Markov Subsampling Based on Huber Criterion

Tieliang Gong, Yuxin Dong, Hong Chen, Bo Dong, *Member, IEEE*, and Chen Li

***Abstract*— Subsampling is an important technique to tackle the computational challenges brought by big data. Many subsampling procedures fall within the framework of importance sampling, which assigns high sampling probabilities to the samples appearing to have big impacts. When the noise level is high, those sampling procedures tend to pick many outliers and thus often do not perform satisfactorily in practice. To tackle this issue, we design a new Markov subsampling strategy based on Huber criterion (HMS) to construct an informative subset from the noisy full data; the constructed subset then serves as refined working data for efficient processing. HMS is built upon a Metropolis–Hasting procedure, where the inclusion probability of each sampling unit is determined using the Huber criterion to prevent over scoring the outliers. Under mild conditions, we show that the estimator based on the subsamples selected by HMS is statistically consistent with a sub-Gaussian deviation bound. The promising performance of HMS is demonstrated by extensive studies on large-scale simulations and real data examples.**

***Index Terms*— Markov chain, regression, robust inference, subsampling.**

I. Introduction

RAPID advancement in modern science and technology introduces data with extraordinary size and complexity, which brings great challenges to conventional machine learning and statistical methods. In the literature, two fundamental approaches have emerged to tackle the challenges of big data: one is the divide-and-conquer strategy [1], which involves partitioning the data into manageable segments, implementing a particular algorithm on these data segments in parallel, and synthesizing a global output by aggregating the segmental outputs; the other approach is the subsampling strategy [2], which involves selecting a representative subset from the full data as a surrogate and obtaining an output through further analyzation of the surrogate. The divide-and-conquer strategy usually relies on high computational power with computing clusters and is particularly effective when a dataset is too big to fit in one computer. However, it still consumes considerable computational resources and the access of distributed computational platforms is restricted by high cost. As a computationally cheaper alternative, subsampling gains its merit for the situation when the computational resources are limited.

The key task of subsampling is to effectively identify important samples in order to maintain the essential information of the full data. This task is particularly challenging for big data, which often comes with poor quality (high noise level) due to the uncontrolled collecting process. In the literature, informative sampling strategies are commonly adopted, where important samples are given high probabilities to be selected. During the last two decades, extensive studies have been concluded on informative sampling, e.g., statistical leverage score method [3]–[6], gradient method [7], and influence function method [8]. Leverage score subsampling assigns the sampling probabilities proportionally to a distance measure within the covariates. It does not consider the response and hence is sensitive to outliers. Both gradient-based subsampling and influence function-based subsampling are using the response together with the covariates to design sampling patterns, in which the probabilities are computed proportionally to the quadratic loss gradient or influence function. Although they do avoid the interference of outliers to some extent, the estimators calculated upon the associated subsamples are highly dependent on a reliable pilot model, which may be difficult to obtain in highly noisy setup.

Huber criterion [9] provides an effective way to deal with this situation. It is a hybrid of square loss for relatively small errors and absolute loss for relatively large ones and hence is robust to heavy-tailed errors and outliers. Recent studies have shown the great potential of Huber criterion for robust estimation and inference. For example, Lambert-Lacroix *et al.* [10] proposed to combine the Huber criterion and adaptive penalty as lasso and showed that the resulting estimator is more robust than adaptive LASSO in prediction and variable selection tasks. Wang *et al.* [11] developed data-driven Huber-type methods for regression tasks and established sub-Gaussian-type concentration bounds for the Huber-type estimator. In [12], the adaptive Huber regression method was proposed, which significantly outperforms least squares both in terms of mean and standard deviation. Besides, it admits exponential-type concentration bounds when the error variables have finite moments. Fan *et al.* [13] investigated the nonasymptotic consistency of ℓ_1 regularized robust M-estimator with Huber loss under the Markov chain

Manuscript received 13 May 2021; revised 25 February 2022; accepted 4 July 2022. Date of publication 14 July 2022; date of current version 6 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0110700; in part by the National Natural Science Foundation of China under Grant 62106191, Grant 12071166, Grant 62192781, and Grant 61721002; in part by the Innovation Research Team of Ministry of Education under Grant IRT 17R86; and in part by the Project of China Knowledge Centre for Engineering Science and Technology and Project of Chinese Academy of Engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China." *(Corresponding author: Tieliang Gong.)*

Tieliang Gong, Yuxin Dong, and Chen Li are with the Key Laboratory of Intelligent Networks and Network Security, Ministry of Education, and the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: adidasgtl@gmail.com; dongyuxin@stu.xjtu.edu.cn; cli@xjtu.edu.cn).

Hong Chen is with the College of Science, Huazhong Agriculture University, Wuhan 430070, China (e-mail: chenh@mail.hzau.edu.cn).

Bo Dong is with the School of Continuing Education, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: dong.bo@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2022.3189069

setting. Chen *et al.* [14] additionally investigated collinearity and explored the grouping effect in the Huber regression. Meyer [15] proposed an alternative probabilistic interpretation of minimizing the Huber loss, which is equivalent to minimizing an upper bound on the Kullback–Leibler (KL) divergence in Laplacian settings. Wang *et al.* [16] achieved robust forecasting based on the Huber criterion for both non-Gaussian and nonstationary data.

In light of these advances, we aim to design a robust subsampling procedure by adopting the Huber criterion. To this end, this article proposes a Markov subsampling strategy based on Huber criterion (HMS) for linear regression. The procedure is given as follows. We first obtain a rough estimator $\boldsymbol{\beta}_0$ based on a simple pilot selection, which determines the importance of each sample by calculating the Huber loss. We then perform subsampling from the full data $\mathcal{D}$ to generate a subset $\mathcal{D}_S$ through the Metropolis–Hasting (MH)-type procedure, where the sampling probability is assigned according to the Huber loss. By doing so, samples with large Huber loss are unlikely to be selected, and hence, the noisy samples and outliers are ruled out with high probability. Moreover, the MH sampling procedure and its variants require a proposal distribution to specify the sample importance, which is crucial to the success (e.g., fast convergence rate) of these algorithms, as improper selection of proposal distribution may result in misleading estimates. Different from MH-type algorithms, HMS determines the sample importance directly by the Huber criterion, where the turning parameter is prespecified through the data-driven strategy and hence avoids such a problem.

Our contributions are summarized as follows.

1) We develop a distribution-free HMS to construct an informative subset from the noisy full data, which further enables robust statistical inference and prediction.
2) Theoretically, we establish the statistical consistency for the regression estimator based on the subsample suggested by HMS in terms of Bahadur-type representation [17], [18]. Our results indicate that, with an appropriate robust parameter, the HMS-based estimator achieves a nearly optimal convergence rate. The theoretical results also extend the error analysis of Huber estimator under independent and identically distributed (i.i.d.) samples to a Markov-dependent setup.
3) Extensive empirical studies verify our theoretical findings. The promising performance of HMS estimator is also supported by both large-scale simulations and real data examples.

The rest of this article is organized as follows. Section II sets the notations and problem statement. Section III introduces the proposed Markov subsampling algorithm based on the Huber criterion. Section IV establishes the asymptotic analysis and the corresponding error bounds of the subsampling estimator. Section V demonstrates the experimental results on both simulation studies and real data examples. Section VI concludes our work.

II. Notations and Preliminaries

A. Notations

To make our arguments in the following, some concepts and notations being used throughout this article are introduced.

Let $\mathbf{u} = (u_1, u_2, \ldots, u_d)^\top \in \mathbb{R}^d$ and $p \geq 1$, and we denote the ℓ_p-norm and ℓ_∞-norm of $\mathbf{u}$ as $\|\mathbf{u}\|_p = (\sum_{i=1}^d |u_i|^p)^{1/p}$ and $\|\mathbf{u}\|_\infty = \max_{j\in[1,d]} |u_j|$. For any $\mathbf{w} \in \mathbb{R}^d$, $\langle \mathbf{u}, \mathbf{w}\rangle = \mathbf{u}^\top \mathbf{w}$. For two scalars a and b, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the corresponding spectral norm is defined by $\|\mathbf{A}\| = \max_{\mathbf{u}\in\mathbb{S}^{n-1}} \|\mathbf{A}\mathbf{u}\|_2$, where $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. If $\mathbf{A} \in \mathbb{R}^{n\times n}$, we denote the minimum and maximum eigenvalue of $\mathbf{A}$ by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$, respectively. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its gradient vector by $\nabla f \in \mathbb{R}^d$.

Definition 1 [19]: A random variable $X \in \mathbb{R}$ is said to be sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{s^2\sigma^2}{2}\right) \quad \forall s \in \mathbb{R}. \tag{1}$$

The following concepts are important in our theoretical analysis. Let $\{X_i\}_{i\geq 1}$ be a Markov chain on a general space $\mathcal{X}$ with invariant probability distribution π. Let $P(x, \mathrm{d}y)$ be a Markov transition kernel on a general space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and P^* be its adjoint. Denote $\mathcal{L}_2(\pi)$ by the Hilbert space consisting of square integrable functions with respect to π. For any function $h : \mathcal{X} \to \mathbb{R}$, we write $\pi(h) := \int h(x)\pi(\mathrm{d}x)$. Define the norm of $h \in \mathcal{L}_2(\pi)$ as $\|h\|_\pi = (\langle h, h\rangle)^{1/2}$. Let $P^t(x, \mathrm{d}y), (t \in \mathbb{N})$ be the t-step Markov transition kernel corresponding to P; then, for $i \in \mathbb{N}, x \in \mathcal{X}$ and a measurable set S, $P^t(x, S) = \Pr(X_{t+i} \in S|X_i = x)$. Following the above notations, we introduce the definitions of ergodicity and spectral gap for a Markov chain.

Definition 2: Let $M(x)$ be a nonnegative function. For an initial probability measure $\rho(\cdot)$ on $\mathcal{B}(\mathcal{X})$, a Markov chain is uniformly ergodic if

$$\|P^t(\rho, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq T(x)t^n \tag{2}$$

for some $T(x) < \infty$ and $t < 1$, where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm.

A Markov chain is geometrically ergodic if (2) holds for some $t < 1$, which eliminates the bounded assumption on $T(x)$. The dependence of a Markov chain can be characterized by the absolute spectral gap, defined as follows.

Definition 3 (Absolute Spectral Gap): A Markov operator P has a $\mathcal{L}_2$ spectral gap $1 - \lambda$ if

$$\lambda(P) := \sup\{\|Ph\|_\pi : \|h\|_\pi = 1, \pi(h) = 0\} < 1. \tag{3}$$

The quantity $1 - \lambda$ measures the convergence speed of a Markov chain toward its stationary distribution π [20]. A smaller λ usually implies faster convergence speed and less variable dependence.

B. Huber Regression

In this article, we consider the data generated from the following linear regression model:

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\beta}^*\rangle + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{4}$$

where y_i is the response, $\mathbf{x}_i \in \mathbb{R}^d$ is the covariate, ε_i is the error, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is the regression coefficient. It is well known that the ordinary least-squares estimator $\boldsymbol{\beta}_{\text{ols}}$ for (4) has a suboptimal polynomial-type deviation bound, which makes it inappropriate for large-scale estimation and inference. The key factor lies in the sensitivity of square loss to outliers [21]. To overcome this drawback, the Huber loss [9], [12] is proposed for achieving robust estimation. The Huber loss is defined by

$$\ell_\tau(x) = \begin{cases} x^2/2, & \text{if } |x| \le \tau \\ \tau|x| - \tau^2/2, & \text{if } |x| > \tau \end{cases} \tag{5}$$

where $\tau > 0$ is the robustification parameter that controls the bias and robustness. This function is quadratic with small values of τ while growing linearly for large values of τ. The specification of τ is critical in practical applications. Some recent studies on deviation bounds of Huber regression [11], [12] suggest that τ should be adaptive with the dimension of input space, the moment condition of the noise distribution, and the sample size to achieve robustness and unbiasedness estimate. Specifically, Sun *et al.* [12] obtained near-optimal deviation bounds of Huber regression for both low- and high-dimensional cases. These observations will motivate us to derive optimal bounds for HMS estimation.

Define the empirical loss function $L_\tau(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \ell_\tau(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle)$. The object of Huber regression is to find an optimizer of the following convex optimization problem:

$$\boldsymbol{\beta}^*_\tau = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} L_\tau(\boldsymbol{\beta}) \tag{6}$$

which can be easily solved via the iteratively reweighted least-squares method [22]. Denote the derivative of Huber loss $\ell_\tau(x)$ as φ_τ, i.e.,

$$\varphi_\tau = \operatorname{sign}(x)\min\{|x|, \tau\}, \quad x \in \mathbb{R}. \tag{7}$$

In this article, we focus on the setting that $n \gg d$. Denote $\mathbf{X}_S$ by the subsample matrix produced by HMS and $\bar{\mathbf{x}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$. Suppose that $\boldsymbol{\Sigma} = \mathbb{E}_\pi(\mathbf{X}_S\mathbf{X}_S^\top)$ is positive definite, and the regression errors ε_i satisfy $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0$ and $\upsilon_{i,\delta} = \mathbb{E}(|\varepsilon_i|^{1+\delta}|\mathbf{x}_i) < \infty$. With this setup, we write

$$\upsilon_\delta = \frac{1}{n}\sum_{i=1}^n \upsilon_{i,\delta} \quad \text{and} \quad u_\delta = \min\left\{\upsilon_\delta^{1/(1+\delta)}, \sqrt{\upsilon_1}\right\}, \quad \delta > 0.$$

III. Markov Subsampling Based on Huber Criterion

As discussed before, the currently used informative measures (leverage score, gradient, and influence function) in subsampling may not reflect the real contribution of each sample in highly noisy settings, and hence, the resulting estimator can be misleading. To alleviate this issue, we develop a HMS to achieve robust estimation. The core idea is to select the samples with small errors based on the Huber criterion by the Markov chain Monte Carlo (MCMC) method. Concretely, HMS consists of three steps: 1) pilot estimation; 2) Huber loss calculation; and 3) Markov subsampling.

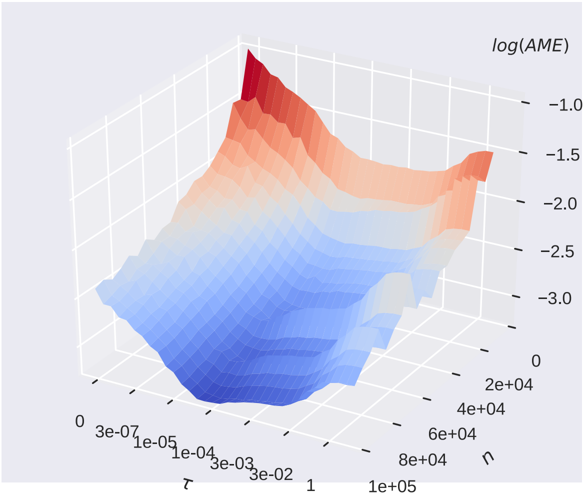

Fig. 1. log(AME) versus τ and n [AME: averaged mean error, defined in (43)]. Here, we generate the data by (4), where $n = 1M$, $d = 500$ and ε_i are i.i.d. from Student-t distribution with degree of freedom 2.

1) *Pilot Estimation:* The idea of pilot is widely applied in the subsampling procedure [7], [8], [23], [24], where the sampling probability is specified by a pilot estimation. A popular way for calculating pilot is uniform subsampling. To avoid bringing additional computational burden, we suggest the pilot $\boldsymbol{\beta}_0$ to be calculated by least-squares criterion based on a small random subset with user preference size $d < r \ll n$, i.e., $\boldsymbol{\beta}_0 = (\mathbf{X}_r^\top\mathbf{X}_r)^{-1}\mathbf{X}_r^\top\mathbf{y}_r$. It only takes additional $\mathcal{O}(rd^2)$ CPU time. We empirically demonstrate that the HMS estimator does not rely heavily on the quality of $\boldsymbol{\beta}_0$.
2) *Huber Loss Calculation:* The robustification parameter τ in the Huber criterion plays a tradeoff role between bias and robustness. In practical, τ is usually set to be fixed through 95% asymptotic efficiency rule [9], [18], [25], [26]. However, a fixed value may not guarantee a good estimator, especially in highly noisy cases. As shown in Fig. 1, τ should be adapted with n and d (consider that $n \gg d$, we ignore the effect of d). It can be seen that there exists some τ such that the absolute mean error (AME) of $\boldsymbol{\beta}_0$ achieves minimum for a fixed sample size n. In practice, we first restrict τ in a reasonable range and select the optimal value then according to the minimal AME principal. After specifying τ, the importance of a sample $(\mathbf{x}_i, y_i)$ can be measured by the corresponding Huber loss $\ell_\tau(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}_0)$. The greater importance of a sample often comes with smaller Huber loss.
3) *Markov Subsampling:* It has been shown that the Markov chain samples may lead to more robust estimation than i.i.d. counterparts in machine learning [27] and optimization tasks [28], [29]. With this in mind, we tend to implement probabilistic sampling through an MH-type procedure. The core step, probabilistic acceptance rule, is designed based on the Huber criterion. Concretely, at some current sample $\mathbf{z}_t$, a randomly selected candidate

Algorithm 1 Huber Regression With Markov Subsampling

```
Input: Dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$, subset $\mathcal{D}_S = \emptyset$, robustification parameter $\tau$, burn-in period: $t_0$, subsample size $n_{sub} \ll n$.
Train a pilot estimator $\boldsymbol{\beta}_0$ by $\boldsymbol{\beta}_0 = (\mathbf{X}_r^\top \mathbf{X}_r)^{-1}\mathbf{X}_r^\top \mathbf{y}_r$, where $(\mathbf{X}_r, \mathbf{y}_r)$ are the random subsamples with size $n_0 = n_{sub}$.
Randomly select a sample $\mathbf{z}_1$ from $\mathcal{D}$, and set $\mathcal{D}_s = \mathbf{z}_1$.
for $2 \le T \le (n_{sub} + t_0)$ do
  while $|\mathcal{D}_S| < T$ do
    Randomly draw a candidate $\mathbf{z}^* = (\mathbf{x}^*, y^*)$
    Calculate the acceptance probability by
        $$p = \min\left\{1, \frac{\ell_\tau(y_T - \langle \mathbf{x}_T, \boldsymbol{\beta}_0 \rangle)}{\ell_\tau(y^* - \langle \mathbf{x}^*, \boldsymbol{\beta}_0 \rangle)}\right\} \qquad (8)$$
    Set $\mathcal{D}_S = \mathcal{D}_S \cup \mathbf{z}^*$ with probability $p$
    If $\mathbf{z}^*$ is accepted, set $\mathbf{z}_{t+1} = \mathbf{z}^*$
  end while
end for
Denote the last $n_{sub}$ samples as $\mathcal{D}_S = (\mathbf{X}_S, \mathbf{y}_S) = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{sub}}$.
Solve $\boldsymbol{\beta}_\tau$ by Huber regression (6) based on $\mathcal{D}_S$.
Output: $\boldsymbol{\beta}_\tau$.
```

sample $\mathbf{z}^*$ is accepted with probability defined in (8). If $\mathbf{z}^*$ is accepted, we set $\mathcal{D}_S = \mathcal{D}_S \cup \mathbf{z}^*$ and $\mathbf{z}^* = \mathbf{z}_{t+1}$; otherwise, we randomly select a sample as a candidate and repeat this process. Finally, we accept the last n_{sub} elements generated by this procedure after a user-specified burn-in period.

The detailed procedures are summarized in Algorithm 1. Note that the probabilistic acceptance rule (8) tends to select the samples with small Huber loss with high probability. Moreover, the subsamples generated by Algorithm 1 constitute an irreducible Markov chain and therefore are uniformly ergodic [30], [31]. Computationally, HMS takes $\mathcal{O}(n_0 d^2)$ time for pilot estimation, $\mathcal{O}((n_{\text{sub}} + t_0)d)$ time for MH sampling procedure, and $\mathcal{O}(n_{\text{sub}} d^2)$ time for optimizing (6) (the L-BFGS-B optimization strategy [32] is adopted). Hence, the total time complexity is $\mathcal{O}((2n_{\text{sub}} + t_0)d^2)$, which is much saving computational cost since $n_{\text{sub}}, t_0 \ll n$.

IV. Theoretical Assessments of HMS Estimator

In this section, we provide theoretical support for the proposed HMS. In particular, we aim at bounding the difference between the HMS estimator $\boldsymbol{\beta}_\tau$ and the oracle $\boldsymbol{\beta}^*$. Previous theoretical studies on subsample estimator are based on least squares [2], [5], [7], which has a closed-form solution. However, the HMS estimator does not admit an explicit closed-form representation and the robustification parameter τ is not fixed, and all these pose the difficulties in analyzing its statistical properties. To overcome these issues, we adopt the Lepski-type method developed in [12]. We first present several necessary assumptions as follows.

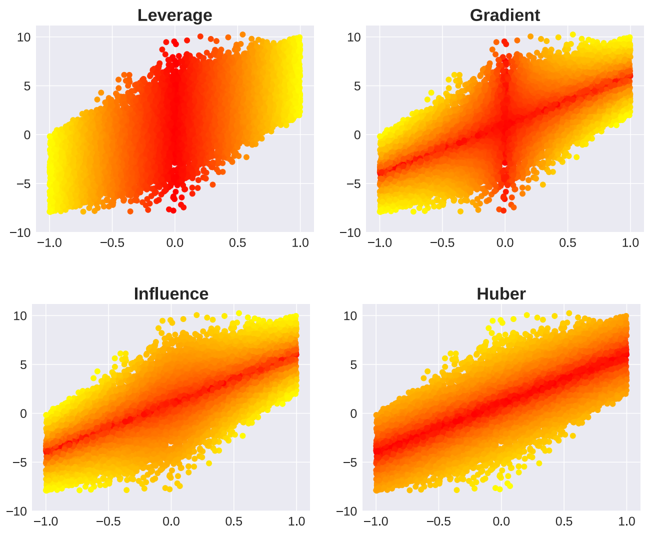

Fig. 2. Illustration of sampling probabilities with different importance measures. The data are generated by $y = 5x + 1 + \varepsilon$ where $n = 10\,000$ and ε is a mixture of Gaussian and uniform distribution. The bright yellow and red represent high and low sampling probability, respectively. Both leverage score and gradient are inclined to select the points with large residuals near the center. The influence tends to balance the regression design and the residual. The HMS can further enhance the effect of the influence.

Assumption 1 [30] (Nonzero Spectral Gap Markov Chain): The underlying Markov chain $\{X_i\}_{i=1}^n$ is stationary with unique invariant measure π and admits an absolute spectral gap $1-\lambda$.

Assumption 2 (Bounded Covariates): There exists an envelope function $M : X \to \mathbb{R}$, such that for any function f, $\max |f(X)| \le M(X)$ for π-almost every X.

Assumption 3 (Bounded $(1+\delta)$-Moments of Errors): $\mathbb{E}(\varepsilon_i|X_i) = 0$. For some $\delta > 0$ and $\upsilon_\delta > 0$, $\mathbb{E}[|\varepsilon_i|^{1+\delta}|X_i] < \upsilon_\delta$.

The absolute spectral gap $1-\lambda$ in Assumption 1 usually involves in spectral radius and geometrical ergodicity. Given the transition kernel P of a Markov chain, denote its spectral radius by $\lambda_\infty(P) = \lim_{k\to\infty} \|P^k - \pi(\cdot)\|_\pi^{1/k}$. It is known that $\lambda_\infty(P) \le \lambda(P)$ [30], where the equality holds for reversible Markov chain. The condition $1-\lambda(P) > 0$ implies geometrical ergodicity. A nonzero spectral gap is closely related to other convergence criterion of Markov chains [33]. Assumption 2 requires that the covariates are bounded by an envelope function, which can be a function of time, space, or any forms of random variable. The boundedness assumption is quite common in statistics and learning theory analysis [34], [35]. Assumption 3 requires errors to be with finite conditional $(1+\delta)$-moments, which covers a broad range of heavy-tailed noises, including the Student-t, the Pareto, log Normal, and log Gamma. Now, we are ready to present the main results for HMS estimator.

The following lemmas play an important role to prove our main theoretical results, where Lemma 1 is the Bernstein inequality within Markov-dependent setting, Lemma 2 gives the localized analysis on bounding β_η, and Lemma 3 presents the upper bound on the ℓ_2 error between an estimation $\boldsymbol{\beta}$ from a d-dimensional hypersphere and $\boldsymbol{\beta}^*$.

Lemma 1 [36]: Let $\{X_i\}_{i\ge 1}$ be a stationary Markov chain with invariant distribution π and right L_2-spectral gap $1-\lambda \in$

$(0, 1]$. Let $f_i : \mathcal{X} \to [-c, c]$ be a bounded function with $\pi(f_i) = 0$ and $\sigma^2 = \sum_{i=1}^{n} \pi(f_i^2)/n$. Then, for any $0 \leq t \leq (1-\lambda)/5c$, we have for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2(A_1\sigma^2 + A_2 c\epsilon)}\right) \tag{9}$$

where $A_1 = ((1+\lambda)/(1-\lambda))$ and $A_2 = (\mathbf{1}/3)\mathbf{1}_{\lambda=0} + (5/(1-\lambda))\mathbf{1}_{\lambda>0}$.

Lemma 2 [37]: Suppose that L is a convex function. Let $D_L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = L(\boldsymbol{\beta}_1) - L(\boldsymbol{\beta}_2) - \langle \nabla L(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle$ and $\bar{D}_L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = D_L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + D_L(\boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$. For $\boldsymbol{\beta}_\eta = \boldsymbol{\beta}^* + \eta(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ with $\eta \in (0, 1]$

$$\bar{D}_L(\boldsymbol{\beta}_\eta, \boldsymbol{\beta}^*) \leq \eta \bar{D}_L(\boldsymbol{\beta}, \boldsymbol{\beta}^*). \tag{10}$$

Lemma 3 [12]: Suppose that $\upsilon_\delta < \infty$ for some $0 < \delta \leq 1$ and $(\mathbb{E}\langle \mathbf{u}, \bar{\mathbf{x}}\rangle^4)^{1/4} \leq C\|\mathbf{u}\|_2$ for all $\mathbf{u} \in \mathbb{R}^d$ and some constant $C > 0$. Moreover, let $\tau, r > 0$ satisfy $\tau \geq 2\max\{(4\upsilon_\delta)^{1/(1+\delta)}, 4C^2 r\}$ and $n \geq (\tau/r)^2(d+t)$. Then, with probability at least $1 - e^{-t}$

$$\langle \nabla L_\tau(\boldsymbol{\beta}) - \nabla L_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \geq \frac{1}{4}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 \tag{11}$$

uniformly over

$$\boldsymbol{\beta} \in \mathbf{B}_0(r) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2 \leq r\}.$$

Proposition 1 provides a concentration inequality for $\|\boldsymbol{\Sigma}^{-1/2}\nabla L_\tau(\boldsymbol{\beta}^*)\|_2$, which is fundamental to our theoretical analysis.

Proposition 1: Suppose that the Markov chain samples generated by Algorithm 1 are with invariant distribution π and satisfy Assumptions 1–3; then, for any $0 < \delta \leq 1$

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla L_\tau(\boldsymbol{\beta}^*)\|_2 \leq \frac{4\sqrt{\pi}C_0 A_2(d+t)\tau}{n_{\text{sub}}} + 4C_0\sqrt{\frac{A_1\upsilon_\delta\tau^{1-\delta}(d+t)}{n_{\text{sub}}}} + \upsilon_\delta\tau^{-\delta} \tag{12}$$

holds with confidence at least $1 - 2e^{-t}$, where $A_1 = ((1+\lambda)/(1-\lambda))$ and $A_2 = (\mathbf{1}/3)\mathbf{1}(\lambda \leq 0) + (5/(1-\lambda))\, 1(\lambda > 0)$.

Proof: To bound $\|\boldsymbol{\Sigma}^{-1/2}\nabla L_\tau(\boldsymbol{\beta}^*)\|_2$, we first define a random vector

$$\begin{aligned}\boldsymbol{\zeta}^* &= \boldsymbol{\Sigma}^{-1/2}\{\nabla L_\tau(\boldsymbol{\beta}^*) - \nabla\mathbb{E}L_\tau(\boldsymbol{\beta}^*)\}\\ &= -\frac{1}{n_{\text{sub}}}\sum_{i=1}^{n_{\text{sub}}}\{\zeta_i\bar{\mathbf{x}}_i - \mathbb{E}(\zeta_i\bar{\mathbf{x}}_i)\}\end{aligned} \tag{13}$$

where $\zeta_i = \varphi_\tau(\varepsilon_i)$ and $\bar{\mathbf{x}}_i = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i$ with $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ being positive. Assume that there exists a 1/2-Net $\mathcal{N}_{1/2}$ of the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$ with $|\mathcal{N}_{1/2}| \leq 2^d$ such that $\|\boldsymbol{\zeta}^*\|_2 \leq 2\max_{\mathbf{u}\in\mathcal{N}_{1/2}}|\langle u, \boldsymbol{\zeta}^*\rangle|$. Without loss of generality, we assume that $\mathbf{x}_i$'s are centralized. By Assumption 2, we know that $\mathbf{x}_i$ are sub-Gaussian vectors, i.e.,

$$\mathbb{P}(|\langle \mathbf{u}, \bar{\mathbf{x}}\rangle| \geq p) \leq \exp\left(-p^2\|u\|_2^2/C_0^2\right) \tag{14}$$

for any $\mathbf{u} \in \mathbb{S}^{d-1}$ and $p \in \mathbb{R}$, where C_0 is a positive constant. We then have

$$\mathbb{E}|\langle \mathbf{u}, \bar{\mathbf{x}}_i\rangle|^k \leq C_0^k k\Gamma(k/2), \quad k \geq 1. \tag{15}$$

It immediately implies

$$\begin{aligned}\sum_{i=1}^{n_{\text{sub}}}\mathbb{E}(\zeta_i\langle \mathbf{u}, \bar{\mathbf{x}}_i\ \rangle)^2 &\leq 2C_0^2\tau^{1-\delta}\sum_{i=1}^{n_{\text{sub}}}\upsilon_{i,1} = 2C_0^2 n\upsilon_\delta\tau^{1-\delta}\\ \sum_{i=1}^{n_{\text{sub}}}\mathbb{E}(\zeta_i\langle \mathbf{u}, \bar{\mathbf{x}}_i\ \rangle)^k &\leq \frac{k!}{2}(C_0\tau/2)^{k-2}2C_0^2 n_{\text{sub}}\upsilon_\delta\tau^{1-\delta}\end{aligned} \tag{16}$$

for $k \geq 3$. Furthermore, by Assumption 3, we have

$$\mathbb{E}[\varphi_\tau(\varepsilon)] = -\mathbb{E}[(\varepsilon - \tau)\mathbf{1}(\varepsilon > \tau)] + \mathbb{E}[(-\varepsilon - \tau)\mathbf{1}(\varepsilon < -\tau)]. \tag{17}$$

Thus, for any $k > 2$

$$|\mathbb{E}\varphi_\tau(\varepsilon)| \leq \mathbb{E}[(|\varepsilon| - \tau)\mathbf{1}(|\varepsilon| > \tau)] \leq \tau^{1-k}\mathbb{E}[|\varepsilon|^k].$$

It follows from Lemma 1 with $c = \sqrt{\pi}C_0\tau$ and $\sigma^2 = 2C_0^2\upsilon_\delta\tau^{1-\delta}$ that:

$$\mathbb{P}\left\{|\langle \mathbf{u}, \boldsymbol{\zeta}^*\rangle| \leq \frac{2\sqrt{\pi}C_0A_2\omega\tau}{n_{\text{sub}}} + 2C_0\sqrt{\frac{A_1\upsilon_\delta\tau^{t-1}\omega}{n_{\text{sub}}}}\right\} \geq 1 - 2e^{-\omega} \tag{18}$$

for $\forall\omega > 0$. By taking the union bound over $\mathbf{u} \in \mathcal{N}_{1/2}$, the following inequality:

$$\|\boldsymbol{\zeta}^*\|_2 \leq \frac{4\sqrt{\pi}C_0A_2\omega\tau}{n} + 4C_0\sqrt{\frac{A_1\upsilon_\delta\tau^{1-\delta}\omega}{n}} \tag{19}$$

holds with confidence at least $1 - 2^{d+1}\cdot e^{-\omega}$. Then, we consider the deterministic part $\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathbb{E}L_\tau(\boldsymbol{\beta}^*)\|_2$, by direct calculation

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathbb{E}L_\tau(\boldsymbol{\beta}^*)\|_2 \leq \sup_{\mathbf{u}\in\mathbb{S}^{d-1}}\frac{1}{n_{\text{sub}}}\sum_{i=1}^{n_{\text{sub}}}\mathbb{E}|\zeta_i\langle \mathbf{u}, \bar{\mathbf{x}}_i\rangle| \leq \upsilon_\delta\tau^{-\delta}.$$

Let $\omega = d + t$, and by combining above inequality and (19), we obtain the stated result. ■

Theorem 1: Suppose that the Markov chain samples generated by Algorithm 1 are with invariant distribution π and satisfy Assumptions 1–3; then, for any $t > 0$, with confidence at least $1 - 2e^{-t}$, the HMS estimator $\boldsymbol{\beta}_\tau$ with $\tau = (1/A_\lambda)(n_{\text{sub}}/(d+t))^{\max\{(1/(1+\delta)),(1/2)\}}$ satisfies

$$\|\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*\|_2 \leq C_1\lambda_{\max}(\boldsymbol{\Sigma}^{1/2})A_\lambda\left(\frac{d+t}{n_{\text{sub}}}\right)^{\min\left\{\frac{\delta}{1+\delta},\frac{1}{2}\right\}} \tag{20}$$

provided that $n_{\text{sub}} \geq C_2(d+t)$, where $C_1, C_2 > 0$ are the constants independent of n and d, $A_\lambda = \max\{((1+\lambda)/(1-\lambda))^{1/2}, (1/3)\mathbf{1}_{\lambda=0} + (5/(1-\lambda))\mathbf{1}_{\lambda>0}\}$.

Proof: To begin with, recall that $\mathbf{B}_0(r) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2 \leq r\}$ for some $r > 0$. Define $\boldsymbol{\beta}_{\tau,\eta} := \boldsymbol{\beta}^* + \eta(\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*) \in \mathbf{B}_0(r)$, where $\eta \in (0, 1]$. Then, we know from Lemma 2 that

$$\begin{aligned}&\langle \nabla L_\tau(\boldsymbol{\beta}_{\tau,\eta}) - \nabla L_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta}_{\tau,\eta} - \boldsymbol{\beta}^*\rangle\\ &\quad\leq \eta\langle \nabla L_\tau(\boldsymbol{\beta}_\tau) - \nabla L_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*\rangle.\end{aligned} \tag{21}$$

It is easy to see $\nabla L_\tau(\boldsymbol{\beta}_\tau) = 0$ due to the Karush–Kuhn–Tucker (KKT) condition. According to the mean value theorem

$$\nabla L_\tau(\boldsymbol{\beta}_{\tau,\eta}) - \nabla L_\tau(\boldsymbol{\beta}^*)$$

$$= \int_0^1 \nabla^2 L_\tau(t\boldsymbol{\beta}_{\tau,\eta} + (1-t)\boldsymbol{\beta}^*)\,\mathrm{d}t(\boldsymbol{\beta}_{\tau,\eta} - \boldsymbol{\beta}^*). \quad (22)$$

Assume that there exist a constant $c > 0$ such that

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le r} \lambda_{\min}(\nabla^2 L_\tau(\boldsymbol{\beta})) \ge C_0$$

and, hence, $C_0\|\boldsymbol{\beta}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2^2 \le \|\nabla L_\tau(\boldsymbol{\beta}^*)\|_2 \cdot \|\boldsymbol{\beta}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2$; reducing the result yields

$$\|\boldsymbol{\beta}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2 \le C_0^{-1}\|\nabla L_\tau(\boldsymbol{\beta}^*)\|_2. \quad (23)$$

Since $\boldsymbol{\beta}_{\tau,\eta} \in \mathbf{B}_0(r)$, according to Lemma 3 with $r = \tau/(4C_0^2)$, we get

$$\langle \nabla L_\tau(\boldsymbol{\beta}_{\tau,\eta}) - \nabla L_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \ge \frac{1}{4}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_{\tau,\eta} - \boldsymbol{\beta}^*)\|_2^2 \quad (24)$$

with confidence at least $1 - e^{-t}$. Then, by Proposition 1,

$$\begin{aligned}\|\boldsymbol{\Sigma}^{-1/2}\nabla L_\tau(\boldsymbol{\beta}^*)\|_2 \le&\ \frac{4\sqrt{\pi}C_0 A_2(d+t)\tau}{n_{\text{sub}}} + \upsilon_\delta \tau^{-\delta}\\ &+ 4C_0\sqrt{\frac{A_1\upsilon_\delta\tau^{1-\delta}(d+t)}{n_{\text{sub}}}}\\ :=&\ r^* \end{aligned} \quad (25)$$

holds with confidence at least $1 - e^{-t}$. Combining (24) and (25), we know that with confidence at least $1 - 2e^{-t}$,

$$\|\boldsymbol{\beta}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2 \le 4\ r^* \quad (26)$$

provided that $n \ge C_1(d + t)$, where $C_1 > 0$ is a constant depending only on C_0. The constructed estimator $\boldsymbol{\beta}_{\tau,\eta}$ lies in the interior of the ball with radius r. By the construction in the beginning of the proof, this enforces $\eta = 1$, and thus, $\boldsymbol{\beta}_\tau = \boldsymbol{\beta}_{\tau,\eta}$. This completes the proof. ■

Remark 1: Theorem 1 indicates that the HMS estimator $\boldsymbol{\beta}_\tau$ is consistent under moderate conditions, i.e., $\|\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*\| \to 0$ as $n_{\text{sub}} \to \infty$. The founding condition requires that the Markov chain generated by Algorithm 1 has absolute spectral gap. HMS almost trivially meets this condition since the corresponding Markov chain is uniformly ergodic and hence geometrically ergodic. Moreover, the error bound of HMS only requires finite moments of error ε_i, which is weaker than sub-Gaussian error condition in linear regression models for subsampling [2], [7], [38]. We find that τ should adapt with subsample size n_{sub}, the input dimension d, the moments of error term, and the dependence of underlying Markov chain. In particular, with an appropriate choice of τ, the convergence rate of HMS estimator is with $\mathcal{O}((d/n_{\text{sub}})^{\min\{(\delta/(1+\delta)),(1/2)\}})$ decay, which matches the near-optimal deviations in the i.i.d. case [12]. Note that the Markov dependence impacts on τ in the way that the subsample size n_{sub} is discounted by a factor A_λ. In other words, in order to achieve τ-adaptation effect, the required subsample size increases with A_λ when transferring from i.i.d. sample setup to Markov dependence setup. Furthermore, a small value for λ implies a fast convergence rate of HMS estimator.

Theorem 2: Under the same conditions with Theorem 1, for any $t > 0$, the HMS estimator $\boldsymbol{\beta}_\tau$ with $\tau = ((1-\lambda)/(1+\lambda))^{1/2}(n_{\text{sub}}/((d+t)\log d))^{(1/(2(1+\delta)))}$ satisfies

$$\begin{aligned}\mathbb{P}\Bigg\{\Bigg\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*) - \frac{1}{n_{\text{sub}}}\sum_{i=1}^{n_{\text{sub}}}\varphi_\tau(\varepsilon_i)\bar{\mathbf{x}}_i\Bigg\|_2 \\ \ge C_3\sqrt{\frac{1+\lambda}{1-\lambda}}\sqrt{\frac{(d+t)\log d}{n_{\text{sub}}}}\Bigg\} \le 3\ e^{-t}\end{aligned} \quad (27)$$

provided that $n_{\text{sub}} \ge C_4(d + t)$, where C_3 and C_4 are the constants independent n and d.

Proof: If $r_1 = 4\ r^*$, we know from the proof of Theorem 1 that

$$\mathbb{P}\big\{\boldsymbol{\beta}_\tau \in \mathbf{B}_0(r_1)\big\} \ge 1 - 2\ e^{-t} \quad (28)$$

provided that $n_{\text{sub}} \ge C_1(d + t)$. Define the random process $\Phi(\boldsymbol{\beta}) = L_\tau(\boldsymbol{\beta}) - \mathbb{E}L_\tau(\boldsymbol{\beta})$ and

$$\Psi(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla L_\tau(\boldsymbol{\beta}) - \nabla L_\tau(\boldsymbol{\beta}^*)\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \quad (29)$$

Our goal is to bound $\|\Psi(\boldsymbol{\beta}_\tau)\|_2 = \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*) + \boldsymbol{\Sigma}^{-1/2}\nabla L_\tau(\boldsymbol{\beta}^*)\|_2$, and the key step lies in bounding the supremum of empirical process $\{\Psi(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbf{B}_0(r)\}$. To achieve this goal, we need to bound $\mathbb{E}\Psi(\boldsymbol{\beta})$ and $\Psi(\boldsymbol{\beta}) - \mathbb{E}\Psi(\boldsymbol{\beta})$.

Denote $\hat{\boldsymbol{\beta}}$ as the convex combination of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. By the mean value theorem, we see that

$$\begin{aligned}\mathbb{E}\Psi(\boldsymbol{\beta}) &= \boldsymbol{\Sigma}^{-1/2}\{\nabla\mathbb{E}L_\tau(\boldsymbol{\beta}) - \nabla\mathbb{E}L_\tau(\boldsymbol{\beta}^*)\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\\ &= \{\boldsymbol{\Sigma}^{-1/2}\nabla^2\mathbb{E}L_\tau(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_d\}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\end{aligned} \quad (30)$$

hence

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbf{B}_0(r)} &\|\mathbb{E}\Psi(\boldsymbol{\beta})\|_2\\ &\le r \times \sup_{\boldsymbol{\beta}\in\mathbf{B}_0(r)} \|\boldsymbol{\Sigma}^{-1/2}\nabla^2\mathbb{E}L_\tau(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_d\|.\end{aligned} \quad (31)$$

We know from Assumption 2 that $\|\mathbf{x}_i\|_\infty \le M(\mathbf{x})$, where $M(\mathbf{x}) : \mathbf{x} \to \mathbb{R}$ is a envelope function. Consider $\boldsymbol{\beta} \in \mathbf{B}_0(r)$ and $\mathbf{u} \in \mathbb{S}^{d-1}$, and we have

$$\begin{aligned}&|\mathbf{u}^\top\{\boldsymbol{\Sigma}^{-1/2}\nabla^2\mathbb{E}L_\tau(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_d\}\mathbf{u}|\\ &= \frac{1}{n_{\text{sub}}}\sum_{i=1}^{n_{\text{sub}}}\mathbb{E}\big\{\mathbf{1}\{y_i - \langle\mathbf{x}_i, \boldsymbol{\beta}\rangle \ge \tau\}\langle\mathbf{u}, \bar{\mathbf{x}}_i\rangle^2\big\}\\ &\le \frac{1}{n_{\text{sub}}}\sum_{i=1}^{n_{\text{sub}}}\mathbb{E}\Big\{(\mathbf{1}\{|\varepsilon_i| \ge \tau/2\}\\ &\qquad + \mathbf{1}\{\mathbf{x}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*) > \tau/2\})\langle\mathbf{u}, \bar{\mathbf{x}}_i\rangle^2\Big\}\\ &\le \frac{1}{n_{\text{sub}}}\sum_{i=1}^{n_{\text{sub}}}\mathbb{E}\big\{(\mathbf{1}\{|\varepsilon_i| \ge \tau/2\} + \mathbf{1}\{\|\mathbf{x}_i\|_\infty > \tau/2r\})\langle\mathbf{u}, \bar{\mathbf{x}}_i\rangle^2\big\}\\ &\le \frac{1}{n_{\text{sub}}}\sum_{i=1}^{n_{\text{sub}}}\mathbb{E}\big\{(\mathbf{1}\{|\varepsilon_i| \ge \tau/2\} + \mathbf{1}\{\|M(\mathbf{x}) > \tau/2r\})\langle\mathbf{u}, \bar{\mathbf{x}}_i\rangle^2\big\}\\ &\le 2^{1+\delta}\sigma^2\tau^{-1-\delta}\upsilon_\delta + \sqrt{\frac{A_1\log d}{n_{\text{sub}}}} + 4CA_1\sigma^4 r^2\end{aligned} \quad (32)$$

which implies

$$\sup_{\boldsymbol{\beta}\in\mathbf{B}_0(r)} \|\mathbb{E}\Psi(\boldsymbol{\beta})\|_2 \le 2^{1+\delta}\sigma^2\tau^{-1-\delta}\upsilon_\delta + \sqrt{\frac{A_1\log d}{n_{\text{sub}}}} + 4CA_1\sigma^4 r^2. \quad (33)$$

Next, we focus on bounding $\Psi(\boldsymbol{\beta}) - \mathbb{E}\Psi(\boldsymbol{\beta})$. To this end, we first rewrite

$$\Psi(\boldsymbol{\beta}) - \mathbb{E}\Psi(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla\Phi(\boldsymbol{\beta}) - \nabla\Phi(\boldsymbol{\beta}^*)\}. \quad (34)$$

Set

$$\Delta = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \quad (35)$$

and define the empirical process

$$\bar{\Psi}(\Delta) := \Psi(\boldsymbol{\beta}) - \mathbb{E}\Psi(\boldsymbol{\beta}). \quad (36)$$

It is easy to check that $\bar{\Psi}(0) = 0$ and $\mathbb{E}\bar{\Psi}(\Delta) = 0$. For any $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$ and $m \in \mathbb{R}$

$$\begin{aligned}
&\mathbb{E}\{m\sqrt{n}\mathbf{u}^\top\nabla_\Delta\bar{\Psi}(\Delta)\mathbf{v}\}\\
&\le \prod_{i=1}^{n_{\text{sub}}}\Bigg\{1 + \frac{m^2}{n_{\text{sub}}}\mathbb{E}\Big[\Big(\langle\mathbf{u},\bar{\mathbf{x}}_i\rangle^2\langle\mathbf{v},\bar{\mathbf{x}}_i\rangle^2 + \mathbb{E}|\langle\mathbf{u},\bar{\mathbf{x}}\rangle^2\langle\mathbf{v},\bar{\mathbf{x}}\rangle|^2\Big)\\
&\qquad\qquad \times e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}\left(|\langle\mathbf{u},\bar{\mathbf{x}}_i\rangle\langle\mathbf{v},\bar{\mathbf{x}}_i\rangle| + \mathbb{E}|\langle\mathbf{u},\bar{\mathbf{x}}\rangle^2\langle\mathbf{v},\bar{\mathbf{x}}\rangle|\right)}\Big]\Bigg\}\\
&\le \prod_{i=1}^{n_{\text{sub}}}\Bigg\{1 + e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}}\frac{m^2}{n_{\text{sub}}}\mathbb{E}\Big[e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}}|\langle\mathbf{u},\bar{\mathbf{x}}_i\rangle\langle\mathbf{v},\bar{\mathbf{x}}_i\rangle|\Big]\\
&\qquad\qquad + e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}}\frac{m^2}{n_{\text{sub}}}\mathbb{E}\Big[\langle\mathbf{u},\bar{\mathbf{x}}_i\rangle^2\langle\mathbf{v},\bar{\mathbf{x}}_i\rangle^2 e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}|\langle\mathbf{u},\bar{\mathbf{x}}_i\rangle\langle\mathbf{v},\bar{\mathbf{x}}_i\rangle|}\Big]\Bigg\}\\
&\le \prod_{i=1}^{n}\Bigg\{1 + e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}}\frac{m^2}{n_{\text{sub}}}\max_{\mathbf{w}\in\mathbb{S}^{d-1}}\mathbb{E}\Big[e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}\langle\mathbf{w},\bar{\mathbf{x}}\rangle^2}\Big]\\
&\qquad\qquad + e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}}\frac{m^2}{n_{\text{sub}}}\max_{\mathbf{w}\in\mathbb{S}^{d-1}}\mathbb{E}\Big[\langle\mathbf{w},\bar{\mathbf{x}}\rangle^4 e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}\langle\mathbf{w},\bar{\mathbf{x}}\rangle^2}\Big]\Bigg\}\\
&\le \exp\Bigg\{m^2 e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}}\Bigg(\max_{\mathbf{w}\in\mathbb{S}^{d-1}}\mathbb{E}\Big[e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}\langle\mathbf{w},\bar{\mathbf{x}}\rangle^2}\Big]\\
&\qquad\qquad + \max_{\mathbf{w}\in\mathbb{S}^{d-1}}\mathbb{E}\Big[\langle\mathbf{w},\bar{\mathbf{x}}\rangle^4 e^{\frac{|m|}{\sqrt{n_{\text{sub}}}}\langle\mathbf{w},\bar{\mathbf{x}}\rangle^2}\Big]\Bigg)\Bigg\}. \quad (37)
\end{aligned}$$

Recall that each $\mathbf{x}_i$ is sub-Gaussian random variable, and hence, there exist constants A_3 and A_4 that depend only on C_0 such that for any $|m| \le (n_{\text{sub}}/A_3)^{1/2}$

$$\sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{d-1}}\mathbb{E}\{m\sqrt{n_{\text{sub}}}\mathbf{u}^\top\nabla_\Delta\bar{\Psi}(\Delta)\mathbf{v}\} \le \exp\{A_4 m^2/2\}. \quad (38)$$

In [39, Th. A.3], we see that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\beta}\in\mathbf{B}_0(r)}\|\Psi(\boldsymbol{\beta}) - \mathbb{E}\Psi(\boldsymbol{\beta})\|_2 \le 6A_4 r\sqrt{8d+2t}\right\} \ge 1 - e^{-t} \quad (39)$$

when $n_{\text{sub}} \ge A_4(8d+2t)$. Combining (33) and (39) together, we get

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbf{B}_0(r_1)}\|\boldsymbol{\Sigma}^{-1/2}\{\nabla L_\tau(\boldsymbol{\beta}) - \nabla L_\tau(\boldsymbol{\beta}^*)\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2\\
&\le 2^{1+\delta}\sigma^2\tau^{-1-\delta}\upsilon_\delta + \sqrt{\frac{A_1\log d}{n_{\text{sub}}}}\\
&\quad + 4CA_1\sigma^4 r_1^2\tau^{-2} + 6A_4\sqrt{\frac{8d+2t}{n_{\text{sub}}}}r_1 \quad (40)
\end{aligned}$$

with confidence at least $1 - e^{-t}$. This together with (28) yields the final result. ■

Remark 2: Theorem 2 provides a nonasymptotic Bahadur representation [18] for HMS estimator $\boldsymbol{\beta}_\tau$ when the error terms have finite $(1+\delta)$th moments. It further implies that the approximation of $\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*$ has a subexponential tail. For the truncated random variable $\varphi_\tau(\varepsilon)$, we can see that

$$\begin{aligned}
|\mathbb{E}\varphi_\tau(\varepsilon)| &= -\mathbb{E}[(\varepsilon - \tau)\mathbf{1}(\varepsilon > \tau)] + \mathbb{E}[(-\varepsilon - \tau)\mathbf{1}(\varepsilon < -\tau)]\\
&\le \mathbb{E}[(|\varepsilon| - \tau)\mathbf{1}(|\varepsilon| > \tau)]\\
&\le \tau^{1-\delta}\mathbb{E}(|\varepsilon|^\delta). \quad (41)
\end{aligned}$$

This together with (27) shows that the HMS estimator $\boldsymbol{\beta}_\tau$ achieves nonasymptotic robustness against heavy-tailed noise. Specifically, by taking

$$t = \log(n_{\text{sub}}), \quad \tau \asymp \sqrt{\frac{1-\lambda}{1+\lambda}}\sqrt{\frac{n_{\text{sub}}}{d + \log(n_{\text{sub}})}}$$

TABLE I
STATISTICS OF REAL-WORLD DATASETS

Datasets	# Sample size	# Features
Appliances Energy Prediction	19735	29
Poker Hand	25010	11
Gas Turbine CO and NOx Emission	36733	11
Wave Energy Converters	288000	32
PPPTS	45730	9
Beijing Multi-Site Air-Quality	382168	14

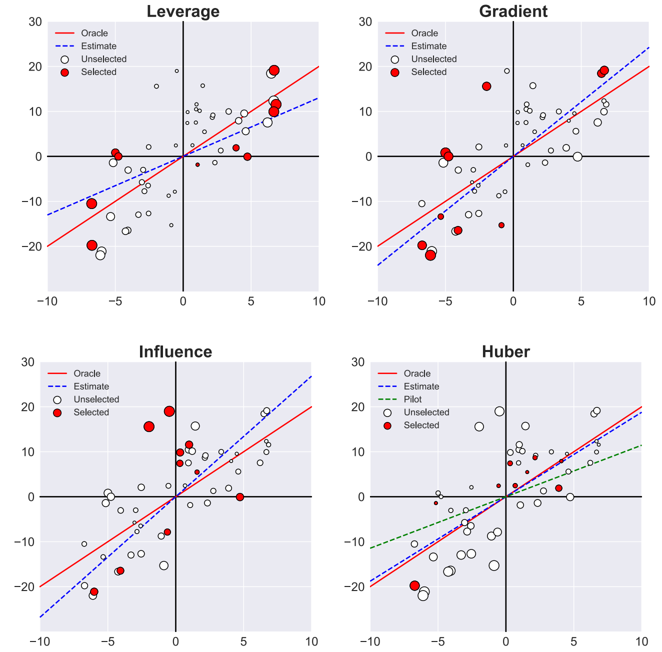

Fig. 3. Comparisons on different sampling patterns. The oracle, pilot, and subsampled estimator are denoted by the red real line, the green dashed line, and the blue dashed line, respectively.

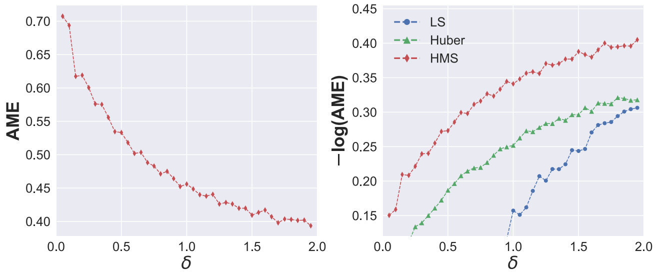

Fig. 4. Left: AME curve of HMS. Right: comparisons on $-\log(\text{AME})$ of different sampling procedures.

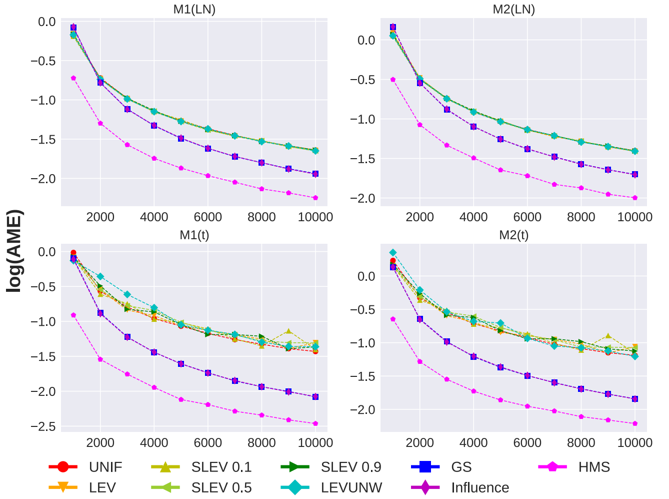

Fig. 5. Comparisons on AME of different sampling procedures. In all settings, we vary the subsample size $n_{\text{sub}} = sr * n$ with $sr = [0.002, 0.004, 0.006, 0.008, 0.01]$.

we have

$$\left\| \boldsymbol{\beta}_\tau - \boldsymbol{\beta}^* - \frac{1}{n_{\text{sub}}} \sum_{i=1}^{n_{\text{sub}}} \varphi_\tau(\varepsilon_i) \boldsymbol{\Sigma}^{-1} \mathbf{x}_i \right\|_2 = \mathcal{O}\left(\sqrt{\frac{1+\lambda}{1-\lambda}} \sqrt{\frac{d + \log(n_{\text{sub}})}{n_{\text{sub}}}} \right) \tag{42}$$

with confidence at least $1 - \mathcal{O}(n_{\text{sub}}^{-1})$. From an asymptotic viewpoint, it implies that if $d = o(n_{\text{sub}})$ as $n_{\text{sub}} \to \infty$, then for any deterministic vector $\mathbf{u} \in \mathbb{R}^d$, $\langle \mathbf{u}, \boldsymbol{\beta}_\tau - \boldsymbol{\beta}^* \rangle$ converges to $n_{\text{sub}}^{-1} \sum_{i=1}^{n_{\text{sub}}} \varphi_\tau(\varepsilon_i) \boldsymbol{\Sigma}^{-1} \mathbf{x}_i$ in distribution.

V. Experimental Results

This section aims to evaluate the empirical performance of the proposed HMS procedure. All numerical studies are implemented with Python 3.8 under Ubuntu 16.04 operation system with 2.2-GHz CPUs and 256-GB memory.

A. Sampling Pattern

We first investigate the performance of HMS by comparing the sampling pattern to leverage sampling, gradient-based sampling (GS), and influence-based sampling (IS). The toy data are generated by $y = 2\,x + \varepsilon$ with $n = 50$ and $d = 1$, where noise term comes from the Student's $\mathbf{t}$ distribution with two degrees of freedom, i.e., $\varepsilon \sim \mathbf{t}(2)$. Considering that both GS and IS require a pilot to determine the sampling probability, here, we fix the pilot (marked by a green dashed line) for a fair comparison. The pilot is specified by uniform sampling $n_0 = 10$ points. The turning parameter τ of HMS is set to 0.1. We plot $n_{\text{sub}} = 10$ data points (marked in red) selected by different sampling approaches, where the size denotes the corresponding assigned sampling probability. The estimators of four sampling approaches are then calculated based on the subsampled data. As shown in Fig. 3, we see that the selected data points of HMS are more close to the oracle (marked by red real line) than competitors, and hence the subsampled estimator (marked by blue dashed line) can better recover the ground-truth estimator. Moreover, it can be observed that HMS can return a reliable estimator even if the pilot is deviated from the oracle, which implies its great potential on selecting informative data from the noisy data.

B. Phase Transition

Theorem (1) implies that

$$-\log(\|\boldsymbol{\beta}_\tau - \boldsymbol{\beta}^*\|) \asymp \frac{\delta}{1+\delta}\log(n_{\text{sub}}) - \frac{\delta}{1+\delta}\log(A_\lambda v_\delta) \qquad 0 < \delta \le 1.$$

In order to validate the phase transition behavior of HMS estimator, we generate the data by (4) with $n = 10K$ and $d = 50$ and sample independent noise from $\mathbf{t}(df)$, which has finite $(1+\delta)$th moments provided $\delta < df - 1$ and infinite dfth moment. The oracle β^* is generated from the discrete uniform distribution $\{\pm 3, \pm 2, \pm 1, 0\}$. Following the setting in [12], we set $n_{\text{sub}} = 1000$ and $\delta = df - 1 - 0.05$. The turning parameter τ is specified by $\tau = \sigma(n_{\text{sub}}/t)^{1/2}$, where $\sigma^2 = (1/n)\sum_{i=1}^n (y_i - \bar{y})$ with $\bar{y} = (1/n)\sum_{i=1}^n y_i$. The quality of the fit is measured by the AME

$$\text{AME} = \frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{\beta}_{\tau k} - \boldsymbol{\beta}^*\|. \tag{43}$$

Fig. 4 shows the AME comparisons for HMS, least square with uniform sampling, and Huber regression with uniform sampling. One can observe that the AME of HMS estimator is decreasing with the increase of δ. In particular, HMS can achieve lower AME than Huber and LS with varying degrees of freedom. This further exhibits the significant advantages of HMS in robust regression.

C. Simulation Studies

We generate the data by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ [7], where the $n \times d$ design matrix $\mathbf{X}$ is constructed by a mixture of Gaussian $(1/2)N(\mu_1, \sigma_1^2) + (1/2)N(\mu_2, \sigma_2^2)$ in two different ways: $(M1)\mu_1 = -2, \sigma_1 = 3, \mu_2 = 2, \sigma_2 = 10$; $(M2)\mu_1 = 0, \sigma_1 = 3, \mu_2 = 0$, and $\sigma_2 = 10$. We generate two different types of i.i.d. noise, including log-normal distribution $\varepsilon_i \sim \text{Lognormal}(0, 1)$ and Student-t distribution $\varepsilon_i \sim \boldsymbol{t}(2)$, and both of them are heavy tailed and produce outliers with large variance. We denote the models combining these design matrices and noise distributions as follows: $M1(\boldsymbol{LN})$, $M1(\boldsymbol{t})$, $M2(\boldsymbol{LN})$, and $M2(\boldsymbol{t})$.

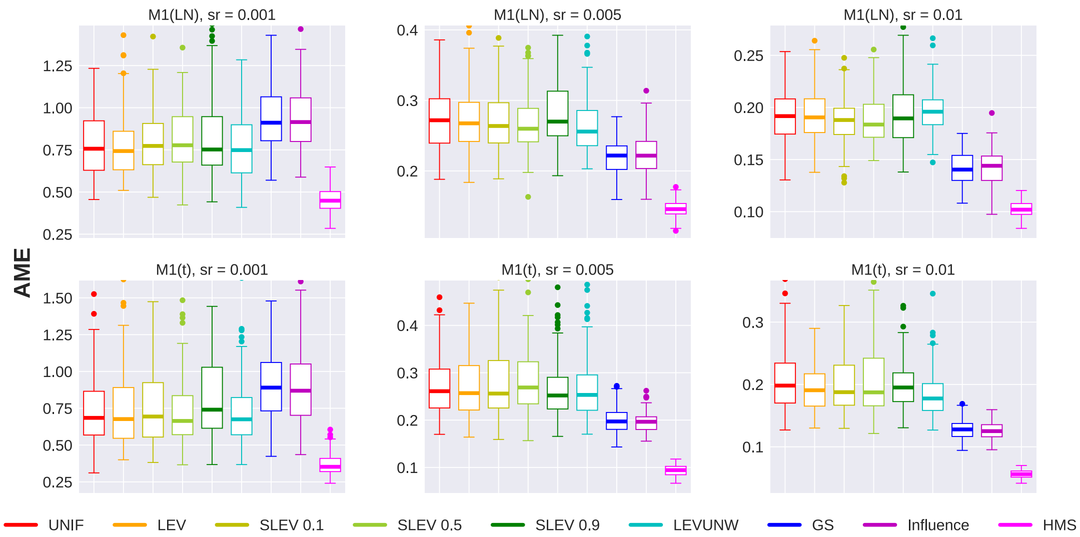

Fig. 6. Boxplots of AME for different subsampling methods ($n = 10K$ and $d = 50$).

TABLE II

COMPARISONS ON TIME COST (MILLISECONDS) FOR DIFFERENT SAMPLING METHODS. THE TIME COST OF HMS CONSISTS OF TWO PARTS: SELECTION OF τ (LEFT) + SAMPLING (RIGHT)

Methods	$n = 100K, d = 50$			$n = 500K, d = 250$			$n = 1M, d = 500$		
	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$
LEV	55.9	56.1	57.6	1515.1	1597.5	1624.7	12345.4	12068.4	12947.8
SLEV	54.1	56.8	62.9	1543.8	1610.9	1640.5	9932.8	9106.2	9545.5
LEVUNW	53.8	55.6	61.1	1550.0	1580.1	1625.4	8235.8	8984.8	8918.5
GS	33.7	34.0	37.6	818.2	804.2	879.2	3505.3	3473.6	3526.8
IS	64.8	63.5	79.1	1618.5	1645.9	1794.8	8852.1	9243.0	10051.1
HMS	5.9 + 23.2	6.6 + 34.5	7.5 + 44.4	141.1 + 229.2	167.1 + 302.5	189.0 + 389.2	557.9 + 780.8	662.4 + 910.0	754.9 + 1076.4

TABLE III

APE COMPARISONS (MEAN ± STANDARD DEVIATION) FOR DIFFERENT SAMPLING METHODS FOR REAL DATASETS

Methods	Appliances Energy Prediction			Poker Hand			Gas Turbine CO and NOx Emission		
	$sr = 0.2\%$	$sr = 0.5\%$	$sr = 1\%$	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$
UNIF	37.375(757.056)	17.544(150.132)	14.091(10.515)	20.628(39.454)	16.494(3.870)	16.145(1.733)	1.515(2.641)	1.212(0.268)	1.187(0.116)
LEV	977.049(1113.369)	30.047(832.387)	14.793(34.945)	21.687(86.551)	16.464(4.402)	16.121(1.558)	1.401(1.556)	1.198(0.147)	1.182(0.074)
SLEV1	43.437(1145.047)	32.101(786.459)	17.922(145.850)	20.199(33.124)	16.462(3.625)	16.143(1.602)	1.477(2.216)	1.204(0.170)	1.187(0.127)
SLEV5	716.760(1126.215)	39.196(1336.134)	17.582(199.610)	21.613(52.904)	16.501(3.955)	16.118(1.350)	1.428(2.594)	1.198(0.155)	1.181(0.072)
SLEV9	759.980(1476.744)	19.909(286.266)	18.714(242.235)	20.688(41.378)	16.435(2.933)	16.163(1.864)	1.386(1.148)	1.198(0.129)	1.182(0.095)
LEVUNW	23.865(118.949)	15.038(9.115)	13.849(3.739)	20.002(33.140)	16.448(3.846)	16.172(1.984)	**1.355**(0.935)	1.209(0.182)	1.192(0.083)
GS	959.747(1093.527)	20.077(239.837)	21.219(754.344)	22.568(66.602)	16.425(3.497)	16.057(1.138)	1.523(3.019)	**1.195**(0.145)	1.180(0.082)
IS	773.265(1675.704)	79.019(3602.502)	57.206(2274.425)	22.089(55.309)	16.391(2.900)	16.044(0.968)	1.513(2.485)	1.203(0.175)	1.180(0.069)
HMS	**21.906**(106.737)	**14.267**(15.968)	**13.549**(6.247)	**18.594**(15.623)	**16.217**(1.871)	**16.022**(1.044)	1.365(1.164)	1.197(0.161)	**1.178**(0.052)

TABLE IV

APE COMPARISONS (MEAN ± STANDARD DEVIATION) FOR DIFFERENT SAMPLING METHODS FOR REAL DATASETS

Methods	Wave Energy Converters			Physicochemical Properties of Protein Tertiary Structure			Beijing Multi-Site Air-Quality Data		
	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$	$sr = 0.1\%$	$sr = 0.5\%$	$sr = 1\%$
UNIF	55.611(9.120)	52.892(1.638)	52.594(0.680)	22.946(33.494)	19.074(8.027)	18.647(4.511)	27.930(9.177)	27.207(1.000)	27.125(0.330)
LEV	55.553(9.456)	52.875(1.724)	52.585(0.783)	20.328(11.959)	18.494(2.109)	18.298(0.891)	27.535(2.552)	27.162(0.455)	27.109(0.205)
SLEV1	55.305(8.762)	52.866(1.634)	52.585(0.665)	22.336(49.774)	18.509(3.508)	18.299(0.915)	27.698(4.926)	27.176(0.541)	27.118(0.350)
SLEV5	55.528(8.142)	52.853(1.439)	52.591(0.826)	20.207(13.829)	18.399(1.568)	18.269(0.842)	27.586(2.942)	27.156(0.467)	27.108(0.301)
SLEV9	55.425(9.317)	52.885(1.679)	52.579(0.795)	20.024(12.033)	18.446(1.916)	18.295(0.834)	27.523(2.037)	27.158(0.457)	27.106(0.208)
LEVUNW	55.365(6.857)	52.906(1.462)	52.625(0.786)	19.562(6.400)	18.700(1.997)	18.558(1.131)	27.832(6.767)	27.338(1.295)	27.259(0.736)
GS	55.905(10.413)	**52.753**(1.170)	**52.504**(0.528)	21.165(16.045)	18.463(1.907)	18.265(0.872)	27.346(1.187)	27.100(0.157)	**27.076**(0.066)
IS	55.738(9.943)	52.756(1.259)	52.517(0.581)	21.364(20.183)	18.474(1.561)	18.427(2.645)	27.343(1.236)	27.104(0.172)	27.078(0.082)
HMS	**54.982**(6.715)	52.789(1.323)	52.560(0.683)	**19.472**(15.802)	**18.299**(1.580)	**18.089**(0.871)	**27.185**(0.479)	**27.085**(0.075)	27.089(0.063)

We compare the proposed HMS with several representative methods, including uniform sampling (UNIF), leverage subsampling (LEV) [3], unweighted leverage subsampling (LEVUNW), shrinkage leverage subsampling (SLEV) [5], GS [7], and IS [8]. The sampling probability of SLEV is a convex combination of leverage and uniform distribution, i.e., $\pi_i^{\text{SLEV}} = \alpha\pi_i^{\text{LEV}} + (1 - \alpha)\pi_i^{\text{UNIF}}$. Here, we consider three different shrinkage factors $\alpha = 0.1, 0.5$, and 0.9 for SLEV, denoted by SLEV0.1, SLEV0.5, and SLEV0.9, respectively. LEVUNW performs the same sampling procedure as LEV but

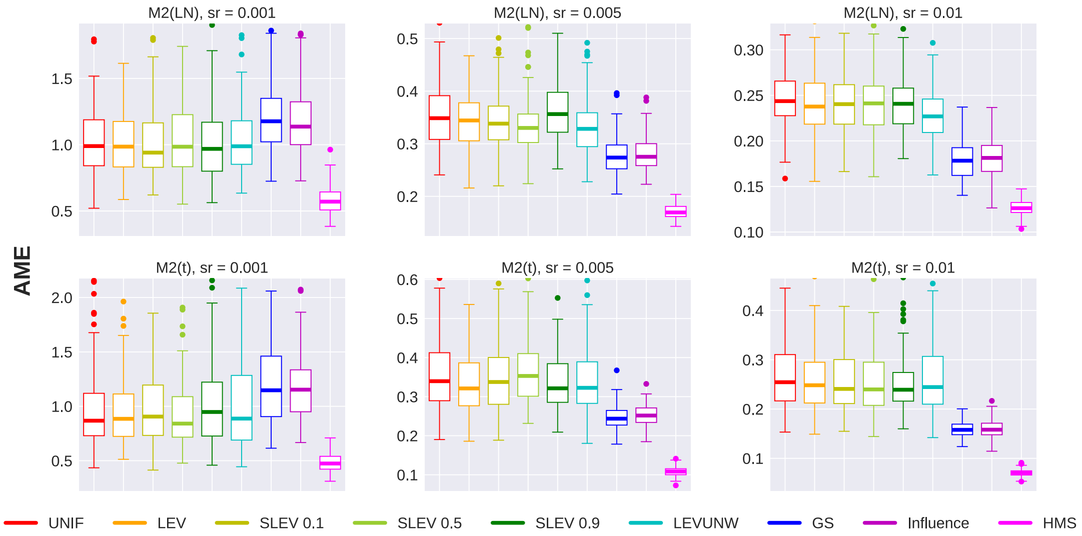

Fig. 7. Boxplots of AME for different subsampling methods ($n = 10K$ and $d = 50$).

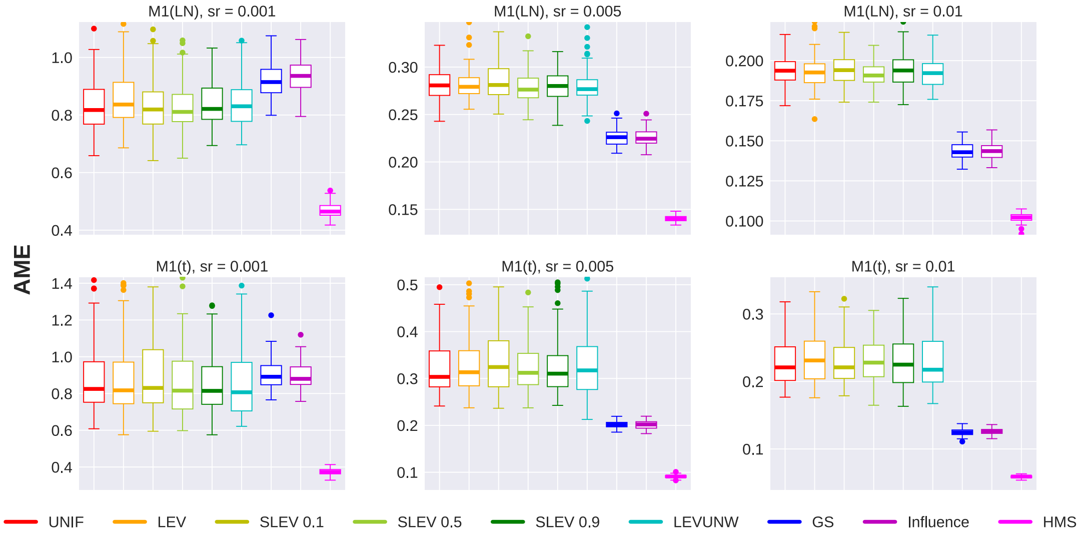

Fig. 8. Boxplots of AME for different subsampling methods ($n =$ 1M and $d = 500$).

solves the unweighted least-squares problem instead. For IS, the sampling weight for $(\mathbf{x}_i \cdot y_i)$ is proportional to $\|\psi_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i)\|$, where $\psi_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i) = (y_i - \mathbf{x}_i \boldsymbol{\beta}) \sum_n^{-1} \mathbf{x}_i$ is the influence function. For GS, IS, and HMS, the pilot is calculated by uniform sampling with size $n_0 = n_{\text{sub}}$, and the parameter τ in HMS is specified through a grid search strategy.

For each model, we set $n = 100K$ and $1M$ and corresponding $d = 50$ and 500. Denote sr by the sampling ratio, and we set subsample size by $n_{\text{sub}} = \text{sr}*n$ with $\text{sr} = 0.001, 0.005$, and 0.01. Each result is reported over $K = 100$ runs repeatedly and the mean error is calculated.

The AME comparisons for different sampling procedures are shown in Figs. 5–9, and the corresponding running time comparison is shown in Table II. Several observations can be made about the reported results.

1) Leverage-based sampling procedures perform slightly worse than uniform sampling when data are corrupted by heavy-tailed noises, and this is because leverage cannot exactly reflect the true importance of each sample in such cases.
2) GS and IS behave similarly in different settings. The reason is that the design matrix $\mathbf{X}$ consists of a mixture of i.i.d. Gaussian entries, leading to the covariance matrix Σ_n that approximates a diagonal matrix, which makes influence function assigns similar sampling probability as gradient does.
3) GS and IS perform worse than leverage-based approaches and uniform sampling when the sampling ratio is small. The main reason is that both of them need a pilot to guide sampling, and inefficient training for the pilot will deteriorate their performance. However, HMS performs significantly better than GS and IS with the same pilot. This demonstrates the tolerance of HMS to imperfect pilots.
4) HMS performs much better than the other competitors in almost all settings, both in AME and running time. The

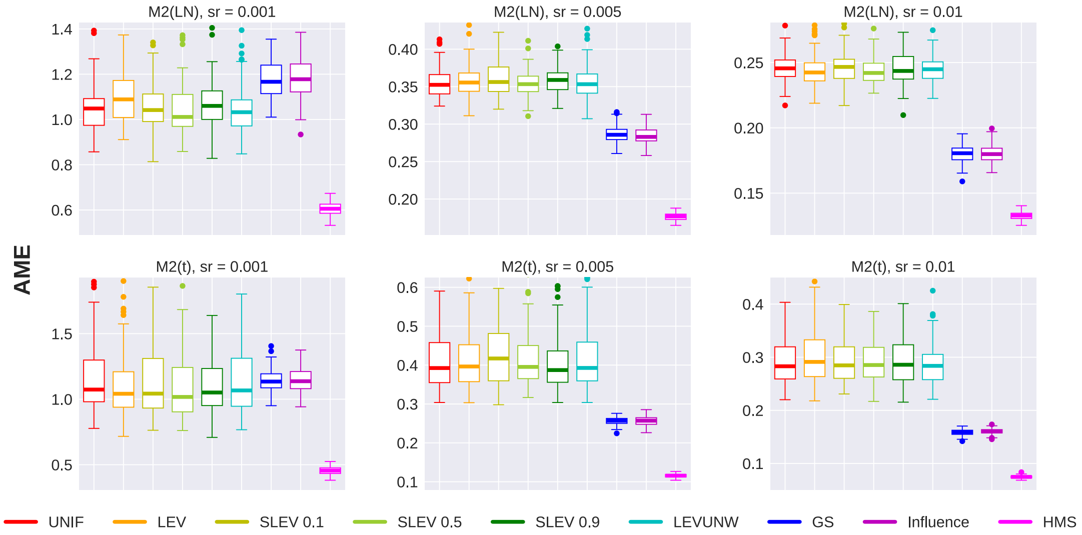

Fig. 9. Boxplots of AME for different subsampling methods ($n = 1\text{M}$ and $d = 500$).

efficiency improvement of HMS is still prominent even considering the time for hyperparameter (τ) selection, which implies the great advantage of HMS on selecting the informative samples under high-level noise settings.

D. Real Data Examples

We further evaluate the proposed HMS on six real-world datasets, including Appliances Energy Prediction, Poker Hand, Gas Turbine CO and NOx Emission , Wave Energy Converters, Physicochemical Properties of Protein Tertiary Structure (PPPTS), and Beijing Multi-Site Air-Quality. All these datasets come from the UCI machine learning repository https://archive.ics.uci.edu/ml/datasets.php, covering various prediction tasks. For Poker Hand dataset, we only use the training set. For Wave Energy Converters dataset, we remove 16 columns due to collinearity. For Beijing Multi-Site Air-Quality dataset, we remove four text-valued columns and take PM2.5 as the prediction target. The results are averaged over $K = 100$ runs of each experiment, and the average prediction errors (APEs)

$$\text{APE} = \frac{1}{K}\sum_{k=1}^{K}\|\hat{\mathbf{y}}_k - \mathbf{y}\|$$

are reported in Tables III and IV. It can be observed that HMS can achieve superior performance in these regression tasks. Specifically, HMS almost always reach the lowest error and standard deviation when the sampling ratio remains small, and this shows the great potential of applying HMS to deal with big data. For the Gas Emission, Wave Energy, and Air-Quality datasets, HMS sometimes yields suboptimal results compared to other methods. This is because in real-world scenarios, the properties of the noise are unknown, and some of the assumptions are not guaranteed to be hold, i.e., bounded covariates or bounded $1 + \delta$ order error moments. The convergence of HMS is thus influenced and results in suboptimal samples. However, HMS still achieves the highest performance in most conditions, which demonstrates its outstanding robustness over other methods.

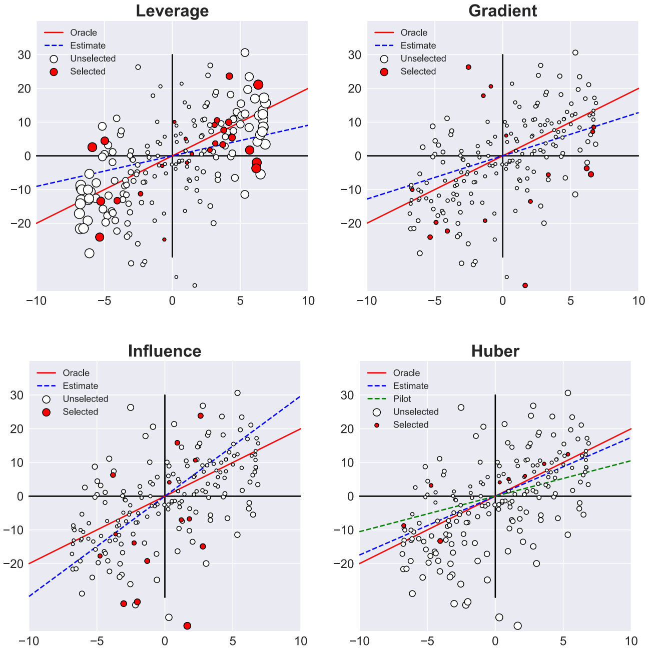

Fig. 10. Comparisons on different sampling patterns with $n = 200$ and $\varepsilon \sim \mathbf{t}(2)$.

VI. Discussion and Future Research

In this article, we propose a HMS to achieve robust estimation. The deviation bounds of HMS estimator are established. We find that the HMS estimator exhibits a similar phase transition to that in the independent setup. The only difference is up to a factor $((1 - \lambda)/(1 + \lambda))^{1/2}$, defined by the absolute spectral gap λ of underlying Markov chain. Extensive studies on large-scale simulations and real data examples demonstrate the effectiveness of HMS. There are many opportunities along

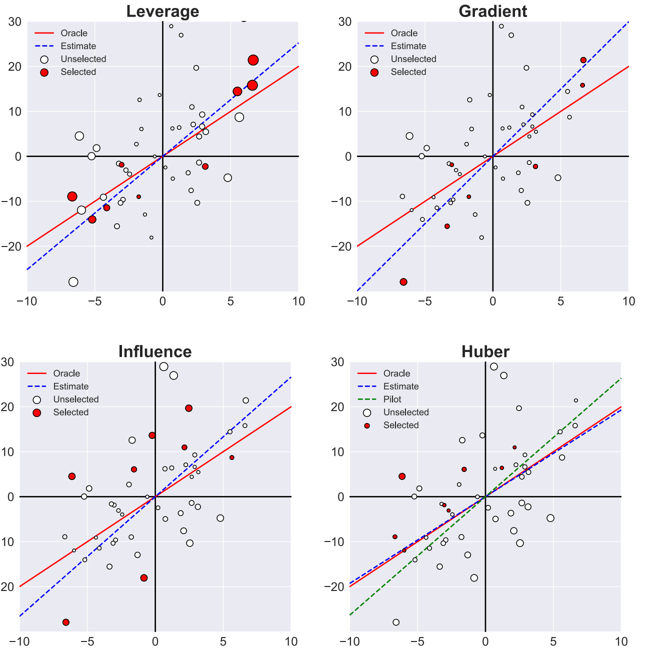

Fig. 11. Comparisons on different sampling patterns with $n = 50$ and $\varepsilon \sim \mathbf{Lognormal}(0, 1)$.

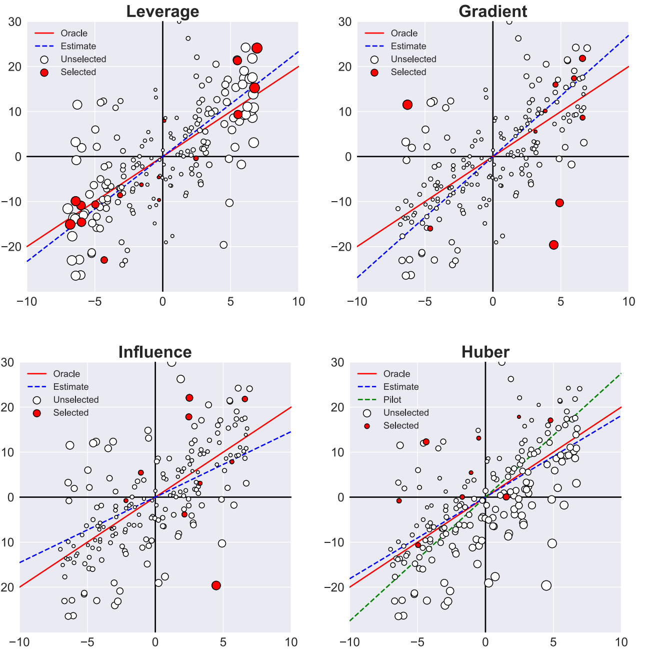

Fig. 12. Comparisons on different sampling patterns with $n = 200$ and $\varepsilon \sim \mathbf{Lognormal}(0, 1)$.

the line of current research, such as how to deduce the lower bounds for HMS estimator and how to perform HMS in high-dimensional cases. All these problems deserve further research.

Appendix

In this section, we add supplementary experiments on different data scales. In Figs. 10–12, we give additional experiment results of sampling patterns with different subsampling strategies. We keep the same settings as "Sampling Pattern" and set number of samples $n = \{50, 200\}$, distribution of noises $\varepsilon \sim \{\mathbf{t}(2), \mathbf{Lognormal}(0, 1)\}$. In all of these settings, we can derive the same conclusion that HMS achieves the lowest estimation error even when the pilot is deviated from the oracle.

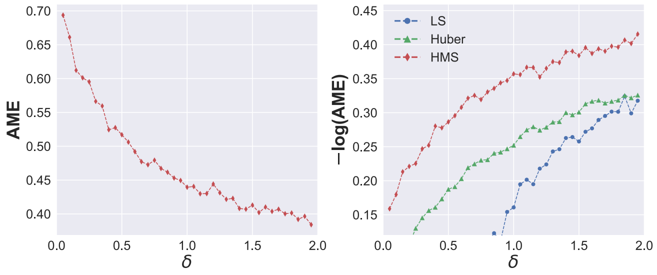

Fig. 13. Comparisons on AME of different sampling procedures with $n = 5000$, $d = 25$, and $n_{\mathrm{sub}} = 50$.

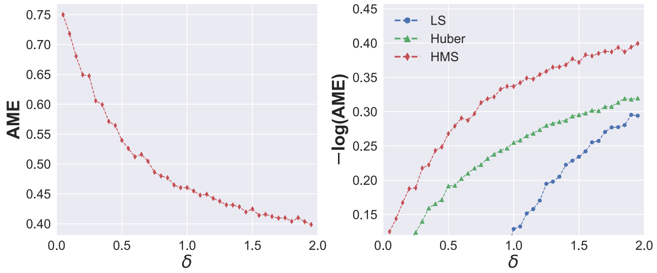

Fig. 14. Comparisons on AME of different sampling procedures $n = 20\,000$, $d = 100$, and $n_{\mathrm{sub}} = 200$.

In Figs. 13 and 14, we give additional experiment results of the phase transition behavior. Again, we keep the same parameter settings as "Phase Transition" and alter the simulation data size to $n = 5000, d = 25$, and $n_{\mathrm{sub}} = 50$, and $n = 20\,000, d = 100$, and $n_{\mathrm{sub}} = 200$. As can be seen, HMS still achieves lower AME than Huber and LS consistently under various data scales.

References

[1] Y. Zhang, J. Duchi, and M. Wainwright, "Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 3299–3340, Jan. 2015.

[2] P. S. Dhillon, Y. Lu, D. Foster, and L. Ungar, "New subsampling algorithms for fast least squares regression," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 360–368.

[3] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, "Fast approximation of matrix coherence and statistical leverage," *J. Mach. Learn. Res.*, vol. 13, pp. 3475–3506, Dec. 2012.

[4] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, "Faster least squares approximation," *Numer. Math.*, vol. 117, no. 2, pp. 219–249, Feb. 2011.

[5] P. Ma, M. W. Mahoney, and B. Yu, "A statistical perspective on algorithmic leveraging," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 861–911, 2015.

[6] A. Rudi, D. Calandriello, L. Carratino, and L. Rosasco, "On fast leverage score sampling and optimal learning," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5677–5687.

[7] R. Zhu, "Gradient-based sampling: An adaptive importance sampling for least-squares," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Princeton, NJ, USA: Citeseer, 2016, pp. 406–414.

[8] D. Ting and E. Brochu, "Optimal subsampling with influence functions," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3654–3663.

[9] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 4, pp. 73–101, 1964.

[10] S. Lambert-Lacroix and L. Zwald, "Robust regression through the Huber's criterion and adaptive lasso penalty," *Electron. J. Statist.*, vol. 5, pp. 1015–1053, 2011.

[11] L. Wang, C. Zheng, W. Zhou, and W.-X. Zhou, "A new principle for tuning-free Huber regression," *Stat. Sinica*, vol. 31, pp. 2153–2177, Jan. 2021.

[12] Q. Sun, W.-X. Zhou, and J. Fan, "Adaptive Huber regression," *J. Amer. Stat. Assoc.*, vol. 115, no. 529, pp. 254–265, Jan. 2020.

[13] J. Fan, Y. Guo, and B. Jiang, "Adaptive Huber regression on Markov-dependent data," *Stochastic Processes their Appl.*, vol. 150, pp. 802–818, Aug. 2022.

[14] B. Chen, W. Zhai, and Z. Huang, "Low-rank elastic-net regularized multivariate Huber regression model," *Appl. Math. Model.*, vol. 87, pp. 571–583, Nov. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0307904X20302389

[15] G. P. Meyer, "An alternative probabilistic interpretation of the Huber loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5261–5269.

[16] Y. Wang, X. Zhong, F. He, H. Chen, and D. Tao, "Huber additive models for non-stationary time series analysis," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–30. [Online]. Available: https://openreview.net/forum?id=9kpuB2bgnim

[17] R. R. Bahadur, "A note on quantiles in large samples," *Ann. Math. Statist.*, vol. 37, no. 3, pp. 577–580, Jun. 1966.

[18] X. He and Q.-M. Shao, "A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs," *Ann. Statist.*, vol. 24, no. 6, pp. 2608–2630, Dec. 1996.

[19] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[20] D. Rudolf, "Explicit error bounds for Markov chain Monte Carlo," 2011, *arXiv:1108.3201*.

[21] O. Catoni, "Challenging the empirical mean and empirical variance: A deviation study," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 48, no. 4, pp. 1148–1185, Nov. 2012.

[22] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Commun. Statist. Theory Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[23] J. Yu, H. Wang, M. Ai, and H. Zhang, "Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data," *J. Amer. Statist. Assoc.*, vol. 117, no. 537, pp. 265–276, 2022.

[24] H. Wang, "More efficient estimation for logistic regression with optimal subsamples," *J. Mach. Learn. Res.*, vol. 20, no. 132, pp. 1–59, 2019.

[25] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, vol. 589. Hoboken, NJ, USA: Wiley, 2005.

[26] P.-L. Loh, "Statistical consistency and asymptotic normality for high-dimensional robust M-estimators," *Ann. Statist.*, vol. 45, no. 2, pp. 866–896, 2017.

[27] T. Gong, B. Zou, and Z. Xu, "Learning with ℓ_1-regularizer based on Markov resampling," *IEEE Trans. Cybern.*, vol. 46, no. 5, pp. 1189–1201, May 2015.

[28] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. A. Simões, "Gradient sampling methods for nonsmooth optimization," in *Numerical Nonsmooth Optimization: State of the Art Algorithms*, A. M. Bagirov, M. Gaudioso, N. Karmitsa, M. M. Mäkelä, and S. Taheri, Eds. Cham, Switzerland: Springer, 2020, pp. 201–225, doi: 10.1007/978-3-030-34910-3_6.

[29] T. Sun, Y. Sun, and W. Yin, "On Markov chain gradient descent," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9918–9927.

[30] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. London, U.K.: Springer, 1993, doi: 10.1007%2F978-1-4471-3267-7.

[31] D. Down, S. P. Meyn, and R. L. Tweedie, "Exponential and uniform ergodicity of Markov processes," *Ann. Probab.*, vol. 23, no. 4, pp. 1671–1691, Oct. 1995.

[32] J. L. Morales and J. Nocedal, "Remark on 'algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization,'" *ACM Trans. Math. Softw.*, vol. 38, no. 1, pp. 1–4, Nov. 2011.

[33] G. Roberts and J. Rosenthal, "Geometric ergodicity and hybrid Markov chains," *Electron. Commun. Probab.*, vol. 2, no. none, pp. 13–25, Jan. 1997.

[34] J. Fan, B. Jiang, and Q. Sun, "Hoeffding's lemma for Markov chains and its applications to statistical learning," 2018, *arXiv:1802.00211*.

[35] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, vol. 24. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[36] B. Jiang, Q. Sun, and J. Fan, "Bernstein's inequality for general Markov chains," 2018, *arXiv:1801.00341*.

[37] J. Fan, H. Liu, Q. Sun, and T. Zhang, "I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error," *Ann. Statist.*, vol. 46, no. 2, p. 814, Apr. 2018.

[38] B. McWilliams, G. Krummenacher, M. Lucic, and J. M. Buhmann, "Fast and robust least squares estimation in corrupted linear models," in *Proc. 27th Int. Conf. Adv. Neural Process. Syst.*, 2014, pp. 415–423.

[39] V. Spokoiny, "Bernstein–von Mises theorem for growing parameter dimension," 2013, *arXiv:1302.3430*.

Tieliang Gong received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2018.

From September 2018 to October 2020, he was a Post-Doctoral Researcher with the Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada. He is currently an Assistant Professor with the School of Computer Science and Technology, Xi'an Jiaotong University. His current research interests include statistical learning theory, machine learning, and information theory.

Yuxin Dong received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology.

His research interests include information theory, statistical learning theory, and bioinformatics.

Hong Chen received the B.S., M.S., and Ph.D. degrees from Hubei University, Wuhan, China, in 2003, 2006, and 2009, respectively.

From February 2016 to August 2017, he was a Post-Doctoral Researcher with the Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX, USA. He is currently a Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan. His current research interests include machine learning, statistical learning theory, and approximation theory.

Bo Dong (Member, IEEE) received the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2014.

He did post-doctoral research at the MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, from 2014 to 2017. He is currently the Research Director of the School of Continuing Education, Xi'an Jiaotong University. His research interests focus on data mining and intelligent e-learning.

Chen Li received the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2014.

From June 2014 to March 2016, he was a Post-Doctoral Researcher with the Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Professor with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. His research interests include natural language processing, biological text mining, digital pathology, and bioinformatics.