On the Consistency of Feature Selection With Lasso for Non-linear Targets

Yue Zhang[†] YUE.ZHANG13@CASE.EDU
Soumya Ray[‡] SRAY@CASE.EDU
Weihong Guo[†] WEIHONG.GUO@CASE.EDU
[†]Department of Mathematics, Applied Mathematics and Statistics
[‡]Department of Electrical Engineering and Computer Science
Case Western Reserve University, Cleveland, OH 44106, USA

Abstract

An important question in feature selection is whether a selection strategy recovers the "true" set of features, given enough data. We study this question in the context of the popular Least Absolute Shrinkage and Selection Operator (Lasso) feature selection strategy. In particular, we consider the scenario when the model is misspecified so that the learned model is linear while the underlying real target is nonlinear. Surprisingly, we prove that under certain conditions, Lasso is still able to recover the correct features in this case. We also carry out numerical studies to empirically verify the theoretical results and explore the necessity of the conditions under which the proof holds.

1. Introduction

Feature selection is an extremely important part of machine learning algorithms. Finding a good set of features reduces overfitting, improves robustness to noise and enables faster convergence to the target. Various strategies have been proposed in the literature for feature selection, including filter-based, wrapper-based and embedded methods. In this work, we are interested in embedded feature selection methods. Here, the learning objective for a classifier is modified by introducing an extra term which typically encourages "sparsity" in the solution. If the hypothesis space being explored is linear, this means that many coefficients associated with the features will be zero. Thus they will be eliminated from the learned model.

One of the most well known examples of the embedded approach is the Least Absolute Shrinkage and Selection Operator, or Lasso (Tibshirani, 1996). In the Lasso, feature sparsity is encouraged through the addition of a $\lambda\|w\|_1$ term, where λ is a coefficient that trades off the importance of the loss and Lasso terms. This L_1 norm term is the best smooth approximation to the L_0 norm, which directly counts nonzero feature coefficients. Further, the resulting optimization problem remains convex if the loss term is convex, so that there is a well-defined global optimum. Finally, recent advances in convex optimization have made great strides in designing efficient and effective solution methods for objective functions extended with the Lasso term (Boyd, 2010). As a result, the Lasso is one of the most widely used embedded feature selection strategies in machine learning, and has consistently shown good empirical results.

In this paper, we are interested in theoretically characterizing the behavior of the Lasso. For any feature selection strategy, a key question is whether, given enough data, it can recover the "true" underlying set of features for different target concepts. This question can be expressed through *selection consistency*: given enough data, can we be sure that the set of nonzero coefficients in the learned solution will be the same as that in the target? Selection consistency is clearly a good property to be able to guarantee for a feature selection strategy.

Typically, we are interested in the selection consistency of a feature selection technique under *model misspecification.* After all, in general we do not know what the target concept looks like. Thus in general, the hypothesis space explored during learning may not contain the target. A key question then is whether a feature selection strategy can be selection consistent even in this case.

In this work, we focus on the Lasso feature selection operator, when the hypothesis space is linear (commonly used in many machine learning applications), and ask if the procedure is selection consistent when the underlying target is *nonlinear*. Building on prior work (Brillinger, 1982; Plan & Vershynin, 2015), we prove that, for certain data dis-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

tributions and nonlinear targets, the answer is affirmative. This seems a surprising result and may further explain the empirical success of the Lasso. This result extends to the group Lasso as well. We then perform numerical studies to support the theoretical results and investigate to what extent the theoretical assumptions we make to prove our result are necessary.

The paper is organized as follows: in section 2, we set up our model and describe our basic assumptions, and then discuss related work. Section 3 contains the main result. Numerical experiments are shown in section 4, and we discuss the results and conclude in section 5.

2. Definitions, Notation and Related Work

Model setup. Assume we are given n independent observations (x_i, y_i), $i = 1, 2, ..., n$ which are generated by some non-linear regression model:

$$y_i = g(x_i^T w) + \epsilon_i, \quad i = 1, 2, ..., n,$$

where $\epsilon_i's$ are i.i.d. Gaussian random variables, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, function $g : R \to R$ is a non-linear mapping function which is not known *a priori*, $x_i's \in R^p$ are i.i.d. feature vectors. In this work, following prior work in this area (Brillinger, 1982), we assume that the x_i's are generated from an underlying Gaussian distribution, $x_i \sim \mathcal{N}(0, \Sigma)$. $w \in \mathbb{R}^p$ is the weight vector we want to recover. We assume w has unit l_2-norm, $\|w\|_2 = 1$. Without loss of generality, we assume $w = (w_1, w_2, ..., w_q, w_{q+1}, ..., w_p)^T$, where $w_j \neq 0$ for $j = 1, ..., q$, and $w_j = 0$ for $j = q+1, ..., p$. Let $w_r = (w_1, w_z, ..., w_q)^T$ and $w_z = (w_{q+1}, ..., w_p)^T$ denote the non-zero and zero parts of w respectively. Here the subscripts r and z can be read as "representable parts" and "zeros".

Let $y = (y_1, y_2, ..., y_n)^T$, $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ and X be the $n \times p$ data matrix whose i^{th} row is x_i^T. We consider the following feature selection model:

$$\min_w \|Xw - y\|_2^2 + \lambda f(w).$$

f: $R^p \to R$ is a convex regularization function, which is $\|\cdot\|_1$ for classical Lasso and $\|\cdot\|_{1,2}$ for group Lasso. Here, for a given vector $w \in \mathbb{R}^p$ parsed into m groups (not necessarily equal size), the $l_{1,2}$ norm is defined as $\|w\|_{1,2} := \sum_j^m \|w_j\|_2$.

The solution $\hat{w}^n$ to this model is defined as:

$$\hat{w}^n = \operatorname*{argmin}_w \|Xw - y\|_2^2 + \lambda^n f(w). \quad (*)$$

We use the superscript n to emphasize that the solution of the Lasso may depend on the number of the observations. Likewise, the regularization parameter λ may depend on n as well, and we use λ^n when we wish to make this dependence explicit. We next formalize the notion of *consistency*, which is used to evaluate the goodness of the technique.

Definition 2.1 (Estimation Consistency)**.** *The solution $\hat{w}^n$ obtained from (*) is called estimation consistent if*

$$\|\hat{w}^n - w\|_2 \to_p 0, \quad n \to \infty.$$

Here $\to_p$ means converges in probability.

Definition 2.2 (Selection Consistency)**.** *The solution $\hat{w}^n$ obtained from (*) is called selection consistent if*

$$P(supp(\hat{w}^n) = supp(w)) \to 1, \quad n \to \infty,$$

where $supp(w) = \{i | w_i \neq 0\}$ is the support of w.

Note that one consistency result does not necessarily imply the other. For example, consider $\hat{w}^n = (1, 2, \frac{1}{n}, \frac{1}{n^2}, ...)$ and $w = (1, 2, 0, 0, ...)$. Then $\|w - \hat{w}^n\|_2$ can be small enough such that $\hat{w}^n$ is estimation consistent with w, but the supports are different for any arbitrary large n. One can also easily construct an example that is selection consistent but not estimation consistent.

We now introduce a set of sufficient conditions that will allow us to guarantee selection consistency for the Lasso even though the underlying model is generated by some unknown nonlinear link function.

Assumptions. Using the notation of representable parts and zeros above, we write the data covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ as:

$$\Sigma = \begin{pmatrix} \Sigma_{rr} & \Sigma_{rz} \\ \Sigma_{zr} & \Sigma_{zz} \end{pmatrix}$$

We assume Σ is invertible and Σ_{rr} has bounded positive eigenvalues away from 0, that is, $0 < \Lambda_{min} \leq \Lambda(\Sigma_{rr}) \leq \Lambda_{max} < \infty$, for some constants Λ_{min} and Λ_{max}. Here $\Lambda(\Sigma_{rr})$ denotes the eigenvalues of Σ_{rr}. Furthermore, We assume the following:

- y has finite fourth moment $E(y)^4 < \infty$;
- The link function g is differentiable almost everywhere and $E(|g(t)|) < \infty$ and $E(|g'(t)|) < \infty$, for $t \sim \mathcal{N}(0, 1)$;
- $E(x_j^T x_j |g(x_i^T w)|^2) < \infty$, for $j = 1, 2, ..., n$.

The last two assumptions are closely related to the existence of a practical solution as well as sufficient for accomplishing the desired result. We show in Section 4 that for some experimental functions that violate these assumptions, the Lasso fails to select the right features.

2.1. Related Work

The Lasso selection model (Tibshirani, 1996) has been intensively studied in the past two decades. In the context of signal processing, this approach corresponds to basis pursuit, pioneered by (Chen et al., 1998). When the data has no noise, a solid theoretical foundation has been built by many researchers, e.g. (Feuer & Nemirovski, 2003; Donoho & Tanner, 2005; Cands et al., 2006; Donoho, 2006).

The closest line of work to ours starts with (Knight & Fu, 2000), where the authors analyze the *estimation consistency* conditions for various Lasso type models: the regularization term $f(w)$ is not only l_1 norm, but any l_q norm with $q \in (0, 2]$. In particular, they show that by choosing $\lambda \sim n^{1/2}$, Lasso tends to select the true model with non-vanishing probability as n grows. A sufficient condition to guarantee the selection consistency of the Lasso, namely *irrepresentable condition*, was independently proposed by (Zhao & Yu, 2006) and (Meinshausen & Bhlmann, 2006). The necessity of this condition was proved in later work (Zou, 2006; Yuan & Lin, 2007). A precise characterization of the relation between the regularization parameter λ, n and p (the dimension of w) to guarantee consistent selection is shown in (Wainwright, 2009). In order to select the true features with high probability, one should expect $\lambda = \Omega((\log p^{-1/2})n^{1/2})$. While these results are substantially based on sparse linear regression, consistency results on other generalized linear models such as logistic regression have also been developed (Bunea, 2008; Ravikumar et al., 2010). In the context of non-linear regression, (Tateishi et al., 2010) show the advantage of the Lasso through numerical experiments when the mapping function is linearized by finite Gaussian basis functions.

Prior work (Cands & Recht, 2013; Negahban et al., 2012) provides a systematic approach to analyze the estimation consistency of general sparse models. (Negahban et al., 2012) introduces the notion of *restricted strong convexity*, a property that guarantees nice curvature structure of the loss function near the true features, and establishes a series of estimation error bounds on sparse regression models. This framework is extended to selection consistency analysis by (Lee et al., 2015).

Much of the analysis mentioned above relies on the target regression function being known to be a linear or logistic regression relation, even though obviously the Lasso model itself does not impose any prior knowledge of such. Recent work (Thrampoulidis et al., 2015) has investigated the case of nonlinear link functions and presented consistency results for the Lasso. However, their results are only precise in the context of the estimation consistency which is not directly applicable in feature selection perspective. The work we present does not impose any prior knowledge on the form of the regression function beyond the qualifications outlined above, and is applicable to the standard, fixed-dimensional feature selection case with a fixed and unknown subset of relevant features.

3. Selection Consistency of the Lasso

Before we proceed to our main results, we first introduce some concepts and present several useful lemmas.

3.1. Theoretical preparations

In prior work (Brillinger, 1982) [in section 3, theorem 1], it is shown that even when the observed $y_i' s$ are generated by some unknown link function g, under certain assumptions, the least squares estimator with linear regression fit is *asymptotically* centered around the true predictor times a scaling constant.

Theorem 3.1. *(Brillinger, 1982) Suppose the assumptions in Section 2 hold. Let $y_i = g(x_i^T w) + \epsilon_i$, $i = 1, ..., n$, $x_i' s$ are independent normals with mean 0 and non-singular covariance matrix Σ, $\epsilon_i' s$ are independent of $x_i' s$ and have finite variance σ^2. Let $\hat{w}$ be the ordinary least squares estimator, i.e.,*

$$\hat{w} = \arg\min_w \|Xw - y\|_2^2.$$

Then $\sqrt{n}(\hat{w} - \mu w)$ is asymptotically normal with mean 0 and covariance matrix

$$\sigma^2\Sigma^{-1} + \Sigma^{-1}E\{h(x)^2 xx^T\}\Sigma^{-1},$$

where $h(x) = g(x^T w) - \mu x^T w - \gamma$, $\gamma = E\{g(x^T w) - \mu x^T w\}$ and $\mu = Cov\{g(x^T w), x^T w\}/Var\{x^T w\}$. Furthermore, if w is scaled properly such that $\|\sqrt{\Sigma} w\| = 1$, then $\mu = E[tg(t)]$, $t \sim \mathcal{N}(0, 1)$.

This implies the following result.

Corollary 3.1.1. *For the ordinary least squares with linear observations, $\mu = E(t^2) = 1$ with $t \sim \mathcal{N}(0, 1)$ and $h(x) = 0$, we then have the classical asymptotic estimation on least squares solution*

$$\sqrt{n}(\hat{w} - w) \to_d \mathcal{N}(0, \sigma^2\Sigma^{-1}),$$

where '$\to_d$' means convergence in distribution.

In order to be able to select the right features, the intuition is that the irrelevant features cannot be highly correlated to the relevant features. The following definition is the same as the *strong irrepresentable condition* proposed by (Zhao & Yu, 2006) which describes this property quantitatively. To be consistent with the framework in this paper, we call this strong $\mu-$irrepresentable condition:

Definition 3.1. *(Strong $\mu-$irrepresentable condition). We say that strong $\mu-$irrepresentable condition holds, if there* *exists a constant $s \in (0,1]$, such that*

$$|\Sigma_{zr}(\Sigma_{rr})^{-1}\,\mathbf{sign}(\mu w)| \leq 1-s.$$

Here Σ_{zr} and Σ_{rr} are the sub covariance matrices of $x_i' s$ defined in Section 2 and the "$\leq$" holds element-wise.

Here the $\mathbf{sign}(\cdot)$ denotes an element-wise application of the standard sign function.

The following proposition (Plan & Vershynin, 2015) [Section 4] indicates that after appropriate scaling, even though the discrepancy $y - \mu X w$ may generally depend on both X and w, the projection of it onto X is well behaved in the sense that the expectation of the projection is zero.

Proposition 3.2. *(Plan & Vershynin, 2015) Let $\tilde{X} = XQ^{-1}$, $\tilde{z}_i = g(\tilde{x}_i^T w) - \tilde{\mu}\tilde{x}_i^T w$, $i = 1, 2, ..., n$ where $Q^TQ = \Sigma$, $\tilde{\mu} = Cov(g(\tilde{x}^T w), \tilde{x}^T w)/Var(\tilde{x}^T w)$, $\tilde{x}_i$ is the i^{th} row of $\tilde{X}$, then*

$$E(\tilde{X}^T\tilde{z}) = 0.$$

Note that in the linear setting when $y = Xw + \epsilon$, z becomes ϵ which is independent of the columns of X, and thus proposition 3.2 holds naturally. It can be verified that the proposition fails to hold without the transformation on X. [1]

Note that since we require $E(y)^4 < \infty$, we will be able to use a bootstrap strategy to estimate the covariance matrix by re-sampling the data $x_i' s$ and thus make confidence intervals on $\hat{w}$ for large enough n.

3.2. Main Result

With the above preparation, given our prior assumptions in Section 2, we consider the solution $\hat{w}^n$ to the generalized Lasso, where in this paper we assume $f(x)$ is either $\|\cdot\|_1$ for the classical Lasso or $\|\cdot\|_{1,2}$ for the group Lasso:

$$\hat{w}^n = \underset{w}{\operatorname{argmin}}\, \|Xw - y\|^2 + \lambda^n f(w), \qquad (*)$$

where $y_i = g(x_i^T w) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, X is the data matrix with i.i.d. rows and the i^{th} row x_i^T satisfies $x_i \sim \mathcal{N}(0, \Sigma)$, $i = 1, 2, ..., n$.

First, consider the case $f(w) = \|w\|_1$. Let X_r be the first q columns of X corresponding to the non-zero entries in w and X_z be the rest $p - q$ columns. Then the following probability events hold:

Proposition 3.3. *Assume the strong $\mu-$ irrepresentable condition holds for some constant $s > 0$. Then for large enough n, we have*

$$P(\mathbf{sign}(\hat{w}) = \mathbf{sign}(\mu w)) \geq P(\Omega_1 \cap \Omega_2),$$

with

$$\Omega_1 = \{|\Sigma_{rr}^{-1} D_r| < \sqrt{n}(|\mu w_r| - \frac{\lambda^n}{2n}\Sigma_{rr}^{-1}\,\mathbf{sign}(\mu w_r))\},$$
$$\Omega_2 = \{|\Sigma_{zr}\Sigma_{rr}^{-1} D_r - D_z| \leq \frac{\lambda^n}{2\sqrt{n}} s\},$$

where

$$D_r = X_r^T(y - \mu X w)/\sqrt{n},$$
$$D_z = X_z^T(y - \mu X w)/\sqrt{n},$$

and μ is a constant defined in Section 3.1.

Here Ω_1 and Ω_2 come directly from the first order optimality conditions of $(*)$. The proof is shown in the appendix.

After combining this proposition with the proper concentration inequalities, we are able to establish the selection consistency of the Lasso. In other words, we will be able to select significant features with high probability:

Theorem 3.4. *($\mu-$sign consistency) Under the assumptions in Section 2 and the strong $\mu-$irrepresentable condition, if λ^n is chosen such that $\lambda^n \sim n^{k_3}$, with some constant k_3 satisfying $\max\{\frac{1+k_1}{2}, \frac{1+k2}{2}\} < k_3 < 1$, $0 \leq k_1 < 1$, $0 \leq k_2 \leq 1$, we have*

$$P(\mathbf{sign}(\hat{w}) = \mathbf{sign}(\mu w)) \geq 1 - O(e^{-n^{k_1}})$$
$$- O(\frac{1}{n^{2k_3 - 1 - k_2} s^2}).$$

Here k_2 depends on the choice of g. In particular, $k_2 = 0$ when g is linear.

Proof. First, according to proposition 3.3, we have

$$P(\mathbf{sign}(\hat{w}) = \mathbf{sign}(\mu w)) \geq P(\Omega_1 \cap \Omega_2).$$

Since the first q entries are non-zeros, we have

$$1 - P(\Omega_1 \cap \Omega_2) \leq P(\Omega_1^c) + P(\Omega_2^c)$$
$$\leq \sum_{i=1}^{q} P(|\alpha_i| \geq \sqrt{n}(|\mu w_r| - \frac{\lambda^n}{2n}|\Sigma_{rr}^{-1}\,\mathbf{sign}(\mu w_r)|)) +$$
$$\sum_{i=q+1}^{p} P(|\beta_i| \geq \frac{\lambda^n}{2\sqrt{n}}(1 - |\Sigma_{zr}\Sigma_{rr}^{-1}\,\mathbf{sign}(\mu w_r)|)),$$

where $\alpha = \Sigma_{rr}^{-1} D_r$ and $\beta = \Sigma_{zr}\Sigma_{rr}^{-1} D_r - D_z$.

First we analyze the distribution of α. By definition,

$$D_r = X_r^T(y - \mu X w)/\sqrt{n}, \quad D_z = X_z^T(y - \mu X w)/\sqrt{n}.$$

Furthermore, as $w = [w_r, w_z]$, $w_z = 0$, one can verify that $Xw = X_r w_r$. Therefore,

$$\alpha = \Sigma_{rr}^{-1} D_r = n(X_r^T X_r)^{-1} X_r^T(y - \mu X w)/\sqrt{n}$$
$$= \sqrt{n}(X_r^T X_r)^{-1} X_r^T(y - \mu X_r w_r)$$
$$= \sqrt{n}(X_r^{\dagger} y - \mu w_r).$$

[1] Further details can be found from https://filer.case.edu/wxg49/.

Since $X_r^\dagger y$ is the least squares estimate coming from $\min \|y - X_r w_r\|_2^2 = \|y - Xw\|_2^2$, by theorem 3.1, the least squares estimator has the following asymptotic behavior

$$\sqrt{n}(\hat{w} - \mu w) \sim \mathcal{N}(0, \sigma^2\Sigma^{-1} + \Sigma^{-1}E\{h(x)^2 xx^T\}\Sigma^{-1}),$$

where $h(x) = g(x^T w) - \mu x^T w - \gamma$, $\gamma = E\{g(x^T w) - \mu x^T w\}$, $\mu = Cov\{g(x^T w), x^T w\}/Var\{x^T w\}$.

Therefore α behaves asymptotically as:

$$\alpha \sim \mathcal{N}(0, \sigma^2\Sigma_{rr}^{-1} + \Sigma_{rr}^{-1}E\{h(x_r)^2 x_r x_r^T\}\Sigma_{rr}^{-1}).$$

By our assumption, the covariance matrix of α is well defined, thus α behaves as a Gaussian variable with mean 0 and bounded variance element-wise. The standard Gaussian tail estimation shows that if $\lambda^n/n^{(1+k_1)/2} \to \infty$ and $\lambda^n/n \to 0$, for some constant $0 \leq k_1 < 1$ then

$$\begin{aligned}\sum_{i=1}^{q} P(|\alpha_i| \geq \sqrt{n}(|\mu w_r| - \frac{\lambda^n}{2n}|\Sigma_{rr}^{-1}\,\mathbf{sign}(\mu w_r)|)) \\ = O(\exp(-n^{-k_1})).\end{aligned}$$

Now we estimate β:

$$\begin{aligned}\beta &= \Sigma_{zr}\Sigma_{rr}^{-1}D_r - D_z \\ &= (\Sigma_{zr}\Sigma_{rr}^{-1}X_r^T - X_z^T)G(X_r w_r)/\sqrt{n} \\ &\quad + (\Sigma_{zr}\Sigma_{rr}^{-1}X_r^T - X_z^T)\epsilon/\sqrt{n} + O(1),\end{aligned}$$

where mapping function $G : \mathbb{R}^n \to \mathbb{R}$ is defined as a vector version of g, that is, $G(w) := (g(w_1), g(w_z), ..., g(w_n))^T$. The $O(1)$ term comes from the empirical estimation rate of Σ. As ϵ is independent of X, the second term of β has 0 mean and bounded variance element-wise. The first term of β characterizes the 'linearity' of function $G(w)$: if $G(w)$ is linear, then the first term vanishes to $O(1)$. One should be aware that X_z is not independent of X_r and hence in general $E(X_z^T G(X_r w_r)) \neq 0$. However, we will show that the expectation of β is still under control with our assumptions.

If $\Sigma = I_{p\times p}$, then from proposition 3.2, $E(X^T z) = 0$. So we have:

$$E(X^T z) = E\begin{pmatrix} X_r^T y \\ X_z^T y \end{pmatrix} - n\mu\begin{pmatrix} \Sigma_{rr}w_r \\ \Sigma_{zr}w_r \end{pmatrix} = 0.$$

This leads to:

$$E(\beta) = n\mu(\Sigma_{zr}\Sigma_{rr}^{-1}\Sigma_{rr}w_r - \Sigma_{zr}w_r) = 0.$$

For more general Σ, we use the multivariate Stein's lemma (Stein, 1972) (Liu, 1994) [Lemma 1]:

Lemma 3.5. *Let $x = (x_1, ..., x_n)$ be multivariate normally distributed with mean vector μ and covariance matrix Σ. For any function $h(x_1, ..., x_n)$ such that $\partial h/\partial x_i$ exists almost everywhere and $E|(\frac{\partial}{\partial x_i}h(x))| < \infty$, $i = 1, 2, ...n$, the following fact holds:*

$$Cov(x_1, h(x)) = \sum_{i=1}^{n} Cov(x_1, x_i)E(\frac{\partial}{\partial x_i}h(x)).$$

Using this lemma, the expectation of β:

$$\begin{aligned}E(\beta) &= (\Sigma_{zr}\Sigma_{rr}^{-1}E(X_r^T G(X_r w_r)) - \\ &\qquad E(X_z^T G(X_r w_r)))/\sqrt{n} + O(1) \\ &= (\Sigma_{zr}w_r E(\sum_{j=1}^{n} g'(\sum_{i=1}^{q} w_i x_{ji})) - \\ &\qquad \Sigma_{zr}w_r E(\sum_{j=1}^{n} g'(\sum_{i=1}^{q} w_i x_{ji})))/\sqrt{n} + O(1) \\ &= O(1).\end{aligned}$$

Finally, in order to apply the concentration inequality, we estimate the variance of β. First notice that by the strong $\mu-$irrepresentable condition,

$$\|\Sigma_{zr}\Sigma_{rr}^{-1}X_r^T G(X_r w_r)\|_2 < 2\sqrt{p-q}\|X_r^T G(X_r w_r)\|_2.$$

Notice that $X_r w_r$ follows $\mathcal{N}(0, q^2 w_r^T \Sigma_{rr} w_r)$ distribution element-wise. Thus each element of $X_r w_r$ is bounded between $[-4q^2\Lambda_{max}, 4q^2\Lambda_{max}]$ with probability 1. Based on our assumption, $g(w)$ is differentiable a.e. and $E(|g'(w)|) < \infty$. Expanding the product of $X_r^T G(X_r w_r)$ leads to

$$Var(X_r^T G(X_r w_r)/\sqrt{n}) < c(g, \Lambda(\Sigma_{rr}), p, q)n,$$

where $c(g, \Lambda(\Sigma_{rr}), p, q)$ is a constant that depends on $g, \Lambda(\Sigma_{rr}), p$ and q. The same argument can be applied to $X_z^T G(X_r w_r)$. The second term of β can be derived from a classical result (Knight & Fu, 2000):

$$Var((\Sigma_{rr}^{-1}X_r^T - X_z^T)\epsilon/\sqrt{n}) = \Sigma_{zz} - \Sigma_{zr}\Sigma_{rr}^{-1}\Sigma_{rz}.$$

We get $Var(\beta) = O(n)$. Indeed, this is the worst case analysis. For general functions that satisfy our assumptions, $Var(\beta)$ can be smaller. More precisely, let $Var(\beta) = \Theta(n^{k_2})$, for some constant $0 \leq k_2 \leq 1$.

Finally, using Chebyshev's inequality and by choosing λ^n such that $\lambda^n \sim n^{k_3}$ for some constant k_3 such that $\max\{\frac{k_1+1}{2}, \frac{k_2+1}{2}\} < k_3 < 1$ and $0 \leq k_1 < 1$, k_2 is a constant depending on $Var(\beta)$, we get

$$\begin{aligned}\sum_{i=q+1}^{p} P(|\beta_i| \geq \frac{\lambda^n}{2\sqrt{n}}(1 - |\Sigma_{zr}^n(\Sigma_{rr}^n)^{-1}\,\mathbf{sign}(\mu w_r)|)) \\ = O(\frac{4nVar(\beta)}{(\lambda^n)^2 s^2}) = O(\frac{1}{n^{2k_3-1-k_2}s^2}).\end{aligned}$$

Combining the results of α and β completes the proof.

□

What this theorem shows is that, by maintaining regularization parameter λ^n at a certain level, the sign of the solution of model (*) will be the same as the sign of a constant times the true solution with high probability with n large enough. The constant k_2 depends on the non-linearity of link function g. The theorem matches the intuition that in order to select the features with high probability, the worse the g is (k_2 is closer to 1), the more regularization we need (a larger k_3 so that the last big-O term vanishes at a certain rate).

The practical use of this theorem is that, given enough data, by proper sampling, one can estimate the sign of μ by its definition using an empirical procedure. This can verify whether $\hat{w}$ and w have same or opposite signs.

By applying theorem 3.4, we can derive the classical convergence rate for the probability of sparse linear selection:

Corollary 3.5.1. *When $g(w)$ is linear, we have $\mu = 1$ and $k_2 = 0$. By choosing $\lambda \sim n^{k_3}$, for some $\frac{k_1+1}{2} \leq k_3 < 1$, $0 \leq k_1 < 1$, the probability of successful selection is:*

$$P(\mathbf{sign}(\hat{w}) = \mathbf{sign}(w)) \geq 1 - O(e^{-n^{k_1}}) - O(\frac{1}{n^{2k_3-1}s^2}).$$

The classical results (Zhao & Yu, 2006; Meinshausen & Bhlmann, 2006; Wainwright, 2009) show that in the linear setting, we have $P(\mathbf{sign}(\hat{w}) = \mathbf{sign}(w)) = 1 - O(e^{-n^c})$ for some constant $c > 0$. Thus our lower bound result is consistent with the classical version. By further looking into the details of the proof, one can verify that the difference comes from the different inequalities used: Gaussian tail inequality (in classical analysis) and Chebyshev's inequality (in our setting). The former achieves an exponential decay and thus the two big-Os are combined.

Extension to Group Lasso. A natural extension of the Lasso is the group Lasso, where the regularization $f(x)$ becomes $\|\cdot\|_{1,2}$. Given data $X = (X_1, X_2, ..., X_m)$, each $X_j \in \mathbb{R}^{n \times p_j}$ represents data in group j. A formal definition of the group Lasso corresponds to:

$$\min \|\sum_{j=1}^{m} X_j w_j - y\|^2 + \lambda^n \sum_{j=1}^{m} d_j \|w_j\|$$

Here $d_j > 0$ is a fixed weight for each group. In our setting, we consider $y_i = g(\sum_{j=1}^{m} X_{ij} w_j) + \epsilon_i$, $\epsilon_i \in \mathcal{N}(0, \sigma^2)$, $X_j \sim \mathcal{N}(0, \Sigma_j)$, $i = 1, ..., n$, $j = 1, ..., m$. For any minimizer of $E(Xw - y)^2$, we assume that $E((Xw - y)^2|X)$ is almost surely greater than some constant $c > 0$. We will further assume that w is normalized with Σ_j such that $\|\Sigma_j w_j\|_2 = 1$, for $j = 1, 2, ..., m$. Similar to the Lasso, we have the strong irrepresentable condition for the group Lasso (Bach, 2008):

Definition 3.2. *(Strong irrepresentable condition for the Group Lasso) The group Lasso is said to satisfy the strong irrepresentable condition if there exists $s \in (0, 1]$, such that*

$$\max_{i \in J^c} \frac{1}{d_j} \|\Sigma_{X_i X_J} \Sigma_{X_J X_J}^{-1} Diag(d_j/\|w_j\|) w_J\| \leq 1 - s.$$

Here $Diag(d_j/\|w_j\|$ denotes a block diagonal matrix with each block $d_j/\|w_j\| I_{p_j}$, J denotes the index set of the groups with all w_J non-zeros, J^c denotes its compliment. (Bach, 2008) shows that with properly chosen λ^n, one gets successful group selection with probability tending to 1 with large enough n. The following corollary continues this in general:

Corollary 3.5.2. *If the strong irrepresentability condition for the group Lasso holds, and if $\lambda^n \sim n^{k_3}$ such that $\max\{\frac{1+k_1}{2}, \frac{1+k2}{2}\} < k_3 < 1$, $0 \leq k_1 < 1$, $0 \leq k_2 \leq 1$, with k_2 depending on the choice of g, the probability of successful group selection: $P(\mathbf{sign}(\hat{w}) = \mathbf{sign}(\mu w)) \to 1$, with n large enough.*

The proof is similar to that in (Bach, 2008) in which the author does not assume any regression relation between X and y which naturally suits our setting. Combining with theorem 3.1, we get the desired result. [2]

4. Numerical Experiments

In this section, we present some numerical studies exploring the theoretical results described above. We describe experiments both to verify and illustrate our theoretical results as well as to test some of the assumptions we make.

Our setup for these studies is as follows. We generate data using several nonlinear targets and then solve a least squares problem with a *linear* hypothesis extended with a Lasso term:

$$\text{Data: } y_i = g(x_i^T w) + \epsilon_i,$$
$$\text{Model: } \hat{w} = \underset{w}{\text{argmin}} \|Xw - y\|_2^2 + \lambda\|w\|_1.$$

The data matrix X is generated in a way that each row $x_i \sim \mathcal{N}(0, \Sigma)$. Here Σ is either an identity matrix I or has power-decay entries $\Sigma_{ij} = \rho^{|i-j|}$, $0 < \rho < 1$. We denote the latter as $\Sigma(\rho)$. Both of these choices satisfy the strong irrepresentability condition. Other types of satisfying covariance matrices are shown in previous work (Zhao & Yu, 2006). The noise $\epsilon \sim \mathcal{N}(0, 0.04I)$ and $w \in \mathbb{R}^{100}$ with the first 10 entries non-zeros. The optimization problem is solved using the ADMM algorithm (Boyd, 2010).

To make the experiments comprehensive, we select a variety of functions as well as one negative example of which

[2] Further details can be found from https://filer.case.edu/wxg49/.

the link function fails to satisfy our assumptions:

$$\begin{aligned} \text{Polynomials:} \quad & g_1(u) = u^m + u^{m-1} + ... + 1 \ (m > 2), \\ \text{Sine:} \quad & g_2(u) = sin(mu) + ... + sin(u) + 1, \\ \text{Mixed:} \quad & g_3(u) = cos(u) + \frac{1}{1+e^{-u}} - u^3 - 2, \\ \text{False:} \quad & g_4(u) = 1/u, \\ \text{Test:} \quad & g_5(u) = (u+1)^2 - 2u + \exp(u) + 1. \end{aligned}$$

Figure 1 shows two examples illustrating the 'scaling' phenomenon: the solution of the selection model $\hat{w}$ is close to the underlying truth w up to a scaling factor μ. In this figure, we plot the coefficients for $g_1(u)$ with $m = 3$ (left) and $g_3(u)$ (right). The training sample has 2000 examples generated with $\Sigma = \Sigma(0.8)$. The regularization parameter λ^n is set to 0.8 during optimization. As the figure shows, the found solution $\hat{w}$ is linearly related to w, though the sign may be different. The left panel and the right panel corresponds to positive and negative μ respectively ($\mu_1 \approx 3.73$ (left) , $\mu_2 \approx -2.53$ (right)).

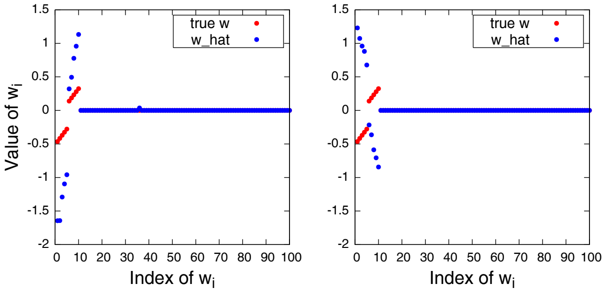

Figure 1. Illustration of scaling in the recovered model. Left: $g_1(u)$ with $m = 3$, Right: $g_3(u)$.

We next verify our main result. In Figure 2 we plot the probability of a relevant feature being successfully selected versus the sample size n for different link functions g and covariance matrices Σ. Each mark on the curves corresponds to an average over 200 trials using different random training samples.

To overcome the difficulty of not knowing the constants in theorem 3.4, we uniformly sample λ from $n^{0.6}$ to $n^{1.2}$. The reason this upper-bound exceeds n here is for experimental selection. As we only know $\lambda = \Theta(n^k)$, $k \in (\frac{1}{2}, 1)$, we use this setting to search for the unknown constant factor. The numerical results in figure 2 are consistent with our theoretical results. The Lasso model is able to select the right features with a large enough sample size even for highly non-linear functions such as polynomials and sine functions. Even though $w \in \mathbb{R}^{100}$, figure 2 also indicates that 2400 to 4000 of samples are sufficient to achieve successful selection almost surely while in the linear case this reduces to only 200 to 500. Thus under the right conditions we may expect the Lasso to be very successful at selecting relevant features even with modest sample sizes.

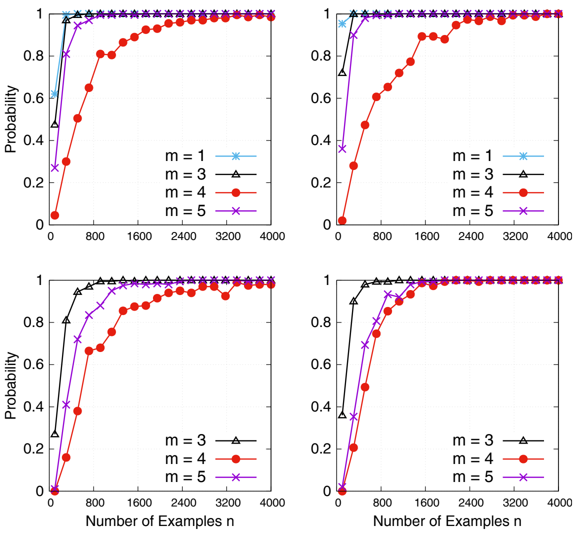

Figure 2. Numerical simulations on the probability of successful selection with different sample size n and Σ. The first row are simulations on $g_1(w)$, $m = 1$ corresponds to the linear mapping function, second row corresponds to $g_2(w)$. The first column has $\Sigma = I$, second column has $\Sigma = \Sigma(0.2)$.

Our theorem indicates that a consistent successful selection probability can be achieved by setting λ in the order of n^k, where k is a constant depending on the noise. Figure 3 shows the numerical demonstration of this statement. The weight vector $w \in R^{100}$ has been normalized as $\|w\|_2 = 1$. The y-axis is obtained from $\log(\lambda)/\log(n)$ which is equivalent to k with a shifting constant. The black pixels indicates successful selections with probability over 95%. This figure indicates that we can consistently choose λ as n^k with k as a constant to achieve satisfying selection performance.

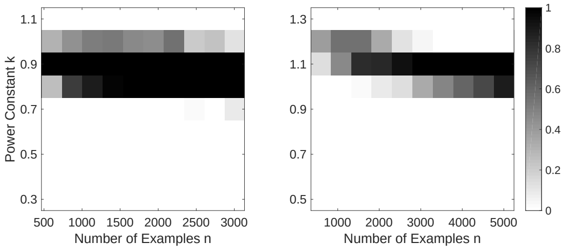

Figure 3. Numerical tests on the relation between power constant k and the sample size n with fixed amount of noise ϵ. Left: $g_1(u)$ with $m = 3$, $\epsilon \sim \mathcal{N}(0, 2I)$. Right: $g_1(u)$ with $m = 5$, $\epsilon \sim \mathcal{N}(0, I)$.

Figure 4 shows the results of consistency tests for the group Lasso. For simplicity, we only show this for the polynomial families. In this test, $w \in \mathbb{R}^{80}$ has eight groups of ten. The last four groups are set to be zeros. The w in the first four groups are randomly chosen to be ± 1. The results are consistent with our theory. Note that a relative large sample size may be required for successful group selection.

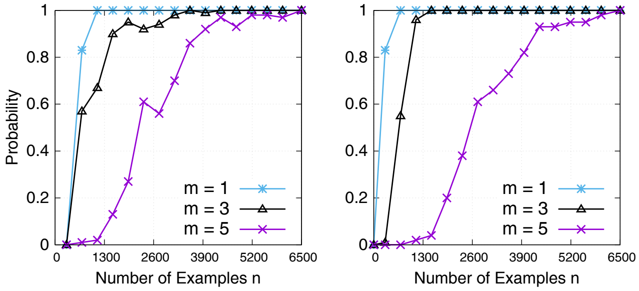

Figure 4. Numerical simulations on the probability of successful group selection with different sample size n and Σ for polynomial families. The first row are simulations on $g_1(w)$, $m = 1$ corresponds to the linear mapping function, second row corresponds to $g_2(w)$. The first column has $\Sigma = I$, second column has $\Sigma = \Sigma(0.2)$.

Next, we study the necessity of some of our assumptions. Throughout the paper we have used the assumption that the data follows a Gaussian distribution, following related work. To get more insight into the Lasso as well as the necessity of this assumption, we run numerical tests on data X generated from a variety of distributions. We use a test function g_5, a randomly chosen function that does not have any special patterns. Figure 5 shows the test results. We also test our assumptions on the link function from Section 2 through the target g_4, which fails to satisfy our expectation assumptions on the link function. In particular, $E(g_4'(t))$ does not exist when $t \sim \mathcal{N}(0, 1)$.

From Figure 5 we observe the following. First, feature selection from g_4, which violates our conditions, fails completely. Second, for the given function g_5, along with the Gaussian, the uniform distribution as well as Weibull distribution with parameter 1 and 1.5 also lead to successful feature selection. The former maintains thinner tails than the Gaussian while the latter has heavier tails. However, the convergence rate of the probability for Gaussian distribution outperforms the other two. Interestingly, two very similar distributions show different behavior. The $Beta(2, 2)$ distribution also has finite support as the uniform and $t(\nu = 1)$ maintains heavier tails than the Gaussian, but both of them fail to achieve selection consistency. These results indicate that there may be some room to relax the requirements we assume, but not much. Understanding the gap between the necessary and sufficient conditions is a direction for future work.

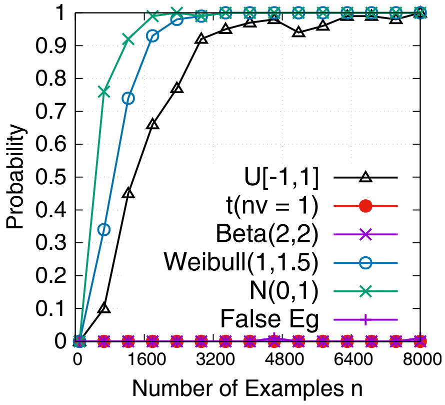

Figure 5. Numerical tests on the necessity of assumptions. $U[-1, 1]$ is uniform distribution on $[-1, 1]$, $t(nv = 1)$ is t-distribution with freedom $\nu = 1$, $Beta(2, 2)$ corresponds to Beta distribution with $\alpha = \beta = 2$, $Weibull(1, 1.5)$ is Weibull distribution with scale 1, shape 1.5, $N(0, 1)$ is standard normal for comparison. The "False Eg" line corresponds to tests on g_4 with data from Gaussian distribution $\Sigma = I$.

5. Conclusion

In this paper, we have studied the selection consistency of the Lasso when the observations are generated from some unknown link function that might be nonlinear while the learning happens with a linear hypothesis class. We prove that under suitable assumptions, the Lasso model is still able to select the right features, though the recovered coefficients may either be dampened or amplified by an unknown constant. We have described the asymptotic probability behavior of the selection consistency of the Lasso solution and derived the classical consistency results as a special case. These results extend to the group Lasso as well. Our numerical studies verify the predicted behavior and also indicate the necessity of our assumptions on the link function, though there may be room to relax some other assumptions on the data distribution. In future work, we plan to investigate this and study the necessary conditions that lead to consistent selection, the Lasso with non-Gaussian noise, as well as the case when the number of features are larger than the number of samples.

Acknowledgements

We thank the reviewers for the helpful comments. The work of Y. Zhang and W. Guo are partially supported by NSF grant DMS-1521582.

References

Bach, Francis R. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

Boyd, Stephen. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000016.

Brillinger, David R. A generalized linear model with gaussian regressor variables. A Festschrift for Erich L. Lehmann, 1982.

Bunea, Florentina. Honest variable selection in linear and logistic regression models via L1 and L1+ L2 penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

Cands, Emmanuel and Recht, Benjamin. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589, 2013.

Cands, Emmanuel J., Romberg, Justin, and Tao, Terence. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

Chen, Scott Shaobing, Donoho, David L., and Saunders, Michael A. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20:33–61, 1998.

Donoho, David L. and Tanner, Jared. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.

Donoho, D.L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.871582.

Feuer, Arie and Nemirovski, Arkadi. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579–1581, 2003.

Knight, Keith and Fu, Wenjiang. Asymptotics for lasso-type estimators. *Annals of statistics*, pp. 1356–1378, 2000.

Lee, Jason D., Sun, Yuekai, and Taylor, Jonathan E. On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics*, 9(1):608–642, 2015. ISSN 1935-7524. doi: 10.1214/15-EJS1013.

Liu, Jun S. Siegel's formula via Stein's identities. *Statistics & Probability Letters*, 21(3):247–251, October 1994. ISSN 0167-7152. doi: 10.1016/0167-7152(94)90121-X.

Meinshausen, Nicolai and Bhlmann, Peter. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, June 2006. ISSN 0090-5364. doi: 10.1214/009053606000000281.

Negahban, Sahand N., Ravikumar, Pradeep, Wainwright, Martin J., and Yu, Bin. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, November 2012. ISSN 0883-4237. doi: 10.1214/12-STS400.

Plan, Yaniv and Vershynin, Roman. The generalized Lasso with non-linear observations. *arXiv preprint arXiv:1502.04071*, 2015.

Ravikumar, Pradeep, Wainwright, Martin J., Lafferty, John D., and others. High-dimensional Ising model selection using L1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Stein, Charles. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. The Regents of the University of California, 1972.

Tateishi, Shohei, Matsui, Hidetoshi, and Konishi, Sadanori. Nonlinear regression modeling via the lasso-type regularization. *Journal of Statistical Planning and Inference*, 140(5):1125 – 1134, 2010. ISSN 0378-3758.

Thrampoulidis, Christos, Abbasi, Ehsan, and Hassibi, Babak. LASSO with Non-linear Measurements is Equivalent to One With Linear Measurements. In *Advances in Neural Information Processing Systems*, pp. 3402–3410, 2015.

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

Wainwright, M.J. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using -Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009. ISSN 0018-9448. doi: 10.1109/TIT.2009.2016018.

Yuan, Ming and Lin, Yi. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007.

Zhao, Peng and Yu, Bin. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7: 2541–2563, 2006.

Zou, Hui. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006.