Journal of Machine Learning Research 17 (2016) 1-52 Submitted 2/16; Revised 7/16; Published 11/16

Newton-Stein Method: An Optimization Method for GLMs via Stein's Lemma

Murat A. Erdogdu ERDOGDU@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305-4065, USA

Editor: Qiang Liu

Abstract

We consider the problem of efficiently computing the maximum likelihood estimator in *Generalized Linear Models* (GLMs) when the number of observations is much larger than the number of coefficients ($n \gg p \gg 1$). In this regime, optimization algorithms can immensely benefit from approximate second order information. We propose an alternative way of constructing the curvature information by formulating it as an estimation problem and applying a *Stein-type lemma*, which allows further improvements through sub-sampling and eigenvalue thresholding. Our algorithm enjoys fast convergence rates, resembling that of second order methods, with modest per-iteration cost. We provide its convergence analysis for the general case where the rows of the design matrix are samples from a sub-Gaussian distribution. We show that the convergence has two phases, a quadratic phase followed by a linear phase. Finally, we empirically demonstrate that our algorithm achieves the highest performance compared to various optimization algorithms on several data sets.

Keywords: Optimization, Generalized Linear Models, Newton's method, Sub-sampling

1. Introduction

Generalized Linear Models (GLMs) play a crucial role in numerous statistical and machine learning problems. GLMs formulate the natural parameter in exponential families as a linear model and provide a miscellaneous framework for statistical methodology and supervised learning tasks. Celebrated examples include linear, logistic, multinomial regressions and applications to graphical models (Nelder and Baker, 1972; McCullagh and Nelder, 1989; Koller and Friedman, 2009).

In this paper, we focus on how to solve the maximum likelihood problem efficiently in the GLM setting when the number of observations n is much larger than the dimension of the coefficient vector p, i.e., $n \gg p \gg 1$. GLM optimization task is typically expressed as a minimization problem where the objective function is the negative log-likelihood that is denoted by $\ell(\beta)$ where $\beta \in \mathbb{R}^p$ is the coefficient vector. Many optimization algorithms are available for such minimization problems (Bishop, 1995; Boyd and Vandenberghe, 2004; Nesterov, 2004). However, only a few uses the special structure of GLMs. In this paper, we consider updates that are specifically designed for GLMs, which are of the from

$$\beta \leftarrow \beta - \gamma \mathbf{Q} \nabla_\beta \ell(\beta)\,, \tag{1}$$

where γ is the step size and $\mathbf{Q}$ is a scaling matrix which provides curvature information.

For the updates of the form Equation 1, the performance of the algorithm is mainly determined by the scaling matrix $\mathbf{Q}$. Classical *Newton's method* and *natural gradient descent* can be recovered by simply taking $\mathbf{Q}$ to be the inverse Hessian and the inverse Fisher's information at the current iterate, respectively (Amari, 1998; Nesterov, 2004). Second order methods may achieve quadratic convergence rate, yet they suffer from excessive cost of computing the scaling matrix at every iteration. On the other hand, if we take $\mathbf{Q}$ to be the identity matrix, we recover the standard *gradient descent* which has a linear convergence rate. Although the convergence rate of gradient descent is considered slow compared to that of second order methods such as Newton's method, modest per-iteration cost makes it practical for large-scale optimization.

The trade-off between convergence rate and per-iteration cost has been extensively studied (Bishop, 1995; Boyd and Vandenberghe, 2004; Nesterov, 2004). In $n \gg p \gg 1$ regime, the main objective is to construct a scaling matrix $\mathbf{Q}$ that is computational feasible which also provides sufficient curvature information. For this purpose, several Quasi-Newton methods have been proposed (Bishop, 1995; Nesterov, 2004). Updates given by Quasi-Newton methods satisfy an equation which is often called the *Quasi-Newton relation*. A well-known member of this class of algorithms is the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

In this paper, we propose a Newton-type algorithm that utilizes the special structure of GLMs by relying on a Stein-type lemma (Stein, 1981). It attains fast convergence rates with low per-iteration cost. We call our algorithm *Newton-Stein* method which we abbreviate as *NewSt* . Our contributions can be summarized as follows:

- We recast the problem of constructing a scaling matrix as an estimation problem and apply a Stein-type lemma along with the sub-sampling technique to form a computationally feasible $\mathbf{Q}$.

- Newton-Stein method allows further improvements through eigenvalue shrinkage, eigenvalue thresholding, sub-sampling and various other techniques that are available for covariance estimation.

- Excessive per-iteration cost of $\mathcal{O}(np^2+p^3)$ of Newton's method is replaced by $\mathcal{O}(np+p^2)$ per-iteration cost and a one-time $\mathcal{O}(|S|p^2)$ cost, where $|S|$ is the sub-sample size.

- Assuming that the rows of the design matrix are i.i.d. and have bounded support (or sub-Gaussian), and denoting the iterates of Newton-Stein method by $\{\hat{\beta}^t\}_t$, we prove a bound of the form

$$\big\|\hat{\beta}^{t+1}-\beta_*\big\|_2 \leq \tau_1\big\|\hat{\beta}^t-\beta_*\big\|_2 + \tau_2\big\|\hat{\beta}^t-\beta_*\big\|_2^2, \tag{2}$$

  where β_* is the true minimizer and τ_1, τ_2 are the convergence coefficients. The above bound implies that the local convergence starts with a quadratic phase and transitions into linear as the iterate gets closer to the true minimizer. We further establish a global convergence result of Newton-Stein method coupled with a line search algorithm.

- We demonstrate the performance of Newton-Stein method on real and synthetic data sets by comparing it to commonly used optimization algorithms.

The rest of the paper is organized as follows: Section 1.1 surveys the related work and Section 1.2 introduces the notations we use throughout the paper. Section 2 briefly discusses the GLM framework and its relevant properties. In Section 3, we introduce Newton-Stein method, develop its intuition, and discuss the computational aspects. Section 4 covers the theoretical results and in Section 4.4 we discuss how to choose the algorithm parameters. Section 5 provides the empirical results where we compare the proposed algorithm with several other methods on four data sets. Finally, in Section 6, we conclude with a brief discussion along with a few future research directions.

1.1 Related Work

There are numerous optimization techniques that can be used to find the maximum likelihood estimator in GLMs. For moderate values of n and p, the classical second order methods such as Newton's method (also referred to as Newton-Raphson) are commonly used. In large-scale problems, data dimensionality is the main factor while determining the optimization method, which typically falls into one of two major categories: online and batch methods. Online methods use a gradient (or sub-gradient) of a single, randomly selected observation to update the current iterate (Robbins and Monro, 1951). Their per-iteration cost is independent of n, but the convergence rate might be extremely slow. There are several extensions of the classical stochastic descent algorithms, providing significant improvement and improved stability (Bottou, 2010; Duchi et al., 2011; Schmidt et al., 2013; Kolte et al., 2015).

On the other hand, batch algorithms enjoy faster convergence rates, though their per-iteration cost may be prohibitive. In particular, second order methods enjoy quadratic convergence, but constructing the Hessian matrix generally requires excessive amount of computation. To remedy this issue, most research is focused on designing an approximate and cost-efficient scaling matrix. This idea lies at the core of Quasi-Newton methods such as BFGS (Bishop, 1995; Nesterov, 2004).

Another approach to construct an approximate Hessian makes use of sub-sampling techniques (Martens, 2010; Byrd et al., 2011; Vinyals and Povey, 2011; Erdogdu and Montanari, 2015; Roosta-Khorasani and Mahoney, 2016a,b). Many contemporary learning methods rely on sub-sampling as it is simple and it provides significant boost over the first order methods. Further improvements through conjugate gradient methods and Krylov sub-spaces are available. Sub-sampling can also be used to obtain an approximate solution, with certain large deviation guarantees (Dhillon et al., 2013).

There are many composite variants of the aforementioned methods, that mostly combine two or more techniques. Well-known composite algorithms are the combinations of sub-sampling and Quasi-Newton (Schraudolph et al., 2007; Byrd et al., 2016), stochastic and deterministic gradient descent (Friedlander and Schmidt, 2012), natural gradient and Newton's method (Le Roux and Fitzgibbon, 2010), natural gradient and low-rank approximation (Le Roux et al., 2008), sub-sampling and eigenvalue thresholding (Erdogdu and Montanari, 2015).

Lastly, algorithms that specialize on certain types of GLMs include coordinate descent methods for the penalized GLMs (Friedman et al., 2010), trust region Newton-type methods (Lin et al., 2008), and approximation methods (Erdogdu et al., 2016b,a).

1.2 Notation

Let $[n] = \{1, 2, ..., n\}$ and denote by $|S|$, the size of a set S. The gradient and the Hessian of f with respect to β are denoted by $\nabla_\beta f$ and $\boldsymbol{\nabla}^2_\beta f$, respectively. The j-th derivative of a function $f(w)$ is denoted by $f^{(j)}(w)$. For a vector x and a symmetric matrix $\mathbf{X}$, $\|x\|_2$ and $\|\mathbf{X}\|_2$ denote the ℓ_2 and spectral norms of x and $\mathbf{X}$, respectively. $\|x\|_{\psi_2}$ denotes the sub-Gaussian norm, which will be defined later. S^{p-1} denotes the p-dimensional sphere. $\mathcal{P}_\mathcal{C}$ denotes the projections onto the set $\mathcal{C}$, and $B_p(R) \subset \mathbb{R}^p$ denotes the p-dimensional ball of radius R. For a random variable x and density f, $x \sim f$ means that the distribution of x follows the density f. Multivariate Gaussian density with mean $\mu \in \mathbb{R}^p$ and covariance $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is denoted as $\mathsf{N}_p(\mu, \mathbf{\Sigma})$. For random variables x, y, $d(x, y)$ and $\mathfrak{D}(x, y)$ denote probability metrics (will be explicitly defined) measuring the distance between the distributions of x and y. $\mathcal{N}_{[]}(\cdots)$ and T_ϵ denote the bracketing number and ϵ-net.

2. Generalized Linear Models

Distribution of a random variable $y \in \mathbb{R}$ belongs to an exponential family with natural parameter $\eta \in \mathbb{R}$ if its density can be written as

$$f(y|\eta) = e^{\eta y - \phi(\eta)} h(y),$$

where ϕ is the *cumulant generating function* and h is the *carrier density*. Let $y_1, y_2, ..., y_n$ be independent observations such that $\forall i \in [n]$, $y_i \sim f(y_i|\eta_i)$. Denoting $\eta = (\eta_1, ..., \eta_n)^T$, the joint likelihood can be written as

$$f(y_1, y_2, ..., y_n|\eta) = \exp\left\{\sum_{i=1}^n [y_i\eta_i - \phi(\eta_i)]\right\} \prod_{i=1}^n h(y_i). \tag{3}$$

We consider the problem of learning the maximum likelihood estimator in the above exponential family framework, where the vector $\eta \in \mathbb{R}^n$ is modeled through the linear relation,

$$\eta = \mathbf{X}\beta,$$

for some design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with rows $x_i \in \mathbb{R}^p$, and a coefficient vector $\beta \in \mathbb{R}^p$. This formulation is known as *Generalized Linear Models* (GLMs) with canonical links. The cumulant generating function ϕ determines the class of GLMs, i.e., for ordinary least squares (OLS) $\phi(z) = z^2/2$, for logistic regression (LR) $\phi(z) = \log(1+e^z)$, and for Poisson regression (PR) $\phi(z) = e^z$.

Finding the maximum likelihood estimator in the above formulation is equivalent to minimizing the negative log-likelihood function $\ell(\beta)$,

$$\ell(\beta) = \frac{1}{n}\sum_{i=1}^n \left[\phi(\langle x_i, \beta\rangle) - y_i\langle x_i, \beta\rangle\right], \tag{4}$$

where $\langle x, \beta\rangle$ is the inner product between the vectors x and β. The relation to OLS and LR can be seen much easier by plugging in the corresponding $\phi(z)$ in Equation 4. The gradient

and the Hessian of $\ell(\beta)$ can be written as:

$$\nabla_\beta \ell(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\phi^{(1)}(\langle x_i,\beta\rangle)x_i - y_i x_i\right], \qquad \boldsymbol{\nabla}^2_\beta \ell(\beta) = \frac{1}{n}\sum_{i=1}^{n}\phi^{(2)}(\langle x_i,\beta\rangle)x_i x_i^T. \tag{5}$$

For a sequence of scaling matrices $\{\mathbf{Q}^t\}_{t>0} \in \mathbb{R}^{p\times p}$, we consider iterations of the form

$$\hat{\beta}^{t+1} = \hat{\beta}^t - \gamma_t \mathbf{Q}^t \nabla_\beta \ell(\hat{\beta}^t)$$

where γ_t is the step size. The above iteration is our main focus, but with a new approach on how to compute the sequence of matrices $\{\mathbf{Q}^t\}_{t>0}$. We will formulate the problem of finding a scalable $\mathbf{Q}^t$ as an estimation problem and apply a Stein-type lemma that provides us with a computationally efficient update rule.

3. Newton-Stein Method

Classical Newton-Raphson (or simply Newton's) method is the standard approach for training GLMs for moderately large date sets. However, its per-iteration cost makes it impractical for large-scale optimization. The main bottleneck is the computation of the Hessian matrix that requires $\mathcal{O}(np^2)$ flops which is prohibitive when $n \gg p \gg 1$. Numerous methods have been proposed to achieve the fast convergence rate of Newton's method while keeping the per-iteration cost manageable. To this end, a popular approach is to construct a scaling matrix $\mathbf{Q}^t$, which approximates the inverse Hessian at every iteration t.

The task of constructing an approximate Hessian can be viewed as an estimation problem. Assuming that the rows of $\mathbf{X}$ are i.i.d. random vectors, the Hessian of the negative log-likelihood of GLMs with a cumulant generating function ϕ has the following sample average form

$$\left[\mathbf{Q}^t\right]^{-1} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \phi^{(2)}(\langle x_i,\beta\rangle) \approx \mathbb{E}[xx^T\phi^{(2)}(\langle x,\beta\rangle)]\,.$$

We observe that $\left[\mathbf{Q}^t\right]^{-1}$ is just a sum of i.i.d. matrices. Hence, the true Hessian is nothing but a sample mean estimator to its expectation. Another natural estimator would be the sub-sampled Hessian method which is extensively studied by Martens, 2010; Byrd et al., 2011; Erdogdu and Montanari, 2015; Roosta-Khorasani and Mahoney, 2016a. Therefore, our goal is to propose an estimator for the population level Hessian that is also computationally efficient. Since n is large, the proposed estimator will be close to the true Hessian.

We use the following Stein-type lemma to find a more efficient estimator to the expectation of the Hessian.

Lemma 1 (Stein-type lemma) *Assume that $x \sim \mathsf{N}_p(0,\boldsymbol{\Sigma})$ and $\beta \in \mathbb{R}^p$ is a constant vector. Then for any function $f:\mathbb{R}\to\mathbb{R}$ that is twice* "weakly" *differentiable, we have*

$$\mathbb{E}\left[xx^T f(\langle x,\beta\rangle)\right] = \mathbb{E}\left[f(\langle x,\beta\rangle)\right]\boldsymbol{\Sigma} + \mathbb{E}\left[f^{(2)}(\langle x,\beta\rangle)\right]\boldsymbol{\Sigma}\beta\beta^T\boldsymbol{\Sigma}\,. \tag{6}$$

Proof The proof will follow from integration by parts. Let $g(x|\boldsymbol{\Sigma})$ denote the density of a multivariate normal random variable x with mean 0 and covariance $\boldsymbol{\Sigma}$. We recall the basic

Algorithm 1 Newton-Stein Method

Input: $\hat{\beta}^0, |S|, \epsilon, \{\gamma_t\}_{t\geq 0}$.

1. Estimate the covariance using a random sub-sample $S \subset [n]$:

   $\widehat{\mathbf{\Sigma}}_S = \frac{1}{|S|}\sum_{i\in S} x_i x_i^T$.

2. **while** $\left\|\hat{\beta}^{t+1} - \hat{\beta}^t\right\|_2 > \epsilon$ **do**

   $\hat{\mu}_2(\hat{\beta}^t) = \frac{1}{n}\sum_{i=1}^n \phi^{(2)}(\langle x_i, \hat{\beta}^t\rangle), \qquad \hat{\mu}_4(\hat{\beta}^t) = \frac{1}{n}\sum_{i=1}^n \phi^{(4)}(\langle x_i, \hat{\beta}^t\rangle),$

   $$\mathbf{Q}^t = \frac{1}{\hat{\mu}_2(\hat{\beta}^t)}\left[\widehat{\mathbf{\Sigma}}_S^{-1} - \frac{\hat{\beta}^t[\hat{\beta}^t]^T}{\hat{\mu}_2(\hat{\beta}^t)/\hat{\mu}_4(\hat{\beta}^t) + \langle \widehat{\mathbf{\Sigma}}_S\hat{\beta}^t, \hat{\beta}^t\rangle}\right],$$

   $\hat{\beta}^{t+1} = \hat{\beta}^t - \gamma_t\, \mathbf{Q}^t \nabla_\beta \ell(\hat{\beta}^t)$,

   $t \leftarrow t+1$.

3. **end while**

Output: $\hat{\beta}^t$.

identity $xg(x|\mathbf{\Sigma})dx = -\mathbf{\Sigma} dg(x|\mathbf{\Sigma})$ and write

$$\begin{aligned}
\mathbb{E}[xx^T f(\langle x,\beta\rangle)] =& \int xx^T f(\langle x,\beta\rangle) g(x)dx,\\
=&\mathbf{\Sigma}\left\{\int f(\langle x,\beta\rangle) g(x|\mathbf{\Sigma})dx + \int \beta x^T f^{(1)}(\langle x,\beta\rangle) g(x|\mathbf{\Sigma})dx\right\},\\
=&\mathbf{\Sigma}\left\{\mathbb{E}[f(\langle x,\beta\rangle)] + \int \beta\beta^T f^{(2)}(\langle x,\beta\rangle) g(x|\mathbf{\Sigma})dx\mathbf{\Sigma}\right\},\\
=&\mathbb{E}[f(\langle x,\beta\rangle)]\mathbf{\Sigma} + \mathbb{E}\left[f^{(2)}(\langle x,\beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}.
\end{aligned}$$

■

The right hand side of Equation 6 is a rank-1 update to the first term. Hence, its inverse can be computed with $\mathcal{O}(p^2)$ cost. Quantities that change at each iteration are the ones that depend on β, i.e.,

$$\mu_2(\beta) = \mathbb{E}[\phi^{(2)}(\langle x,\beta\rangle)], \quad \text{and} \quad \mu_4(\beta) = \mathbb{E}[\phi^{(4)}(\langle x,\beta\rangle)].$$

Note that $\mu_2(\beta)$ and $\mu_4(\beta)$ are scalar quantities and they can be estimated by their corresponding sample means $\hat{\mu}_2(\beta)$ and $\hat{\mu}_4(\beta)$ (explicitly defined at Step 2 of Algorithm 1) respectively, with only $\mathcal{O}(np)$ computation.

To complete the estimation task suggested by Equation 6, we need an estimator for the covariance matrix $\mathbf{\Sigma}$. A natural estimator is the sample mean where, we only use a sub-sample of the indices $S \subset [n]$ so that the cost is reduced to $\mathcal{O}(|S|p^2)$ from $\mathcal{O}(np^2)$. Sub-sampling based sample mean estimator is denoted by $\widehat{\mathbf{\Sigma}}_S = \frac{1}{|S|}\sum_{i\in S} x_i x_i^T$, which is

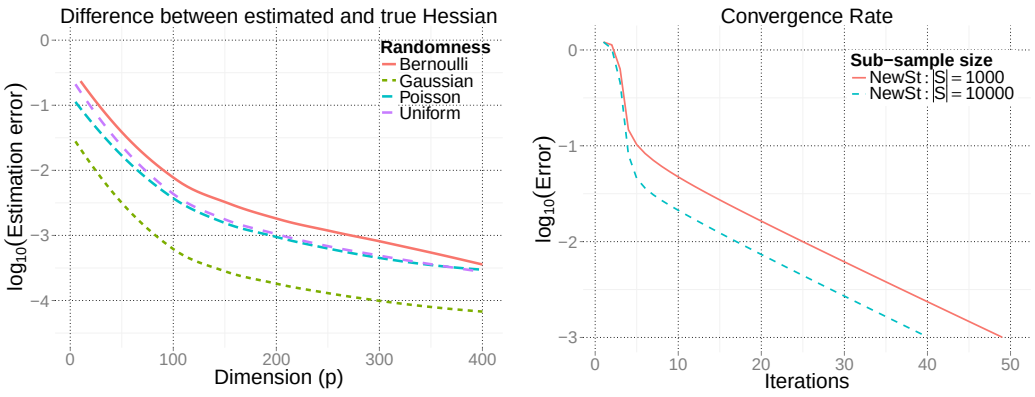

Figure 1: The left plot demonstrates the accuracy of proposed Hessian estimation over different distributions. Number of observations is set to be $n = \mathcal{O}(p\log(p))$. The right plot shows the phase transition in the convergence rate of Newton-Stein method (NewSt). Convergence starts with a quadratic rate and transitions into linear. Plots are obtained using *Covertype* data set.

widely used in large-scale problems (Vershynin, 2010). We highlight the fact that Lemma 1 replaces $\mathcal{O}(np^2)$ per-iteration cost of Newton's method with a one-time cost of $\mathcal{O}(np^2)$. We further use sub-sampling to reduce this one-time cost to $\mathcal{O}(|S|p^2)$, and obtain the following Hessian estimator at β

$$\underbrace{\left[\mathbf{Q}^t\right]^{-1}}_{\in\mathbb{R}^{p\times p}} = \underbrace{\hat{\mu}_2(\beta)}_{\in\mathbb{R}}\underbrace{\widehat{\mathbf{\Sigma}}_S}_{\in\mathbb{R}^{p\times p}} + \underbrace{\hat{\mu}_4(\beta)}_{\in\mathbb{R}}\overbrace{\underbrace{\widehat{\mathbf{\Sigma}}_S\beta\beta^T\widehat{\mathbf{\Sigma}}_S}_{\in\mathbb{R}^{p\times p}}}^{\text{rank-1 update}} \tag{7}$$

We emphasize that any covariance estimation method can be applied in the first step of the algorithm. There are various estimation techniques most of which rely on the concept of *shrinkage* (Cai et al., 2010; Donoho et al., 2013). This is because, important curvature information is generally contained in the largest few spectral features (Erdogdu and Montanari, 2015). In particular, for a given threshold r, we suggest to use the largest r eigenvalues of the sub-sampled covariance estimator $\widehat{\mathbf{\Sigma}}_S$, and setting rest of them to $(r+1)$-th eigenvalue. This operation helps denoising and provides additional computational benefits when inverting the covariance estimator (Erdogdu and Montanari, 2015).

Inverting the constructed Hessian estimator can make use of the low-rank structure. First, notice that the updates in Equation 7 are based on rank-1 matrix additions. Hence, we can simply apply Sherman–Morrison inversion formula to Equation 7 and obtain an explicit equation for the scaling matrix $\mathbf{Q}^t$ (Step 2 of Algorithm 1). This formulation would impose another inverse operation on the covariance estimator. We emphasize that this operation is performed once. Therefore, instead of $\mathcal{O}(p^3)$ per-iteration cost of Newton's method due to inversion, Newton-Stein method (NewSt) requires $\mathcal{O}(p^2)$ per-iteration and a one-time cost of $\mathcal{O}(p^3)$. Assuming that Newton-Stein and Newton methods converge in T_1 and T_2 iterations respectively, the overall complexity of Newton-Stein is $\mathcal{O}\left(npT_1 + p^2T_1 + (|S|+p)p^2\right) \approx$

$\mathcal{O}\left(npT_1 + p^2T_1 + |S|p^2\right)$ whereas that of Newton is $\mathcal{O}(np^2T_2 + p^3T_2)$. We show both empirically and theoretically that the quantities T_1 and T_2 are close to each other.

The convergence rate of Newton-Stein method has two phases. Convergence starts quadratically and transitions into linear rate when it gets close to the true minimizer. The phase transition behavior can be observed through the right plot in Figure 1. This is a consequence of the bound provided in Equation 2, which is the main result of our theorems on the local convergence (given in Section 4).

Even though Lemma 1 assumes that the covariates are multivariate Gaussian random vectors, in Section 4, the only assumption we make on the covariates is either bounded support or sub-Gaussianity, both of which cover a wide class of random variables including Bernoulli, elliptical distributions, bounded variables etc. The left plot of Figure 1 shows that the estimation is accurate for many distributions. This is a consequence of the fact that the proposed estimator in Equation 7 relies on the distribution of x only through inner products of the form $\langle x, v\rangle$, which in turn results in an approximate normal distribution due to the central limit theorem. To provide more intuition, we explain this through *zero-biased transformations* which is a general version of Stein's lemma for arbitrary distributions (Goldstein and Reinert, 1997).

Definition 2 *Let z be a random variable with mean 0 and variance σ^2. Then, there exists a random variable z^* that satisfies $\mathbb{E}[zf(z)] = \sigma^2\mathbb{E}[f^{(1)}(z^*)]$, for all differentiable functions f. The distribution of z^* is said to be the z-zero-bias distribution.*

The normal distribution is the unique distribution whose zero-bias transformation is itself (i.e. the normal distribution is a fixed point of the operation mapping the distribution of z to that of z^*). The distribution of z^* is referred to as z-zero-bias distribution and is entirely determined by the distribution of z. Properties such as existence can be found, for example, in Chen et al., 2010.

To provide some intuition behind the usefulness of Lemma 1 even for arbitrary distributions, we use zero-bias transformations. For simplicity, assume that the covariate vector x has i.i.d. entries from an arbitrary distribution with mean 0, and variance 1. Then the zero-bias transformation applied twice to the entry (i, j) of matrix $\mathbb{E}[xx^Tf(\langle x, \beta\rangle)]$ yields

$$\mathbb{E}[x_ix_jf(\langle x, \beta\rangle)] = \begin{cases} \mathbb{E}[f(\beta_ix_i^* + \sum_{k\neq i} x_k\beta_k)] + \beta_i^2\mathbb{E}\big[f^{(2)}(\beta_ix_i^{**} + \sum_{k\neq i} x_k\beta_k)\big] & \text{if} \quad i = j, \\ \beta_i\beta_j\mathbb{E}\big[f^{(2)}(\beta_ix_i^* + \beta_jx_j^* + \sum_{k\neq i,j} x_k\beta_k)\big] & \text{if} \quad i \neq j, \end{cases}$$

where x_i^* and x_i^{**} have x_i-zero-bias and x_i^*-zero-bias distributions, respectively. For each entry (i, j) at most two summands of $\langle x, \beta\rangle = \Sigma_k x_k\beta_k$ change their distributions. Therefore, if β is well spread and p is sufficiently large, the sums inside the expectations will behave similar to the inner product $\langle x, \beta\rangle$. Correspondingly, the above equations will be close to their Gaussian counterpart as given in Equation 6.

4. Theoretical Results

We start by introducing the terms that will appear in the theorems. Then we will provide two technical results on bounded and sub-Gaussian covariates. The proofs of the theorems are technical and provided in Appendix.

4.1 Preliminaries

Hessian estimation described in the previous section relies on a Gaussian approximation. For theoretical purposes, we use the following probability metric to quantify the gap between the distribution of x_i's and that of a normal vector.

Definition 3 *Given a family of functions $\mathcal{H}$, and random vectors $x, y \in \mathbb{R}^p$, for $\mathcal{H}$ and any $h \in \mathcal{H}$, define*

$$d_{\mathcal{H}}(x, y) = \sup_{h \in \mathcal{H}} d_h(x, y) \qquad \text{where} \qquad d_h(x, y) = \big|\mathbb{E}\left[h(x)\right] - \mathbb{E}\left[h(y)\right]\big|.$$

Many probability metrics can be expressed as above by choosing a suitable function class $\mathcal{H}$. Examples include *Total Variation* (TV), *Kolmogorov* and *Wasserstein* metrics (Gibbs and Su, 2002; Chen et al., 2010). Based on the second and the fourth derivatives of the cumulant generating function, we define the following function classes:

$$\mathcal{H}_1 = \left\{h(x) = \phi^{(2)}(\langle x, \beta\rangle) : \beta \in \mathcal{C}\right\}, \qquad \mathcal{H}_2 = \left\{h(x) = \phi^{(4)}(\langle x, \beta\rangle) : \beta \in \mathcal{C}\right\},$$
$$\mathcal{H}_3 = \left\{h(x) = \langle v, x\rangle^2 \phi^{(2)}(\langle x, \beta\rangle) : \beta \in \mathcal{C}, \|v\|_2 = 1\right\},$$

where $\mathcal{C} \in \mathbb{R}^p$ is a closed, convex set that is bounded by the radius R. Exact calculation of such probability metrics are often difficult. The general approach is to upper bound the distance by a more intuitive metric. In our case, we observe that $d_{\mathcal{H}_j}(x, y)$ for $j = 1, 2, 3$, can be easily upper bounded by $d_{\mathrm{TV}}(x, y)$ up to a scaling constant, when the covariates have bounded support.

In our theoretical results, we rely on projected updates onto a closed convex set $\mathcal{C}$, which are of the form

$$\hat{\beta}^{t+1} = \mathcal{P}_{\mathcal{C}}^t\left(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla_\beta \ell(\hat{\beta}^t)\right)$$

where the projection is defined as $\mathcal{P}_{\mathcal{C}}^t(\beta) = \mathrm{argmin}_{w \in \mathcal{C}} \frac{1}{2}\|w - \beta\|^2_{\mathbf{Q}^{t-1}}$, with $\mathcal{C}$ bounded by R. This is a special case of proximal Newton-type algorithms and further generalization is straightforward (See Lee et al., 2014). We will further assume that the covariance matrix has full rank and its smallest eigenvalue is lower bounded by a positive constant.

4.2 Bounded Covariates

We have the following per-step bound for the iterates generated by the Newton-Stein method, when the covariates are supported on a ball.

Theorem 4 (Local convergence) *Assume that the covariates $x_1, x_2, ..., x_n$ are i.i.d. random vectors supported on a ball of radius $\sqrt{K}$ with*

$$\mathbb{E}[x_i] = 0 \qquad \text{and} \qquad \mathbb{E}\big[x_i x_i^T\big] = \mathbf{\Sigma}.$$

Further assume that the cumulant generating function ϕ has bounded 2nd-5th derivatives and that the set $\mathcal{C}$ is bounded by R. For $\left\{\hat{\beta}^t\right\}_{t>0}$ given by the Newton-Stein method for $\gamma = 1$, define the event

$$\mathcal{E} = \left\{\inf_{\|u\|_2=1} \left|\mu_2(\hat{\beta}^t)\langle u, \mathbf{\Sigma} u\rangle + \mu_4(\hat{\beta}^t)\langle u, \mathbf{\Sigma}\hat{\beta}^t\rangle^2\right| > 2\kappa^{-1} \quad \forall t, \quad \beta_* \in \mathcal{C}\right\} \tag{8}$$

for some positive constant κ, and the optimal value β_*. If $n, |S|$ and p are sufficiently large, then there exist constants c, c_1, c_2 depending on the radii K, R, $\mathbb{P}(\mathcal{E})$ and the bounds on $\phi^{(2)}$ and $|\phi^{(4)}|$ such that conditioned on the event $\mathcal{E}$, with probability at least $1 - c/p^2$, we have

$$\left\|\hat{\beta}^{t+1} - \beta_*\right\|_2 \leq \tau_1 \left\|\hat{\beta}^t - \beta_*\right\|_2 + \tau_2 \left\|\hat{\beta}^t - \beta_*\right\|_2^2, \tag{9}$$

where the coefficients τ_1 and τ_2 are deterministic constants defined as

$$\tau_1 = \kappa \mathfrak{D}(x, z) + c_1 \kappa \sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}}, \qquad \tau_2 = c_2 \kappa, \tag{10}$$

and $\mathfrak{D}(x, z)$ is defined as

$$\mathfrak{D}(x, z) = \|\mathbf{\Sigma}\|_2 \; d_{\mathcal{H}_1}(x, z) + \|\mathbf{\Sigma}\|_2^2 R^2 \; d_{\mathcal{H}_2}(x, z) + d_{\mathcal{H}_3}(x, z), \tag{11}$$

for a multivariate Gaussian random variable z with the same mean and covariance as x_i's.

The bound in Equation 9 holds with high probability, and the coefficients τ_1 and τ_2 are deterministic constants which will describe the convergence behavior of the Newton-Stein method. Observe that the coefficient τ_1 is sum of two terms: $\mathfrak{D}(x, z)$ measures how accurate the Hessian estimation is, and the second term depends on the sub-sampling size $|S|$ and the data dimensions n, p.

Theorem 4 shows that the convergence of Newton-Stein method can be upper bounded by a compositely converging sequence, that is, the squared term will dominate at first providing us with a quadratic rate, then the convergence will transition into a linear phase as the iterate gets close to the optimal value. The coefficients τ_1 and τ_2 govern the linear and quadratic terms, respectively. The effect of sub-sampling appears in the coefficient of linear term. In theory, there is a threshold for the sub-sampling size $|S|$, namely $\mathcal{O}(n/\log(n))$, beyond which further sub-sampling has no effect. The transition point between the quadratic and the linear phases is determined by the sub-sampling size and the properties of the data. The phase transition behavior can be observed through the right plot in Figure 1.

Using the above theorem, we state the following corollary.

Corollary 5 *Assume that the assumptions of Theorem 4 hold. For a constant $\delta \geq \mathbb{P}\left(\mathcal{E}^C\right)$, and a tolerance ϵ satisfying*

$$\epsilon \geq 20R\left\{c/p^2 + \delta\right\},$$

and for an iterate satisfying $\mathbb{E}\big[\|\hat{\beta}^t - \beta_*\|_2\big] > \epsilon$, the following inequality holds for the iterates of Newton-Stein method,

$$\mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2\right] \leq \tilde{\tau}_1 \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2\right] + \tau_2 \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2^2\right],$$

where $\tilde{\tau}_1 = \tau_1 + 0.1$ and τ_1, τ_2 are as in Theorem 4.

The bound stated in the above corollary is an analogue of composite convergence (given in Equation 9) in expectation. Note that our results make strong assumptions on the derivatives of the cumulant generating function ϕ. We emphasize that these assumptions

are valid for linear and logistic regressions. An example that does not fit in our scheme is *Poisson regression* with $\phi(z) = e^z$. However, we observed empirically that the algorithm still provides significant improvement.

The following theorem characterizes the local convergence behavior of a compositely converging sequence.

Theorem 6 *Assume that the assumptions of Theorem 4 hold with $\tau_1 < 1$ and for $\vartheta = \left\|\hat{\beta}^0 - \beta_*\right\|_2$ define the interval $\Xi = \left(\frac{\tau_1\vartheta}{1-\tau_2\vartheta}, \vartheta\right)$. Conditioned on the event $\mathcal{E} \cap \left\{\vartheta < (1-\tau_1)/\tau_2\right\}$, there exists a constant c such that with probability at least $1 - c/p^2$, the number of iterations to reach a tolerance of ϵ cannot exceed*

$$\inf_{\xi \in \Xi} \mathcal{J}(\xi) \coloneqq \log_2\left(\frac{\log\left(\tau_1 + \tau_2\xi\right)}{\log\left((\tau_1/\xi + \tau_2)(1-\tau_1)/\tau_2\right)}\right) + \frac{\log(\epsilon/\xi)}{\log(\tau_1 + \tau_2\xi)}, \tag{12}$$

where the constants τ_1 and τ_2 are as in Theorem 4.

The expression in Equation 12 has two terms: the first one is due to the quadratic phase whereas the second one is due to the linear phase. To obtain the properties of local convergence, a locality constraint is required. We note that $\tau_1 < 1$ is a necessary assumption, which is satisfied for sufficiently large n and $|S|$.

In the following, we establish the global convergence of the Newton-Stein method coupled with a backtracking line search—which is explicitly given in Section 4.4.

Theorem 7 (Global Convergence) *Assume that the assumptions of Theorem 4 hold and at each step, the step size γ_t of the Newton-Stein method is determined by the backtracking line search with parameters a and b. Then conditioned on the event $\mathcal{E}$, there exists a constant c such that with probability at least $1 - c/p^2$, the sequence of iterates $\{\hat{\beta}^t\}_{t>0}$ generated by the Newton-Stein method converges globally.*

4.3 Sub-Gaussian Covariates

In this section, we carry our analysis to the more general case, where the covariates are sub-Gaussian vectors.

Theorem 8 (Local convergence) *Assume that $x_1, x_2, ..., x_n$ are i.i.d. sub-Gaussian random vectors with sub-Gaussian norm K such that*

$$\mathbb{E}[x_i] = 0, \qquad \mathbb{E}[\|x_i\|_2] = \mu \qquad \text{and} \qquad \mathbb{E}\left[x_i x_i^T\right] = \mathbf{\Sigma}.$$

Further assume that the cumulant generating function ϕ is uniformly bounded and has bounded 2nd-5th derivatives and that $\mathcal{C}$ is bounded by R. For $\left\{\hat{\beta}^t\right\}_{t>0}$ given by the Newton-Stein method and the event $\mathcal{E}$ in Equation 8, if we have $n, |S|$ and p sufficiently large and

$$n^{0.2}/\log(n) \gtrsim p,$$

then there exist constants c_1, c_2, c_3, c_4 depending on the eigenvalues of $\mathbf{\Sigma}$, the radius R, μ, $\mathbb{P}(\mathcal{E})$ and the bounds on $\phi^{(2)}$ and $|\phi^{(4)}|$ such that conditioned on the event $\mathcal{E}$, with probability at least $1 - c_1 e^{-c_2 p}$, the bound given in Equation 9 holds for constants

$$\tau_1 = \kappa\mathfrak{D}(x, z) + c_3\kappa\sqrt{\frac{p}{\min\left\{|S|, n^{0.2}/\log(n)\right\}}}, \qquad \tau_2 = c_4\kappa p^{1.5}, \tag{13}$$

where $\mathfrak{D}(x, z)$ defined as in Equation 11.

The above theorem is more restrictive than Theorem 4. We require n to be much larger than the dimension p. Also note that a factor of $p^{1.5}$ appears in the coefficient of the quadratic term. We also notice that the threshold for the sub-sample size reduces to $n^{0.2}/\log(n)$.

We have the following analogue of Corrolary 5.

Corollary 9 *Assume that the assumptions of Theorem 8 hold. For a constant $\delta \geq \mathbb{P}\left(\mathcal{E}^C\right)$, and a tolerance ϵ satisfying*

$$\epsilon \geq 20R\sqrt{c_1 e^{-c_2 p} + \delta},$$

and for an iterate satisfying $\mathbb{E}\big[\|\hat{\beta}^t - \beta_*\|_2\big] > \epsilon$, the iterates of Newton-Stein method will satisfy,

$$\mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2\right] \leq \tilde{\tau}_1 \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2\right] + \tau_2 \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2^2\right],$$

where $\tilde{\tau}_1 = \tau_1 + 0.1$ and τ_1, τ_2 are as in Theorem 8.

When the covariates are in fact multivariate normal, we have $\mathfrak{D}(x, z) = 0$ which implies that the coefficient τ_1 is smaller. Correspondingly, the quadratic phase lasts longer providing better performance.

We conclude this section by noting that the global convergence properties of the sub-Gaussian case is very similar to the previous section where we had bounded covariates.

4.4 Algorithm Parameters

Newton-Stein method takes two input parameters and for those, we suggest near-optimal choices based on our theoretical results. We further discuss the choice of a covariance estimation method which provides additional improvements to the proposed algorithm.

- *Sub-sample size:* Newton-Stein method uses a subset of indices to approximate the covariance matrix $\mathbf{\Sigma}$. Corollary 5.50 of Vershynin, 2010 proves that a sample size of $\mathcal{O}(p)$ is sufficient for sub-Gaussian covariates and that of $\mathcal{O}(p\log(p))$ is sufficient for arbitrary distributions supported in some ball to estimate a covariance matrix by its sample mean estimator. In the regime we consider, $n \gg p$, we suggest to use a sample size of $|S| = \mathcal{O}(p\log(p))$ for this task.

- *Covariance estimation method:* Many methods have been suggested to improve the estimation of the covariance matrix and almost all of them rely on the concept of *shrinkage* (Cai et al., 2010; Donoho et al., 2013). Therefore, we suggest to use a thresholding based approach suggested by Erdogdu and Montanari, 2015. For a given threshold r, we take the largest r eigenvalues of the sub-sampled covariance estimator, setting rest of them to $(r+1)$-th eigenvalue. Eigenvalue thresholding can be considered as a shrinkage operation which will retain only the important second order information. Choosing the rank threshold r can be simply done on the sample mean estimator of $\mathbf{\Sigma}$. After obtaining the sub-sampled estimate of the mean, one can either plot the spectrum and choose manually or use an optimal technique from Donoho

et al., 2013. The suggested method requires a single time $\mathcal{O}(rp^2)$ computation and reduces the cost of inversion from $O(p^3)$ to $O(rp^2)$. We highlight that the Newton-Stein method was originally presented with the eigenvalue thresholding in an early version of this paper (Erdogdu, 2015).

- *Step size:* Step size choices for the Newton-Stein method are quite similar to those of Newton-type methods (i.e., see Boyd and Vandenberghe, 2004). In the *damped phase*, one should use a line search algorithm such as *backtracking* with parameters $a \in (0, 0.5)$ and $b \in (0, 1)$. Defining the modified gradient (or composite gradient Lee et al., 2014) $D_\gamma(\hat{\beta}^t) = \frac{1}{\gamma}\big\{\hat{\beta}^t - \mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla \ell(\hat{\beta}^t))\big\}$, we compute the step size via

$$\gamma = \bar{\gamma}; \quad \textbf{while: } \ell\Big(\hat{\beta}^t - \gamma D_\gamma(\hat{\beta}^t)\Big) > \ell(\hat{\beta}^t) - a\gamma\langle \nabla \ell(\hat{\beta}^t), D_\gamma(\hat{\beta}^t)\rangle, \quad \gamma \leftarrow \gamma b.$$

  The above line search algorithm leads to global convergence with high probability as stated in Theorem 7.

  The step size choice for the local phase depends on the use of eigenvalue thresholding. If no shrinkage method is applied, line search algorithm should be initialized with $\bar{\gamma} = 1$. If a shrinkage method (e.g. eigenvalue thresholding) is applied, then choosing a larger local step size may provide faster convergence. If the data follows the r-spiked model, the optimal step size will be close to 1 if there is no sub-sampling. However, due to fluctuations resulting from sub-sampling, starting with $\bar{\gamma} = 1.2$ will provide faster local rates. This case has been explicitly studied in a preliminary version of this work (Erdogdu, 2015). A heuristic derivation and a detailed discussion can also be found in Section E in the Appendix.

5. Experiments

In this section, we validate the performance of Newton-Stein method through extensive numerical studies. We experimented on two commonly used GLM optimization problems, namely, *Logistic Regression* (LR) and *Linear Regression* (OLS). LR minimizes Equation 4 for the logistic function $\phi(z) = \log(1 + e^z)$, whereas OLS minimizes the same equation for $\phi(z) = z^2/2$. In the following, we briefly describe the algorithms that are used in the experiments:

- *Newton's Method* (NM) uses the inverse Hessian evaluated at the current iterate, and may achieve local quadratic convergence. NM steps require $\mathcal{O}(np^2 + p^3)$ computation which makes it impractical for large-scale data sets.

- *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) forms a curvature matrix by cultivating the information from the iterates and the gradients at each iteration. Under certain assumptions, the convergence rate is locally super-linear and the per-iteration cost is comparable to that of first order methods.

- *Limited Memory BFGS* (L-BFGS) is similar to BFGS, and uses only the recent few iterates to construct the curvature matrix, gaining significant performance in terms of memory usage.

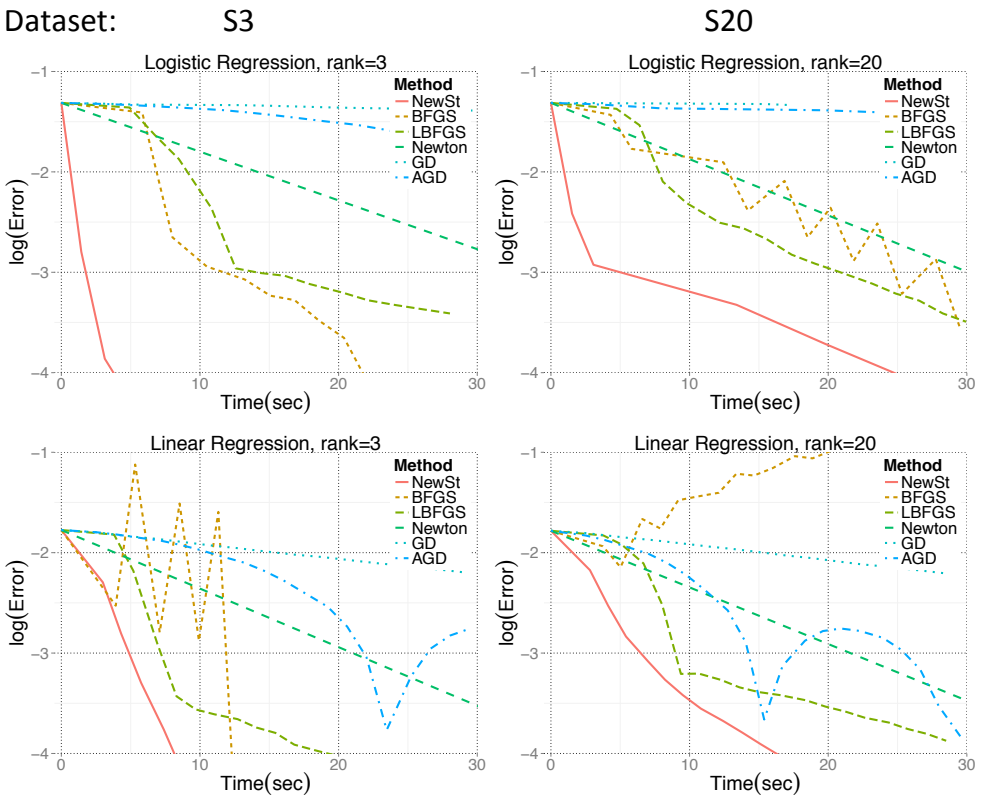

Figure 2: Performance of various optimization methods on two different simulated data sets. Red straight line represents the Newton-Stein method (NewSt). y and x axes denote $\log_{10}(\|\hat{\beta}^t - \beta_*\|_2)$ and time elapsed in seconds, respectively.

- *Gradient Descent* (GD) update is proportional to the negative of the full gradient evaluated at the current iterate. Under smoothness assumptions, GD achieves a locally linear convergence rate, with $\mathcal{O}(np)$ per-iteration cost.
- *Accelerated Gradient Descent* (AGD) is proposed by Nesterov (Nesterov, 1983), which improves over the gradient descent by using a momentum term. Performance of AGD strongly depends of the smoothness of the function.

For all the algorithms, we use a constant step size that provides the fastest convergence. We use the Newton-Stein method with eigenvalue thresholding as described in Section 4.4. The parameters such as sub-sample size $|S|$, and rank r are selected by following the guidelines described in Section 4.4. The rank threshold r (which is an input to the eigenvalue thresholding) is specified at the title of each plot.

5.1 Simulations With Synthetic Data Sets

Synthetic data sets, S3, S10, and S20 are generated through a multivariate Gaussian distribution where the covariance matrix follows r-spiked model, i.e., $r = 3$ for S3 and $r = 20$

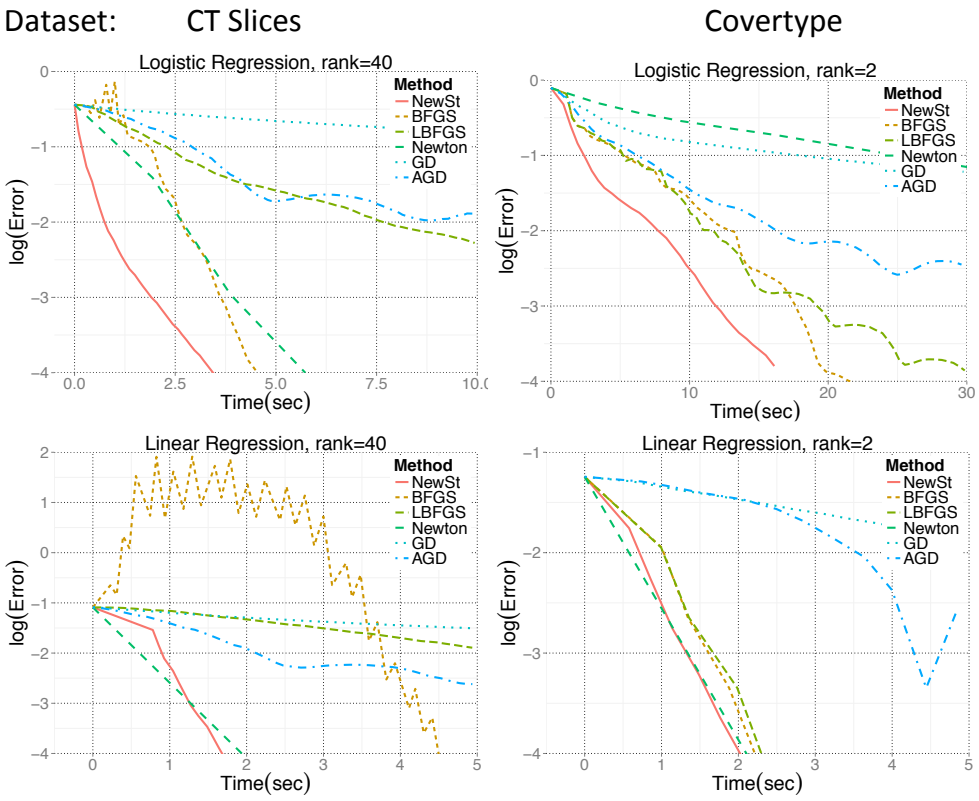

Figure 3: Performance of various optimization methods on two different real data sets obtained from Lichman, 2013. Red straight line represents the Newton-Stein method (NewSt). y and x axes denote $\log_{10}(\|\hat{\beta}^t - \beta_*\|_2)$ and time elapsed in seconds, respectively.

for S20. To generate the covariance matrix, we first generate a random orthogonal matrix, say $\mathbf{M}$. Next, we generate a diagonal matrix $\mathbf{\Lambda}$ that contains the eigenvalues, i.e., the first r diagonal entries are chosen to be large, and rest of them are equal to 1. Then, we let $\mathbf{\Sigma} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^T$. For dimensions of the data sets, see Table 2. We also emphasize that the data dimensions are chosen so that Newton's method still does well.

The simulation results are summarized in Figure 2. Further details regarding the experiments can be found in Table 1. We observe that Newton-Stein method (NewSt) provides a significant improvement over the classical techniques.

Observe that the convergence rate of NewSt has a clear phase transition point in the top left plot in Figure 2. As argued earlier, this point depends on various factors including sub-sampling size $|S|$ and data dimensions n, p, the rank threshold r and structure of the covariance matrix. The prediction of the phase transition point is an interesting line of research. However, our convergence guarantees are conservative and we believe that they cannot be used for this purpose.

5.2 Experiments With Real Data Sets

We experimented on two real data sets where the data sets are downloaded from UCI repository (Lichman, 2013). Both data sets satisfy $n \gg p$, but we highlight the difference between the proportions of dimensions n/p. See Table 2 for details.

We observe that Newton-Stein method performs better than classical methods on real data sets as well. More specifically, the methods that come closer to NewSt is Newton's method for moderate n and p and BFGS when n is large.

The optimal step-size for Newton-Stein method will typically be larger than 1 which is mainly due to eigenvalue thresholding operation. This feature is desirable if one is able to obtain a large step-size that provides convergence. In such cases, the convergence is likely to be faster, yet more unstable compared to the smaller step size choices. We observed that similar to other second order algorithms, Newton-Stein method is also susceptible to the step size selection. If the data is not well-conditioned, and the sub-sample size is not sufficiently large, algorithm might have poor performance. This is mainly because the sub-sampling operation is performed only once at the beginning. Therefore, it might be good in practice to sub-sample once in every few iterations.

Data set	S3				S20			
Type	LR		LS		LR		LS	
Method	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter
NewSt	10.637	2	8.763	4	23.158	4	16.475	10
BFGS	22.885	8	13.149	6	40.258	17	54.294	37
LBFGS	46.763	19	19.952	11	51.888	26	33.107	20
Newton	55.328	2	38.150	1	47.955	2	39.328	1
GD	865.119	493	155.155	100	1204.01	245	145.987	100
AGD	169.473	82	65.396	42	182.031	83	56.257	38

Data set	CT Slices				Covertype			
Type	LR		LS		LR		LS	
Method	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter
NewSt	4.191	32	1.799	11	16.113	31	2.080	5
BFGS	4.638	35	4.525	37	21.916	48	2.238	3
LBFGS	26.838	217	22.679	180	30.765	69	2.321	3
Newton	5.730	3	1.937	1	122.158	40	2.164	1
GD	96.142	1156	61.526	721	194.473	446	22.738	60
AGD	96.142	880	45.864	518	80.874	186	32.563	77

Table 1: Details of the experiments presented in Figures 2 and 3.

Data set	n	p	Reference, UCI repo (Lichman, 2013)
CT slices	53500	386	Graf et al., 2011
Covertype	581012	54	Blackard and Dean, 1999
S3	500000	300	3-spiked model, (Donoho et al., 2013)
S10	500000	300	10-spiked model, (Donoho et al., 2013)
S20	500000	300	20-spiked model, (Donoho et al., 2013)

Table 2: Data sets used in the experiments.

5.3 Analysis of Number of Iterations

We provide additional plots to better understand the convergence behavior of the algorithms. Plots in Figure 4 show the decrease in $\log_{10}(\|\hat{\beta}^t - \beta_0\|_2)$ error over iterations (instead of time elapsed).

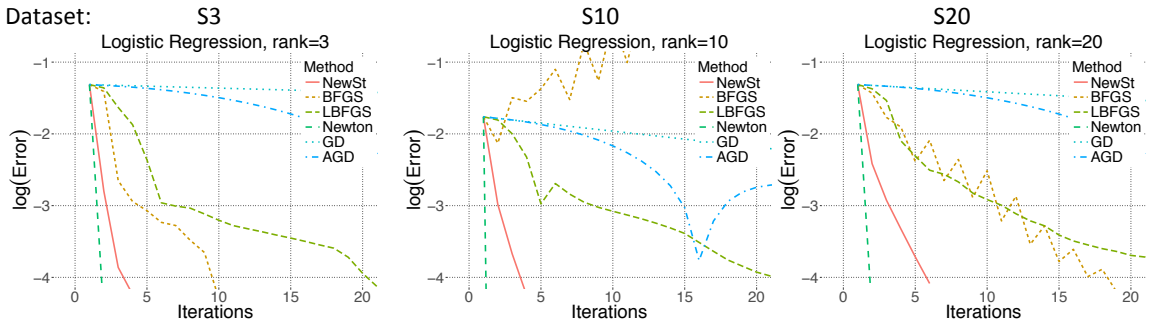

Figure 4: Figure shows the convergence behavior over the number of iterations. y and x axes denote $\log_{10}(\|\hat{\beta}^t - \beta_*\|_2)$ and the number iterations, respectively.

We observe from the plots that Newton's method enjoys the fastest convergence rate as expected. The one that is closest to Newton's method is the Newton-Stein method. This is because the Hessian estimator used by Newton-Stein method better approximates the true Hessian as opposed to Quasi-Newton methods. We emphasize that x axes in Figure 4 denote the number of iterations whereas in figures shown previously in this section x axes were the time elapsed.

6. Discussion

In this paper, we proposed an efficient algorithm for training GLMs. We call our algorithm Newton-Stein method (NewSt) as it takes a Newton-type step at each iteration relying on a Stein-type lemma. The algorithm requires a one time $\mathcal{O}(|S|p^2)$ cost to estimate the covariance structure and $\mathcal{O}(np)$ per-iteration cost to form the update equations. We observe that the convergence of Newton-Stein method has a phase transition from quadratic rate to linear rate. This observation is justified theoretically along with several other guarantees for the bounded as well as the sub-Gaussian covariates such as per-step convergence bounds, conditions for local rates and global convergence with line search, etc. Parameter selection guidelines of Newton-Stein method are based on our theoretical results. Our experiments show that Newton-Stein method provides significant improvement over the classical optimization methods.

Relaxing some of the theoretical constraints is an interesting line of research. In particular, strong assumptions on the cumulant generating functions might be loosened. Another interesting direction is to determine when the phase transition point occurs, which would provide a better understanding of the effects of sub-sampling and eigenvalue thresholding.

Acknowledgments

The author is grateful to Mohsen Bayati and Andrea Montanari for stimulating conversations on the topic of this work. The author would like to thank Bhaswar B. Bhattacharya and Qingyuan Zhao for carefully reading this article and providing valuable feedback.

Appendix A. Preliminary Concentration Inequalities

In this section, we provide several concentration bounds that will be useful throughout the proofs. We start by defining a special class of random variables.

Definition 10 (Sub-Gaussian) *A random variable $x \in \mathbb{R}$ is called* sub-Gaussian *if it satisfies*

$$\mathbb{E}[|x|^m]^{1/m} \leq K\sqrt{m}, \qquad m \geq 1,$$

for some finite constant K. The smallest such K is the sub-Gaussian norm of x and it is denoted by $\|x\|_{\psi_2}$. Similarly, a random vector $y \in \mathbb{R}^p$ is called sub-Gaussian if there exists a constant $K' > 0$ such that

$$\sup_{v \in S^{p-1}} \|\langle y, v\rangle\|_{\psi_2} \leq K'.$$

Definition 11 (Sub-exponential) *A random variable $x \in \mathbb{R}$ is called* sub-exponential *if it satisfies*

$$\mathbb{E}[|x|^m]^{1/m} \leq Km, \qquad m \geq 1,$$

for some finite constant K. The smallest such K is the sub-exponential norm of x and it is denoted by $\|x\|_{\psi_1}$. Similarly, a random vector $y \in \mathbb{R}^p$ is called sub-exponential if there exists a constant $K' > 0$ such that

$$\sup_{v \in S^{p-1}} \|\langle y, v\rangle\|_{\psi_1} \leq K'.$$

We state the following Lemmas from Vershynin, 2010 for the convenience of the reader (i.e., See Theorem 5.39 and the following remark for sub-Gaussian distributions, and Theorem 5.44 for distributions with arbitrary support):

Lemma 12 (Vershynin, 2010) *Let S be an index set and $x_i \in \mathbb{R}^p$ for $i \in S$ be i.i.d. sub-Gaussian random vectors with*

$$\mathbb{E}[x_i] = 0, \qquad \mathbb{E}[x_i x_i^T] = \mathbf{\Sigma}, \qquad \|x_i\|_{\psi_2} \leq K.$$

There exists constants c, C depending only on the sub-Gaussian norm K such that with probability $1 - 2e^{-ct^2}$,

$$\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 \leq \max\left(\delta, \delta^2\right) \quad \text{where} \quad \delta = C\sqrt{\frac{p}{|S|}} + \frac{t}{\sqrt{|S|}}.$$

Remark 13 *We are interested in the case where $\delta < 1$, hence the right hand side becomes $\max\left(\delta, \delta^2\right) = \delta$. In most cases, we will simply let $t = \sqrt{p}$ and obtain a bound of order $\sqrt{p/|S|}$ on the right hand side. For this, we need $|S| = \mathcal{O}(C^2 p)$ which is a reasonable assumption in the regime we consider.*

The following lemma is an analogue of Lemma 12 for covariates sampled from arbitrary distributions with bounded support.

Lemma 14 (Vershynin, 2010) *Let S be an index set and $x_i \in \mathbb{R}^p$ for $i \in S$ be i.i.d. random vectors with*

$$\mathbb{E}[x_i] = 0, \qquad \mathbb{E}\left[x_i x_i^T\right] = \mathbf{\Sigma}, \qquad \|x_i\|_2 \leq \sqrt{K} \ \ a.s.$$

Then, for some absolute constant c, with probability $1 - pe^{-ct^2}$, we have

$$\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 \leq \max\left(\|\mathbf{\Sigma}\|_2^{1/2}\delta, \delta^2\right) \quad \text{where} \quad \delta = t\sqrt{\frac{K}{|S|}}.$$

Remark 15 *We will choose $t = \sqrt{3\log(p)/c}$ which will provide us with a probability of $1 - 1/p^2$. Therefore, if the sample size is sufficiently large, i.e.,*

$$|S| \geq \frac{3K\log(p)}{c\|\mathbf{\Sigma}\|_2} = \mathcal{O}(K\log(p)/\|\mathbf{\Sigma}\|_2),$$

we can estimate the true covariance matrix quite well for arbitrary distributions with bounded support. In particular, with probability $1 - 1/p^2$, we obtain

$$\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 \leq c'\sqrt{\frac{\log(p)}{|S|}},$$

where $c' = \sqrt{3K\|\mathbf{\Sigma}\|_2/c}$.

In the following, we will focus on empirical processes and obtain uniform bounds for proposed Hessian approximation. To that extent, we provide a few basic definitions which will be useful later in the proofs. For a more detailed discussion on the machinery used throughout the next section, we refer reader to Van der Vaart, 2000.

Definition 16 *On a metric space (X, d), for $\epsilon > 0$, $T_\epsilon \subset X$ is called an ϵ-net over X if $\forall x \in X$, $\exists t \in T_\epsilon$ such that $d(x, t) \leq \epsilon$.*

In the following, we will use L_1 distance between two functions f and g, namely $d(f, g) = \int |f - g|$. Note that the same distance definition can be carried to random variables as they are simply real measurable functions. The integral takes the form of expectation.

Definition 17 *Given a function class $\mathcal{F}$, and any two functions l and u (not necessarily in $\mathcal{F}$), the bracket $[l, u]$ is the set of all $f \in \mathcal{F}$ such that $l \leq f \leq u$. A bracket satisfying $l \leq u$ and $\int |u - l| \leq \epsilon$ is called an ϵ-bracket in L_1. The bracketing number $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_1)$ is the minimum number of different ϵ-brackets needed to cover $\mathcal{F}$.*

The preliminary tools presented in this section will be utilized to obtain the concentration results in Section B.

Appendix B. Main Lemmas

B.1 Concentration of Covariates With Bounded Support

Lemma 18 *Let $x_i \in \mathbb{R}^p$, for $i = 1, 2, ..., n$, be i.i.d. random vectors supported on a ball of radius $\sqrt{K}$, with mean 0, and covariance matrix $\mathbf{\Sigma}$. Further, let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Then, for sufficiently large n, there exist constants c_1, c_2, c_3 such that*

$$\mathbb{P}\left(\sup_{\beta \in B_p(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)] \right| > c_1 \sqrt{\frac{p\log(n)}{n}}\right) \leq c_2 e^{-c_3 p},$$

where the constants depend only on the bound B.

Proof We start by using the Lipschitz property of the function f, i.e., $\forall \beta, \beta' \in B_p(R)$,

$$\begin{aligned} \left|f(\langle x, \beta\rangle) - f(\langle x, \beta'\rangle)\right| \leq & L\|x\|_2 \|\beta - \beta'\|_2, \\ \leq & L\sqrt{K}\|\beta - \beta'\|_2, \end{aligned}$$

where the first inequality follows from Cauchy-Schwartz. Now let T_Δ be a Δ-net over $B_p(R)$. Then $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$ such that the right hand side of the above inequality is smaller than $\Delta L\sqrt{K}$. Then, we can write

$$\left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)] \right| \leq \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta'\rangle) - \mathbb{E}[f(\langle x, \beta'\rangle)] \right| + 2\Delta L\sqrt{K}. \tag{14}$$

By choosing

$$\Delta = \frac{\epsilon}{4L\sqrt{K}},$$

and taking supremum over the corresponding β sets on both sides, we obtain the following inequality

$$\sup_{\beta \in B_n(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)] \right| \leq \max_{\beta \in \mathcal{T}_\Delta} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)] \right| + \frac{\epsilon}{2}.$$

Now, since we have $\|f\|_\infty \leq B$ and for a fixed β and $i = 1, 2, ..., n$, the random variables $f(\langle x_i, \beta\rangle)$ are i.i.d., by the Hoeffding's concentration inequality, we have

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)] \right| > \epsilon/2 \right) \leq 2\exp\left(-\frac{n\epsilon^2}{8B^2}\right).$$

Combining Equation 14 with the above result and a union bound, we easily obtain

$$\begin{aligned} &\mathbb{P}\left(\sup_{\beta \in B_n(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)] \right| > \epsilon \right) \\ &\leq \mathbb{P}\left(\max_{\beta \in \mathcal{T}_\Delta} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)] \right| > \epsilon/2 \right) \leq 2|\mathcal{T}_\Delta| \exp\left(-\frac{n\epsilon^2}{8B^2}\right), \end{aligned}$$

where $\Delta = \epsilon/4L\sqrt{K}$.

Next, we apply Lemma 33 and obtain that

$$|\mathcal{T}_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta}\right)^p = \left(\frac{R\sqrt{p}}{\epsilon/4L\sqrt{K}}\right)^p.$$

We require that the probability of the desired event is bounded by a quantity that attains an exponential decay with rate $\mathcal{O}(p)$. This can be attained if

$$\epsilon^2 \geq \frac{8B^2p}{n}\log\left(4eLR\sqrt{K}\sqrt{p}/\epsilon\right).$$

Assuming that n is sufficiently large, and using Lemma 34 with $a = 8B^2p/n$ and $b = 4eLR\sqrt{Kp}$, we obtain that ϵ should be

$$\epsilon = \sqrt{\frac{4B^2p}{n}\log\left(\frac{30L^2R^2Kn}{B^2}\right)} = \mathcal{O}\left(\sqrt{\frac{p\log(n)}{n}}\right).$$

When $n > 30L^2R^2K/B^2$, we obtain

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > 3B\sqrt{\frac{p\log(n)}{n}}\right) \leq 2e^{-p}.$$

■

In the following, we state similar bounds on functions of the following form

$$x \to f(\langle x,\beta\rangle)\langle x,v\rangle^2,$$

which appear in the summation that form the Hessian matrix.

Lemma 19 *Let $x_i \in \mathbb{R}^p$, for $i = 1,...,n$, be i.i.d. random vectors supported on a ball of radius $\sqrt{K}$, with mean 0, and covariance matrix $\mathbf{\Sigma}$. Also let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Then, for $v \in S^{p-1}$ and sufficiently large n, there exist constants c_1, c_2, c_3 such that*

$$\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > c_1\sqrt{\frac{p\log(n)}{n}}\right) \leq c_2e^{-c_3p},$$

where the constants depend only on the bound B and the radius $\sqrt{K}$.

Proof As in the proof of Lemma 18, we start by using the Lipschitz property of the function f, i.e., $\forall\beta,\beta' \in B_p(R)$,

$$\begin{aligned}\|f(\langle x,\beta\rangle)\langle x,v\rangle^2 - f(\langle x,\beta'\rangle)\langle x,v\rangle^2\|_2 \leq& L\|x\|_2^3\|\beta-\beta'\|_2,\\ \leq& LK^{1.5}\|\beta-\beta'\|_2.\end{aligned}$$

For a net T_Δ, $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$ such that right hand side of the above inequality is smaller than $\Delta L K^{1.5}$. Then, we can write

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta'\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta'\rangle)\langle x,v\rangle^2]\right| + 2\Delta L K^{1.5}. \tag{15}$$

This time, we choose

$$\Delta = \frac{\epsilon}{4LK^{1.5}},$$

and take the supremum over the corresponding feasible β-sets on both sides,

$$\begin{aligned}&\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| \\ &\leq \max_{\beta\in\mathcal{T}_\Delta}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| + \frac{\epsilon}{2}.\end{aligned}$$

Now, since we have $\|f\|_\infty \leq B$ and for fixed β and v, $i = 1, 2, ..., n$, $f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2$ are i.i.d. random variables. By the Hoeffding's concentration inequality, we write

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon/2\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8B^2K^2}\right).$$

Using Equation 15 and the above result combined with the union bound, we easily obtain

$$\begin{aligned}&\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon\right) \\ &\leq \mathbb{P}\left(\max_{\beta\in\mathcal{T}_\Delta}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon/2\right) \\ &\leq 2|\mathcal{T}_\Delta|\exp\left(-\frac{n\epsilon^2}{8B^2K^2}\right),\end{aligned}$$

where $\Delta = \epsilon/4LK^{1.5}$. Using Lemma 33, we have

$$|\mathcal{T}_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta}\right)^p = \left(\frac{R\sqrt{p}}{\epsilon/4LK^{1.5}}\right)^p.$$

As before, we require that the right hand side of above inequality gets a decay with rate $\mathcal{O}(p)$. Using Lemma 34 with $a = 8B^2K^2p/n$ and $b = 100LRK^{1.5}\sqrt{p}$, we obtain that ϵ should be

$$\epsilon = \sqrt{\frac{4B^2K^2p}{n}\log\left(\frac{50^2L^2R^2Kn}{B^2}\right)} = \mathcal{O}\left(\sqrt{\frac{p\log(n)}{n}}\right).$$

When $n > 50LRK^{1/2}/B$, we obtain

$$\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > 4BK\sqrt{\frac{p\log(n)}{n}}\right) \leq 2e^{-3.2p}.$$

The rate $-3.2p$ will be important later. ■

B.2 Concentration of Sub-Gaussian Covariates

In this section, we derive the analogues of the Lemmas 18 and 19 for sub-Gaussian covariates. Note that the Lemmas in this section are more general in the sense that they also cover the case where the covariates have bounded support. As a result, the resulting convergence coefficients are worse compared to the previous section.

Lemma 20 *Let $x_i \in \mathbb{R}^p$, for $i = 1, ..., n$, be i.i.d. sub-Gaussian random vectors with mean 0, covariance matrix $\mathbf{\Sigma}$ and sub-Gaussian norm K. Also let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Then, there exists absolute constants c_1, c_2, c_3 such that*

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > c_1\sqrt{\frac{p\log(n)}{n}}\right) \leq c_2 e^{-c_3 p},$$

where the constants depend only on the eigenvalues of $\mathbf{\Sigma}$, bound B and radius R and sub-Gaussian norm K.

Proof We start by defining the brackets of the form

$$\begin{aligned} l_\beta(x) =& f(\langle x,\beta\rangle) - \epsilon\frac{\|x\|_2}{4\mathbb{E}\left[\|x\|_2\right]}, \\ u_\beta(x) =& f(\langle x,\beta\rangle) + \epsilon\frac{\|x\|_2}{4\mathbb{E}\left[\|x\|_2\right]}. \end{aligned}$$

Observe that the size of bracket $[\ell_\beta, u_\beta]$ is $\epsilon/2$, i.e., $\mathbb{E}[u_\beta - \ell_\beta] = \epsilon/2$. Now let T_Δ be a Δ-net over $B_p(R)$ where we use $\Delta = \epsilon/(4L\mathbb{E}\left[\|x\|_2\right])$. Then $\forall\beta \in B_p(R)$, $\exists\beta' \in T_\Delta$ such that $f(\langle\cdot,\beta\rangle)$ falls into the bracket $[\ell_{\beta'}, u_{\beta'}]$. This can be seen by writing out the Lipschitz property of the function f. That is,

$$\begin{aligned} |f(\langle x,\beta\rangle) - f(\langle x,\beta'\rangle)| \leq& L\|x\|_2\|\beta - \beta'\|_2, \\ \leq& \Delta L\|x\|_2, \end{aligned}$$

where the first inequality follows from Cauchy-Schwartz. Therefore, we conclude that

$$\mathcal{N}_{[]}(\epsilon/2, \mathcal{F}, L_1) \leq |T_\Delta|$$

for the function class $\mathcal{F} = \{f(\langle \cdot, \beta \rangle) : \beta \in B_p(R)\}$. We further have $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$ such that

$$\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \leq \frac{1}{n}\sum_{i=1}^{n} u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)] + \frac{\epsilon}{2},$$
$$\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \geq \frac{1}{n}\sum_{i=1}^{n} l_{\beta'}(x_i) - \mathbb{E}[l_{\beta'}(x)] - \frac{\epsilon}{2}.$$

Using the above inequalities, we have, $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$

$$\left\{\left[\frac{1}{n}\sum_{i=1}^{n} u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)]\right] > \epsilon/2\right\} \cup \left\{\left[-\frac{1}{n}\sum_{i=1}^{n} l_{\beta'}(x_i) + \mathbb{E}[l_{\beta'}(x)]\right] > \epsilon/2\right\} \supset$$
$$\left\{\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)]\right| > \epsilon\right\}.$$

By the union bound, we obtain

$$\mathbb{P}\left(\max_{\beta \in \mathcal{T}_\Delta}\left[\frac{1}{n}\sum_{i=1}^{n} u_\beta(x_i) - \mathbb{E}[u_\beta(x)]\right] > \epsilon/2\right) + \mathbb{P}\left(\max_{\beta \in \mathcal{T}_\Delta}\left[-\frac{1}{n}\sum_{i=1}^{n} l_\beta(x_i) + \mathbb{E}[l_\beta(x)]\right] > \epsilon/2\right)$$
$$\geq \mathbb{P}\left(\sup_{\beta \in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)]\right| > \epsilon\right). \tag{16}$$

In order to complete the proof, we need concentration inequalities for u_β and l_β. We state the following lemma.

Lemma 21 *There exists a constant C depending on the eigenvalues of $\mathbf{\Sigma}$ and B such that, for each $\beta \in B_p(R)$ and for some $0 < \epsilon < 1$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} u_\beta(x_i) - \mathbb{E}[u_\beta(x)]\right| > \epsilon/2\right) \leq 2e^{-Cn\epsilon^2},$$
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} l_\beta(x_i) - \mathbb{E}[l_\beta(x)]\right| > \epsilon/2\right) \leq 2e^{-Cn\epsilon^2},$$

where

$$C = \frac{c}{\left(B + \frac{\sqrt{2}K}{4\mu/\sqrt{p}}\right)^2}$$

for an absolute constant c.

Remark 22 *Note that $\mu = \mathbb{E}[\|x\|_2] = \mathcal{O}(\sqrt{p})$ and hence $\mu/\sqrt{p} = \mathcal{O}(1)$.*

Proof By the relation between sub-Gaussian and sub-exponential norms, we have

$$\begin{aligned} \|\|x\|_2\|_{\psi_2}^2 \leq \|\|x\|_2^2\|_{\psi_1} \leq & \sum_{i=1}^{p} \|x_i^2\|_{\psi_1}, \\ \leq & 2\sum_{i=1}^{p} \|x_i\|_{\psi_2}^2, \\ \leq & 2K^2 p. \end{aligned} \tag{17}$$

Therefore $\|x\|_2 - \mathbb{E}[\|x\|_2]$ is a centered sub-Gaussian random variable with sub-Gaussian norm bounded above by $2K\sqrt{2p}$. We have,

$$\mathbb{E}[\|x\|_2] = \mu.$$

Note that μ is actually of order $\sqrt{p}$. Assuming that the left hand side of the above equality is equal to $\sqrt{p}K'$ for some constant $K' > 0$, we can conclude that the random variable $u_\beta(x) = f(\langle x, \beta\rangle) + \epsilon\frac{\|x\|_2}{4\mathbb{E}[\|x\|_2]}$ is also sub-Gaussian with

$$\begin{aligned} \|u_\beta(x)\|_{\psi_2} \leq & B + \frac{\epsilon}{4\mathbb{E}\left[\|x\|_2\right]}\|\|x\|_2\|_{\psi_2} \\ & \leq B + \frac{\epsilon}{4\sqrt{p}K'}K\sqrt{2p} \\ & \leq B + C' \end{aligned}$$

where $C' = \sqrt{2}K/4K'$ is a constant and we also assumed $\epsilon < 1$. Now, define the function

$$g_\beta(x) = u_\beta(x) - \mathbb{E}[u_\beta(x)].$$

Note that $g_\beta(x)$ is a centered sub-Gaussian random variable with sub-Gaussian norm

$$\|g_\beta(x)\|_{\psi_2} \leq 2B + 2C'.$$

Then, by the Hoeffding-type inequality for the sub-Gaussian random variables, we obtain

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} g_\beta(x_i)\right| > \epsilon/2\right) \leq 2e^{-cn\epsilon^2/(B+C')^2}$$

where c is an absolute constant. The same argument also holds for $l_\beta(x)$. ■

Using the above lemma with the union bound over the set T_Δ, we can write

$$\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)]\right| > \epsilon\right) \leq 4|T_\Delta|e^{-Cn\epsilon^2}.$$

Since we can also write, by Lemma 33

$$|T_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta}\right)^p \leq \left(\frac{4RL\mathbb{E}[\|x\|_2]\sqrt{p}}{\epsilon}\right)^p, \\ \leq \left(\frac{4\sqrt{2}RLKp}{\epsilon}\right)^p,$$

and we observe that, for the constant $c' = 4\sqrt{2}RLK$,

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > \epsilon\right) \leq 4\left(\frac{4\sqrt{2}RLKp}{\epsilon}\right)^p e^{-Cn\epsilon^2}, \\ = 4\exp\left\{p\log(c'p/\epsilon) - Cn\epsilon^2\right\}.$$

We will obtain an exponential decay of order p on the right hand side. For some constant h depending on n and p, if we choose $\epsilon = hp$, we need

$$h^2 \geq \frac{1}{Cnp}\log(c'/h).$$

By the Lemma 34, choosing $h^2 = \log(2c'^2Cnp)/(2Cnp)$, we satisfy the above requirement. Note that for n large enough, the condition of the lemma is easily satisfied. Hence, for

$$\epsilon^2 = \frac{p\log(2c'^2Cnp)}{2Cn} = \mathcal{O}\left(\frac{p\log(n)}{n}\right),$$

we obtain that there exists constants c_1, c_2, c_3 such that

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > c_1\sqrt{\frac{p\log(n)}{n}}\right) \leq c_2 e^{-c_3 p},$$

where

$$c_1 = \frac{3\left(B + \frac{\sqrt{2}K}{4\sqrt{\mathrm{Tr}(\mathbf{\Sigma})/p - 16K^2}}\right)^2}{2c}, \\ c_2 = 4, \\ c_3 = \frac{1}{2}\log(7) \leq \frac{1}{2}\log(\log(64R^2L^2K^2C) + 6\log(p)).$$

when $p > e$ and $64R^2L^2K^2C > e$.

■

In the following, we state the concentration results on the unbounded functions of the form

$$x \to f(\langle x,\beta\rangle)\langle x,v\rangle^2.$$

Functions of this type form the summands of the Hessian matrix in GLMs.

Lemma 23 *Let x_i, for $i = 1, ..., n$, be i.i.d sub-Gaussian random variables with mean 0, covariance matrix $\mathbf{\Sigma}$ and sub-Gaussian norm K. Also let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Further, let $v \in \mathbb{R}^p$ such that $\|v\|_2 = 1$. Then, for n, p sufficiently large satisfying*

$$n^{0.2}/\log(n) \gtrsim p,$$

there exist constants c_1, c_2 depending on L, B, R and the eigenvalues of $\mathbf{\Sigma}$ such that, we have

$$\mathbb{P}\left(\sup_{\beta \in B_p(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \beta\rangle)\langle x, v\rangle^2] \right| > c_1\sqrt{\frac{p}{n^{0.2}}\log(n)}\right) \leq c_2 e^{-p}.$$

Proof We define the brackets of the form

$$\begin{aligned} l_\beta(x) =& f(\langle x, \beta\rangle)\langle x, v\rangle^2 - \epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]}, \\ u_\beta(x) =& f(\langle x, \beta\rangle)\langle x, v\rangle^2 + \epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]}, \end{aligned} \tag{18}$$

and we observe that the bracket $[\ell_\beta, u_\beta]$ has size $\epsilon/2$ in L_1, that is,

$$\mathbb{E}\left[|u_\beta(x) - l_\beta(x)|\right] = \epsilon/2.$$

Next, for the following constant

$$\Delta = \frac{\epsilon}{4L\mathbb{E}\left[\|x\|_2^3\right]},$$

we define a Δ-net over $B_p(R)$ and call it $\mathcal{T}_\Delta$. Then, $\forall \beta \in B_p(R)$, $\exists \beta' \in \mathcal{T}_\Delta$ such that $f(\langle \cdot, \beta\rangle)\langle \cdot, v\rangle^2$ belongs to the bracket $[\ell_{\beta'}, u_{\beta'}]$. This can be seen by writing the Lipschitz continuity of the function f, i.e.,

$$\begin{aligned} \left|f(\langle x, \beta\rangle)\langle x, v\rangle^2 - f(\langle x, \beta'\rangle)\langle x, v\rangle^2\right| =& \langle x, v\rangle^2 \left|\left\{f(\langle x, \beta\rangle) - f(\langle x, \beta'\rangle)\right\}\right|, \\ \leq& L\|x\|_2^2 \ \|v\|_2^2 \ \left|\langle x, \beta - \beta'\rangle\right|, \\ \leq& L\|x\|_2^3\|\beta - \beta'\|_2, \\ \leq& \Delta L\|x\|_2^3, \end{aligned}$$

where we used Cauchy-Schwartz to obtain the above inequalities. Hence, we may conclude that for the bracketing functions given in Equation 18, the corresponding bracketing number of the function class

$$\mathcal{F} = \{f(\langle \cdot, \beta\rangle)\langle \cdot, v\rangle^2 : \beta \in B_p(R)\}$$

is bounded above by the covering number of the ball of radius R for the given scale $\Delta = \epsilon/(4L\mathbb{E}\left[\|x\|_2^3\right])$, i.e.,

$$\mathcal{N}_{[]}(\epsilon/2, \mathcal{F}, L_1) \leq |\mathcal{T}_\Delta|.$$

Next, we will upper bound the target probability using the bracketing functions u_β, l_β. We have $\forall \beta \in B_p(R)$, $\exists \beta' \in \mathcal{T}_\Delta$ such that

$$\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2] \le \frac{1}{n}\sum_{i=1}^n u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)] + \frac{\epsilon}{2},$$
$$\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2] \ge \frac{1}{n}\sum_{i=1}^n l_{\beta'}(x_i) - \mathbb{E}[l_{\beta'}(x)] - \frac{\epsilon}{2}.$$

Using the above inequalities, $\forall \beta \in B_p(R)$, $\exists \beta' \in \mathcal{T}_\Delta$, we can write

$$\left\{\left[\frac{1}{n}\sum_{i=1}^n u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)]\right] > \epsilon/2\right\} \cup \left\{\left[-\frac{1}{n}\sum_{i=1}^n l_{\beta'}(x_i) + \mathbb{E}[l_{\beta'}(x)]\right] > \epsilon/2\right\} \supset$$
$$\left\{\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon\right\}.$$

Hence, by the union bound, we obtain

$$\mathbb{P}\left(\max_{\beta\in\mathcal{T}_\Delta}\left[\frac{1}{n}\sum_{i=1}^n u_\beta(x_i) - \mathbb{E}[u_\beta(x)]\right] > \epsilon/2\right) + \mathbb{P}\left(\max_{\beta\in\mathcal{T}_\Delta}\left[-\frac{1}{n}\sum_{i=1}^n l_\beta(x_i) + \mathbb{E}[l_\beta(x)]\right] > \epsilon/2\right)$$
$$\ge \mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon\right). \tag{19}$$

In order to complete the proof, we need one-sided concentration inequalities for u_β and l_β. Handling these functions is somewhat tedious since $\|x\|_2^3$ terms do not concentrate nicely. We state the following lemma.

Lemma 24 *For given $\alpha, \epsilon > 0$, and n sufficiently large such that, $\nu(n^\alpha, p, \epsilon, B, K, \mathbf{\Sigma}) < \epsilon/4$ where*

$$\nu(n^\alpha, p, \epsilon, B, K, \mathbf{\Sigma}) =: 2\left(n^\alpha + \frac{6BK^2p}{c}\right)\exp\left(-c\frac{n^\alpha}{6BK^2p}\right) + 2\Bigg\{n^\alpha + \frac{3K^2p}{c\mathrm{Tr}(\mathbf{\Sigma})}n^{\alpha/3}\epsilon^{2/3}$$
$$+ \frac{3K^4p^2}{c^2\mathrm{Tr}(\mathbf{\Sigma})^2}\epsilon^{4/3}n^{-\alpha/3}\Bigg\}\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right).$$

Then, there exists constants c', c'', c''' depending on the eigenvalues of $\mathbf{\Sigma}$, B and K such that $\forall \beta$, we have,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n u_\beta(x_i) - \mathbb{E}[u_\beta(x)] > \epsilon/2\right) \le 2\exp\left(-c'n^\alpha/p\right)$$
$$+ 2\exp\left(-c''n^{2\alpha/3}\epsilon^{-2/3}\right) + \exp\left(-c'''n^{1-2\alpha}\epsilon^2\right),$$

and

$$\mathbb{P}\left(-\frac{1}{n}\sum_{i=1}^n l_\beta(x_i) + \mathbb{E}[l_\beta(x)] > \epsilon/2\right) \le 2\exp\left(-c'n^\alpha/p\right) +$$
$$2\exp\left(-c''n^{2\alpha/3}\epsilon^{-2/3}\right) + \exp\left(-c'''n^{1-2\alpha}\epsilon^2\right).$$

Proof For the sake of simplicity, we define the functions

$$\tilde{u}_\beta(w) = u_\beta(w) - \mathbb{E}[u_\beta(x)],$$
$$\tilde{l}_\beta(w) = l_\beta(w) - \mathbb{E}[l_\beta(x)].$$

We will derive the result for the upper bracket, $\tilde{u}$, and skip the proof for the lower bracket $\tilde{l}$ as it follows from the same steps. We write,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i) > \epsilon/2\right) \leq & \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i) > \epsilon/2,\ \max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| < n^\alpha\right) \\ & + \mathbb{P}\left(\max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| \geq n^\alpha\right). \end{aligned} \tag{20}$$

We need to bound the right hand side of the above equation. For the second term, since $\tilde{u}_\beta(x_i)$'s are i.i.d. centered random variables, we have

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| \geq n^\alpha\right) =& 1 - \mathbb{P}\left(\max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| < n^\alpha\right), \\ =& 1 - \mathbb{P}\left(|\tilde{u}_\beta(x)| < n^\alpha\right)^n, \\ =& 1 - \left(1 - \mathbb{P}\left(|\tilde{u}_\beta(x)| \geq n^\alpha\right)\right)^n, \\ \leq& n\mathbb{P}\left(|\tilde{u}_\beta(x)| \geq n^\alpha\right).\end{aligned}$$

Also, note that

$$\begin{aligned}|\tilde{u}_\beta(x)| \leq& B\|x\|_2^2 + \epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} + \mathbb{E}[u_\beta(x)], \\ \leq& B\|x\|_2^2 + \epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} + B\lambda_{\max}(\mathbf{\Sigma}) + \epsilon/4.\end{aligned}$$

Therefore, if $t > 3B\lambda_{\max}(\mathbf{\Sigma})$ and for ϵ small, we can write

$$\{|\tilde{u}_\beta(x)| > t\} \subset \left\{B\|x\|_2^2 > t/3\right\} \cup \left\{\epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} > t/3\right\}. \tag{21}$$

Since x is a sub-Gaussian random variable with $\|x\|_{\psi_2} = K$, we have

$$K = \sup_{w\in S^{p-1}}\|\langle w, x\rangle\|_{\psi_2} = \|x\|_{\psi_2}.$$

Using this and the relation between sub-Gaussian and sub-exponential norms as in Equation 17, we have $\|\|x\|_2\|_{\psi_2}^2 \leq 2K^2p$. This provides the following tail bound for $\|x\|_2$,

$$\mathbb{P}\left(\|x\|_2 > s\right) \leq 2\exp\left(-\frac{cs^2}{2pK^2}\right), \tag{22}$$

where c is an absolute constant. Using the above tail bound, we can write,

$$\mathbb{P}\left(\|x\|_2^2 > \frac{1}{3B}t\right) \leq 2\exp\left(-c\frac{t}{6BK^2p}\right).$$

For the next term in Equation 21, we need a lower bound for $\mathbb{E}\left[\|x\|_2^3\right]$. We use a modified version of the Hölder's inequality and obtain

$$\mathbb{E}\left[\|x\|_2^3\right] \geq \mathbb{E}\left[\|x\|_2^2\right]^{3/2} = \mathrm{Tr}(\mathbf{\Sigma})^{3/2}.$$

Using the above inequality, we can write

$$\begin{aligned}
\mathbb{P}\left(\epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} > t/3\right) \leq& \mathbb{P}\left(\|x\|_2^3 > \frac{4}{3\epsilon}\mathrm{Tr}(\mathbf{\Sigma})^{3/2}t\right), \\
=& \mathbb{P}\left(\|x\|_2 > \left(\frac{4t}{3\epsilon}\right)^{1/3}\mathrm{Tr}(\mathbf{\Sigma})^{1/2}\right), \\
\leq& 2\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right),
\end{aligned}$$

where c is the same absolute constant as in Equation 22.

Now for $\alpha > 0$ such that $t = n^\alpha > 3B\lambda_{\max}(\mathbf{\Sigma})$ (we will justify this assumption for a particular choice of α later), we combine the above results,

$$\mathbb{P}\left(|\tilde{u}_\beta(x)| > t\right) \leq 2\exp\left(-c\frac{t}{6BK^2p}\right) + 2\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right). \tag{23}$$

Next, we focus on the first term in Equation 20. Let $\mu = \mathbb{E}[\tilde{u}_\beta(x)\mathbb{I}_{\{|\tilde{u}_\beta(x)|<n^\alpha\}}]$, and write

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \tilde{u}_\beta(x_i) > \frac{\epsilon}{2};\ \max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| < n^\alpha\right) \leq& \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \tilde{u}_\beta(x_i)\mathbb{I}_{\{|\tilde{u}_\beta(x_i)|<n^\alpha\}} > \frac{\epsilon}{2}\right), \\
=& \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \tilde{u}_\beta(x_i)\mathbb{I}_{\{|\tilde{u}_\beta(x_i)|<n^\alpha\}} - \mu > \frac{\epsilon}{2} - \mu\right) \\
\leq& \exp\left\{-\frac{n^{1-2\alpha}}{2}\left(\frac{\epsilon}{2} - \mu\right)^2\right\},
\end{aligned}$$

where we used the Hoeffding's concentration inequality for the bounded random variables. Further, note that

$$0 = \mathbb{E}[\tilde{u}_\beta(x)] = \mu + \mathbb{E}\left[\tilde{u}_\beta(x)\mathbb{I}_{\{|\tilde{u}_\beta(x)|>n^\alpha\}}\right].$$

By Lemma 30, we can write

$$|\mu| = \left|\mathbb{E}\left[\tilde{u}_\beta(x)\mathbb{I}_{\{|\tilde{u}_\beta(x)|>n^\alpha\}}\right]\right| \leq n^\alpha\mathbb{P}(|\tilde{u}_\beta(x)| > n^\alpha) + \int_{n^\alpha}^\infty \mathbb{P}(|\tilde{u}_\beta(x)| > t)dt.$$

The first term on the right hand side can be easily bounded by using Equation 23, i.e.,

$$n^\alpha \mathbb{P}(|\tilde{u}_\beta(x)| > n^\alpha) \leq 2n^\alpha \exp\left(-c\frac{n^\alpha}{6BK^2p}\right) + 2n^\alpha \exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right).$$

For the second term, using Equation 23 once again, we obtain

$$\begin{aligned}\int_{n^\alpha}^\infty \mathbb{P}(|\tilde{u}_\beta(x)| > t)dt \leq& 2\int_{n^\alpha}^\infty \exp\left(-c\frac{t}{6BK^2p}\right)dt + 2\int_{n^\alpha}^\infty \exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right)dt,\\ =&\frac{12BK^2p}{c}\exp\left(-c\frac{n^\alpha}{6BK^2p}\right) + 2\int_{n^\alpha}^\infty \exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right)dt.\end{aligned}$$

Next, we apply Lemma 31 to bound the second term on the right hand side. That is, we have

$$\begin{aligned}&\int_{n^\alpha}^\infty \exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right)dt\\ &\leq \left\{\frac{3K^2p}{c\mathrm{Tr}(\mathbf{\Sigma})}n^{\alpha/3}\epsilon^{2/3} + \frac{3K^4p^2}{c^2\mathrm{Tr}(\mathbf{\Sigma})^2}\epsilon^{4/3}n^{-\alpha/3}\right\}\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right).\end{aligned}$$

Combining the above results, we can write

$$\begin{aligned}|\mu| \leq& 2\left(n^\alpha + \frac{6BK^2p}{c}\right)\exp\left(-c\frac{n^\alpha}{6BK^2p}\right)\\ &+ 2\left\{n^\alpha + \frac{3K^2p}{c\mathrm{Tr}(\mathbf{\Sigma})}n^{\alpha/3}\epsilon^{2/3} + \frac{3K^4p^2}{c^2\mathrm{Tr}(\mathbf{\Sigma})^2}\epsilon^{4/3}n^{-\alpha/3}\right\}\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right),\\ =:&\nu(n^\alpha, p, \epsilon, B, K, \mathbf{\Sigma}).\end{aligned}$$

Notice that, the upper bound on $|\mu|$, namely $\nu(n^\alpha, p, \epsilon, B, K, \mathbf{\Sigma})$, is close to 0 when n is large. This is because of exponentially decaying functions that dominates the other terms. We assume that n is sufficiently large that the upper bound for $|\mu|$ is less than $\epsilon/4$. For the value of α, we will choose $\alpha = 0.4$ later in the proof.

Applying this bounds in Equation 20, we obtain

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \tilde{u}_\beta(x_i) > \epsilon/2\right) \leq& 2\exp\left(-c\frac{n^\alpha}{6BK^2p}\right)\\ &+ 2\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right) + \exp\left(-\frac{n^{1-2\alpha}}{32}\epsilon^2\right),\\ =&2\exp\left(-c'n^\alpha/p\right) + 2\exp\left(-c''n^{2\alpha/3}\epsilon^{-2/3}\right) + \exp\left(-c'''n^{1-2\alpha}\epsilon^2\right),\end{aligned}$$

where

$$\begin{aligned} c' =& \frac{c}{6BK^2}, \\ c'' =& \frac{c\mathrm{Tr}(\mathbf{\Sigma})/p}{2K^2} \geq \frac{c\lambda_{\min}(\mathbf{\Sigma})}{2K^2}, \\ c''' =& \frac{1}{32}. \end{aligned}$$

Hence, the proof is completed for the upper bracket.

The proof for the lower brackets $l_\beta(x)$ follows from exactly the same steps and omitted here. ■

Applying the above lemma on Equation 19, for $\alpha > 0$, we obtain

$$\begin{aligned} &\mathbb{P}\left(\sup_{\beta \in B_n(R)} \left| \frac{1}{n} \sum_{i=1}^n f(\langle x_i, \beta \rangle) \langle x_i, v \rangle^2 - \mathbb{E}[f(\langle x, \beta \rangle) \langle x, v \rangle^2] \right| > \epsilon \right) \\ &\leq 4|T_\Delta| \exp\left(-c' n^\alpha / p\right) + 4|T_\Delta| \exp\left(-c'' n^{2\alpha/3} \epsilon^{-2/3}\right) + 2|T_\Delta| \exp\left(-c''' n^{1-2\alpha} \epsilon^2\right). \end{aligned} \tag{24}$$

Observe that we can write, by Lemma 33

$$|T_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta} \right)^p = \left(\frac{4\sqrt{p} RL \mathbb{E}[\|x\|_2^3]}{\epsilon} \right)^p.$$

Also, recall that $\|x\|_2$ was a sub-Gaussian random variable with $\|\|x\|_2\|_{\psi_2} \leq K\sqrt{2p}$. Using the definition of sub-Gaussian norm, we have

$$\frac{1}{\sqrt{3}} \mathbb{E}[\|x\|_2^3]^{1/3} \leq \|\|x\|_2\|_{\psi_2} \leq \sqrt{2p}\, K, \implies \mathbb{E}[\|x\|_2^3] \leq 15 K^3 p^{3/2}.$$

Therefore, we have $\mathbb{E}[\|x\|_2^3] = \mathcal{O}(p^{3/2})$ (recall that we had a lower bound of the same order). We define a constant K', and as ϵ is small, we have

$$|T_\Delta| \leq \left(\frac{60 RLK^3 p^2}{\epsilon} \right)^p = \left(\frac{K' p^2}{\epsilon} \right)^p,$$

where we let $K' = 60RLK^3$. We will show that each term on the right hand side of Equation 24 decays exponentially with a rate of order p. For the first term, for $s > 0$, we write

$$\begin{aligned} |T_\Delta| \exp\left(-c' n^\alpha / p\right) =& \exp\left(-c' n^\alpha / p + p \log(K') + 2p \log(p) + p \log(\epsilon^{-1})\right), \\ \leq& \exp\left(-c' n^\alpha / p + 2p \log(K' p / \epsilon)\right). \end{aligned} \tag{25}$$

Similarly for the second and third terms, we write

$$\begin{aligned} |T_\Delta| \exp\left(-c'' n^{2\alpha/3} \epsilon^{-2/3}\right) \leq& \exp\left(-c'' n^{2\alpha/3} \epsilon^{-2/3} + 2p \log(K' p / \epsilon)\right), \\ |T_\Delta| \exp\left(-c''' n^{1-2\alpha} \epsilon^2\right) \leq& \exp\left(-c''' n^{1-2\alpha} \epsilon^2 + 2p \log(K' p / \epsilon)\right). \end{aligned} \tag{26}$$

We will seek values for ϵ and α to obtain an exponential decay with rate p on the right sides of Equations 25 and 26. That is, we need

$$\begin{aligned} c' n^{\alpha}/p \geq & 2p \log(K'' p/\epsilon), \\ c'' n^{2\alpha/3} \geq & 2p \log(K'' p/\epsilon) \epsilon^{2/3}, \\ c''' n^{1-2\alpha} \epsilon^2 \geq & 2p \log(K'' p/\epsilon), \end{aligned} \tag{27}$$

where $K'' = eK'$.

We apply Lemma 34 for the last inequality in Equation 27. That is,

$$\begin{aligned} \epsilon^2 = & \frac{p}{c''' n^{1-2\alpha}} \log\left(c''' K''^2 p n^{1-2\alpha} \right), \\ = & \mathcal{O}\left(\frac{p}{n^{1-2\alpha}} \log(n) \right). \end{aligned} \tag{28}$$

where we assume that n is sufficiently large. The above statement holds for $\alpha < 1/2$.

In the following, we choose $\alpha = 0.4$ and use the assumption that

$$n^{0.2}/\log(n) \gtrsim p, \tag{29}$$

which provides $\epsilon < 1$. Note that this choice of α also justifies the assumption used to derive Equation 23. One can easily check that $\alpha = 0.4$ implies that the first and the second statements in Equation 27 are satisfied for sufficiently large n.

It remains to check whether $\nu(n^{\alpha}, p, \epsilon, B, K, \mathbf{\Sigma}) < \epsilon/4$ (in Lemma 24) for this particular choice of α and ϵ. It suffices to consider only the dominant terms in the definition of ν. We use the assumption on n, p and write

$$\begin{aligned} \nu(n^{0.4}, p, \epsilon, B, K, \mathbf{\Sigma}) \lesssim & n^{0.4} \exp\left(-\frac{c n^{0.4}}{6BK^2 p} \right) + n^{0.4} \exp\left(-\frac{c \mathrm{Tr}(\mathbf{\Sigma})/p \; n^{0.8/3}}{2K^2} \right), \\ \lesssim & n^{0.4} \exp\left(-\frac{c}{6BK^2} n^{0.2} \right) + n^{0.4} \exp\left(-\frac{c \lambda_{\min}(\mathbf{\Sigma})}{2K^2} n^{0.8/3} \right). \end{aligned} \tag{30}$$

For n sufficiently large, due to exponential decay in $n^{0.2}$, the above quantity can be made arbitrarily small. Hence, for some constants c_1, c_2, we obtain

$$\mathbb{P}\left(\sup_{\beta \in B_p(R)} \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) \langle x_i, v \rangle^2 - \mathbb{E}[f(\langle x, \beta \rangle) \langle x, v \rangle^2] \right| > c_1 \sqrt{\frac{p}{n^{0.2}} \log(n)} \right) \leq c_2 e^{-p}.$$

■

Appendix C. Proofs of Theorems 4 and 8

We will provide the proofs of Theorems 4 and 8 in parallel as they follow from similar steps. The only difference is the application of the lemmas that are provided in the previous

sections. On the event $\mathcal{E}$, we write,

$$\begin{aligned}\hat{\beta}^t - \beta_* - \gamma \mathbf{Q}^t \nabla_\beta \ell(\hat{\beta}^t) &= \hat{\beta}^t - \beta_* - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi (\hat{\beta}^t - \beta_*), \qquad (31)\\ &= \left(I - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right) (\hat{\beta}^t - \beta_*)\,.\end{aligned}$$

In the following, we will work on the event that $\widehat{\boldsymbol{\Sigma}}_S$ is invertible and that $[\mathbf{Q}^t]^{-1}$ is positive definite. We later show that conditioned on $\mathcal{E}$, this event holds with very high probability when $|S|$ is sufficiently large.

We use the nonexpensiveness of the projection $\mathcal{P}^t_{\mathcal{C}}$, i.e., for any $u, u' \in \mathbb{R}^p$ and $v = \mathcal{P}^t_{\mathcal{C}}(u)$, $v' = \mathcal{P}^t_{\mathcal{C}}(u')$ we have $\langle u - u', [\mathbf{Q}^t]^{-1}(u - u')\rangle \geq \langle v - v', [\mathbf{Q}^t]^{-1}(v - v')\rangle$. This simply means that the projection decreases the distance. Therefore, we can write

$$\begin{aligned}\left\| \hat{\beta}^{t+1} - \beta_* \right\|_{\mathbf{Q}^{t^{-1}}} &\leq \left\| \hat{\beta}^t - \beta_* - \gamma \mathbf{Q}^t \nabla_\beta \ell(\hat{\beta}^t) \right\|_{\mathbf{Q}^{t^{-1}}} \\ &\leq \left\| [\mathbf{Q}^t]^{-1/2} - \gamma [\mathbf{Q}^t]^{1/2} \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right\|_2 \left\| \hat{\beta}^t - \beta_* \right\|_2 . \qquad (32)\end{aligned}$$

The coefficient of $\|\hat{\beta}^t - \beta_*\|_2$ in Equation 32 determines the convergence behavior of the algorithm. Switching back to ℓ_2 norm, we obtain an upper bounded of the form

$$\left\| \hat{\beta}^{t+1} - \beta_* \right\|_2 \leq \left\| \mathbf{Q}^t \right\|_2 \left\| [\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right\|_2 \left\| \hat{\beta}^t - \beta_* \right\|_2 ,$$

where we have set step size $\gamma = 1$. First, we will bound the second term on the right hand side. We define the following,

$$\mathfrak{E}(\beta) = \mathbb{E}\left[\phi^{(2)}(\langle x, \beta \rangle) \right] \boldsymbol{\Sigma} + \mathbb{E}\left[\phi^{(4)}(\langle x, \beta \rangle) \right] \boldsymbol{\Sigma} \beta \beta^T \boldsymbol{\Sigma} .$$

Note that for a function f and fixed β, $\mathbb{E}[f(\langle x, \beta \rangle)] = h(\beta)$ is a function of β. With a slight abuse of notation, we write $\mathbb{E}[f(\langle x, \hat{\beta} \rangle)] = h(\hat{\beta})$ as a random variable. We have

$$\begin{aligned}\left\| [\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right\|_2 &\leq \left\| [\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t) \right\|_2 \qquad (33)\\ &+ \left\| [\mathbb{E}[x x^T \phi^{(2)}(\langle x, \hat{\beta}^t \rangle)] - \mathfrak{E}(\hat{\beta}^t) \right\|_2 \\ &+ \left\| \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi - \mathbb{E}\left[x x^T \int_0^1 \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*) \rangle) d\xi \right] \right\|_2 \\ &+ \left\| \mathbb{E}[x x^T \phi^{(2)}(\langle x, \hat{\beta}^t \rangle)] - \mathbb{E}\left[x x^T \int_0^1 \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*) \rangle) d\xi \right] \right\|_2 .\end{aligned}$$

For the first term on the right hand side, we state the following lemma.

Lemma 25 *When the covariates are sub-Gaussian, there exist constants C_1, C_2 such that, with probability at least $1 - C_1/p^2$,*

$$\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \leq C_2\sqrt{\frac{p}{\min\left\{|S|p/\log(p), n/\log(n)\right\}}}.$$

Similarly, when the covariates are sampled from a distribution with bounded support, there exist constants C_1', C_2', C_3' such that, with probability $1 - C_1'e^{-C_2'p}$,

$$\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \leq C_3'\sqrt{\frac{p}{\min\left\{|S|, n/\log(n)\right\}}},$$

where the constants depend on K, B and the radius R.

Proof In the following, we will only provide the proof for the bounded support case. The proof for the sub-Gaussian covariates follows from the same steps, by only replacing Lemma 14 with Lemma 12, and Lemma 18 with Lemma 20.

Using a uniform bound on the feasible set, we write

$$\begin{aligned}
&\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \\
&\leq \sup_{\beta\in\mathcal{C}} \left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S + \hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T - \mathbb{E}[\phi^{(2)}(\langle x,\beta\rangle)]\mathbf{\Sigma} - \mathbb{E}[\phi^{(4)}(\langle x,\beta\rangle)]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2.
\end{aligned}$$

We will find an upper bound for the quantity inside the supremum. By denoting the expectations of $\hat{\mu}_2(\beta)$ and $\hat{\mu}_4(\beta)$, with $\mu_2(\beta)$ and $\mu_4(\beta)$ respectively, we write

$$\begin{aligned}
&\left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S + \hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T - \mathbb{E}[\phi^{(2)}(\langle x,\beta\rangle)]\mathbf{\Sigma} - \mathbb{E}[\phi^{(4)}(\langle x,\beta\rangle)]\mathbf{\Sigma}\beta(\mathbf{\Sigma}\beta)^T\right\|_2 \\
&\leq \left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S - \mu_2(\beta)\mathbf{\Sigma}\right\|_2 + \left\|\hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T - \mu_4(\beta)\mathbf{\Sigma}\beta(\mathbf{\Sigma}\beta)^T\right\|_2.
\end{aligned}$$

For the first term on the right hand side, we have

$$\begin{aligned}
\left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S - \mu_2(\beta)\mathbf{\Sigma}\right\|_2 \leq& |\hat{\mu}_2(\beta)| \left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 + \|\mathbf{\Sigma}\|_2 \left|\hat{\mu}_2(\beta) - \mu_2(\beta)\right|, \\
\leq& B_2 \left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 + K\left|\hat{\mu}_2(\beta) - \mu_2(\beta)\right|.
\end{aligned}$$

By the Lemmas 14 and 18, for an absolute constant c, we have with probability $1 - 1/p^2$,

$$\begin{aligned}
\sup_{\beta\in\mathcal{C}} \left\|\hat{\mu}_2(\beta)\zeta_r(\widehat{\mathbf{\Sigma}}_S) - \mu_2(\beta)\mathbf{\Sigma}\right\|_2 \leq& B_2c\sqrt{K\|\mathbf{\Sigma}\|_2}\sqrt{\frac{\log(p)}{|S|}} + 3B_2K\sqrt{\frac{p\log(n)}{n}}, \\
\leq& 3cB_2K\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}}, \\
=& \mathcal{O}\left(\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}}\right).
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\left\|\hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T-\mu_4(\beta)\mathbf{\Sigma}\beta(\mathbf{\Sigma}\beta)^T\right\|_2\\
&\leq|\hat{\mu}_4(\beta)|\left\|\widehat{\mathbf{\Sigma}}_S\beta\beta^T\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2+|\hat{\mu}_4(\beta)-\mu_4(\beta)|\left\|\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2,\\
&\leq B_4R^2\left\{\|\widehat{\mathbf{\Sigma}}_S\|_2+\|\mathbf{\Sigma}\|_2\right\}\left\|\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}\right\|_2+R^2\|\mathbf{\Sigma}\|_2^2|\hat{\mu}_4(\beta)-\mu_4(\beta)|,\\
&\leq B_4R^2\left\{\|\widehat{\mathbf{\Sigma}}_S\|_2+K\right\}\left\|\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}\right\|_2+R^2K^2|\hat{\mu}_4(\beta)-\mu_4(\beta)|.
\end{aligned}$$

Again, by the Lemmas 14 and 18, for an absolute constant c, we have with probability $1-1/p^2$,

$$\begin{aligned}
B_4R^2\left\{\|\widehat{\mathbf{\Sigma}}_S\|_2+K\right\}\left\|\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}\right\|_2\leq&cKB_4R^2\left\{2K+cK\sqrt{\frac{\log(p)}{|S|}}\right\}\sqrt{\frac{\log(p)}{|S|}},\\
\leq&2cK^2B_4R^2\sqrt{\frac{\log(p)}{|S|}}+c^2K^2B_4R^2\frac{\log(p)}{|S|},\\
\leq&2cK^2B_4R^2\left(1+c\sqrt{\frac{\log(p)}{|S|}}\right)\sqrt{\frac{\log(p)}{|S|}},\\
\leq&4cK^2B_4R^2\sqrt{\frac{\log(p)}{|S|}},\\
=&\mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right),
\end{aligned}$$

for sufficiently large $|S|$, i.e., $|S|\geq c^2\log(p)$.

Further, by Lemma 18, we have with probability $1-2e^{-p}$,

$$\sup_{\beta\in\mathcal{C}}|\hat{\mu}_4(\beta)-\mu_4(\beta)|\leq 3B_4\sqrt{\frac{p\log(n)}{n}}=\mathcal{O}\left(\sqrt{\frac{p\log(n)}{n}}\right).$$

Combining the above results, for sufficiently large $p,|S|$, we have with probability at least $1-1/p^2-2e^{-p}$,

$$\begin{aligned}
&\sup_{\beta\in\mathcal{C}}\left\|\hat{\mu}_2(\beta)\zeta_r(\widehat{\mathbf{\Sigma}}_S)-\mu_2(\beta)\mathbf{\Sigma}\right\|_2+\sup_{\beta\in\mathcal{C}}\left\|\hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T-\mu_4(\beta)\mathbf{\Sigma}\beta(\mathbf{\Sigma}\beta)^T\right\|_2\\
&\leq 3B_2Kc\sqrt{\frac{p}{\min\left\{p/\log(p)|S|,n/\log(n)\right\}}}+4cK^2B_4R^2\sqrt{\frac{\log(p)}{|S|}}+3B_4R^2K^2\sqrt{\frac{p\log(n)}{n}},
\end{aligned}$$

$$
\begin{aligned}
&\leq 3B_2Kc\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}} + 4cK^2B_4R^2\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}},\\
&\leq CK\max\{B_2, B_4KR^2\}\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}},\\
&= \mathcal{O}\left(\sqrt{\frac{p}{\min\left\{|S|p/\log(p), n/\log(n)\right\}}}\right).
\end{aligned}
$$

Hence, for some constants C_1, C_2, with probability $1 - C_1/p^2$, we have

$$
\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \leq C_2\sqrt{\frac{p}{\min\left\{|S|p/\log(p), n/\log(n)\right\}}},
$$

where the constants depend on $K, B = \max\{B_2, B_4\}$ and the radius R. ■

Lemma 26 *The bias term can be upper bounded by*

$$
\left\|\mathbb{E}[xx^T\phi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \leq d_{\mathcal{H}_3}(x, z) + \|\mathbf{\Sigma}\|_2\ d_{\mathcal{H}_1}(x, z) + \|\mathbf{\Sigma}\|_2^2R^2\ d_{\mathcal{H}_2}(x, z),
$$

for both sub-Gaussian and bounded support cases.

Proof For a random variable $z \sim \mathsf{N}_p(0, \mathbf{\Sigma})$, by the triangle inequality, we write

$$
\begin{aligned}
&\left\|\mathbb{E}[xx^T\phi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2\\
&\leq \left\|\mathbb{E}[xx^T\phi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathbb{E}[zz^T\phi^{(2)}(\langle z, \hat{\beta}^t\rangle)]\right\|_2 + \left\|\mathbb{E}[zz^T\phi^{(2)}(\langle z, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2
\end{aligned}
$$

For the first term on the right hand side, we have

$$
\begin{aligned}
&\left\|\mathbb{E}[xx^T\phi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathbb{E}[zz^T\phi^{(2)}(\langle z, \hat{\beta}^t\rangle)]\right\|_2\\
&\leq \sup_{\beta\in\mathcal{C}}\sup_{\|v\|_2=1}\left|\mathbb{E}\left[\langle v, x\rangle^2\phi^{(2)}(\langle x, \beta\rangle)\right] - \mathbb{E}\left[\langle v, z\rangle^2\phi^{(2)}(\langle z, \beta\rangle)\right]\right|,\\
&\leq d_{\mathcal{H}_3}(x, z).
\end{aligned}
$$

For the second term, we write

$$\begin{aligned}
&\left\|[\mathbb{E}[zz^T\phi^{(2)}(\langle z,\hat{\beta}^t\rangle)]-\mathfrak{E}(\hat{\beta}^t)\right\|_2\\
&\leq \sup_{\beta\in\mathcal{C}}\left\|\mathbb{E}[zz^T\phi^{(2)}(\langle z,\beta\rangle)]-\mathbb{E}[\phi^{(2)}(\langle x,\beta\rangle)]\mathbf{\Sigma}+\mathbb{E}\left[\phi^{(4)}(\langle x,\beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2,\\
&\leq \sup_{\beta\in\mathcal{C}}\Big\|\mathbb{E}[\phi^{(2)}(\langle z,\beta\rangle)]\mathbf{\Sigma}+\mathbb{E}\left[\phi^{(4)}(\langle z,\beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\\
&\qquad\qquad-\mathbb{E}[\phi^{(2)}(\langle x,\beta\rangle)]\mathbf{\Sigma}-\mathbb{E}\left[\phi^{(4)}(\langle x,\beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\Big\|_2,\\
&\leq \sup_{\beta\in\mathcal{C}}\left\|\mathbb{E}[\phi^{(2)}(\langle z,\beta\rangle)]\mathbf{\Sigma}-\mathbb{E}[\phi^{(2)}(\langle x,\beta\rangle)]\mathbf{\Sigma}\right\|_2,\\
&\qquad\qquad+\sup_{\beta\in\mathcal{C}}\left\|\mathbb{E}\left[\phi^{(4)}(\langle z,\beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}-\mathbb{E}\left[\phi^{(4)}(\langle x,\beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2,\\
&\leq \|\mathbf{\Sigma}\|_2\sup_{\beta\in\mathcal{C}}\left|\mathbb{E}[\phi^{(2)}(\langle z,\beta\rangle)]-\mathbb{E}[\phi^{(2)}(\langle x,\beta\rangle)]\right|\\
&\qquad\qquad+\|\mathbf{\Sigma}\|_2^2R^2\sup_{\beta\in\mathcal{C}}\left|\mathbb{E}[\phi^{(4)}(\langle z,\beta\rangle)]-\mathbb{E}[\phi^{(4)}(\langle x,\beta\rangle)]\right|,\\
&\leq \|\mathbf{\Sigma}\|_2 d_{\mathcal{H}_1}(x,z)+\|\mathbf{\Sigma}\|_2^2R^2d_{\mathcal{H}_2}(x,z).
\end{aligned}$$

Hence, we conclude that

$$\left\|\mathbb{E}[xx^T\phi^{(2)}(\langle x,\hat{\beta}^t\rangle)]-\mathfrak{E}(\hat{\beta}^t)\right\|_2\leq d_{\mathcal{H}_3}(x,z)+\|\mathbf{\Sigma}\|_2\ d_{\mathcal{H}_1}(x,z)+\|\mathbf{\Sigma}\|_2^2R^2\ d_{\mathcal{H}_2}(x,z).$$

■

Lemma 27 *There exists constants c_1,c_2,c_3 depending on the eigenvalues of $\mathbf{\Sigma}$, B,L and R such that, with probability at least $1-c_2e^{-c_3p}$*

$$\left\|\frac{1}{n}\sum_{i=1}^n x_ix_i^T\int_0^1\phi^{(2)}(\langle x_i,\beta_*+\xi(\hat{\beta}^t-\beta_*)\rangle)d\xi-\mathbb{E}\left[xx^T\int_0^1\phi^{(2)}(\langle x,\beta_*+\xi(\hat{\beta}^t-\beta_*)\rangle)d\xi\right]\right\|_2\leq\delta,$$

where $\delta=c_1\sqrt{\frac{p}{n^{0.2}}\log(n)}$ for sub-Gaussian covariates, and $\delta=c_1\sqrt{\frac{p}{n}\log(n)}$ for covariates with bounded support.

Proof We provide the proof for bounded support case. The proof for sub-Gaussian case can be carried by replacing Lemma 19 with Lemma 23.

By the Fubini's theorem, we have

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \int_0^1 \phi^{(2)}(\langle x_i, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi - \mathbb{E}\left[xx^T \int_0^1 \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi \right] \right\|_2, \\
&= \left\| \int_0^1 \left\{ \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \phi^{(2)}(\langle x_i, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) - \mathbb{E}\left[xx^T \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) \right] \right\} d\xi \right\|_2, \\
&\leq \int_0^1 \left\| \left\{ \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \phi^{(2)}(\langle x_i, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) - \mathbb{E}\left[xx^T \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) \right] \right\} \right\|_2 d\xi, \\
&\leq \sup_{\beta \in \mathcal{C}} \left\| \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \phi^{(2)}(\langle x_i, \beta\rangle) - \mathbb{E}\left[xx^T \phi^{(2)}(\langle x, \beta\rangle) \right] \right\|_2.
\end{aligned}$$

Using the properties of operator norm, the above bound can be written as

$$\begin{aligned}
&\sup_{\beta \in \mathcal{C}} \left\| \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \phi^{(2)}(\langle x_i, \beta\rangle) - \mathbb{E}\left[xx^T \phi^{(2)}(\langle x, \beta\rangle) \right] \right\|_2 \\
&= \sup_{\beta \in \mathcal{C}} \sup_{v \in S^{p-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \phi^{(2)}(\langle x_i, \beta\rangle) \langle x_i, v\rangle^2 - \mathbb{E}\left[\phi^{(2)}(\langle x, \beta\rangle) \langle x, v\rangle^2 \right] \right|,
\end{aligned}$$

where S^{p-1} denotes the p-dimensional unit sphere.
For $\Delta = 0.25$, let T_Δ be an Δ-net over S^{p-1}. Using Lemma 32, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\beta \in \mathcal{C}} \sup_{v \in S^{p-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \phi^{(2)}(\langle x_i, \beta\rangle) \langle x_i, v\rangle^2 - \mathbb{E}\left[\phi^{(2)}(\langle x, \beta\rangle) \langle x, v\rangle^2 \right] \right| > \epsilon \right), \\
&\leq \mathbb{P}\left(\sup_{\beta \in \mathcal{C}} \sup_{v \in T_\Delta} \left| \frac{1}{n}\sum_{i=1}^{n} \phi^{(2)}(\langle x_i, \beta\rangle) \langle x_i, v\rangle^2 - \mathbb{E}\left[\phi^{(2)}(\langle x, \beta\rangle) \langle x, v\rangle^2 \right] \right| > \epsilon/2 \right), \\
&\leq |T_\Delta| \mathbb{P}\left(\sup_{\beta \in \mathcal{C}} \left| \frac{1}{n}\sum_{i=1}^{n} \phi^{(2)}(\langle x_i, \beta\rangle) \langle x_i, v\rangle^2 - \mathbb{E}\left[\phi^{(2)}(\langle x, \beta\rangle) \langle x, v\rangle^2 \right] \right| > \epsilon/2 \right), \\
&= 9^p \mathbb{P}\left(\sup_{\beta \in \mathcal{C}} \left| \frac{1}{n}\sum_{i=1}^{n} \phi^{(2)}(\langle x_i, \beta\rangle) \langle x_i, v\rangle^2 - \mathbb{E}\left[\phi^{(2)}(\langle x, \beta\rangle) \langle x, v\rangle^2 \right] \right| > \epsilon/2 \right).
\end{aligned}$$

By applying Lemma 19 to the last line above, we obtain

$$\mathbb{P}\left(\sup_{\beta \in \mathcal{C}} \left| \frac{1}{n}\sum_{i=1}^{n} \phi^{(2)}(\langle x_i, \beta\rangle) \langle x_i, v\rangle^2 - \mathbb{E}[\phi^{(2)}(\langle x, \beta\rangle) \langle x, v\rangle^2] \right| > 4B_2 K \sqrt{\frac{p}{n}\log(n)} \right) \leq 2e^{-3.2p}.$$

Notice that $3.2 - \log(9) > 1$. Therefore, by choosing n large enough, on the set $\mathcal{E}$, we obtain that with probability at least $1 - 2e^{-p}$

$$\sup_{\beta \in \mathcal{C}} \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \phi^{(2)}(\langle x_i, \beta \rangle) - \mathbb{E}\left[x x^T \phi^{(2)}(\langle x, \beta \rangle) \right] \right\|_2 \leq 8 B_2 K \sqrt{\frac{p}{n} \log(n)}.$$

■

Lemma 28 *There exists a constant C depending on K and L such that,*

$$\left\| \mathbb{E}[x x^T \phi^{(2)}(\langle x, \hat{\beta}^t \rangle)] - \mathbb{E}\left[x x^T \int_0^1 \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*) \rangle) d\xi \right] \right\|_2 \leq \tilde{C} \|\hat{\beta}^t - \beta_*\|_2,$$

where $\tilde{C} = C$ for the bounded support case and $\tilde{C} = C p^{1.5}$ for the sub-Gaussian case.

Proof By the Fubini's theorem, we write

$$\begin{aligned}
&\left\| \mathbb{E}[x x^T \phi^{(2)}(\langle x, \hat{\beta}^t \rangle)] - \mathbb{E}\left[x x^T \int_0^1 \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*) \rangle) d\xi \right] \right\|_2, \\
&= \left\| \int_0^1 \mathbb{E}\left[x x^T \left\{ \phi^{(2)}(\langle x, \hat{\beta}^t \rangle) - \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*) \rangle) \right\} \right] d\xi \right\|_2.
\end{aligned}$$

Moving the integration out, right hand side of the above equation is smaller than

$$\begin{aligned}
&\int_0^1 \left\| \mathbb{E}\left[x x^T \left\{ \phi^{(2)}(\langle x, \hat{\beta}^t \rangle) - \phi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*) \rangle) \right\} \right] \right\|_2 d\xi, \\
&\leq \int_0^1 \left\| \mathbb{E}\left[x x^T L |\langle x, (1-\xi)(\hat{\beta}^t - \beta_*) \rangle| \right] \right\|_2 d\xi, \\
&\leq \mathbb{E}\left[\|x\|_2^3 \|\hat{\beta}^t - \beta_*\|_2 \right] L \int_0^1 (1-\xi) d\xi, \\
&= \frac{L \mathbb{E}[\|x\|_2^3]}{2} \|\hat{\beta}^t - \beta_*\|_2.
\end{aligned}$$

We observe that, when the covariates are supported in the ball of radius $\sqrt{K}$, we have $\mathbb{E}[\|x\|_2^3] \leq K^{3/2}$. When they are sub-Gaussian random variables with norm K, we have $\mathbb{E}[\|x\|_2^3] \leq K^3 6^{1.5} p^{1.5}$.

■

By combining the above results, for bounded covariates we obtain

$$\begin{aligned}
&\left\| [\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}_\beta^2 \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right\|_2 \\
&\quad \leq \mathfrak{D}(x, z) + c_1 \sqrt{\frac{p}{\min\{|S| p / \log(p), n / \log(n)\}}} + c_2 \|\hat{\beta}^t - \beta_*\|_2,
\end{aligned}$$

and for sub-Gaussian covariates, we obtain

$$\begin{aligned}&\left\|[\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_{\beta}\ell(\beta_* + \xi(\hat{\beta}^t - \beta_*))d\xi\right\|_2 \\ &\qquad \leq \mathfrak{D}(x,z) + c_1\sqrt{\frac{p}{\min\left\{|S|, n^{0.2}/\log(n)\right\}}} + c_2 p^{1.5}\|\hat{\beta}^t - \beta_*\|_2\,,\end{aligned}$$

where

$$\mathfrak{D}(x,z) = d_{\mathcal{H}_3}(x,z) + \|\boldsymbol{\Sigma}\|_2\; d_{\mathcal{H}_1}(x,z) + \|\boldsymbol{\Sigma}\|_2^2 R^2\; d_{\mathcal{H}_2}(x,z)\,.$$

In the following, we will derive an upper bound for $\left\|\mathbf{Q}^t\right\|_2$, which is equivalent to proving the positive definiteness of $[\mathbf{Q}^t]^{-1}$ and finding a lower bound for $\|[\mathbf{Q}^t]^{-1}\|_2$. The sub-Gaussian case is more restrictive than the bounded support case. Therefore we derive the bound for the sub-Gaussian case. We have

$$\begin{aligned}\lambda_{\min}\left([\mathbf{Q}^t]^{-1}\right) &= \inf_{\|u\|_2=1}\left\{\hat{\mu}_2(\hat{\beta}^t)\langle u, \widehat{\boldsymbol{\Sigma}}_S u\rangle + \hat{\mu}_4(\hat{\beta}^t)\langle u, \widehat{\boldsymbol{\Sigma}}_S\hat{\beta}^t\rangle^2\right\}, \\ &\geq \inf_{\|u\|_2=1}\left\{\hat{\mu}_2(\hat{\beta}^t)\langle u, \boldsymbol{\Sigma} u\rangle + \hat{\mu}_4(\hat{\beta}^t)\langle u, \boldsymbol{\Sigma}\hat{\beta}^t\rangle^2\right\} \\ &\quad - B_2\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}\|_2 - B_4R^2\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}\|_2\|\widehat{\boldsymbol{\Sigma}}_S + \boldsymbol{\Sigma}\|_2.\end{aligned}$$

On the event $\mathcal{E}$, the first term on the right hand side is lower bounded by κ^{-1}. For the other terms, we use Lemma 12 and write

$$\begin{aligned}\lambda_{\min}\left([\mathbf{Q}^t]^{-1}\right) &\leq 2\kappa^{-1} - \|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}\|_2\left\{B_2 + B_4R^2\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}\|_2 + 2B_4R^2\|\boldsymbol{\Sigma}\|_2\right\}, \\ &\leq 2\kappa^{-1} - C\sqrt{\frac{p}{|S|}}\left\{B_2 + B_4R^2C\sqrt{\frac{p}{|S|}} + 2B_4R^2\|\boldsymbol{\Sigma}\|_2\right\}\end{aligned}$$

with probability $1 - 2e^{-cp}$. When $|S| > 4pC^2\max\left\{1, 2C(B_2 + 3B_4R^2\lambda_{\max}(\boldsymbol{\Sigma}))\kappa\right\}^2$, with probability $1 - 2e^{-cp}$, we obtain

$$\lambda_{\min}\left([\mathbf{Q}^t]^{-1}\right) \geq \kappa^{-1}.$$

This proves that, with high probability, on the event $\mathcal{E}$, $[\mathbf{Q}^t]^{-1}$ is positive definite and consequently we obtain

$$\|\mathbf{Q}^t\|_2 \leq \kappa.$$

Finally, we take into account the conditioning on the event $\mathcal{E}$. Since we worked on the event $\mathcal{E}$, the probability of a desired outcome is at least $\mathbb{P}(\mathcal{E}) - \delta$, where δ is either c/p^2 or ce^{-p} depending on the distribution of the covariates. Hence, conditioned on the event $\mathcal{E}$, the probability becomes $1 - \delta/\mathbb{P}(\mathcal{E})$, which completes the proof.

C.1 Proof of Corollaries 5 and 9

In the following, we provide the proof for Corollary 5. The proof for Corollary 9 follows from the exact same steps.

The statement of Theorem 4 holds on the probability space with a probability lower bounded by $\mathbb{P}(\mathcal{E}) - c/p^2$ for some constant c (See previous section). Let $\mathcal{Q}$ denote this set, on which the statement of the theorem holds without the conditioning on the event $\mathcal{E}$. Note that $\mathcal{Q} \subset \mathcal{E}$ and we also have

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{Q}) \geq \mathbb{P}(\mathcal{E}) - c/p^2. \tag{34}$$

This suggests that the difference between $\mathcal{Q}$ and $\mathcal{E}$ is small. By taking expectations on both sides over the set $\mathcal{Q}$, we obtain,

$$\begin{aligned}\mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}\right] \leq \kappa \left\{\mathfrak{D}(x,z) + c_1\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}}\right\}\mathbb{E}\left[\|\hat{\beta}^{t} - \beta_*\|_2\right] \\ +\kappa c_2 \mathbb{E}\left[\|\hat{\beta}^{t} - \beta_*\|_2^2\right]\end{aligned}$$

where we used

$$\mathbb{E}\left[\|\hat{\beta}^{t} - \beta_*\|_2^l; \mathcal{Q}\right] \leq \mathbb{E}\left[\|\hat{\beta}^{t} - \beta_*\|_2^l\right], \qquad l = 1, 2.$$

Similarly for the iterate $\hat{\beta}^{t+1}$, we write

$$\begin{aligned}\mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2\right] =& \mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}\right] + \mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}^C\right], \\ \leq& \mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}\right] + 2R\mathbb{P}(\mathcal{Q}^C), \\ \leq& \mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}\right] + 2R\left(\mathbb{P}(\mathcal{E}^C) + \frac{c}{p^2}\right), \\ \leq& \mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}\right] + \frac{\epsilon}{10}, \\ \leq& \mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}\right] + \frac{\mathbb{E}\left[\|\hat{\beta}^{t} - \beta_*\|_2\right]}{10}.\end{aligned}$$

Combining these two inequalities, we obtain

$$\begin{aligned}&\mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2\right] \leq \left\{0.1 + \kappa\mathfrak{D}(x,z) + c_1\kappa\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}}\right\}\mathbb{E}\left[\|\hat{\beta}^{t} - \beta_*\|_2\right] \\ &+ c_2\kappa\mathbb{E}\left[\|\hat{\beta}^{t} - \beta_*\|_2^2\right].\end{aligned}$$

Hence the proof follows.

C.2 Proof of Theorem 6

The iterates generated by the Newton-Stein method satisfy the following inequality,

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \left(\tau_1 + \tau_2 \|\hat{\beta}^t - \beta_*\|_2\right) \|\hat{\beta}^t - \beta_*\|_2,$$

on the event $\mathcal{Q}$ where $\mathcal{Q}$ is defined in the previous section. We have observed that $\mathbb{P}(\mathcal{Q}) \geq \mathbb{P}(\mathcal{E}) - c/p^2$ in Equation 34. Since the coefficients τ_1 and τ_2 are obtained by uniform bounds on the feasible set, the above inequality holds for every t on $\mathcal{Q}$. On the event we consider, $\mathcal{Q} \cap \{\vartheta < (1-\tau_1)/\tau_2\}$, the starting point satisfies the following

$$\tau_1 + \tau_2 \|\hat{\beta}^0 - \beta_*\|_2 < 1, \tag{35}$$

which implies that the sequence of iterates converges. Let $\xi \in (\epsilon, \vartheta)$ and t_ξ be the last iteration that $\|\hat{\beta}^t - \beta_*\|_2 > \xi$. Then, for $t > t_\xi$

$$\begin{aligned}\|\hat{\beta}^{t+1} - \beta_*\|_2 &\leq \left(\tau_1 + \tau_2 \|\hat{\beta}^t - \beta_*\|_2\right) \|\hat{\beta}^t - \beta_*\|_2, \\ &\leq (\tau_1 + \tau_2 \xi) \|\hat{\beta}^t - \beta_*\|_2.\end{aligned}$$

This convergence behavior describes a linear rate and requires at most

$$\frac{\log(\epsilon/\xi)}{\log(\tau_1 + \tau_2 \xi)}$$

iterations to reach a tolerance of ϵ. For $t \leq t_\xi$, we have

$$\begin{aligned}\|\hat{\beta}^{t+1} - \beta_*\|_2 &\leq \left(\tau_1 + \tau_2 \|\hat{\beta}^t - \beta_*\|_2\right) \|\hat{\beta}^t - \beta_*\|_2, \\ &\leq (\tau_1/\xi + \tau_2) \|\hat{\beta}^t - \beta_*\|_2^2.\end{aligned}$$

This describes a quadratic rate and the number of iterations to reach a tolerance of ξ can be upper bounded by

$$\log_2 \left(\frac{\log\left(\xi \left(\tau_1/\xi + \tau_2\right)\right)}{\log\left(\tau_1/\xi + \tau_2\right) \|\hat{\beta}^0 - \beta_*\|_2} \right) \leq \log_2 \left(\frac{\log\left(\tau_1 + \tau_2 \xi\right)}{\log\left((\tau_1/\xi + \tau_2)(1-\tau_1)/\tau_2\right)} \right).$$

Therefore, the overall number of iterations to reach a tolerance of ϵ is upper bounded by

$$\log_2 \left(\frac{\log\left(\tau_1 + \tau_2 \xi\right)}{\log\left((\tau_1/\xi + \tau_2)(1-\tau_1)/\tau_2\right)} \right) + \frac{\log(\epsilon/\xi)}{\log(\tau_1 + \tau_2 \xi)}$$

which is a function of ξ. Therefore, we take the minimum over the feasible set and conclude that on $\mathcal{E} \cap \{\vartheta < (1-\tau_1)/\tau_2\}$, the number of iterations to reach a tolerance of ϵ is upper bounded by $\inf_\xi \mathcal{J}(\xi)$ with a bad event probability of c/p^2. By conditioning on the event $\mathcal{E} \cap \{\vartheta < (1-\tau_1)/\tau_2\}$, we conclude that with probability at least $1 - c'/p^2$, the statement of the theorem holds for $c' = c/\mathbb{P}(\mathcal{E} \cap \{\vartheta < (1-\tau_1)/\tau_2\})$.

Appendix D. Proof of Theorem 7

We have the following projected updates

$$\hat{\beta}^{t+1} = \mathcal{P}_{\mathcal{C}}\left(\hat{\beta}^t - \gamma_t \mathbf{Q}^t \nabla \ell(\hat{\beta}^t); \mathbf{Q}^t\right) = \hat{\beta}^t - \gamma_t D_{\gamma_t}(\hat{\beta}^t),$$

where we define

$$D_\gamma(\hat{\beta}^t) = \frac{1}{\gamma}\left(\hat{\beta}^t - \mathcal{P}_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla \ell(\hat{\beta}^t); \mathbf{Q}^t)\right).$$

For simplicity, we only consider the projection onto a convex set, i.e.,

$$\begin{aligned} \mathcal{P}^t_{\mathcal{C}}(\beta^+) = \mathcal{P}_{\mathcal{C}}(\beta^+; \mathbf{Q}^t) =& \operatorname*{argmin}_{w \in \mathcal{C}} \frac{1}{2}\|w - \beta^+\|^2_{\mathbf{Q}^{t-1}}, \\ =& \operatorname*{argmin}_{w \in \mathbb{R}^p} \frac{1}{2}\|w - \beta^+\|^2_{\mathbf{Q}^{t-1}} + \mathbb{I}_{\mathcal{C}}(w), \end{aligned} \tag{36}$$

where $\mathbb{I}_{\mathcal{C}}(w)$ is the indicator function for the convex set $\mathcal{C}$, i.e.

$$\mathbb{I}_{\mathcal{C}}(w) = \begin{cases} 0 & \text{if } w \in \mathcal{C}, \\ \infty & \text{otherwise.} \end{cases}$$

We note that other projection methods (such as proximal mappings) are also applicable to our update rule.

Defining the decrement $\lambda^t = \langle \nabla \ell(\hat{\beta}^t), D_\gamma(\hat{\beta}^t)\rangle$, we consider the following form of backtracking line search with update parameters $a \in (0, 0.5)$ and $b \in (0, 1)$:

$$\gamma = \bar{\gamma}; \quad \textbf{while: } \ell\left(\hat{\beta}^t - \gamma D_\gamma(\hat{\beta}^t)\right) > \ell(\hat{\beta}^t) - a\gamma\lambda^t, \quad \gamma \leftarrow \gamma b.$$

Depending on the projection choice, there are various other search methods that can be applied. Before we move on to the convergence analysis, we first establish some properties of the modified gradient D_γ.

For a given point $w \in \mathcal{C}$, the sub-differential of the indicator function is the normal cone. This together with Equation 36 implies that

$$\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla \ell(\hat{\beta}^t) - \mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla \ell(\hat{\beta}^t)) \in \mathbf{Q}^t \, \partial \mathbb{I}_{\mathcal{C}}(\mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla \ell(\hat{\beta}^t))),$$

which in turn implies

$$\gamma[\mathbf{Q}^t]^{-1}\left\{D_\gamma(\hat{\beta}^t) - \mathbf{Q}^t \nabla \ell(\hat{\beta}^t),\right\} \in \partial \mathbb{I}_{\mathcal{C}}(\mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla \ell(\hat{\beta}^t))),$$

and correspondingly for any $\beta \in \mathcal{C}$

$$\langle [\mathbf{Q}^t]^{-1} D_\gamma(\hat{\beta}^t) - \nabla \ell(\hat{\beta}^t), \mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla \ell(\hat{\beta}^t)) - \beta\rangle \geq 0.$$

For $\beta = \hat{\beta}^t \in \mathcal{C}$, this yields

$$\kappa^{-1}\|D_\gamma(\hat{\beta}^t)\|_2^2 \leq \langle D_\gamma(\hat{\beta}^t), [\mathbf{Q}^t]^{-1} D_\gamma(\hat{\beta}^t)\rangle \leq \langle \nabla \ell(\hat{\beta}^t), D_{\gamma_t}(\hat{\beta}^t)\rangle, \tag{37}$$

with probability at least $P(\mathcal{E}) - c/p^2$. Also note that the Hessian of the GLM problem can be upper bounded by

$$\left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \phi^{(2)}(\langle x_i, \hat{\beta}^t \rangle \right\|_2 \leq B_2 \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right\|_2 \leq B_2 K.$$

Now we move to the convergence analysis. For a step size γ, by the convexity of the negative log-likelihood, we can write almost surely

$$\begin{aligned} \ell(\hat{\beta}^t - \gamma D_\gamma(\hat{\beta}^t)) \leq & \ell(\hat{\beta}^t) - \gamma \langle \nabla \ell(\hat{\beta}^t), D_\gamma(\hat{\beta}^t) \rangle + \frac{\gamma^2 B_2 K}{2} \|D_\gamma(\hat{\beta}^t)\|_2^2, \\ \leq & \ell(\hat{\beta}^t) - \gamma \langle \nabla \ell(\hat{\beta}^t), D_\gamma(\hat{\beta}^t) \rangle \left\{ 1 - \frac{\gamma}{2} B_2 K \kappa \right\} \end{aligned}$$

and notice that the exit condition for the backtracking line search algorithm is satisfied when $\gamma \leq (\kappa B_2 K)^{-1}$. Hence, the line search returns a step size satisfying

$$\gamma_t \geq \min\{\bar{\gamma}, b/(\kappa B_2 K)\}.$$

Using the line search condition, we have

$$\ell\left(\hat{\beta}^t - \gamma_t D_{\gamma_t}(\hat{\beta}^t)\right) - \ell(\hat{\beta}^t) \leq - a\gamma_t \lambda^t,$$

with probability at least $\mathbb{P}(\mathcal{E}) - c/p^2$ which implies that the sequence $\{\ell(\hat{\beta}^t)\}_t$ is decreasing. We note that this event is independent of the iteration number due to uniform positive definite condition given in $\mathcal{E}$. Since ℓ is continuous and $\mathcal{C}$ is closed, ℓ is a closed function. Hence, the sequence $\{\ell(\hat{\beta}^t)\}_t$ must converge to a limit. This implies that $a\gamma_t\lambda^t \to 0$. But we have $a > 0$ and $\gamma_t > \min\{\bar{\gamma}, b/(\kappa B_2 K)\} > 0$. Therefore, we conclude that $\lambda^t \to 0$. Using the inequality provided in Equation 37, we conclude that $\|D_\gamma(\hat{\beta}^t)\|_2$ converges to 0 which implies that the algorithm converges with probability at least $1 - \frac{c}{\mathbb{P}(\mathcal{E})} p^{-2}$, where in the last step we conditioned on $\mathcal{E}$.

Appendix E. Local Step Size Selection

This section provides a heuristic calculation for choosing a local step size when eigenvalue thresholding is applied to the Newton-Stein method. We carry our analysis from Equation 32. The optimal local step size would be

$$\gamma_* = \text{argmin}_\gamma \left\| I - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right\|_2.$$

Defining the following matrix,

$$\boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t) = \int_0^1 \boldsymbol{\nabla}^2_\beta \ell(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi,$$

and we write the governing term as

$$\left\| I - \gamma \mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t) \right\|_2.$$

The above function is piecewise linear in γ and it can be minimized by setting

$$\gamma_* = \frac{2}{\lambda_1\left(\mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)\right) + \lambda_p\left(\mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)\right)}.$$

Since we don't have access to the optimal value β_*, we cannot determine the exact value of $\boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)$. Hence, we will assume that $\boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)$ and the current estimate are close.

In the regime $n \gg p$, and by our construction of the scaling matrix $\mathbf{Q}^t$, we have

$$\mathbf{Q}^t \approx \left[\mathbb{E}[xx^T \phi^{(2)}(\langle x, \hat{\beta}^t\rangle)]\right]^{-1} \quad \text{and} \quad \boldsymbol{\nabla}^2_\beta \ell(\hat{\beta}^t) \approx \mathbb{E}[xx^T \phi^{(2)}(\langle x, \hat{\beta}^t\rangle)].$$

The crucial observation is that the eigenvalue thresholding suggested in Erdogdu and Montanari, 2015 estimates the smallest eigenvalue with $(r+1)$-th eigenvalue (say $\hat{\sigma}^2$) which overestimates true value (say σ^2) in general. Even though the largest eigenvalue of $\mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)$ will be close to 1, the smallest value will be $\sigma^2/\hat{\sigma}^2$. This will make the optimal step size larger than 1. Hence, we suggest

$$\gamma = \frac{2}{1 + \sigma^2/\hat{\sigma}^2},$$

if σ^2 were known. We also have, by the Weyl's inequality,

$$\left|\hat{\sigma}^2 - \sigma^2\right| \leq \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2 \leq C\sqrt{\frac{p}{|S|}},$$

with high probability. Whenever r is less than $p/2$, we suggest to use

$$\gamma = \frac{2}{1 + \frac{\hat{\sigma}^2 - \mathcal{O}(\sqrt{p/|S|})}{\hat{\sigma}^2}},$$

if σ^2 is unknown.

Appendix F. Useful Lemmas

Lemma 29 *Let Γ denote the* Gamma *function. Then, for $r \in (0,1)$, we have*

$$z^{1-r} < \frac{\Gamma(z+1)}{\Gamma(z+r)} < (1+z)^{1-r}.$$

Lemma 30 *Let Z be a random variable with a density function f and cumulative distribution function F. If $F^C = 1 - F$, then,*

$$\left|\mathbb{E}[Z\mathbb{I}_{\{|Z|>t\}}]\right| \leq t\mathbb{P}(|Z| > t) + \int_t^\infty \mathbb{P}(|Z| > z)dz.$$

Proof We write,

$$\mathbb{E}[Z\mathbb{I}_{\{|Z|>t\}}] = \int_t^\infty z f(z)dz + \int_{-\infty}^{-t} z f(z)dz.$$

Using integration by parts, we obtain

$$\begin{aligned}\int zf(z)dz =& - zF^C(z) + \int F^C(z)dz,\\ =& zF(z) - \int F(z)dz.\end{aligned}$$

Since $\lim_{z\to\infty} zF^C(z) = \lim_{z\to-\infty} zF(z) = 0$, we have

$$\begin{aligned}\int_t^\infty zf(z)dz =& tF^C(t) + \int_t^\infty F^C(z)dz,\\ \int_{-\infty}^{-t} zf(z)dz =& - tF(-t) - \int_{-\infty}^{-t} F(z)dz,\\ =& - tF(-t) - \int_t^\infty F(-z)dz.\end{aligned}$$

Hence, we obtain the following bound,

$$\begin{aligned}\left|\mathbb{E}[Z\mathbb{I}_{\{|Z|>t\}}]\right| =& \left|tF^C(t) + \int_t^\infty F^C(z)dz - tF(-t) - \int_t^\infty F(-z)dz\right|,\\ \leq& t\left(F^C(t) + F(-t)\right) + \left(\int_t^\infty F^C(z) + F(-z)dz\right),\\ \leq& t\mathbb{P}(|Z|>t) + \int_t^\infty \mathbb{P}(|Z|>z)dz.\end{aligned}$$

■

Lemma 31 *For positive constants c_1, c_2, we have*

$$\int_{c_1}^\infty e^{-c_2t^{2/3}}dt \leq \left\{\frac{3c_1^{1/3}}{2c_2} + \frac{3}{4c_2^2c_1^{1/3}}\right\}e^{-c_2c_1^{2/3}}$$

Proof By the change of variables $t^{2/3} = x^2$, we get

$$\int_{c_1}^\infty e^{-c_2t^{2/3}}dt = 3\int_{c_1^{1/3}}^\infty x^2e^{-c_2x^2}dx.$$

Next, we notice that

$$de^{-c_2x^2} = -2c_2xe^{-c_2x^2}dx.$$

Hence, using the integration by parts, we have

$$\int_{c_1}^\infty e^{-c_2t^{2/3}}dt = \frac{3}{2c_2}\left\{c_1^{1/3}e^{-c_2c_1^{2/3}} + \int_{c_1^{1/3}}^\infty e^{-c_2x^2}dx\right\}.$$

We will find an upper bound on the second term. Using the change of variables, $x = y + c_1^{1/3}$, we obtain

$$\begin{aligned}\int_{c_1^{1/3}}^{\infty} e^{-c_2 x^2} dx &= \int_0^{\infty} e^{-c_2\left(y+c_1^{1/3}\right)^2} dy, \\ &\leq e^{-c_2 c_1^{2/3}} \int_0^{\infty} e^{-2c_2 y c_1^{1/3}} dy, \\ &= \frac{e^{-c_2 c_1^{2/3}}}{2c_2 c_1^{1/3}}.\end{aligned}$$

Combining the above results, we complete the proof. ■

Lemma 32 (Vershynin, 2010) *Let X be a symmetric $p \times p$ matrix, and let T_ϵ be an ϵ-net over S^{p-1}. Then,*

$$\|X\|_2 \leq \frac{1}{1-2\epsilon} \sup_{v \in T_\epsilon} |\langle Xv, v \rangle| .$$

Lemma 33 *Let $B_p(R) \subset \mathbb{R}^p$ be the ball of radius R centered at the origin and T_ϵ be an ϵ-net over $B_p(R)$. Then,*

$$|T_\epsilon| \leq \left(\frac{R\sqrt{p}}{\epsilon}\right)^p .$$

Proof A similar proof appears in (Van der Vaart, 2000). The set $B_p(R)$ can be contained in a p-dimensional cube of size $2R$. Consider a grid over this cube with mesh width $2\epsilon/\sqrt{p}$. Then $B_p(R)$ can be covered with at most $(2R/(2\epsilon/\sqrt{p}))^p$ many cubes of edge length $2\epsilon/\sqrt{p}$. If ones takes the projection of the centers of such cubes onto $B_p(R)$ and considers the circumscribed balls of radius ϵ, we may conclude that $B_p(R)$ can be covered with at most

$$\left(\frac{2R}{2\epsilon/\sqrt{p}}\right)^p$$

many balls of radius ϵ. ■

Lemma 34 *For $a, b > 0$, and ϵ satisfying*

$$\epsilon = \left\{\frac{a}{2}\log\left(\frac{2b^2}{a}\right)\right\}^{1/2} \qquad \text{and} \qquad \frac{2}{a}b^2 > e,$$

we have $\epsilon^2 \geq a\log(b/\epsilon)$. Moreover, the gap in the inequality can be written as

$$\epsilon^2 - a\log(b/\epsilon) = \frac{a}{2}\log\log\left(\frac{2b^2}{a}\right).$$

Proof Since $a, b > 0$ and $x \to e^x$ is a monotone increasing function, the above inequality condition is equivalent to

$$\frac{2\epsilon^2}{a} e^{\frac{2\epsilon^2}{a}} \geq \frac{2b^2}{a}.$$

Now, we use the function $f(w) = we^w$ for $w > 0$ (in fact this function is well-known by the name Lambert W function). f is continuous and invertible on $[0, \infty)$. Note that f^{-1} is also a continuous and increasing function for $w > 0$. Therefore, we have

$$\epsilon^2 \geq \frac{a}{2} f^{-1}\left(\frac{2b^2}{a}\right)$$

Observe that the smallest possible value for ϵ would be simply the square root of $af^{-1}\left(2b^2/a\right)/2$. For simplicity, we will obtain a more interpretable expression for ϵ. By the definition of f^{-1}, we have

$$\log(f^{-1}(y)) + f^{-1}(y) = \log(y).$$

Since the condition on a and b enforces $f^{-1}(y)$ to be larger than 1, we obtain the simple inequality that

$$f^{-1}(y) \leq \log(y).$$

Using the above inequality, if ϵ satisfies

$$\epsilon^2 = \frac{a}{2} \log\left(\frac{2b^2}{a}\right),$$

we obtain the desired inequality. ■

References

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., NY, USA, 1995. ISBN 0198538642.

Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, pages 131–151, 1999.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

Charles G Broyden. The convergence of a class of double-rank minimization algorithms 2. the new algorithm. *IMA Journal of Applied Mathematics*, 6(3):222–231, 1970.

Richard H Byrd, Gillian M Chin, Will Neveitt, and Jorge Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.

Richard H Byrd, SL Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science, 2010.

Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. In *Advances in Neural Information Processing Systems 26*, pages 360–368. Curran Associates, Inc., 2013.

David L Donoho, Matan Gavish, and Iain M Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851*, 2013.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul): 2121–2159, 2011.

Murat A Erdogdu. Newton-Stein Method: A Second Order Method for GLMs via Stein's Lemma. In *Advances in Neural Information Processing Systems 28*, pages 1216–1224. Curran Associates, Inc., 2015.

Murat A Erdogdu and Andrea Montanari. Convergence rates of sub-sampled newton methods. In *Advances in Neural Information Processing Systems 28*, pages 3034–3042. Curran Associates, Inc., 2015.

Murat A Erdogdu, Mohsen Bayati, and Lee H Dicker. Scalable approximations for generalized linear problems. *arXiv preprint arXiv:1611.06686*, 2016a.

Murat A Erdogdu, Lee H Dicker, and Mohsen Bayati. Scaled least squares estimator for glms in large-scale problems. In *Advances In Neural Information Processing Systems*, pages 3324–3332, 2016b.

Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13 (3):317–322, 1970.

Michael P Friedlander and Mark Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26, 1970.

Larry Goldstein and Gesine Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7(4):935–952, 1997.

Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Pölsterl, and Alexander Cavallaro. 2d image registration in ct images using radial image descriptors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 607–614. Springer, 2011.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Ritesh Kolte, Murat A Erdogdu, and Ayfer Ozgur. Accelerating svrg via second-order information. In *NIPS Workshop on Optimization for Machine Learning*, 2015.

Nicolas Le Roux and Andrew W Fitzgibbon. A fast natural newton method. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 623–630, 2010.

Nicolas Le Roux, Manzagol Pierre-antoine, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 849–856. Curran Associates, Inc., 2008.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

Moshe Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Chih-Jen Lin, Ruby C Weng, and S Sathiya Keerthi. Trust region newton method for logistic regression. *Journal of Machine Learning Research*, 9(Apr):627–650, 2008.

James Martens. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 735–742, 2010.

Peter McCullagh and John A Nelder. *Generalized linear models*, volume 2. Chapman and Hall London, 1989.

John A Nelder and R. Jacob Baker. *Generalized linear models*. Wiley Online Library, 1972.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/k2). In *Doklady AN SSSR*, volume 269, pages 543–547, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods i: Globally convergent algorithms. *arXiv preprint arXiv:1601.04737*, 2016a.

Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods ii: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016b.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.

Nicol N Schraudolph, Jin Yu, Simon Günter, et al. A stochastic quasi-newton method for online convex optimization. In *International Conference on Artificial Intelligence and Statistics*, volume 7, pages 436–443, 2007.

David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, pages 1135–1151, 1981.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2010.

Oriol Vinyals and Daniel Povey. Krylov subspace descent for deep learning. *arXiv preprint arXiv:1111.4259*, 2011.