Scalable Approximations for Generalized Linear Problems

Murat A. Erdogdu   ERDOGDU@CS.TORONTO.EDU
Department of Computer Science
Department of Statistical Sciences
University of Toronto
Toronto, ON M5S 3G3, Canada

Mohsen Bayati   BAYATI@STANFORD.EDU
Graduate School of Business
Stanford University
Stanford, CA 94305, USA

Lee H. Dicker   LDICKER@STAT.RUTGERS.EDU
Department of Statistics and Biostatistics
Rutgers University
Piscataway, NJ 08854, USA

Editor: Qiang Liu

Abstract

In stochastic optimization, the population risk is generally approximated by the empirical risk which is in turn minimized by an iterative algorithm. However, in the large-scale setting, empirical risk minimization may be computationally restrictive. In this paper, we design an efficient algorithm to approximate the population risk minimizer in generalized linear problems such as binary classification with surrogate losses and generalized linear regression models. We focus on large-scale problems where the iterative minimization of the empirical risk is computationally intractable, i.e., the number of observations n is much larger than the dimension of the parameter p ($n \gg p \gg 1$). We show that under random sub-Gaussian design, the true minimizer of the population risk is approximately proportional to the corresponding ordinary least squares (OLS) estimator. Using this relation, we design an algorithm that achieves the same accuracy as the empirical risk minimizer through iterations that attain up to a quadratic convergence rate, and that are computationally cheaper than any batch optimization algorithm by at least a factor of $\mathcal{O}(p)$. We provide theoretical guarantees for our algorithm, and analyze the convergence behavior in terms of data dimensions. Finally, we demonstrate the performance of our algorithm on well-known classification and regression problems, through extensive numerical studies on large-scale datasets, and show that it achieves the highest performance compared to several other widely used optimization algorithms.

Keywords: Generalized Linear Problems, Stochastic optimization, Subsampling, Dimension reduction in optimization.

1. Introduction

We consider the following stochastic optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ R(\beta) \coloneqq \mathbb{E}\left[\Psi\left(\langle x, \beta \rangle\right) - y\langle x, \beta \rangle\right], \tag{1}$$

where $\Psi : \mathbb{R} \to \mathbb{R}$ is a non-linear function, $y \in \mathcal{Y} \subset \mathbb{R}$ denotes the response variable, $x \in \mathcal{X} \subset \mathbb{R}^p$ denotes the predictor (or covariate), and the expectation is over the joint distribution of (y, x). The above minimization is called a generalized linear problem in its canonical representation, and it is commonly encountered in statistical learning. Celebrated examples include binary classification with smooth surrogate losses (Buja et al., 2005; Reid and Williamson, 2010), and generalized linear models (GLMs) such as Poisson regression, logistic regression, ordinary least squares, multinomial regression with many applications involving graphical models (Nelder and Baker, 1972; McCullagh and Nelder, 1989; Wainwright and Jordan, 2008; Koller and Friedman, 2009). These methods play a crucial role in numerous machine learning and statistics problems, and they provide a miscellaneous framework for many regression and classification tasks.

The exact minimization of the stochastic optimization problem (1) requires the knowledge of the underlying distribution of the variables (y, x). In practice, however, the joint distribution of the pair is not available; therefore, after observing n independent data points (y_i, x_i), the standard approach is to minimize the following surrogate of (1), often referred to as empirical risk approximation

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \widehat{R}(\beta) := \frac{1}{n} \sum_{i=1}^{n} \Psi \left(\langle x_i, \beta \rangle \right) - y_i \langle x_i, \beta \rangle. \tag{2}$$

In the case of GLMs, the empirical risk minimization given in (2) is equivalent to the maximum likelihood estimation (MLE), whereas in the case of binary classification, it is generally referred to as surrogate loss minimization. Due to the non-linear structure of the optimization task given in (2), minimizing the empirical risk requires iterative methods. The first-order approximation of the non-linear risk yields the gradient descent algorithm, which attains a (local) linear convergence rate under certain conditions with $\mathcal{O}(np)$ per-iteration cost. Although its convergence rate is slow compared to that of the second-order methods, its modest per-iteration cost makes it practical for large-scale problems. Regardless of the problem formulation, the most commonly used optimization method for computing the MLE is the Newton-Raphson method, which may be viewed as a reweighted least squares algorithm (McCullagh and Nelder, 1989; Buja et al., 2005). This method uses a second-order approximation to benefit from the curvature of the log-likelihood and achieves locally quadratic convergence. A drawback of this approach is its excessive per-iteration cost of $\mathcal{O}(np^2)$. To remedy this, Hessian-free Krylov sub-space based methods such as conjugate gradient can be used, but the resulting direction is imprecise (Hestenes and Stiefel, 1952; Paige and Saunders, 1975; Martens, 2010). In the regime $n \gg p$, another popular optimization technique is the class of Quasi-Newton methods (Bishop, 1995; Nesterov, 2004), which can attain a per-iteration cost of $\mathcal{O}(np)$, and the convergence rate is locally superlinear; a well-known member of this class of methods is the BFGS algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

In this paper, we take an alternative approach for minimizing (1), based on an identity that is well-known in some areas of statistics, but appears to have received relatively little attention for its computational implications in large-scale problems. Let β^{pop} denote the true minimizer of the population risk given in (1), and let β^{ols} denote the corresponding ordinary least squares (OLS) coefficients defined as $\beta^{\mathsf{ols}} = \mathbb{E}\left[xx^T\right]^{-1} \mathbb{E}\left[xy\right]$. Then, under

certain random predictor (design) models,

$$\beta^{\mathsf{pop}} \propto \beta^{\mathsf{ols}}. \tag{3}$$

For logistic regression with Gaussian design (which is equivalent to Fisher's discriminant analysis), (3) was noted by Fisher in the 1930s (Fisher, 1936); a more general formulation for models with Gaussian design is given in Brillinger (1982). The relationship (3) suggests that if the constant of proportionality is known, then β^{pop} can be estimated by computing the OLS estimator, which may be substantially simpler than minimizing the empirical risk. In fact, in some applications like binary classification, it may not be necessary to find the constant of proportionality in (3). Our work in this paper builds on this idea.

Our contributions can be summarized as follows.

1. We show that β^{pop} is approximately proportional to β^{ols} in the random design setting, regardless of the covariate (predictor) distribution. That is, we prove

$$\left\| \beta^{\mathsf{pop}} - c_\Psi \times \beta^{\mathsf{ols}} \right\|_\infty \lesssim c \frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}},$$

   for some $c_\Psi \in \mathbb{R}$ which depends on the non-linearity Ψ. We note that above rate is the *relative* decay, and it is obtained under the assumption that $\mathbb{E}[\langle x, \beta^{\mathsf{pop}} \rangle^2] \lesssim c^2$. Our generalization uses zero-bias transformations (Goldstein and Reinert, 1997). We also show that the relation (3) still holds under certain types of regularization.

2. We design a computationally efficient estimator for β^{pop} by first estimating the OLS coefficients, and then estimating the proportionality constant c_Ψ via line search. We refer to the resulting estimator as the Scaled Least Squares (SLS) estimator and denote it by $\hat{\beta}^{\mathsf{sls}}$. After estimating the OLS coefficients, the second step of our algorithm involves finding a root of a real valued function; this can be accomplished using iterative methods with up to a quadratic convergence rate and only $\mathcal{O}(n)$ per-iteration cost. This is computationally cheaper than the classical batch methods such as gradient descent by at least a factor of $\mathcal{O}(p)$.

3. For random design with sub-Gaussian predictors and bounded $\Psi^{(2)}$, we show that

$$\left\| \hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}} \right\|_\infty \lesssim \frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}} + \sqrt{\frac{p}{n}}.$$

   This bound characterizes the performance of the proposed estimator in terms of data dimensions, and justifies the use of the algorithm in the large-scale setting where $n \gg p \gg 1$. We also provide theoretical guarantees when subsampling is utilized in the algorithm for further efficiency.

4. We demonstrate how to transform a binary classification problem with smooth surrogate loss into a generalized linear problem, and how our methods can be applied to obtain a computationally efficient optimization scheme.

3

5. We propose a scalable optimization method for converting one generalized linear problem to another by exploiting the proportionality relation (3). The proposed algorithm requires only $\mathcal{O}(n)$ computations per each iteration, with no additional one-time cost. We further discuss the canonicalization of the square loss, which may be of independent interest for non-convex optimization.

6. We empirically study the statistical and computational performance of $\hat{\beta}^{\mathsf{sls}}$, and compare it to that of the empirical risk minimizer (using several well-known implementations), on a variety of large-scale datasets.

The rest of the paper is organized as follows: Section 1.1 surveys the related work and Section 2 introduces the required background and the notation. In Section 3, we provide the intuition behind the relationship (3), which are based on exact calculations for the Gaussian design setting. In Section 4, we propose our algorithm and discuss its computational properties. Theoretical results are given in Section 5. In Section 6, we propose an algorithm to convert one GLM type to another. We discuss how a binary classification problem can be cast as a generalized linear problem in Section 7, and in Section 7.1 we propose a method to canonicalize the square loss. Section 8 provides a thorough comparison between the proposed algorithm and other existing methods. Finally, we conclude with a brief discussion in Section 9.

1.1. Related work

As mentioned in Section 1, the relationship (3) is well-known in several forms in statistics literature. Brillinger (1982) derived (3) for models with Gaussian predictors using Stein's lemma. Li and Duan (1989) studied model misspecification problems in statistics and derived (3) when the predictor distribution has linear conditional means (this is a slight generalization of Gaussian predictors). The relation (3) has led to various techniques for dimension reduction (Li, 1991; Li and Dong, 2009), and more recently, it has been studied by Plan and Vershynin (2016); Thrampoulidis et al. (2015) in the context of compressed sensing. It has been shown that the standard lasso estimator may be very effective when used in models where the relationship between the expected response and the signal is nonlinear, and the predictors (i.e. the design or sensing matrix) are Gaussian. A common theme for all of this previous work is that it focuses solely on settings where (3) holds exactly and the predictors (covariates) are Gaussian (or, in the case of Li and Duan (1989), very nearly Gaussian).

The proportionality relation is solely based on Stein's lemma and its variants. There are recent studies using Stein's lemma in various machine learning and optimization tasks such as second-order optimization (Erdogdu, 2015, 2016, 2017), Bayesian inference (Liu and Wang, 2016), measuring sample quality and model evaluation tasks (Gorham and Mackey, 2015; Liu et al., 2016).

Two key novelties of the present paper are (i) our focus on the computational benefits following from (3) for large scale problems with $n \gg p \gg 1$; and (ii) our rigorous finite sample analysis of models with non-Gaussian predictors, where (3) is shown to be approximately valid. To the best of our knowledge, the present paper and its earlier version Erdogdu et al. (2016) are the first to consider the relation (3) in the context of optimization.

2. Preliminaries and Notation

We assume a random design setting, where the observed data consists of n random iid pairs $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$; $y_i \in \mathcal{Y} \subset \mathbb{R}$ is the response variable and $x_i = (x_{i1}, \cdots, x_{ip})^T \in \mathcal{X} \subset \mathbb{R}^p$ is the vector of predictors or covariates. We focus on problems where the minimization (1) is desirable, but we allow model misspecification, and do not need to assume that (y_i, x_i) are actually drawn from a particular parametric model. In other words, the joint distribution of (y_i, x_i) can be arbitrary.

$$\beta^{\mathsf{pop}} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \ \mathbb{E}\left[\Psi(\langle x_i, \beta \rangle) - y_i \langle x_i, \beta \rangle\right]. \tag{4}$$

While we make no assumptions on Ψ beyond smoothness, note that when the optimization problem is GLM, and Ψ is the cumulant generating function for $y_i \mid x_i$, then the problem reduces to the standard GLM with canonical link and regression parameters β^{pop} (McCullagh and Nelder, 1989). Examples of GLMs in this form include logistic regression with $\Psi(w) = \log\{1 + e^w\}$, Poisson regression with $\Psi(w) = e^w$, and linear regression (least squares) with $\Psi(w) = w^2/2$.

Our objective is to find a computationally efficient estimator for β^{pop}. The alternative estimator for β^{pop} proposed in this paper is related to the OLS coefficient vector, which is defined by $\beta^{\mathsf{ols}} := \mathbb{E}[x_i x_i^T]^{-1} \mathbb{E}[x_i y_i]$; the corresponding OLS estimator is $\hat{\beta}^{\mathsf{ols}} := (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$, where $\mathbf{X} = (x_1, \ldots, x_n)^T$ is the $n \times p$ design matrix and $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$.

Additionally, throughout the text we let $[m] = \{1, 2, ..., m\}$, for positive integers m, and we denote the size of a set S by $|S|$. The m-th derivative of a function $g : \mathbb{R} \to \mathbb{R}$ is denoted by $g^{(m)}$. For a vector $u \in \mathbb{R}^p$ and a $n \times p$ matrix $\mathbf{U}$, we let $\|u\|_q$ and $\|\mathbf{U}\|_q$ denote the ℓ_q-vector and -operator norms, respectively. If $S \subseteq [n]$, let $\mathbf{U}_S$ denote the $|S| \times p$ matrix obtained from $\mathbf{U}$ by extracting the rows that are indexed by S. For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{p \times p}$, $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the maximum and minimum eigenvalues, respectively, and $\rho_q(\mathbf{M})$ for $q \in \{1, 2, \infty\}$ denotes the condition number of $\mathbf{M}$ with respect to ℓ_q-norm. We denote by N_p the p-variate normal distribution, and all expectations are over all randomness inside the brackets. Finally, we use $a \lesssim b$ and $a \leq \mathcal{O}(b)$ interchangeably, whichever is convenient (where $\mathcal{O}(\cdot)$ refers to the big O notation).

3. From OLS to the True Minimizer: Gaussian Case

To motivate our methodology, we assume in this section that the covariates are multivariate normal, as in Brillinger (1982). These distributional assumptions will be relaxed to a certain extent in Section 5.

Proposition 1 *Assume that the covariates are multivariate normal with mean 0 and covariance matrix Σ, i.e. $x_i \sim \mathsf{N}_p(0, \Sigma)$. Then β^{pop} can be written as*

$$\beta^{\mathsf{pop}} = c_\Psi \times \beta^{\mathsf{ols}}, \tag{5}$$

where $c_\Psi \in \mathbb{R}$ is the fixed point of the mapping

$$z \to \mathbb{E}\left[\Psi^{(2)}(\langle x_i, \beta^{\mathsf{ols}} \rangle z)\right]^{-1}. \tag{6}$$

Proof [of Proposition 1] The stationary point of the minimization problem in (4) satisfies the following normal equations,

$$\mathbb{E}\left[y_i x_i\right] = \mathbb{E}\left[x_i \Psi^{(1)}(\langle x_i, \beta \rangle)\right] \quad \text{for} \quad i = 1, 2, ..., p. \tag{7}$$

Now, denote by $\phi(x \mid \boldsymbol{\Sigma})$ the multivariate normal density with mean 0 and covariance matrix $\boldsymbol{\Sigma}$. We recall the well-known property of Gaussian density $\mathrm{d}\phi(x \mid \boldsymbol{\Sigma})/\mathrm{d}x = -\boldsymbol{\Sigma}^{-1} x \phi(x \mid \boldsymbol{\Sigma})$. Using this and the integration by parts on the right hand side of the above equation, we obtain for $i = 1, 2, ..., p$

$$\mathbb{E}\left[x_i \Psi^{(1)}(\langle x_i, \beta \rangle)\right] = \int x \Psi^{(1)}(\langle x, \beta \rangle) \phi(x \mid \boldsymbol{\Sigma}) \, \mathrm{d}x, \tag{8}$$
$$= \boldsymbol{\Sigma}\beta \underbrace{\mathbb{E}\left[\Psi^{(2)}(\langle x_i, \beta \rangle)\right]}_{\in \mathbb{R}},$$

which is basically the Stein's lemma. Combining this with the normal equations (7) and multiplying both side with $\boldsymbol{\Sigma}^{-1}$, we obtain the desired result.

■

Proposition 1 and its proof provide the main intuition behind our proposed method. Observe that in our derivation, we only worked with the right hand side of the normal equations (7) which does not depend on the response variable y_i. Therefore, the equivalence will hold regardless of the joint distribution of (y_i, x_i). This is the main difference from the proof of Brillinger (1982) where y_i is assumed to follow a single index model. In Section 5, where we extend the method to non-Gaussian predictors, the identity (8) is generalized via the zero-bias transformations from Goldstein and Reinert (1997).

3.1. Regularization

A version of Proposition 1 incorporating regularization — an important tool for datasets where p is large relative to n or the predictors are highly collinear — is also possible, as outlined briefly in this section. We focus on ℓ^2-regularization (ridge regression) in this section; some connections with lasso (ℓ^1-regularization) are discussed in Section 5 and Corollary 6.

For $\lambda \geq 0$, define the ℓ_2-regularized empirical risk minimizer,

$$\beta_\lambda^{\mathsf{pop}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \mathbb{E}\left[\Psi(\langle x_i, \beta \rangle) - y_i \langle x_i, \beta \rangle\right] + \frac{\lambda}{2} \|\beta\|_2^2 \tag{9}$$

and the corresponding ℓ^2-regularized OLS coefficients $\beta_\lambda^{\mathsf{ols}} = \left(\mathbb{E}\left[x_i x_i^T\right] + \lambda \mathbf{I}\right)^{-1} \mathbb{E}\left[x_i y_i\right]$ (so $\beta^{\mathsf{pop}} = \beta_0^{\mathsf{pop}}$ and $\beta^{\mathsf{ols}} = \beta_0^{\mathsf{ols}}$). The same argument as above implies that

$$\beta_\lambda^{\mathsf{pop}} = c_\Psi \times \beta_\gamma^{\mathsf{ols}}, \quad \text{where} \quad \gamma = \lambda c_\Psi. \tag{10}$$

This suggests that the ordinary ridge regression for the linear model can be used to estimate the ℓ^2-regularized empirical risk minimizer $\beta_\lambda^{\mathsf{pop}}$. Further pursuing these ideas for problems where regularization is a critical issue may be an interesting area for future research.

Algorithm 1 SLS: Scaled Least Squares Estimator

Input: Data $(y_i, x_i)_{i=1}^n$

Step 1. Compute the least squares estimator: $\hat{\beta}^{\mathsf{ols}}$ and $\hat{y} = \mathbf{X}\hat{\beta}^{\mathsf{ols}}$.

For a sub-sampling based OLS estimator, let $S \subset [n]$ be a random subset and take $\hat{\beta}^{\mathsf{ols}} = \frac{|S|}{n}(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}^Ty$.

Step 2. Solve the following equation for $c \in \mathbb{R}$: $1 = \frac{c}{n}\sum_{i=1}^n \Psi^{(2)}(c\,\hat{y}_i)$.

Use Newton's root-finding method:

Initialize c;

Repeat until convergence:

$$c \leftarrow c - \frac{c\,\frac{1}{n}\sum_{i=1}^n \Psi^{(2)}(c\,\hat{y}_i) - 1}{\frac{1}{n}\sum_{i=1}^n \left\{\Psi^{(2)}(c\,\hat{y}_i) + c\,\hat{y}_i\Psi^{(3)}(c\,\hat{y}_i)\right\}}.$$

Output: $\hat{\beta}^{\mathsf{sls}} = c \times \hat{\beta}^{\mathsf{ols}}$.

---

4. SLS: Scaled Least Squares Estimator

Motivated by the results in the previous section, we design a computationally efficient algorithm that approximates the stochastic optimization problem (1) that is as simple as solving the least squares problem; it is described in Algorithm 1. The algorithm has two basic steps. First, we estimate the OLS coefficients, and then in the second step we estimate the proportionality constant via a simple root-finding algorithm.

There are numerous fast optimization methods to solve the least squares problem, and even a superficial review of these could go beyond the page limits of this paper. We emphasize that this step (finding the OLS estimator) does not have to be iterative and it is the main computational cost of the proposed algorithm. We suggest using a sub-sampling based estimator for β^{ols}, where we only use a subset of the observations to estimate the covariance matrix. Let $S \subset [n]$ be a random sub-sample and denote by $\mathbf{X}_S$ the sub-matrix formed by the rows of $\mathbf{X}$ in S. Then the sub-sampled OLS estimator is given as $\hat{\beta}^{\mathsf{ols}} = \left(\frac{1}{|S|}\mathbf{X}_S^T\mathbf{X}_S\right)^{-1}\frac{1}{n}\mathbf{X}^Ty$. Properties of sub-sampling and sketching based estimators have been well-studied (Vershynin, 2010; Dhillon et al., 2013; Erdogdu and Montanari, 2015; Pilanci and Wainwright, 2015; Roosta-Khorasani and Mahoney, 2016a,b). For sub-Gaussian covariates, it suffices to use a sub-sample size of $\mathcal{O}\left(p\log(p)\right)$ (Vershynin, 2010). Hence, this step requires a single time computational cost of $\mathcal{O}\left(|S|p^2 + p^3 + np\right) \approx \mathcal{O}\left(p\max\{p^2\log(p), n\}\right)$. For other approaches, we refer reader to Rokhlin and Tygert (2008); Drineas et al. (2011); Dhillon et al. (2013); Erdogdu and Montanari (2015) and the references therein.

The second step of Algorithm 1 involves solving a simple root-finding problem. As with the first step, there are numerous methods available for completing this task. Newton's root-finding method with quadratic convergence or Halley's method with cubic convergence may be appropriate choices. We highlight that this step costs only $\mathcal{O}(n)$ per-iteration and that we can attain fast convergence rates of higher order algorithms. The resulting per-iteration cost is cheaper than commonly used batch algorithms by at least a factor of $\mathcal{O}(p)$ — indeed, the cost of computing the gradient is $\mathcal{O}(np)$. For simplicity, we use Newton's root-finding method. We also note that in certain classification problems, only the direction of β^{pop} is needed to make a prediction, so there is no need to compute the scaling constant.
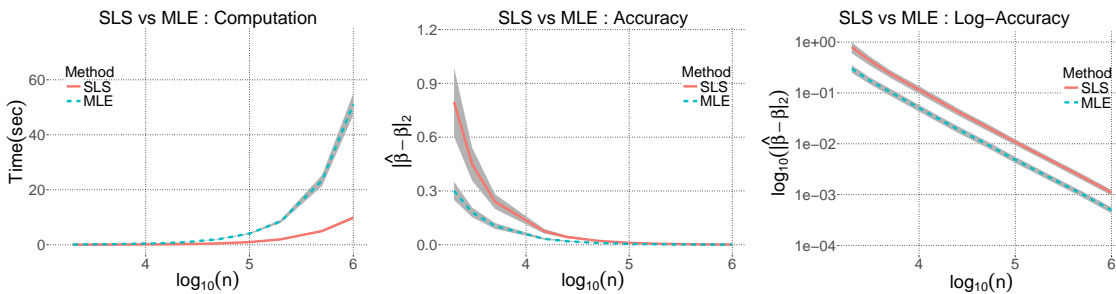
Figure 1: Logistic regression with iid standard Gaussian design. The left plot shows the computational cost (time) for finding the MLE and SLS as n grows and $p = 200$. The middle and the right plots depict the accuracy of the estimators in standard and log scales, respectively. In the regime where the MLE is expensive to compute, the SLS is found much more rapidly and has the same accuracy. `R`'s built-in functions are used to find the MLE.

Correct initialization of the scaling constant c depends on the optimization problem. For example, in the case of GLM problems, assuming that the GLM is a good approximation to the true conditional distribution, by the law of total variance and basic properties of GLMs, we have

$$\text{Var}\left(y_i\right) = \mathbb{E}\left[\text{Var}\left(y_i \mid x_i\right)\right] + \text{Var}\left(\mathbb{E}\left[y_i \mid x_i\right]\right) \approx c_\Psi^{-1} + \text{Var}\left(\Psi^{(1)}(\langle x_i, \beta \rangle)\right). \tag{11}$$

It follows that the initialization $c = 2/\text{Var}\left(y_i\right)$ is reasonable as long as we have $c_\Psi^{-1} \approx \mathbb{E}\left[\text{Var}\left(y_i \mid x_i\right)\right]$ not much smaller than $\text{Var}\left(\Psi^{(1)}(\langle x_i, \beta \rangle)\right)$. Our experiments show that SLS is very robust to initialization.

In Figure 1, we compare the performance of our SLS estimator to that of the MLE in a GLM optimization problem, when both are used to analyze synthetic data generated from a logistic regression model under general Gaussian design with randomly generated covariance matrix. The left plot shows the computational cost of obtaining both estimators as n increases for fixed $p = 200$. The middle and the right plots show the accuracy of the estimators where the latter is in log-scale. In the regime $n \gg p \gg 1$ — where the MLE is hard to compute — the MLE and the SLS achieve the same accuracy, yet SLS has significantly smaller computation time. We refer the reader to Section 5 for theoretical results characterizing the finite sample behavior of the SLS.

5. Theoretical Results

In this section, we use the zero-bias transformations (Goldstein and Reinert, 1997) to generalize the equivalence relation given in the previous section to the settings where the covariates are non-Gaussian.

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

The existence of z^* in Definition 2 is a consequence of Riesz representation theorem (Goldstein and Reinert, 1997). 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^* – which is basically Stein's lemma).

To provide some intuition behind the usefulness of the zero-bias transformation, we refer back to the proof of Proposition 1. For simplicity, assume that the covariate vector x_i has iid entries with mean 0, and variance 1. Then the zero-bias transformation applied to the j-th normal equation in (7) yields

$$\underbrace{\mathbb{E}\left[y_i x_{ij}\right] = \mathbb{E}\left[x_{ij}\Psi^{(1)}\left(x_{ij}\beta_j + \Sigma_{k\neq j}x_{ik}\beta_k\right)\right]}_{j\text{-th normal equation}} = \underbrace{\beta_j\mathbb{E}\left[\Psi^{(2)}\left(x_{ij}^*\beta_j + \Sigma_{k\neq j}x_{ik}\beta_{ik}\right)\right]}_{\text{Zero-bias transformation}}. \qquad (12)$$

The distribution of x_{ij}^* is the x_{ij}-zero-bias distribution and is entirely determined by the distribution of x_{ij}; general properties of x_{ij}^* can be found, for example, in Chen et al. (2010). If β is well spread, it turns out that taken together, with $j = 1, \ldots, p$, the far right-hand side in (12) behaves similar to the right side of (8), with $\boldsymbol{\Sigma} = \mathbf{I}$; that is, the behavior is similar to the Gaussian case, where the proportionality relationship given in Proposition 1 holds. This argument leads to an approximate proportionality relationship for problems with non-Gaussian predictors, which, when carried out rigorously, yields the following result.

Theorem 3 *Suppose that the whitened covariates $w_i = \boldsymbol{\Sigma}^{-1/2}x_i$ are independent random vectors with mean 0, covariance $\mathbf{I}$, and sub-Gaussian norm κ_x. Furthermore, w_i's have constant first and second conditional moments, i.e., $\forall j \in [p]$ and $\tilde{\beta} = \boldsymbol{\Sigma}^{1/2}\beta^{\mathsf{pop}}$, $\mathbb{E}[w_{ij}|\Sigma_{k\neq j}\tilde{\beta}_k w_{ik}]$ and $\mathbb{E}[w_{ij}^2|\Sigma_{k\neq j}\tilde{\beta}_k w_{ik}]$ are deterministic, and the function $\Psi^{(2)}$ is Lipschitz continuous with constant k. Then, for $c_{\Psi} = 1/\mathbb{E}\left[\Psi^{(2)}(\langle x_i, \beta^{\mathsf{pop}}\rangle)\right]$, we have*

$$\left\|\frac{1}{c_{\Psi}}\times\beta^{\mathsf{pop}} - \beta^{\mathsf{ols}}\right\|_{\infty} \leq \eta\|\boldsymbol{\Sigma}^{1/2}\beta^{\mathsf{pop}}\|_{\infty}\|\beta^{\mathsf{pop}}\|_{\infty}, \quad where \quad \eta = 8k\kappa_x^3\rho_{\infty}. \qquad (13)$$

Theorem 3 is proved in Section 10 of the Appendix. It implies that the population parameters β^{ols} and β^{pop} are approximately equivalent up to a scaling factor, with a *relative* error bound of $\mathcal{O}(\|\boldsymbol{\Sigma}^{1/2}\beta^{\mathsf{pop}}\|_{\infty})$. For the above analysis to provide a decreasing bound in the dimension p, one needs a tractability assumption on either one of the terms in the error bound. The most common way to achieve this is to introduce a structure on the inner product between x and β^{pop}, e.g. $\mathbb{E}[\langle x, \beta^{\mathsf{pop}}\rangle^2]^{1/2} = \mathcal{O}\left(\sqrt{p}\right)$ suffices to get a decay in (13). We emphasize that it is common practice to standardize the response and the features before training any model; therefore this preliminary procedure can provide the required order and consequently the desired decay in p.

Below, we discuss two commonly encountered settings in the literature as two corollaries.

Corollary 4 *Let the assumptions of Theorem 3 hold. If we further assume that the covariates x_i are supported on a ball of radius r, and $\Sigma^{1/2}$ is diagonally dominant, we have*

$$\left\|\frac{1}{c_{\Psi}}\times\beta^{\mathsf{pop}} - \beta^{\mathsf{ols}}\right\|_{\infty} \leq \eta\frac{\|\beta^{\mathsf{pop}}\|_{\infty}^2}{\sqrt{p}}, \quad where \quad \eta = 16kr\kappa_x^3\sqrt{\rho_2}\rho_{\infty}. \qquad (14)$$

Note that we do not require an underlying model assumption between y and x. Therefore, one can always perform under the finite support assumption by first scaling the design

matrix appropriately, and then using the proposed estimator. This assumption is common in many fields, yet can appear in different forms. For example in the machine learning literature, the standard convention (the same setting in Corollary 4) is to assume that $\|x_i\|_2 \le r$, and that $\|\beta^{\mathsf{pop}}\|_2 = \mathcal{O}\left(\sqrt{p}\right)$ (Kalai and Sastry, 2009; Shalev-Shwartz et al., 2009; Kakade et al., 2011), or the inner product is supported on an interval $|\langle x, \beta^{\mathsf{pop}}\rangle| \le b$ almost surely (Alquier and Biau, 2013), where the latter is a more strict assumption. In the compressed sensing and other areas in statistics, a similar assumption is imposed on the covariance matrix $\mathbb{E}\left[xx^T\right] = \tilde{\boldsymbol{\Sigma}}/p$ which provides us with $\mathbb{E}[\|x\|_2^2] = \mathcal{O}(1)$ (Donoho et al., 2009; Bayati and Montanari, 2012; Bayati et al., 2013; Su and Candes, 2016).

On the other hand, the standard convention in the dimension reduction techniques is to assume that the covariates have norm of order $\sqrt{p}$, i.e. $\mathbb{E}\left[\|x\|_2\right] = \mathcal{O}\left(\sqrt{p}\right)$, and the coefficients have norm of order 1, i.e. $\|\beta\|_2 = \tau = \mathcal{O}(1)$ (Duan and Li, 1991; Hall and Li, 1993; Hristache et al., 2001). This setting again yields that the inner product between those two vectors are tractable and provides us with a decay in the latter term, i.e. $\|\beta^{\mathsf{pop}}\|_\infty = \mathcal{O}\left(1/\sqrt{p}\right)$. In this setting, Theorem 3 yields the following corollary.

Corollary 5 *Let the assumptions of Theorem 3 hold. If we further assume that β^{pop} is r-well-spread in the sense that for $\|\beta^{\mathsf{pop}}\|_2 = \tau$, we have $\tau/\|\beta^{\mathsf{pop}}\|_\infty = r\sqrt{p}$ for some $r \le 1$,*

$$\left\|\frac{1}{c_\Psi} \times \beta^{\mathsf{pop}} - \beta^{\mathsf{ols}}\right\|_\infty \le \eta\frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}}, \quad where \ \ \eta = 8k\kappa_x^3\rho_\infty(\tau/r)\|\boldsymbol{\Sigma}^{1/2}\|_\infty. \tag{15}$$

In the above bound, the decay is in fact $1/p$, but we chose to state the *relative decay*. We emphasize that both settings in Corollaries 4 and 5 are equivalent and both assumptions lead a tractable inner product $\langle x, \beta^{\mathsf{pop}}\rangle$. The constants κ_x, ρ_k are invariant to the scalings considered in both settings.

The assumption that β^{pop} is well-spread can be relaxed with minor modifications. For example, if we have a sparse coefficient vector, where $\mathrm{supp}(\beta^{\mathsf{pop}}) = \{j; \ \beta_j^{\mathsf{pop}} \ne 0\}$ is the support set of β^{pop}, then Corollary 5 holds with p replaced by the size of the support set.

The assumptions on the conditional moments are the relaxed versions of assumptions that are commonly encountered in dimension reduction techniques. For example, sliced inverse regression methods assume that the first conditional moment $\mathbb{E}\left[x\big|\langle x, \beta\rangle\right]$ is linear in x for all β (Li and Duan, 1989; Li, 1991), which is satisfied by elliptically distributed random vectors. An important case that is not covered by these methods is the independent coordinate case, i.e., when the whitened covariates have independent, but not necessarily identical entries. This is made possible by introducing additional approximation error (i.e. suboptimality due to non-Gaussian design). We refer reader to Li and Dong (2009), for a good review of dimension reduction techniques and their corresponding assumptions. We also highlight that our moment assumptions can be relaxed further, at the expense of introducing some additional complexity into the results.

An interesting consequence of Theorem 3 and the corollaries following the theorem is that whenever an entry of β^{pop} is zero, the corresponding entry of β^{ols} has to be small, and conversely. For $\lambda \ge 0$, define the lasso coefficients

$$\beta_\lambda^{\mathsf{lasso}} = \underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}} \ \frac{1}{2}\mathbb{E}\left[(y_i - \langle x_i, \beta\rangle)^2\right] + \lambda\|\beta\|_1. \tag{16}$$

Corollary 6 *Let* $\text{supp}(\beta)$ *denote the support of* β, *i.e., the set* $\{i \in [p] : \beta_i \neq 0\}$. *Given* $\|\frac{1}{c_\Psi}\beta^{\text{pop}} - \beta^{\text{ols}}\|_\infty \leq \delta$, *and for any* $\lambda \geq \delta$, *if* $\mathbb{E}[x_i] = 0$ *and* $\mathbb{E}[x_i x_i^T] = \mathbf{I}$, *we have*

$$\text{supp}(\beta^{\text{lasso}}) \subset \text{supp}(\beta^{\text{pop}}).$$

Further, if λ *and* β^{pop} *also satisfy that* $\forall j \in \text{supp}(\beta^{\text{pop}})$, $|\beta_j^{\text{pop}}| > c_\Psi(\lambda + \delta)$, *then we have* $\text{supp}(\beta^{\text{lasso}}) = \text{supp}(\beta^{\text{pop}})$.

We note that δ above can be obtained using Theorem 3 or its corollaries. So far in this section, we have only discussed properties of the population parameters, such as β^{pop} and β^{ols}. In the remainder of this section, we turn our attention to the results on the estimators that are the main focus of this paper; these results ultimately build on our earlier results, i.e. Theorem 3. We will be considering the setting in Corollary 5 in the rest of this section, i.e., $\|x\|_2 = \mathcal{O}\left(\sqrt{p}\right)$ and $\|\beta^{\text{pop}}\|_2 = \mathcal{O}(1)$. Results presented below can be easily generalized to the setting in Corollary 4 as well.

In order to precisely describe the performance of $\hat{\beta}^{\text{sls}}$, we first need bounds on the OLS estimator. The OLS estimator has been studied extensively in the literature; however, for our purposes, we find it convenient to derive a new bound on its accuracy. While we have not seen this exact bound elsewhere, it is very similar to Theorem 5 of Dhillon et al. (2013).

Proposition 7 *Assume that* $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^T] = \Sigma$, *and that* $\Sigma^{-1/2}x_i$ *and* y_i *are sub-Gaussian with norms* κ *and* γ, *respectively. For* λ_{min} *denoting the smallest eigenvalue of* Σ, *and* $|S| > \eta p$,

$$\left\|\hat{\beta}^{\text{ols}} - \beta^{\text{ols}}\right\|_2 \leq \eta \lambda_{min}^{-1/2}\sqrt{\frac{p}{|S|}}, \tag{17}$$

with probability at least $1 - 3e^{-p}$, *where* η *depends only on* γ *and* κ.

Proposition 7 is proved in Section 10. Our main result on the performance of $\hat{\beta}^{\text{sls}}$ is given next.

Theorem 8 *Let the assumptions of Theorem 3, Corollary 5, and Proposition 7 hold. Further assume that* $\left|\Psi^{(2)}\right| \leq b$ *and for some* $\bar{c}$ *and* ϵ *the function* $f(z) = z\mathbb{E}\left[\Psi^{(2)}(\langle x, \beta^{\text{ols}}\rangle z)\right]$ *satisfies* $f(\bar{c}) > 1 + \epsilon$, *and the derivative of* f *in the interval* $[0, \bar{c}]$ *does not change sign, i.e., its absolute value is lower bounded by* $\upsilon > 0$. *Then, for* n *and* $|S|$ *sufficiently large, with probability at least* $1 - 4e^{-p}$, *we have*

$$\left\|\hat{\beta}^{\text{sls}} - \beta^{\text{pop}}\right\|_\infty \leq \eta_1 \frac{\|\beta^{\text{pop}}\|_\infty}{\sqrt{p}} + \eta_2\sqrt{\frac{p}{|S|}}, \tag{18}$$

where the constants η_1 *and* η_2 *are defined by*

$$\eta_1 = \eta k \bar{c} \rho_\infty(\tau/r)\|\Sigma^{1/2}\|_\infty \tag{19}$$

$$\eta_2 = \eta\bar{c}\left(\lambda_{min}^{-1/2} + k\upsilon^{-1}\max\left\{(\|\Sigma^{1/2}\beta^{\text{ols}}\|_2 + 1)\|\Sigma\|_2 + b/k, \bar{c}\right\}\right), \tag{20}$$

and $\eta > 0$ *is a constant depending on* κ *and* γ.

Note that the convergence rate of the upper bound in (18) depends on the sum of the two terms, both of which are functions of the data dimensions n and p. The first term on the right in (18) comes from Theorem 3 and Corollary 5, which bounds the discrepancy between $c_\Psi \times \beta^{\mathsf{ols}}$ and β^{pop}. This term is small when p is large, and it does not depend on the number of observations n. If n and p grow together, then p has to grow at a smaller rate to achieve convergence. We note that the actual decay of this term is $\mathcal{O}\left(1/p\right)$ (since $\|\beta^{\mathsf{pop}}\|_\infty = \mathcal{O}\left(1/\sqrt{p}\right)$), but we state the relative decay which is $\mathcal{O}\left(1/\sqrt{p}\right)$.

The second term in the upper bound (18) comes from estimating β^{ols} and c_Ψ. This term is increasing in p, which reflects the fact that estimating β^{pop} is more challenging when p is large. As expected, this term is decreasing in n and $|S|$, i.e. larger sample size yields better estimates. When the full OLS solution is used ($|S| = n$), the second term becomes $\mathcal{O}(\sqrt{p/n})$, which suggests that the sample size should be at least of order p for good performance. We note that in the case of Gaussian covariates, $\sqrt{n}$-consistency follows since we don not have a bias term.

In certain loss functions, the uniform boundedness assumption on $\Psi^{(2)}$ may be too strict. However, this assumption can be removed by making an additional assumption that the covariates have bounded support. This yields the same decay rate $\mathcal{O}\left(\sqrt{p/n}\right)$ in the second term. However, if one needs to preserve the sub-Gaussian assumption on the covariates, then the above analysis – when carried out without the uniform bound assumption on $\Psi^{(2)}$ – yields the following result. We note that we still require the uniform Lipschitz assumption but one can always utilize the local Lipschitz properties of the function $\Psi^{(2)}$.

Theorem 9 *Let the assumptions of Theorem 3, Corollary 5, and Proposition 7 hold with $\mathbb{E}[\|\mathbf{\Sigma}^{-1/2}x\|_2] = \tilde{\mu}\sqrt{p}$. Further assume that $\sup_{\beta:\|\beta-\Sigma^{1/2}\beta^{\mathsf{ols}}\|_2\leq 1} \|\Psi^{(2)}(\langle x_i, \beta\rangle)\|_{\psi_2} \leq \kappa_g$ and for some $\bar{c}$ and ϵ the function $f(z) = z\mathbb{E}\left[\Psi^{(2)}(\langle x, \beta^{\mathsf{ols}}\rangle z)\right]$ satisfies $f(\bar{c}) > 1 + \epsilon$, and the derivative of f in the interval $[0, \bar{c}]$ does not change sign, i.e., its absolute value is lower bounded by $\upsilon > 0$. Then, for n and $|S|$ sufficiently large, with probability at least $1 - 5e^{-p}$, we have*

$$\left\|\hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}}\right\|_\infty \leq \eta_1 \frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}} + \eta_2 \sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}, \tag{21}$$

where the constants η_1 and η_2 are defined by

$$\eta_1 = \eta k\bar{c}\rho_\infty(\tau/r)\|\mathbf{\Sigma}^{1/2}\|_\infty \tag{22}$$

$$\eta_2 = \eta\bar{c}\left(\lambda_{min}^{1/2} + \upsilon^{-1}\|\beta^{\mathsf{ols}}\|_\infty \max\left\{(\kappa_g + k/\tilde{\mu}), k\bar{c}\kappa_x\right\}\right), \tag{23}$$

and $\eta > 0$ is a constant depending on κ_x and γ.

We observe that this time when the full OLS solution is used ($|S| = n$), the second term becomes $\mathcal{O}(\sqrt{p\log(n)/n})$, which is worse than the rate given in Theorem 8 by a logarithmic factor. In this setting, we require that $n/\log(n)$ should be at least of order p for good performance. Also, note that there is a theoretical threshold for the sub-sampling size $|S|$, namely $\mathcal{O}\left(n/\log(n)\right)$, beyond which further sub-sampling provides no improvement.

When the covariates are Gaussian, the Lipschitz assumption on the function $\Psi^{(2)}$ is enough the guarantee the sub-Gaussianity of $\Psi^{(2)}(\langle x_i, \beta\rangle)$. However, we note that the condition given in the Theorem is more general.

5.1. Convergence in ℓ_2-norm

In the sequel, we assume that the full OLS solution is used ($|S| = n$). Using Proposition 7 and Lemma 16, we can show that the ℓ_2-distance between the estimator $\hat{\beta}^{\mathsf{sls}}$ and the population parameter $c_\Psi \beta^{\mathsf{ols}}$ converges to 0 with rate $\sqrt{n}$ when p is fixed. It is straightforward to see that under Gaussian design, this leads to consistency in ℓ_2-norm with convergence rate $\sqrt{n}$. However, the quantity $c_\Psi \beta^{\mathsf{ols}}$ is in general different from β^{pop} under general sub-Gaussian design; thus, in such cases consistency cannot be achieved in the classical setting where p is fixed (see i.e. Claim 1 in Chen and Banerjee (2017)). In such a scenario, Theorem 3 can be used to obtain an upper bound on the suboptimality resulting from general sub-Gaussian design.

In order to talk about consistency when $n, p \to \infty$ under ℓ_2-norm, a normalization by one of the growing dimensions is required, see e.g. Bayati and Montanari (2012); Bayati et al. (2013). Hence, in this regime, we focus on the quantity $\frac{1}{\sqrt{p}}\|\hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}}\|_2$ under the assumptions of Corollary 4 with $\mathbf{\Sigma} = \mathbf{I}$. Since we have $\sqrt{n}$-convergence of our estimator to the population parameter $c_\Psi \beta^{\mathsf{ols}}$, consistency reduces to showing that the suboptimality term $\frac{1}{\sqrt{p}}\left\|c_\Psi \beta^{\mathsf{ols}} - \beta^{\mathsf{pop}}\right\|_2$ converges to 0. This term can be controlled with additional sparsity-type structural assumptions on the problem. Below, we discuss two scenarios that may lead to such a setting.

Sparsity. Assume that β^{pop} has s non-zero terms, i.e., $\|\beta^{\mathsf{pop}}\|_0 = s$. Following the same proof technique of Theorem 3, we can show that

$$\frac{1}{\sqrt{p}}\left\|c_\Psi \beta^{\mathsf{ols}} - \beta^{\mathsf{pop}}\right\|_2 \leq C\sqrt{\frac{s}{p}}\|\beta^{\mathsf{pop}}\|_\infty, \tag{24}$$

for some constant $C > 0$. As long as the non-sparsity level satisfies $s = \mathcal{O}(p)$, we can attain consistency in ℓ_2-norm. The rate of convergence will be determined by the growing rates of $n, p(n)$ and $s(n)$. For example, to attain $\sqrt{n}$-convergence, one needs $s = \mathcal{O}(1)$ and $p = \mathcal{O}(n)$, but milder growth conditions can still lead to consistency at the expense of slower convergence.

Normality. Similarly, one can assume that a subset of features with size s follow a non-Gaussian distribution, and the remaining $p - s$ features are normally distributed. The same argument above can also provides consistency in ℓ_2-norm – again in this case and the convergence rate will be determined by the ratio $\sqrt{s/p}$.

5.2. Verifying the key assumptions

The assumption on $f(z) = z\mathbb{E}\left[\Psi^{(2)}(\langle x, \beta^{\mathsf{ols}}\rangle, z)\right]$ appears in both theorems and depends both on the function Ψ and the underlying covariate distribution. Therefore, in order to verify this assumption rigorously, one needs to know the distribution of the covariates which is not always available. However, a straightforward way to verify this assumption is to use the sample mean estimate $\hat{f}(z) = z\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i, \hat{\beta}^{\mathsf{ols}}\rangle z)$ using the dataset at hand to estimate $f(z)$. In Figure 2, we use this technique to verify that our assumptions are valid for two commonly used loss functions, namely, boosting and logistic regression when the covariates are normally distributed $x \sim \mathsf{N}(0, \frac{1}{p}\mathbf{I})$ and $\|\beta^{\mathsf{ols}}\|_2 = \sqrt{p}/4$. We observe through the figure that for both loss functions, one can easily verify that the conditions are satisfied.
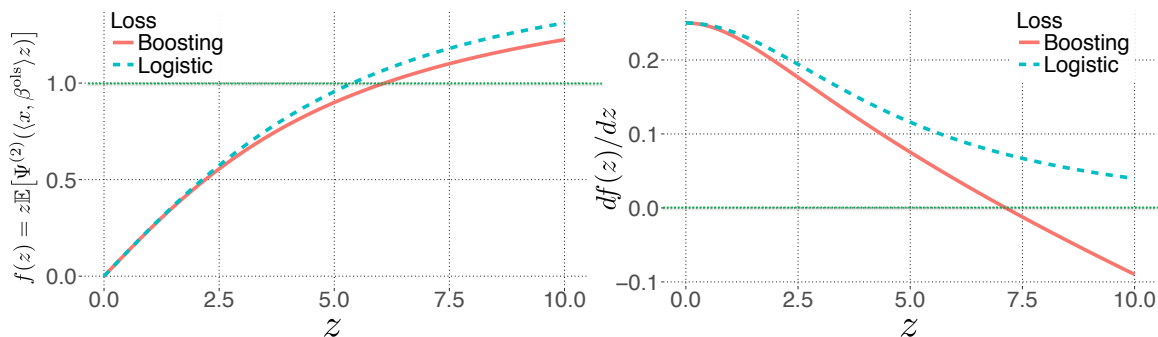
Figure 2: We demonstrate how to validate our assumption on $f(z) = z\mathbb{E}\left[\Psi^{(2)}(\langle x, \beta^{\mathsf{ols}}\rangle z)\right]$ on two commonly used loss functions. Left plot shows $f(z)$, whereas right plots shows its derivative. In these plots, covariates are generated from $x \sim \mathsf{N}(0, \mathbf{I}/p)$, and $\|\beta^{\mathsf{ols}}\|_2 = \sqrt{p}/4$.

The assumption on $f(z)$ can be verified by a more rigorous treatment. For simplicity, we again consider the compressed sensing setting where we assume $x \sim \mathsf{N}(0, \frac{1}{p}\mathbf{I})$, and this time we let $\|\beta^{\mathsf{pop}}\|_2 = c_\Psi \|\beta^{\mathsf{ols}}\|_2 \approx \sqrt{p}$. In order to obtain a scaling constant $c_\Psi \leq 20$, assume that we have $\|\beta^{\mathsf{ols}}\|_2 = \sqrt{p}/20$.

Boosting loss: In this case, we have $\Psi(z) = z/2 + \sqrt{1 + z^2/4}$ which yields

$$\Psi^{(2)}(z) = \frac{1}{4}(1 + z^2/4)^{-3/2} \quad \text{and} \quad \Psi^{(4)}(z) = \frac{3}{64}\frac{5z^2(1 + z^2/4)^{-2} - 4}{(1 + z^2/4)^{5/2}}. \tag{25}$$

We notice that both of these functions are even, i.e., $\Psi^{(2)}(z) = \Psi^{(2)}(-z)$, and $\Psi^{(4)}(z) = \Psi^{(4)}(-z)$. We use the local convexity for $z \geq 0$ around $z = 2$, and obtain $\Psi^{(2)}(z) \geq a - bz$ where $a = \Psi^{(2)}(2) - 2\Psi^{(3)}(2)$ and $b = -\Psi^{(3)}(2)$. For $W \sim \mathsf{N}(0, 1)$ and ϕ denoting the standard normal density, we write

$$f(z) = z\mathbb{E}\left[\Psi^{(2)}(\langle x, \beta^{\mathsf{ols}}\rangle z)\right] = z\mathbb{E}\left[\Psi^{(2)}(Wz/20)\right], \tag{26}$$

$$= 2z\int_0^\infty \Psi^{(2)}(wz/20)\phi(w)dw \geq 2z\int_0^{20a/bz}(a - bwz/20)\phi(w)dw,$$

$$= 2z\left\{a\Phi(20a/bz) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1 - e^{-200a^2/b^2z^2})\right\}.$$

Using the above bound, we observe that for $\bar{c} = 6$ and $\delta = .23$, we have $f(\bar{c}) > 1 + \delta$.

Next, we verify the derivative condition on f, that is, $f^{(1)}$ does not change sign in the interval $[0, \bar{c}]$. By the Stein's lemma, we write

$$f^{(1)}(z) = \mathbb{E}\left[\Psi^{(2)}(Wz/20)\right] + (z/20)^2\mathbb{E}\left[\Psi^{(4)}(Wz/20)\right]. \tag{27}$$

Using the previous inequalities together with the above expression, we have

$$f^{(1)}(z) \geq \mathbb{E}\left[\Psi^{(2)}(Wz/20)\right] - \frac{9}{100}\left|\Psi^{(4)}\right| \tag{28}$$

$$\geq 2\left\{a\Phi(20a/bz) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1 - e^{-200a^2/b^2z^2})\right\} - \frac{27}{1600} \geq 0.188,$$

14

which verifies the derivative assumption on $f(z)$.

Logistic regression: In this case, we have $\Psi(z) = \log(1 + e^z)$ which yields

$$\Psi^{(2)}(z) = \frac{e^z}{(1 + e^z)^2} \quad \text{and} \quad \Psi^{(4)}(z) = \frac{e^z(1 - 4e^z + e^{2z})}{(1 + e^z)^4}. \tag{29}$$

As before, both $\Psi^{(2)}$ and $\Psi^{(4)}$ are even functions. Again, appealing to the local convexity properties, for $z \geq 0$ around $z = 2.5$, we obtain $\Psi^{(2)}(z) \geq a - bz$ where $a = \Psi^{(2)}(2.5) - 2.5\Psi^{(3)}(2.5)$ and $b = -\Psi^{(3)}(2.5)$. Using (26), we obtain

$$f(z) \geq 2z \left\{ a\Phi(20a/bz) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1 - e^{-200a^2/b^2z^2}) \right\}. \tag{30}$$

Using the above bound, we see that for $\bar{c} = 6$, and $\delta = .22$ we have $f(\bar{c}) > 1 + \delta$.

For the derivative bound, we again use (27), and obtain that for $z \in [0, \bar{c}]$,

$$f^{(1)}(z) \geq \mathbb{E}\left[\Psi^{(2)}(Wz/20)\right] - \frac{9}{100}\left|\Psi^{(4)}\right| \geq \mathbb{E}\left[\Psi^{(2)}(Wz/20)\right] - \frac{9}{800} \geq 0.19, \tag{31}$$

which verifies the assumption.

6. Converting One GLM to Another

In this section, we assume that one has already obtained a set of coefficients by minimizing a certain GLM loss, but wants to minimize another GLM loss using the same dataset. It is often the case that a practitioner would like to change the loss function (or equivalently the model) based on its performance. However, when the dataset is large, training a new model from scratch is extremely time consuming. In the following, we will use the proportionality relation to transition between different loss functions.

Assume that a practitioner fitted a GLM using the loss function (or cumulant generating function) Ψ_1, but they would like to train a new model based on the second loss function Ψ_2. Instead of minimizing the new loss based on Ψ_2, one can exploit the proportionality relation and obtain the coefficients for the new GLM problem. Denote by β_1^{pop} and β_2^{pop} the GLM coefficients corresponding to the loss functions Ψ_1 and Ψ_2, respectively. Using the results in the previous sections, we have

$$\frac{1}{c_{\Psi_1}}\beta_1^{\mathsf{pop}} = \frac{1}{c_{\Psi_2}}\beta_2^{\mathsf{pop}} = \beta^{\mathsf{ols}},$$

for the normal case, that is, both coefficients are proportional to the OLS coefficients which does not depend on the loss function. Therefore, these coefficients β_1^{pop} and β_2^{pop} are also proportional to each other and we can write

$$\beta_2^{\mathsf{pop}} = \frac{c_{\Psi_2}}{c_{\Psi_1}} \beta_1^{\mathsf{pop}} := \rho \, \beta_1^{\mathsf{pop}}, \tag{32}$$

where the proportionality constant between two different GLM types turns out to be the ratio between c_{Ψ_1} and c_{Ψ_2}, i.e. $\rho = c_{\Psi_2}/c_{\Psi_1}$. Using the definition of c_{Ψ_2}, we write

$$1 = c_{\Psi_2} \, \mathbb{E}\left[\Psi_2^{(2)}(\langle x, \beta_2^{\mathsf{pop}}\rangle)\right],$$
$$= c_{\Psi_1}\rho \, \mathbb{E}\left[\Psi_2^{(2)}(\langle x, \beta_1^{\mathsf{pop}}\rangle\rho)\right].$$

15

Algorithm 2 Conversion from one GLM to another

Input: Data $(y_i, x_i)_{i=1}^n$, and $\hat{\beta}_1^{\text{glm}}$

Step 1. Compute $\hat{y} = \mathbf{X}\hat{\beta}_1^{\text{glm}}$, and $\kappa = \frac{1}{n}\sum_{i=1}^n \Psi_1^{(2)}(\hat{y}_i)$.

Step 2. Solve the following equation for $\rho \in \mathbb{R}$: $\kappa = \frac{\rho}{n}\sum_{i=1}^n \Psi_2^{(2)}(\hat{y}_i\rho)$
    Use Newton's root-finding method:
      Initialize $\rho = 1$;
      Repeat until convergence:

$$\rho \leftarrow \rho - \frac{\rho\frac{1}{n}\sum_{i=1}^n \Psi_2^{(2)}(\rho\,\hat{y}_i) - \kappa}{\frac{1}{n}\sum_{i=1}^n \left\{ \Psi_2^{(2)}(\rho\,\hat{y}_i) + \rho\,\hat{y}_i\Psi_2^{(3)}(\rho\,\hat{y}_i) \right\}}.$$

Output: $\hat{\beta}_2^{\text{glm}} = \rho \times \hat{\beta}_1^{\text{glm}}$.

Dividing the both sides by c_{Ψ_1} and using the equality $c_{\Psi_1}^{-1} = \mathbb{E}\left[\Psi_1^{(2)}(\langle x, \beta_1^{\text{pop}}\rangle) \right]$, we obtain

$$\mathbb{E}\left[\Psi_1^{(2)}(\langle x, \beta_1^{\text{pop}}\rangle) \right] = \rho\,\mathbb{E}\left[\Psi_2^{(2)}(\langle x, \beta_1^{\text{pop}}\rangle\rho) \right].$$

The above equation only involves β_1^{pop} as the coefficients (which is assumed to be already known/obtained by the practitioner). Therefore, if we solve it for the ratio ρ, we can estimate β_2^{pop} by simply using the proportionality relation given in (32).

The procedure described above is summarized as Algorithm 2. We emphasize that this procedure does not require the computation of the OLS estimator which was the main cost of SLS, and it only requires a per-iteration cost of $\mathcal{O}(n)$. In other words, conversion from one GLM type to another is easier compared to obtaining the GLM coefficients from scratch.

7. Binary Classification with Proper Scoring Rules

In this section, we assume that for $i \in [n]$, the response is binary $y_i \in \{0, 1\}$. The binary classification problem can be described by the following minimization of an empirical risk

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \; \frac{1}{n}\sum_{i=1}^n \ell(y_i; q(\langle x_i, \beta\rangle)), \tag{33}$$

where ℓ and q are referred to as the loss and the link functions, respectively. There are various loss functions that are used in practice. Examples include log-loss, boosting loss, square loss etc (See Table 1). As before, we constrain our analysis on the canonical links. The concept of canonical links for binary classification is introduced by Buja et al. (2005), and it is quite similar to the generalized linear problems.

For any given loss function, we define the partial losses $\ell_k(\cdot) = \ell(y = k; \cdot)$ for $k \in \{0, 1\}$. Since we have a binary response variable, we can write any loss in the following format

$$\begin{aligned} \ell(y; q) &= y\ell_1(q) + (1 - y)\ell_0(q), \\ &= y\left(\ell_1(q) - \ell_0(q)\right) + \ell_0(q). \end{aligned}$$

Table 1: Common loss functions and their inverse canonical links

NAME	LOSS FUNCTION: $\ell(y; q)$	WEIGHT: $w(q)$	INVERSE CANONICAL LINK: q
LOG-LOSS	$-y \log(q) - (1-y) \log(1-q)$	$\frac{1}{q(1-q)}$	$\frac{1}{1+\exp(-z)}$
BOOSTING LOSS	$y(q^{-1}-1)^{1/2} + (1-y)(q^{-1}-1)^{-1/2}$	$\frac{1}{[q(1-q)]^{3/2}}$	$\frac{1}{2} + \frac{z/4}{2(z^2/4+1)^{1/2}}$
SQUARE LOSS	$y(1-q)^2 + (1-y)q^2$	1	$\frac{1+z}{2}$

The above formulation is of the form of a generalized linear problem. Before moving forward, we recall the concept of proper scoring in binary classification, which is sometimes referred to as Fisher consistency.

Definition 10 (Proper scoring rules) *Assume that $y \sim Bernoulli(\eta)$. If the expected loss $\mathbb{E}\left[\ell(y, q)\right]$ is minimized by $q = \eta$ for all $\eta \in (0, 1)$, we call the loss function a proper scoring rule.*

The following theorem by Schervish (1989) provides a methodology for constructing a loss function for the proper scoring rules.

Theorem 11 (Schervish (1989)) *Let $w(dt)$ be a positive measure on $(0, 1)$ that is finite on interval $(\epsilon, 1 - \epsilon) \ \forall \epsilon > 0$. Then the following defines a proper scoring rule*

$$\ell_1(q) = \int_q^1 (1 - t)w(dt), \ and \ \ell_0(q) = \int_0^q tw(dt).$$

The measure $w(dt)$ uniquely defines the loss function (generally referred to as the weight function, since all losses can be written as weighted average of cost weighted misclassification error (Buja et al., 2005; Reid and Williamson, 2010)). Examples of weight functions is given in Table 1. The above theorem has many interesting interpretations; one that is most useful to us is that $\ell_0^{(1)}(q) = qw(q)$.

The notion of canonical links for proper scoring rules are introduced by Buja et al. (2005), which corresponds to the notion of matching loss (Helmbold et al., 1999; Reid and Williamson, 2010). The derivation of canonical links stems from the Hessian of the above minimization, which remedies two potential problems: non-convexity and asymptotic variance inflation. It turns out that by setting $w(q)q^{(1)}$ as constant, one can remedy both problems (Buja et al., 2005). We will skip the derivation and, without loss of generality, assume that the canonical link-loss pair satisfies $w(q)q^{(1)} = 1$. Note that any loss function has a natural canonical link. The following Theorem summarizes this concept.

Theorem 12 (Buja et al. (2005)) *For proper scoring rules with $w > 0$, there exists a canonical link function which is unique up to addition and multiplication by constants. Conversely, any link function is canonical for a unique proper scoring rule.*

The canonical link for a given loss function can be explicitly derived from the equation $w(q)q^{(1)} = 1$. We have provided some examples in Table 1. Using the definition of canonical

link for proper scoring rules, we write the normal equations $\frac{d}{d\beta}\mathbb{E}\left[\ell(y, q(\langle x, \beta \rangle))\right] = 0$ as

$$
\mathbb{E}\left[xq^{(1)}(\langle x, \beta \rangle)\ell_0^{(1)}(q(\langle x, \beta \rangle))\right] = \mathbb{E}\left[yxq^{(1)}(\langle x, \beta \rangle)\left(\ell_0^{(1)}(q(\langle x, \beta \rangle)) - \ell_1^{(1)}(q(\langle x, \beta \rangle))\right)\right],
$$

$$
\implies \mathbb{E}\left[xq^{(1)}(\langle x, \beta \rangle)q(\langle x, \beta \rangle)w(q(\langle x, \beta \rangle))\right] = \mathbb{E}\left[yxq^{(1)}(\langle x, \beta \rangle)w(q(\langle x, \beta \rangle))\right],
$$

$$
\implies \mathbb{E}\left[xq(\langle x, \beta \rangle)\right] = \mathbb{E}\left[yx\right],
$$

$$
\implies \boldsymbol{\Sigma}\beta\mathbb{E}\left[q^{(1)}(\langle x, \beta \rangle)\right] = \mathbb{E}\left[yx\right].
$$

The last equation provides us with the analog of the proportionality relation we observed in generalized linear problems. In this case, we observe that the proportionality constant becomes $1/\mathbb{E}\left[q^{(1)}(\langle x, \beta \rangle)\right]$. Therefore, our algorithm can be used to obtain a fast training procedure for the binary classification problems under canonical links as well.

7.1. Canonicalization of the Square Loss

Consider a minimization problem of the following form

$$
\underset{\beta}{\text{minimize}} \; \frac{1}{n}\sum_{i=1}^{n}[y_i - f(\langle x_i, \beta \rangle)]^2. \tag{34}
$$

The above problem is commonly encountered in many machine learning tasks – specifically, in the context of neural networks, the function f is called the activation function. Here, we consider a toy example to demonstrate how our methodology can be useful in a minimization problem of the above form.

We first use Taylor series expansion around a point θ (which should be close to $\langle x, \beta \rangle$), in order to approximate the function $f(z)$ with a linear function around $f(\theta)$. We write

$$
\min_{\beta} \mathbb{E}\left[(y - f(\langle x, \beta \rangle))^2\right] \approx \min_{\beta} \mathbb{E}\left[f(\langle x, \beta \rangle)^2 - 2y\langle x, \beta \rangle f'(\theta)\right]. \tag{35}
$$

The above approximation yields

$$
\Psi(z) = \frac{f(z)^2}{2f'(\theta)}, \tag{36}
$$

and the proportionality relation given in previous sections would hold for the approximation, which will be accurate when the activation function is smooth around the user-specified point θ. This method can be used to derive proportionality relations for GLMs with non-canonical links (conditional on the link being nice), and also may be of interest in non-convex optimization, e.g., reliable initialization.

8. Experiments

8.1. Numerical Analysis of Convergence

This section contains the results of numerical studies that demonstrate the convergence results presented in Section 5. We first demonstrate the convergence of the population
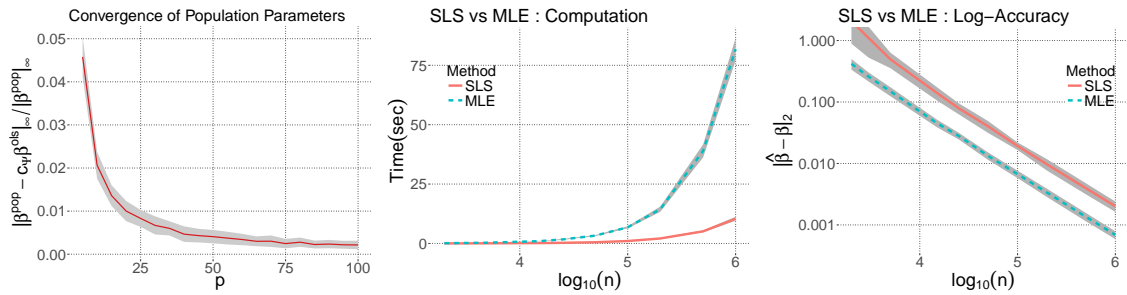
Figure 3: Logistic regression with iid Bernoulli ± 1 design. The left plot shows the relative error between the population parameters. The middle plot shows the computational cost (time) for finding the MLE and SLS as n grows and $p = 200$. The right plot depict the accuracy of the estimators in log scale. In the regime where the MLE is expensive to compute, the SLS is found much more rapidly and has the same accuracy. `R`'s built-in functions are used to find the MLE.

parameters with respect to the dimension. In this experiment, we estimate the relative error $\left\| \beta^{\mathsf{pop}} - c_\Psi \beta^{\mathsf{ols}} \right\|_\infty / \left\| \beta^{\mathsf{pop}} \right\|_\infty$ using a Monte-Carlo simulation. The dimension takes values from $p \in \{5, 10, 15, ..., 100\}$, and for each fixed p, we choose the sample size $n(p) = Kp$ where $K = 10,000$. $\beta^{\mathsf{pop}} = (1, 1, ..., 1)/\sqrt{p}$ is fixed throughout this experiment. Covariates are generated from a Bernoulli distribution $\mathbb{P}(x_{ij} = \pm 1) = 0.5$, and the response is generated from the logistic model. We estimate β^{ols} using coordinatewise average of the ordinary least squares estimator over the whole dataset. The scaling constant is estimated by the inverse of $\frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle x_i, \beta^{\mathsf{pop}} \rangle)$. We repeat this experiment 100 times for each p, and the results are reported in left plot in Figure 3. We observe that the relative error decreases with increasing dimension.

The second experiment demonstrates the effect of sample size. In this experiment, we fix $p = 200$, and the sample size varies $n \in 10^3 \times \{2, 3, 5, 15, 25, 50, 100, 200, 500, 1000\}$. Covariates are again generated from Bernoulli distribution $\mathbb{P}(x_{ij} = \pm 1) = 0.5$, and the response is generated from the logistic model. Maximum likelihood estimator is obtained through `R`'s built-in function `glm`. We report the computation time of these estimators as well as their estimation error in the middle and the right plots in Figure 3, respectively. We observe that MLE performs consistently better in terms of test error, yet it is significantly expensive to compute when n gets larger. In the large-scale regime, the difference between the achieved test errors becomes negligible as seen in the right plot. This is the regime that SLS provides significant advantage. We also emphasize that MLE and SLS has the same convergence rate as can be observed in the right plot.

8.2. Comparisons With Other Optimization Methods

This section contains the results of a variety of numerical studies, which show that the Scaled Least Squares estimator reaches its minimum achievable test error substantially faster than commonly used batch algorithms for finding the MLE. Both logistic and Poisson regression models (two types of GLMs) are utilized in our analyses, which are based on several synthetic and real datasets.

Below, we briefly describe the optimization algorithms for the MLE that were used in the experiments.
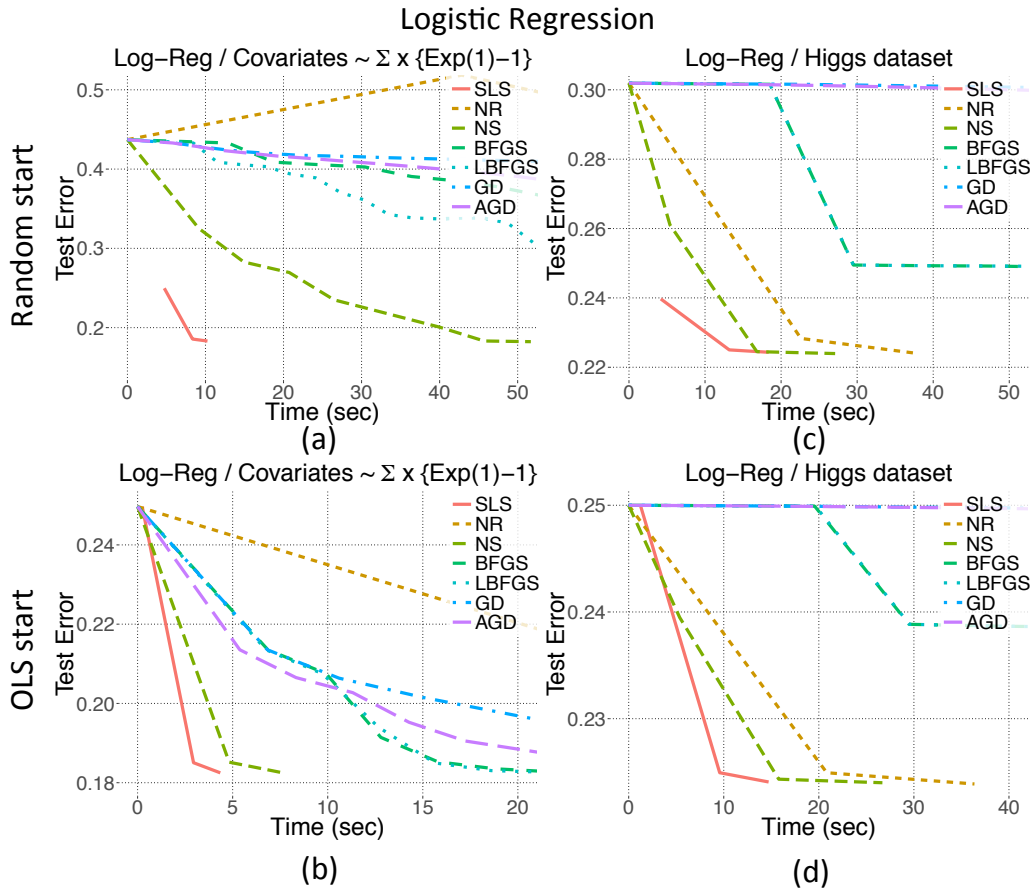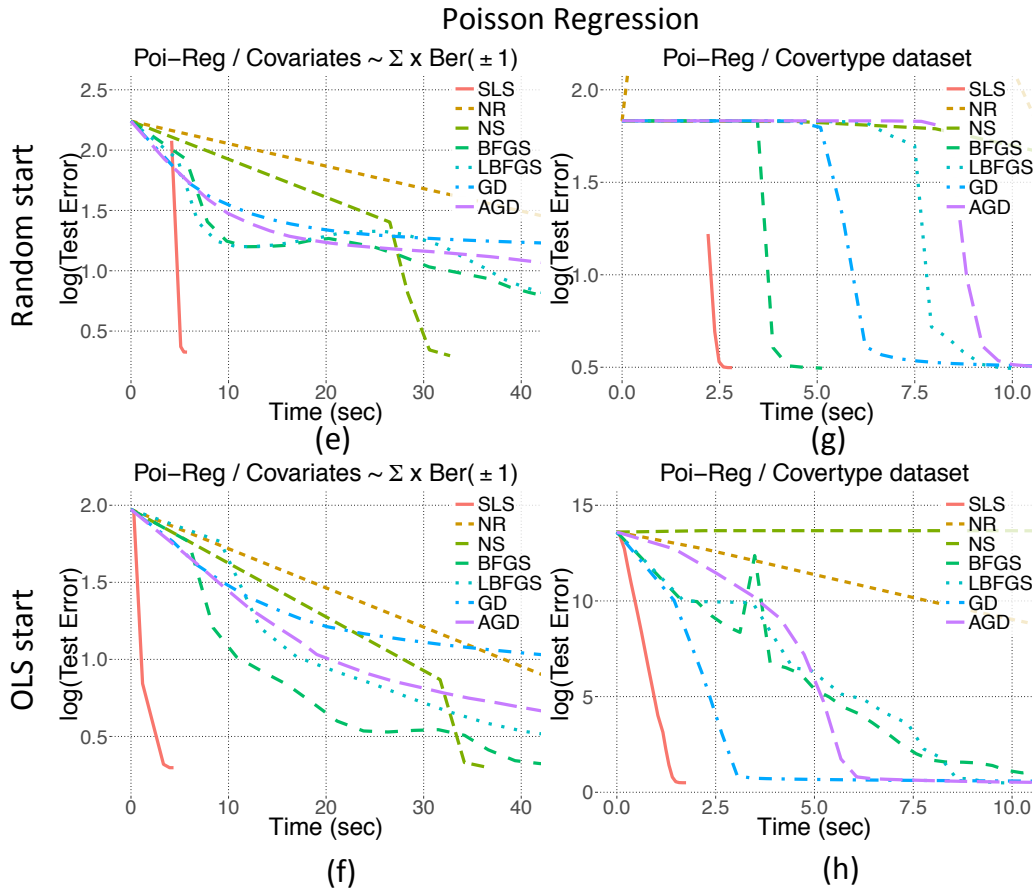
Figure 4: We compared the performance of SLS to that of MLE for the logistic regression problem on several datasets. MLE optimization is solved by various optimization algorithms. SLS is represented with red straight line. The details are provided in Table 2.

1. **Newton-Raphson (NR)** achieves locally quadratic convergence by scaling the gradient by the inverse of the Hessian evaluated at the current iterate. Computing the Hessian has a per-iteration cost of $\mathcal{O}\left(np^2\right)$, which makes it impractical for large-scale datasets.

2. **Newton-Stein (NS)** is a recently proposed second-order batch algorithm specifically designed for GLMs (Erdogdu, 2015, 2016). The algorithm uses Stein's lemma and sub-sampling to efficiently estimate the Hessian with a cost of $\mathcal{O}\left(np\right)$ per-iteration, achieving near quadratic rates.

3. **Broyden-Fletcher-Goldfarb-Shanno (BFGS)** is the most popular and stable quasi-Newton method (Nesterov, 2004). At each iteration, the gradient is scaled by a matrix that is formed by accumulating information from previous iterations and gradient computations. The convergence is locally super-linear with a per-iteration cost of $\mathcal{O}\left(np\right)$.

Figure 5: We compared the performance of SLS to that of MLE for the Poisson regression problem on several datasets. MLE optimization is solved by various optimization algorithms. SLS is represented with red straight line. Details are given in Table 2.

4. **Limited memory BFGS (LBFGS)** is a variant of BFGS, which uses only the recent iterates and gradients to approximate the Hessian, providing significant improvement in terms of memory usage. LBFGS has many variants; we use the formulation given in Bishop (1995).

5. **Gradient descent (GD)** takes a step in the opposite direction of the gradient, evaluated at the current iterate. Its performance strongly depends on the condition number of the design matrix. Under certain assumptions, the convergence is linear with $\mathcal{O}(np)$ per-iteration cost.

6. **Accelerated gradient descent (AGD)** is a modified version of gradient descent with an additional "momentum" term (Nesterov, 1983). Its per iteration cost is $\mathcal{O}(np)$ and its performance strongly depends on the smoothness of the objective function.

For all the algorithms for computing the MLE, the step size at each iteration is chosen via the backtracking line search (Boyd and Vandenberghe, 2004).

Table 2: Details of the experiments shown in Figures 4 and 5 with simulated and real datasets HIGGS (Baldi et al., 2014), and Covertype (Blackard and Dean, 1999)

Model	Logistic regression				Poisson regression			
Dataset	$\Sigma \times \{\text{Exp}(1)\text{-}1\}$		Higgs		$\Sigma \times \text{Ber}(\pm 1)$		Covertype	
Size	$n = 6.0 \times 10^5,\ p = 300$		$n = 1.1 \times 10^7,\ p = 29$		$n = 6.0 \times 10^5,\ p = 300$		$n = 5.8 \times 10^5,\ p = 53$	
Initial	Rnd	Ols	Rnd	Ols	Rnd	Ols	Rnd	Ols
Plot	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Method↓	Time in seconds / number of iterations (to reach min test error)							
Sls	8.34/4	2.94/3	13.18/3	9.57/3	5.42/5	3.96/5	2.71/6	1.66/20
Nr	301.06/6	82.57/3	37.77/3	36.37/3	170.28/5	130.1/4	16.7/8	32.48/18
Ns	51.69/8	7.8/3	27.11/4	26.69/4	32.71/5	36.82/4	21.17/10	282.1/216
Bfgs	148.43/31	24.79/8	660.92/68	701.9/68	67.24/29	72.42/26	5.12/7	22.74/59
Lbfgs	125.33/39	24.61/8	6368.1/651	6946.1/670	224.6/106	357.1/88	10.01/14	10.05/17
Gd	669/138	134.91/25	100871/10101	141736/13808	1711/513	1364/374	14.35/25	33.58/87
Agd	218.1/61	35.97/12	2405.5/251	2879.69/277	103.3/51	102.74/40	11.28/15	11.95/25

Recall that the proposed Algorithm 1 is composed of two steps; the first finds an estimate of the OLS coefficients. This up-front computation is not needed for any of the MLE algorithms described above. On the other hand, each of the MLE algorithms requires some initial value for β, but no such initialization is needed to find the OLS estimator in Algorithm 1. This raises the question of how the MLE algorithms should be initialized, in order to compare them fairly with the proposed method. We consider two scenarios in our experiments: first, we use the OLS estimator computed for Algorithm 1 to initialize the MLE algorithms; second, we use a random initial value.

On each dataset, the main criterion for assessing the performance of the estimators is how rapidly the minimum test error is achieved. The test error is measured as the mean squared error of the estimated mean using the current parameters at each iteration on a test dataset, which is a randomly selected (and set-aside) 10% portion of the entire dataset. As noted previously, the MLE is more accurate for small n (see Figure 1). However, in the regime considered here ($n \gg p \gg 1$), the MLE and the SLS perform very similarly in terms of their error rates; for instance, on the Higgs dataset, the SLS and MLE have test error rates of 22.40% and 22.38%, respectively. For each dataset, the minimum achievable test error is set to be the maximum of the final test errors, where the maximum is taken over all of the estimation methods. Let $\Sigma^{(1)}$ and $\Sigma^{(2)}$ be two randomly generated covariance matrices. The datasets we analyzed were: (i) a synthetic dataset generated from a logistic regression model with iid {exponential(1)$-$1} predictors scaled by $\Sigma^{(1)}$; (ii) the Higgs dataset (logistic regression) Baldi et al. (2014); (iii) a synthetic dataset generated from a Poisson regression model with iid binary($\pm$1) predictors scaled by $\Sigma^{(2)}$; (iv) the Covertype dataset (Poisson regression) Blackard and Dean (1999).

In all cases, the SLS outperformed the alternative algorithms for finding the MLE by a large margin, in terms of computation. Detailed results may be found in Figures 4 and 5, and Table 2. We provide additional experiments with different datasets in Section 12.

9. Discussion

In this paper, we showed that the true minimizer of a generalized linear problem and the OLS estimator are approximately proportional under the general random design setting. Using this relation, we proposed a computationally efficient algorithm for large-scale problems that achieves the same accuracy as the empirical risk minimizer by first estimating the OLS coefficients and then estimating the proportionality constant through iterations that can attain quadratic or cubic convergence rate, with only $\mathcal{O}(n)$ per-iteration cost.

We briefly mentioned that the proportionality between the coefficients holds even when there is regularization in Section 3.1. Further pursuing this idea may be interesting for large-scale problems where regularization is crucial. Another interesting line of research is to find similar proportionality relations between the parameters in other large-scale optimization problems such as support vector machines. Such relations may reduce the problem complexity significantly.

10. Proof of Main Results

In this section, we provide the details and the proofs of our technical results. For convenience, we briefly state the following definitions.

Definition 13 (Sub-Gaussian) *For a given constant κ, a random variable $x \in \mathbb{R}$ is said to be* sub-Gaussian *if it satisfies $\sup_{m \geq 1} m^{-1/2} \mathbb{E}\left[|x|^m\right]^{1/m} \leq \kappa$. Smallest such κ 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 a* sub-Gaussian vector *if there exists a constant κ' such that $\sup_{v \in S^{p-1}} \|\langle y, v \rangle\|_{\psi_2} \leq \kappa'$.*

Definition 14 (Sub-exponential) *For a given constant κ, a random variable $x \in \mathbb{R}$ is called* sub-exponential *if it satisfies $\sup_{m \geq 1} m^{-1} \mathbb{E}\left[|x|^m\right]^{1/m} \leq \kappa$. Smallest such κ 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 a* sub-exponential vector *if there exists a constant κ' such that $\sup_{v \in S^{p-1}} \|\langle y, v \rangle\|_{\psi_1} \leq \kappa'$.*

The proof of Theorem 3 is given next.
Proof [of Theorem 3] For simplicity, we denote the whitened covariate by $w = \mathbf{\Sigma}^{-1/2} x$. Since w is sub-Gaussian with norm κ, its j-th entry w_j has bounded third moment. That is,

$$\kappa = \sup_{\|u\|_2 = 1} \|\langle u, w \rangle\|_{\psi_2} \geq \|w_j\|_{\psi_2} = \sup_{m \geq 1} m^{-1/2} \mathbb{E}\left[|w_j|^m\right]^{1/m} \geq \frac{1}{\sqrt{3}} \mathbb{E}\left[|w_j|^3\right]^{1/3}, \qquad (37)$$

where in the first step, we used $u = e_j$, the j-th standard basis vector. Hence, we obtain a bound on the third moment, i.e,

$$\max_j \mathbb{E}\left[|w_j|^3\right] \leq 3^{3/2} \kappa^3. \qquad (38)$$

Using the normal equations, we write

$$\mathbb{E}[yx] = \mathbb{E}\left[x \Psi^{(1)}(\langle x, \beta \rangle)\right] = \mathbf{\Sigma}^{1/2} \mathbb{E}\left[w \Psi^{(1)}(\langle w, \mathbf{\Sigma}^{1/2} \beta \rangle)\right], \qquad (39)$$

$$= \mathbf{\Sigma}^{1/2} \mathbb{E}\left[w \Psi^{(1)}(\langle w, \tilde{\beta} \rangle)\right],$$

23

where we defined $\tilde{\beta} = \mathbf{\Sigma}^{1/2}\beta$. By multiplying both sides with $\mathbf{\Sigma}^{-1}$, we obtain

$$\beta^{\mathsf{ols}} = \mathbf{\Sigma}^{-1/2}\mathbb{E}\left[w\Psi^{(1)}(\langle w, \tilde{\beta}\rangle)\right]. \tag{40}$$

Now we define the partial sums $W_{-i} = \sum_{j \neq i} \tilde{\beta}_j w_j = \langle \tilde{\beta}, w\rangle - \tilde{\beta}_i w_i$. We will focus on the i-th entry of the above expectation given in (40). Denoting the zero biased transformation of w_i conditioned on W_{-i} by w_i^*, we have

$$\begin{aligned}
\mathbb{E}\left[w_i\Psi^{(1)}(\langle w, \tilde{\beta}\rangle)\right] =& \mathbb{E}\left[\mathbb{E}\left[w_i\Psi^{(1)}\left(\tilde{\beta}_i w_i + W_{-i}\right)\Big|W_{-i}\right]\right], \\
=& \tilde{\beta}_i\mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i w_i^* + W_{-i})\right], \\
=& \tilde{\beta}_i\mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i(w_i^* - w_i) + \langle w, \tilde{\beta}\rangle)\right],
\end{aligned} \tag{41}$$

where in the second step, we used the assumption on conditional moments. Let $\mathbf{D}$ be a diagonal matrix with diagonal entries $\mathbf{D}_{ii} = \mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i(w_i^* - w_i) + \langle w, \tilde{\beta}\rangle)\right]$. Using (40) together with (41), we obtain the equality

$$\beta^{\mathsf{ols}} = \mathbf{\Sigma}^{-1/2}\mathbf{D}\tilde{\beta} = \mathbf{\Sigma}^{-1/2}\mathbf{D}\mathbf{\Sigma}^{1/2}\beta. \tag{42}$$

Now, using the Lipschitz continuity assumption of the variance function, we have

$$\left|\mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i(w_i^* - w_i) + \langle w, \tilde{\beta}\rangle)\right] - \mathbb{E}\left[\Psi^{(2)}(\langle w, \tilde{\beta}\rangle)\right]\right| \leq k|\tilde{\beta}_i|\mathbb{E}\left[|w_i^* - w_i|\right]. \tag{43}$$

In the following, we will use the properties of zero-biased transformations. Consider the quantity

$$r = \sup \frac{\mathbb{E}\left[|w_i^* - w_i|\big|W_{-i}\right]}{\mathbb{E}\left[|w_i|^3\big|W_{-i}\right]} \tag{44}$$

where w_i^* has w_i-zero biased distribution (conditioned on W_{-i}) and the supremum is taken with respect to all random variables with mean 0, standard deviation 1 and finite third moment, and w_i^* is achieving the minimal ℓ_1 coupling to w_i conditioned on W_{-i}. It is shown in Goldstein (2007) that the above bound holds for $r = 1.5$ for the unconditional zero-bias transformations. Here, we take a similar approach to show that the same bound holds for the conditional case as well. By using the triangle inequality, we have

$$\begin{aligned}
\mathbb{E}\left[|w_i^* - w_i|\big|W_{-i}\right] \leq& \mathbb{E}\left[|w_i^*|\big|W_{-i}\right] + \mathbb{E}\left[|w_i|\big|W_{-i}\right] \\
\leq& \frac{1}{2}\mathbb{E}\left[|w_i|^3\big|W_{-i}\right] + \mathbb{E}\left[|w_i|^3\big|W_{-i}\right]^{1/3}.
\end{aligned} \tag{45}$$

Since $\mathbb{E}\left[|w_i|^2\big|W_{-i}\right]$ is constant, it is equal to $\mathbb{E}\left[|w_i|^2\right] = 1$. This yields that the second term in the last line is upper bounded by $\mathbb{E}\left[|w_i|^3\big|W_{-i}\right]$. Consequently, by taking expectations over both hand sides we obtain that

$$\mathbb{E}\left[|w_i^* - w_i|\right] \leq 1.5\,\mathbb{E}\left[|w_i|^3\right]. \tag{46}$$

Then the right hand side of (43) can be upper bounded by

$$k|\tilde{\beta}_i|\mathbb{E}\left[|w_i^* - w_i|\right] \leq rk \max_i \left\{|\tilde{\beta}_i|\mathbb{E}\left[|w_i|^3\right]\right\} \leq 8k\kappa^3\|\mathbf{\Sigma}^{1/2}\beta\|_\infty, \tag{47}$$

where in the second step we used the bound on the third moment given in (38). The last inequality provides us with the following result,

$$\max_i \left|\mathbf{D}_{ii} - \frac{1}{c_\Psi}\right| \leq 8k\kappa^3\|\mathbf{\Sigma}^{1/2}\beta\|_\infty. \tag{48}$$

Finally, combining this with (40) and (42), we obtain

$$\left\|\beta^{\mathsf{ols}} - \frac{1}{c_\Psi}\beta\right\|_\infty = \left\|\mathbf{\Sigma}^{-1/2}\mathbf{D}\mathbf{\Sigma}^{1/2}\beta - \frac{1}{c_\Psi}\beta\right\|_\infty \tag{49}$$

$$= \left\|\mathbf{\Sigma}^{-1/2}\left(\mathbf{D} - \frac{1}{c_\Psi}\mathbf{I}\right)\mathbf{\Sigma}^{1/2}\beta\right\|_\infty,$$

$$\leq \max_i\left|\mathbf{D}_{ii} - \frac{1}{c_\Psi}\right|\left\|\mathbf{\Sigma}^{1/2}\right\|_\infty\left\|\mathbf{\Sigma}^{-1/2}\right\|_\infty\|\beta\|_\infty,$$

$$\leq 8k\kappa^3\rho_\infty\|\mathbf{\Sigma}^{1/2}\|_\infty\|\beta^{\mathsf{pop}}\|_\infty^2,$$

which completes the proof. ∎

Proof [of Corollary 4] Due to diagonal dominance property, we have

$$\|\Sigma^{1/2}\|_\infty = \max_i \sum_{j=1}^p \left|\Sigma_{ij}^{1/2}\right| \leq 2\max_i \Sigma_{ii}^{1/2} \leq 2\|\Sigma\|_2^{1/2}. \tag{50}$$

Since we have $\|x\|_2 \leq r$, we write

$$r^2 \geq \mathbb{E}\left[\|x\|_2^2\right] = \mathrm{Tr}\left(\Sigma\right) \geq p\|\Sigma\|_2/\rho_2. \tag{51}$$

Combining this with the previous inequality we obtain

$$\|\Sigma^{1/2}\|_\infty \leq 2r\sqrt{\rho_2/p}. \tag{52}$$

Finally, the result follows from using the last bound in Theorem 3. ∎

Proof [of Corollary 5] The result directly follows from using the r-well-spreadness assumption in Theorem 3. ∎

Proof [of Proposition 7] For convenience, we denote the whitened covariates by $w_i = \mathbf{\Sigma}^{-1/2}x_i$. We have $\mathbb{E}\left[w_i\right] = 0$, $\mathbb{E}\left[w_i w_i^T\right] = \mathbf{I}$, and $\|w_i\|_{\psi_2} \leq \kappa$. Also denote the sub-sampled covariance matrix with $\widehat{\mathbf{\Sigma}} = \frac{1}{|S|}\sum_{i\in S}x_i x_i^T$, and its whitened version as $\widetilde{\mathbf{\Sigma}} = \frac{1}{|S|}\sum_{i\in S}w_i w_i^T$. Further, define $\hat{\zeta} = \frac{1}{n}\sum_{i=1}^n w_i y_i$ and $\zeta = \mathbb{E}\left[wy\right]$. Then, we have

$$\hat{\beta}^{\mathsf{ols}} = \widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}^{1/2}\hat{\zeta} \quad \text{and} \quad \beta^{\mathsf{ols}} = \mathbf{\Sigma}^{-1/2}\zeta.$$

For now, we work on the event that $\widehat{\boldsymbol{\Sigma}}$ is invertible. We will see that this event holds with very high probability. We write

$$\left\| \boldsymbol{\Sigma}^{1/2}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}}) \right\|_2 = \left\| \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2}\hat{\zeta} - \boldsymbol{\Sigma}^{-1/2}\zeta \right\|_2, \tag{53}$$

$$= \left\| \widetilde{\boldsymbol{\Sigma}}^{-1} \left\{ \hat{\zeta} - \zeta + \left(\mathbf{I} - \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} \right) \zeta \right\} \right\|_2,$$

$$\leq \left\| \widetilde{\boldsymbol{\Sigma}}^{-1} \right\|_2 \left\{ \left\| \hat{\zeta} - \zeta \right\|_2 + \left\| \mathbf{I} - \widetilde{\boldsymbol{\Sigma}} \right\|_2 \|\zeta\|_2 \right\},$$

where we used the triangle inequality and the properties of the operator norm.

For the first term on the right hand side of (53), we write

$$\left\| \widetilde{\boldsymbol{\Sigma}}^{-1} \right\|_2 = \frac{1}{\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}})} \leq \frac{1}{1 - \delta},$$

where we assumed that such a $\delta > 0$ exists. In fact, when $\delta < 0.5$, we obtain a bound of 2 on the right hand side, which also justifies the invertibility assumption of $\widehat{\boldsymbol{\Sigma}}$. By Corollary 5.50 of Vershynin (2010) and the following remark, we have with probability at least $1 - 2\exp\{-p\}$,

$$\left\| \widetilde{\boldsymbol{\Sigma}} - \mathbf{I} \right\|_2 \leq c\sqrt{\frac{p}{|S|}},$$

where c is a constant depending only on κ. When $|S| > 4c^2 p$, we obtain

$$\left| \lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}) - 1 \right| \leq \left\| \widetilde{\boldsymbol{\Sigma}} - \mathbf{I} \right\|_2 \leq 0.5,$$

where the first inequality follows from the Lipschitz property of the eigenvalues.

Next, we bound the difference between $\hat{\zeta}$ and its expectation ζ. We write the bounds on the sub-exponential norm

$$\|wy\|_{\psi_1} = \sup_{\|v\|_2=1} \sup_{m \geq 1} m^{-1}\mathbb{E}\left[|\langle v, w\rangle y|^m\right]^{1/m}, \tag{54}$$

$$\leq \sup_{\|v\|_2=1} \sup_{m \geq 1} m^{-1}\mathbb{E}\left[|\langle v, w\rangle|^{2m}\right]^{1/2m}\mathbb{E}\left[|y|^{2m}\right]^{1/2m},$$

$$\leq \sup_{\|v\|_2=1} \sup_{m \geq 1} m^{-1/2}\mathbb{E}\left[|\langle v, w\rangle|^{2m}\right]^{1/2m} \sup_{m \geq 1} m^{-1/2}\mathbb{E}\left[|y|^{2m}\right]^{1/2m},$$

$$\leq 2\|w\|_{\psi_2}\|y\|_{\psi_2} = 2\gamma\kappa.$$

Hence, we have $\max_i \|w_i y_i - \mathbb{E}[w_i y_i]\|_{\psi_1} \leq 4\gamma\kappa$. Further, let e_j denote the j-th standard basis, and notice that each entry of w is also sub-Gaussian with norm upper bounded by κ, i.e.,

$$\kappa = \|w\|_{\psi_2} = \sup_{\|u\|_2=1} \|\langle u, w\rangle\|_{\psi_2} \geq \|\langle e_j, w\rangle\|_{\psi_2} = \|w_j\|_{\psi_2}.$$

Also, we can write

$$
2\gamma\kappa \geq \|wy\|_{\psi_1} = \sup_{\|u\|_2=1} \sup_{m \geq 1} m^{-1}\mathbb{E}\left[|\langle u,w\rangle y|^m\right]^{1/m}, \tag{55}
$$

$$
\geq \sup_{\|u\|_2=1} \mathbb{E}\left[|\langle u,w\rangle y|\right],
$$

$$
\geq \sup_{\|u\|_2=1} \mathbb{E}\left[\langle u,w\rangle y\right] = \sup_{\|u\|_2=1} \langle u,\zeta\rangle = \|\zeta\|_2,
$$

where in the last step, we used the fact that dual norm of ℓ_2 norm is itself.

Next, we apply Lemma 17 to $\hat{\zeta} - \zeta$, and obtain with probability at least $1 - \exp\{-p\}$

$$
\left\|\hat{\zeta} - \zeta\right\|_2 \leq c\gamma\kappa\sqrt{\frac{p}{n}},
$$

whenever $n > c^2 p$ for an absolute constant c.

Combining the above results in (53), we obtain with probability at least $1 - 3\exp\{-p\}$

$$
\left\|\mathbf{\Sigma}^{1/2}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}})\right\|_2 \leq 2\left\{c_1\gamma\kappa\sqrt{\frac{p}{n}} + c_2\gamma\kappa\sqrt{\frac{p}{|S|}}\right\} \leq \eta\sqrt{\frac{p}{|S|}} \tag{56}
$$

where η depends only on κ and γ, and $|S| > \eta p$. Finally, we write

$$
\left\|\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}}\right\|_2 \leq \lambda_{\min}^{-1/2}\left\|\mathbf{\Sigma}^{1/2}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}})\right\|_2,
$$

$$
\leq \eta\lambda_{\min}^{-1/2}\sqrt{\frac{p}{|S|}},
$$

with probability at least $1 - 3\exp\{-p\}$, whenever $|S| > \eta p$. ∎

The following lemma – combined with the Proposition 7 – provides the necessary tools to prove Theorem 9.

Lemma 15 *For a given function $\Psi^{(2)}$ that is Lipschitz continuous with constant k, we define the function $f : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ as $f(c,\beta) = c\,\mathbb{E}\left[\Psi^{(2)}(\langle x,\beta\rangle c)\right]$, and its empirical counterpart as*

$$
\hat{f}(c,\beta) = c\,\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle c).
$$

For the ball centered around β^{ols} with radius δ,

$$
\mathcal{B}^{\delta}(\tilde{\beta}^{\mathsf{ols}}) = \left\{\beta : \left\|\beta - \tilde{\beta}^{\mathsf{ols}}\right\|_2 \leq \delta\right\}, \quad \text{with} \ \ \tilde{\beta}^{\mathsf{ols}} = \Sigma^{1/2}\beta^{\mathsf{ols}},
$$

assume that $\sup_{\beta \in \mathcal{B}^{\delta}(\tilde{\beta}^{\mathsf{ols}})}\left\|\Psi^{(2)}(\langle x,\beta\rangle)\right\|_{\psi_2} \leq \kappa_g$, and $\left\|\Sigma^{-1/2}x_i\right\|_{\psi_2} \leq \kappa_x$. Further, for some $\epsilon, \bar{c} > 0$, assume that $f(\bar{c}, \beta^{\mathsf{ols}}) \geq 1+\epsilon$. Then, $\exists c_{\Psi} > 0$ satisfying the equation $1 = f(c_{\Psi}, \beta^{\mathsf{ols}})$.

27

Further, assume that for some $\tilde{\epsilon} > 0$, we have $\epsilon = \tilde{\epsilon}\sqrt{p}$, and n and $|S|$ sufficiently large, i.e.,

$$\min\left\{\frac{n}{\log(n)}, |S|\right\} > K^2/\tilde{\epsilon}^2$$

for $K = \eta\bar{c}\max\{b + \kappa/\tilde{\mu}, k\bar{c}\kappa\}$. Then, with probability $1 - 5\exp\{-p\}$, there exists a constant $\hat{c}_\Psi \in (0, \bar{c})$ satisfying the equation

$$1 = \hat{c}_\Psi \frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i, \hat{\beta}^{\mathsf{ols}}\rangle\hat{c}_\Psi).$$

Moreover, if the derivative of $z \to f(z, \beta^{\mathsf{ols}})$ is bounded below in absolute value (i.e. does not change sign) by $\upsilon > 0$ in the interval $z \in [0, \bar{c}]$, then with probability $1 - 5\exp\{-p\}$, we have

$$|\hat{c}_\Psi - c_\Psi| \leq C\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}},$$

where $C = K/\upsilon$.

Proof [of Lemma 15] First statement is obvious. We notice that $f(c, \beta^{\mathsf{ols}})$ is a continuous function in its first argument with $f(0, \beta^{\mathsf{ols}}) = 0$ and $f(\bar{c}, \beta^{\mathsf{ols}}) \geq 1 + \delta$. Hence, there exists $c_\Psi > 0$ such that $f(c_\Psi, \beta^{\mathsf{ols}}) = 1$. If there are many solutions to the above equation, we choose the one that is closest to zero. The condition on the derivative will guarantee the uniqueness of the solution.

Next, we will show the existence of $\hat{c}_\Psi$ using a uniform concentration given by Lemma 18. Define the event $\mathcal{E}$ that $\hat{\beta}^{\mathsf{ols}}$ falls into $\mathcal{B}_{\boldsymbol{\Sigma}}^\delta(\beta^{\mathsf{ols}})$, i.e., $\mathcal{E} = \left\{\hat{\beta}^{\mathsf{ols}} \in \mathcal{B}_{\boldsymbol{\Sigma}}^\delta(\beta^{\mathsf{ols}})\right\}$. By Proposition 7 and the inequality given in (56), whenever $|S| > \eta p$, we obtain

$$\mathbb{P}\left(\mathcal{E}^C\right) \leq 3\exp\{-p\},$$

where $\mathcal{E}^C$ denotes the complement of the event $\mathcal{E}$, and η is a constant depending only on κ_x and γ, and $\delta = \eta\sqrt{p/|S|}$. For any $c \in [0, \bar{c}]$, on the event $\mathcal{E}$, we have

$$\left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| \leq \sup_{\beta \in \mathcal{B}_{\boldsymbol{\Sigma}}^\delta(\beta^{\mathsf{ols}})}\left|\hat{f}(c, \beta) - f(c, \beta)\right|.$$

Hence, we obtain the following inequality

$$\mathbb{P}\left(\sup_{c \in [0, \bar{c}]}\left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| > \epsilon\right)$$

$$\leq \mathbb{P}\left(\sup_{c \in [0, \bar{c}]}\left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| > \epsilon; \mathcal{E}\right) + \mathbb{P}\left(\mathcal{E}^C\right),$$

$$\leq \mathbb{P}\left(\sup_{c \in [0, \bar{c}]}\sup_{\beta \in \mathcal{B}_{\boldsymbol{\Sigma}}^\delta(\beta^{\mathsf{ols}})}\left|\hat{f}(c, \beta) - f(c, \beta)\right| > \epsilon\right) + 3\exp\{-p\}.$$

28

In the following, we will use Lemma 18 for the first term in the last line above. Denoting by w, the whitened covariates, we have $\langle x, \beta \rangle = \langle w, \mathbf{\Sigma}^{1/2} \beta \rangle$. Therefore,

$$\sup_{c \in [0,\bar{c}]} \sup_{\beta \in \mathcal{B}_{\mathbf{\Sigma}}^{\delta}(\beta^{\mathsf{ols}})} \left| \hat{f}(c, \beta) - f(c, \beta) \right|$$

$$\leq \bar{c} \sup_{c \in [0,\bar{c}]} \sup_{\beta \in \mathcal{B}_{\mathbf{\Sigma}}^{\delta}(\beta^{\mathsf{ols}})} \left| \frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle w_i, \mathbf{\Sigma}^{1/2} \beta \rangle c) - \mathbb{E}\left[\Psi^{(2)}(\langle w, \mathbf{\Sigma}^{1/2} \beta \rangle c) \right] \right|.$$

Next, define the ball centered around $\tilde{\beta}^{\mathsf{ols}} = \mathbf{\Sigma}^{1/2} \beta^{\mathsf{ols}}$, with radius δ as $\mathcal{B}_{\delta}(\tilde{\beta}^{\mathsf{ols}}) = \mathbf{\Sigma}^{1/2} \mathcal{B}_{\mathbf{\Sigma}}^{\delta}(\beta^{\mathsf{ols}})$. We have $\beta \in \mathcal{B}_{\mathbf{\Sigma}}^{\delta}(\beta^{\mathsf{ols}})$ if and only if $\mathbf{\Sigma}^{1/2} \beta \in \mathcal{B}_{\delta}(\tilde{\beta}^{\mathsf{ols}})$. Then, the right hand side of the above inequality can be written as

$$\bar{c} \sup_{c \in [0,\bar{c}]} \sup_{\beta \in \mathcal{B}_{\delta}(\tilde{\beta}^{\mathsf{ols}})} \left| \frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle w_i, \beta \rangle c) - \mathbb{E}\left[\Psi^{(2)}(\langle w, \beta \rangle c) \right] \right|,$$

$$= \bar{c} \sup_{\beta \in \mathcal{B}_{\bar{c}\delta}(\tilde{\beta}^{\mathsf{ols}})} \left| \frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle w_i, \beta \rangle) - \mathbb{E}\left[\Psi^{(2)}(\langle w, \beta \rangle) \right] \right|.$$

Then, by Lemma 18, we obtain

$$\mathbb{P}\left(\sup_{c \in [0,\bar{c}]} \left| \hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}}) \right| > c' \bar{c}(\kappa_g + \kappa_x/\tilde{\mu}) \sqrt{\frac{p}{n/\log(n)}} \right) \leq 5 \exp\{-p\} \qquad (57)$$

whenever $np > 51 \max\{\chi, \chi^{-1}\}$ where $\chi = (\kappa_g + \kappa_x/\tilde{\mu})^2/(c' \delta^2 k^2 \bar{c}^2 \tilde{\mu}^2)$.

Also, by the Lipschitz condition for $\Psi^{(2)}$, we have for any $c \in [0, \bar{c}]$, and β_1, β_2,

$$|f(c, \beta_1) - f(c, \beta_2)| \leq k c^2 \mathbb{E}\left[\left| \langle w, \mathbf{\Sigma}^{1/2}(\beta_1 - \beta_2) \rangle \right| \right]$$

$$\leq k \bar{c}^2 \kappa_x \left\| \mathbf{\Sigma}^{1/2}(\beta_1 - \beta_2) \right\|_2.$$

Applying the above bound for $\beta_1 = \hat{\beta}^{\mathsf{ols}}$ and $\beta_2 = \beta^{\mathsf{ols}}$, we obtain with probability $1 - 3 \exp\{-p\}$

$$\left| f(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \beta^{\mathsf{ols}}) \right| \leq \eta k \bar{c}^2 \kappa_x \sqrt{\frac{p}{|S|}}, \qquad (58)$$

where the last step follows from Proposition 7 and the inequality given in (56).

Combining this with the previous bound, and taking into account that $\mu = \tilde{\mu} \sqrt{p}$, for any $c \in [0, \bar{c}]$, with probability $1 - 5 \exp\{-p\}$, we obtain

$$\left| \hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \beta^{\mathsf{ols}}) \right| \leq c' \bar{c}(\kappa_g + \kappa_x/\tilde{\mu}) \sqrt{\frac{p}{n/\log(n)}} + \eta k \bar{c}^2 \kappa_x \sqrt{\frac{p}{|S|}}$$

$$\leq K \sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}$$

where $K = \eta \bar{c} \max\{\kappa_g + \kappa_x/\tilde{\mu}, k \bar{c} \kappa_x\}$. Here, η depends only on κ_x and γ.

In particular, for $c = \bar{c}$ we observe that

$$
\begin{aligned}
\hat{f}(\bar{c}, \hat{\beta}^{\mathsf{ols}}) &\geq f(\bar{c}, \beta^{\mathsf{ols}}) - K\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}} \\
&\geq 1 + \epsilon - K\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}.
\end{aligned}
$$

Therefore, for sufficiently large n and $|S|$ satisfying $\min\left\{\frac{n}{\log(n)}, |S|\right\} > K^2/\tilde{\epsilon}^2$ we obtain $\hat{f}(\bar{c}, \hat{\beta}^{\mathsf{ols}}) > 1$. Since this function is continuous and $\hat{f}(0, \hat{\beta}^{\mathsf{ols}}) = 0$, we obtain the existence of $\hat{c}_\Psi \in [0, \bar{c}]$ with probability at least $1 - 5\exp\{-p\}$.

Now, since $\hat{c}_\Psi$ and c_Ψ satisfy the equations $\hat{f}(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) = f(c_\Psi, \beta^{\mathsf{ols}}) = 1$ (with high probability), by the inequality given in (57), with probability at least $1 - 5\exp\{-p\}$, we obtain

$$
\left|1 - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}})\right| = \left|\hat{f}(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}})\right| \leq c'\bar{c}(\kappa_g + \kappa_x/\tilde{\mu})\sqrt{\frac{p}{n/\log(n)}}.
$$

Also, by the same argument in (60), and Proposition 7, we get

$$
\left|f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - f(\hat{c}_\Psi, \beta^{\mathsf{ols}})\right| \leq k\bar{c}^2\kappa \left\|\boldsymbol{\Sigma}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}})\right\|_2 \leq \eta k\bar{c}^2\kappa_x\sqrt{\frac{p}{|S|}}.
$$

Now, using the Taylor's series expansion of $c \to f(c, \beta^{\mathsf{ols}})$ around c_Ψ, and the assumption on the derivative of f with respect to its first argument, we obtain

$$
\begin{aligned}
\upsilon\left|\hat{c}_\Psi - c_\Psi\right| &\leq \left|f(\hat{c}_\Psi, \beta^{\mathsf{ols}}) - f(c_\Psi, \beta^{\mathsf{ols}})\right| \\
&\leq \left|f(\hat{c}_\Psi, \beta^{\mathsf{ols}}) - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}})\right| + \left|f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - 1\right| \\
&\leq \eta k\bar{c}^2\kappa_x\sqrt{\frac{p}{|S|}} + c'\bar{c}(\kappa_g + \kappa_x/\tilde{\mu})\sqrt{\frac{p}{n/\log(n)}} \\
&\leq K\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}
\end{aligned}
$$

with probability at least $1 - 5\exp\{-p\}$. Here, the constant K is the same as before $K = \eta\bar{c}\max\{\kappa_g + \kappa_x/\tilde{\mu}, k\bar{c}\kappa_x\}$.

$\blacksquare$

Lemma 16 *For a given function $\Psi^{(2)}$ that is Lipschitz continuous with constant k, and uniformly bounded by b, we define the functions f, $\hat{f}$, and the constants $\epsilon, c_\Psi, \bar{c}$ as in Lemma 15. Assume that x_i's are i.i.d. sub-Gaussian with $\|x_i\|_{\psi_2} \leq \kappa_x$. Then, with probability $1 - 4\exp\{-p\}$, there exists a constant $\hat{c}_\Psi \in (0, \bar{c})$ satisfying the equation*

$$
1 = \hat{c}_\Psi \frac{1}{n}\sum_{i=1}^n \Psi^{(2)}(\langle x_i, \hat{\beta}^{\mathsf{ols}}\rangle\hat{c}_\Psi).
$$

Moreover, if the derivative of $z \to f(z, \beta^{\mathsf{ols}})$ is bounded below in absolute value (i.e. does not change sign) by $\upsilon > 0$ in the interval $z \in [0, \bar{c}]$, then with probability $1 - 4\exp\{-p\}$, we have

$$|\hat{c}_\Psi - c_\Psi| \leq K\sqrt{\frac{p}{|S|}},$$

where $K = \eta \upsilon^{-1} \bar{c} k \max\left\{(1 + \|\tilde{\beta}\|_2)\|\Sigma\|_2 + b/k, \bar{c}\right\}$, η depends only on γ and κ_x, and $|S|$ is sufficiently large.

Proof [of Lemma 16] The proof follows from the same steps in Lemma 15. First statement is obvious.

Next, we will show the existence of $\hat{c}_\Psi$ using a uniform concentration given by Lemma 19. Let the ellipsoids $\mathcal{B}_\Sigma^\delta(\beta^{\mathsf{ols}})$ and $\mathcal{B}^\delta(\tilde{\beta}^{\mathsf{ols}})$, and the event $\mathcal{E}$ be as in Lemma 15. By Proposition 7 and the inequality given in (56), whenever $|S| > \eta p$, we obtain

$$\mathbb{P}\left(\mathcal{E}^C\right) \leq 3\exp\{-p\},$$

where $\mathcal{E}^C$ denotes the complement of the event $\mathcal{E}$, and η is a constant depending only on κ_x and γ. For any $c \in [0, \bar{c}]$, on the event $\mathcal{E}$, we have

$$\left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| \leq \sup_{\beta \in \mathcal{B}_\Sigma^\delta(\beta^{\mathsf{ols}})} \left|\hat{f}(c, \beta) - f(c, \beta)\right|.$$

Using the same arguments as in the previous lemma that led to (57), we obtain the following inequality

$$\mathbb{P}\left(\sup_{c \in [0, \bar{c}]} \left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| > \epsilon\right)$$

$$\leq \mathbb{P}\left(\sup_{c \in [0, \bar{c}]} \sup_{\beta \in \mathcal{B}_\Sigma^\delta(\beta^{\mathsf{ols}})} \left|\hat{f}(c, \beta) - f(c, \beta)\right| > \epsilon\right) + 3\exp\{-p\},$$

$$\leq \mathbb{P}\left(\sup_{\beta \in \mathcal{B}_{\bar{c}\delta}(\tilde{\beta}^{\mathsf{ols}})} \left|\frac{1}{n}\sum_{i=1}^n \Psi^{(2)}(\langle w_i, \beta\rangle) - \mathbb{E}\left[\Psi^{(2)}(\langle w, \beta\rangle)\right]\right| > \epsilon/\bar{c}\right) + 3\exp\{-p\}$$

where $\mathcal{B}_{\bar{c}\delta}(\tilde{\beta}^{\mathsf{ols}})$ is the ball centered around $\tilde{\beta}^{\mathsf{ols}} = \Sigma^{1/2}\beta^{\mathsf{ols}}$ with radius $\delta = \eta\sqrt{p/|S|}$.

Then, using Lemma 19, we obtain

$$\mathbb{P}\left(\sup_{c \in [0, \bar{c}]} \left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| > 2\bar{c}\left(k\left\{\|\tilde{\beta}\|_2 + \eta\sqrt{\frac{p}{|S|}}\right\}\|\Sigma\|_2 + b\right)\sqrt{\frac{p}{n}}\right) \leq 4\exp\{-p\}.$$

(59)

By the Lipschitz property of $\Psi^{(2)}$ as well as sub-Gaussian property of x, we have for any $c \in [0, \bar{c}]$, and β_1, β_2,

$$|f(c, \beta_1) - f(c, \beta_2)| \leq kc^2 \mathbb{E}\left[\left|\langle w, \Sigma^{1/2}(\beta_1 - \beta_2)\rangle\right|\right]$$

$$\leq k\bar{c}^2\kappa_x \left\|\Sigma^{1/2}(\beta_1 - \beta_2)\right\|_2.$$

Applying the above bound for $\beta_1 = \hat{\beta}^{\mathsf{ols}}$ and $\beta_2 = \beta^{\mathsf{ols}}$, we obtain with probability $1 - 3\exp\{-p\}$

$$\left| f(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \beta^{\mathsf{ols}}) \right| \le \eta k \bar{c}^2 \kappa_x \sqrt{\frac{p}{|S|}}, \tag{60}$$

where the last step follows from Proposition 7 and the inequality given in (56).

Combining this with the previous bound, for any $c \in [0, \bar{c}]$, with probability $1 - 4\exp\{-p\}$, we obtain

$$\left| \hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \beta^{\mathsf{ols}}) \right| \le 2\bar{c}\left(k\|\tilde{\beta}\|_2 \|\Sigma\|_2 + \eta k \|\Sigma\|_2 \sqrt{\frac{p}{|S|}} + b \right) \sqrt{\frac{p}{n}} + \eta k \bar{c}^2 \kappa_x \sqrt{\frac{p}{|S|}}$$

$$= 2\bar{c}k\left\{ \|\tilde{\beta}\|_2 \|\Sigma\|_2 \sqrt{\frac{p}{n}} + \eta \|\Sigma\|_2 \frac{p}{\sqrt{|S|n}} + b\sqrt{\frac{p}{n}} + \eta \bar{c}\kappa_x \sqrt{\frac{p}{|S|}} \right\}$$

$$\le \eta \bar{c}k \max\left\{ (1 + \|\tilde{\beta}\|_2) \|\Sigma\|_2 + b/k, \bar{c} \right\} \sqrt{\frac{p}{|S|}} = K\sqrt{\frac{p}{|S|}}$$

where $K = \eta \bar{c}k \max\left\{ (1 + \|\tilde{\beta}\|_2) \|\Sigma\|_2 + b/k, \bar{c} \right\}$, and $|S| > \eta p$. Here, η depends only on the constants κ_x and γ. In particular, for $c = \bar{c}$ we observe that

$$\hat{f}(\bar{c}, \hat{\beta}^{\mathsf{ols}}) \ge f(\bar{c}, \beta^{\mathsf{ols}}) - K\sqrt{\frac{p}{|S|}}$$

$$\ge 1 + \epsilon - K\sqrt{\frac{p}{|S|}}.$$

Therefore, for sufficiently large $|S|$ satisfying $|S| > K^2 p$, we obtain $\hat{f}(\bar{c}, \hat{\beta}^{\mathsf{ols}}) > 1$. Since this function is continuous and $\hat{f}(0, \hat{\beta}^{\mathsf{ols}}) = 0$, we obtain the existence of $\hat{c}_\Psi \in [0, \bar{c}]$ with probability at least $1 - 4\exp\{-p\}$.

Now, since $\hat{c}_\Psi$ and c_Ψ satisfy the equations $\hat{f}(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) = f(c_\Psi, \beta^{\mathsf{ols}}) = 1$ (with high probability), by the inequality given in (59), with probability at least $1 - 4\exp\{-p\}$, we obtain

$$\left| 1 - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) \right| = \left| \hat{f}(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) \right| \le 2\bar{c}\left(k\left\{ \|\tilde{\beta}\|_2 + \eta\sqrt{\frac{p}{|S|}} \right\} \|\Sigma\|_2 + b \right) \sqrt{\frac{p}{n}}.$$

Now, using the Taylor's series expansion of $c \to f(c, \beta^{\mathsf{ols}})$ around c_Ψ, and the assumption on the derivative of f with respect to its first argument, we obtain

$$\upsilon |\hat{c}_\Psi - c_\Psi| \le \left| f(\hat{c}_\Psi, \beta^{\mathsf{ols}}) - f(c_\Psi, \beta^{\mathsf{ols}}) \right|$$

$$\le \left| f(\hat{c}_\Psi, \beta^{\mathsf{ols}}) - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) \right| + \left| f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - 1 \right|$$

$$\le \eta k \bar{c}^2 \kappa_x \sqrt{\frac{p}{|S|}} + 2\bar{c}\left(k\left\{ \|\tilde{\beta}\|_2 + \eta\sqrt{\frac{p}{|S|}} \right\} \|\Sigma\|_2 + b \right) \sqrt{\frac{p}{n}}$$

$$\le K\sqrt{\frac{p}{|S|}}$$

with probability at least $1 - 4\exp\{-p\}$. Here, the constant K is the same as before $K = \eta\bar{c}k \max\left\{(1 + \|\tilde{\beta}\|_2)\|\Sigma\|_2 + b/k, \bar{c}\right\}$.

$\blacksquare$

Proof [of Theorem 8] We have

$$\left\|\hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}}\right\|_\infty \leq \left\|c_\Psi\beta^{\mathsf{ols}} - \beta^{\mathsf{pop}}\right\|_\infty + \left\|\hat{c}_\Psi\hat{\beta}^{\mathsf{ols}} - c_\Psi\beta^{\mathsf{ols}}\right\|_\infty, \tag{61}$$

where we used the triangle inequality for the ℓ_∞ norm. The first term on the right hand side can be bounded using Theorem 3 and Corollary 5. We write

$$\left\|c_\Psi\beta^{\mathsf{ols}} - \beta^{\mathsf{pop}}\right\|_\infty \leq \eta_1 \frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}}, \tag{62}$$

for $\eta_1 = 8k\bar{c}\kappa^3\rho_\infty\|\mathbf{\Sigma}^{1/2}\|_\infty(\tau/r)$.

For the second term, we write

$$\left\|\hat{c}_\Psi\hat{\beta}^{\mathsf{ols}} - c_\Psi\beta^{\mathsf{ols}}\right\|_\infty = \left\|\hat{c}_\Psi\hat{\beta}^{\mathsf{ols}} \pm \hat{c}_\Psi\beta^{\mathsf{ols}} - c_\Psi\beta^{\mathsf{ols}}\right\|_\infty, \tag{63}$$

$$\leq |\hat{c}_\Psi| \left\|\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}}\right\|_\infty + |\hat{c}_\Psi - c_\Psi| \left\|\beta^{\mathsf{ols}}\right\|_\infty,$$

where the first step follows from triangle inequality. By Lemma 16, for sufficiently large n and $|S|$, with probability $1 - 4\exp\{-p\}$, the constant $\hat{c}_\Psi$ exists and it is in the interval $(0, \bar{c}]$. By the same lemma, with probability $1 - 4\exp\{-p\}$, we have

$$|\hat{c}_\Psi - c_\Psi| \leq \eta_4\sqrt{\frac{p}{|S|}}, \tag{64}$$

where $\eta_4 = \eta'\upsilon^{-1}\bar{c}k \max\left\{(\|\Sigma^{1/2}\beta^{\mathsf{ols}}\|_2 + 1)\|\Sigma\|_2 + b/k, \bar{c}\right\}$ for some constant η' depending on the sub-Gaussian norms κ and γ.

Also, by the norm equivalence and Proposition 7, we have with probability $1 - 4\exp\{-p\}$

$$\left\|\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}}\right\|_\infty \leq \eta_3\sqrt{\frac{p}{|S|}}, \tag{65}$$

for $\eta_3 = \eta''\lambda_{\min}^{-1/2}$, where η'' is constant depending only on γ and κ.

Finally, combining all these inequalities with the last line of (61), we have with probability $1 - 4\exp\{-p\}$,

$$\left\|\hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}}\right\|_\infty \leq \eta_1\frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}} + \eta_3\bar{c}\sqrt{\frac{p}{|S|}} + \eta_4\left\|\beta^{\mathsf{ols}}\right\|_\infty\sqrt{\frac{p}{|S|}}, \tag{66}$$

$$\leq \eta_1\frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}} + \left(\eta_3\bar{c} + \eta_4\left\|\beta^{\mathsf{ols}}\right\|_\infty\right)\sqrt{\frac{p}{|S|}},$$

$$= \eta_1\frac{\|\beta^{\mathsf{pop}}\|_\infty}{\sqrt{p}} + \eta_2\sqrt{\frac{p}{|S|}},$$

where

$$\eta_1 = \eta k \bar{c} \rho_\infty (\tau / r) \| \mathbf{\Sigma}^{1/2} \|_\infty \tag{67}$$

$$\eta_2 = \eta_3 \bar{c} + \eta_4 \left\| \beta^{\mathsf{ols}} \right\|_\infty = \eta \bar{c} \left(\lambda_{\min}^{-1/2} + k \upsilon^{-1} \max \left\{ (\| \Sigma^{1/2} \beta^{\mathsf{ols}} \|_2 + 1) \| \Sigma \|_2 + b/k, \bar{c} \right\} \right)$$

∎

Proof [of Theorem 9] Proof is essentially the same as that of Theorem 8, where this time we use Lemma 16 instead of Lemma 15.

∎

Proof [of Corollary 6] The normal equations for the lasso minimization yields

$$\mathbb{E} \left[x x^T \right] \beta_\lambda^{\mathsf{lasso}} - \beta^{\mathsf{ols}} + \lambda s = 0,$$

where $s \in \partial \left\| \beta_\lambda^{\mathsf{lasso}} \right\|_1$. It is well-known that under the orthogonal design where the covariates have i.i.d. entries, the above equation reduces to $\mathsf{soft}(\beta^{\mathsf{ols}}; \lambda) = \beta_\lambda^{\mathsf{lasso}}$, where $\mathsf{soft}(\cdot ; \lambda)$ denotes the soft thresholding operator at level λ. For any $\beta \in \mathbb{R}^p$, let $\mathrm{supp}(\beta)$ denote the support of β, i.e., the set $\{ i \in [p] : \beta_i \neq 0 \}$. We have

$$\mathrm{supp}(\beta_\lambda^{\mathsf{lasso}}) = \{ i \in [p] : \beta_{\lambda, i}^{\mathsf{lasso}} \neq 0 \} = \{ i \in [p] : |\beta_i^{\mathsf{ols}}| > \lambda \}$$

By Theorem 3, we have $|\beta_i^{\mathsf{ols}}| \leq \frac{1}{c_\Psi} |\beta_i^{\mathsf{pop}}| + \delta$, which implies that

$$\mathrm{supp}(\beta_\lambda^{\mathsf{lasso}}) \subset \left\{ i \in [p] : \frac{1}{c_\Psi} |\beta_i^{\mathsf{pop}}| + \delta \geq \lambda \right\}.$$

Hence, whenever $\lambda \geq \delta$, we have $\mathrm{supp}(\beta_\lambda^{\mathsf{lasso}}) \subset \mathrm{supp}(\beta^{\mathsf{pop}})$. Further, we have by Theorem 3

$$\frac{1}{c_\Psi} |\beta_i^{\mathsf{pop}}| \leq |\beta_i^{\mathsf{ols}}| + \delta.$$

Hence, whenever $|\beta_i^{\mathsf{pop}}| > c_\Psi (\lambda + \delta)$, we get $|\beta_i^{\mathsf{ols}}| > \lambda$. If this condition is satisfied for any entry in the support of β^{pop}, the corresponding lasso coefficient will be non-zero. Therefore, we get $\mathrm{supp}(\beta^{\mathsf{pop}}) \subset \mathrm{supp}(\beta_\lambda^{\mathsf{lasso}})$ under this assumption. Combining this with the previous result, we conclude the proof. ∎

11. Auxiliary Lemmas

Lemma 17 (Sub-exponential vector concentration) *Let $x_1, x_2, ..., x_n$ be independent centered sub-exponential random vectors with $\max_i \| x_i \|_{\psi_1} = \kappa$. Then we have*

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_{i=1}^n x_i \right\|_2 > c \kappa \sqrt{\frac{p}{n}} \right) \leq \exp \{ -p \}. \tag{68}$$

whenever $n > 4c^2 p$ for an absolute constant c.

Proof [of Lemma 17] For a vector $z \in \mathbb{R}^p$, we have $\|z\|_2 = \sup_{\|u\|_2=1} \langle u, z \rangle$ since the dual of ℓ_2 norm is itself. Therefore, we write

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n x_i \right\|_2 > t \right) = \mathbb{P}\left(\sup_{\|u\|_2=1} \frac{1}{n} \sum_{i=1}^n \langle u, x_i \rangle > t \right).$$

Now, let $\mathcal{N}_\epsilon$ be an ϵ-net over $\mathcal{S}^{p-1} = \{u \in \mathbb{R}^p : \|u\|_2 = 1\}$, and observe that

$$\max_{u \in \mathcal{N}_\epsilon} \langle u, x \rangle \geq (1-\epsilon) \sup_{\|u\|_2=1} \langle u, x \rangle = (1-\epsilon)\|x\|_2,$$

with $|\mathcal{N}_\epsilon| \leq (1 + 2/\epsilon)^p$. Hence, we may write

$$\mathbb{P}\left(\sup_{\|u\|_2=1} \frac{1}{n} \sum_{i=1}^n \langle u, x_i \rangle > t \right) \leq \mathbb{P}\left(\max_{u \in \mathcal{N}_\epsilon} \frac{1}{n} \sum_{i=1}^n \langle u, x_i \rangle > t(1-\epsilon) \right),$$

$$\leq |\mathcal{N}_\epsilon| \, \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \langle u, x_i \rangle > t(1-\epsilon) \right).$$

For any $u \in \mathcal{S}^{p-1}$, we have $\|\langle u, x_i \rangle\|_{\psi_1} \leq \kappa$. Then, by the Bernstein-type inequality for sub-exponential random variables (Vershynin, 2010), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \langle u, x_i \rangle > t(1-\epsilon) \right) \leq \exp\left\{ -cn \min\left\{ \frac{t^2(1-\epsilon)^2}{\kappa^2}, \frac{t(1-\epsilon)}{\kappa} \right\} \right\},$$

for an absolute constant c. Therefore, the probability on the left hand side of (68) can be bounded by

$$\left(1 + \frac{2}{\epsilon} \right)^p \exp\left\{ -cn \frac{t^2(1-\epsilon)^2}{\kappa^2} \right\} = \exp\left\{ -cn \frac{t^2(1-\epsilon)^2}{\kappa^2} + p \log\left(1 + \frac{2}{\epsilon} \right) \right\},$$

whenever $t < \kappa/(1-\epsilon)$. Choosing $\epsilon = 0.5$ and for an absolute constant $c' > 3.24/c$ and letting $t = c'\kappa\sqrt{\frac{p}{n}}$, we conclude the proof. ∎

Lemma 18 *Let $B(\tilde{\beta})$ denote the ball centered around $\tilde{\beta}$ with radius δ, i.e.,*

$$B(\tilde{\beta}) = \left\{ \beta : \left\| \beta - \tilde{\beta} \right\|_2 \leq \delta \right\}.$$

For $i = 1, ..., n$, let $x_i \in \mathbb{R}^p$ be i.i.d. sub-Gaussian random vectors with $\|x_i\|_{\psi_2} \leq \kappa_x$. Given a function $g : \mathbb{R} \to \mathbb{R}$ that is Lipschitz continuous with k, and satisfying $\sup_{\beta \in B(\tilde{\beta})} \|g(\langle x_i, \beta \rangle)\|_{\psi_2} \leq \kappa_g$, we have

$$\mathbb{P}\left(\sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^n g(\langle x_i, \beta \rangle) - \mathbb{E}\left[g(\langle x, \beta \rangle) \right] \right| > c(\kappa_g + \kappa_x/\tilde{\mu}) \sqrt{\frac{p}{n/\log(n)}} \right) \leq 2 \exp\{-p\},$$

whenever $np > 51 \max\{\chi, \chi^{-1}\}$ for $\chi = (\kappa_g + \kappa_x/\tilde{\mu})^2/(c\delta^2 k^2 \tilde{\mu}^2)$. Above, c is an absolute constant.

Proof [of Lemma 18] Let $\mathbb{E}\left[\|x\|_2\right] = \mu = \tilde{\mu}\sqrt{p}$ and for $\epsilon > 0$, $\beta \in B(\tilde{\beta})$ and $w \in \mathbb{R}^p$ define the bounding functions

$$l_\beta(w) = g(\langle w, \beta \rangle) - \epsilon \|w\|_2/4\mu, \quad u_\beta(w) = g(\langle w, \beta \rangle) + \epsilon \|w\|_2/4\mu.$$

Let $\mathcal{N}_\Delta$ be a net over $B(\tilde{\beta})$ in the sense that for any $\beta_1 \in B(\tilde{\beta})$, $\exists \beta_2 \in \mathcal{N}_\Delta$ such that $\|\beta_1 - \beta_2\|_2 \leq \Delta$. We fix $\Delta_* = \epsilon/(4k\mu)$ and write $\forall \beta_1 \in B$, $\exists \beta_2 \in \mathcal{N}_{\Delta_*}$,

1. an upper bound of the form:

$$\begin{aligned} g(\langle w, \beta_1 \rangle) &\leq g(\langle w, \beta_2 \rangle) + k \left| \langle w, \beta_1 - \beta_2 \rangle \right|, \\ &\leq g(\langle w, \beta_2 \rangle) + k \|w\|_2 \Delta_* = u_{\beta_2}(w), \end{aligned}$$

2. and a lower bound of the form:

$$\begin{aligned} g(\langle w, \beta_1 \rangle) &\geq g(\langle w, \beta_2 \rangle) - k \left| \langle w, \beta_1 - \beta_2 \rangle \right|, \\ &\geq g(\langle w, \beta_2 \rangle) - k \|w\|_2 \Delta_* = l_{\beta_2}(w), \end{aligned}$$

where the second steps in the above inequalities follow from the Cauchy-Schwarz inequality. These functions are called *bracketing functions* in the context of empirical process theory.

Hence, we can write that $\forall \beta_1 \in B(\tilde{\beta})$, $\exists \beta_2 \in \mathcal{N}_{\Delta_*}$ such that

$$\frac{1}{n} \sum_{i=1}^n l_{\beta_2}(x_i) - \mathbb{E}\left[l_{\beta_2}(x)\right] - \epsilon/2 \leq \frac{1}{n} \sum_{i=1}^n g(\langle x_i, \beta_1 \rangle) - \mathbb{E}\left[g(\langle x, \beta_1 \rangle)\right],$$

$$\leq \frac{1}{n} \sum_{i=1}^n u_{\beta_2}(x_i) - \mathbb{E}\left[u_{\beta_2}(x)\right] + \epsilon/2.$$

The above inequalities translate to the following conclusion: Whenever the following event happens,

$$\left\{ \left| \frac{1}{n} \sum_{i=1}^n g(\langle x_i, \beta_1 \rangle) - \mathbb{E}\left[g(\langle x, \beta_1 \rangle)\right] \right| > \epsilon \right\},$$

at least one of the following events happens

$$\left\{ \frac{1}{n} \sum_{i=1}^n u_{\beta_2}(x_i) - \mathbb{E}\left[u_{\beta_2}(x)\right] > \epsilon/2 \right\} \quad \text{or} \quad \left\{ \frac{1}{n} \sum_{i=1}^n l_{\beta_2}(x_i) - \mathbb{E}\left[l_{\beta_2}(x)\right] < -\epsilon/2 \right\}.$$

Therefore, using the union bound on the above events, we may obtain

$$\mathbb{P}\left(\sup_{\beta \in B(\tilde{\beta})} \left| \frac{1}{n} \sum_{i=1}^n g(\langle x_i, \beta \rangle) - \mathbb{E}\left[g(\langle x, \beta \rangle)\right] \right| > \epsilon \right) \tag{69}$$

$$\leq \mathbb{P}\left(\max_{\beta \in \mathcal{N}_{\Delta_*}} \frac{1}{n} \sum_{i=1}^n u_\beta(x_i) - \mathbb{E}\left[u_\beta(x)\right] > \epsilon/2 \right)$$

$$+ \mathbb{P}\left(\max_{\beta \in \mathcal{N}_{\Delta_*}} \frac{1}{n} \sum_{i=1}^n l_\beta(x_i) - \mathbb{E}\left[l_\beta(x)\right] < -\epsilon/2 \right).$$

Note that the right hand side of the above inequality has two terms both of which are of the same form. For simplicity, we bound only the first one. The bound for the second one follows from the exact same steps.

The relation between sub-Gaussian and sub-exponential norms (Vershynin, 2010) allows us to write

$$\| \|x\|_2 \|_{\psi_2}^2 \leq \| \|x\|_2^2 \|_{\psi_1} \leq \sum_{i=1}^{p} \|x_i^2\|_{\psi_1} \leq 2\sum_{i=1}^{p} \|x_i\|_{\psi_2}^2 \leq 2\kappa_x^2 p, \tag{70}$$

where the second step follows from the triangle inequality. Hence, we conclude that $\|x\|_2 - \mathbb{E}[\|x\|_2]$ is a centered sub-Gaussian random variable with norm upper bounded by $3\kappa_x\sqrt{p}$.

For $\epsilon < 4/3$, we notice that the random variable $u_\beta(x) = g(\langle x, \beta \rangle) + \epsilon\|x\|_2/4\mu$ is also sub-Gaussian with norm

$$\|u_\beta(x)\|_{\psi_2} \leq \kappa_g + \frac{\epsilon}{4\tilde{\mu}}3\kappa_x \leq \kappa_g + \kappa_x/\tilde{\mu},$$

and consequently, the centered random variable $u_\beta(x) - \mathbb{E}[u_\beta(x)]$ has the sub-Gaussian norm upper bounded by $2\kappa_g + 2\kappa_x/\tilde{\mu}$.

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

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} u_\beta(x_i) - \mathbb{E}[u_\beta(x)] > \epsilon/2\right) \leq \exp\left\{-cn\frac{\epsilon^2}{(\kappa_g + \kappa_x/\tilde{\mu})^2}\right\}$$

for an absolute constant $c > 0$.

By the same argument above, one can obtain the same result for the function $l_\beta(x)$. Using Hoeffding bounds in (69) along with the union bound over the net, we immediately obtain

$$\mathbb{P}\left(\sup_{\beta \in B(\tilde{\beta})}\left|\frac{1}{n}\sum_{i=1}^{n} g(\langle x_i, \beta \rangle) - \mathbb{E}[g(\langle x, \beta \rangle)]\right| > \epsilon\right) \leq 2|\mathcal{N}_{\Delta_*}|\exp\left\{-cn\frac{\epsilon^2}{(\kappa_g + \kappa_x/\tilde{\mu})^2}\right\}$$

for some absolute constant c.

Using a standard covering argument over the net $\mathcal{N}_{\Delta_*}$ as given in Erdogdu (2015), we have

$$|\mathcal{N}_{\Delta_*}| \leq \left(\frac{\delta\sqrt{p}}{\Delta_*}\right)^p = \left(\frac{4\delta k\tilde{\mu}p}{\epsilon}\right)^p.$$

Combining this with the previous bound, and choosing

$$\epsilon^2 = \frac{p}{n}\frac{(\kappa_g + \kappa_x/\tilde{\mu})^2}{2c}\log\left(\frac{32c\delta^2 k^2\tilde{\mu}^2 pn}{(\kappa_g + \kappa_x/\tilde{\mu})^2}\right)$$

we get

$$2\left(\frac{4\delta k\tilde{\mu}p}{\epsilon}\right)^p\exp\left\{-cn\frac{\epsilon^2}{(\kappa_g + \kappa_x/\tilde{\mu})^2}\right\}$$

$$= 2\exp\left\{-\frac{p}{2}\log\log\left(\frac{32c\delta^2 k^2\tilde{\mu}^2 pn}{(\kappa_g + \kappa_x/\tilde{\mu})^2}\right)\right\} \leq 2\exp\{-p\},$$

whenever $np > 51 \max\{\chi, \chi^{-1}\}$ for $\chi = (\kappa_g + \kappa_x/\tilde{\mu})^2/(c\delta^2 k^2 \tilde{\mu}^2)$.

∎

Lemma 19 *Let $B(\tilde{\beta})$ denote the ball centered around $\tilde{\beta}$ with radius δ, i.e.,*

$$B(\tilde{\beta}) = \left\{ \beta : \left\| \beta - \tilde{\beta} \right\|_2 \leq \delta \right\}.$$

For $i = 1, ..., n$, let $x_i \in \mathbb{R}^p$ be i.i.d. random vectors with a covariance matrix Σ. Given a function $g : \mathbb{R} \to \mathbb{R}$ that is uniformly bounded by $b > 0$, and Lipschitz continuous with k,

$$\mathbb{P}\left(\sup_{\beta \in B(\tilde{\beta})} \left| \frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle x_i, \beta \rangle) - \mathbb{E}\left[\Psi^{(2)}(\langle x, \beta \rangle) \right] \right| \geq 2 \left\{ k(\|\tilde{\beta}\|_2 + \delta) \|\Sigma\|_2 + b \right\} \sqrt{\frac{p}{n}} \right) \leq e^{-p}. \tag{71}$$

Proof [of Lemma 19] This is a standard result in the theory of Rademacher complexity – see for example Bartlett and Mendelson (2002). Define the following empirical process

$$G(x_1, ..., x_n) = \sup_{\beta \in B(\tilde{\beta})} \left| \frac{1}{n} \sum_{i=1}^{n} g(\langle x_i, \beta \rangle) - \mathbb{E}\left[g(\langle x, \beta \rangle) \right] \right|. \tag{72}$$

We notice that for any i, we have

$$\left| G(.., x_i, ..) - G(.., x_i', ..) \right| \leq \frac{2b}{n}, \tag{73}$$

due to boundedness of the function g. By the McDiarmid's inequality (Azuma), we have

$$\mathbb{P}\left(G > \epsilon \right) \leq \exp\left\{ -\frac{(\epsilon - \mathbb{E}\left[G \right])^2 n}{2b^2} \right\}. \tag{74}$$

The above inequality – when stated differently – reads that with probability at least $1 - \exp\{-p\}$, we have

$$G < \mathbb{E}\left[G \right] + b\sqrt{\frac{2p}{n}}. \tag{75}$$

For $\sigma_1, ..., \sigma_n$ i.i.d. random variables with $\mathbb{P}\left(\sigma_i = \pm 1 \right) = 0.5$, we have

$$\mathbb{E}\left[G \right] \leq \frac{2}{n} \mathbb{E}\left[\sup_{\beta \in B(\tilde{\beta})} \sum_{i=1}^{n} \sigma_i g(\langle x_i, \beta \rangle) \right] = \mathcal{R}(g \circ \mathcal{B}(\tilde{\beta})), \tag{76}$$

where $\mathcal{R}$ denoting the Rademacher complexity. Since g is Lipschitz continuous with constant k, we have $\mathcal{R}(g \circ B(\tilde{\beta})) \leq k\mathcal{R}(B(\tilde{\beta}))$ by the Ledoux-Talagrand contraction inequality. We

write

$$\mathcal{R}(B(\tilde{\beta})) = \frac{2}{n}\mathbb{E}\left[\sup_{\beta\in B(\tilde{\beta})}\sum_{i=1}^{n}\sigma_i\langle x_i,\beta\rangle\right] \le \frac{2}{n}\mathbb{E}\left[\sup_{\beta\in B(\tilde{\beta})}\left\|\sum_{i=1}^{n}\sigma_i x_i\right\|_2\|\beta\|_2\right], \qquad (77)$$

$$\le(\|\tilde{\beta}\|_2+\delta)\frac{2}{n}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\sigma_i x_i\right\|_2^2\right]^{1/2} = (\|\tilde{\beta}\|_2+\delta)\frac{2}{n}\left(\sum_{i=1}^{n}\mathbb{E}\left[\|x_i\|_2^2\right]\right)^{1/2},$$

$$\le 2(\|\tilde{\beta}\|_2+\delta)\|\Sigma\|_2\sqrt{\frac{p}{n}}.$$

Combining this with the previous results, we obtain

$$\mathbb{P}\left(\sup_{\beta\in B(\tilde{\beta})}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle) - \mathbb{E}\left[\Psi^{(2)}(\langle x,\beta\rangle)\right]\right| \ge 2\left\{k(\|\tilde{\beta}\|_2+\delta)\|\Sigma\|_2+b\right\}\sqrt{\frac{p}{n}}\right) \le e^{-p}.$$

$\blacksquare$

12. Additional Experiments

In this section, we provide additional experiments. The overall setting is the same as Section 8. The only difference is that we change the sampling distribution of the datasets, which are stated in the title of each plot. As in Section 8, SLS estimator outperforms its competitors by a large margin in terms of the computation time.

The results are provided in Figures 7 and 6, and Table 3.
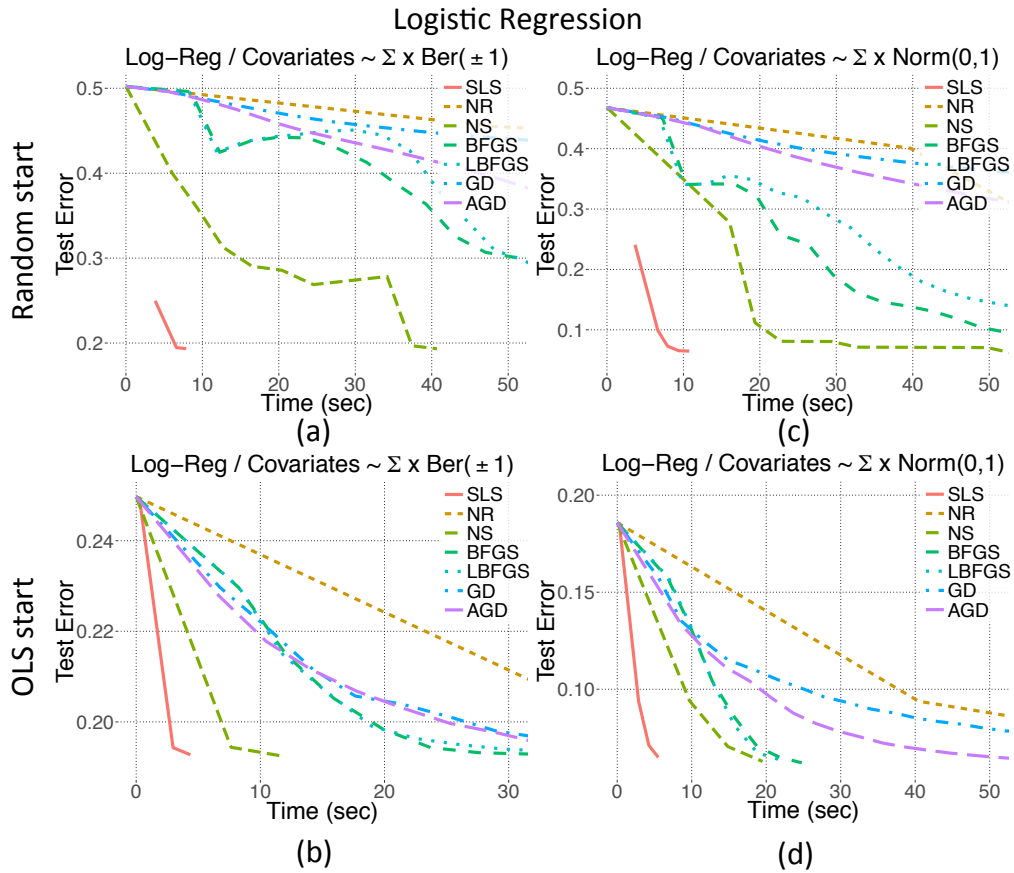
Figure 6: Additional experiments comparing the performance of SLS to that of MLE obtained with various optimization algorithms on several datasets. SLS is represented with red straight line. The details are provided in Table 3

Table 3: Details of the experiments shown in Figures 7 and 6.

| MODEL	LOGISTIC REGRESSION				POISSON REGRESSION			
DATASET	$\mathbf{\Sigma}\times\text{BER}(\pm1)$		$\mathbf{\Sigma}\times\text{NORM}(0,1)$		$\mathbf{\Sigma}\times\{\text{EXP}(1)\text{-}1\}$		$\mathbf{\Sigma}\times\text{NORM}(0,1)$	
SIZE	$n=6.0\times10^5$, $p=300$		$n=6.0\times10^5$, $p=300$		$n=6.0\times10^5$, $p=300$		$n=6.0\times10^5$, $p=300$	
INIT	RND	OLS	RND	OLS	RND	OLS	RND	OLS
PLOT	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
METHOD↓	TIME IN SECONDS / NUMBER OF ITERATIONS (TO REACH MIN TEST ERROR)							
SLS	6.61/3	2.97/3	9.38/5	4.25/4	14.68/4	2.99/4	6.66/10	4.13/10
NR	222.21/6	84.08/3	186.33/6	115.76/4	218.1/6	218.9/4	364.63/9	363.4/9
NS	40.68/10	11.57/3	53.06/9	19.52/4	39.22/6	59.61/4	51.48/10	39.8/10
BFGS	125.83/33	35.41/9	155.3/48	24.78/8	46.61/20	48.71/12	92.84/36	74.22/38
LBFGS	142.09/38	44.41/12	444.62/143	21.79/7	96.53/39	50.56/12	296.4/111	228.1/117
GD	409.9/134	79.45/22	1773.1/509	135.62/44	569.1/211	124.31/48	792.3/344	1041.1/366
AGD	177.3/159	43.76/12	359.56/95	53.73/18	157.9/57	63.16/16	74.74/32	62.21/32
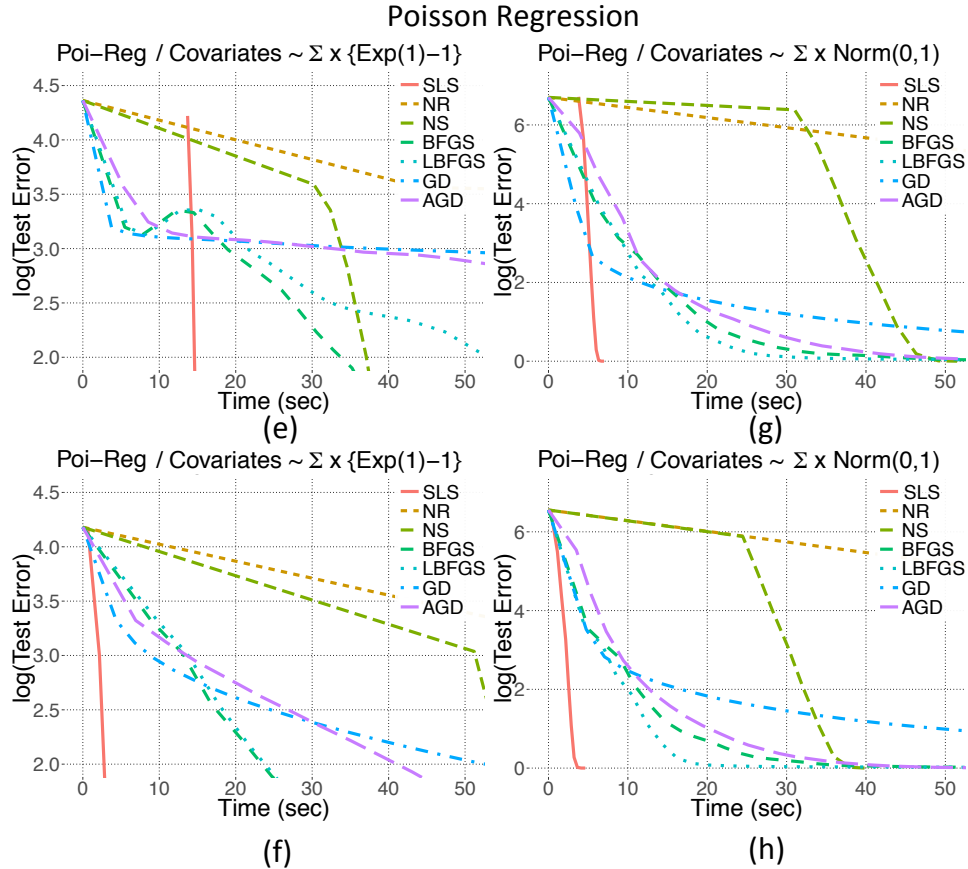
Figure 7: Additional experiments comparing the performance of SLS to that of MLE obtained with various optimization algorithms on several datasets. SLS is represented with red straight line. The details are provided in Table 3

References

Pierre Alquier and Gérard Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(Jan):243–280, 2013.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5, 2014.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari. Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, pages 944–952, 2013.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Comput. Electron. Agr.*, 24:131–151, 1999.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

David R Brillinger. A generalized linear model with "Gaussian" regressor variables. In *A Festschrift For Erich L. Lehmann*, pages 97–114. CRC Press, 1982.

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.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. 2005.

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

Sheng Chen and Arindam Banerjee. Robust structured estimation with single-index models. In *International Conference on Machine Learning*, pages 712–721, 2017.

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*, pages 360–368, 2013.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Petros Drineas, Michael W Mahoney, S Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

Naihua Duan and Ker-Chau Li. Slicing regression: a link-free regression method. *The Annals of Statistics*, pages 505–530, 1991.

Murat A Erdogdu. Newton-stein method: A second order method for glms via stein's lemma. In *Advances in Neural Information Processing Systems*, pages 1216–1224, 2015.

Murat A. Erdogdu. Newton-stein method: an optimization method for glms via stein's lemma. *The Journal of Machine Learning Research*, 17(1):7565–7616, 2016.

Murat A Erdogdu. *Stein's lemma and subsampling in large-scale optimization*. PhD thesis, Stanford University, 2017.

Murat A Erdogdu and Andrea Montanari. Convergence rates of sub-sampled newton methods. In *Advances in Neural Information Processing Systems*, pages 3052–3060, 2015.

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

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugenic*, 7:179–188, 1936.

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

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

Larry Goldstein. l^1 bounds in normal approximation. *Annals Probability*, 35:1888–1930, 2007.

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

Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

Peter Hall and Ker-Chau Li. On almost linearity of low dimensional projections from high dimensional data. *The annals of Statistics*, pages 867–889, 1993.

David P Helmbold, Jyrki Kivinen, and Manfred K Warmuth. Relative loss bounds for single neurons. *IEEE Transactions on Neural Networks*, 10(6):1291–1304, 1999.

Magnus R Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems'. *Journal of Research of the National Bureau of Standards*, 49(6), 1952.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

Bing Li and Yuexiao Dong. Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics*, pages 1272–1298, 2009.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *Annals of Statistics*, 17:1009–1052, 1989.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2370–2378, 2016.

Qiang Liu, Jason D Lee, and Michael I Jordan. A kernelized stein discrepancy for goodness-of-fit tests and model evaluation. *arXiv preprint arXiv:1602.03253*, 2016.

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*. Chapman and Hall, 2nd edition, 1989.

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

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27:372–376, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Christopher C Paige and Michael A Saunders. Solution of sparse indefinite systems of linear equations. *SIAM journal on numerical analysis*, 12(4):617–629, 1975.

Mert Pilanci and Martin J Wainwright. Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250*, 2015.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422, 2010.

Vladimir Rokhlin and Mark Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.

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 J Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, pages 1856–1879, 1989.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Conference on Learning Theory (COLT)*, 2009.

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

Weijie Su and Emmanuel Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

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

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.