(\cdot | Z_i).$$

If $\{Z_i\}_{i=1}^n$ are independently and identically distributed (i.i.d.) then this regression setup reduces to the usual one in which $\{(\boldsymbol{x}_i, \varepsilon_i)\}_{i=1}^n$ are independent.

For the task of estimating the true s-sparse coefficients $\boldsymbol{\beta}^\star$, consider the ℓ_1-regularized robust M-estimator with the Huber loss as follows.

$$\widehat{\boldsymbol{\beta}}_{\tau,\lambda} = \arg\min_{\boldsymbol{\beta}} H_\tau(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \quad H_\tau(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n h_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \tag{1}$$

where h_τ is the Huber loss (**?**)

$$h_\tau(w) = \begin{cases} w^2/2 & \text{if } |w| \le \tau \\ \tau|w| - \tau^2/2 & \text{if } |w| > \tau, \end{cases} \tag{2}$$

4

with the so-called robustification parameter τ (**?**), and λ is the regularization parameter encouraging the sparsity of $\boldsymbol{\beta}$ (**?**).

Heuristically, a larger robustification parameter τ reduces the bias of $\widehat{\boldsymbol{\beta}}_{\tau,\lambda}$ at the cost of less robustness. The extreme case of $\tau = \infty$ corresponds to the ordinary least squares estimation. We find that τ should adapt to the sample size n, the dimensionality d, the heaviness of the tails of the errors and the dependence of the Markov chain. Suppose the heavy-tailed errors admit finite $(1 + \delta)$-moments (conditionally on Z_i's) for some $\delta > 0$. A large δ indicates light tails of errors. The dependence of the Markov chain is measured by a quantity $\gamma \in [0, 1]$, denoting the norm of the Markov operator (induced by transition kernel) acting on the Hilbert space of all squared-integrable and mean-zero functions with respect to the invariant distribution. A small γ indicates a fast convergence of the Markov chain towards its stationarity from a non-stationary initial distribution (**?**). Our analyses show under moderate conditions that the choice of

$$\tau \asymp \left(\frac{1 - \gamma}{1 + \gamma} \cdot \frac{n}{\log d} \right)^{1/(1 + \min\{\delta, 1\})}.$$

achieves the optimal trade-off between bias and robustness in the Markov dependent setup.

3 Assumptions and Theorems

This section presents our main result under four assumptions. The first assumption is on the convergence speed of the underlying Markov chain.

Assumption 1 (Markov chain with non-zero spectral gap)**.** *The underlying Markov chain $\{Z_i\}_{i=1}^n$ is stationary with its unique invariant measure π and admits a non-zero spectral gap $1 - \gamma$.*

Recall that the quantity γ is defined as the norm of the Markov operator (induced by transition kernel) acting on the Hilbert space of all π-squared-integrable and π-mean-zero functions. $1 - \gamma$ is called spectral gap of the Markov chain. It has been involved as constants in mean squared error bound for Markov chain Monte Carlo (**?**), Hoeffding-type and Bernstein-type inequalities for Markov chains (**???????**). A non-zero spectral gap is closely related to other convergence criteria of Markov chains (**???**).

Next two assumptions allow both covariates and heavy-tailed errors to be heteroskedastic, but impose on them some moment conditions. For the covariates, the bounded fourth moment of some envelop function is required. For the heavy-tailed errors, the bounded (conditional) $(1 + \delta)$-moment is required.

Assumption 2 (Covariates with bounded fourth moments). *There exists an envelop function $M : z \mapsto \mathbb{R}$ for functions f_{ij}'s, i.e., $M(z) \geq \max_{1 \leq i \leq n, 1 \leq j \leq d} |f_{ij}(z)|$ for π-almost every z. And, $\sigma^4 := \int M^4(z)\pi(dz) < \infty$.*

Assumption 3 (Errors with bounded $(1+\delta)$-moments). *$\mathbb{E}[\varepsilon_i | Z_i] = 0$ almost surely, i.e., $\int \varepsilon g(\varepsilon|z)d\varepsilon = 0$ for π-almost every z. And, for some $\delta > 0$ and $v_\delta > 0$, $\mathbb{E}[|\varepsilon_i|^{1+\delta}|Z_i] < v_\delta$ almost surely, i.e., $\int |\varepsilon|^{1+\delta} g(\varepsilon|z)d\varepsilon < v_\delta$ for π-almost every z.*

The last assumption is on the restricted eigenvalue of the (aggregated) covariance matrix of covariates. It is a unified condition in the literature of high-dimensional regression, see e.g., **?** and **?**. Let $S = \{j : \beta_j^\star \neq 0\}$ be the index set of active covariates. Define the ℓ_1-cone

$$\mathcal{C} := \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}_{S^c}\|_1 \leq 3\|\boldsymbol{u}_S\|_1\}, \tag{3}$$

where $\boldsymbol{u}_{S^c}$ is the subvector assembling $\{\boldsymbol{u}_j : j \in S^c\}$ and $\boldsymbol{u}_S$ is the subvector assembling $\{\boldsymbol{u}_j : j \in S\}$. The constant 3 in the definition of $\mathcal{C}$ has no specific meaning. It can be replaced by other constant larger than 1. Write the (aggregated) covariance matrix of covariates as

$$\boldsymbol{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}]. \tag{4}$$

Assumption 4 (Restricted eigenvalue of covariance matrix). *There exists constant $\kappa > 0$ such that, for sufficiently large n,*

$$\inf\{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_n\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1, \boldsymbol{u} \in \mathcal{C}\} \geq 2\kappa.$$

It is not hard to see that this condition holds if the smallest eigenvalue of $\boldsymbol{\Sigma}_n$ is strictly bounded away from 0 for sufficiently large n. Furthermore, if vectorial functions $\boldsymbol{f}_i = \boldsymbol{f}$ do not vary with i then $\{\boldsymbol{x}_i\}_{i=1}^n$ is a stationary time series as $\boldsymbol{x}_i = \boldsymbol{f}(Z_i)$ are functions of the underlying Markov chain $\{Z_i\}_{i=1}^n$. In this case, $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^{\mathrm{T}}]$, and Assumption **??** holds if the smallest eigenvalue of $\boldsymbol{\Sigma}$ is strictly bounded away from 0.

Now we are ready to present the main result of this paper. Note that "w.h.p." stands for "with high probability $1 - o(1)$".

Theorem 1. *Suppose Assumptions **??**-**??** hold and*

$$s\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}} = o(1).$$

*Then the AHR estimator $\widehat{\boldsymbol{\beta}}_{\tau,\lambda}$ in (**??**) with robustification and regularization parameters*

$$\tau \asymp \left(\frac{1-\gamma}{1+\gamma} \cdot \frac{n}{\log d}\right)^{1/(1+\min\{\delta,1\})}, \quad \lambda \asymp \left(\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}\right)^{\min\{\delta,1\}/(1+\min\{\delta,1\})}$$

achieves estimation errors

$$\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^{\star}\|_1 \lesssim s\lambda, \quad \|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^{\star}\|_2 \lesssim \sqrt{s}\lambda \quad w.h.p..$$

Similar to the discovery of **?** in the independent setup, there is a smooth phase transition of τ-adaption and $\boldsymbol{\beta}^{\star}$-estimation. A small $0 < \delta < 1$ suffices for AHR to consistently estimate $\boldsymbol{\beta}^{\star}$, albeit the rates $s(\log d/n)^{\delta/(1+\delta)}$ or $s^{1/2}(\log d/n)^{\delta/(1+\delta)}$ (given fixed $\gamma < 1$) are slower than those for the case of $\delta = 1$. In the latter case of $\delta = 1$, AHR achieves rates $s(\log d/n)^{1/2}$ for ℓ_1-error and $s^{1/2}(\log d/n)^{1/2}$ for ℓ_2-error like classical high-dimensional regression methods; but AHR only requires bounded second moments of error ε_i, which is weaker than the sub-Gaussian error condition in classical high-dimensional regression methods. A larger $\delta > 1$ gains no more estimation accuracy than $\delta = 1$.

The Markov dependence impacts on the adaption of the robustification parameter and the estimation of regression coefficients in the way that the sample size n is discounted by a factor $(1 - \gamma)/(1 + \gamma) < 1$. In other words, to achieve comparable τ-adaption and $\boldsymbol{\beta}^{\star}$-estimation, the required sample size increases by $(1+\gamma)/(1-\gamma)$ when moving from the independent setup to the Markov-dependent setup. Furthermore, this theorem allows γ to approach to 1 as n increases, so long as $s\sqrt{(1+\gamma)/(1-\gamma) \cdot \log d/n} \to 0$. Even though the spectral gap of the Markov chain is difficult to accurately compute in practice (**?**), Theorem **??** also apply if one replaces γ with an inaccurate overestimate $\gamma' \geq \gamma$.

4 Proof of Theorem ??

We break the proof of Theorem **??** into three propositions. Proposition **??** bounds the ℓ_1- and ℓ_2-errors of a generic ℓ_1-regularized M-estimator $\widehat{\boldsymbol{\beta}}$ minimizing $\mathcal{L}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Roughly speaking, given a *localized restricted eigenvalue* (LRE) condition, it establishes an ℓ_2-error bound

$$\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^{\star}\|_2 \lesssim \sqrt{s}\|\nabla H_\tau(\boldsymbol{\beta}^{\star})\|_\infty,$$

and a similar ℓ_1-error bound with an additional factor $\sqrt{s}$, by applying Proposition 1 to $(\mathcal{L}, \widehat{\boldsymbol{\beta}}) = (H_\tau, \widehat{\boldsymbol{\beta}}_{\tau,\lambda})$. This LRE condition requires strictly positive restricted eigenvalues over a local ℓ_1-neighborhood. It is a simplified version of (**?**, Definition 2). Another LRE condition in ℓ_2-neighborhood has found applications in (**?**, Definition 4.1).

Proposition 1. *Consider a ℓ_1-regularized minimizer $\widehat{\boldsymbol{\beta}} = \arg\min \mathcal{L}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$ of a convex, twice differentiable function $\mathcal{L} : \mathbb{R}^d \mapsto \mathbb{R}$. Suppose the following two conditions hold:*

(i) (localized restricted eigenvalue) Recall that S is the support of $\boldsymbol{\beta}^\star$ and that $\mathcal{C}$ is defined as (**??**). There exists a constant $\kappa > 0$ such that

$$\inf\{\boldsymbol{u}^{\mathrm{T}}\nabla^2\mathcal{L}(\boldsymbol{\beta})\boldsymbol{u} : \ \boldsymbol{u} \in \mathcal{C}, \ \|\boldsymbol{u}\|_2 = 1, \ \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 \le 48s\lambda/\kappa\} \ge \kappa.$$

(ii) A s-sparse $\boldsymbol{\beta}^\star$ satisfies $\|\nabla\mathcal{L}(\boldsymbol{\beta}^\star)\|_\infty \le \lambda/2$.

Then

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \le 48s\lambda/\kappa, \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le 12\sqrt{s}\lambda/\kappa.$$

To facilitate the proof of Theorem **??** as an application of Proposition **??** to $(\mathcal{L}, \widehat{\boldsymbol{\beta}}) = (H_\tau, \widehat{\boldsymbol{\beta}}_{\tau,\lambda})$, we establish the LRE condition for the Hessian matrix of the Huber loss $\nabla^2 H_\tau(\boldsymbol{\beta})$ in Proposition **??** and bound its gradient $\nabla H_\tau(\boldsymbol{\beta}^\star)$ in Proposition **??**.

Proposition 2. *Under Assumptions* **??**-**??**, *we have*

$$\inf\{\boldsymbol{u}^{\mathrm{T}}\nabla^2 H_\tau(\boldsymbol{\beta})\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1, \boldsymbol{u} \in \mathcal{C}, \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 \le r\}$$
$$= 2\kappa - sO_p\left(\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}} + \frac{1}{\tau^{1+\delta}} + \frac{r^2}{\tau^2}\right).$$

Proposition 3. *Under Assumptions* **??**-**??**, *for some constant* $C > 0$

$$\|\nabla H_\tau(\boldsymbol{\beta}^\star)\|_\infty \le \sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{2\sigma^2 v_{\min\{\delta,1\}}\tau^{1-\min\{\delta,1\}}\log d}{n}}$$
$$+ \frac{1+\gamma}{1-\gamma} \cdot \frac{20\tau\log d}{n} + C\tau^{-\min\{\delta,1\}}, \quad w.h.p..$$

Since $\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^\star\|_2 \lesssim \sqrt{s}\|\nabla H_\tau(\boldsymbol{\beta}^\star)\|_\infty$, optimizing the bound of $\|\nabla H_\tau(\boldsymbol{\beta}^\star)\|_\infty$ in Proposition **??** over τ get the optimal ℓ_2-error bound for $\widehat{\boldsymbol{\beta}}_{\tau,\lambda}$. Collecting these pieces together proves Theorem **??**.

Proof of Theorem **??**. Setting

$$\tau = \left(\frac{1-\gamma}{1+\gamma} \cdot \frac{n}{\log d}\right)^{1/(1+\min\{\delta,1\})},$$
$$\lambda = 2(\sqrt{2\sigma^2 v_{\min\{\delta,1\}}} + 20 + C)\tau^{-\min\{\delta,1\}},$$
$$r = 48s\lambda/\kappa$$

in Propositions **??**-**??** yields

$$\inf\{\boldsymbol{u}^{\mathrm{T}}\nabla^2 H_\tau(\boldsymbol{\beta})\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1, \boldsymbol{u} \in \mathcal{C}, \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 \le r\} \ge \kappa, \quad w.h.p..$$
$$\|\nabla H_\tau(\boldsymbol{\beta}^\star)\|_\infty \le \lambda/2, \quad w.h.p..$$

Applying Proposition 1 to $(\mathcal{L}, \widehat{\boldsymbol{\beta}}) = (H_\tau, \widehat{\boldsymbol{\beta}}_{\tau,\lambda})$ completes the proof. $\qquad\square$

5 Other Technical Proofs

5.1 Proof of Proposition ??

The proof of Proposition **??** consists of two steps.

  (a) $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \in \mathcal{C}$, where $\mathcal{C}$ is defined in (**??**).

  (b) $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \le 4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le r := 48s\lambda/\kappa$.

Statement (a) asserts that $\widehat{\boldsymbol{\beta}}$ belongs to the ℓ_1 convex cone $\mathcal{C}$ around the estimand $\boldsymbol{\beta}^\star$. Its proof uses merely the convexity of $\mathcal{L}$, the optimality of $\widehat{\boldsymbol{\beta}}$, and Hölder's inequality for vectors, all of which are routine in the literature of high-dimensional linear regression. **?**, Lemma 8 is a slightly stronger version of statement (a). Here we give a neat proof of statement (a) for the sake of readability. Statement (b) provides similar but general results compared to **?**, Theorem 8: the former allows Markov-dependent covariate vectors with bounded forth moments (see Assumption **??**) to fulfill the LRE condition in Proposition **??**, while the latter requires a stronger condition that covariate vectors are i.i.d. and sub-Gaussian (see **?**, Condition 5 and Theorem 8). Moreover, our proof of statement (b) is simpler than that of **?**, Theorem 8, as the latter relies on an auxiliary result of the symmetric Bregman divergence (see **?**, Lemma 2).

*Proof of Proposition **??**(a).* By the optimality of $\widehat{\boldsymbol{\beta}}$,

$$\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\boldsymbol{\beta}^\star) \le \lambda(\|\boldsymbol{\beta}^\star\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1).$$

By the convexity of $\mathcal{L}$, Hölder's inequality, and the condition that $\|\nabla\mathcal{L}(\boldsymbol{\beta}^\star)\|_\infty \le \lambda/2$,

$$\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\boldsymbol{\beta}^\star) \ge \langle \nabla\mathcal{L}(\boldsymbol{\beta}^\star), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \rangle \ge -\|\nabla\mathcal{L}(\boldsymbol{\beta}^\star)\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \ge -\lambda\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1/2.$$

It follows that

$$2(\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^\star\|_1) \le \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1.$$

On the left-hand side, using the fact that $\boldsymbol{\beta}^\star_{S^c} = \mathbf{0}$,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^\star\|_1 &= \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 + \|\widehat{\boldsymbol{\beta}}_S\|_1 - \|\boldsymbol{\beta}^\star_S\|_1 \\
&= \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_{S^c}\|_1 + \|\boldsymbol{\beta}^\star_S + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_S\|_1 - \|\boldsymbol{\beta}^\star_S\|_1 \\
&\ge \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_{S^c}\|_1 - \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_S\|_1
\end{aligned}$$

On the right-hand side,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 = \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_{S^c}\|_1 + \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_S\|_1.$$

Putting the last three displays together and rearranging terms completes the proof. $\qquad\square$

Proof of Proposition **??**(b). First, it follows from step (a) that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \le 4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2$. It is left to show $4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le r$. Suppose for the sake of contradiction that $4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 > r$. By the optimality of $\widehat{\boldsymbol{\beta}}$ and the integral form of the Taylor expansion,

$$
\begin{aligned}
0 &\ge [\mathcal{L}(\widehat{\boldsymbol{\beta}}) + \lambda\|\widehat{\boldsymbol{\beta}}\|_1] - [\mathcal{L}(\boldsymbol{\beta}^\star) + \lambda\|\boldsymbol{\beta}^\star\|_1] \\
&= \lambda(\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^\star\|_1) + \langle \nabla\mathcal{L}(\boldsymbol{\beta}^\star), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \rangle \\
&\quad + \int_0^1 (1-t)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^{\mathrm{T}} \nabla^2 \mathcal{L}(\boldsymbol{\beta}^\star + t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) dt
\end{aligned}
$$

For the first term,

$$
\lambda(\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^\star\|_1) \ge -\lambda\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \ge -4\sqrt{s}\lambda\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2.
$$

For the second term, from the condition that $\|\nabla\mathcal{L}(\boldsymbol{\beta}^\star)\|_\infty \le \lambda/2$, it follows that

$$
\begin{aligned}
|\langle \nabla\mathcal{L}(\boldsymbol{\beta}^\star), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \rangle| &\le \|(\nabla\mathcal{L}(\boldsymbol{\beta}^\star))_S\|_2 \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_S\|_2 + \|(\nabla\mathcal{L}(\boldsymbol{\beta}^\star))_{S^c}\|_\infty \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_{S^c}\|_1 \\
&\le \sqrt{s}\|(\nabla\mathcal{L}(\boldsymbol{\beta}^\star))_S\|_\infty \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_S\|_2 + \|(\nabla\mathcal{L}(\boldsymbol{\beta}^\star))_{S^c}\|_\infty \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_{S^c}\|_1 \\
&\le \frac{\sqrt{s}\lambda}{2}\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_S\|_2 + \frac{3\lambda}{2}\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)_S\|_1 \le 2\sqrt{s}\lambda\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2.
\end{aligned}
$$

Proceed to lower bound the third term. To this end, note that $r/4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 < 1$ by the initial assumption. For any $0 \le t \le r/4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2$,

$$
\|[\boldsymbol{\beta}^\star + t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)] - \boldsymbol{\beta}^\star\|_1 \le t\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \le t \times 4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le r.
$$

Combining it with step (a) that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \in \mathcal{C}$ and the LRE condition yield a lower bound for the third term

$$
\begin{aligned}
&\int_0^1 (1-t)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^{\mathrm{T}} \nabla^2 \mathcal{L}(\boldsymbol{\beta}^\star + t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) dt \\
&\ge \int_0^{r/4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2} (1-t)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^{\mathrm{T}} \nabla^2 \mathcal{L}(\boldsymbol{\beta}^\star + t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) dt \\
&\ge \int_0^{r/4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2} (1-t)\kappa\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2 dt = \frac{\kappa r}{4\sqrt{s}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 - \frac{\kappa r^2}{32s}
\end{aligned}
$$

Putting the lower bounds of three terms together with the scaling of $r = 48s\lambda/\kappa$ yields

$$
0 \ge \left(\frac{\kappa r}{4\sqrt{s}} - 6\sqrt{s}\lambda \right)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 - \frac{\kappa r^2}{32s} = 6\sqrt{s}\lambda\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 - \frac{3\lambda r}{2},
$$

which contradicts to the initial assumption that $4\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 > r$. □

5.2 Proof of Proposition ??

We first present two lemmas, which are useful in the proof of Proposition ??. Proofs of these two lemmas are put at the end of this subsection.

Lemma 1. *Under Assumptions ??-??,*

$$\max_{1\leq j,k\leq d}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_{ij}x_{jk}]\right| = O_p\left(\sqrt{\frac{1+\gamma}{1-\gamma}\cdot\frac{\log d}{n}}\right).$$

Lemma 2. *Under Assumptions ??-??,*

$$\frac{1}{n}\sum_{i=1}^{n}M^2(Z_i)1\{|\varepsilon_i| > \tau/2\} \leq \sigma^2\left(\frac{2}{\tau}\right)^{1+\delta}v_\delta + O_p\left(\sqrt{\frac{1+\gamma}{1-\gamma}\cdot\frac{\log d}{n}}\right).$$

Proof of Proposition ??. For any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1 \leq r$,

$$1\{|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}| > \tau\} \leq 1\{|\varepsilon_i| > \tau/2\} + 1\{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)| > \tau/2\}$$
$$\leq 1\{|\varepsilon_i| > \tau/2\} + 1\{\|\boldsymbol{x}_i\|_\infty > \tau/2r\}$$
$$\leq 1\{|\varepsilon_i| > \tau/2\} + 1\{M(Z_i) > \tau/2r\},$$

where M is the envelop function introduced by Assumption ??. Let

$$\epsilon_{1n} = \max_{1\leq j,k\leq d}\left|\left[\boldsymbol{\Sigma}_n - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]_{j,k}\right| = \max_{1\leq j,k\leq d}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_{ij}x_{ik}]\right|$$

$$\epsilon_{2n} = \max_{1\leq j,k\leq d}\left|\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}1\{|\varepsilon_i| \geq \tau/2\}\right]_{j,k}\right| \leq \frac{1}{n}\sum_{i=1}^{n}M^2(Z_i)1\{|\varepsilon_i| \geq \tau/2\}$$

$$\epsilon_{3n} = \max_{1\leq j,k\leq d}\left|\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}1\{M(Z_i) \geq \tau/2r\}\right]_{j,k}\right| \leq (2r/\tau)^2\times\frac{1}{n}\sum_{i=1}^{n}M^4(Z_i).$$

For any $\boldsymbol{u}\in\mathcal{C}$ such that $\|\boldsymbol{u}\|_2 = 1$, we have $\|\boldsymbol{u}\|_1 \leq 4\sqrt{s}$. Thus,

$$\boldsymbol{u}^{\mathrm{T}}\nabla^2 H_\tau(\boldsymbol{\beta})\boldsymbol{u} = \boldsymbol{u}^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}1\{|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}| \leq \tau\}\right]\boldsymbol{u},$$

$$\geq \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_n\boldsymbol{u} - \boldsymbol{u}^{\mathrm{T}}\left[\boldsymbol{\Sigma}_n - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]\boldsymbol{u}$$

$$- \boldsymbol{u}^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}1\{|\varepsilon_i| > \tau/2\}\right]\boldsymbol{u}$$

$$- \boldsymbol{u}^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}1\{M(Z_i) > \tau/2r\}\right]\boldsymbol{u}.$$

$$\geq \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_n\boldsymbol{u} - \epsilon_{1n}\|\boldsymbol{u}\|_1^2 - \epsilon_{2n}\|\boldsymbol{u}\|_1^2 - \epsilon_{3n}\|\boldsymbol{u}\|_1^2,$$

$$\geq 2\kappa - 16s(\epsilon_{1n} + \epsilon_{2n} + \epsilon_{3n}).$$

Further bounding ϵ_{1n} by Lemma **??**, ϵ_{2n} by Lemma **??** and

$$\epsilon_{3n} \leq (2r/\tau)^2 \times (\sigma^4 + o_p(1))$$

by the law of large number for Markov chains (**?**, Theorem 17.1.2) completes the proof. $\qquad\square$

Proof of Lemma **??**. Define a truncation operator

$$\mathcal{T}_t(w) = \begin{cases} -t & \text{if } w < -t \\ w & \text{if } |w| \leq t \\ +t & \text{if } w > +t. \end{cases} \tag{5}$$

For each $1 \leq j \leq d$ and each $1 \leq k \leq d$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[x_{ij} x_{jk}] \right| \leq D_{1jk} + D_{2jk} + D_{3jk},$$

where

$$D_{1jk} = \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\mathcal{T}_t[x_{ij} x_{ik}] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[x_{ij} x_{ik}] \right|,$$

$$D_{2jk} = \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{T}_t[x_{ij} x_{ik}] - \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \right|,$$

$$D_{3jk} = \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{T}_t[x_{ij} x_{ik}] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\mathcal{T}_t[x_{ij} x_{ik}] \right|.$$

Using the fact that $|\mathcal{T}_t(w) - w| \leq |w| 1\{|w| > t\} \leq |w|^2/t$, Cauchy-Schwarz inequality and the bounds on fourth moment of the envelop function $M : z \mapsto \mathbb{R}$ in Assumption **??**,

$$\max_{j,k} D_{1jk} \leq \max_{j,k} \frac{1}{tn} \sum_{i=1}^{n} \mathbb{E}[|x_{ij} x_{ik}|^2] \leq \max_{j,k} \frac{1}{t} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[x_{ij}^4] \times \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[x_{ik}^4]}$$

$$\leq \frac{1}{tn} \sum_{i=1}^{n} \mathbb{E}[M^4(Z_i)] \leq \frac{\sigma^4}{t}.$$

By a similar argument and the law of large number for Markov chains (**?**, Theorem 17.1.2)

$$\max_{j,k} D_{2jk} \leq \frac{1}{tn} \sum_{i=1}^{n} M^4(Z_i) = \frac{\sigma^4 + o_p(1)}{t}.$$

Noting that $|\mathcal{T}_t[W] - \mathbb{E}\mathcal{T}_t[W]| \leq 2t$ almost surely, we apply the Bernstein's inequality for Markov chains in (**?**, Theorem 1.1) and yields

$$\mathbb{P}\left(D_{3jk} > \epsilon\right) \leq 2 \exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma} \cdot V_{n,t} + 10t\epsilon}\right) \leq 2 \exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma} \cdot \sigma^4 + 10t\epsilon}\right),$$

where
$$V_{n,t} = \frac{1}{n} \sum_{i=1}^{n} \text{Var}\{\mathcal{T}_t[x_{ij}x_{ik}]\} \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[x_{ij}^2 x_{ik}^2] \leq \sigma^4.$$

A union bound delivers that

$$\mathbb{P}\left(\max_{j,k} D_{3jk} > \epsilon\right) \leq 2p^2 \exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma} \cdot \sigma^4 + 10t\epsilon}\right).$$

Let $t = \frac{1}{10c}\sqrt{\frac{1-\gamma}{1+\gamma} \cdot \frac{n}{\log d}}$, and $\epsilon = c\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}}$ for some $c > \sqrt{2(\sigma^4 + 1)}$. Then

$$\max_{j,k} D_{3jk} \geq c\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}}$$

with probability at most

$$2\exp\left(-\left(\frac{c^2}{\sigma^4 + \frac{1-\gamma}{1+\gamma}} - 2\right)\log d\right) \leq 2\exp\left(-\left(\frac{c^2}{\sigma^4 + 1} - 2\right)\log d\right).$$

Collecting these pieces together completes the proof. $\qquad\square$

Proof of Lemma ??. Note that $\{(Z_i, \varepsilon_i)\}_{i=1}^n$ are identically distributed. Write

$$\mathbb{E}\left[M^2(Z_1)\mathbf{1}\{|\varepsilon_1| > \tau/2\}\right] = \mathbb{E}\left[M^2(Z_1)\mathbb{P}(|\varepsilon_1| > \tau/2|Z_1)\right]$$

$$\leq \mathbb{E}\left[M^2(Z_1)\left(\frac{2}{\tau}\right)^{1+\delta}\mathbb{E}[|\varepsilon_1|^{1+\delta}|Z_1]\right]$$

$$\leq \mathbb{E}[M^2(Z_1)]\left(\frac{2}{\tau}\right)^{1+\delta}v_\delta \leq \sigma^2\left(\frac{2}{\tau}\right)^{1+\delta}v_\delta.$$

It is left to show

$$\left|\frac{1}{n}\sum_{i=1}^n M^2(Z_i)\mathbf{1}\{|\varepsilon_i| > \tau/2\} - \mathbb{E}\left[M^2(Z_1)\mathbf{1}\{|\varepsilon_1| > \tau/2\}\right]\right| = O_p\left(\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}}\right).$$

Recall that $\mathcal{T}_t$ is the truncation operator with threshold t defined in (??). Break down the quantity on the left-hand side of the last display into three terms.

$$D_1 + D_2 + D_3,$$

where

$$D_1 = \left|\mathbb{E}\mathcal{T}_t[M^2(Z_1)\mathbf{1}\{|\varepsilon_1| > \tau/2\}] - \mathbb{E}[M^2(Z_1)\mathbf{1}\{|\varepsilon_1| > \tau/2\}]\right|,$$

$$D_2 = \left|\frac{1}{n}\sum_{i=1}^n \mathcal{T}_t[M^2(Z_i)\mathbf{1}\{|\varepsilon_i| > \tau/2\}] - \frac{1}{n}\sum_{i=1}^n M^2(Z_i)\mathbf{1}\{|\varepsilon_i| > \tau/2\}\right|,$$

$$D_3 = \left|\frac{1}{n}\sum_{i=1}^n \mathcal{T}_t[M^2(Z_i)\mathbf{1}\{|\varepsilon_i| > \tau/2\}] - \mathbb{E}\mathcal{T}_t[M^2(Z_1)\mathbf{1}\{|\varepsilon_1| > \tau/2\}]\right|.$$

13

Using the fact that $|\mathcal{T}_t(w) - w| \leq |w|1\{|w| > t\} \leq |w|^2/t$, Cauchy-Schwarz inequality and the bounds on fourth moment of $M : z \mapsto \mathbb{R}$ in Assumption **??**,

$$D_1 \leq \frac{\mathbb{E}[M^4(Z_1)]}{t} = \frac{\sigma^4}{t}.$$

By a similar argument and the law of large number for Markov chains (**?**, Theorem 17.1.2)

$$D_2 \leq \frac{1}{tn} \sum_{i=1}^{n} M^4(Z_i) = \frac{\sigma^4 + o_p(1)}{t}.$$

Noting that $|\mathcal{T}_t[W] - \mathbb{E}\mathcal{T}_t[W]| \leq 2t$ almost surely, we apply the Bernstein's inequality for geometrically ergodic Markov chains in (**?**, Theorem 1.1) and yields

$$\mathbb{P}\left(D_3 > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma} \cdot V_t + 10t\epsilon}\right) \leq 2\exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma} \cdot \sigma^4 + 10t\epsilon}\right),$$

where

$$V_t = \mathrm{Var}\{\mathcal{T}_t[M^2(Z_1)\{|\varepsilon_2| > \tau/2\}]\} \leq \mathbb{E}[M^4(Z_1)] = \sigma^4.$$

Let $t = \frac{1}{10c}\sqrt{\frac{1-\gamma}{1+\gamma} \cdot \frac{n}{\log d}}$, and $\epsilon = c\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}}$ for some $c > 0$ then

$$D_3 \geq c\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}}$$

with probability at most

$$\exp\left(-\frac{c^2 \log d}{\sigma^4 + \frac{1-\gamma}{1+\gamma}}\right).$$

Collecting these pieces together completes the proof. $\qquad\square$

5.3 Proof of Proposition **??**

Proof of Proposition **??**. We merely consider the case of $0 < \delta \leq 1$. The proof for the case of $\delta > 1$ is the same with that for $\delta = 1$. Recall that $\mathcal{T}_t$ is the truncation operator with threshold t defined in (**??**), and h_τ is the Huber loss with the robustification parameter τ defined in (**??**). For each $1 \leq j \leq d$,

$$-\nabla_j H_\tau(\boldsymbol{\beta}^\star) = -\frac{\partial H_\tau(\boldsymbol{\beta}^\star)}{\partial \beta_j} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{T}_\tau(\varepsilon_i)x_{ij},$$

Note that $\mathbb{E}[\epsilon_i|Z_i] = 0$ almost surely in Assumption **??**. Write

$$|\mathbb{E}[\nabla_j H_\tau(\boldsymbol{\beta}^\star)]| \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{T}_\tau(\varepsilon_i)|x_{ij}|] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{E}[\mathcal{T}_\tau(\varepsilon_i)|\boldsymbol{x}_i]|x_{ij}|]$$

$$= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\{\mathbb{E}[\varepsilon_i|\boldsymbol{x}_i] - \mathbb{E}[\epsilon_i|Z_i]\}|x_{ij}|]$$

$$\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{E}[|\varepsilon_i|\mathbb{1}\{|\varepsilon_i| > \tau\}|Z_i]|x_{ij}|]$$

$$\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{E}[|\varepsilon_i|^{1+\delta}\tau^{-\delta}|Z_i]|x_{ij}|] \leq \sigma v_\delta \tau^{-\delta}$$

Next, we use a similar argument to that in Lemmas **??**-**??** to bound the deviations of $\nabla_j H_\tau(\boldsymbol{\beta}^\star)$ from their expectations. Break down the deviation into three terms as follows.

$$|\nabla_j H_\tau(\boldsymbol{\beta}^\star) - \mathbb{E}[\nabla_j H_\tau(\boldsymbol{\beta}^\star)]| \leq D_{1j} + D_{2j} + D_{3j},$$

where

$$D_{1j} = \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\mathcal{T}_t[\mathcal{T}_\tau(\varepsilon_i)x_{ij}] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{T}_\tau(\varepsilon_i)x_{ij}]\right|,$$

$$D_{2j} = \left|\frac{1}{n}\sum_{i=1}^n \mathcal{T}_t[\mathcal{T}_\tau(\varepsilon_i)x_{ij}] - \frac{1}{n}\sum_{i=1}^n \mathcal{T}_\tau(\varepsilon_i)x_{ij}\right|,$$

$$D_{3j} = \left|\frac{1}{n}\sum_{i=1}^n \mathcal{T}_t[\mathcal{T}_\tau(\varepsilon_i)x_{ij}] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\mathcal{T}_t[\mathcal{T}_\tau(\varepsilon_i)x_{ij}]\right|.$$

Using the fact that $|\mathcal{T}_t(w) - w| \leq |w|\mathbb{1}\{|w| > t\} \leq |w|^2/t$, Cauchy-Schwarz inequality, the bounds on moments in Assumptions **??**-**??**,

$$\max_j D_{1j} \leq \frac{1}{tn}\sum_{i=1}^n \mathbb{E}[|\mathcal{T}_\tau(\varepsilon_i)|^2 M^2(Z_i)] \leq \frac{\mathbb{E}[\tau^{1-\delta}|\varepsilon_1|^{1+\delta}M^2(Z_1)]}{t} \leq \frac{\sigma^2 v_\delta \tau^{1-\delta}}{t}.$$

Note that the augmented sequence $\{(Z_i, \varepsilon_i)\}_{i=1}^n$ is a stationary Markov chain. By a similar argument to that in Lemmas **??**-**??** and the law of large number for Markov chains (**?**, Theorem 17.1.2)

$$\max_j D_{2j} \leq \frac{1}{tn}\sum_{i=1}^n \tau^{1-\delta}|\varepsilon_i|^{1+\delta}M^2(Z_i)$$

$$= \frac{\tau^{1-\delta}(\mathbb{E}[|\varepsilon_1|^{1+\delta}M^2(Z_1)] + o_p(1))}{t} \leq \frac{\tau^{1-\delta}(\sigma^2 v_\delta + o_p(1))}{t}.$$

Noting that $|\mathcal{T}_t[W] - \mathbb{E}\mathcal{T}_t[W]| \leq 2t$ almost surely, we apply the Bernstein's inequality for Markov chains in (**?**, Theorem 1.1) and yields

$$\mathbb{P}(D_{3j} > \epsilon) \leq 2\exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma}\cdot V_{n,t} + 10t\epsilon}\right) \leq 2\exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma}\cdot \sigma^2 v_\delta \tau^{1-\delta} + 10t\epsilon}\right),$$

where

$$V_{n,t} = \frac{1}{n} \sum_{i=1}^{n} \text{Var}\{\mathcal{T}_t[\mathcal{T}_\tau(\varepsilon_i)x_{ij}]\} \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[|\mathcal{T}_\tau(\varepsilon_i)|^2 x_{ij}^2] \leq \sigma^2 v_\delta \tau^{1-\delta}.$$

A union bound delivers that

$$\mathbb{P}\left(\max_j D_{3j} > \epsilon\right) \leq 2d \exp\left(-\frac{n\epsilon^2}{\frac{1+\gamma}{1-\gamma} \cdot \sigma^2 v_\delta \tau^{1-\delta} + 10t\epsilon}\right).$$

Let $t = \frac{1+\gamma}{1-\gamma} \cdot \tau$ then

$$\max_j D_{3j} > \sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{2\sigma^2 v_\delta \tau^{1-\delta} \log d}{n}} + \frac{1+\gamma}{1-\gamma} \cdot \frac{20\tau \log d}{n}$$

with probability at most $2/d$. Collecting these pieces together completes the proof. $\square$

6 Discussion

Heavy-tailed errors and dependent observations are two stylized features of many real high-dimensional data. However, the current framework of high-dimensional regression assumes sub-Gaussian tails of errors and independent observations for the convenience of theoretical analyses. Our long-term goal is to generalize the current framework to cover real cases of both heavy-tailed errors and dependent observations. While (?) proposed AHR for the robust estimation against heavy-tailed errors, this paper makes progresses in the vertical direction and deals with dependent observations.

The technical key of AHR is the Bernstein-type inequality to quantitatively characterize the tradeoff between bias and robustness in the Huber regression so that the optimal robustification parameter is adaptively chosen. For the independent setup, Bernstein-type inequalities, especially those for unbounded random variables, have been well established. However, for the Markov-dependent setup, the existing Bernstein-type inequalities in the literature work for bounded functions of Markov chains only (??). We develop a truncation argument to extend the Bernstein-type inequality in (?) to unbounded functions of Markov chains. This truncation argument is potentially useful for other high-dimensional statistical problems in the Markov-dependent setup, e.g., the high-dimensional covariance matrix estimation on Markov-dependent samples.

In the Markov-dependent setup, we find that the τ-adaption and the $\boldsymbol{\beta}$-estimation of AHR exhibits a similar phase transition to that in the independent setup. The only difference is the sample size n is discounted by a factor $(1-\gamma)/(1+\gamma)$, where γ is the norm of Markov operator acting on the Hilbert space $L_2(\pi)$ and measures

the Markov dependence. If $\gamma = 0$ then the conclusion recovers the main result of **?** for the independent data samples. If $\gamma \to 1$ as $n \to \infty$ then the conclusion in Theorem **??** is still valid so long as

$$s\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log d}{n}} \to 0.$$

It is also valid if γ is replaced with an overestimate $\gamma' \geq \gamma$. Inspired by the discussion on the optimal variance proxies of concentration inequalities for functions of Markov chains in (**?**), we conjecture that the discounting factor $(1-\gamma)/(1+\gamma)$ is optimally sharp when translating the theories of high-dimensional linear regression from the independent setup to the Markov-dependent setup.

The theoretical and practical applicabilities of AHR on time series data with more complicated dependence structure are the potential directions of future works. Of particular interest is how the dependence of the time series data get involved in the τ-adaption and the β-estimation of AHR. However, the theoretical analysis of AHR in more complicated setups still hinges on the availability of more general Bernstein-type inequalities. Unfortunately, there are few concentration inequalities fulfilling the goal to theoretically justify AHR and other high-dimensional regression methods on data with a general dependence structure.