Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(12): 28738–28752.
DOI:10.3934/math.20231471
Received: 03 August 2023
Revised: 16 September 2023
Accepted: 27 September 2023
Published: 23 October 2023

Research article

Concentration for multiplier empirical processes with dependent weights

Huiming Zhang[1,2] **and Hengzhen Huang**[3,*]

[1] Institute of Artificial Intelligence, Beihang University, Beijing 100083, China
[2] Zhuhai UM Science & Technology Research Institute, Zhuhai 519000, China
[3] College of Mathematics and Statistics, Guangxi Normal University, Guilin 541004, China

* **Correspondence:** Email: hzhuang@mailbox.gxnu.edu.cn; Tel: +8618290108925.

Abstract: A novel concentration inequality for the sum of independent sub-Gaussian variables with random dependent weights is introduced in statistical settings for high-dimensional data. The random dependent weights are functions of some regularized estimators. We applied the proposed concentration inequality to obtain a high probability bound for the stochastic Lipschitz constant for negative binomial loss functions involved in Lasso-penalized negative binomial regressions. We used this bound to study oracle inequalities for Lasso estimators. Additionally, a similar concentration inequality was derived for a randomly weighted sum of independent centred exponential family variables.

Keywords: high-dimensional data; multiplier empirical processes; sub-Gaussian distributions; statistical inference
Mathematics Subject Classification: 62K05, 62K15

1. Introduction

Over the last two decades, modern data collection techniques have enabled scientists and engineers to access and load vast numbers of variables as random data in their experiments. Probability theory provides the mathematical foundations for statistics and data-driven problems have led to various new advances in statistical research, which in turn contributes new and challenging problems in probability for further study. For instance, the rapid development of high-dimensional statistics has spurred the growth of the probability theory and even pure mathematics, including concentration inequalities, random matrix theory, geometric functional analysis and more [2, 16].

The emergence of high-throughput data has led to a surge in statistical research on complex data, particularly on high-dimensional data and statistical learning [14, 20]. This trend has gained traction in various scientific fields despite the high cost of measurements. Data sets are typically small, with only tens or hundreds of observations, and limited computing power often restricts the size of suitable finite samples. As a result, modern statisticians and data scientists have shifted their focus from asymptotic to non-asymptotic analysis, as it can handle small sample sizes and large model dimensions. Concentration inequalities play a crucial role in high-dimensional statistical inference, as they can derive various explicit non-asymptotic error bounds as a function of the sample size, sparsity level and dimension. When analyzing the various error bounds of the regularized estimator, concentration inequalities are indispensable tools for analysis [3, 20].

When the random variables are unbound, the classcial Hoeffding's inequality [8] falls to do a non-asymptotic analysis. We need the concept of sub-Gaussian random variables [9] to obtain tight Hoeffding-type concentration inequalities for the sum of independent random variables. A centered random variable (r.v.) X is called sub-Gaussian ($X \sim \operatorname{subG}(\sigma^2)$) if $\mathrm{E}e^{sX} \le e^{s^2\sigma^2/2}$ for $\forall\ s \in \mathbb{R}$, where $\sigma > 0$ is the sub-Gaussian parameter. From Chernoff's inequality, the exponential decay of the sub-Gaussian tail is obtained by $P(X \ge t) \le \inf_{s>0} \exp\{-st\}\mathrm{E}\exp\{sX\} \le \inf_{s>0}\exp(-st + \frac{\sigma^2 s^2}{2}) = \exp(-\frac{t^2}{2\sigma^2})$, minimizing the upper bound by putting $s = t/\sigma^2$. Moreover, for independent $\{X_i\}_{i=1}^n$ with $X_i \sim \operatorname{subG}(\sigma_i)$, we have sub-Gaussian concentration inequality

$$P\Big(|\sum\nolimits_{i=1}^n X_i| \ge t\Big) \le 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right\},\ t \ge 0, \tag{1.1}$$

for any variance proxies $\{\sigma_i^2\}_{i=1}^n$ of $\{X_i\}_{i=1}^n$ (see Theorem 1.5 in [1]). Define the L_p-norm of r.v. X as $\|X\|_p := (\mathrm{E}|X|^p)^{1/p}$. An alternative form of the sub-Gaussian parameter is defined by the sub-Gaussian norm $\|\cdot\|_{\theta_2}$ for zero-mean r.v. X is defined as $\|X\|_{\theta_2} := \sup_{p\ge1}[\frac{\mathrm{E}X^{2p}}{(2p-1)!!}]^{1/(2p)}$ (see page 23 in [1]).

Corollary 1.7 in [15] extended the sub-Gaussian concentration inequality (1.1) to the weighted sum of independent sub-Gaussian random variables with fixed weights.

Lemma 1. *[Concentration for weighted sub-Gaussian sum] Let $Y_1, \ldots, Y_n$ be n independent r.v.s with $Y_i \sim \operatorname{subG}(\sigma_i^2)$. Define $\sigma^2 = \max\limits_{1\le i\le n}\sigma_i^2 < \infty$. For any $\boldsymbol{w} := (w_1, \cdots, w_n)^\top$, we have*

$$P\Big(|\sum\nolimits_{i=1}^n w_i Y_i| > t\Big) \le 2\exp\left(-\frac{t^2}{2\sigma^2\|\boldsymbol{w}\|_2^2}\right).$$

However, if w_i's are random in Lemma 1, the story is totally different. The goal of this paper is to obtain novel theoretical results on the concentration inequality for the sum of dependent variables with random weights, under high-dimensional data background. Our theory is motivated from the non-asymptotic oracle inequalities of the regularized estimator in high-dimensional negative binomial regressions [21], and the concentration of random Lipschitz coefficients associated with empirical loss functions [4]. Our setting is different from classical *multiplier empirical processes* serving the multiplier Bootstrap inference, where the multipliers are random variables independent of $\{Y_i\}_{i=1}^n$ (see Chapter 2.9 of [17] and [6, 7]). Mendelson [11] studied the concentration inequalities for the centered multiplier process indexed by a functional class, where the i.i.d. multipliers need not be independent of the original empirical processes. In the analyses of high-dimension continuous data regressions by empirical processes, researches often resort to concentration inequalities of the Lipschitz function of strongly log-concave distributions (see Theorems 2.26 and 3.16 in [18]). For high-dimension count data regressions, our section 3.3 discusses the discrete distributions with strongly log-concave structures, which it is considered hard to check the definition of *discrete strongly log-concave distributions* (see (3.12) below). However, this strong assumption is usually intractable and unverifiable from the data. The sub-Gaussian assumption for the i.i.d. data is testable (see [23]).

In section two, we present the main results of the theory and demonstrate their applications in a class of high-dimensional generalized linear models. Theoretical proofs of the main results and some lemmas and additional results are given in section three. Finally, the conclusions are presented in section four.

2. Concentration for dependent summations

2.1. Main results

When controlling the summation of a function of the random sample indexed by a common estimator $\hat{\theta}$, it is false to use any sort of classical law of large numbers and central limit theorems (or concentration inequality for independent summation).

Formally, let $X_1, \ldots, X_n$ be a random sample independently drawn from P on a measurable space $(\mathcal{X}, \mathscr{A})$. Given an estimator $\hat{\theta}$, we want to study its asymptotic properties for summation of some function $f_{\hat{\theta}}(X_i)$,

$$\frac{1}{n}\sum_{i=1}^{n}[f_{\hat{\theta}}(X_i) - \mathrm{E}f_{\hat{\theta}}(X_i)].$$

A possible solution: Prove a uniform version (the suprema of empirical processes, see [17]) for all possible $\hat{\theta}$ on a set K, which is usually stronger than what is needed.

$$\frac{1}{n}\sum_{i=1}^{n}[f_{\hat{\theta}}(X_i) - \mathrm{E}f_{\hat{\theta}}(X_i)] \le \sup_{\theta\in K}\left|\frac{1}{n}\sum_{i=1}^{n}[f_{\theta}(X_i) - \mathrm{E}f_{\theta}(X_i)]\right|.$$

The summation in the sup enjoy independence.

In following theorem, we extend Lemma 1 with dependent and random weights.

Theorem 1. *[Concentration for weighted dependent sum] Let Y_i's be independent centered sub-Gaussian random variables with* $\max_{1\le i\le n}\|Y_k\|_{\theta_2} < \infty$. *Let $w_i(\hat{\theta})$'s be a series of bounded function of a bounded random variable $\hat{\theta}$ as the weights (can be dependent on all Y_i's), where $\|\hat{\theta}\|_1 \le r < \infty$ and* $\max_{1\le i\le n} w_i(\cdot) \le 1$. *Then, with probability of at least* $1-\delta$*,*

$$\left|\frac{1}{n}\sum_{i=1}^{n} w_i(\hat{\theta})Y_i\right| \le 4\sqrt{\frac{1}{n}\sum_{i=1}^{n}|||Y_i - Y_i'|||_{\theta_2}^2}\sqrt{\frac{\log\delta^{-1}}{n}} + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y_i^2)}\sqrt{\frac{2\log(2p)}{n}}, \tag{2.1}$$

for all $\|\hat{\theta}\|_1 < \infty$, where Y_i' is an independent copy of Y_i.

The first term in (2.1) is due to sub-Gaussian concentration, and the second term in (2.1) is from the upper bound of the expected version of the superma of empirical process $f(\boldsymbol{Y}) := \frac{1}{n}\sup_{\|\theta\|_1\le r}|\sum_{i=1}^{n} w_i(\theta)Y_i|$ (see the proof in section three).

2.2. Applications: Local stochastic Lipschitz conditions in GLMs

The concentration of random Lipschitz coefficients associated with empirical loss functions is crucial for deriving error bounds of Lasso or Elastic-net penalized high-dimensional generalized linear models (GLMs) in high-dimensional regressions. For more information, please refer to [3, 4, 10].

Definition 1. *[Elastic-net or Lasso penalized loss problems] Let $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ be independent identically random variables with values in $\mathbb{R} \times \mathbb{R}^p$, where $\{Y_i\}_{i=1}^n \sim Y$ are response variables and $\{\boldsymbol{X}_i\}_{i=1}^n \sim \boldsymbol{X}$ are covariates. Let $l(y, \boldsymbol{x}, \boldsymbol{\beta})$ be a loss function of parameter $\boldsymbol{\beta}$ and data $(y, \boldsymbol{x})$. The empirical loss function is defined as*

$$\mathbb{P}_n l(Y, \boldsymbol{X}, \boldsymbol{\beta}) := \tfrac{1}{n} \sum_{i=1}^n l(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}).$$

Elastic-net (or Lasso) estimators are given by

$$\hat{\boldsymbol{\beta}} =: \hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\mathbb{P}_n l(Y, \boldsymbol{X}, \boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\}, \tag{2.2}$$

where $\lambda_1 > 0$ and $\lambda_2 \geq 0$ are tuning parameters.

Define the minimizer

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathrm{E}[l(Y, \boldsymbol{X}, \boldsymbol{\beta})] \tag{2.3}$$

as the vector of true coefficients, where $l(Y, \boldsymbol{X}, \boldsymbol{\beta})$ is the loss function. The ℓ_1 ball is denoted as $\mathcal{S}_R(\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq R\}$. Theorem 1 can be used to establish exponential-type concentration inequalities for *the local stochastic Lipschitz (LSL) constant*:

$$\sup_{\boldsymbol{\beta} \in \mathcal{S}_R(\boldsymbol{\beta}^*)} \frac{(\mathbb{P}_n - \mathbb{P})\,[l(Y, \boldsymbol{X}, \boldsymbol{\beta}) - l(Y, \boldsymbol{X}, \boldsymbol{\beta}^*)]}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1}.$$

When study error bounds $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ of Lasso or Elastic-net penalized high-dimensional GLMs, one must bound the dependent empirical processes $\frac{(\mathbb{P}_n - \mathbb{P})[l(Y, \boldsymbol{X}, \hat{\boldsymbol{\beta}}) - l(Y, \boldsymbol{X}, \boldsymbol{\beta}^*)]}{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}$ with the LSL constant as the upper bound.

Next, we provide an example of negative binomial loss in negative binomial regressions [21]. The negative binomial loss function is $l(y, \boldsymbol{x}, \boldsymbol{\beta}) = y\boldsymbol{x}^\top \boldsymbol{\beta} - (\theta + y)\log(\theta + e^{\boldsymbol{x}^\top \boldsymbol{\beta}})$, where θ is called the dispersion parameter. Denote the expected risk function as $\mathbb{P}l(Y, \boldsymbol{X}, \boldsymbol{\beta}) := \mathrm{E}l(Y, \boldsymbol{X}, \boldsymbol{\beta})$. Let $l_1(y, \boldsymbol{x}, \boldsymbol{\beta}) := -y[\boldsymbol{x}^\top \boldsymbol{\beta} - \log(\theta + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\})]$, and $l_2(\boldsymbol{x}, \boldsymbol{\beta}) := \theta \log(\theta + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\})$, then

$$(\mathbb{P}_n - \mathbb{P})\,[l(y, \boldsymbol{x}, \boldsymbol{\beta}) - l(y, \boldsymbol{x}, \boldsymbol{\beta}^*)] = (\mathbb{P}_n - \mathbb{P})\,[l_1(y, \boldsymbol{x}, \boldsymbol{\beta}) - l_1(y, \boldsymbol{x}, \boldsymbol{\beta}^*)] + (\mathbb{P}_n - \mathbb{P})\,[l_2(\boldsymbol{x}, \boldsymbol{\beta}) - l_2(\boldsymbol{x}, \boldsymbol{\beta}^*)].$$

The upper bounds for the first and second parts of the empirical process: $(\mathbb{P}_n - \mathbb{P})(l_m(\boldsymbol{\beta}^*) - l_m(\hat{\boldsymbol{\beta}}))$ for $m = 1, 2$ is paramount to study the error bound of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$.

Let λ be a positive constant that needs to be determined. We have

$$P\left(\sup_{\boldsymbol{\beta} \in \mathcal{S}_R(\boldsymbol{\beta}^*)} \frac{|\,(\mathbb{P}_n - \mathbb{P})\,[l(Y, \boldsymbol{X}, \boldsymbol{\beta}) - l(Y, \boldsymbol{X}, \boldsymbol{\beta}^*)]|}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1} \leq \lambda\right)$$
$$\leq P\left(\sup_{\boldsymbol{\beta} \in \mathcal{S}_R(\boldsymbol{\beta}^*)} \frac{|\,(\mathbb{P}_n - \mathbb{P})\,[l_1(Y, \boldsymbol{X}, \boldsymbol{\beta}) - l_1(Y, \boldsymbol{X}, \boldsymbol{\beta}^*)]|}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1} \leq \frac{\lambda}{2}\right)$$

$$+P\left(\sup_{\beta\in\mathcal{S}_R(\boldsymbol{\beta}^*)}\frac{|(\mathbb{P}_n-\mathbb{P})\left[l_2(\boldsymbol{X},\boldsymbol{\beta})-l_2(\boldsymbol{X},\boldsymbol{\beta}^*)\right]|}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1}\leq\frac{\lambda}{2}\right). \tag{2.4}$$

Here, we assume that both $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are bound, and θ is a known dispersion parameter. The high probability for the second term in (2.4) is easy to deal with if we apply McDiarmid's inequality (see Lemma 4 of [21]). However, the high probability for the first term in (2.4) is hard to control since it contains unbounded negative binomial variables $\{Y_i\}_{i=1}^n$. Zhang and Jia [21] used the concentration inequality for strongly log-concave discrete distributions to solve this problem, but the strongly log-concave property is difficult to check for discrete distribution (see (H.4) in [21]). The sub-Gaussian distribution assumption is easy to verify for negative binomial variables, and this is from the fact that the negative binomial distribution belongs to the exponential family if the dispersion parameter is given. When Θ is compact in (2.7) below, Proposition 3.2 in [20] shows that $\{Y_i\}_{i=1}^n$ is sub-Gaussian.

From Taylor's expansion of continuous functions, one has $\log(\theta+e^x)-\log(\theta+e^a)=\frac{e^{\tilde{a}}}{\theta+e^{\tilde{a}}}(x-a)$, where $\tilde{a}$ is some real number between a and x. Let $\boldsymbol{X}_i^\top\tilde{\boldsymbol{\beta}}$ be some point between $\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^\top\boldsymbol{\beta}^*$, i.e., $\tilde{\boldsymbol{\beta}}=\begin{pmatrix}t_1\hat{\beta}_1\\ \vdots\\ t_p\hat{\beta}_p\end{pmatrix}+\begin{pmatrix}(1-t_1)\beta_1^*\\ \vdots\\ (1-t_p)\beta_p^*\end{pmatrix}$ for $\{t_j\}_{j=1}^p\subset[0,1]$. Observe that

$$\begin{aligned}(\mathbb{P}_n-\mathbb{P})[l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}})]&=\frac{-1}{n}\sum_{i=1}^n(Y_i-\mathrm{E}Y_i)\boldsymbol{X}_i^\top[(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}})-\log(\frac{\theta+\exp\{\mathbf{X}_i^\top\boldsymbol{\beta}^*\}}{\theta+\exp\{\mathbf{X}_i^\top\hat{\boldsymbol{\beta}}\}})]\\&=\frac{-1}{n}\sum_{i=1}^n(Y_i-\mathrm{E}Y_i)\boldsymbol{X}_i^\top[(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}})-\frac{\exp\{\boldsymbol{X}_i^\top\tilde{\boldsymbol{\beta}}\}\boldsymbol{X}_i^\top(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}})}{\theta+\exp\{\boldsymbol{X}_i^\top\tilde{\boldsymbol{\beta}}\}}]\\&=\frac{1}{n}\sum_{i=1}^n\frac{\theta\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)}{\theta+\exp\{\boldsymbol{X}_i^\top\tilde{\boldsymbol{\beta}}\}}\cdot(Y_i-\mathrm{E}Y_i).\end{aligned} \tag{2.5}$$

For a finite M_0, if we have $\hat{\boldsymbol{\beta}}\in\mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$, then $\tilde{\boldsymbol{\beta}}\in\mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$. This is from the fact $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|\leq\sum_{j=1}^p t_j|\hat{\beta}_j-\beta_1^*|\leq\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|\leq M_0$. Suppose $|\boldsymbol{X}_i|_\infty$ is uniformly bounded by $1/M_0$. We have $|(2.5)|:=\frac{1}{n}\sum\limits_{i=1}^n w_i(\hat{\boldsymbol{\beta}})(Y_i-\mathrm{E}Y_i)$ with dependent weights

$$w_i(\hat{\boldsymbol{\beta}}):=\frac{\theta X_i^\top(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)}{\theta+\exp\{X_i^\top\tilde{\boldsymbol{\beta}}\}}\text{ and }|w_i(\hat{\boldsymbol{\beta}})|\leq 1.$$

Thus, the high probability upper bound in Theorem 2.1 is applicable to determine $\frac{\lambda}{2}$, i.e.,

$$\frac{\lambda}{2}=4\sqrt{\frac{1}{n}\sum_{k=1}^n|||Y_k-Y_k'|||_{\theta_2}^2}\sqrt{\frac{\log\delta^{-1}}{n}}+2\sqrt{\frac{1}{n}\sum_{i=1}^n\mathrm{E}(Y_i-\mathrm{E}Y_i)^2}\sqrt{\frac{2\log 2p}{n}}. \tag{2.6}$$

2.3. Berstein-type concentration inequalities

In this section, let $\{Y_i\}_{i=1}^n$ be exponential family random variables with density

$$f(y_i;\theta_i)=c(y_i)\exp\{y_i\eta_i-b(\eta_i)\},\ \eta_i\in\Theta. \tag{2.7}$$

Here, $\mathrm{E}(Y_i) = \dot{b}(\eta_i)$ and $\mathrm{Var}(Y_i) = \ddot{b}(\eta_i)$. It should be noted that Proposition 3.2 in [20] shows that $\{Y_i\}_{i=1}^n$ is sub-Gaussian if Θ is compact.

Under distributional assumption (2.7), we will study the Berstein-type concentration inequalities for the randomly weighted sum of centered exponential family random variables (with different parameters θ_i's):

$$\sum_{i=1}^n \{w_i(\hat{\theta})Y_i - \mathrm{E}[w_i(\hat{\theta})Y_i]\},$$

where the $\{w_i(\hat{\theta})\}_{i=1}^n$ are called the multipliers (or random weights), and $\hat{\theta}$ is independent of $\{Y_i\}_{i=1}^n$.

Theorem 2. *[Concentration inequalities for randomly weighted sum of exponential family r.v.s] If* $\{Y_i\}_{i=1}^n$ *has density* (2.7) *with moment conditions:* $\mathrm{E}|Y_i|^k \le k!C_Y^k$, *where* $C_Y > 0$ *is a constant. We assume that*

(i) Bounded weights: $\boldsymbol{W}_{\hat{\theta}} := (w_1(\hat{\theta}), \cdots, w_n(\hat{\theta}))^\top$ is a random vector s.t. $\max_{1\le i\le n} |w_i(\hat{\theta})| \le w < \infty$;

(ii) Let $\mathrm{E}[|Y_i|^k|\boldsymbol{W}_{\hat{\theta}}] = \rho_{i,k}\mathrm{E}[|Y_i|^k]$ with $\mathrm{E}\rho_{i,k} = 1$;

(iii) There exists a non-decreasing sequence $\{u_n\}$ and constant C_ρ s.t. $P(\max\limits_{k\ge1,1\le i\le n} \rho_{i,k} > u_n) \le C_\rho/u_n$. Then,

$$P(|\sum_{i=1}^n [w_i(\hat{\theta})Y_i - \mathrm{E}[w_i(\hat{\theta})Y_i)]| \ge t) \le 2\exp\left\{-\frac{t^2}{16nu_n(wC_T)^2 + 4wC_Tt}\right\} + \frac{C_\rho}{u_n}, \tag{2.8}$$

where $\{w_i(\hat{\theta})Y_i\}$ is dependent since each $w_i(\hat{\theta})Y_i$ depends on a common estimator $\hat{\theta}$ from the data $\{Y_i\}_{i=1}^n$.

It should be noted that our work here is related to Proposition 3.2 in [20] and it is about the concentration inequalities for the non-random weighted sum of exponential family random variables. For condition (ii), suppose that the estimator $\hat{\theta}$ converges to a true parameter θ^* almost surely, and one has $\mathrm{E}[|Y_i|^k|W_{\theta^*}] = \mathrm{E}[|Y_i|^k]$ since W_{θ^*} is non-random. The difference between conditional expectation and unconditional expectation is:

$$\max_{k\ge1,1\le i\le n} \left|\rho_{i,k} - 1\right| = \max_{k\ge1,1\le i\le n} \frac{|\mathrm{E}[|Y_i|^k|\boldsymbol{W}_{\hat{\theta}}] - \mathrm{E}[|Y_i|^k|\boldsymbol{W}_{\theta^*}]|}{\mathrm{E}[|Y_i|^k|\boldsymbol{W}_{\theta^*}]} \le O_p(\|\hat{\theta} - \theta^*\|_1),$$

if $|\mathrm{E}[|Y_i|^k|\boldsymbol{W}_{\boldsymbol{\beta}}]$ is a ℓ_1-Lipschitz function of $\boldsymbol{\beta}$. We call $P(\max\limits_{k\ge1,1\le i\le n} \rho_{i,k} \ge u_n) \le C_\rho/u_n$ in assumption (iii) a high level condition. Intuitionally, due to the dependence summation, the random weighted summation will lose the rate of convergence in the exponential inequalities (addition term $4wC_Tt$ is added), compared to the case of non-random weighted summation. The assumption of compact parameter space for the exponential family is key to obtaining the sub-Gaussian type concentration inequalities.

Our multiplier concentration inequality here is different from [11], which studies the concentration upper bounds for centered multiplier empirical processes $\frac{1}{\sqrt{n}}\sum\limits_{i=1}^n [W_iY_i - \mathrm{E}(W_iY_i)]$ (the random weights $\{W_i\}$ and random variables $\{Y_i\}$ need not be independent); however, they assume that $\{W_i\}$ is i.i.d. To the best of our knowledge, Theorem 2 is a new concentration inequality that is suitable for the weighted sum of dependent random variables.

3. Theoretical proofs

3.1. Proofs of main results

Proof of Theorem 1: Let $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^\top$ be a vector of independent r.v.s in a space $\mathcal{Y}$, and define $(Y_1', \cdots, Y_n')^\top$ as an independent copy of $(Y_1, \cdots, Y_n)^\top$. For any function $f : \mathcal{Y}^n \to \mathbb{R}$, it is of interest to study the concentration for $f(\boldsymbol{Y})$ about its expectation. In case of Theorem 1,

$$f(\boldsymbol{Y}) := \tfrac{1}{n} \sup_{\|\theta\|_1 \le r} |\sum_{i=1}^{n} w_i(\theta) Y_i|.$$

For $z \in \mathcal{Y}$ and $k \in \{1, \ldots, n\}$, define the substitution operator

$$S_z^k : \mathcal{Y}^n \to \mathcal{Y}^n \text{ by } S_z^k y := (y_1, \ldots, y_{k-1}, z, y_{k+1}, \ldots, y_n)$$

and the centered conditional version of f

$$\begin{aligned} D_{f,Y_k}(y) &:= f(y_1, \ldots, y_{k-1}, Y_k, y_{k+1}, \ldots, y_n) - \mathrm{E} f(y_1, \ldots, y_{k-1}, Y_k', y_{k+1}, \ldots, y_n) \\ &= f(S_{Y_k}^k y) - \mathrm{E} f(S_{Y_k'}^k y) = \mathrm{E}[f(S_{Y_k}^k y) - f(S_{Y_k'}^k y) \mid Y_k]. \end{aligned} \tag{3.1}$$

Next, we use a constant-sharper sub-Gaussian concentration for $f(Z)$ in Corollary 4 in [22], which requires the $\|\cdot\|_{\theta_2}$-norm condition of r.v. $\{D_{f,Z_i}(z)\}_{i=1}^n$.

Lemma 2. *If $\{D_{f,Z_i}(z)\}_{i=1}^n$ has finite $\|\cdot\|_{\theta_2}$-norm for $z \in \mathcal{Z}$, then $f(Z) - \mathrm{E} f(Z) \sim$ $\operatorname{subG}(8 \sup_{z\in\mathcal{Z}} \sum_{i=1}^n \left\|D_{f,Z_i}(z)\right\|_{\theta_2}^2)$ and*

$$P\{f(Z) - \mathrm{E} f(Z) > t\} \le e^{-t^2/(16 \sup_{z\in\mathcal{Z}} \sum_{i=1}^n \|D_{f,Z_i}(z)\|_{\theta_2}^2)},\ t \ge 0.$$

From the identity in (3.1), we have

$$\begin{aligned} \|D_{f,Y_k}(y)\|_{\theta_2} &= \|f(y_1, \ldots, y_{k-1}, Y_k, y_{k+1}, \ldots, y_n) - \mathrm{E} f(y_1, \ldots, y_{k-1}, Y_k', y_{k+1}, \ldots, y_n)\|_{\theta_2} \\ &= \|\frac{1}{n} \sup_{\|\theta\|_1 \le r} |w_1(\theta) y_1 + \cdots + w_{k-1}(\theta) y_{k-1} + w_k(\theta) Y_k + w_{k+1}(\theta) y_{k+1} + \cdots + w_n(\theta) y_n| \\ &\quad - \frac{1}{n} \sup_{\|\theta\|_1 \le r} |w_1(\theta) y_1 + \cdots + w_{k-1}(\theta) y_{k-1} + w_k(\theta) Y_k' + w_{k+1}(\theta) y_{k+1} + \cdots + w_n(\theta) y_n| \mid Y_k]\|_{\theta_2} \\ &\le \frac{1}{n} \|\mathrm{E}[\sup_{\|\theta\|_1 \le r} w_k(\theta) |Y_k - Y_k'| |Y_k]\|_{\theta_2} \le \frac{1}{n} \|\mathrm{E}[|Y_k - Y_k'| |Y_k]\|_{\theta_2}. \end{aligned} \tag{3.2}$$

The conditional Jensen's inequality gives

$$\begin{aligned} \mathrm{E}\left[\left|\mathrm{E}\left[|Y_k - Y_k'| \mid X_k\right]\right|^p\right] &\le \mathrm{E}\left[\mathrm{E}\{|Y_k - Y_k'| \mid X_k\}^p\right] = \mathrm{E}\left[\mathrm{E}\{(|Y_k - Y_k'|^p)^{1/p} \mid X_k\}^p\right] \\ &\le \mathrm{E}\left\{\mathrm{E}\left[|Y_k - Y_k'|^p \mid X_k\right]\right\} = \mathrm{E}|Y_k - Y_k'|^p,\ p \ge 1. \end{aligned} \tag{3.3}$$

The definition $\|X\|_{\theta_2} = \sup_{k\ge1} [\frac{2^k k!}{(2k)!} \mathrm{E} X^{2k}]^{1/(2k)}$ shows $\|D_{f,Y_k}(y)\|_{\theta_2} \le \frac{1}{n} \||Y_k - Y_k'|\|_{\theta_2}$ by (3.2) and (3.3). Hence, we have $\sup_{z\in\mathcal{Z}} \sum_{i=1}^n \|D_{f,Z_i}(z)\|_{\theta_2}^2 = \frac{1}{n^2} \sum_{k=1}^n \||Y_k - Y_k'|\|_{\theta_2}^2$ in Lemma 2, which leads to

$$\mathrm{P}\{f(\boldsymbol{Y}) - \mathrm{E} f(\boldsymbol{Y}) > t\} \le e^{-(nt)^2/16 \sum_{k=1}^n \||Y_k - Y_k'|\|_{\theta_2}^2},\ t \ge 0.$$

Let $\delta = e^{-(nt)^2/16\sum_{k=1}^n |||Y_k - Y_k'|||_{\theta_2}^2}$, and $t = 4\sqrt{\frac{1}{n}\sum_{k=1}^n |||Y_k - Y_k'|||_{\theta_2}^2}\sqrt{\frac{\log\delta^{-1}}{n}}$. We have

$$f(\boldsymbol{Y}) \le t + \mathrm{E}f(\boldsymbol{Y}) = 4\sqrt{\frac{1}{n}\sum_{i=1}^n |||Y_i - Y_i'|||_{\theta_2}^2}\sqrt{\frac{\log\delta^{-1}}{n}} + \mathrm{E}f(\boldsymbol{Y}), \tag{3.4}$$

with probability at least $1-\delta$.

It remains to obtain a bounds on $\mathrm{E}f(\boldsymbol{Y})$, which is upper bounded by the symmetrization theorem from Lemma 3 with different functions. To see this, let $X_i = Y_i$ in Lemma 3 and $g_i(Y_i) = w_i(\theta)Y_i$ for $i = 1,\cdots,n$.

Since $w_i(\theta)$'s are series of bounded functions of a common bounded variable θ where $\|\theta\|_1 \le r$ and $\max_{1\le i\le n} w_i(\cdot) \le 1$, for any vector θ with $\|\theta\|_1 \le r$, there exists a sequence of vectors $\{a_{w_i}\}_{i=1}^n \in \mathbb{R}^p$ with $\left\|a_{w_i}\right\|_\infty \le 1/r$ such that

$$w_i(\theta) = a_{w_i}^\top\theta \le \left\|a_{w_i}\right\|_\infty \|\theta\|_1 \le 1. \tag{3.5}$$

Equation (3.5) and Lemma 3 imply

$$\begin{aligned}
\mathrm{E}f(\boldsymbol{Y}) &\le \frac{2}{n}\mathrm{E}\left(\sup_{\|\theta\|_1\le r} |\sum_{i=1}^n w_i(\theta)\epsilon_i Y_i|\right) = \frac{2}{n}\mathrm{E}\left(\sup_{\|\theta\|_1\le r} |\sum_{i=1}^n\sum_{j=1}^p \epsilon_i Y_i a_{w_ij}\theta_j|\right)\\
&= \frac{2}{n}\mathrm{E}\left(\sup_{\|\theta\|_1\le r} |\sum_{j=1}^p(\sum_{i=1}^n \epsilon_i Y_i a_{w_ij})\theta_j|\right)\\
\text{[by Hölder's inequality]}\quad &\le \frac{2}{n}\mathrm{E}\left(\sup_{\|\theta\|_1\le r}\max_{1\le j\le p}\left|\sum_{i=1}^n \epsilon_i Y_i a_{w_ij}\right| \cdot \|\theta\|_1\right)\\
&\le \frac{2r}{n}\mathrm{E}\left(\max_{1\le j\le p}\left|\sum_{i=1}^n \epsilon_i Y_i a_{w_ij}\right|\right) = \frac{2r}{n}\mathrm{E}\left(\mathrm{E}_\epsilon\max_{1\le j\le p}\left|\sum_{i=1}^n \epsilon_i Y_i a_{w_ij}\right|\right).
\end{aligned}$$

Next, we apply the maximal inequality. By Corollary 7.5 in [20], with $\mathrm{E}[\epsilon_i Y_i a_{w_ij}|\boldsymbol{Y}] = 0$ and $\epsilon_i Y_i a_{w_ij} \le \max_{1\le i\le n}\left\|a_{w_i}\right\|_\infty Y_i = r^{-1}Y_i$, one has

$$\begin{aligned}
\frac{2r}{n}\mathrm{E}\left(\mathrm{E}_\epsilon\max_{1\le j\le p}\left|\sum_{i=1}^n \epsilon_i Y_i a_{w_ij}\right|\right) &\le \frac{2}{n}\sqrt{2\log(2p)}\mathrm{E}\left(\sqrt{\sum_{i=1}^n Y_i^2|\boldsymbol{Y}}\right)\\
\text{[By Jensen's inequality]}\quad &\le \frac{2}{n}\sqrt{2\log(2p)}\sqrt{\mathrm{E}\sum_{i=1}^n Y_i^2} = 2\sqrt{\frac{1}{n}\sum_{i=1}^n \mathrm{E}Y_i^2}\sqrt{\frac{2\log(2p)}{n}}.
\end{aligned}$$

Thus, $\mathrm{E}f(\boldsymbol{Y}) \le 2\sqrt{\frac{1}{n}\sum_{i=1}^n \mathrm{E}Y_i^2}\sqrt{\frac{2\log(2p)}{n}}$. Using (3.4),

$$f(\boldsymbol{Y}) \le t + \mathrm{E}f(\boldsymbol{Y}) = 4\sqrt{\frac{1}{n}\sum_{i=1}^n |||Y_i - Y_i'|||_{\theta_2}^2}\sqrt{\frac{\log\delta^{-1}}{n}} + 2\sqrt{\frac{1}{n}\sum_{i=1}^n \mathrm{E}Y_i^2}\sqrt{\frac{2\log(2p)}{n}}, \tag{3.6}$$

with the probability of at least $1-\delta$.

Proof of Theorem 2: We will adopt the following result, which gives the moments inequality for the exponential family. It is deduced by the analytic properties of the absolute moments of exponential family random variables:

$$\mathrm{E}|Y|^k \le k!C_Y^k,$$

see Proposition 5.2 in [20]. For notation simplicity, let $\boldsymbol{W} := \boldsymbol{W}_{\hat{\theta}}$ and $W_i = w_i(\hat{\theta})$. By using Taylor's expansion and the binomial coefficient formula, we have the following upper bound for the conditional moment generating function of $W_iY_i - \mathrm{E}(W_iY_i)$, conditioning on the event $\{\max_{k\ge1, 1\le i\le n} \rho_{i,k} \le u_n\}$:

$$\begin{aligned}
\mathrm{E}[e^{s(W_iY_i - \mathrm{E}(W_iY_i))}|\boldsymbol{W}] &= 1 + \sum_{m=2}^{\infty} \frac{s^m}{m!}\mathrm{E}[(W_iY_i - \mathrm{E}(W_iY_i))^m|\boldsymbol{W}] \\
&= 1 + \sum_{m=2}^{\infty} \frac{s^m}{m!}\mathrm{E}[\sum_{k=0}^{m} \begin{pmatrix} k \\ m \end{pmatrix} (W_iY_i)^k(-\mathrm{E}(W_iY_i))^{m-k}|\boldsymbol{W}] \\
&\le 1 + \sum_{m=2}^{\infty} \frac{s^m}{m!}[\sum_{k=0}^{m} \begin{pmatrix} k \\ m \end{pmatrix} \mathrm{E}|W_iY_i|^k(\mathrm{E}|W_iY_i|)^{m-k}|\boldsymbol{W}] \\
\text{(Due to } \max_{1\le i\le n}|W_i| \le w) \quad &\le 1 + \sum_{m=2}^{\infty} \frac{s^m}{m!}[w^m \sum_{k=0}^{m} \begin{pmatrix} k \\ m \end{pmatrix} \mathrm{E}(|Y_i|^k|\boldsymbol{W})(\mathrm{E}|Y_i|)^{m-k}] \\
&\le 1 + \sum_{m=2}^{\infty} \frac{s^m}{m!}[(2w)^m \max_{1\le k\le m}\{\mathrm{E}(|Y_i|^k|\boldsymbol{W})(\mathrm{E}|Y_i|)^{m-k}\} \\
\text{(By assumption (iii))} &\le 1 + u_n \sum_{m=2}^{\infty} \frac{(2ws)^m}{m!}[\max_{1\le k\le m}\{\mathrm{E}(|Y_i|^k)(\mathrm{E}|Y_i|)^{m-k}\}, \qquad (3.7)
\end{aligned}$$

for $s \in (0, \delta)$ with some $\delta > 0$.

Therefore, we can assume that $|2swC_T| < 1$, so

$$\begin{aligned}
\mathrm{E}[e^{s(W_iY_i - \mathrm{E}(W_iY_i))}|W] &\le 1 + u_n \sum_{m=2}^{\infty} \frac{(2ws)^m}{m!}[m!C_T^m] = 1 + u_n(2swC_T)^2 \sum_{m=2}^{\infty}(2swC_T)^{m-2} \\
&= 1 + \frac{u_n(2swC_T)^2}{1 - 2swC_T} \le e^{\frac{u_n(2swC_T)^2}{1-2swC_T}}. \qquad (3.8)
\end{aligned}$$

Define the randomly weighted sum $S_n^W =: \sum_{i=1}^n W_iY_i$. By the conditional independence of $\{W_iY_i|\boldsymbol{W}\}_{i=1}^n$, it follows that by (3.8),

$$\mathrm{E}[e^{s(S_n^W - \mathrm{E}S_n^W)}|\boldsymbol{W}] = \prod_{i=1}^{n} \mathrm{E}[\exp\{s[W_iY_i - \mathrm{E}(W_iY_i)]\}|\boldsymbol{W}] \le e^{\frac{nu_n(2swC_T)^2}{1-2swC_T}}. \qquad (3.9)$$

By conditional Markov's inequality and on $\{\max_{k\ge1, 1\le i\le n} \rho_{i,k} < u_n\}$, we have for $a > 0$

$$\begin{aligned}
P(|S_n^W - \mathrm{E}S_n^W| \ge t|\boldsymbol{W}) &\le P(a(S_n^W - \mathrm{E}S_n^W) \ge at|\boldsymbol{W}) + P(a(-S_n^W + \mathrm{E}S_n^W) \ge at|\boldsymbol{W}) \\
&\le \frac{\mathrm{E}[e^{a(S_n^W - \mathrm{E}S_n^W)}|\boldsymbol{W}]}{\exp(at)} + \frac{\mathrm{E}[e^{a(-S_n^W + \mathrm{E}S_n^W)}|\boldsymbol{W}]}{\exp(at)} \\
\text{[Using (3.9) as } a \in (-\delta, \delta)] \quad &\le 2\exp\{\frac{nu_n(2awC_T)^2}{1 - 2awC_T} - at\} = 2\exp\left\{-\frac{t^2}{16nu_n(wC_T)^2 + 4wC_Tt}\right\} \qquad (3.10)
\end{aligned}$$

where the last equality is obtained by setting $a = \frac{t}{8nu_n(wC_T)^2+2wC_T t}$.

Taking expectation w.r.t. $\boldsymbol{W}$ on (3.10), it implies

$$
\begin{aligned}
P(|S_n^W - \mathrm{E}S_n^W| \geq t) &= P\left(|S_n^W - \mathrm{E}S_n^W| \geq t, \max_{k\geq 1, 1\leq i\leq n} \rho_{i,k} > u_n\right) + P\left(|S_n^W - \mathrm{E}S_n^W| \geq t, \max_{k\geq 1, 1\leq i\leq n} \rho_{i,k} \leq u_n\right) \\
&\leq P\left(\max_{k\geq 1, 1\leq i\leq n} \rho_{i,k} > u_n) + P(|S_n^W - \mathrm{E}S_n^W| \geq t, \max_{k\geq 1, 1\leq i\leq n} \rho_{i,k} \leq u_n\}\right) \\
&\leq C_\rho/u_n + \mathrm{E}[P(|S_n^W - \mathrm{E}S_n^W| \geq t, \max\nolimits_{k\geq 1, 1\leq i\leq n} \rho_{i,k} \leq u_n|\boldsymbol{W})] \\
&\leq C_\rho/u_n + 2\exp\left\{-\frac{t^2}{16nu_n(wC_T)^2 + 4wC_T t}\right\}.
\end{aligned}
$$

3.2. Some lemmas

Lemma 3. *[Symmetrization theorem with different functions] Let* $\varepsilon_1, ..., \varepsilon_n$ *be a Rademacher sequence with uniform distribution on* $\{-1, 1\}$*, independent of* $X_1, ..., X_n$ *and* $g_i \in \mathcal{G}_i$*. Then,*

$$
\mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^n [g_i(X_i) - \mathrm{E}\{g_i(X_i)\}]\right|\right) \leq 2\mathrm{E}\left[\mathrm{E}_\epsilon\left\{\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^n \epsilon_i g_i(X_i)\right|\right\}\right],
$$

where $\mathrm{E}_\epsilon\{\cdot\}$ refers to the expectation w.r.t. $\epsilon_1, ..., \epsilon_n$.

Proof: Let $\{X_i'\}_{i=1}^n$ be an independent copy of $\{X_i\}_{i=1}^n$. The E' denotes the expectation w.r.t. $\{X_i'\}_{i=1}^n$, and let $\mathcal{F}_n' = \sigma\left(X_1', \cdots, X_n'\right)$. So,

$$
\begin{aligned}
\mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^n [g_i(X_i) - \mathrm{E}\{g_i(X_i)\}]\right|\right) &= \mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\mathrm{E}'\sum_{i=1}^n [g_i(X_t) - g_i(X_i')]|\mathcal{F}_n'\right|\right) \\
\text{(Jensen's inequality of the absolute function)} &\leq \mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n} \mathrm{E}'\left|\sum_{i=1}^n [g_i(X_t) - g_i(X_i')]\right| |\mathcal{F}_n'\right) \\
\text{(Jensen's inequality of the max function)} &\leq \mathrm{E}\left(\mathrm{E}' \sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^n [g_i(X_t) - g_i(X_i')]\right| |\mathcal{F}_n'\right) \\
&= \mathrm{E}\left(\sup_{f_1,\cdots,f_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^n [g_i(X_t) - g_i(X_i')]\right|\right),
\end{aligned}
$$

where we use the conditional expectation version of Jensen's inequalities.

Since $\varepsilon_i[g_i(X_i) - g_i(X_i')]$ and $g_i(X_i) - g_i(X_i')$ have the same distribution, then,

$$
= \mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^n \varepsilon_i[g_i(X_i) - g_i(X_i')]\right|\right) \leq 2\mathrm{E}\left[\mathrm{E}_\epsilon\left\{\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^n \epsilon_i g_i(X_i)\right|\right\}\right].
$$

If $g_i = f$, $\mathcal{G}_i = \mathcal{F}$ for $i = 1, 2, \cdots, n$, then we have the classical symmetrization theorem.

Lemma 4. *[Symmetrization Theorem, Lemma 2.3.1 in [17]] Let* $\varepsilon_1, ..., \varepsilon_n$ *be a Rademacher sequence with uniform distribution on* $\{-1, 1\}$*, independent of* $X_1, ..., X_n$ *and* $f \in \mathcal{F}$*. Then, we have*

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\left[f(X_i)-\mathrm{E}\left\{f(X_i)\right\}\right]\right|\right]\le 2\mathrm{E}\left[\mathrm{E}_{\epsilon}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right\}\right],$$

where $\mathrm{E}[\cdot]$ refers to the expectation w.r.t. $X_1,...,X_n$ and $\mathrm{E}_{\epsilon}\left\{\cdot\right\}$ w.r.t. $\epsilon_1,...,\epsilon_n$.

3.3. Concentration for strongly log-concave discrete distributions

In this section, we restate the applications of the concentration inequality for a function of the data under the so-called strongly log-concave discrete distribution assumption, which was used in the Supplementary Material of [21]. We utilized the convex geometry approach to establish the tail bounds. In convex geometry, the following discrete version of the Prékopa-Leindler inequality can be found in Theorem 1.2 of [5]. The discrete version of the Prékopa-Leindler inequality is an essential inequality when deriving concentration inequalities of strongly log-concave counting measures. This shares the same idea when we consider the continuous version of Prékopa-Leindler inequality (see Theorem 3.15 of [18]).

Let $\lfloor r\rfloor=\max\{m\in\mathbb{Z};m\le r\}$ be the lower integer part of $r\in\mathbb{R}$, and $\lceil r\rceil=-\lfloor -r\rfloor$ be the upper integer part. Denote $\lfloor \boldsymbol{x}\rfloor=(\lfloor x_1\rfloor,\ldots\lfloor x_n\rfloor)$ and $\lceil \boldsymbol{x}\rceil=(\lceil x_1\rceil,\ldots,\lceil x_n\rceil)$.

Lemma 5. *[Discrete Prékopa-Leindler inequality] Let* $f,g,h,k:\mathbb{Z}^n\to[0,\infty)$ *be functions that satisfy the following inequality:*

$$f(\boldsymbol{x})g(\boldsymbol{y})\le h(\lfloor\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\rfloor)k(\lceil(1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y}\rceil),\quad\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{Z}^n,\quad\forall\lambda\in[0,1].\tag{3.11}$$

Then, we have

$$\left(\sum_{\boldsymbol{x}\in Z^n}f(\boldsymbol{x})\right)\left(\sum_{\boldsymbol{x}\in Z^n}g(\boldsymbol{x})\right)\le\left(\sum_{\boldsymbol{x}\in\mathbb{Z}^n}h(\boldsymbol{x})\right)\left(\sum_{\boldsymbol{x}\in Z^n}k(\boldsymbol{x})\right).$$

From a geometric perspective, the Prékopa-Leindler inequality is a valuable method to prove concentration inequalities under Lipschitz functions of strongly log-concave distributions. From the idea in [12], a distribution P with a density $p(\boldsymbol{x})$ (w.r.t. the counting measure) is said to be strongly discrete log-concave, if $\psi(\boldsymbol{x})=:-\log p(\boldsymbol{x}):\mathbb{Z}^n\to\mathbb{R}$ is *strongly midpoint log-convex* for some $\gamma>0$:

$$\psi(\boldsymbol{x})+\psi(\boldsymbol{y})-\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil)-\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)\ge\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2,\quad\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{Z}^n.\tag{3.12}$$

The inequality (3.12) is an extension of *strongly convexity* for continuous functions on $\mathbb{R}^n$:

$$\lambda\psi(\boldsymbol{x})+(1-\lambda)\psi(\boldsymbol{y})-\psi(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})\ge\frac{\gamma}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2,\quad\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n,\quad\forall\lambda\in[0,1],$$

with modulus of convexity γ [13].

Strongly log-convex property for a discrete density function requires that continuous functions are restricted on a lattice space. If $\gamma=0$, (3.12) turns to the *discrete midpoint convexity* property for $\psi(\boldsymbol{x})$

$$\psi(\boldsymbol{x})+\psi(\boldsymbol{y})\ge\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil)+\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor),\quad\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{Z}^n,$$

see [12]. However, directly restricting a continuous function to some lattice space may not necessarily obtain discrete convex functions. For the corresponding counterexample, see [19].

For one-dimensional P, the probability mass function $p(x)$ is said to be log-concave if the sequence $\{p(x)\}_{x\in\mathbb{Z}}$ is log-concave; that is, for any $\lambda n + (1-\lambda)m \in \mathbb{Z}$ with $m, n \in \mathbb{Z}$ and $\lambda \in (0,1)$, one has

$$p(\lambda n + (1-\lambda)m) \geq p(n)^{\lambda} p(m)^{1-\lambda}.$$

Proposition 1. *[The concentration inequality of strongly log-concave discrete distributions] Consider a strongly log-concave discrete distribution P_γ with index $\gamma > 0$ on $\mathbb{Z}^n$. For $f : \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz w.r.t. Euclidean norm, then,*

$$\Pr\{|f(\boldsymbol{X}) - \mathrm{E}f(\boldsymbol{X})| \geq t\} \leq 2e^{-\frac{\gamma t^2}{4L^2}}. \tag{3.13}$$

Proof: Let h be a zero-mean function with Lipschitz constant L (w.r.t. the Euclidean norm). It remains to prove the upper bound of a moment generating function $\mathrm{E}e^{h(\boldsymbol{X})} \leq e^{\frac{L^2}{\tau}}$. Then, for f with Lipschitz constant K and $\lambda \in \mathbb{R}$, we apply the upper bound to the zero-mean function $h(\boldsymbol{X}) := \lambda(f(\boldsymbol{X}) - \mathrm{E}f(\boldsymbol{X}))$, which has Lipschitz constant $L = \lambda K$. Given $\lambda \in (0,1)$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n$, define the proximity operator of h as

$$l(\boldsymbol{y}) := \inf_{\boldsymbol{x}\in\mathbb{Z}^n}\left\{h(\boldsymbol{x}) + \frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right\}.$$

With this proximity operator, the proof is proceeding by using the discrete Prekopa-Leindler inequality (Lemma 5) with $\lambda = 1/2$, $h(\boldsymbol{t}) = k(\boldsymbol{t}) =: p(\boldsymbol{t}) = e^{-\psi(\boldsymbol{t})}$, $f(\boldsymbol{x}) := e^{-h(\boldsymbol{x})-\psi(\boldsymbol{x})}$ and $g(\boldsymbol{y}) := e^{l(\boldsymbol{y})-\psi(\boldsymbol{y})}$.

We check that

$$e^{\frac{1}{2}[l(\boldsymbol{y})-h(\boldsymbol{x})-\psi(\boldsymbol{y})-\psi(\boldsymbol{x})]} \leq e^{-\frac{1}{2}\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil)} \cdot e^{-\frac{1}{2}\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)} \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n. \tag{3.14}$$

Then, (3.14) satisfies Lemma 5 with $\lambda = 1/2$.

By discrete strong convexity of the function ψ

$$\frac{1}{2}[\psi(\boldsymbol{x}) + \psi(\boldsymbol{y}) - \psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil) - \psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor) \geq \frac{\gamma}{8}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2,$$

and the proximity operator of h, we have

$$\begin{aligned}
&-\frac{1}{2}\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil) - \frac{1}{2}\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)\\
\geq\ &\frac{1}{2}\left\{l(\boldsymbol{y}) - h(\boldsymbol{x}) - \frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right\} - \frac{1}{2}\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil) - \frac{1}{2}\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)\\
\geq\ &\frac{1}{2}\{l(\boldsymbol{y}) - h(\boldsymbol{x})\} - \frac{1}{2}\psi(\boldsymbol{y}) - \frac{1}{2}\psi(\boldsymbol{x}),
\end{aligned}$$

which verifies (3.14).

By $\sum_{\boldsymbol{x}\in\mathbb{Z}^n} h(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in\mathbb{Z}^n} k(\boldsymbol{x}) = 1$, we know that Lemma 5 gives

$$\mathrm{E}e^{l(\boldsymbol{Y})}\mathrm{E}e^{-h(\boldsymbol{X})} = \sum_{\boldsymbol{x}\in\mathbb{Z}^n} e^{-h(\boldsymbol{x})-\psi(\boldsymbol{x})} \sum_{\boldsymbol{y}\in\mathbb{Z}^n} e^{l(\boldsymbol{y})-\psi(\boldsymbol{y})} \leq 1.$$

Then, Jensen's inequality implies

$$\mathrm{E}e^{l(\boldsymbol{Y})} \le (\mathrm{E}e^{-h(\boldsymbol{X})})^{-1} \le (e^{\mathrm{E}[-h(\boldsymbol{X})]})^{-1} = 1,$$

where in the last equality we use $\mathrm{E}[-h(\boldsymbol{X})] = \mathrm{E}[\lambda(f(\boldsymbol{X}) - \mathrm{E}f(\boldsymbol{X}))] = 0$. The definition of the proximity operator shows

$$\begin{aligned} 1 \ge \mathrm{E}e^{l(\boldsymbol{y})} = \mathrm{E}e^{\inf_{\boldsymbol{x}\in\mathbb{Z}^n}\left\{h(\boldsymbol{x})+\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{Y}\|_2^2\right\}} &= \mathrm{E}e^{\inf_{\boldsymbol{x}\in\mathbb{Z}^n}\left\{h(\boldsymbol{Y})+[h(\boldsymbol{x})-h(\boldsymbol{Y})]+\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{Y}\|_2^2\right\}} \\ &\ge \mathrm{E}e^{h(\boldsymbol{Y})+\inf_{\boldsymbol{x}\in\mathbb{R}^n}\left\{-L\|\boldsymbol{x}-\boldsymbol{Y}\|_2+\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{Y}\|_2^2\right\}} = \mathrm{E}e^{h(\boldsymbol{Y})-L^2/\gamma}, \end{aligned}$$

where the second last inequality due to L-Lipschitz of h, i.e., $|h(\boldsymbol{x}) - h(\boldsymbol{Y})| \le L\|\boldsymbol{x} - \boldsymbol{Y}\|_2$.

Then, we have $\mathrm{E}e^{\lambda(f(\boldsymbol{X})-\mathrm{E}f(\boldsymbol{X})]} \le e^{\frac{1}{2}\cdot\lambda^2\cdot\frac{2L^2}{\gamma}}$ for all $\lambda \in \mathbb{R}$. This means that $f(\boldsymbol{X}) - \mathrm{E}f(\boldsymbol{X}) \sim \mathrm{subG}(\frac{2L^2}{\gamma})$, hence the tail bound is (3.13).

4. Conclusions

Non-asymptotic statistical inference on high-dimensional data is important for many fields, such as data mining and machine learning. In this paper, we derived a novel concentration inequality for the sum of independent sub-Gaussian variables with random dependent weights in high-dimensional regression settings. We applied the proposed concentration inequality to obtain a high probability bound for the stochastic Lipschitz constant for negative binomial loss functions involved in Lasso-penalized negative binomial regressions, and used this bound to study oracle inequalities for the Lasso estimators. The usefulness of the proposed concentration inequality in applications was justified by solid theoretical proofs.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the reviewers and Dr. Pengfei Wang for their valuable suggestions that have significantly improved the quality of this paper. This work was supported in part by the National Natural Science Foundation of China under grant numbers 12101630 and 12261011, and the "double first-class" construction projects of Chinese universities under grant number ZG216S2348.

Conflict of interest

The authors declare no conflict of interest.

References

1. V. V. Buldygin, Y. V. Kozachenko, *Metric characterization of random variables and random processes*, Providence: American Mathematical Society, 2000.

2. S. Boucheron, G. Lugosi, P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford: Oxford University Press, 2013.

3. P. Bühlmann, S. A. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*, Berlin: Springer, 2011. https://doi.org/10.1007/978-3-642-20192-9

4. Z. Chi, A local stochastic Lipschitz condition with application to Lasso for high dimensional generalized linear models, arXiv:1009.1052. https://doi.org/10.48550/arXiv.1009.1052

5. D. Halikias, B. Klartag, B. A. Slomka, Discrete variants of Brunn-Minkowski type inequalities, *Annales de la Faculté des Sciences de Toulouse Mathématiques*, **30** (2021), 267–279. https://doi.org/10.5802/afst.1674

6. Q. Han, J. A. Wellner, Convergence rates of least squares regression estimators with heavy-tailed errors, *Ann. Statist.*, **47** (2019), 2286–2319. https://doi.org/10.1214/18-AOS1748

7. Q. Han, Multiplier U-processes: sharp bounds and applications, *Bernoulli*, **28** (2022), 87–124. https://doi.org/10.3150/21-BEJ1334

8. W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.*, **58** (1963), 13–30. https://doi.org/10.1080/01621459.1963.10500830

9. J. Kahane, Propriétés locales des fonctions à séries de Fourier aléatoires, *Stud. Math.*, **19** (1960), 1–25. https://doi.org/10.4064/sm-19-1-1-25

10. S. Li, H. Wei, X. Lei, Heterogeneous overdispersed count data regressions via double-penalized estimations, *Mathematics*, **10** (2022), 1700. https://doi.org/10.3390/math10101700

11. S. Mendelson, Upper bounds on product and multiplier empirical processes, *Stoch. Proc. Appl.*, **126** (2016), 3652–3680. https://doi.org/10.1016/j.spa.2016.04.019

12. S. Moriguchi, K. Murota, A. Tamura, F. Tardella, Discrete midpoint convexity, *Math. Oper. Res.*, **45** (2020), 99–128. https://doi.org/10.1287/moor.2018.0984

13. M. W. Mahoney, J. C. Duchi, A. C. Gilbert, *The mathematics of data*, Providence: American Mathematical Society, 2018.

14. P. Massart, Some applications of concentration inequalities to statistics, *Annales de la Facult des Sciences de Toulouse Mathmatiques*, **9** (2000), 245–303. https://doi.org/10.5802/afst.961

15. P. Rigollet, J. C. Hütter, *High dimensional statistics*, New York: Spring, 2019.

16. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, arXiv:1011.3027. https://doi.org/10.48550/arXiv.1011.3027

17. A. W. Vaart, J. A. Wellner, *Weak convergence and empirical processes: with applications to statistics*, New York: Springer, 1996. https://doi.org/10.1007/978-1-4757-2545-2

18. M. J. Wainwright, *High-dimensional statistics: a non-asymptotic viewpoint*, Cambridge: Cambridge University Press, 2019.

19. Ü. Yüceer, Discrete convexity: convexity for functions defined on discrete spaces, *Discrete Appl. Math.*, **119** (2002), 297–304. https://doi.org/10.1016/S0166-218X(01)00191-3

20. H. Zhang, S. Chen, Concentration inequalities for statistical inference, *Commun. Math. Res.*, **37** (2021), 1–85 https://doi.org/10.4208/cmr.2020-0041

21. H. Zhang, J. Jia, Elastic-net regularized high-dimensional negative binomial regression: consistency and weak signals detection, *Stat. Sinica*, **32** (2022), 181–207. https://doi.org/10.5705/SS.202019.0315

22. H. Zhang, X. Lei, Growing-dimensional partially functional linear models: non-asymptotic optimal prediction error, *Phys. Scr.*, **98** (2023), 095216. https://doi.org/10.1088/1402-4896/aceac0

23. H. Zhang, H. Wei, G. Cheng, Tight non-asymptotic inference via sub-Gaussian intrinsic moment norm, arXiv:2303.07287. https://doi.org/10.48550/arXiv.2303.07287