Vanilla Lasso for sparse classification under single index models

Jiyi Liu and **Jinzhu Jia**

School of Mathematical Sciences and Center of Statistical Science, LMAM, LMEQF
Peking University
Beijing 100871, China
e-mail: nbljy111@163.com   jzjia@math.pku.edu.cn

Abstract: This paper study sparse classification problems. We show that under single-index models, vanilla Lasso could give good estimate of unknown parameters. With this result, we see that even if the model is not linear, and even if the response is not continuous, we could still use vanilla Lasso to train classifiers. Simulations confirm that vanilla Lasso could be used to get a good estimation when data are generated from a logistic regression model.

1. Introduction

Classification problem is important in a lot of fields such as pattern recognition, bioinformatics etc. There are a few classic classification methods, including LDA (linear discriminant analysis), logistic regression, naive bayes, SVM (support vector machine). When the number of features (or predictors) denoted by p is fixed, under some regularity conditions, LDA is proved to be optimal (see standard statistical text books such as Anderson et al. (1958)). Logistic regression is very popular because it has no distribution assumption on X. It has been proved that LDA is equivalent to least squares Fisher (1936). Using this connection, in fact, one could solve the LDA problem by directly using the vanilla ℓ_1 penalized least squares, i.e. the Lasso (Tibshirani (1996)). Using Lasso to solve sparse LDA has already been proposed by Mai et al. (2012), in which they showed that under irrepresentable condition and some other regularity conditions, the Lasso could select the important features (or predictors) for linear discriminant analysis. ℓ_1 penalized logistic regression has been widely used for high-dimensional classification problems (Koh et al., 2007; Ravikumar et al., 2010).

LDA and logistic regression model are parametric models. Both models require a few assumptions on the data collected. To make less assumptions, non-parametric or semi-parametric models could be used. To control model complexity, single index model is a great choice for semi-parametric models (Powell et al., 1989; Klein and Spady, 1993; Ichimura, 1993). A single-index model is defined as follows.

$$E(y|x) = F(x^T\beta),$$

where F is an unknown function. F is usually estimated via Nadaraya-Watson nonparametric estimator

$$\tilde{F}(x_i^T\beta) = \sum_{j\neq i} y_j K\left(\frac{x_i^T\beta - x_j^T\beta}{h}\right) / \sum_{j\neq i} K\left(\frac{x_i^T\beta - x_j^T\beta}{h}\right).$$

Once the form of F is given, a maximum (quasi)likelihood estimation for β could be given via maximizing the following quasi-likelihood:

$$\sum_{i=1}^{n}(y_i \log(\tilde{F}(x_i^T\beta)) + (1-y_i)\log(1-\tilde{F}(x_i^T\beta))).$$

Nonlinear least squares or Nonlinear weighted least square could also be used to estimated β. Nonlinear weighted least squares estimator is defined as follows.

$$\hat{\beta} = \arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}w(x_i)[y_i - \tilde{F}(x_i^T\beta)]^2,$$

when $w(x_i) = 1$, for all $i = 1, 2, \ldots, n$, it is a nonlinear least squares estimator.

For high-dimensional classification problems, a very demand is to select very few important features. AIC or BIC could be used. Their convex relaxation – ℓ_1 constrained optimization could also be used. In this paper, we propose a new feature selection method for high-dimensional classification problems. We apply vanilla Lasso to this high-dimensional classification problems. We show that if data are from a single index model (Hardle et al., 1993; Cui et al., 2009) with F unknown, then vanilla Lasso could be used to select important features.

The rest of the paper is organized as follows. In Section 2, we give the main results of the paper. Section 3 provides proofs. We use a simulate study to support our theories in Section 4 and finally we conclude in Section 5.

2. Vanilla Lasso for single index models

Suppose observations $(x_i^T, y_i) \in \mathbb{R}^p \times \{-1, 1\}, i = 1, 2, \ldots, n$ i.i.d. follow some unknown distribution. y_i and x_i have the following single index model:

$$E(y_i|x_i) = F(x_i^T\beta), \tag{2.1}$$

where $F(\cdot)$ is an unknown function and β is the unknown parameter that could be used to construct a classifier. If $F(x) = \frac{e^x}{1+e^x}$, then we have logistic regression model; if $F(x) = P(Z \leq x)$, where Z is a standard Gaussian random variable, then we have a probit model. Our goal is to estimate the unknown coefficient β and then construct a classifier. For both logistic regression and probit model the classifier is $sign(x_i^T\beta)$. Plan and Vershynin (2013) first considered this problem as a one-bit compressed sensing problem. They proposed to estimate β by maximizing $\sum_{i=1}^{n} y_i x_i^T\beta$ and gave a few constraints on the solution of β. We argue that a more natural objective function like minimizing least squares could be used. We state our main results as follows.

Theorem 1. *Let* $\beta^* \in \mathbb{R}^n$ *with* $\|\beta^*\|_1 \leq \sqrt{s}, \|\beta^*\|_2 = 1$. *The single index model* (2.1) *is satisfied. Let* $\lambda = E[F(Z)Z]$, *where* Z *is a standard Gaussian random variable. Denote* $\hat{\beta}$ *as the solution of following Lasso problem:*

$$\hat{\beta} = \arg\min_{\beta}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2, \quad s.t. \ \|\beta\|_1 \leq s.$$

Suppose $n \geq C\delta^{-2} s \log(\frac{\lambda^2 p}{2s})$, then with probability $P \geq 1 - 13e^{-c\delta\lambda\sqrt{n}}$:

$$\left\| \frac{\hat{\beta}}{||\hat{\beta}||_2} - \beta^* \right\|_2 \leq \delta.$$

Remark 1. Plan and Vershynin (2015) has a similar result. Here we use a quite different proof skill to prove this result. Our proof depends on standard concentration inequalities.

Remark 2. It is easy to see that parameter β is not identifiable, because $F(\cdot)$ is an unknown function, any scaling factor could be absorbed into $F(\cdot)$. It is reasonable that Theorem 1 controls the error for scaled β. What is worthy noticing is that the restrict of original signal β^* is basically $\frac{||\beta^*||_1}{||\beta^*||_2} \leq \sqrt{s}$, which is called effective sparsity in previous work. We can replace β^* with $\frac{\beta^*}{\|\beta^*\|_2}$ to make it normalized.

3. Proofs

We denote $\lambda = E[F(Z)Z]$, where Z is a standard Gaussian random variable. In any where in the following context, we have to ensure that $\lambda > 0$. Without loss of generality, we assume $\lambda < 1$ if not mentioned.

We can combine the following statements to prove our main result stated in Theorem 1.

Theorem 2. *Let $\beta^* \in \mathbb{R}^n$ with $||\beta^*||_1 \leq \sqrt{s}, ||\beta^*||_2 = 1$, and the single index model (2.1) is satisfied. Let $\hat{\beta}\left(\frac{2}{\lambda}, 1\right)$ to be the solution of the following constrained optimization problem.*

$$\hat{\beta}\left(\frac{2}{\lambda}, 1\right) = \arg\min_{\beta} \|y - X\beta\|_2 \qquad s.t. \ \|\beta\|_1 \leq \frac{2\sqrt{s}}{\lambda}, \|\beta\|_2 = 1.$$

Suppose $n \geq C\delta^{-2} s \log(\frac{\lambda^2 p}{2s})$, then with probability $P \geq 1 - 5e^{-c\delta\lambda\sqrt{n}}$, we have

$$\left\| \hat{\beta}\left(\frac{2}{\lambda}, 1\right) - \beta^* \right\|_2 \leq \delta$$

Theorem 3. *Let $\beta^* \in \mathbb{R}^n$ with $\|\beta^*\|_1 \leq \sqrt{s}, \|\beta^*\|_2 = 1$, and the single index model (2.1) is satisfied. Let Let $\hat{\beta}\left(\frac{2}{\lambda}, 1\right)$ and $\hat{\beta}(1, k)$ to be the solution of the following constrained optimization problems for $k \in [\lambda - \varepsilon, \lambda + \varepsilon]$ and $\varepsilon \leq \frac{\lambda}{2}$.*

$$\hat{\beta}\left(\frac{2}{\lambda}, 1\right) = \arg\min_{\beta} ||y - X\beta||_2 \qquad s.t. \ \|\beta\|_1 \leq \frac{2\sqrt{s}}{\lambda}, \|\beta\|_2 = 1.$$

$$\hat{\beta}(1, k) = \arg\min_{\beta} ||y - X\beta||_2 \qquad s.t. \ \|\beta\|_1 \leq \sqrt{s}, \|\beta\|_2 = k.$$

Then when $n > C\delta^{-2}\lambda^{-2} s \log(\frac{2(\lambda-\varepsilon)^2 p}{s})$, with probability $P \geq 1 - 6e^{-c\delta\lambda\sqrt{n}}$,

$$\left\| \hat{\beta}\left(\frac{2}{\lambda}, 1\right) - \frac{\hat{\beta}(1, k)}{k} \right\|_2 \leq \delta$$

Theorem 4. *Let $\beta^* \in \mathbb{R}^n$ with $\|\beta^*\|_1 \leq \sqrt{s}, \|\beta^*\|_2 = 1$, and the single index model* (2.1) *is satisfied. Let Let $\hat{\beta}$ to be the solution of the following constrained optimization problem.*

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2, \quad s.t. \ \|\beta\|_1 \leq s.$$

Then with probability $P \geq 1 - e^{-\frac{\varepsilon^2 n}{32}} - e^{-\frac{(1-\lambda-\varepsilon)^2 \varepsilon^2}{8} n}$,

$$\lambda - \varepsilon \leq ||\hat{\beta}||_2 \leq \lambda + \varepsilon$$

Here, we show how these three combined can prove Thm 1. We abbreviate $\hat{\beta}_1$ for $\hat{\beta}\left(\frac{2}{\lambda}, 1\right)$ and set $\varepsilon = \min\{\frac{\lambda}{2}, \frac{1-\lambda}{2}\}$. From theorem 4, with $P \geq 1 - e^{-\frac{\varepsilon^2 n}{32}} - e^{-\frac{(1-\lambda-\varepsilon)^2 \varepsilon^2}{8} n}$, the solution $\hat{\beta}_0$ in (2.1) satisfies $\lambda - \varepsilon \leq \|\hat{\beta}_0\|_2 \leq \lambda + \varepsilon$. In this case, the condition in the theorem 3 holds. Combined with the result of theorem 2 setting $k = \frac{1}{\lambda - \varepsilon}$, we conclude

$$\left\| \beta - \frac{\hat{\beta}_0}{\|\hat{\beta}_0\|_2} \right\|_2 \leq \|\beta - \hat{\beta}_1\|_2 + \|\hat{\beta}_1 - \frac{\hat{\beta}_0}{\|\hat{\beta}_0\|_2}\|_2 \leq 2\delta \tag{3.1}$$

The probability for this assertion is

$$\begin{aligned} P &\geq 1 - (1 - P_1) - (1 - P_2) - (1 - P_3) \\ &\geq 1 - 5e^{-c\delta\lambda\sqrt{n}} - 6e^{-c\delta\lambda\sqrt{n}} - e^{-\frac{\varepsilon^2 n}{32}} - e^{-\frac{(1-\lambda-\varepsilon)^2 \varepsilon^2}{8} n} \\ &= 1 - 11e^{-c\delta\lambda\sqrt{n}} - e^{-\frac{\varepsilon^2 n}{32}} - e^{-\frac{(1-\lambda-\varepsilon)^2 \varepsilon^2}{8} n} \end{aligned}$$

Notice in (3.10) we set $\varepsilon = \min\{\frac{\lambda}{2}, \frac{1-\lambda}{2}\}$

$$P \geq 1 - 13e^{-c\delta\lambda\sqrt{n}}$$

The lower bound of n in Theorem 2 and 3 can both be rewritten as $C\delta^{-2} s \log(\frac{\lambda^2 p}{2s})$ under this circumstances, and thus, this bound suffices to conclude our results. Finally, replace δ with $\frac{\delta}{2}$, result in (3.1) is exactly what we want, and thus our proof is completed. We now prove Theorems 2, 3 and 4.

3.1. Proof of Thm 2

Here we give the proof of Theorem 2.

Denote $f(\beta') = \frac{1}{n}(\sum_{i=1}^{n} y_i \langle x_i, \beta' \rangle - \frac{1}{2} \sum_{i=1}^{n} \langle x_i, \beta' \rangle^2)$. According to the main theorem in Plan and Vershynin (2013), set K=$\{\beta \in \mathbb{R}^n : \|\beta\|_2 = 1, \|\beta\|_1 \leq j\sqrt{s}\}$ there, and we can conclude that when $\|\beta - \beta'\|_2 \geq \delta$, with $P \geq 1 - 4e^{-2\gamma^2}$,

$$\frac{1}{n}(\sum_{i=1}^{n} y_i \langle x_i, \beta \rangle - \sum_{i=1}^{n} y_i \langle x_i, \beta' \rangle) \geq \frac{\lambda\delta}{2} - \frac{\sqrt{Cj^2 s \log(\frac{2p}{j^2 s})} + 4\gamma}{\sqrt{n}} \tag{3.2}$$

Noticing that $\sum_{i=1}^{n} \langle x_i, \beta' \rangle^2 \sim \|\beta'\|_2^2 \chi^2(n)$, which belongs to sub-exponential distribution defined as follows.

Definition 1 (Sub-exponential). *A random variable X is sub-exponential with parameter (σ, b), if for all $|\lambda| < 1/b$,*

$$E\left[\exp\left[\lambda\left(X - E(X)\right)\right]\right] \leq \exp\left(\frac{1}{2}\sigma^2\lambda^2\right) \tag{3.3}$$

It is not hard to check that the parameter (a, b) for $\chi^2(n)$ can be chosen as $(\sqrt{8n}, 4)$, and the rescaling of a variable will not influence the parameter a. We state the following concentration result, which could be found from Boucheron et al. (2013).

Theorem 5. *For sub-exponential random variable X with parameter (a,b),*

$$P(X - EX \geq t) \leq e^{max\{-\frac{t^2}{2a^2}, -\frac{t}{2b}\}} \tag{3.4}$$

Here we donote $X = x'^T A^T A x'$ and $Y = x^T A^T A x$, then $EX = EY = n$, for some ε

$$
\begin{aligned}
P(X - Y \geq -\varepsilon n^{\frac{3}{4}}) &= P(Y - X \leq \varepsilon n^{\frac{3}{4}}) \\
&= 1 - P(Y - X \geq \varepsilon n^{\frac{3}{4}}) \\
(3.4) &\geq 1 - e^{\max\{-\frac{\varepsilon n^{\frac{3}{4}}}{8}, -\frac{\varepsilon^2 n^{\frac{1}{2}}}{16}\}}
\end{aligned}
$$

So with high probability,

$$
\begin{aligned}
f(\beta) - f(\beta') &= \frac{1}{n}\left(\sum_{i=1}^{n} y_i \langle x_i, \beta \rangle - \sum_{i=1}^{n} y_i \langle x_i, \beta' \rangle\right) + \frac{1}{2n}(Y - X) \\
&\geq \frac{\lambda\delta}{2} - \frac{\sqrt{Cj^2 s \log(\frac{2p}{j^2 s})} + 4\gamma}{\sqrt{n}} - \frac{\varepsilon}{2\sqrt[4]{n}} \\
&:= \frac{\lambda\delta}{4} - \frac{a}{\sqrt{n}} - \frac{\varepsilon}{2\sqrt[4]{n}}.
\end{aligned} \tag{3.5}
$$

There, for convenience's sake, we set $\gamma = \frac{\lambda\delta\sqrt{n}}{8}$.

When $n > \left(\frac{\varepsilon + \sqrt{\varepsilon^2 + 4\lambda\delta a}}{\lambda\delta}\right)^4$, the expression in (3.5) > 0, so such x' could not be a optimizer since x is also a feasible point. Thus, our original assumption that $\|\beta - \beta'\|_2 \geq \delta$ cannot hold. This proves what we want.

Now we calculate the so-called "high probability". We take $\varepsilon = \sqrt{\lambda\delta a}$ here. This statement holds when both parts in (3.2) and (3.4) hold. So P$\geq 1 - 4e^{-2\gamma^2} - e^{\max\{-\frac{\varepsilon n^{\frac{3}{4}}}{8}, -\frac{\varepsilon^2 n^{\frac{1}{2}}}{16}\}}$. It is sufficient to set

$$n \geq 4\left(\frac{\varepsilon + \sqrt{\lambda\delta a}}{\lambda\delta}\right)^4 = C\lambda^{-2}\delta^{-2}a^2$$

and in this case, we have

$$P \geq 1 - 4e^{-2\gamma^2} - e^{\max\{-c\gamma^{\frac{3}{2}}, -c'\gamma\}}$$
$$\geq 1 - 4e^{-2\gamma^2} - e^{-c\gamma}$$
$$\geq 1 - 5e^{-c\gamma}$$
$$= 1 - 5e^{-c\delta\lambda\sqrt{n}}.$$

3.2. Proof of Theorems 3

Second, we come to Theorem 3.
In this section, we abbreviate $\hat{\beta}_1$ as $\hat{\beta}(\frac{2}{\lambda}, 1)$, and $\hat{\beta}_k$ as $\hat{\beta}(1, k)$. First, notice that $\frac{1}{k}\hat{\beta}_k$ is feasible for problem

$$\min_{\beta} ||y - X\beta||_2 \qquad s.t. \ ||\beta||_1 \leq \frac{2\sqrt{s}}{\lambda}, ||\beta||_2 = 1.$$

and $k\hat{\beta}_1$ is feasible for

$$\min_{\beta} ||y - X\beta||_2 \qquad s.t. \ ||\beta||_1 \leq \sqrt{s}, ||\beta||_2 = k.$$

So, from the definition of optimizer, we can conclude

$$f(\hat{\beta}_k) \geq f(k\hat{\beta}_1) \tag{3.6}$$

$$f(\hat{\beta}_1) \geq f(\frac{1}{k}\hat{\beta}_k) \tag{3.7}$$

Denote $a = \frac{1}{n}\sum_{i=1}^{n} y_i \langle x_i, \frac{\hat{\beta}_k}{k} \rangle$ and $b = \frac{1}{2n}\sum_{i=1}^{n} \langle x_i, \frac{\hat{\beta}_k}{k} \rangle^2$ here. And, we similarly define a', b' for $\hat{\beta}_1$. Now $f(\hat{\beta}_k) = ka - k^2 b$, $f(k\hat{\beta}_1) = ka' - k^2 b'$.

Then the conditions shown above can be translated to

$$ka - k^2 b \geq ka' - k^2 b'$$

$$a' - b' \geq a - b$$

We deduce as following:
$$a - b = a - kb + (k-1)b$$
$$\geq a' - kb' + (k-1)b \tag{3.8}$$
$$= a' - b' + (1-k)(b' - b)$$

Which indicates
$$0 \leq (a' - b') - (a - b) \leq (1-k)(b - b')$$
$$\Rightarrow \quad b \geq b'$$

Since $b - b'$ is sub-exponential, it cannot be too large, which indicates $\|f(\hat{\beta}_1) - f(\frac{\hat{\beta}_k}{k})\|_2$ could not be too large. Thus, the same as Section 3.1, $\hat{\beta}_1$ and $\frac{\hat{\beta}_k}{k}$ cannot be far away. We take this idea down to the ground.

According to (3.4) and (3.5), with $P \geq 1 - e^{\max\{-\frac{\varepsilon n^{\frac{3}{4}}}{8}, -\frac{\varepsilon^2 n^{\frac{1}{2}}}{16}\}}$, we have

$$b - b' \leq \frac{\varepsilon}{2\sqrt[4]{n}} \tag{3.9}$$

When $\|\hat{\beta}_1 - \frac{\hat{\beta}_k}{k}\| \geq \delta$, as shown in Section 3.1, with $P \geq 1 - 5e^{-c\delta\lambda\sqrt{n}}$

$$\|\beta - \hat{\beta}_1\|_2 \leq \frac{\delta}{2}$$

$$\Rightarrow \|\beta - \frac{\hat{\beta}_k}{k}\|_2 \geq \frac{\delta}{2}$$

According to (3.5)

$$f(\beta) - f(\frac{\hat{\beta}_k}{k}) \geq \frac{\lambda\delta}{8} - \frac{a}{\sqrt{n}} - \frac{\varepsilon}{2\sqrt[4]{n}}$$

We choose

$$\varepsilon = \min\{\frac{\lambda}{2}, \frac{1-\lambda}{2}\} \leq \frac{\lambda\delta}{8}\sqrt[4]{n} - \frac{a}{\sqrt[4]{n}} \tag{3.10}$$

combined with (3.9)

$$\frac{\lambda\delta}{8} - \frac{a}{\sqrt{n}} - \frac{\varepsilon}{2\sqrt[4]{n}} \geq \frac{\varepsilon}{2\sqrt[4]{n}} > (1-k)(b-b') \tag{3.11}$$

Thus the assumption of $\|\hat{\beta}_1 - \frac{\hat{\beta}_k}{k}\| \geq \delta$ could not hold with $P \geq 1 - 5e^{-c\delta\lambda\sqrt{m}}$. On the inverse overally, we can conclude $\|\hat{\beta}_1 - \frac{\hat{\beta}_k}{k}\| \leq \delta$ with probability

$$P \geq 1 - e^{\max\{-\frac{\varepsilon n^{\frac{3}{4}}}{8}, -\frac{\varepsilon^2 n^{\frac{1}{2}}}{16}\}} - 5e^{-c\delta\lambda\sqrt{n}}$$
$$\geq 1 - 6e^{-c\delta\lambda\sqrt{n}}$$

3.3. Proof of Theorems 4

Here in Section 3.3, for every β', we denote $\beta' = \mu p$, where $\mu = \|\beta'\|_2$, and $g(\mu) = f(\beta')$ for a fixed p.

First, we prove $\|\hat{x_0}\|_2 \leq \lambda + \varepsilon$. The basic idea of our proof is that, since the optimal function is quadratic, it has a maximum point. When searching for the overall max point, the 2-norm of it will spontaneously come to a given point-λ. We would like to search for another point with larger function value if not so.

Notice $g'(\mu) = \frac{1}{n}(\sum\limits_{i=1}^{n} y_i\langle x_i, p\rangle - \mu \sum\limits_{i=1}^{n}\langle x_i, p\rangle^2)$. When $\mu > \lambda + \varepsilon$, noticing that $ng'(\mu)$ is sub-exponential, we conclude through concentration inequality:

$$
\begin{aligned}
P(g'(\mu) \leq -c) &= P(ng'(\mu) \leq -cn) \\
&= 1 - P(ng'(\mu) \geq -cn) \\
&= 1 - P\big(ng'(\mu) \geq E(g'(\mu)) - cn - E(\sum_{i=1}^{n} y_i\langle a_i, p\rangle - \mu \sum_{i=1}^{n}\langle a_i, p\rangle^2)\big) \\
&= 1 - P\big(ng'(\mu) \geq E(g'(\mu)) - cn - (\lambda - \mu)n\big) \\
&= 1 - P\big(ng'(\mu) - E(g'(\mu)) \geq (\mu - \lambda - c)n\big) \\
&\geq 1 - e^{\max\{-\frac{(\mu-\lambda-c)^2 n}{8}, -\frac{(\mu-\lambda-c)n}{8}\}} \\
&= 1 - e^{-\frac{(\mu-\lambda-c)^2 n}{8}} \\
&\geq 1 - e^{-\frac{\varepsilon^2 n}{32}}
\end{aligned}
$$

In the final row, we actually choose $c = \frac{\varepsilon}{2}$.

With this probability, for any β' with $\|\beta'\|_2 \geq \lambda + \varepsilon$, we can find another new feasible point through the line: $y = t\beta'$. Specifically, we can find $\bar{\beta} = \mu'p$, where $\mu' < \mu$, to satisfy $\bar{\beta}$ satisfies the restriction and $f(\bar{\beta}) > f(\beta')$. The feasibility of $\bar{\beta}$ is quite obvious. So we have $\|\hat{\beta}_0\|_2 \leq \lambda + \varepsilon$.

On the other hand, we prove $\|\hat{\beta}_0\|_2 \geq \lambda - \varepsilon$. This time, we go to the following proposition

with $P \geq 1 - e^{\frac{(1-\lambda-\varepsilon)^2 \varepsilon^2}{8} n}$, $f(\hat{\beta}_\lambda) > f(\beta)$, where β is the solution to

$$
\beta = \arg\min_{\beta'} \sum_{i=1}^{n}(y_i - x_i^T \beta')^2, \quad s.t. \ \|\beta'\|_1 \leq s, \ \|\beta'\|_2 \leq \lambda - \varepsilon
$$

then, with $P \geq 1 - e^{\frac{(1-\lambda-\varepsilon)^2\varepsilon^2}{8} n}$, the solution of non$-$scale cannot lie into the disk of $\lambda - \varepsilon$

Assume $\|\beta\|_2 = k$, *then* $k \leq \lambda - \varepsilon$ and according to (3.4).

$$
\begin{aligned}
P(f(\hat{\beta}_\lambda) - f(\beta) > c) &= P(f(\hat{\beta}_\lambda) - f(\beta) - E(f(\hat{\beta}_\lambda) - f(\beta)) > c - E(f(\hat{\beta}_\lambda) - f(\beta))) \\
&= 1 - P(f(\beta) - f(\hat{\beta}_\lambda) - E(f(\beta) - f(\hat{\beta}_\lambda)) \geq E(f(\hat{\beta}_\lambda) - f(\beta)) - c) \\
&= 1 - P(f(\beta) - f(\hat{\beta}_\lambda) - E(f(\beta) - f(\hat{\beta}_\lambda)) \geq Ef(\hat{\beta}_\lambda) - Ef(\beta) - c) \\
&= 1 - P(f(\beta) - f(\hat{\beta}_\lambda) - E(f(\beta) - f(\hat{\beta}_\lambda)) \geq \lambda - k - \frac{1}{2}(\lambda^2 - k^2) - c) \\
&= 1 - P(nf(\beta) - nf(\hat{\beta}_\lambda) - E(nf(\beta) - nf(\hat{\beta}_\lambda)) \geq (\lambda - k)(1 - \frac{\lambda + k}{2})n - cn) \\
&\geq 1 - e^{\max\{-\frac{((\lambda-k)(1-\frac{\lambda+k}{2})-c)^2}{8}n, -\frac{(\lambda-k)(1-\frac{\lambda+k}{2})-c}{8}n\}} \\
&= 1 - e^{-\frac{((\lambda-k)(1-\frac{\lambda+k}{2})-c)^2}{8}n} \\
&\geq 1 - e^{-\frac{(\varepsilon(2-2\lambda-\varepsilon)-2c)^2}{32}n}
\end{aligned}
$$

Thus, we set $c = \frac{\varepsilon^2}{4}$ and conclude

$$
P(f(\hat{\beta}_\lambda) - f(\beta) > \frac{\varepsilon^2}{4}) \geq 1 - e^{-\frac{(1-\lambda-\varepsilon)^2\varepsilon^2}{8}n}
$$

So, with this probability, the optimizer of non-scale cannot lie inside the $\lambda - \varepsilon$ disc. Finally, we combine these two parts to calculate the overall probability

$$
P \geq 1 - e^{-\frac{\varepsilon^2 n}{32}} - e^{-\frac{(1-\lambda-\varepsilon)^2\varepsilon^2}{8}n}.
$$

Till now, all proof is finished. □

4. Simulation

In this section, we use simulation study to illustrate our theoretical findings. We first demonstrate that when data are from a logistic model which is a special single index model by choosing $F(t) = \frac{e^t}{1+e^t}$ in model (2.1) . In this special model, the classifier is defined as

$$
\ell(x) = sign(x^T \hat{\beta}),
$$

where $\hat{\beta}$ could be estimated via vanilla Lasso – ℓ_1 constrained least squares. For a new data point with observed features x, we calculate $\ell(x)$ and claim it as labelled +1 if $\ell(x) = 1$ or else, we claim it as labelled -1. From this classifier, we see that the scale $\hat{\beta}$ does not make any contribution to the classifier. In fact, we will show later that β^* could not be estimated correctly, but $\frac{\beta^*}{\|\beta^*\|_2}$ could.

Let β^* be a fixed s-sparse vector in $\mathbb{R}^p$, and $X_{n \times p}$ is an $n \times p$ matrix with entries from i.i.d Gaussian distribution where n is the number of samples. Note that in this section, we set β^* as any s-sparse vector, so the 2-norm of it is probably not 1. We generate a response vector y from the following model:

$$
y_i \in \{-1, 1\}
$$

$$P(y_i = 1 | x_i) = \frac{1}{1 + \exp\{x_i^T \beta\}}$$

$$P(y_i = -1 | x_i) = 1 - P(y_i = 1 | x_i)$$

Here, we set $p = 1200$, $s = 10$, and n from 200 to 3000. Matrices X are generated for each n. Results are showed below, in which x-axis stands for $200n$ and y-axis for error $||\frac{\beta^*}{\|\beta^*\|_2} - \frac{\hat{\beta}}{||\hat{\beta}||_2}||_2$:
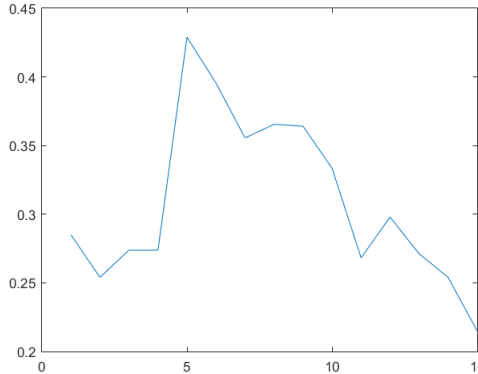

FIG 1. *Relationship of n and error of β.*

As we can see, the error declines fairly sharply as sample size n grows. We also show the differences between $\hat{\beta}$ and β^*. From Figure 1 we see that even when n is very big, the difference does not vanish, which suggests that $\hat{\beta}$ is not a good estimator for β^*, while Figure 1 suggests that $\hat{\beta}/\|\hat{\beta}\|_2$ is a good estimator of $\beta^*/\|\beta^*\|_2$.

5. Conclusions

In this paper, we studied how Lasso performs for a sparse classification problem when data from a single-index model. Vanilla Lasso is very simple and is computationally very fast. For high-dimensional data, it is much more convenient than using semi-parametric method directly. We show that it performs quite well for high-dimensional binary classification problems. If one classifier like logistic regression or probit regression model does not depend on the scale of β, vanilla Lasso could give direct classifiers. If one does want to use more complex models, vanilla lasso could be used as a first step to remove huge amount of redundant features.

References

Anderson, T. W., T. W. Anderson, T. W. Anderson, and T. W. Anderson (1958). *An introduction to multivariate statistical analysis*, Volume 2. Wiley New York.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence.* OUP Oxford.

Cui, X., W. Härdle, and L. Zhu (2009). Generalized single-index models.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics 7*(2), 179–188.

Hardle, W., P. Hall, H. Ichimura, et al. (1993). Optimal smoothing in single-index models. *The annals of Statistics 21*(1), 157–178.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics 58*(1), 71–120.

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, 387–421.

Koh, K., S.-J. Kim, and S. P. Boyd (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research 8*(8), 1519–1555.

Mai, Q., H. Zou, and M. Yuan (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42.

Plan, Y. and R. Vershynin (2013). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on 59*(1), 482–494.

Plan, Y. and R. Vershynin (2015). The generalized lasso with non-linear observations. *arXiv preprint arXiv:1502.04071*.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 1403–1430.

Ravikumar, P., M. J. Wainwright, J. D. Lafferty, et al. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics 38*(3), 1287–1319.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.