The Impact of Regularization on High-dimensional Logistic Regression

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi*
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA, USA.

Abstract

Logistic regression is commonly used for modeling dichotomous outcomes. In the classical setting, where the number of observations is much larger than the number of parameters, properties of the maximum likelihood estimator in logistic regression are well understood. Recently, Sur and Candes [26] have studied logistic regression in the high-dimensional regime, where the number of observations and parameters are comparable, and show, among other things, that the maximum likelihood estimator is biased. In the high-dimensional regime the underlying parameter vector is often structured (sparse, block-sparse, finite-alphabet, etc.) and so in this paper we study regularized logistic regression (RLR), where a convex regularizer that encourages the desired structure is added to the negative of the log-likelihood function. An advantage of RLR is that it allows parameter recovery even for instances where the (unconstrained) maximum likelihood estimate does not exist. We provide a precise analysis of the performance of RLR via the solution of a system of six nonlinear equations, through which any performance metric of interest (mean, mean-squared error, probability of support recovery, etc.) can be explicitly computed. Our results generalize those of Sur and Candes and we provide a detailed study for the cases of ℓ_2^2-RLR and sparse (ℓ_1-regularized) logistic regression. In both cases, we obtain explicit expressions for various performance metrics and can find the values of the regularizer parameter that optimizes the desired performance. The theory is validated by extensive numerical simulations across a range of parameter values and problem instances.

1 Introduction

Logistic regression is the most commonly used statistical model for predicting dichotomous outcomes [11]. It has been extensively employed in many areas of engineering and applied sciences, such as in the medical [3, 32] and social sciences [14]. As an example, in medical studies logistic regression can be used to predict the risk of developing a certain disease (e.g. diabetes) based on a set of observed characteristics from the patient (age, gender, weight, etc.)

Linear regression is a very useful tool for predicting a quantitive response. However, in many situations the response variable is qualitative (or categorical) and linear regression is no longer appropriate [12]. This is mainly due to the fact that least-squares often succeeds under the assumption that the error components are independent with normal distribution. In categorical predictions, however, the error components are neither inependent nor normally distributed [19].

In logistic regression we model the probability that the label, Y, belongs to a certain category. When no prior knowledge is available regarding the structure of the parameters, maximum likelihood is often used for fitting the model. Maximum likelihood estimation (MLE) is a special case of maximum

*This work was supported in part by the National Science Foundation under grants CNS-0932428, CCF-1018927, CCF-1423663 and CCF-1409204, by a grant from Qualcomm Inc., by a grant from Futurewei Inc., by NASA's Jet Propulsion Laboratory through the President and Director's Fund, and by King Abdullah University of Science and Technology.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

a posteriori estimation (MAP) that assumes a uniform prior distribution on the parameters.
In many applications in statistics, machine learning, signal processing, etc., the underlying parameter obeys some sort of *structure* (sparse, group-sparse, low-rank, finite-alphabet, etc.). For instance, in modern applications where the number of features far exceeds the number of observations, one typically enforces the solution to contain only a few non-zero entries. To exploit such structural information, inspired by the Lasso [31] algorithm for linear models, researchers have studied regularization methods for generalized linear models [24, 9]. From a statistical viewpoint, adding a regularization term provides a MAP estimate with a non-uniform prior distribution that has higher densities in the set of structured solutions.

1.1 Prior work

Classical results in logistic regression mainly concern the regime where the sample size, n, is overwhelmingly larger than the feature dimension p. It can be shown that in the limit of large samples when p is fixed and $n \to \infty$, the maximum likelihood estimator provides an efficient estimate of the underlying parameter, i.e., an unbiased estimate with covariance matrix approaching the inverse of the the Fisher information [34, 17]. However, in most modern applications in data science, the datasets often have a huge number of features, and therefore, the assumption $\frac{n}{p} \gg 1$ is not valid. Sur and Candes [5, 26, 27] have recently studied the performance of the maximum likeliood estimator for logistic regression in the regime where n is proportional to p. Their findings challenge the conventional wisdom, as they have shown that in the linear asymptotic regime the maximum likelikehood estimate is not even unbiased. Their analysis provides the precise performance of the maximum likelihood estimator.
There have been many studies in the literature on the performance of regularized (penalized) logistic regression, where a regularizer is added to the negative log-likelihood function (a partial list includes [4, 13, 33]). These studies often require the underlying parameter to be heavily structured. For example, if the parameters are sparse the sparsity is taken to be $o(p)$. Furthermore, they provide orderwise bounds on the performance but do not give a precise characterization of the quality of the resulting estimate. A major advantage of adding a regularization term is that it allows for recovery of the parameter vector even in regimes where the maximum likelihood estimate does not exist (due to an insufficient number of observations.)

1.2 Summary of contributions

In this paper, we study regularized logistic regression (RLR) for parameter estimation in high-dimensional logistic models. Inspired by recent advances in the performance analysis of M-estimators for linear models [7, 8, 28], we precisely characterize the assymptotic performance of the RLR estimate. Our characterization is through a system of six nonlinear equations in six unknowns, through whose solution all locally-Lipschitz performance measures such as the mean, mean-squared error, probability of support recovery, etc., can be determined. In the special case when the regularization term is absent, our 6 nonlinear equations reduce to the 3 nonlinear equations reported in [26]. When the regularizer is quadratic in parameters, the 6 equations also simplifies to 3. When the regularizer is the ℓ_1 norm, which corresponds to the popular sparse logistic regression [15, 16], our equations can be expressed in terms of q-functions, and quantities such as the probability of correct support recovery can be explicitly computed. Numerous numerical simulations validate the theoretical findings across a range of problem settings. To the extent of our knowledge, this is the first work that precisely characterizes the performance of the regularized logistic regression in high dimensions.
For our analysis, we utilize the recently developed **C**onvex **G**aussian **M**in-max **T**heorem (**CGMT**) [29] which is a strengthened version of a classical Gaussian comparison inequality due to Gordon [10], and whose origins are in [25]. Previously, the CGMT has been successfully applied to derive the precise performance in a number of applications such as regularized M-estimators [28], analysis of the generalized lasso [18, 29], data detection in massive MIMO [1, 2, 30], and PhaseMax in phase retrieval [6, 23, 22].

2 Preliminaries

2.1 Notations

We gather here the basic notations that are used throughout this paper. $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. $X \sim p_X$ implies that the random variable X has a density p_X. $\xrightarrow{\text{P}}$ and $\xrightarrow{\text{d}}$ represent convergence in probability and in distribution, respectively. Lower letters are reserved for vectors and upper letters are for matrices. $\mathbf{1}_d$, and $\mathbf{I}_d$ respectively denote the all-one

vector and the identity matrix in dimension d. For a vector $\mathbf{v}$, v_i denotes its i^{th} entry, and $||\mathbf{v}||_p$ (for $p \geq 1$), is its ℓ_p norm, where we remove the subscript when $p = 2$. A function $f : \mathbb{R}^p \to \mathbb{R}$ is called *(invariantly) separable* if $f(\mathbf{w}) = \sum_{i=1}^{p} \tilde{f}(w_i)$ for all $\mathbf{w} \in R^p$, where $\tilde{f}(\cdot)$ is a real-valued function. For a function $\Phi : \mathbb{R}^d \to \mathbb{R}$, the Moreau envelope associated with $\Phi(\cdot)$ is defined as,

$$M_{\Phi}(\mathbf{v}, t) = \min_{\mathbf{x} \in \mathbb{R}^d} \ \frac{1}{2t}||\mathbf{v} - \mathbf{x}||^2 + \Phi(\mathbf{x}) \ , \tag{1}$$

and the proximal operator is the solution to this optimization, i.e.,

$$\text{Prox}_{t\Phi(\cdot)}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathbb{R}^d} \ \frac{1}{2t}||\mathbf{v} - \mathbf{x}||^2 + \Phi(\mathbf{x}) \ . \tag{2}$$

2.2 Mathematical Setup

Assume we have n samples from a logistic model with parameter $\boldsymbol{\beta}^* \in \mathbb{R}^p$. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the set of samples (a.k.a. the training data), where for $i = 1, 2, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector and the label $y_i \in \{0, 1\}$ is a Bernouli random variable with,

$$\mathbb{P}[y_i = 1 | \mathbf{x}_i] = \rho'(\mathbf{x}_i^T \boldsymbol{\beta}^*) \ , \ \text{ for } \ i = 1, 2, \ldots, n \ , \tag{3}$$

where $\rho'(t) := \frac{e^t}{1+e^t}$ is the standard logistic function. The goal is to compute an estimate for $\boldsymbol{\beta}^*$ from the training data $\mathcal{D}$. The maximum likelihood estimator, $\hat{\boldsymbol{\beta}}_{ML}$, is defined as,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ML} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \prod_{i=1}^{n} \mathbb{P}_{\boldsymbol{\beta}}(y_i | \mathbf{x}_i) &= \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \prod_{i=1}^{n} \frac{e^{y_i(\mathbf{x}_i^T \boldsymbol{\beta})}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \\ &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(\mathbf{x}_i^T \boldsymbol{\beta}) - y_i(\mathbf{x}_i^T \boldsymbol{\beta}) \ . \end{aligned} \tag{4}$$

Where $\rho(t) := \log(1 + e^t)$ is the *link function* which has the standard logistic function as its derivative. The last optimization is simply minimization over the negative log-likelihood. This is a convex optimization program as the log-likelihood is concave with respect to $\boldsymbol{\beta}$.
As explained earlier in Section 1, in many interesting settings the underlying parameter possesses cerain structure(s) (sparse, low-rank, finite-alphabet, etc.). In order to exploit this structure we assume $f : \mathbb{R}^p \to \mathbb{R}$ is a *convex* function that measures the (so-called) "complexity" of the structured solution. We fit this model by the regularized maximum (binomial) likelihood defined as follows,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \ \frac{1}{n} \cdot \big[\sum_{i=1}^{n} \rho(\mathbf{x}_i^T \boldsymbol{\beta}) - y_i(\mathbf{x}_i^T \boldsymbol{\beta})\big] + \frac{\lambda}{p} \, f(\boldsymbol{\beta}) \ . \tag{5}$$

Here, $\lambda \in \mathbb{R}_+$ is the regularization parameter that must be tuned properly. In this paper, we study the linear asymptotic regime in which the problem dimensions p, n grow to infinity at a proportional rate, $\delta := \frac{n}{p} > 0$. Our main result characterizes the performance of $\hat{\boldsymbol{\beta}}$ in terms of the ratio, δ, and the signal strength, $\kappa = \frac{||\boldsymbol{\beta}^*||}{\sqrt{p}}$. For our analysis we assume that the regularizer $f(\cdot)$ is separable, $f(\mathbf{w}) = \sum_i \tilde{f}(w_i)$, and the data points are drawn independently from the Gaussian distribution, $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$. Note that the assumptions considered in the analysis of the We further assume that the entries of $\boldsymbol{\beta}^*$ are drawn from a distribution Π. Our main result characterizes the performance of the resulting estimator through the solution of a system of six nonlinear equations with six unknowns. In particular, we use the solution to compute some common descriptive statistics of the estimate, such as the mean and the variance.

3 Main Results

In this section, we present the main result of the paper, that is the characterization of the asymptotic performance of regularized logistic regression (RLR). When the estimation performance is measured via a locally- Lipschitz function (e.g. mean-squared error), Theorem 1 precisely predicts the asymptotic behavior of the error. The derived expression captures the role of the regularizer, $f(\cdot)$, and the particular distribution of $\boldsymbol{\beta}^*$, through a set of scalars derived by solving a system of nonlinear equations. In Section 3.1 we present this system of nonlinear equations along with some insights on how to numerically compute its solution. After formally stating our result in Section 3.2, we use that to predict the general behavior of $\hat{\boldsymbol{\beta}}$. In particular, in Section 3.3 we compute its correlation with the true signal as well as its mean-squared error.

3.1 A nonlinear system of equations

As we will see in Theorem 1, given the signal strength κ, and the ratio δ, the asymptotic performance of RLR is characterized by the solution to the following system of nonlinear equations with six unknowns $(\alpha, \sigma, \gamma, \theta, \tau, r)$.

$$\begin{cases} \kappa^2\alpha = \mathbb{E}\big[\beta \operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta\beta + \frac{r}{\sqrt{\delta}}Z)\big)\big] , \\ \gamma = \frac{1}{r\sqrt{\delta}}\,\mathbb{E}\big[Z \operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta\beta + \frac{r}{\sqrt{\delta}}Z)\big)\big] , \\ \kappa^2\alpha^2 + \sigma^2 = \mathbb{E}\left[\operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta\beta + \frac{r}{\sqrt{\delta}}Z)\big)^2\right] , \\ \gamma^2 = \frac{2}{r^2}\,\mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - \operatorname{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big] , \\ \theta\gamma = -2\,\mathbb{E}\big[\rho''(-\kappa Z_1)\operatorname{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big] , \\ 1 - \frac{\gamma}{\sigma\tau} = \mathbb{E}\big[\frac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''\big(\operatorname{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\big] . \end{cases} \tag{6}$$

Here Z, Z_1, Z_2 are standard normal variables, and $\beta \sim \Pi$, where Π denotes the distribution on the entries of $\boldsymbol{\beta}^*$. The following remarks provide some insights on solving the nonlinear system.

Remark 1 (Proximal Operators)**.** *It is worth noting that the equations in* (6) *include the expectation of functionals of two proximal operators. The first three equations are in terms of Prox$_{\tilde{f}(\cdot)}$, which can be computed explicitly for most widely used regularizers. For instance, in ℓ_1-regularization, the proximal operator is the well-known shrinkage function defined as $\eta(x,t) := \frac{x}{|x|}(|x| - t)_+$. The remaining equations depend on computing the proximal operator of the link function $\rho(\cdot)$. For $x \in \mathbb{R}$, Prox$_{t\rho(\cdot)}(x)$ is the unique solution of $z + t\rho'(z) = x$.*

Remark 2 (Numerical Evaluation)**.** *Define $\mathbf{v} := [\alpha, \sigma, \gamma, \theta, \tau, r]^T$ as the vector of unknonws. The nonlinear system* (6) *can be reformulated as $\mathbf{v} = S(\mathbf{v})$ for a properly defined $S : \mathbb{R}^6 \to \mathbb{R}^6$. We have empirically observed in our numerical simulations that a fixed-point iterative method, $\mathbf{v}_{t+1} = S(\mathbf{v}_t)$, converges to $\mathbf{v}^*$, such that $\mathbf{v}^* = S(\mathbf{v}^*)$.*

3.2 Asymptotic performance of regularized logistic regression

We are now able to present our main result. Theorem 1 below describes the average behavior of the entries of $\hat{\boldsymbol{\beta}}$, the solution of the RLR. The derived expression is in terms of the solution of the nonlinear system (6), denoted by $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$. An informal statement of our result is that as $n \to \infty$, the entries of $\hat{\boldsymbol{\beta}}$ converge as follows,

$$\hat{\boldsymbol{\beta}}_j \xrightarrow{d} \Gamma(\boldsymbol{\beta}^*_j, Z) , \quad \text{for } j = 1, 2, \ldots, p , \tag{7}$$

where Z is a standard normal random variable, and $\Gamma : \mathbb{R}^2 \to \mathbb{R}$ is defined as,

$$\Gamma(c, d) := \operatorname{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}c + \frac{\bar{r}}{\sqrt{\delta}}d)\big) . \tag{8}$$

In other words, the RLR solution has the same behavior as applying the proximal operator on the "perturbed signal", i.e., the true signal added with a Gaussian noise.

Theorem 1. *Consider the optimization program* (5)*, where for $i = 1, 2, \ldots, n$, $\mathbf{x}_i$ has the multivariate Gaussian distribution $\mathcal{N}(0, \frac{1}{p}\mathbf{I}_p)$, and $y_i = Ber(\mathbf{x}_i^T\boldsymbol{\beta}^*)$, and the entries of $\boldsymbol{\beta}^*$ are drawn independently from a distribution Π. Assume the parameters δ, κ, and λ are such that the nonlinear system* (6) *has a unique solution $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$. Then, as $p \to \infty$, for any locally-Lipschitz[2] function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we have,*

$$\frac{1}{p}\sum_{j=1}^{p} \Psi(\hat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}^*_j) \xrightarrow{P} \mathbb{E}\big[\Psi\big(\Gamma(\beta, Z), \beta\big)\big] , \tag{9}$$

where $Z \sim \mathcal{N}(0,1)$, $\beta \sim \Pi$ is independent of Z, and the function $\Gamma(\cdot,\cdot)$ is defined in (8)*.*

[2] A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is said to be *locally-Lipschitz* if,

$$\forall M > 0,\ \exists L_M \geq 0,\ \text{such that } \forall \mathbf{x}, \mathbf{y} \in \big[-M, +M\big]^d \ : \quad |\Phi(\mathbf{x}) - \Phi(\mathbf{y})| \leq L_M||\mathbf{x} - \mathbf{y}|| .$$

We defer the detailed proof to the Appendix. In short, to show this result we first represnt the optimization as a bilinear form $\mathbf{u}^T\mathbf{X}\mathbf{v}$, where $\mathbf{X}$ is the measurement matrix. Applying the CGMT to derive an equivalent optimization, we then simplify this optimization to obtain an unconstrained optimization with six scalar variables. The nonlinear system (6) represents the first-order optimality condition of the resulting scalar optimization.
Before stating the consequences of this result, a few remarks are in order.

Remark 3 (Assumptions)**.** *The assumptions in Theorem 1 are chosen in a conservative manner. In particular, we could relax the separability condition on $f(\cdot)$, to some milder condition in terms of asymptotic convergence of its proximal operator. Furthermore, one can relax the assumption on the entries of $\boldsymbol{\beta}^*$ being i.i.d. to a weaker assumption on the empirical distribution of its entries. However, for the applications of this paper, the theorem in its current form is adequate.*

Remark 4 (Choosing Ψ)**.** *The performance measure in Theorem 1 is computed in terms of evaluation of a locally-Lipschitz function, $\Psi(\cdot,\cdot)$. As an example, $\Psi(u,v)=(u-v)^2$ can be used to compute the mean-squared error. Later on, we will appeal to this theorem with various choices of Ψ to evaluate different performance measures on $\hat{\boldsymbol{\beta}}$.*

3.3 Correlation and variance of the RLR estimate

As the first application of Theorem 1 we compute common descriptive statistics of the estimate $\hat{\boldsymbol{\beta}}$. In the following corollaries, we establish that the parametrs $\bar{\alpha}$, and $\bar{\sigma}$ in (6) correspond to the correlation and the mean-squared error of the resulting estimate.

Corollary 1. *As $p\to\infty$, $\frac{1}{||\boldsymbol{\beta}^*||^2}\,\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^* \xrightarrow{P} \bar{\alpha}$.*

Proof. Recall that $||\boldsymbol{\beta}^*||^2 = p\kappa^2$. Applying Theorem 1 with $\Psi(u,v)=uv$ gives,

$$\frac{1}{||\boldsymbol{\beta}^*||^2}\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^* = \frac{1}{\kappa^2 p}\sum_{j=1}^{p}\hat{\boldsymbol{\beta}}_j\boldsymbol{\beta}_j^* \xrightarrow{P} \frac{1}{\kappa^2}\mathbb{E}\big[\beta\,\mathrm{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}\beta+\frac{\bar{r}}{\sqrt{\delta}}Z)\big)\big] = \bar{\alpha}\ , \tag{10}$$

where the last equality is derived from the first equation in the nonlinear system (6), along with the fact that $(\bar{\alpha},\bar{\sigma},\bar{\gamma},\bar{\theta},\bar{\tau},\bar{r})$ is a solution to this system. □

Corollary 1 states that upon centering $\hat{\boldsymbol{\beta}}$ around $\bar{\alpha}\boldsymbol{\beta}^*$, it becomes decorrelated from $\boldsymbol{\beta}^*$. Therefore, we define a new estimate $\tilde{\boldsymbol{\beta}} := \frac{\hat{\boldsymbol{\beta}}}{\bar{\alpha}}$ and compute its mean-squared error in the following corollary.

Corollary 2. *As $p\to\infty$, $\frac{1}{p}||\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*||^2 \xrightarrow{P} \frac{\bar{\sigma}^2}{\bar{\alpha}^2}$.*

Proof. We appeal to Theorem 1 with $\Psi(u,v)=(u-\bar{\alpha}v)^2$,

$$\frac{1}{p}||\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*||^2 = \frac{1}{\bar{\alpha}^2}\big(\frac{1}{p}||\hat{\boldsymbol{\beta}}-\bar{\alpha}\boldsymbol{\beta}^*||^2\big) \xrightarrow{P} \frac{1}{\bar{\alpha}^2}\mathbb{E}\big[\big(\mathrm{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}\beta+\frac{\bar{r}}{\sqrt{\delta}}Z)\big)-\bar{\alpha}\beta\big)^2\big] = \frac{\bar{\sigma}^2}{\bar{\alpha}^2}\ , \tag{11}$$

where the last equality is derived from the third equation in the nonlinear system (6) together with the result of Corollary 1. □

In the next two sections, we investigate other properties of the estimate $\hat{\boldsymbol{\beta}}$ under ℓ_1 and ℓ_2 regularization.

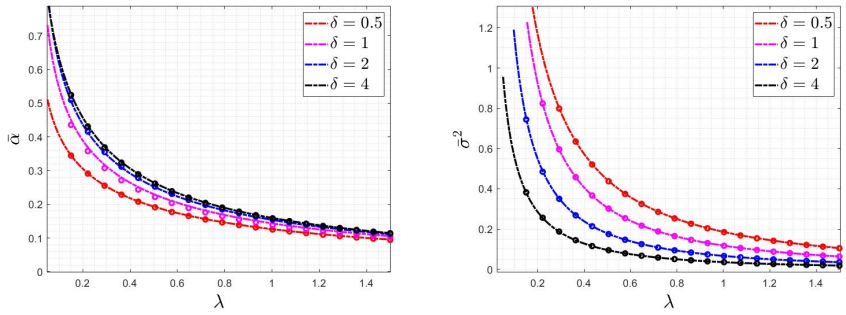
4 RLR with ℓ_2^2-regularization

The ℓ_2 norm regularization is commonly used in machine learning applications to stabilize the model. Adding this regularization would simply shrink all the parameters toward the origin and hence decrease the variance of the resulting model. Here, we provide a precise performance analysis of the RLR with ℓ_2^2-regularization, i.e.,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\ \frac{1}{n}\cdot\big[\sum_{i=1}^{n}\rho(\mathbf{x}_i^T\boldsymbol{\beta})-y_i(\mathbf{x}_i^T\boldsymbol{\beta})\big]+\frac{\lambda}{2p}\sum_{i=1}^{p}\boldsymbol{\beta}_i^2\ . \tag{12}$$

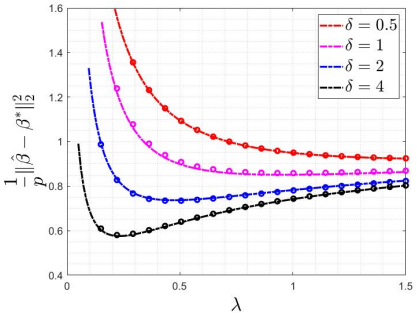
To analyze (12), we use the result of Theorem 1. It can be shown that in the nonlinear system (6), $\bar{\theta}$, $\bar{\tau}$, $\bar{r}$ can be derived explicitely from solving the first three equations. This is due to the fact that the proximal operator of $\tilde{f}(\cdot)=\frac{1}{2}(\cdot)^2$ can be expressed in the following closed-form,

$$\mathrm{Prox}_{t\tilde{f}(\cdot)}(x) = \arg\min_{y\in\mathbb{R}}\ \frac{1}{2t}(y-x)^2+\frac{1}{2}y^2 = \frac{x}{1+t}\ . \tag{13}$$



(a)

(b)



(c)

Figure 1: The performance of the regularized logistic regression under ℓ_2^2 penalty (a) the correlation factor $\bar{\alpha}$ (b) the variance $\bar{\sigma}^2$, and (c) the mean-squared error $\frac{1}{p}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||^2$. The dashed lines depict the theoretical result derived from Theorem 2, and the dots are the result of empirical simulations. The empirical results is the average over 100 independent trials with $p = 250$ and $\kappa = 1$.

This indicates that the proximal operator in this case is just a simple rescaling. Substituting (13) in the nonlinear system (6), we can rewrite the first three equations as follows,

$$\begin{cases} \theta = \dfrac{\alpha}{\gamma\delta}\,, \\ \tau = \dfrac{\delta\gamma}{\sigma\big(1-\lambda\delta\gamma\big)}\,, \\ r = \dfrac{\sigma}{\gamma\sqrt{\delta}}\,. \end{cases} \tag{14}$$

Therefore we can state the following Theorem for ℓ_2^2-regularization:

Theorem 2. *Consider the optimization* (12) *with parameters κ, δ, and γ, and the same assumptions as in Theorem 1. As $p \to \infty$, for any locally-Lipschitz function $\Psi(\cdot,\cdot)$, the following convergence holds,*

$$\frac{1}{p}\sum_{j=1}^{p} \Psi(\hat{\boldsymbol{\beta}}_j - \bar{\alpha}\boldsymbol{\beta}_j^*, \boldsymbol{\beta}_j^*) \xrightarrow{P} \mathbb{E}\big[\Psi\big(\bar{\sigma}Z, \beta\big)\big]\,, \tag{15}$$

where Z is standard norma, $\beta \sim \Pi$, and $\bar{\alpha}$,$\bar{\sigma}$ are the unique solution to the following nonlinear system of equations,

$$\begin{cases} \dfrac{\sigma^2}{2\delta} = \mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - \mathit{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big]\,, \\ -\dfrac{\alpha}{2\delta} = \mathbb{E}\big[\rho''(-\kappa Z_1)\mathit{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big]\,, \\ 1 - \dfrac{1}{\delta} + \lambda\gamma = \mathbb{E}\big[\dfrac{2\rho'(-\kappa Z_1)}{1+\gamma\rho''\big(\mathit{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\big]\,. \end{cases} \tag{16}$$

The proof is deferred to the Appendix. Theorem 2 states that upon centering the estimate $\hat{\boldsymbol{\beta}}$, it becomes decorrelated from $\boldsymbol{\beta}^*$ and the distribution of the entries approach a zero-mean Gaussian distribution with variance $\bar{\sigma}^2$.
Figure 1 depicts the performance of the regularized estimate for different values of λ. As observed in the figure, increasing the value of λ reduces the correlation factor $\bar{\alpha}$ (Figure 1a) and the variance $\bar{\sigma}^2$ (Figure 1b). Figure 1c shows the mean-squared-error of the estimate as a function of λ . It indicates that for different values of δ there exist an optimal value λ_{opt} that achieves the minimum mean-squared error.

4.1 Unstructured case

When $\lambda = 0$ in (12), we obtain the optimization with no regularization, i.e., the maximum likelihood estimate. When we set λ to zero in (16), Theorem 2 gives the same result as Sur and Candes reported in [26]. In their analysis, they have also provided an interesting interpretation of $\bar{\gamma}$ in terms of the likelihood ratio statistics. Studying the likelihood ratio test is beyond the scope of this paper.

5 Sparse Logistic Regression

In this section we study the performance of our estimate when the regularizer is the ℓ_1 norm. In modern machine learning applications the number of features, p, is often overwhelmingly large. Therefore, to avoid overfitting one typically needs to perform feature selection, that is, to exclude irrelevent variables from the regression model [12]. Adding an ℓ_1 penalty to the loss function is the most popular approach for feature selection.
As a natural consequence of the result of Theorem 1, we study the performance of RLR with ℓ_1 regularizer (referred to as "sparse LR") and evaluate its success in recovery of the sparse signals. In Section 5.1, we extend our general analysis to the case of sparse LR. In other words, we will precisely analyze the performance of the solution of the following optimization,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \ \frac{1}{n}\cdot\big[\sum_{i=1}^{n} \rho(\mathbf{x}_i^T\boldsymbol{\beta}) - y_i(\mathbf{x}_i^T\boldsymbol{\beta})\big] + \frac{\lambda}{p}||\boldsymbol{\beta}||_1 \ . \tag{17}$$

In Section 5.1, we explicitly describe the expectations in the nonlinear system (6) using two q-functions[3]. In Section 5.2, we analyze the support recovery in the resulting estimate and show that the two q-functions represent the probability of on and off support recovery.

5.1 Convergence behavior of sparse LR

For our analysis in this section, we assume each entry $\boldsymbol{\beta}_i^*$, for $i = 1, \dots, p$, is sampled i.i.d. from a distribution,

$$\Pi(\beta) = (1-s)\cdot\delta_0(\beta) + s\cdot\big(\frac{\phi(\frac{\beta}{\frac{\kappa}{\sqrt{s}}})}{\frac{\kappa}{\sqrt{s}}}\big), \tag{18}$$

where $s \in (0,1)$ is the *sparsity factor*, $\phi(t) := \frac{e^{-t^2/2}}{\sqrt{2\pi}}$ is the density of the standard normal distribution, and $\delta_0(\cdot)$ is the Dirac delta function. In other words, entries of $\boldsymbol{\beta}^*$ are zero with probability $1-s$, and the non-zero entries have a Gaussian distribution with appropriately defined variance. Although our analysis can be extended further, here we only present the result for a Gaussian distribution on the non-zero entries. The proximal operator of $\tilde{f}(\cdot) = |\cdot|$ is the soft-thresholding operator defined as, $\eta(x,t) = \frac{x}{|x|}(x-t)_+$. Therefore, we are able to explicitly compute the expectations with respect to $\tilde{f}(\cdot)$ in the nonlinear system (6). To streamline the representation, we define the following two proxies,

$$t_1 = \frac{\lambda}{\sqrt{\frac{r^2}{\delta} + \frac{\theta^2\kappa^2}{s}}} \ , \ \ t_2 = \frac{\lambda}{\frac{r}{\sqrt{\delta}}} \ . \tag{19}$$

In the next section, we provide an interpretation for t_1 and t_2. In particular, we will show that $Q(\bar{t_1})$, and $Q(\bar{t_2})$ are related to the probabilities of on and off support recovery. We can rewrite the first three

[3]The q-function is the tail distribution of the standard normal r.v. defined as, $Q(t) := \int_t^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx$.

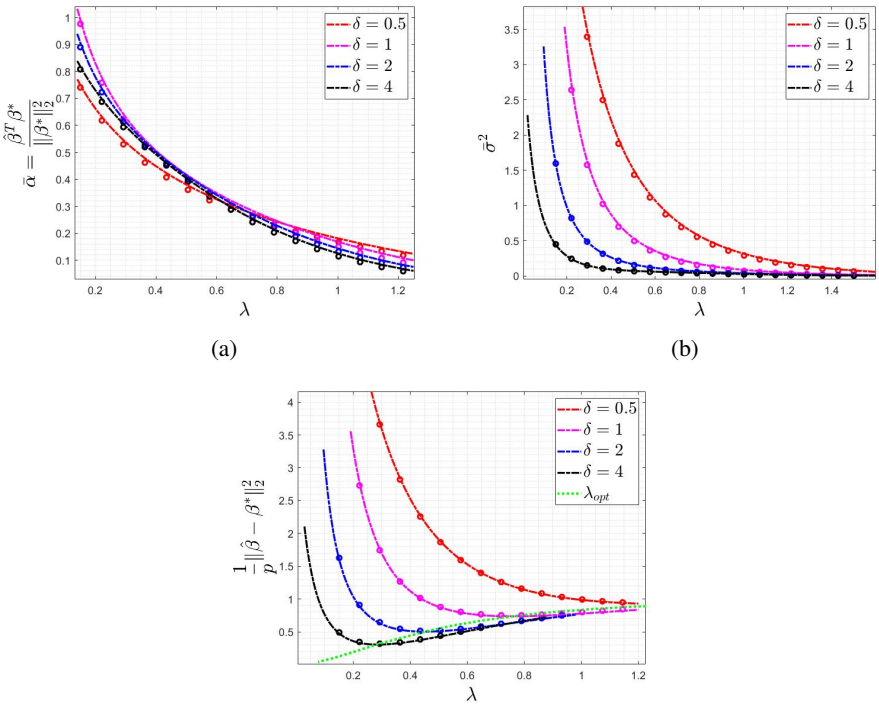


Figure 2: The performance of the regularized logistic regression under ℓ_1 penalty (a) the correlation factor $\bar{\alpha}$ (b) the variance $\bar{\sigma}^2$, and (c) the mean-squared error $\frac{1}{p}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||^2$. The dashed lines are the theoretical result derived from Theorem 1, and the dots are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independent trials with $p = 250$ and $\kappa = 1$.

equations in (6) as follows,

$$\begin{cases} \dfrac{\alpha}{2\sigma\tau} = \theta \cdot Q(t_1) \ , \\ \dfrac{\delta\gamma}{2\sigma\tau} = s \cdot Q\big(t_1\big) + (1-s) \cdot Q\big(t_2\big) \ , \\ \dfrac{\kappa^2\alpha^2 + \sigma^2}{2\sigma^2\tau^2} = \dfrac{\delta\gamma\lambda^2}{2\sigma\tau} + \dfrac{\gamma r^2}{2\sigma\tau} + \kappa^2\theta^2 \cdot Q\big(t_1\big) - \lambda^2(s \cdot \dfrac{\phi(t_1)}{t1} + (1-s) \cdot \dfrac{\phi(t_2)}{t_2}) \ . \end{cases} \tag{20}$$

Appending the three equations in (20) to the last three equations in (6) gives the nonlinear system for sparse LR. Upon solving these equations, we can use the result of Theorem 1 to compute various performance measure on the estimate $\hat{\boldsymbol{\beta}}$. Figure 2 shows the performance of our estimate as a function of λ. It can be seen that the bound derived from our theoretical result matches the empirical simulations. Also, it can be inferrred from Figure 2c that the optimal value of λ (λ_{opt} that achieves the minimum mean-squared error) is a decreasing function of δ.

5.2 Support recovery

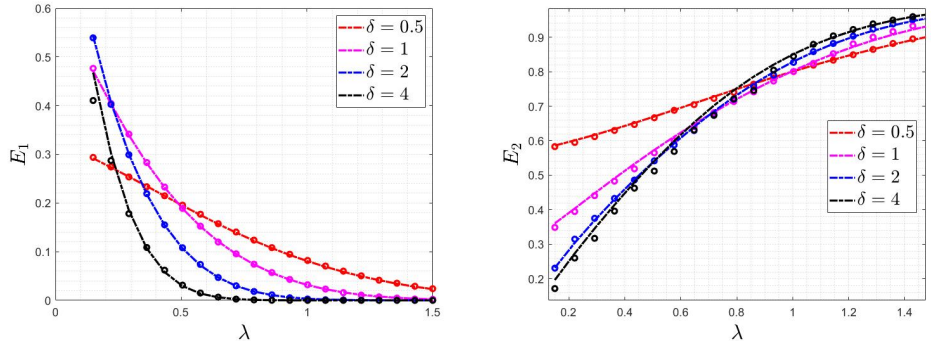
In this section, we study the support recovery in sparse LR. As mentioned earlier, sparse LR is often used when the underlying paramter has few non-zero entries. We define the support of $\boldsymbol{\beta}^*$ as $\Omega := \{j | 1 \leq j \leq p, \boldsymbol{\beta}^*_j \neq 0\}$. Here, we would like to compute the probability of success in recovery of the support of $\boldsymbol{\beta}^*$.

Let $\hat{\boldsymbol{\beta}}$ denote the solution of the optimization (17). We fix the value $\epsilon > 0$ as a hard-threshold based on which we decide whether an entry is on the support or not. In other words, we form the following set as our estimate of the support given $\hat{\boldsymbol{\beta}}$,

$$\hat{\Omega} = \{j | 1 \leq j \leq p, |\hat{\beta}_j| > \epsilon\} \tag{21}$$

In order to evaluate the success in support recovery, we define the following two error measures,

$$E_1(\epsilon) = \text{Prob}\{j \in \hat{\Omega} | j \notin \Omega\} \ \ , \ \ E_2(\epsilon) = \text{Prob}\{j \notin \hat{\Omega} | j \in \Omega\} \ . \tag{22}$$



(a) (b)

Figure 3: The support recovery in the regularized logistic regression with ℓ_1 penalty for (a) E_1: the probability of false detection, (b) E_2: the probability of missing an entry of the support. The dashed lines are the theoretical results derived from Lemma 1, and the dots are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independent trials with $p = 250$ and $\kappa = 1$ and $\epsilon = 0.001$.

In our estimation, E_1 represents the probability of false alarm, and E_2 is the probability of misdetection of an entry of the support. The following lemma indicates the asymptotic behavior of both errors as ϵ approcahes zero .

Lemma 1 (Support Recovery)**.** *Let $\hat{\boldsymbol{\beta}}$ be the solution to the optimization* (17)*, and the entries of $\boldsymbol{\beta}^*$ have distribution* Π *defined in* (18)*. Assume λ is chosen such that the nonlinear system* (6) *has a unique solution $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$. As $p \to \infty$ we have,*

$$\begin{aligned} \lim_{\epsilon \downarrow 0} E_1(\epsilon) &\xrightarrow{p} 2\,Q\big(\bar{t}_1\big) \;\; \textit{where,} \;\; \bar{t}_1 = \frac{\lambda}{\frac{\bar{r}}{\sqrt{\delta}}}, \;\; \textit{and,} \\ \lim_{\epsilon \downarrow 0} E_2(\epsilon) &\xrightarrow{p} 1 - \, 2\,Q\big(\bar{t}_2\big) \;\; \textit{where,} \;\; \bar{t}_2 = \frac{\lambda}{\sqrt{\frac{\bar{r}^2}{\delta} + \frac{\bar{\theta}^2 \kappa^2}{s}}} \, . \end{aligned} \tag{23}$$

6 Conclusion and Future Directions

In this paper, we analyzed the performance of the regularized logistic regression (RLR), which is often used for parameter estimation in binary classification. We considered the setting where the underlying parameter has certain structure (e.g. sparse, group-sparse, low-rank, etc.) that can be enforced via a convex penalty function $f(\cdot)$. We precisely characterized the performance of the regularized maximum likelihood estimator via the solution to a nonlinear system of equations. Our main results can be used to measure the performance of RLR for a general convex penalty function $f(\cdot)$. In particular, we apply our findings to two important special cases, i.e., ℓ_2^2-RLR and ℓ_1-RLR. When the regularizer is quadratic in parameters, we have shown that the nonlinear system can be simplified to three equations. When the regularization parameter, λ, is set to zero, which corresponds to the maximum likelihood estimator, we simply derived the results reported by Sur and Candes [26]. For sparse logistic regression, we established that the nonlinear system can be represented using two q-functions. We further show that these two q-functions represent the probability of the support recovery.

For our analysis, we assumed the datapoints are drawn independently from a gaussian distribution and utilized the CGMT framework. An interesting future work is to extend our analysis to non-gaussian distributions. To this end, we can exploit the techniques that have been used to establish the universality law (see [20, 21] and the references therein). As mentioned earlier in Section 1, an advantage of RLR is that it allows parameter recovery even for instances where the (unconstrained) maximum likelihood estimate does not exist. Therefore, another interesting future direction is to analyze the conditions on λ (as a function of δ and κ) that guarantees the existence of the solution to the RLR optimization (5). In the unstructured setting, this has been studied in a recent work by Candes and Sur [5].

References

[1] Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. Performance analysis of convex data detection in mimo. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4554–4558. IEEE, 2019.

[2] Ismail Ben Atitallah, Christos Thrampoulidis, Abla Kammoun, Tareq Y Al-Naffouri, Babak Hassibi, and Mohamed-Slim Alouini. Ber analysis of regularized least squares for bpsk recovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4262–4266. IEEE, 2017.

[3] Carl R Boyd, Mary Ann Tolson, and Wayne S Copes. Evaluating trauma care: the triss method. trauma score and the injury severity score. *The Journal of trauma*, 27(4):370–378, 1987.

[4] Florentina Bunea et al. Honest variable selection in linear and logistic regression models via 1 and 1+ 2 penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

[5] Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

[6] Oussama Dhifallah, Christos Thrampoulidis, and Yue M Lu. Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *arXiv preprint arXiv:1805.09555*, 2018.

[7] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[8] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

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

[10] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[11] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[12] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[13] Sham Kakade, Ohad Shamir, Karthik Sindharan, and Ambuj Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 381–388, 2010.

[14] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[15] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(Jul):1519–1555, 2007.

[16] Balaji Krishnapuram, Lawrence Carin, Mario AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968, 2005.

[17] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[18] Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

[19] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[20] Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2017.

[21] Ashkan Panahi and Babak Hassibi. A universal analysis of large-scale regularized least squares solutions. In *Advances in Neural Information Processing Systems*, pages 3381–3390, 2017.

[22] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. Learning without the phase: Regularized phasemax achieves optimal sample complexity. In *Advances in Neural Information Processing Systems*, pages 8641–8652, 2018.

[23] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. A precise analysis of phasemax in phase retrieval. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 976–980. IEEE, 2018.

[24] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[25] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[26] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

[27] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, pages 1–72, 2017.

[28] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[29] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709, 2015.

[30] Christos Thrampoulidis, Ilias Zadik, and Yury Polyanskiy. A simple bound on the ber of the map decoder for massive mimo systems. *arXiv preprint arXiv:1903.03949*, 2019.

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

[32] Jack V Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231, 1996.

[33] Sara A Van de Geer et al. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

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