THEORY OF HIGH-DIMENSIONAL ℓ_1-PENALIZED LOGISTIC REGRESSION

by

Emmanuel Otubo

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 5, 2023

Keywords: Penalized Logistic Regression, Generalized Approximate Message Passing (GAMP), State Evolution (SE), Phase Transition, Large-n-large-p asymptotics

Approved by

Peng Zeng, Chair, Associate Professor of Statistics
Guangun (Vivian) Cao, Associate Professor of Statistics
Mark Carpenter, Professor of Mathematics and Statistics
Elvan Ceyhan, Associate Professor of Statistics
Xiaoying (Maggie) Han, Professor of Mathematics

Abstract

In the setting where sample size n is sufficiently large relative to the number of features p, a classical result is that fitting a logistic model by means of maximum likelihood produces estimates that are approximately normal, unbiased and efficient. The usual claim is that these estimations are approximately valid if there are about 5-10 observations per unknown parameter. Sur and Candès (2019) shows in the context of the logistic regression that in the modern setting where the sample size and number of features are large and comparable, this claim is misleading and untrue, and hence, inferences based upon the results of common software packages can be unreliable. This dissertation considers the logistics regression with ℓ_1- penalty and extends the results of Sur and Candès (2019) to the asymptotic regime where the dimension p of the covariates, and the sample size n grow together to infinity in such a way that $n/p \to \delta \in (0, \infty)$. There are two major contributions made here. First, it explicitly characterizes the asymptotic mean square error of the ℓ_1-penalized logistic regression estimators. Secondly, it provides empirical evidence of the existence and the location of a phase transition in the accuracy of signal recovery of the logistic lasso estimator in the two-dimensional sparsity-undersampling phase space. The formalism here is based on the asymptotic analysis of the GAMP algorithm. The findings offer theoretical insights into high-dimensional regression methods. For example, it can be used to tune the regularization parameter since it provides explicit characterization of the asymptotic MSE. Also, the phase transition result provides a guide for when the ℓ_1-regularized estimator is reliable in the context of the logistic regression.

Acknowledgments

Firstly, I give glory to God who is my source and who has always guided my path in all of life's endeavors.

Next, to my advisor, Dr. Peng Zeng, I wish to take this moment to express my deepest appreciation for your guidance and support throughout my journey as a PhD student. Your expertise, patience, generosity and dedication have been invaluable to me, and I am truly grateful for the opportunity to work under your supervision. Your mentorship has not only shaped my research skills but has also helped me grow as an individual. Your ability to provide constructive feedback and encourage critical thinking has pushed me to explore new ideas and challenge myself intellectually. Your unwavering belief in my abilities has given me the confidence to overcome obstacles and pursue ambitious goals. I am particularly grateful for the countless hours you have spent discussing research ideas, reviewing my work, and providing insightful suggestions. Your commitment to excellence and attention to detail have inspired me to strive for the highest standards in my work. Your mentorship has not only enhanced my technical skills but has also taught me the importance of perseverance, resilience, and continuous learning. Thank you once again for your unwavering support, guidance, and mentorship. I am truly grateful for everything you have done for me.

I would also like to thank my dissertation committee members Dr. Guangun (Vivian) Cao, Dr. Mark Carpenter, Dr. Elvan Ceyhan and Dr. Xiaoying (Maggie) Han, and also my dissertation reader Dr. Jingjing Qian, for accommodating me in their very busy schedules and taking time to read my work. I am grateful for the time and effort they have each invested in me, and I am honored to have had the opportunity to work with such a distinguished committee.

Also, the story of my graduate life does not even begin, save for several other wonderful interactions I have had with some faculties of our department. I am indebted to Dr. Govil who has remained like a father figure to me from the first day I arrived Auburn. I am also particularly grateful to Dr. Nedret Billor for being absolutely inspiring and supportive. That I

had the opportunity to do a data science internship in the course of my PhD is all thank to her.. Moreover, her multivariate statistical analysis class was game changing for me. I am immensely thankful to Dr. Overtoun Jenda. Consistently being part of his GABBR STEM academies every summer all through my PhD program and having the opportunity to sit close and be inspired by his impeccable work ethics have been a major influence on my personality formation. I cannot thank him enough. My gratitude also goes to Dr, Ash Abebe. His linear models class and teaching style totally changed my paradigm in statistics and I am really grateful. My gratitude extends to rest of the faculties who in one way or the other have contributed in making my graduate school experience worthwhile.

I owe an immense gratitude to my exceptional support network of friends. The entire Nigerian family in the Maths and Stats department, Sr. Maria, Dr. Asogwa, Dr. Aroh, Dr. Izuchukwu, everyone, thank you all for making Auburn fill like home. I also want to appreciate all my other friends in the department, Jordan, Masuzyo, Padmini and all, too many to mention for all the ways you supported me. I am glad I came to graduate school the same time you all did.

With a great sense of joy, I would also like to take a moment to express my most profound appreciation to my parents, Chief Linus Elum Otubo (The Bish!) and Lolo Virginia Uche Otubo (a.k.a. Enyi wa uzo), and my dearest siblings for their unwavering support throughout my dissertation journey. Their love, encouragement, and belief in me have been instrumental in my success. They have been my pillars of strength, always there to lend a helping hand or provide words of wisdom when I needed them most. I am truly grateful for their constant presence in my life and for the countless sacrifices they have made to ensure my happiness and success. I could not have achieved this milestone without them, and I am forever indebted to them for their love and support. Thank you, from the bottom of my heart.

My wife, Oly, and our children, Uma, Elum, and Olisanuru, have consistently been my unwavering support system, constantly cheering me on and celebrating every achievement. Their love and support have brought immense happiness and meaning to my life, and I am eternally grateful for their understanding and devotion. This dissertation serves as a testament to their selflessness and unwavering loyalty, and I dedicate it to them wholeheartedly.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMP	Approximate Message Passing
GAMP	Generalized Approximate Message Passing
GLMs	Generalized Linear Models

Chapter 1

Introduction

1.1 Big data era and high-dimensional problem

Not too long ago, a gigabyte of data was considered quite large. However, with recent advancements in data technologies, organizations are now managing hundreds of petabytes of data. The world is currently witnessing a rapid data explosion, rendering previous data creations, acquisitions, and storage trivial in comparison. A vast amount of data is now routinely generated in various fields, including scientific research, medical imaging, satellite imagery, climate studies, social media, surveillance videos, and omics data. Today, data has become one of the most crucial assets for businesses, playing a central role in transformative innovations such as artificial intelligence (AI) and machine learning. According to a report by IDC (International Data Corporation) Reinsel et al. (2017), the global data volume was predicted to grow exponentially from 4.4 zettabytes in 2013 to 44 zettabytes in 2020, and is projected to reach an astonishing 163 zettabytes by 2025.

The value offered by the largest companies in the world now largely stems from their data, which they continuously analyze to develop new products and enhance efficiency. The meaning and utilization of data have been completely transformed by today's cutting-edge computers, sensors, tablets, mobile devices, and similar technologies. These highly advanced computing and electronic devices have the ability to perform tasks and generate data across multiple platforms simultaneously. It is now possible to capture numerous features of observations, resulting in a situation where the number of observed features (p) often greatly surpasses the number of observations (n). Instead of dealing with a small number of variables and a few hundred or fewer observations as in the past, scientists and researchers now grapple with the challenge of big data. But, while having access to vast amounts of data is generally beneficial, it is only valuable when its true worth is uncovered.

The term“Big data” has emerged to describe extremely large data sets that are characterized by their variety, velocities, and volumes. The definition of what constitutes ”big data” can vary from organization to organization depending on their available resources. For some, encountering hundreds of gigabytes of data may prompt a reevaluation of their data management strategies, while for others, it may take significantly larger amounts of data before size becomes a significant concern.

Although big data offers many opportunities, it also presents challenges. Data volumes are doubling every few years, and organizations struggle to keep up with the pace and to find effective ways to store it. Additionally, data cleanliness is a concern, as data must be stored in a way that allows for meaningful analysis. Classical statistical theory and methods are often inefficient and infeasible in this context. Dealing with big data involves three levels of challenges: acquisition, management, and analysis. The contribution in this work focuses on the area of analysis.

There is an urgent need for new statistical ideas, scalable algorithms, parsimonious models, and accurate theory to analyze and interpret such data. The primary objective of this work is to gain a deeper understanding of the ℓ_1-penalized logistics regression by examining its asymptotic behavior as the sample size n and the number of predictors p increase together at a fixed rate, i.e., $n/p \to \delta$. This scenario is known as the *large-n-large-p asymptotics*.

The logistic regression model is a specific instance of a broader class known as *Generalized Linear Models* (GLMs) which will now be presented.

1.2 Generalized Linear Models

A problem that naturally occurs with high dimensional data is that, with increasing dimension, the relationship between a response y, and a vector of covariates $x \in \mathbb{R}^p$, compounds very quickly in complexity and applicable procedure of analysis are nontrivial even for traditional techniques. Under certain mild assumptions, the linear regression model has been successfully applied for high dimensional problems. When the assumptions of a normally distributed response variable with constant variance are not met, methods such as weighted least squares and data transformation have also proven effective for modeling purposes. But it turns out, however, that linear regression techniques, are restrictive as they are limited only to settings where some form of linearity assumption is reasonable. The GLMs generalise the linear model idea to a broader framework that includes both the linear and nonlinear regression models and also allows the incorporation of non-normal response distributions. A key assumption in the GLMs is that the distribution of the response variable y, must belong to the exponential family which includes many popular examples like the binomial, Poisson, exponential, gamma and normal distributions. McCullagh and Nelder (1983) provide an in-depth coverage on the algorithms, statistical inference, and theory on GLMs.

The central principle of the GLMs is to establish a linear model for an appropriate function of the expected value of y, the response variable, i.e., find $h : \mathbb{R} \to \mathbb{R}$ such that

$$\eta \ = h\Big(E[Y|X] \Big) = h(\mu) = X\beta, \quad X = (X_1, X_2, ..., X_p). \tag{1.1}$$

Notice here that the covariates $X_i \in \mathbb{R}$, affect the distribution of the response $y \in \{0, 1\}$, only through the linear combination, $X_i^T \beta$. The function h, is called the **link function**. There are several possible choices of the link function and we give a few of the most popular ones in the Table 1.1 below.

Overall, GLMs have three basic components:

- **Random Component** - this sets out the probability distribution of the response variable Y, such as the normal distribution in the case of classical regression model, or binomial distribution in the binary logistic regression model. This is the only random component in the model, and there is not a separate error term.

- **Systematic Component** - this sets out the explanatory variables in the model in the form of their linear combination, $x_i^T \beta$.

- **Link Function:** η - The relationship between the random and systematic components is specified via a a monotonic and differentiable link function h.

The table below shows the link functions for some very popular cases.

Table 1.1: **Link Functions for the GLMs**

Distribution	Link Function
Normal	$\eta = \mu$
Logistic	$\eta = \log\left(\frac{\pi_i}{1-\pi_i}\right)$
Probit	$\eta = \Phi^{-1}(\mu)$
Complementary-log-log	$\eta = \log\{-\log(1-\mu)\}$

Φ represents the cdf of the standard normal distribution function. In the case of the **logistic regression**, the *random component* relates the response variable is assumed to follow the binomial distribution with a single trial and success probability $E(Y) = \pi$. The *Systematic component* is the linear combination of the explanatory variable by the regression parameters. The **Probit Model**, like the logistic regression, is also used to estimate the probability that an observation with specific attributes falls into a specific one of two categories. Though it was primarily developed as a way of estimating probabilities of a binary outcome, it has also seen successful application in binary classification where observations are classified based on their predicted probabilities. The **Complementary-log-log Model** is a third option to the Probit and Logit models. A major difference between the complementary-log-log model and logit/probit models is that while complementary-log-log is asymmetrical, the other two are symmetrical. Complementary log-log models are typically employed when the probability of an event falls in one extreme, i.e., very small or very large.

We note that while linear regression is a very useful apparatus for predicting the values of a quantitative response variables, there are many important situations where the response is not quantitative but qualitative or categorical, and there, the linear regression is not applicable and a different set of tools is needed. For example, consider the problem of predicting a dichotomous outcomes, such as in medical science to predict the risk of a patient developing a certain disease. The logistic regression is perhaps the most widely used statistical model for performing such tasks. In more general terms, unlike with linear regression models, GLMs do not assume a linear relationship between the response variable and the explanatory variables. The linear relationship assumption is between the transformed expected response in terms of the link function and the explanatory variables. This allows GLMs to extend to a broader framework as in the examples above and many other important application.

Several interesting problem formulations have been considered by researchers under the framework of the GLMs, but in this work, we will limit our study to the logistic regression model. We begin by first highlighting a few important building block and results.

1.2.1 The Setup

Binary data are ubiquitous in application across a broad range of subject areas such as Finance, Biology, Medicine, Social Sciences, etc. Logistic regression [Cox (1958)] is arguably the most well known parametric statistical model for fitting binary outcome, Y with a family of covariates $X = (X_1.X_2, ..., X_p)$, and assessing the significance of their coefficients. For example, logistic regression may be used to predict: the chance of an online shopper buying or not buying a particular product based on certain characteristics of the shopper; patients survival or not from a disease; that an individual will have heart disease or not, and so on. Logistic regression models are mostly used as a data analysis and inference tool, where the aim is to estimate the contribution of the input variables in explaining the outcome. Given the frequent occurrence of this model in applications, graduate students in statistics and other fields involving data analysis are usually introduced to logistic regression before any other nonlinear multivariate model.

Logistic regression tries to model how the odds of "success" for a binary response variable Y depend on a set of explanatory variables: Specifically, the model is as follows: let (y_i, X_i) be n independent observations where $y_i \in \{0, 1\}$ is the response variable and $X_i \in \mathbb{R}^p$ the vector of predictor variables. The logistic regression model computes the conditional probability of a case given the covariates via

$$\mathbb{P}(y_i = 1|X_i) = \rho'(X_i^T \beta) \tag{1.2}$$

where $\beta \in \mathbb{R}^p$ is the unknown regression vector. The Logistic regression model is fit via the method of maximum likelihood and in this case, the maximum likelihood estimate (MLE) is the minimizer of the negative log-likelihood given by:

$$\begin{aligned} \hat{\beta} &= \text{arg}\min_\beta \textstyle\sum_{i=1}^n \ell(X_i^T\beta, y_i), \\ \text{where} \quad \ell(z,y) &= \rho(z) - yz, \qquad \rho(t) = \log\left(1+e^t\right). \end{aligned} \tag{1.3}$$

Most standard statistical software have built-in packages that produce p-values for assessing the significance of their coefficients.

1.2.2 Classical Results

Statistical inference for GLMs is justified by asymptotics. The large-n-fixed-p asymptotics is the classic setting and dominant in past literature. It makes sense when there are only a few predictors and the sample size $n \gg p$. The main mathematical tool is the Law of Large Numbers and Central Limit Theorem. The asymptotic analysis starts with the consistency of an estimator and then goes on to the asymptotic normality (McCullagh and Nelder (1983).

Classical statistical inference for a host of parametric models including logistic regression relies on maximum likelihood theory. An important fundamental result in classical statistics is about the asymptotic properties of the MLE which states that under some mild regularity conditions on the underlying model, the MLE $\hat{\beta}$ is normally distributed around the true parameter, with variance given as inverse Fisher's information scaled by root n (the sample size). i.e.,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, I_\beta^{-1}), \tag{1.4}$$

where I_β^{-1} is the $p \times p$ Fisher information matrix evaluated at the true β (Lehmann and Romano (2006)). Explicating, (1.4) says that the MLE has the following large sample properties:

1. the MLE is consistent: $\hat{\beta} \xrightarrow{p} \beta$,

2. the MLE is Asymptotically normal,

3. the MLE is asymptotically efficient, i.e., it achieves minimum variance, the Rao-Cramer lower bound.

Property (1.4) also gives that we can calculated p-values for performing hypotheses testing to assess the significance of regression coefficients, β_j. Software packages commonly rely on (1.4) to produce inferential results for most parametric models, and in particular, the logistic regression. Most students in statistics and related fields know how to interpret such computer outputs.

1.3 Background of Study

Classification and regression problems that involve a large number of candidate feature variables are prevalent in various scientific fields. With the advancements in data collection technologies, it has become the norm to observe data in a more detailed manner, resulting in a larger number of recordable features for individual observations. In genomics, for instance, the number of gene features for a single individual often exceeds the number of available individuals for a study. These types of problems are known as high-dimensional problems and will be the main focus of this work. Before delving into the details, let's introduce the concept of *regularization*, which will aid in understanding the subsequent discussions.

1.3.1 Regularized Estimators

As stated above, GLMs are usually estimated via the method of *Maximum Likelihood (ML)*, which tries to calculate the most likely values of the population parameters, β, say, given the observed data. But, with small to modest sample sizes and complex models, ML estimation of most statistical models can show serious estimation problems such as non-convergence, parameter estimates outside the admissible parameter space, and/or over-fitting as more and more feature variables are added in the model.

In this modern data era, large and complex datasets have become everyday norm for statisticians and data analysts. Large data size affords researchers the ability to pursue non-parametric estimation techniques for "unstable" quantities and "discontinuous" functions of the underlying data distribution. Complexity on the other hand refers to high dimensionality of observations, and is the basis for the reliance on complex models to fit data. But, with complex models come many challenges including over-fitting to radom effect and poor prediction. Regularization procedure is one approach to mitigate these problems. Regularization helps solve over-fitting by penalizing high-valued regression coefficients, thereby, reducing parameters and shrinking/simplifying the model. The general idea is to adds penalties to more complex models and then calculate over-fitting scores for potential models; The model with the lowest over-fitting score usually yields the most predictive power.

In practice, a large number of predictors are usually introduced at the initial stage of modeling to attenuate possible modeling biases. To enhance predictability and to select significant variables, statisticians traditionally used stepwise deletion and subset selection (see A. (1960), Hocking (1976)). Although they are practically useful, these selection procedures ignore stochastic errors inherited in the stages of variable selections. Hence, their theoretical properties are somewhat hard to understand. To solve this problem, many other variable selection procedures have been studied; least squares regression methods with various kinds of penalties aimed at tackling the complexity problem, and increasing prediction accuracy have collected much interest, and a substantial amount of research efforts have been poured in that direction by statisticians; Hoerl and Kennard (1970) introduced the ridge regression which proves effective for handling the problem of multicollinearity and prediction performance, however, ridge regression leaves the variable selection problem unattended. Tibshirani (1996) introduced the Least Absolute Shrikage Selection Operator (Lasso) which uses an ℓ_1-penalized likelihood for linear regression with independent Gaussian noise. The major attraction of the lasso and related methods is that they offer interpretable, stable models, and an efficient prediction at a reasonable cost. Lasso provides a way to simultaneously performs variable selection and coefficient estimation. In short, lasso provides a simple, data-guided approach for choosing the optimal level of model complexity, in essence, the degree of regularization of the model.

In a general context, regularization is a process of imposing constraint on model parameters to solve an inference problem, such as MLE, that is unstable or not solvable by regular methods. In other words, regularization introduces some bias in exchange for a larger reduction in variance and hence minimizes over-fitting. The regularized solutions are more stable and the resultant models are usually simpler. To put this in Math terms, a regularization approach is intended to estimate the parameter, β via

$$\hat{\beta}_h = \arg\min_{\beta} \sum_{i=1}^{n} \ell(\boldsymbol{\beta}) + h(\boldsymbol{\beta}) \tag{1.5}$$

where ℓ is some loss function, a common example of which is the negative log-likelihood function, and $h(\boldsymbol{\beta}) = \sum_{j=1}^{p} h(\beta_j)$ for some convex penalty function $h(\cdot)$. The particular case of interest for us in this work is the LASSO, where $h(\beta_j) = \lambda|\beta_j|$, and $\lambda \geq 0$ is the tuning parameter. Given the negative log-likelihood loss function, for $\lambda = 0,\ \hat{\beta}_h$ is exactly the maximum likelihood estimator, while when $\lambda \to \infty$, we usually have $\hat{\beta}_h \to 0$. Hence, the value of λ controls the sparsity of the estimator $\hat{\beta}_h$. There are many existing algorithms for solving (1.5); we will do an depth review in section 1.3.4.

The seminal works of Tibshirani (1996) for lasso and Chen et al. (1998) for basis pursuit spark the interest of researchers in ℓ_1-regularization. Fu and Knight (2000) showed that λ should be selected adaptively to the sample size n in order to get the consistency of $\hat{\beta}(\lambda)$. Zhao and Yu (2006) discussed the variable selection inconsistency of lasso and identified a sufficient condition that guarantees the consistency of variable selection. Fan and Li (2011) promoted penalty with the oracle properties, where zero components can be selected with high probability and nonzero components can be estimated with the same efficiency as if they are known in advance. Several other authors, even within the last decade have continued to study the distribution of M estimators in linear models under various conditions. For example, Karoui et al. (2013); Bean et al. (2013); Donoho and Montanari (2013); Karoui and Noureddine (2018), and a host of others have studied several topical issues on the M estimators.

Unfortunately, the results from the above literature do not carry forward directly, to the GLMs, as they all require a loss function that is strongly convex, and this is not the case for many GLMs. Hence, statisticians have been interested in recovering and developing their analogues in the larger context of the GLMs. There is a large literature on the asymptotic behavior of M estimators for exponential families. For example, Portnoy (1988) studied M estimators for exponential families in the regime where $p^2/n \to 0$. He and Shao (2000) considered M-estimators of general parametric models that include the GLMs.

For various objectives, researchers have continued to study different forms of penalized regression. The notion of effective dimension was introduced by Spokoiny (2012), and they extended classical maximum likelihood results to maximum likelihood estimates with quadratic penalization. They achieve analogues of asymptotic normality and also the Wilk's phenomenon for the penalized MLE in the case of parametric models when the effective dimension is low-dimensional. The other major interest stemming from classical statistics pertain to the consistency of estimators. And to get a consistent estimator for β, some prior knowledge of β is usually needed. The most common assumption is the sparsity of β, in essence, most components of β are zeros. Existing research shows that provided β is sufficiently sparse, the consistency of $\hat{\beta}(\lambda)$ is achievable for a proper choice of λ, even when $p \gg n$ (see e.g., Candes and Tao (2007), Donoho et al. (2006), Meinshausen and Bühlmann (2006), Tropp (2004), Wainwright (2006), Zou (2006)).

1.3.2 *diverging-p* asymptotics

It is worth noting that most of the earlier literature focuses on the regime of *large-n-fixed-p asymptotics*, where the number of predictors, p, is fixed and the sample size, n, tends to infinity. However, in modern data scenarios, it is common to encounter situations where p is close to n, p is greater than n, or even p is much larger than n. In these cases, the classical maximum likelihood estimation results are known to be inadequate. Therefore, it becomes necessary to study the asymptotics for increasing values of p, where p grows alongside n.

Because of the high dimensionality in modern data, it is not appropriate to assume fixed p anymore. The asymptotics on regularization approaches has gradually shifted to the scenario of diverging p. The main goal is to identify suitable conditions that allow p to grow to infinity and at the same time maintain consistency and asymptotic normality of the estimators. The discrepancy between the theory and practical performance was noticed a long time ago, particularly when p is large. One common solution is to explore higher-order asymptotics, where p is still kept fixed as $n \to \infty$. Examples include the Edgeworth expansion of the distribution of MLE and Bartlett's correction for the p-value of the likelihood-ratio test.

The diverging-p asymptotics dates back to the work of Huber (1973), Portnoy (1984), Portnoy (1985), and Mammen (1989) on M-estimation. For regularization approaches, Fan and Peng (2004) showed that if p does not grow too fast, then there exists a penalized likelihood estimator enjoying the oracle properties with a proper choice of λ. They also showed the asymptotic normality of the estimators when $p^4/n \to 0$ and the Wilk's theorem for the likelihood ratio test when $p^5/n \to 0$ as $n \to \infty$. Zou and Zhang (2009) weakened the assumption to $p \sim n^v$ for $0 \leq v < 1$.

In recent researches, p can grow as fast as $\log p = o(n)$. There is a thread of literature on *oracle inequalities*, which provide non-asymptotic upper bounds on the MSE of $\hat{\beta}(\lambda)$ and can be used to prove its consistency. One important problem is to identify easy-to-verify conditions that guarantee the validity of the oracle inequalities. Feuer and Nemirovski (2003) and Cohen et al. (2009) introduced the restricted nullspace property as a sufficient and necessary condition for the exact recovery of coefficients for a k-sparse system. Other sufficient and easy-to-verify conditions include incoherence property (Donoho and Huo (2001)) and restricted isometry property (Candes and Tao (2007, 2005)). Bickel et al. (2009) introduced the weaker restricted eigenvalue condition to discuss the problem of the noisy case. van de Geer and Bühlmann (2009) studied the connection among these different conditions and showed that the compatibility condition (van de Geer (2007)) is the least restrictive. Refer to Bühlmann and van de Geer (2011) for more details.

1.3.3 Large-n-large-p setting

The *large-n-large-p* asymptotics considers the scenario of $n/p \to \delta$, that is, p grows with n at a fixed rate. The performance of $\hat{\beta}(\lambda)$ is studied for fixed λ and when the number of nonzero components of β grows at the same rate as n. Some researchers (e.g. Bayati and Montanari (2012), and a host of others) have derived explicit formulas to evaluate the asymptotic mean squared error (AMSE) of $\hat{\beta}$ for linear regression in different settings. It has also been shown that $\hat{\beta}(\lambda)$ demonstrates a sharp phase transition in the space (δ, ϵ), where ϵ is the sparsity ratio (defined below), see Amelunxen et al. (2013) and the refernces contained therein. Heuristically, these researches precisely characterize the impact and interaction of the following issues on the performance of the regression estimate $\hat{\beta}(\lambda)$.

- *Under-sampling ratio*, defined as $\delta = n/p$, measures the number of observations per predictor
- *Sparsity ratio*, defined as $\epsilon = ||\beta||_0/p$, measures the proportion of nonzero components of β, where $|| \cdot ||_0$ is the number of nonzero components of a vector.
- *Signal strength*, defined as $\gamma^2 = \text{var}(\beta^T x_i)/n$, measures the spread of predictors along the true parameter β.
- *Regularization*, measured by λ, controls the amount of desired sparsity in an estimator

The methods for statistical inference for the large-p domain are usually justified when $\log p = o(n)$. To get consistency and asymptotic normality of $\hat{\beta}(\lambda)$, they need assumptions such as the covariance matrix of x_i is not singular or diverging, nonzero components of β are significantly different from 0 at a certain rate, the magnitude of the sparsity of β affects how fast p can grow, λ needs to be selected at a certain rate of n, etc. Negahban et al. (2012) provide a unified framework for establishing consistency and convergence rates for regularized M-estimators under high-dimensional scaling. The main message of these researches is that when $\log(p) \cdot ||\beta||_0 \ll n$, the consistency and asymptotic normality of 3(1) can be obtained for a proper choice of λ under suitable regularity conditions.

The mathematical tools include the theory of Approximate Message Passing (AMP) and random matrices. We will explore more details later.

1.3.4 Computing Algorithms

In this section we explore the current literature on GLMs and the Approximate Message Passing (AMP) technique. Roughly speaking, the regularization approach for GLMs was primarily used for variable selection and prediction in the past because $\hat{\beta}(\lambda)$ is not consistent for $\lambda > 0$. In the last decade, research on statistical inference such as confidence interval and hypothesis testing based on $\hat{\beta}(\lambda)$ began to emerge. The theoretical justification is mainly from the perspective of diverging-p asymptotics and assumes that λ is selected adaptively to the sample size n.

There are several algorithms that have been proposed to compute $\hat{\beta}(\lambda)$ by solving (1.5). The difficulty lies in that the penalty term, $J(\cdot)$, which is usually non-differentiable, thus making standard algorithms such as Newton-Raphson algorithm not applicable. The coordinate descent algorithm cyclically optimizes one component of β while keeping the remaining fixed (Friedman et al. (2007); Wu and Lange (2008); Friedman et al. (2010); Zhao et al. (2014)). The solution path algorithm is intended to compute the whole solution path of $\hat{\beta}(\lambda)$ for all $\lambda \in [0, \infty)$ (Efron et al. (2004); Park and Hastie (2007); Rosset and Zhu (2007); Tibshirani and Taylor (2011)). The alternating direction method of multipliers (Boyd et al. (2004)) and proximal algorithm (Parikh and Boyd (2014)) are two classes of powerful convex optimization algorithms that are applicable to solve (1.5) in general. For further discussions, reader is refered to Boyd et al. (2004), Lange et al. (2014), and Hastie et al. (2015).

Tibshirani (1996) showed that some components of a lasso estimator can be exactly zero, which implies variable selection and parameter estimation simultaneously. This property has been preserved by almost all regularization approaches proposed later. Refer to Desboulets (2018) for a recent review on variable selection in regression. Early research on statistical inference based on $\hat{\beta}(\lambda)$ includes Zhang and Zhang (2011), Javanmard and Montanari (2014a), Javanmard and Montanari (2014b), and van de Geer et al. (2014). They approved different algorithms to debias $\hat{\beta}(\lambda)$, construct confidence intervals, and perform hypothesis testing. Recent developments include Ning and Liu (2017), Shi et al. (2019), Zhu et al. (2020), Xia et al. (2020), Janková et al. (2020), Ma et al. (2021), Guo et al. (2021), and Cai et al. (2021) among others. With these efforts, the regularization approach for GLMs has become a full-fledged tool for statistical inference.

More recently, one technique that has become extremely popular as a result of it's successful application in various statistical estimation task is the class of iterative algorithm referred to as Approximate Message Passing (AMP) algorithms. It has been successfully applied in linear regression, GLMs, and low-rank matrix estimation (Donoho et al. (2009); Bayati and

Montanari (2011); Fan and Li (2011); Rangan (2011); Schniter and Rangan (2012); Mondelli et al. (2021); Matsushita and Tanaka (2013); Deshpande et al. (2016)). AMP algorithms have some features that make them specially attractive; they are easily adapted to leverage prior information on the signal structure, such as sparsity or other constraints, and under certain assumptions on a design matrix, AMP theory precisely specifies statistical procedures in the high-dimensional regime where the dimensionality parameter p/n converges to a constant (Bayati and Montanari (2012); Donoho et al. (2013)). In high dimensional regime where p is not negligible compared to n, AMP theory provides a precise expression for the asymptotic MSE of the LASSO estimator and not just an upper bound.

The most striking and critical attribute of the AMP recursion is that in large dimension, the empirical distribution of the coordinates of each iterate is approximately normal, with asymptotic variance given by a system of scalar equations called 'state evolution'.

Rangan (2011) proposed the class of *generalized* AMP (GAMP) algorithms as an extension of the AMP to incorporate arbitrary distribution on both the input and output variables. Krzakala et al. (2012) demonstrated the application of a GAMP algorithm for solving general convex optimization problem.

GAMP (Generalized Approximate Message Passing) algorithms are another class of iterative algorithms used for signal recovery in compressed sensing and sparse signal processing problems. These algorithms are based on the principles of belief propagation and approximate message passing. The connection between GAMP and AMP algorithms lies in their underlying principles and iterative procedures. Both algorithms aim to estimate sparse signals from noisy measurements by iteratively updating the estimates based on the observed data and prior information.

GAMP algorithms were initially proposed by Rangan (2011) as a generalization of AMP algorithms. They provide a framework for solving a wide range of signal recovery problems, including compressed sensing, channel estimation, and sparse signal reconstruction. The key idea behind GAMP algorithms is to use approximate message passing techniques to iteratively update the estimates of the sparse signal and the noise variance. These updates are based on the observed measurements and prior information about the signal sparsity and noise statistics.

The connection between GAMP and AMP algorithms can be understood by considering the iterative update equations used in both algorithms. Both algorithms use similar update equations that involve the computation of posterior mean and variance estimates. However, GAMP algorithms incorporate additional steps to handle non-Gaussian noise and non-linear measurement models.

Several variants and extensions of GAMP and AMP algorithms have been proposed in the literature. These include adaptive GAMP algorithms, distributed GAMP algorithms, and GAMP algorithms for structured signal recovery.

1.4 Motivation

1.4.1 Surprising Results of Candes

Sur et al. (2017); Sur and Candès (2019); Candès and Sur (2020) provide new insights on the high dimensional behaviour of logistic regression. They show that for commensurately large n and p, the MLE is biased, contradicting the expectation of classical theory expressed in (1.4), that the MLE is asymptotically unbiased for $n \gg p$. Their result also show that the variability of the MLE is greater than commonly predicted. Hence, the commonly used procedure for significance test of the regression coefficients needs to be adjusted for improved accuracy.The result in Sur and Candès (2019) provides explicit expressions for the bias and variance of the maximum likelihood estimate and describes the asymptotic distribution of the likelihood-ratio statistic given some assumptions.

1.4.2 Problem

Sur et al. (2017); Sur and Candès (2019); Candès and Sur (2020) discuss the asymptotic properties of MLE for unregularized logistic regression in high dimension. It is our interest in this project, to extend their results to the more general setting of ℓ_1-penalized logistic regression in high dimension. Specifically, our target is to estimate the solution to the following optimization problem: for $\lambda > 0$,

$$\begin{aligned} \hat{\beta}_{\ell_1} \quad &= \arg\min_\beta \left\{ \sum_{i=1}^n \ell(x_i^T\beta, y_i) + \lambda \sum_{j=1}^p |\beta_j| \right\}, \\ \text{where} \quad & \ell(z,y) = \rho(z) - yz, \qquad \rho(t) = \log\left(1+e^t\right) \end{aligned} \tag{1.6}$$

The ultimate goal is to explore the impact and interaction of under-sampling, sparsity, signal strength, and regularization on the performance of $\hat{\beta}_{\ell_1}$.

1.4.3 Our Contribution

This dissertation starts by explaining the generalized approximate message passing (GAMP) algorithm, which is used to compute the ℓ_1-penalized logistic regression estimator. It specifies the state evolution, which is a scalar recursion that governs the behavior of both the GAMP algorithm and the logistic lasso estimate. The state evolution provides insights into the operating characteristics of these estimators. Once formulated, the GAMP recursion is then used in the main theorem for the derivation and explicit characterization of the asymptotic limits of the estimator. Some of the consequences of this characterization is the derivation of the asymptotic mean squared error of the logistic lasso estimator, and the asymptotic selection error rate to name a few. The results from rigorous numerical experiments in finite size systems with the tuning parameter λ chosen to minimize the MSE show that there is almost perfect agreement between theoretical predictions and actual values from simulated samples.

Furthermore, by mean of carefully designed numerical experiments, this dissertation provides clear evidence that the logistic lasso estimator undergoes a phase transition in the two dimensional sparsity-undersanpling phase space, $0 \leq \epsilon, \delta \leq 1$. It presents graphical evidence of a phase transition curve that partitions the phase space into two regions. The region above the curve, is the success region where the estimator succeeds in perfectly recovering the signal with high probability. And the region under the curve is the faillure region where the estimator fails with high probability.

Lastly, following the results from earlier works involving the lasso penalty, this dissertation conjectures that the phase transition curve for the logistic lasso estimator will be identical to the previously established phase transition curve in the problem of the reconstruction of underdetermined linear systems in compressed sensing in the k-sparse noiseless case (see Donoho et al. (2011)).

Chapter 2

AMP and related techniques

2.1 Introduction to Generalized Approximate Message Passing

2.1.1 Approximate Message Passing (AMP)

In simple terms, AMP refers to a class of fast iterative algorithms that decouples matrix problems to scalar channel denoising based on belief propagation. Consider $\beta \in \mathbb{R}^p$, a noise vector $\epsilon \in \mathbb{R}^n$, a random design matrix X, and the linear model given by: $y = X\beta + \epsilon$. Donoho et al. (2009) introduced the following AMP algorithm for reconstructing β given $X,\ y$. Starting with an initial guess $\beta^0 = 0$ and recursively obtaining the sequences $\{\beta^t\}_{t\geq 0}$ and $\{\theta^t\}_{t\geq 0}$ by

$$\begin{cases} \beta^{t+1} & = \eta\left(X^T\theta^t + \beta^t\right) \\ \theta^t & = y - X^T\beta^t + \frac{1}{\delta}\theta^{t-1}\left\langle \eta'_{t-1}\left(X^T\theta^{t-1} + \beta^{t-1}\right)\right\rangle \end{cases} \tag{2.1}$$

where η_t are component-wise scalar dinoising functions, $\beta^t \in \mathbb{R}^p$ is the iteration t^{th} estimate of β, and $\theta^t \in \mathbb{R}^p$ the current residual. The notation $\eta'(\cdot)$ refers to the first partial derivative of $\eta(\cdot)$ w.r.t. the first argument, and for any vector $v \in \mathbb{R}^p$, $\langle v \rangle = 1/p\sum_{i=1}^p v_i$.

The AMP algorithm (2.1) is a special case of the general iterative procedure introduced by Bai and Silverstein (2010), which takes the following form: for each $t \geq 0$, let f_t, g_t : $\mathbb{R}^2 \to \mathbb{R}$ be Lipschitz continuous functions, and define the vector sequences $h^t,\ q^t \in \mathbb{R}^p$ and $z^t,\ m^t \in \mathbb{R}^n$, by fixing initial condition q^0, and obtaining $\{b^t\}_{t\geq 0},\ \{m^t\}_{t\geq 0},\ \{h^t\}_{t\geq 1}$, and

$\{q^t\}_{t\geq 1}$ through

$$\begin{aligned} h^{t+1} &= A^T m^t - \xi_t q^t, \quad m^t = g_t(b^t, \epsilon) \\ b^t &= A q^t - \lambda_t m^{t-1}, \quad q^t = f_t(h^t, \beta) \end{aligned} \tag{2.2}$$

where $\xi_t = \langle g_t'(b^t, w)\rangle,\ \lambda_t = \frac{1}{\delta}\langle f_t'(h^t, \beta)\rangle$; and $g_t'(u, \cdot)$ and $f_t'(v, \cdot)$ are $\frac{\partial}{\partial u} g_t$ and $\frac{\partial}{\partial v} f_t$ respectively, and by definition $m^{-1} = 0$.

AMP scheme (2.1) is recovered from (2.2) by defining:

$$\begin{aligned} h^{t+1} &= \beta - (X^T\theta^t + \beta^t), \qquad & q^t &= \beta^t - \beta \\ b^t &= \epsilon - \theta^t, & m^t &= -\theta^t \end{aligned}$$

and the function f_t and g_t are given by

$$f_t(s, \beta) = \eta_{t-1}(\beta - s) - \beta, \qquad g_t(s, \epsilon) = s - \epsilon. \tag{2.3}$$

The main difference between (2.1) and (2.2) is with the data matrix X in place of the theoretical noise matrix A. The term $\frac{1}{\delta}\theta^{t-1}\left\langle \eta_{t-1}'\left(X^T\theta^{t-1} + \beta^{t-1}\right)\right\rangle$, referred to as the *Onsager term* is crucial for ensuring that the AMP iterates θ^t have the desired asymptotic distributional properties depending on the choice of Lipschitz functions η_k.

Effectively, the AMP paradigm introduced by Donoho et al. (2009) in the context of the compressed sensing makes a simple modification to iterative thresholding so that the sparsity-under-sampling tradeoff of the new algorithms is equivalent to that of the corresponding convex optimization procedures. The especially impressive feature of AMP algorithms is that their high-dimensional behavior admits an exact description. In high dimension, the empirical distribution of the coordinates of each iterate is approximately Gaussian, and the variance can be computed via a scalar recursion called *state evolution (SE)*, which we now formally introduce.

2.1.2 State Evolution

Suppose the limit

$$\sigma_0^2 \equiv \lim_{p\to\infty} \frac{1}{p\delta} \left\|q^0\right\|^2$$

exists, is positive and finite. State evolution refers to the iterations $\{\tau_t^2\}_{t\geq 0}$ and $\{\sigma_t^2\}_{t\geq 0}$ defined via

$$\tau_t^2 = \mathbb{E}\left[g_t(\sigma_t Z, \epsilon)^2\right], \qquad \sigma_t^2 = \frac{1}{\delta}\mathbb{E}\left[f_t(\tau_{t-1} Z, \beta)^2\right] \tag{2.4}$$

where $\epsilon \sim \pi_\epsilon$ and $\beta \sim \pi_\beta$ are independent of $Z \sim N(0.1)$.

In the abstract form, the AMP recursion (2.4) is not planned as an algorithm for statistical estimations. To make statement about the statistical properties, we introduce the *state evolution* (SE) formalism which is a deterministic recursion that tracks the behavior of the AMP iterates. SE tells us something about the properties and large system of the AMP. We present this now. Given a probability distribution p_β, let $\tau_0^2 \equiv \sigma^2 + \mathbb{E}(\beta^2)/\delta$, and for $t \geq 0$,

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta_t(\beta + \tau_t Z) - \beta]^2\right\} \tag{2.5}$$

with $\beta \sim \pi_\beta$ and $Z \sim N(0,1)$ independent from β. The recursion (2.5) is termed the *state evolution*. With this, we are now ready to present the main theoretical result. But first, we define the following term: For $k \geq 1$, a function $\psi : \mathbb{R}^m \mapsto \mathbb{R}$ is said to be *pseudo-Lipschitz* of order k if there exists a constant $L > 0$ such that $|\psi(x) - \psi(y)| \leq L(1 + \|x\|_2^{k-1} + \|y\|_2^{k-1})\|x - y\|_2$ for any $x, y \in \mathbb{R}^m$.

We make the following assumptions which are needed for the *SE* to be valid:

(A1) $n, p \to \infty,\ n/p \to \delta \in (0, \infty)$.

(A2) The components of X are iid $N(0, 1/n)$, and are independent of β.

(A3) As $p \to \infty$, the empirical distribution of the components of β converges weakly to a probability measure π_β on $\mathbb{R}$ with bounded second moment. Also, $1/p \sum_{i=1}^{p} \beta_i^2 \to \mathbb{E}_\beta(\beta^2)$.

(A4) The empirical distribution of the entries of $\boldsymbol{\epsilon}$ converges weakly to a probability measure π_ϵ on $\mathbb{R}$ with bounded second moment. Also, $1/n \sum_{i=1}^{n} \epsilon_i^2 \to \mathbb{E}_\epsilon(\epsilon^2)$.

Next, we now present the following AMP *master theorem*

Theorem 2.1 (Bayati and Montanari (2011)). *For any pseudo-Lipschitz function $\psi : \mathbb{R}^2 \mapsto \mathbb{R}$, of order k and all $t \geq 0$, almost surely*

$$\lim \frac{1}{p} \sum_{i=1}^{p} \psi(\beta_i^{t+1}, \beta_i) = \mathbb{E}\Big[\psi\left(\eta_t(\beta + \tau_t Z), \beta\right)\Big]. \tag{2.6}$$

Remark 1. Theorem 2.1 says that if we have a pseudo Lipschitz function that we are interested in, then we can determine the high dimensional properties of the iterates of the AMP algorithm based on the state evolution values.

Remark 2. Another interesting way to think about the convergence in theorem 2.1 is through statements about empirical distribution of the elements. What theorem 2.1 shows is that if π_{β^t} denotes the empirical distribution of the elements of β^t then $\pi_{\beta^t} \to \mathcal{N}(0, \tau_t^2)$. The variances τ_t^2 are determined via the SE recursion (2.5), which depends on the choice of Lipschitz functions $\{\eta_t : t \in \mathbb{N}_0\}$ which in this particular case is the soft threshold function.

The AMP has some features that make them attractive, namely

1. they can be easily tailored to exploit prior knowledge on the signal, e.g. sparsity

2. they achieve faster convergence relative to comparable techniques

3. they have precise asymptotic performance guarantees, in the regime where $n, p \to \infty$ such that $n/p \to \delta > 0$.

2.2 Generalized AMP - (GAMP)

Our study in this work considers a different type of generalization of the AMP that was motivated by recent developments. Likelihood-based inference for the parameter β in (2.9) is justified in large-n-fixed-p-asymptotic regime when n grows much faster than p, Portnoy (1984, 1985). However, in modern large-n-large-p asymptotic regimes where p diverges with n in a non-vanishing rate, different tools are needed to construct and analyse estimators of β, and it is in this context that we present the *generalized Approximate Message Passing (GAMP)* paradigm.

Rangan (2011) proposed the class of generalized approximate message passing (GAMP) algorithms, as an extension of the AMP that is applicable for GLM. GAMP admits application to nonlinear estimation problems wherein $\beta \in \mathbb{R}^p$ is to be estimated given observations $y = (y_1, ..., y_n)$. Examples of this include the logistic, binomial, and Poisson regression, to which the original AMP is not ordinarily applicable. But, before presenting the GAMP algorithm, let us first introduce the *proximal operator* which is used in the discussion.

2.3 Proximal Gradient Method

2.3.1 Proximal Operator

The proximal operator of a convex function $h : \mathbb{R} \to \mathbb{R}$, is given by

$$\text{prox}_h(u, \alpha) = \arg\min_{\beta} \left\{ h(\beta) + \frac{1}{2\alpha} \|\beta - u\|_2^2 \right\}, \qquad \alpha > 0. \tag{2.7}$$

Basically, the operator tries to minimize the value of $h(\cdot)$, but we are penalized if we move to far away from u.

The proximal gradient method is an optimization algorithm and it solves optimization problems that have the following form

$$\min \left\{ L(\beta)_{\text{differentiable}} + h(\beta)_{\text{simple}} \right\}. \tag{2.8}$$

where $L(\cdot)$ is a differentiable function, and $h(\cdot)$ is a simple function, in the sense that it's proximal operator has a closed form formula. Proximal gradient methods provide a general framework for solving regularization problems from statistical learning theory where the regularization penalty may be non-differentiable.

There are many methods to estimate the MLE of the logistic regression, but the proximal gradient method is the most natural choice since the proximal operator is always non-decreasing and 1-Lipschitz, which is a desirable property for AMP/GAMP algorithms.

We briefly introduce GAMP following the notations used in Feng et al. (2021). Assume that $\{(y_i, x_i), y_i \in \mathbb{R}, x_i \in \mathbb{R}^p, i = 1, 2, ..., n\}$ is an iid sample from the following model,

$$y_i = h(\beta^T x_i, \varepsilon_i), \;\; i = 1, 2, ..., n \tag{2.9}$$

where $h : \mathbb{R}^2 \to \mathbb{R}$ is a known function, ε_i is a random error independent of x_i with $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma_\varepsilon^2$, and $\beta \equiv (\beta_1, ..., \beta_p)$ is the goal of inference. For ease of notation, let $(y_1, ..., y_n)^T \in \mathbb{R}^n$ be the response vector and $X = (x_1, ..., x_n)^T \in \mathbb{R}^{n \times p}$ be the design matrix.

2.3.2 GAMP Algorithm

To start, consider problem (1.5) above. For the time being, let $b_k > 0$ and $c_k < 0, \; k = 0, 1, ...$ For $k \geq 0$, define the following proximal operators $f_{k+1} : \mathbb{R} \to \mathbb{R}$ and $g_k^* : \mathbb{R}^2 \to \mathbb{R}$:

$$\begin{aligned} g_k^*(w, v) = &\quad \arg\min_{z\in\mathbb{R}} \left\{ \ell(z, w) + \tfrac{1}{2b_k}(z - v)^2 \right\} = \text{prox}_{\ell(\cdot,w)}(v, b_k) \\ f_{k+1}(u) = &\quad \arg\min_{z\in\mathbb{R}} \left\{ J(z) + \tfrac{c_k}{2}(z + \tfrac{u}{c_k})^2 \right\} = \text{prox}_J\left(-u/c_k, 1/c_k\right) \end{aligned} \tag{2.10}$$

and define $g : \mathbb{R}^2 \to \mathbb{R}$ by

$$g_k(w, \theta) := \frac{g_k^*(w, \theta) - u}{b_k}. \tag{2.11}$$

Note that under the assumption that ℓ and J are convex in their first arguments, $g_k^*(w, v)$, and $f_{k+1}(u)$, are well defined as minimizers of strongly convex functions. Also, by being defined in terms of the proximal operators, g_k^*, g_k and f_{k+1} are all Lipschtz with Lipschitz constants

$1,\ 1/b_k$ and $1/|c_k|$ respectively, and thus, weakly differentiable w.r.t. their first arguments, and we have:

$$g_k^{*'}(w,v) \leq 1, \qquad g_k'(w,v) \leq 0, \qquad f_{k+1}'(u) \geq 0, \tag{2.12}$$

where the derivatives are with respect to their first arguments.

With this, we will now define the GAMP recursion. The GAMP recursion proposed by Rangan (2011), iteratively produces estimates $\theta^k, \hat{\beta}^k$ of $\theta = X\beta \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$, respectively in (2.9), via the following update steps: initialize $\hat{r}^{-1} = 0, b_0 \in \mathbb{R}$ and $\hat{\beta}^0 \in \mathbb{R}^p$, update $\theta^k \in \mathbb{R}^n, \hat{\beta}^k \in \mathbb{R}^p$ and scalars c_k and b_k recursively for $k = 0, 1, 2, ...$ as follows.

$$\begin{aligned} &\theta^k = X\hat{\beta}^k - b_k\hat{r}^{k-1}, &&\hat{r}^k = g_k(\theta^k, y), &&c_k = \tfrac{1}{n}\textstyle\sum_{i=1}^n g_k'(\theta_i^k, y_i), \\ &\beta^{k+1} = X^T\hat{r}^k - c_k\hat{\beta}^k, &&\hat{\beta}^{k+1} = f_{k+1}\left(\beta^{k+1}\right), &&b_{k+1} = \tfrac{1}{n}\textstyle\sum_{j=1}^p f_{k+1}'\left(\beta_j^{k+1}\right), \end{aligned} \tag{2.13}$$

As in the AMP case, the Onsager correction terms $-b_k\hat{r}^{k-1}$ and $-c_k\hat{\beta}^k$ are designed to ensure that in a high-dimensional limiting regime where p is not vanishingly small compared to n, the emperical distribution of the iterates in 2.13 converges to well-defined Wasserstein limits. It turns out that for each $k \in \mathbb{N}_0$, the iterates $\hat{\beta}^{k+1} \in \mathbb{R}^p$ have approximately the same empirical distribution as $f_{k+1}(\mu_k\beta + \sigma_k\xi)$ when p is large; $\beta \in \mathbb{R}^p$ and $\xi \sim N(0, I_p)$ here are the unknown signal and the independent noise vector, respectively, and f_{k+1} can be seen as the denoiser. Considered with the corresponding state evolution which we will describe later, this establishes the basis of a systematic approach to deriving precise performance guarantees for both penalized and unpenalized M-estimators including the Lasso for GLMs in high dimensions.

We will now move to set the ground to introduce the state evolution and state the master theorem for the GAMP. Consider a sequence of recursions (2.13) with $n, p \in \mathbb{N}$, such that $n/p \to \delta > 0$ and assume that:

(A1) The components of X are iid $N(0, 1/n)$, and is independent of $\hat{\beta}^0 \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$, and $\varepsilon \in \mathbb{R}^n$.

(A2) As $p \to \infty$, the empirical distribution of the components of β and the random error ε, respectively converge to the distributions $\Pi_{\bar{\beta}}$ and $,P_{\bar{\varepsilon}}$ with finite second moments, for some random varible $\bar{\beta} \sim \Pi_{\bar{\beta}}$, and $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$.

(A3) There is a non-negative definite $\Sigma_0 \in \mathbb{R}^{2\times 2}$ such that

$$\frac{1}{n}\begin{pmatrix}\beta & \hat{\beta}^0\end{pmatrix}^T\begin{pmatrix}\beta & \hat{\beta}^0\end{pmatrix} = \frac{1}{n}\begin{pmatrix} \beta^T\beta & \beta^T\hat{\beta}^0 \\ (\hat{\beta}^0)^T\beta & (\hat{\beta}^0)^T\hat{\beta}^0 \end{pmatrix} \to \Sigma_0. \tag{2.14}$$

(A4) For each $k \in \mathbb{N}_0$, f_{k+1} is not a constant function on $\mathbb{R}$, and $\tilde{g}_k : (z, w, v) \mapsto g_k(w, h(z, v))$ is Lipschitz on $\mathbb{R}^3$ with the set $K_{z,w} := \{v : (z, w) \mapsto \tilde{g}_k(z, w, v)$ is non-constant$\}$ having nonzero measure.

(A5) For each $k \in \mathbb{N}_0$, if $\Omega_k \in \mathbb{R}^2$ denotes the set of discontinuities of g_k', then $P((Z_k, Y) \in \Omega_k) = 0$, and f_{k+1}' is almost everywhere continuous.

2.3.3 General State Evolution Recursion for GAMP

The limiting empirical distributions of the entries of the GAMP iterates can be decomposed into independent 'signal' and 'noise' components, and the effective signal strength and noise level are determined by a state evolution recursion. With Σ_0 as in (A4), the state evolution recursion can be computed as follows. for $k \in \mathbb{N}_0$,

$$\begin{aligned} \Sigma_k &= \frac{1}{\delta}\begin{pmatrix} E(\bar{\beta}^2) & E\{\bar{\beta} f_k(\mu_k\bar{\beta} + \sigma_k G_k)\} \\ E\{\bar{\beta} f_k(\mu_k\bar{\beta} + \sigma_k G_k)\} & E\{f_k(\mu_k\bar{\beta} + \sigma_k G_k)^2\} \end{pmatrix} \\ \sigma_{k+1}^2 &= E\left[\tilde{g}_k(Z, Z_k, \bar{\varepsilon})^2\right] = E[g_k(Z_k, Y)^2]; \qquad \mu_{k+1} = E[\partial_z \tilde{g}_k(Z, Z_k, \bar{\varepsilon})]. \end{aligned} \tag{2.15}$$

where $(Z, Z_k) \sim N(0, \Sigma_k)$ independent of $\bar{\varepsilon}$, $Y = h(Z, \bar{\varepsilon})$, $G_{k+1} \sim N(0, 1)$ independent of $\bar{\beta}$. An alternative expression for μ_{k+1} which will be useful in the sequel is the following: with

G_{k+1}, Z and $\bar{\varepsilon}$ distributed as above, we have $Z_{k+1} \overset{d}{=} \mu_{Z,k+1} Z + \sigma_{Z,k} G_{k+1}$, with

$$\begin{aligned}
\mu_{Z,k+1} &= \Sigma_{12}/\Sigma_{11} = \frac{E\{\bar{\beta} f_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1} G_{k+1})\}}{E(\bar{\beta}^2)} \\
\sigma^2_{Z,k+1} &= \Sigma_{22} - \Sigma^2_{12}/\Sigma_{11} \\
&= \frac{1}{\delta}\left[E\{f_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1} G_{k+1}\} - \frac{[E\{\bar{\beta} f_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1} G_{k+1}\}]^2}{E(\bar{\beta}^2)} \right]
\end{aligned}$$

we have

$$\sigma^2_{k+1} = E\left[\tilde{g}_k(Z, \mu_{Z,k} Z + \sigma_{Z,k} G_k, \bar{\varepsilon})^2\right] \tag{2.16}$$

and

$$\mu_{k+1} = E[\partial_z \tilde{g}_k(Z, Z_k, \varepsilon)] = \frac{\delta}{E(\bar{\beta}^2)} E[Z g_k(Z_k, Y)] - \mu_{Z,k} \bar{c}_k] \tag{2.17}$$

We have used the following in the above expression of μ_{k+1}.

$$c_k = \langle g'_k(\theta^k, y) \rangle \to E[g'_k(Z_k, Y)] = \bar{c}_k$$

We are now ready to state the **master theorem** for the GAMP.

Theorem 2.2 (Theorem 4.2 in Feng et al. (2021)). *Suppose assumptions (A1)-(A5) hold for a sequence of GAMP recursion (2.13) indexed by n and p, with $n/p \to \delta \in (0, \infty)$ and $\sigma_1 > 0$. Then for each $k \in \mathbb{N}_0$, it follows that for any pseudo-Lipschitz function ψ,*

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\beta_j^{k+1}, \beta_j) \to E[\psi(\mu_{k+1}\bar{\beta} + \sigma_{k+1} G_{k+1}, \bar{\beta})] \tag{2.18}$$

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\theta_i^{k}, \theta_i, \varepsilon_i) \to E[\psi(\mu_{Z,k} Z + \sigma_{Z,k} \tilde{G}_k, Z, \bar{\varepsilon})] \tag{2.19}$$

as $n, p \to \infty$ with $n/p \to \delta$ where in the above expression $\beta^{k+1} = X^T g_k(\theta^k, y) - c_k \hat{\beta}^k$, and $\theta_i \equiv \theta_i(n) = x_i^T \beta$ for $n \in \mathbb{N}$ and $1 \leq i \leq n$.

The master theorem can be interpreted as that the components of β_k when p is large has the same empirical distribution approximately as those of $\mu_k\beta + \sigma_k\epsilon$, where $\epsilon \in N(0; I_p)$ is independent of $\beta \in \mathbb{R}^p$. By comparison with the limiting univariate problem of estimating $\bar{\beta} \sim \pi_{\bar{\beta}}$ via a corrupted observation $\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}$, β_k can be seen as an effective observation and $\rho_k := (\mu_k/\sigma_k)$ as an effective signal-to-noise ratio. In the setting of Theorem 2.2, condition (A6) ensures that

$$b_{k+1} \quad = \frac{1}{\delta}\langle f'_{k+1}(\beta^{k+1})\rangle \to \frac{1}{\delta}E\left[f'_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1})\right] = \bar{b}_{k+1}. \tag{2.20}$$

As a corollary of the master theorem, it follows that since the functions f_k in (2.13) are assumed to be Lipschitz, then for any pseudo-Lipschitz loss function ψ, the **asymptotic estimation error** of $\hat{\beta}^k$ is given by

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j^{k+1}, \beta_j) \to E[\psi(f_k(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}), \bar{\beta})] \tag{2.21}$$

for each $k \in \mathbb{N}$, as $n, p \to \infty$ with $n/p \to \delta$.

2.3.4 Special Cases of GAMP

Linear Models

GAMP has been applied to linear models in various fields. It is particularly useful in scenarios where the number of variables is large and the data is sparse. One of the main advantages of using GAMP in linear models is its ability to handle large-scale problems. Traditional methods, such as least squares or maximum likelihood estimation, can become computationally expensive when dealing with high-dimensional data. GAMP, on the other hand, is designed to handle such scenarios efficiently, making it particularly suitable for applications where computational resources are limited.

Another advantage of GAMP is its ability to handle sparse data. GAMP exploits this sparsity by incorporating a sparsity-promoting prior into the estimation process. This allows GAMP to accurately estimate the non-zero coefficients while effectively shrinking the irrelevant ones towards zero.

Furthermore, GAMP is known for its robustness to noise and model misspecification. It can handle noisy measurements and still provide accurate estimates of the underlying parameters. Additionally, GAMP is flexible and can be easily adapted to different linear models, making it a versatile tool in various applications.

Overall, the application of GAMP to linear models has shown promising results in terms of computational efficiency, handling sparsity, robustness to noise, and adaptability to different models. It has the potential to significantly improve the estimation accuracy and computational efficiency in various fields, such as signal processing, communications, and machine learning.

We now derive the GAMP recursion for the standard linear model

$$y = X\beta + \varepsilon \tag{2.22}$$

obtained by setting $h(z, v) = z + v$ in (2.9). Here $\{\varepsilon_i\}_{i=1}^n \sim P_{\bar{\varepsilon}}$ have finite second moment $\sigma^2 > 0$. Taking $\hat{r}^{-1} = 0 \in \mathbb{R}^n$, $b_0 \in R$, and initializing by some $\hat{\beta}^0 \in \mathbb{R}^p$, the initial AMP algorithm of used in Donoho et al. (2009) and Bayati and Montanari (2011) can be achieved as a special case of the GAMP algorithm (2.13) by choosing $g^k(u, v) := v - u$, thus $c_k = -1$, and so the gamp becomes

$$\begin{aligned} &\theta^k = X\hat{\beta}^k - b_k\hat{r}^{k-1}, \quad \hat{r}^k = y - \theta^k, \\ &\beta^{k+1} = X^T\hat{r}^k + \hat{\beta}^k, \quad \hat{\beta}^{k+1} = f_{k+1}\left(\beta^{k+1}\right), \quad b_{k+1} = \tfrac{1}{n}\textstyle\sum_{j=1}^p f'_{k+1}\left(\beta_j^{k+1}\right), \end{aligned} \tag{2.23}$$

for $k \in \mathbb{N}_0$. Here, β^k is the effective observation at k^{th} iteration, and $\hat{r}^k$ is a corrected residual and has been shown to substantially improves the sparsity–under-sampling tradeoff..

The state evolution recursions (2.15) reduces to

$$\sigma_{k+1}^2 = \sigma^2 + \frac{E\{\left(\bar{\beta} - f_k(\bar{\beta} - \sigma_k G_k)\right)^2\}}{\delta}, \quad \sigma_1^2 = \sigma^2 + E\left[(Z - Z_0)^2\right], \text{ and } \mu_k = 1. \tag{2.24}$$

Under the same setting as the GAMP master Theorem 2.2, the asymptotic performance of $\{\hat{\beta}^k\}$, which is the main result in Bayati and Montanari (2011) is recovered as

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j^{k+1},\beta_j) \to E[\psi(f_k(\bar{\beta}+\sigma_{k+1}G_{k+1}),\bar{\beta})] \tag{2.25}$$

as $n, p \to \infty$ such that $n/p \to \delta$.

GAMP algorithms have been applied by many authors to estimate asymptotic errors of constrained and unconstrained optimization problems under various conditions on the data matrix (see e.g., Donoho and Montanari (2013), Schniter and Rangan (2012), Rangan (2011), Sur and Candès (2019), and the references contained in them).

Logistic Regression (Sur and Candès (2019))

Another advantage of GAMP is its ability to handle non-linear models. While logit models are typically linear in the parameters, GAMP can be extended to handle non-linear models by using appropriate approximations. This allows for more flexibility in modeling complex relationships between the predictors and the response variable.

In addition, GAMP provides a framework for incorporating prior knowledge or constraints into the model. This can be particularly useful in situations where there is limited data available as is common in Medicine where logistic regression is very popular, or when certain assumptions about the model parameters need to be enforced. By incorporating prior knowledge, GAMP can improve the accuracy and interpretability of the model.

A major objectives in this work is to derive a GAMP algorithm for characterizing and estimating the asymptotic MSE of the $\ell_!$-penalized logistic regression. As precursor to that, we first present the result of Sur and Candès (2019) for the unpenalized logistic regression.

Sur and Candès (2019) make the case about the failure of classical MLE results in the high dimensional setting of the logistic regression. In the large n-large-p-asymptotics regime, their theory explicitly characterizes

1. the bias of the MLE

2. the variability of the MLE, and

3. the distribution of the LRT.

Whenever the MLE asymptotically exists with probability one. The case for the existence of MLE here is made in Albert and Anderson (1984), and in the random design setting, Candès and Sur (2020) established a phase transition curve for the existence of MLE. In the setting where the MLE exists, Sur and Candès (2019) gave a GAMP scheme for estimating it.

Sur and Candès (2019) considered the vanilla logistic regression which equivalent to solving a convex optimization problem of the form (1.6) with $J \equiv 0$. The functions f proposed the following scheme: initializing with some $\hat{\beta}^0 \in \mathbb{R}^p$, and setting $U^0 = X\beta^0$, recursively define

$$\begin{aligned} \hat{\beta}^{k+1} &= \beta^k + \tfrac{1}{\kappa} X' \psi_k(y, \theta^k) \\ \theta^{k+1} &= X\beta^{k+1} - \psi_k(y, \theta^k) \end{aligned} \tag{2.26}$$

with

$$\psi_k(y, \theta) = \lambda_k r_k, \qquad r_k = y - \rho'(\mathrm{prox}_{\lambda_k \rho}(\lambda_k y + \theta)). \tag{2.27}$$

The state evolution recursion is given by the following system of nonlinear iterations: starting with initial guesses α_0, σ_0, recursively define the sequence $\{\alpha_s, \sigma_s, \lambda_s\}_{s \geq 0}$ by

$$\left\{ \begin{aligned} 1 - \kappa &= E\left[\tfrac{2\rho'(Z_1^s)}{1+\lambda\rho''(\mathrm{prox}_{\lambda\rho}(Z_2^s))}\right] \\ \alpha_{s+1} &= \alpha + \tfrac{1}{\kappa\gamma^2} E[2\rho'(Q_1^s) Q_1^s \lambda_s \rho'(\mathrm{prox}_{\lambda_s\rho}(Z_2^s))] \\ \sigma_{s+1}^2 &= \tfrac{1}{\kappa^2} E\left[2\rho'(Q_1^s)\left(\lambda_s \rho'(\mathrm{prox}_{\lambda_s\rho}(Z_2^s))\right)^2\right] \end{aligned} \right. \tag{2.28}$$

for $s \geq 1$.

They are argue that whenever the MLE exists, the system (2.28) above converges to a unique fixed point $\{\alpha_*, \sigma_*, \lambda_*\}$. Substituting the corresponding iterations by their fixed point in recursion (2.26), the prove that the following theorem.

Theorem 2.3 (Sur and Candès (2019), Theorem 2). *Assume the logistic model described above where the empirical distribution of β_j converges weakly to a distribution Π with finite second* *moment. Suppose further that the second moment converges in the sense that as* $n \to \infty$, *Ave*$_j(\beta_j^2), \to \mathrm{E}\beta^2, \beta \sim \Pi$. *Then for any pseudo-Lipschitz function* ψ *of order 2, the marginal distributions of the MLE coordinates obey*

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j - \alpha_\star\beta_j) \xrightarrow{a.s.} \mathbb{E}[\psi(\sigma_\star Z, \beta)], \qquad Z \sim \mathcal{N}(0,1) \tag{2.29}$$

where $\beta \sim \Pi$*, independent of* Z*.*

Concerning the distribution of the Likelihood-ratio statistics for testing $\beta_j = 0$, they prove the following theorem.

Theorem 2.4. *Consider the LLR* $\Lambda_j = \min_{\boldsymbol{b}:b_j=0}\ell(\boldsymbol{b}) - \min_{\boldsymbol{b}}\ell(\boldsymbol{b})$ *for testing* $\beta_j = 0$*. In the setting of Theorem 2.29, twice the LLR is asymptotically distributed as a multiple of a* χ^2 *under the null,*

$$2\lambda_j \xrightarrow{d} \frac{\kappa\sigma_\star^2}{\lambda_\star}\chi_1^2. \tag{2.30}$$

Also, the LLR for testing $\beta_{i_1} = \beta_{i_2} = ... = \beta_{i_k} = 0$ *for any finite* k *converges to the rescaled* χ^2 $(\kappa\sigma_\star^2/\lambda_\star)\,\chi_k^2$ *under the null.*

2.4 Inside the GAMP

Here, we look inside the GAMP and explain the functions of its various components. The algorithm begins by initializing $\hat{\beta}^0 = 0$, and then, at every iteration t, proceeds as follows:

1. Calculates the **residual**, $\hat{r}^k = g_k(\theta^k, y) = g_k(X\hat{\beta}^k - b_k\hat{r}^{k-1}, y)$. For example, in the case of the linear model given above, the function $g_k(\cdot,\cdot)$ is taken as $g_k(a,b) = b - a$.

2. Next is the **pseudo data**, $\beta^{k+1} = X^T\hat{r}^k - c_k\hat{\beta}^k$ which has been proved to be equal in distribution to the true β, plus additive white Gaussian noise.

3. And finally, the **denoising** step, $\hat{\beta}^{k+1} = f_{k+1}\left(\beta^{k+1}\right)$, i.e, the estimate $\hat{\beta}^{k+1}$ is a denoising function of the pseudo data.

(a) β^k

(b) $f_k\left(\beta^k\right)$

Figure 2.1: The Denoising Step

The idea of the denoising algorithm is the following: See the blurry image, β^k in Fig (2.1a), what we are observing is the true image, β with additive Gaussian noise, and when the denoising is applied, it utilizes structure that is available within the image to reduce the impact of the noise resulting in the clearer version of the image, $\hat{\beta}^k$ in Fig (2.1b). This process continues over several iterations until convergence. Notice that the denoiser function is not fixed. At every iteration it uses a slightly modified denoiser. Typically, what happens is that the amount of noise in β^k goes down over the first several iterations and evetually converges to a noise floor and the GAMP is somewhat stable. The Onsager correction terms, $-b_k\hat{r}^{k-1}$ and $c_k\hat{\beta}^k$ on the other hand results in two things: (1) the error $(\hat{\beta}^t - \beta)$ in estimating the true signal β will be uncorrelated to the signal, and this helps the denoiser work well; (2) the error will be Gaussian. What these mean is that, without the Onsager terms, after few iterations, the error will quickly be correlated with β, and the denoiser will not work well. Thus, the Onsager term increases the speed on convergence by an appreciable margin.

2.5 More Tools from the G-AMP Algorithm

In this section, result established in Javanmard and Montanari (2013) are presented and they will be principal to the analysis. For simplicity in calculation, the same notation in Javanmard and Montanari (2013) are adopted here.

A G-AMP algorithm takes the form: $\{x^t\}_{t\geq 0}$, where $x_t \in V_{q\times N} \equiv (\mathrm{R}^q)^N$, for some fixed $q \in N$, and N is a function of the sample size n. Define $A = G + G'$, where $G \in \mathrm{R}^{N\times N}$ has i.i.d. entries from $N(0, 1/2N)$.

Let $\mathcal{G} = \{f^t : t \in [N]\}$, such that $f^t : \mathrm{R}^q \times N \to \mathrm{R}^q$, is locally Lipschitz in the first argument for all $t \in [N]$. Then, with some initial condition $x^0 \in V_{q,N}$, a GAMP algorithm updates as follows:

$$x_{\bullet i}^{k+1} = \sum_{j=1}^{N} A_{ij} f^j(x_{\bullet j}^k; k) - \frac{1}{N} \left(\sum_{j=1}^{N} \frac{\partial f^j}{\partial x}(x_{\bullet j}^k; k) \right) f^i(x_{\bullet i}^{k-1}; k-1), \tag{2.31}$$

and terms with negative k-indices are taken to be 0. The notation $\frac{\partial f^j}{\partial x}$, refers to the Jacobian of $f^j(\cdot; k) : \mathrm{R}^q \to \mathrm{R}^q$.

Lemma 2.5 (Javanmard and Montanari (2013), Theorem 1)**.** *For all $t > 1$, each $a \in [q']$, and any pseudo-Lipschitz function $\psi : \mathrm{R}^q \times \mathrm{R}^q \to \mathrm{R}$ of order k, almost surely,*

$$\lim_{N \to \infty} \frac{1}{C_a^N} \sum_{j \in C_a^N} \psi(x_{\bullet j}^t, y_{\bullet j}) = \mathrm{E}\{\psi(Z_a^t, Y_a)\}, \tag{2.32}$$

where $Z_a^t \sim N(0, \Sigma^{(k)})$ is independent of $Y_a \sim P_a$.

Chapter 3

ℓ_1-Penalized Logistic Regression

The GAMP recursion idea comes as a natural choice in considering the ℓ_1-penalized logistic regression, given that the proximal gradient method is specifically designed to handle non-differentiable regularization terms like the ℓ_1 penality. Moreover the proximal operator for the ℓ_1 penalty admits a simple closed form formula which can be leveraged to achieve explicit asymptotic characterizations. The goal of the main result, is to characterise the *Asymptotic Mean Squared Error (AMSE)* of the ℓ_1-penalized logistics regression.

Following the details in Ali and Tibshirani (2018), it can be shown that (1.6) admits a unique solution provided (1.3) has a unique solution. We pursue the estimation of this solution via the application of a GAMP recursion. By showing that the GAMP estimates converge to the corresponding penalized estimators in the large system limit, we derive the asymptotic MSE of the penalized estimator by using state evolution of the corresponding GAMP estimators.

3.1 Main Results

It is now time for the presentation of the main results. However, before diving into them, it is essential to introduce the necessary components that will facilitate the clear and concise presentation of these outcomes.

Starting with some key definitions, the proximal mappings for the ℓ_1 penalty term $(J(\cdot) := |\cdot|)$ and the logistic loss function $\ell(z, y)$ are defined via

$$\text{prox}_J(u; b) = \text{S}(u; \lambda b) := \begin{cases} u - \lambda b, & u > \lambda b \\ 0, & |u| \leq \lambda b \\ u + \lambda b, & u < -\lambda b \end{cases} = \arg\min_{\beta} \left\{ |\beta| + \frac{1}{2\lambda b}(\beta - u)^2 \right\}$$

$$\text{prox}_\ell(\theta; y, b) := \arg\min_z \left\{ \ell(z, y) + \tfrac{1}{2b}(z - \theta)^2 \right\} = \text{prox}_\rho(\theta + by; b).$$

The above definitions are utilized in the fixed point equations that follow. This particular system of equations, which involves four variables $\{c_*, \sigma_*^2, \mu_*, b_*\}$, will soon become evident as the governing factor for the asymptotic behavior of the GAMP estimate.

$$\begin{cases} c_* &= \dfrac{1}{b_*} E\left[\dfrac{1}{1 + b_* \rho''(\text{prox}_\rho(Z_* + b_* Y; b_*))} - 1 \right] \\ \sigma_*^2 &= E\left[(Y - \rho'(\text{prox}_\rho(Z_* + b_* Y; b_*)))^2 \right] \\ \mu_* &= \dfrac{\delta}{E(\bar{\beta}^2)} E[Z\{Y - \rho'(\text{prox}_\rho(Z_* + b_* Y; b_*))\}] - \mu_{Z,*} c_*. \\ b_* &= -\frac{1}{\delta c_*} P(|\mu_* \bar{\beta} + \sigma_* G_*| \geq \lambda). \end{cases} \tag{3.1}$$

where (Z, Z_*) is bivariate Gaussian with mean $\mathbf{0}$, and covariance given by

$$\frac{1}{\delta} \begin{pmatrix} E(\bar{\beta}^2) & E[\bar{\beta}\text{S}(\mu_* \bar{\beta} + \sigma_* G_*)] \\ E[\bar{\beta}\text{S}(\mu_* \bar{\beta} + \sigma_* G_*)] & E[\text{S}(\mu_* \bar{\beta} + \sigma_* G_*)^2] \end{pmatrix}$$

with $G_* \sim N(0, 1)$.

The primary outcome of this study describes the asymptotic average behavior of the ℓ_1-penalized logistic regression estimator, and is present next.

Theorem 3.1. *Suppose δ and Π are such that (1.6) admits a unique solution and let $(c_*, \mu_*, \sigma_*, b_*)$ come from the system (3.1). Then under assumptions (A1) - (A5) above, for any pseudo-Lipschitz function ψ of order 2, it follows that*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi\left(\hat{\beta}_j, \beta_j\right) \overset{a.s.}{=} E\left[\psi\left(-\frac{1}{c_*} S(\mu_* \beta + \sigma_* G, \lambda), \bar{\beta}\right)\right], \tag{3.2}$$

with $G \sim N(0,1)$, and $\bar{\beta} \sim \Pi_{\bar{\beta}}$, independent of G.

Informally, an interpretation of Theorem 3.1 is that as p diverges, the components of $\hat{\beta}^k$ follow approximately the same empirical distribution as those of $-\frac{1}{c_*}\mathrm{S}(\mu_*\bar{\beta} + \sigma_* G, \lambda)$ with $G \sim N(0, 1_p)$ is independent of $\bar{\beta}$.

Theorem 3.1 leads to several important corollaries that are worth noting. For example, Corollary 3.2, below states that it is possible to calculate the exact Asymptotic Mean Squared Error (AM SE) of the GAMP estimate for the ℓ_1-penalized logistic regression. This corollary provides valuable information about the accuracy of the estimation method and allows for the quantification of error in predictions. By understanding the $AMSE$, one can assess the reliability and performance of the GAMP estimate, enabling informed decision-making based on the level of uncertainty in the results.

Corollary 3.2. *Under the assumptions of Theorem 3.1, it follows by setting $\psi(u, v) = |u - v|^2$, that*

$$AMSE = \lim_{p\to\infty} \frac{1}{p} \|\hat{\beta} - \beta\|_2^2 = E\left(\left\|\frac{S(\mu_*\bar{\beta} + \sigma_* G, \lambda)}{-c_*} - \bar{\beta}\right\|_2^2\right).$$

Next is Corollary 3.3 which offers a method to accurately compute the Asymptotic Selection Error Rate $(ASER)$ of the estimator. This means that it is possible to determine the rate at which the estimator makes selection errors as the sample size gets larger and larger.

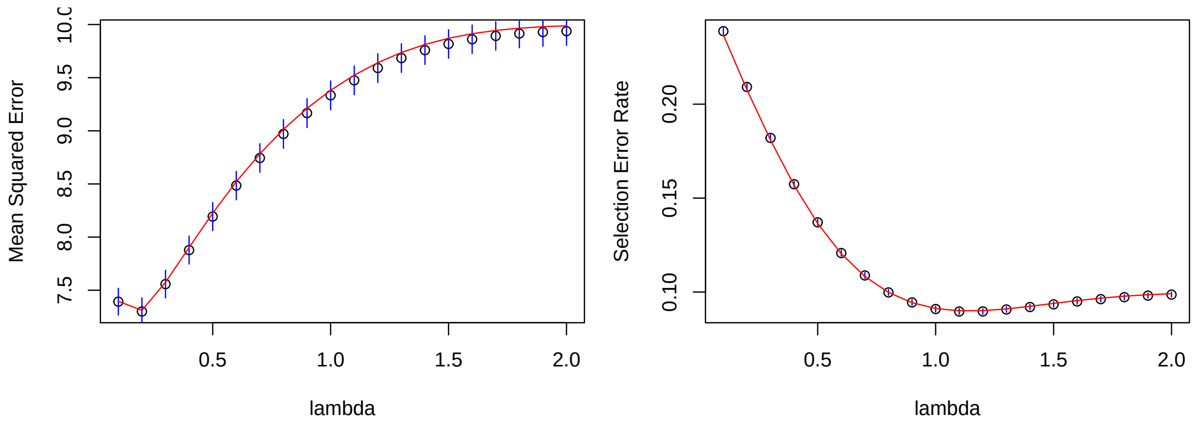


Figure 3.1: Comparison between theoretical estimation and simulation study

Corollary 3.3. *Under the assumptions of Theorem 3.1, setting* $\psi\left(u,v\right) = \boldsymbol{I}\Big(\boldsymbol{I}(u=0) \neq \boldsymbol{I}(v=0)\Big)$*, it follows that*

$$\begin{aligned} ASER &= \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \boldsymbol{I}\Big(\boldsymbol{I}(\beta_j = 0) \neq \boldsymbol{I}(\beta_j = 0)\Big) \\ &= \Big[P\left(\bar{\beta} = 0\right) \cdot FPR\Big] + \Big[P\left(\bar{\beta} \neq 0\right) \cdot (1 - TPR)\Big]. \end{aligned}$$

where $FPR = P(\sigma_*|G| > \lambda)$ *and* $TPR = P(|\mu_*\bar{\beta} + \sigma_* G| > \lambda \mid \bar{\beta} \neq 0)$.

Finally, it is note worthy to realize that by setting $\lambda = 0$, the objective function reduces to the maximum likelihood estimator for logistic regression, in which case, when ever solution exisits, the results of Sur and Candès (2019) are recovered as a special case, i.e.,

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi\left(\hat{\beta}_j - \mu_*\beta_j, \beta_j\right) \overset{a.s.}{=} E\left[\psi\left(\sigma_* G, \bar{\beta}\right)\right], \text{ with } G \sim N(0,1). \tag{3.3}$$

3.2 Simulation Results

In this section, a Monte Carlo simulation is conducted to compare the theoretical predictions with simulated results in finite samples. The reliability of the analytical result, as stated in Theorem 3.1, is confirmed by comparing the MSE calculated using a numerical algorithm with the estimated MSE from theory.

For each setting, the scalars $(c_*, \mu_*, \sigma_*, b_*)$, are estimated from (3.1). Then, the theoretical MSE is obtained using the right hand side of equation (3.2) in Theorem 3.1. The undersampling and sparsity parameters are fixed as $\delta = 0.5$, and $\epsilon = 0.1$. The true signal β is taken to follow a 3-point distribution $\pi_{\bar{\beta}} \sim (1-\epsilon)\delta_0 + \frac{\epsilon}{2}\delta_{-\mu} + \frac{\epsilon}{2}\delta_{\mu}$, with $\mu = 10$. The components of $X \sim N(0, 1/n)$, and $y \sim$ Bernoulli(π_i) with $\pi_i; = 1/(1+e^{-\beta^T x_i})$. The change of MSE versus tuning parameter is plotted. The simulated data has dimension $p = 1000$, and simulation is repeated 200 times for each parameter setting. The R package *glmnet* is used to fit the ℓ_1-logistic regression estimator. The left plot in Figure 3.1 shows the average MSE with 95% confidence intervals based on 200 replicates. The curve is the asymptotic MSE calculated based on the main result. The right plot shows the average SER with 95% confidence intervals based on 200 replicates. The curve is the asymptotic SER calculated based on the main result. Both plots indicate that the theoretical quantities match the empirical values closely.

3.3 Formulation of Algorithm

This section provides a comprehensive overview of the algorithm's development process and gathers essential tools and significant findings that will be necessary in the subsequent sections. It serves as a foundation for understanding the subsequent discussions and analyses.

The following properties, respectively of $\rho'(u)$ and $\text{prox}_\rho(u; b)$ can be verified by their definitions and will be utilized in the later stages of our recursion development.

$$\rho'(-u) = 1 - \rho'(u), \qquad \text{prox}_\rho(u + b; b) = -\text{prox}_\rho(-u; b).$$

The above two identities hold only for $\rho(u) = \log(1 + e^u)$, while the following one holds for any ρ.

$$\text{prox}_\rho(u; b) = u - b\rho'(\text{prox}_\rho(u; b)). \tag{3.4}$$

Now, using the proximal mappings that were defined earlier, we start building a GAMP recursion in the form of equation (2.13). The functions f_{k+1} and g_k are given as follow.

$$\begin{aligned} g_k(u,v) &= \frac{\text{prox}_\ell(u;v,b_k)-u}{b_k} = \frac{\text{prox}_\rho(u+b_kv;b_k)-u}{b_k} = v-\rho'(\text{prox}_\rho(u+b_kv;b_k)) \\ f_{k+1}(w) &= \text{prox}_J\left(-\frac{w}{c_k};-\frac{1}{c_k}\right) = \text{S}\left(-\frac{w}{c_k};-\frac{\lambda}{c_k}\right) = -\frac{\text{S}(w;\lambda)}{c_k} \end{aligned} \tag{3.5}$$

The derivatives are thus given as:

$$\begin{aligned} g_k'(u,v) &= \frac{\text{prox}_\ell'(u;v,b_k)-1}{b_k} = \frac{\text{prox}_\rho'(u+b_kv,b_k)-1}{b_k} = \frac{-\rho''(z_*)}{1+b_k\rho''(z_*)} \\ f_{k+1}'(w) &= -\tfrac{1}{c_k}I(|w|\geq\lambda) = \begin{cases} -1/c_k, & |w|\geq\lambda \\ 0, & |w|<\lambda \end{cases} \end{aligned} \tag{3.6}$$

where $z_* = \text{prox}_\rho(u+by;b)$ and the following relationship is used.

$$z_* = u + b(y-\rho'(z_*)), \qquad \frac{\partial z_*}{\partial u} = \frac{1}{1+b\rho''(z_*)}$$

Continuing with the progression, the final GAMP recursion (3.8) will now be unveiled, and the explanation of its derivation will be postponed for later. Define the following function:

$$\Psi(u,y) = -b\partial_1\ell(\text{prox}_\rho(u+by;b),y). \tag{3.7}$$

where ∂_1 represents the derivative with respect to the first argument of the function. From 3.5, we have

$$\begin{aligned} g_k(\theta_k,y) = y-\rho'(\text{prox}_\rho(\theta_k+b_ky;b_k)) &= -\partial_1\ell(\text{prox}_\rho(\theta_k+b_ky;b_k),y) \\ \Rightarrow b_kg_k(\theta_k,y) &= \Psi(\theta_k,y). \end{aligned}$$

Let $\{b_k,c_k\}$ be the two sequences of negative and non-negative parameters respectively given in (3.9). Starting with the initial condition $\hat{\beta}^0 = 0 \in \text{R}^p,\ b_0 = 1$ and $\Psi(\theta^{-1},b_{-1}) = 0 \in \text{R}^n$, then, the final GAMP recursion is the following:

$$\begin{cases} \theta^k &= \quad X\hat{\beta}^k - \dfrac{b_k}{b_{k-1}}\Psi(\theta_{k-1}, y) \\ \hat{\beta}^{k+1} &= \mathrm{S}\left(\hat{\beta}^k - \dfrac{1}{c_k b_k} X^T \Psi(\theta_k, y), -\dfrac{\lambda}{c_k}\right). \end{cases} \tag{3.8}$$

where,

$$c_k = \left\langle \frac{-\rho''(\mathrm{prox}_\rho(\theta^k + b_k y; b_k))}{1 + b_k \rho''(\mathrm{prox}_\rho(\theta^k + b_k y; b_k))} \right\rangle, \qquad b_{k+1} = -\frac{1}{\delta c_k}\left\langle I\left(\left|\hat{\beta}^{k+1}\right| \geq \lambda\right)\right\rangle \tag{3.9}$$

The GAMP recursion (3.8) mentioned above is a condensed version of the recursion (3.10) described below. This derivation follows the methodology and terminology outlined in Feng et al. (2021). Here are the specific details: initialize $\hat{r}^{-1} = 0$, $\hat{\beta}^0 = 0$, and $b_0 = 1$, define $\{\hat{\beta}^{k+1}, \beta^{k+1}, \theta^k, \hat{r}^k, b_{k+1}, c_k\}$ via

$$\begin{cases} \theta^k &= X\hat{\beta}^k - b_k \hat{r}^{k-1} \\ \hat{r}^k &= \dfrac{\mathrm{prox}_\rho(\theta^k + b_k y; b_k) - \theta^k}{b_k} = y - \rho'(\mathrm{prox}_\rho(\theta^k + b_k y; b_k)) \\ c_k &= \left\langle \dfrac{-\rho''(\mathrm{prox}_\rho(\theta^k + b_k y; b_k))}{1 + b_k \rho''(\mathrm{prox}_\rho(\theta^k + b_k y; b_k))} \right\rangle \\ \beta^{k+1} &= X^T \hat{r}^k - c_k \hat{\beta}^k \\ \hat{\beta}^{k+1} &= -\frac{1}{c_k}\mathrm{S}\left(\beta^{k+1}; \lambda\right) \\ b_{k+1} &= -\frac{1}{\delta c_k}\langle I(|\beta^{k+1}| \geq \lambda)\rangle \end{cases} \tag{3.10}$$

the sequence of functions f_{k+1} and g_k given in (3.5) have been applied in the general GAMP recursion (2.13) to arrive at the above recursion.

3.3.1 State Evolution Recursion

As mentioned earlier, the system (3.1) of scalar equations that describe the asymptotic behavior of the GAMP estimate is obtained from the SE recursion associated with (3.8). The derivation of the SE recursion will now be detailed out.

First, define the following notations.

- $h(z, \varepsilon) = I(\varepsilon \leq \rho'(z))$, where $I(\cdot)$ is the indicator function.

- $\tilde{g}_k(z, u, v) = g_k(u, h(z, v)) = h(z, v) - \rho'(\text{prox}_\rho(u + b_k h(z, v); b_k))$

- $(Z, Z_k) \sim N_2(0, \Sigma_k)$ and Z_k has the same distribution as $\mu_{Z,k} Z + \sigma_{Z,k} \tilde{G}_k$ for an independent $\tilde{G}_k \sim N(0, 1)$.

- $Y = h(Z, \varepsilon)$.

With Σ_0 given as in assumption (A4), the associated state evolution recursion for GAMP algorithm can be computed following (2.15) where the sequences σ_{k+1} and μ_{k+1} are derived as follow.

$$\begin{aligned}
\mu_{Z,k+1} &= \frac{\Sigma_{12}}{\Sigma_{11}} = \frac{E\{\bar{\beta}\text{S}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}; \lambda)\}}{-c_k E(\bar{\beta}^2)} \\
\sigma^2_{Z,k+1} &= \Sigma_{22} - \frac{\Sigma^2_{12}}{\Sigma_{11}} \\
&= \frac{1}{\delta c_k^2}\left[E\{\text{S}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}; \lambda)^2\} - \frac{[E\{\bar{\beta}\text{S}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}; \lambda)\}]^2}{E(\bar{\beta}^2)}\right]
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2_{k+1} &= E\left[(Y - \rho'(\text{prox}_\rho(Z_k + b_k Y; b_k)))^2\right] \\
\mu_{k+1} &= E[\partial_z \tilde{g}_k(Z, Z_k, \varepsilon)] = \frac{\delta}{E(\bar{\beta}^2)} E[Z g_k(Z_k, Y)] - \mu_{Z,k} E[g'_k(Z_k, Y)] \\
&= \frac{\delta}{E(\bar{\beta}^2)} E[Z\{Y - \rho'(\text{prox}_\rho(Z_k + b_k Y; b_k))\}] - \mu_{Z,k} c_k.
\end{aligned}$$

The following expression has been used for c_k in the above expression of μ_{k+1}.

$$c_k = \langle g'_k(\theta^k, y)\rangle \to E[g'_k(Z_k, Y)] = E\left[\frac{-\rho''(\text{prox}_\rho(Z_k + b_k Y; b_k))}{1 + b_k \rho''(\text{prox}_\rho(Z_k + b_k Y; b_k))}\right]$$

Similarly, the following expression has been applied b_{k+1}.

$$b_{k+1} = \tfrac{1}{\delta}\langle f'_{k+1}(\beta^{k+1})\rangle \to \tfrac{1}{\delta} E[f'_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1})] = -\tfrac{1}{\delta c_k} P(|\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}| \geq \lambda).$$

Finally, the SE recursion corresponding to equation (3.8) can be summarized as follows. Initializing $Z_0 = 0$ and $\mu_{Z,0} = 0$. With $b_0 = 1$,

$$\begin{cases} c_k & = E\left[\dfrac{-\rho''(\text{prox}_\rho(Z_k + b_kY; b_k))}{1 + b_k\rho''(\text{prox}_\rho(Z_k + b_kY; b_k))}\right] = \dfrac{1}{b_k}E\left[\dfrac{1}{1 + b_k\rho''(\text{prox}_\rho(Z_k + b_kY; b_k))} - 1\right] \\ \sigma_{k+1}^2 & = E\left[(Y - \rho'(\text{prox}_\rho(Z_k + b_kY; b_k)))^2\right] \\ \mu_{k+1} & = \frac{\delta}{E(\beta^2)}E[Z\{Y - \rho'(\text{prox}_\rho(Z_k + b_kY; b_k))\}] - \mu_{Z,k}c_k. \\ b_{k+1} & = -\frac{1}{\delta c_k}P(|\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}| \geq \lambda). \end{cases} \tag{3.11}$$

The sequences, $\{c_k\}$ and $\{b_k\}$ are essentially the limits of the corresponding c_k and b_k in the above algorithm.

3.3.2 Computation

To demonstrate how to evaluate the expectations used in the SE equations (3.11) above, suppose you need to evaluate an expectation of the form $E[m(Z_k, Y)]$ where $m : \mathbb{R}^2 \to \mathbb{R}$. Notice that $Y = I(\varepsilon < \rho'(Z))$, $Z_k = \mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}$, and ε, Z, and $\tilde{G}$ are mutually independent.

$$\begin{aligned} E[m(Z_k, Y)] &= E[m(Z_k, 1)\rho'(Z) + m(Z_k, 0)(1 - \rho'(Z))] \\ &= E[m(\mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}, 1)\rho'(Z) + m(\mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}, 0)(1 - \rho'(Z))] \end{aligned}$$

Therefore,

$$\begin{aligned} & E[m(\text{prox}_\rho(Z_k + b_kY; b_k))] \\ &= E[m(\text{prox}_\rho(Z_k + b_k; b_k))\rho'(Z)] + E[m(\text{prox}_\rho(Z_k; b_k))(1 - \rho'(Z))] \\ &= E[m(-\text{prox}_\rho(-Z_k; b_k))\rho'(Z)] + E[m(\text{prox}_\rho(Z_k; b_k))\rho'(-Z)] \\ &= E[m(-\text{prox}_\rho(Z_k; b_k))\rho'(-Z)] + E[m(\text{prox}_\rho(Z_k; b_k))\rho'(-Z)] \end{aligned}$$

If $m(\cdot)$ is an even function, then

$$E[m(\text{prox}_\rho(Z_k + b_kY; b_k))] \quad = 2\,E[m(\text{prox}_\rho(Z_k; b_k))\rho'(-Z)]$$

3.4 The Vanilla Logistic Regression (Simplification when $\lambda = 0$)

It is important to recall that when $\lambda = 0$, the problem simplifies to the vanilla logistic regression discussed in Sur and Candès (2019). In this scenario, the recovery of an equivalent GAMP algorithm similar to the one utilized in Sur and Candès (2019) is expected. This demonstrated next.

When $\lambda = 0$, recursion (3.10) reduces to the following. Initial value $\hat{r}^{-1} = 0$, $\hat{\beta}^0 = 0$, and $b_0 \in \mathrm{R}$.

$$\begin{aligned}
\theta^k &= X\hat{\beta}^k - b_k\hat{r}^{k-1}, & c_k &= E\left[\frac{-\rho''(\operatorname{prox}_\rho(\theta^k + b_k y; b_k))}{1 + b_k\rho''(\operatorname{prox}_\rho(\theta^k + b_k y; b_k))}\right], \\
\beta^{k+1} &= X^T\hat{r}^k - c_k\hat{\beta}^k, & \hat{r}^k &== y - \rho'(\operatorname{prox}_\rho(\theta^k + b_k y; b_k)) \\
\hat{\beta}^{k+1} &= -\frac{1}{c_k}\beta^{k+1}, & b_{k+1} &= -\frac{1}{\delta c_k}
\end{aligned}$$

which simplifies to yield

$$\begin{cases}
\theta^k &= X\hat{\beta}^k - b_k\{y - \rho'(\operatorname{prox}_\rho(\theta^{k-1} + b_{k-1}y; b_{k-1}))\} \\
\hat{\beta}^{k+1} &= \hat{\beta}^k + \delta b_{k+1}X^T\{y - \rho'(\operatorname{prox}_\rho(\theta^k + b_k y; b_k))\}
\end{cases} \tag{3.12}$$

Compared with (117) in Feng et al. (2021), there exists a slight difference. In (117) of Feng et al. (2021), $\hat{\beta}^k$ has a factor b_{k+1}/b_k. Further, the corresponding SE recursion is given as follow. Set $Z_0 = 0$ and $\mu_{Z,0} = 0$, then,

$$\begin{cases}
\sigma^2_{k+1} &= E\left[(Y - \rho'(\operatorname{prox}_\rho(Z_k + b_k Y; b_k)))^2\right] \\
\mu_{k+1} &= \frac{\delta}{E(\bar{\beta}^2)}E[Z\{Y - \rho'(\operatorname{prox}_\rho(Z_k + b_k Y; b_k))\}] + \frac{\mu_k b_k}{b_{k+1}}, \\
b_{k+1} &= -\frac{1}{\delta c_k} = \frac{b_k}{\delta}\left\{1 - E\left[\frac{1}{1 + b_k\rho''(\operatorname{prox}_\rho(Z_k + b_k Y; b_k))}\right]\right\}^{-1},
\end{cases} \tag{3.13}$$

with $\mu_{Z,k+1} = \delta\mu_{k+1}b_{k+1}$, $\sigma^2_{Z,k+1} = \delta b^2_{k+1}\sigma^2_{k+1}$, and Z_{k+1} follows the same distribution as $\mu_{Z,k+1}Z + \sigma_{Z,k+1}\tilde{G}$. $Z \sim N(0, \delta^{-1}E(\bar{\beta}^2))$.

3.4.1 Comparison with Sur and Candès (2019)

The following table lists the correspondence of notations.

Sur and Candès (2019)	Manuscript	Description
κ	$1/\delta$	p/n
γ^2	$E(\bar{\beta}^2)/\delta$	signal strength
Q_1, Q_2	Z, Z_k	random variables in SE
λ_k	b_k	
α_k	$\delta b_k \mu_k$	
σ_k	$\delta b_k \sigma_k$	

The expectations are evaluated as follows.

$$\begin{aligned}
E\left[\frac{1}{1+b_k\rho''(\mathrm{prox}_\rho(Z_k+b_kY;b_k))}\right] &= E\left[\frac{2\rho'(-Z)}{1+b_k\rho''(\mathrm{prox}_\rho(Z_k;b_k))}\right] \\
E\left[Z\{Y-\rho'(\mathrm{prox}_\rho(Z_k+b_kY;b_k))\}\right] &= E\left[2(-Z)\rho'(-Z)\rho'(\mathrm{prox}_\rho(Z_k;b_k))\right] \\
E\left[(Y-\rho'(\mathrm{prox}_\rho(Z_k+b_kY;b_k)))^2\right] &= E\left[2\rho'(-Z)(\rho'(\mathrm{prox}_\rho(Z_k;b_k)))^2\right]
\end{aligned}$$

Using the above computation, the state evaluation recursion can be rewritten in the style of Sur and Candès (2019) as follows:

$$\begin{cases}
\sigma_{k+1}^2 &= E\left[2\rho'(-Z)(\rho'(\mathrm{prox}_\rho(Z_k;b_k)))^2\right] \\
\delta b_{k+1}\mu_{k+1} &= \dfrac{\delta^2 b_{k+1}}{E(\bar{\beta}^2)}E\left[2(-Z)\rho'(-Z)\rho'(\mathrm{prox}_\rho(Z_k;b_k))\right]+\delta\mu_k b_k \\
b_{k+1} &= \dfrac{b_k}{\delta}\left\{1-E\left[\dfrac{2\rho'(-Z)}{1+b_k\rho''(\mathrm{prox}_\rho(Z_k;b_k))}\right]\right\}^{-1}
\end{cases} \tag{3.14}$$

where (Z, Z_k) follows normal with mean $\mathbf{0}$ and variance

$$\frac{1}{\delta}\begin{pmatrix} E(\bar{\beta}^2) & \delta b_k\mu_k E(\bar{\beta}^2) \\ \delta b_k\mu_k E(\bar{\beta}^2) & \delta^2 b_k^2\mu_k^2 E(\bar{\beta}^2)+\delta^2 b_k^2\sigma_k^2 \end{pmatrix}$$

The above expressions look similar to those in Sur and Candès (2019). However, it mixed b_k and b_{k+1}, which hopefully has no effect asymptotically. Essentially, Sur and Candès (2019) assumes $b_k = b_{k+1}$ and solves b_k from an equation.

3.5 Technical Proofs

3.5.1 Idea of Proof

In general, the GAMP estimation procedure for M-estimation involves three major steps. In the context of the problem under consideration, they correspond to:

Step 1. Find fixed points $\{\theta, \hat{\beta}, c_*, \sigma_*^2, \mu_*, b_*\}$ of the GAMP recursion (3.8) together with the corresponding state evolution equations (3.11) satisfying

$$\begin{cases} \theta &= X\hat{\beta} - \Psi(\theta, y) \\ \hat{\beta} &= \mathrm{S}\left(\hat{\beta} - \frac{1}{c_* b_*} X^T \Psi(\theta, y); -\frac{\lambda}{c_*}\right). \end{cases} \tag{3.15}$$

and (3.1) respectively.

Step 2. If Step 1 succeeds, then consider the following stationary version of (3.8).

$$\begin{cases} \theta^k &= X\hat{\beta}^k - \Psi(\theta^{k-1}, y) \\ \hat{\beta}^{k+1} &= \mathrm{S}\left(\hat{\beta}^k - \frac{1}{c_* b_*} X^T \Psi(\theta^k, y); -\frac{\lambda}{c_*}\right). \end{cases} \tag{3.16}$$

It is important to note here that recursion (3.16) above is used only as a theoretical device to facilitate proof rather than as a practical algorithm. Thus, for $\lambda > 0$, it is affordable to initialise (3.16) with $\theta^0 = X\hat{\beta}^0$ and $\hat{\beta}^0 = -1/c_*\mathrm{S}(\mu_*\beta + \sigma_*\xi; \lambda)$, the oracle initialiser, with $\xi = (\xi_1, ..., \xi_p)$ such that $\xi_i \sim N(0,1),\ 1 \leq i \leq p$ taken to be independent of the true signal $\beta \in \mathbb{R}^p$.

Step 3. The final step is to show that the iterates (3.16) converge to a fixed point $\hat{\beta} \equiv \hat{\beta}_{\ell_1}$ satisfying (3.15).

3.5.2 State Evolution Analysis

In this section, the asymptotic average behavior of the GAMP iterates $(\hat{\beta}, \theta)$ is characterized. The connection of the GAMP recursions (3.8) and (3.16) to the ℓ_1-penalized logistic regression estimator (1.6) is formalized by the following proposition.

Proposition 3.4. *Let $(\hat{\beta}, \theta)$ be a fixed point of the algorithm (3.16) above, then $\hat{\beta}$ is a minimum of the objective function (1.6).*

Proof. The fixed point condition for recursion (3.16) yields (3.15). The second equation in (3.15) yields that there exists $\beta' \in \partial \|\beta\|_1$ such that

$$\beta - \frac{\lambda}{c_*}\beta' = \beta - \frac{1}{c_* b_*} X^T \Psi(\theta, y)$$

which then gives

$$\lambda \beta' = \frac{1}{b_*} X^T \Psi(\theta, y) = -X^T \partial_1 \ell(\text{prox}_\rho(\theta + b_* y; b_*), y). \tag{3.17}$$

Using the first equation in (3.15), it follows that

$$\begin{aligned} \theta &= X\beta - b_* g_*(\theta, y) = X\beta - \text{prox}_\rho\left(\theta + b_* y; b_*\right) + \theta \\ \Rightarrow X\beta &= \text{prox}_\rho\left(\theta + b_* y; b_*\right). \end{aligned}$$

Plugging this into 3.17 then yields

$$\lambda \beta' = -X^T \partial_1 \ell(X\beta, y),$$

which corresponds to the stationary condition of the regularized estimator (1.6). □

As a consequence of this proposition, there is a guarantee that whenever the estimates $\{\hat{\beta}^k\}_{k\geq 1}$ based on the recursions (3.8 and (3.16) converge, the limit is the solution of the ℓ_1-penalized logistic regression objective function for a fixed choice of λ.

3.5.3 Proof of Main Result

In this section, the main theorem of this project is stated and proved. In simple terms, the theorem says that for each $k > 0$, the empirical distribution of the estimator $\hat{\beta}^k$ from GAMP (3.8) converges to the distribution of $ST(\mu_k\bar{\beta} + \sigma_k G_k, \lambda_*)$, with $G_k \sim N(0,1)$ independent of $\bar{\beta}$.

Theorem 3.5. *Suppose the initial conditions for the GAMP iterative scheme 3.16, and the variance map updates 3.11 satisfy*

$$\mu_0 = \frac{\delta}{E(\bar{\beta}^2)} \lim_{n\to\infty} \frac{\left\langle -\frac{\hat{\beta}^0}{c_*}, \bar{\beta} \right\rangle}{n}, \qquad \sigma_0 = \lim_{n,p\to\infty} \frac{1}{p} \left\| -\frac{\hat{\beta}^0}{c_*} - \mu_0\bar{\beta} \right\|^2, \tag{3.18}$$

then, for any pseudo-Lipschitz function ψ of order 2,

$$\lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi\left(\hat{\beta}_j^k, \beta_j\right) \overset{a.s.}{=} E\left[\psi\left(S(\mu_k\bar{\beta} + \sigma_k G_k), \bar{\beta}\right)\right], \tag{3.19}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(\begin{bmatrix} X_i'\beta \\ \theta_i^k \end{bmatrix}, \begin{bmatrix} \varepsilon_i \\ 0 \end{bmatrix} \right) \overset{a.s.}{=} E\left[\psi\left(\begin{bmatrix} Z \\ Z_k \end{bmatrix}, \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix} \right) \right] \tag{3.20}$$

Proof. Let

$$\begin{aligned} \tilde{\beta}^{k+1} &= -\frac{\beta^{k+1}}{c_*} = \hat{\beta}^k - \frac{1}{b_*c_*} X^T \Psi(\theta^k, y) \\ &= \mathrm{S}(\tilde{\beta}^k, \lambda_*) - \frac{1}{b_*c_*} X^T \Psi(\theta^k, y). \end{aligned}$$

Consider the new sequence $\{v^k, W^k\}$, defined by, $v^0 = \tilde{\beta}^0 - \mu_0\beta$, $W^0 = \theta^0$, and

$$\begin{cases} v^{k+1} &= \mathrm{S}\left[q_k(v^k + \mu_k\beta), \lambda_*\right] - a_{k+1}\beta - \dfrac{1}{b_*c_*} X^T \Psi(W^k, y) \\ W^k &= XST\left(v^k + \mu_k\beta, \lambda_*\right) - \Psi(W^{k-1}, y). \end{cases} \tag{3.21}$$

where

$$\begin{aligned} q_k &= -\frac{\delta}{n}\sum_{i=1}^{n}\Psi'(W_i^k, y_i) \\ a_0 = \mu_0,\ \ a_{k+1} &= \frac{\delta}{n}\sum_{i=1}^{n}\frac{\partial}{\partial a}\Psi'(W_i^k, h(a,\varepsilon_i))\Bigg|_{a=X_i'\beta}, \qquad k \geq 1. \end{aligned} \tag{3.22}$$

Here, Ψ' is the derivative with respect to the first coordinate of Ψ. This recursion differs from 3.16 only by the introduction of the new variables $\{q_k, a_k\}$, plus the regression coefficients β. It turns out that the recursive equations $\{v^k, W^k\}$ fall under the class of G-AMP algorithms. Hence, asymptotic average behavior of $\{v^k, W^k\}$ can be established by appropriately applying Theorem 2.5. This leads to the following lemma.

Lemma 3.6. *Under the assumptions of Theorem 3.5 above, the recursion $\{v^k, W^k\}$, given in (3.21) satisfy: for any $k \geq 1$*

$$\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(v_j^k, \beta_j\right) \overset{a.s.}{=} E\left[\psi\left(\sigma_k G_k, \bar{\beta}\right)\right]$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\psi\left(\begin{bmatrix} X_i'\beta \\ W_i^k \end{bmatrix}, \begin{bmatrix} \varepsilon_i \\ 0 \end{bmatrix}\right) \overset{a.s.}{=} E\left[\psi\left(\begin{bmatrix} Z \\ Z_k \end{bmatrix}, \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}\right)\right]$$

claim: The recursion (3.21) above reduce to the G-AMP form (2.31).

To see, this, fix $q = 2K^0 + 1$ for some arbitrary large $k^0 \in$ N, and let $N = n + p$. Restricting $t \in \{0, \ldots, q\}$, define $x^t \in V_{q,N}$ such that $x^0 = 0$ and for $1 \leq t \leq q$ it follows that: for even iterates $t = 2m,\ m \geq 0$, for each $i = n+1, \cdots, n+p$, define

$$x_{\bullet i}^t := \left[0, v_{i-n}^1, 0, v_{i-n}^2, 0, v_{i-n}^3, ..., v_{i-n}^{\frac{t}{2}}, 0, 0, ...\right]'. \tag{3.23}$$

For odd iterates $t = 2m+1,\ m \geq 0$, for each $i = 1, \cdots, n$, define

$$x_{\bullet i}^t := \left[Z_i, 0, W_i^0, 0, W_i^1, ..., 0, W_i^{\frac{t-1}{2}}, 0, 0, ...\right]', \tag{3.24}$$

and all other entries of x^t are 0. Let $U \in V_{q,N}$ be defined via

$$\begin{bmatrix} U_{1\bullet} \\ \\ U_{2\bullet} \end{bmatrix} = \begin{bmatrix} \varepsilon_1, & \varepsilon_2, & ..., & \varepsilon_n, & \beta_1, & \beta_2, & ..., & \beta_p \\ \\ 0, & 0, & ..., & 0, & v_1^0, & v_2^0, & ..., & v_p^0 \end{bmatrix} \tag{3.25}$$

and the other entries are all 0. Further, for even iterates $t = 2m,\ m \geq 0$, let $f^i(x; 2m) = 0$ for $i = 1, ..., n$. Let $r = \sqrt{N/n}$. For $i = n+1, ..., n+p$, define

$$\begin{aligned} f^i(x; 2m) \ &= r\, [ST(U_{1i}, \lambda_*), 0, ST(U_{2i} + \mu_0 U_{1i}, \lambda_*), 0, ST(x_2 + \mu_1 U_{1i}, \lambda_*), 0, \\ &\quad ST(x_4 + \mu_2 U_{1i}, \lambda_*), 0, ..., ST(x_t + \mu_{t/2} U_{1i}, \lambda_*), 0, 0, ...]'. \end{aligned} \tag{3.26}$$

For the odd iterates $t = 2m+1,\ m \geq 0$, let $f^i(x; 2m+1) = 0$ for $i = n+1, ..., n+p$, and for $i = 1, ..., n$, define

$$\begin{aligned} f^i(x; 2m+1) \ &= \ \delta r\, [0, \Psi_0(x_3, h(x_1, U_{1i})), 0, \Psi_1(x_5, h(x_1, U_{1i})), ..., \\ &\quad \Psi_{\frac{t-1}{2}}(x_{t+2}, h(x_1, U_{1i})), 0, 0, ...\Big]'. \end{aligned} \tag{3.27}$$

Let $\boldsymbol{A} \in \mathrm{R}^{N \times N}$ be a symmetric matrix such that $A_{ii} = 0,\ A_{i,j} = \frac{1}{r} X_{i,j-n}$ for $1 \leq i \leq n$ and $n+1 \leq j \leq n+p$ and the other entries, $A_{ij},\ i < j$ are i.i.d $N(0, 1/N)$. Using these definitions, the following result is established.

Lemma 3.7. *For odd terms of the sequence with column indices $i = 1, ..., n$, and even terms of the sequence with column indices $i = n+1, ..., n+p$, $x^t_\bullet$ defined in (3.23)-(3.24) satisfies the recursion (2.31), with the collection of functions $f^i(\cdot; t)$ given by (3.26)-(3.27), where $\boldsymbol{A}$ is as described above.*

Proof. This follows by matrix multiplication and is, therefore, left out. □

Consider the following new sequence $\hat{x}^t \in V_{q,N}$, with $\hat{x}^0 = 0$. For $1 \leq t \leq q$, set $\hat{x}^t_{\bullet i} = x^t_{\bullet i}$, for corresponding non-zero columns of x^t. For the zero columns of x^t, set the corresponding column of $\hat{x}^t$ as follows: $\hat{x}^1_{\bullet i} = \sum_{j=1}^N A_{ij} f^j(x^0_{\bullet j}; 0)$, and

$$x^{k+1}_{\bullet i} = \sum_{j=1}^N A_{ij} f^j(x^k_{\bullet j}; k) - \frac{1}{N}\left(\sum_{j=1}^N \frac{\partial f^j}{\partial x}(x^k_{\bullet j}; k)\right) f^i(x^{k-1}_{\bullet i}; k-1), \qquad \text{for } k \geq 1,$$

with any negative index equal zero. By using 3.7, the following conclusion can easily be reach

.

Lemma 3.8. *The sequence $\{\hat{x}^t\}_{1\leq t\leq q}$ above satisfies the recursion 2.31 with the functions f^i as specified in 3.26 and 3.27.*

Next, we show that

$$\lim_{n\to\infty}\frac{1}{p}\left\|\tilde{\beta}^k-\mu_k\beta-v^k\right\|^2\overset{a.s.}{=}0, \qquad \text{and} \qquad \lim_{n\to\infty}\frac{1}{n}\left\|\theta^k-W^k\right\|\overset{a.s.}{=}0. \tag{3.28}$$

To see this, let $u^k=\tilde{\beta}^k-\mu_k\beta$. Then. from 3.16 and 3.21, it follows that

$$\begin{aligned}\|W^k-\theta^k\| &= \left\|XST\left(v^k+\mu_k\beta,\lambda_*\right)-\Psi(W^{k-1},y)-X\hat{\beta}^k+\Psi(\theta_{k-1},y)\right\|\\ &\leq \|X\|\left\|ST\left(v^k+\mu_k\beta,\lambda_*\right)-ST\left(u^k+\mu_k\beta,\lambda_*\right)\right\|+\left\|\Psi(\theta_{k-1},y)-\Psi(W^{k-1},y)\right\|.\end{aligned}$$

Using the fact that the function $ST(\cdot,\lambda_*)$ is Lipschitz and

$$\frac{\partial\Psi(s,t)}{\partial s}=b_kg_k'(s,t)=\frac{-b_k\rho''(z_*)}{1+b_k\rho''(z_*)}, \tag{3.29}$$

where $z_*=\text{prox}_\rho(s+bt;b)$, so that, $\Psi(\cdot,t)$ is Lipschitz with Lipschitz constant at most 1, it is the case that

$$\|W^k-\theta^k\|\leq\|X\|\|v^k-u^k\|+\|W^{k-1}-\theta^{k-1}\|. \tag{3.30}$$

Again, from 3.16 and 3.21, it follows that

$$\begin{aligned}
\left\|v^{k+1}-u^{k+1}\right\| &= \Big\|ST\left(q_k(v^k+\mu_k\beta),\lambda_*\right)-a_{k+1}\beta-ST(u^k+\mu_k\beta,\lambda_*)+\mu_{k+1}\beta+\\
&\quad \frac{1}{b_*c_*}X^T\left(\Psi(\theta^k,y)-\Psi(W^k,y)\right)\Big\|\\
&\leq \left\|ST\left(q_k(v^k+\mu_k\beta),\lambda_*\right)-ST(u^k+\mu_k\beta,\lambda_*)\right\|+\|(\mu_{k+1}-a_{k+1})\beta\|+\\
&\quad \frac{1}{\|b_*c_*\|}\|X^T\|\left\|\left(\Psi(\theta^k,y)-\Psi(W^k,y)\right)\right\|\\
&\leq \left\|q_k(v^k+\mu_k\beta)-(u^k+\mu_k\beta)\right\|+\|(\mu_{k+1}-a_{k+1})\beta\|+\\
&\quad \frac{1}{\|b_*c_*\|}\|X^T\|\left\|\left(\Psi(\theta^k,y)-\Psi(W^k,y)\right)\right\|\\
&\leq \left\|(q_k-1)\left(v^k+\mu_k\beta\right)+\left(v^k+\mu_k\beta\right)-(u^k+\mu_k\beta)\right\|+\|(\mu_{k+1}-a_{k+1})\beta\|+\\
&\quad \frac{1}{\|b_*c_*\|}\|X^T\|\left\|\left(\Psi(\theta^k,y)-\Psi(W^k,y)\right)\right\|\\
&\leq \|v^k-u^k\|+|q_k-1|\left\|v^k+\mu_k\beta\right\|+|\mu_{k+1}-a_{k+1}|\,\|\beta\|+ \qquad (3.31)\\
&\quad \frac{1}{|b_*c_*|}\|X^T\|\left\|\theta^k-W^k\right\|.
\end{aligned}$$

Since $v^0 = u^0$, combining 3.30, 3.31, it can be established that there exists a constant M, depending on c_*b_*, such that

$$\left\|v^{k+1}-u^{k+1}\right\| \leq (M\|X\|)^{2(k+1)}\left(\sum_{l=0}^{k}|q_l-1|\left|v^l+\mu_l\beta\right|+\sum_{l=0}^{k}|\mu_{k+1}-a_{k+1}|\,\|\beta\|\right) \tag{3.32}$$

Following similar line of argument as in the proof of Lemma 4 in the supporting document of Sur and Candès (2019), it follows that

$$\lim_{n\to\infty}\frac{1}{\sqrt{p}}\left\|v^k-u^k\right\|=0. \tag{3.33}$$

Further, using 3.30 and the fact that $\lim_{n\to\infty}\|X\|$ is finite almost surely, the same conclusion is reached which states that

$$\lim_{n\to\infty}\frac{1}{n}\left\|\theta^k-W^k\right\|^2 \overset{a.s.}{=} 0$$

thus, establishing 3.28.

Using the fact that ψ is a pseudo-Lipschitz function of order 2, it is found that

$$\begin{aligned}\left|\frac{1}{p}\sum_{j=1}^{p}\psi\left(\tilde{\beta}_j^k-\mu_k\beta_j,\beta_j\right)\right. &- \left.\frac{1}{p}\sum_{j=1}^{p}\psi\left(v_j^k,\beta_j\right)\right| \le \frac{1}{p}\sum_{j=1}^{p}\left|\psi\left(\tilde{\beta}_j^k-\mu_k\beta_j,\beta_j\right)-\psi\left(v_j^k,\beta_j\right)\right| \\ &\le \frac{M}{p}\sum_{j=1}^{p}\left(1+\|(\tilde{\beta}_j^k-\mu_k\beta_j,\beta_j)\|+\|(v_j^k,\beta_j)\|\right)\left|\tilde{\beta}_j^k-\mu_k\beta_j-v_j^k\right| \\ &\le \frac{M}{p}\sqrt{\sum_{j=1}^{p}\left(1+\|(\tilde{\beta}_j^k-\mu_k\beta_j,\beta_j)\|+\|(v_j^k,\beta_j)\|\right)^2}\left\|\tilde{\beta}^k-\mu_k\beta-v^k\right\|.\end{aligned}$$

It follows by definition that $\|\beta\|/\sqrt{p}$ is bounded. Combining 3.28 and Lemma 3.6 yields that $\|\tilde{\beta}^k\|/\sqrt{p}$ is bounded for all k. Thus, using the above inequality and Lemma 3.28, it is evident that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\tilde{\beta}_j^k-\mu_k\beta_j,\beta_j\right) &= \lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(v_j^k,\beta_j\right) \qquad (3.34)\\ \text{i.e.,}\quad \lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\tilde{\beta}_j^k-\mu_k\beta_j,\beta_j\right) &= E\left[\psi\left(\sigma_k G_k,\bar{\beta}\right)\right]\\ \text{which then yields,}\quad \lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\tilde{\beta}_j^k,\beta_j\right) &= E\left[\psi\left(\mu_k\beta+\sigma_k G_k,\bar{\beta}\right)\right].\end{aligned}$$

By the Lipschitz-ness of the function $ST(\cdot,\lambda_*)$, it is then found that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j^k,\beta_j\right) &= \lim_{n\to\infty}\tfrac{1}{p}\textstyle\sum_{j=1}^{p}\psi\left(ST(\tilde{\beta}_j^k,\lambda_*),\beta_j\right)\\ &= E\left[\psi\left(ST(\mu_k\bar{\beta}+\sigma_k G_k,\lambda_*),\bar{\beta}\right)\right] \qquad (3.35)\end{aligned}$$

establishing the first relation in 3.20. A similar argument up to 3.34 holds for the other relation.

□

Chapter 4

Phase Transition

4.1 Introduction and Background

This chapter starts off with a brief historical background and the theoretical framework for the concept known as phase transition. Then, it goes on to give the main result which at this time is a simulation study that provides evidence for the existence of phase transition for ℓ_1-penalized logistic regression estimator based on the asymptotic results of the previous chapter. Specifically, a Monte Carlo simulation is performed and used to shows that the ℓ_1-penalized logistic regression estimator demonstrates some sparsity–under-sampling tradeoff. A parameter space with axes quantifying under-sampling δ, and sparsity ϵ is considered. In the limit of large dimensions, i.e., $n, p \to \infty$ with $n/p = \delta > 0$, and $\|\beta\|_0/p = \epsilon$ ($\|\beta\|_0$ refers to the number of nonzero elements β), the parameter space partitions into two regions: one where the GAMP approach successfully achieves an accurate reconstruction of β and one where it fails.

Phase transitions in Generalized Linear Models (GLMs) are an interesting topic. The term first emanated in statistical physics, where it refers to abrupt changes in the properties of a physical system as a result of small changes in external conditions, such as temperature or pressure. These transitions are characterized by the emergence of new collective behaviors and the breakdown of symmetries.

Mathematically, phase transitions in GLMs have commonly been analyzed using tools from statistical physics, such as replica theory, mean-field theory, and the study of critical phenomena. These frameworks allow us to investigate the behavior and characteristic properties of

GLMs near the phase transition point and understand the emergence of complex patterns and phenomena.

In the context of GLMs, phase transitions refers to a situation when there is an abrupt change in the behavior of the model as a parameter crosses a critical threshold. Such parameter could be related to the sparsity of the data matrix, the strength of the input signals, or other factors. Phase transitions can manifest as sudden changes in performance metrics of a model, such as the model's predictive power, or the accuracy of parameter estimation. The comprehension and characterization of these phase transitions can enable researchers to identify the critical regions in the parameter space where the model's behavior qualitatively alters.

It is important to note that the study of phase transitions in GLMs is an active area of research, and there is still a multitude of unanswered questions and challenges that need to be tackled. However, by exploring the theoretical background and mathematical frameworks underlying phase transitions in GLMs, we can gain valuable insights into the behavior of these models and their applications in various fields.

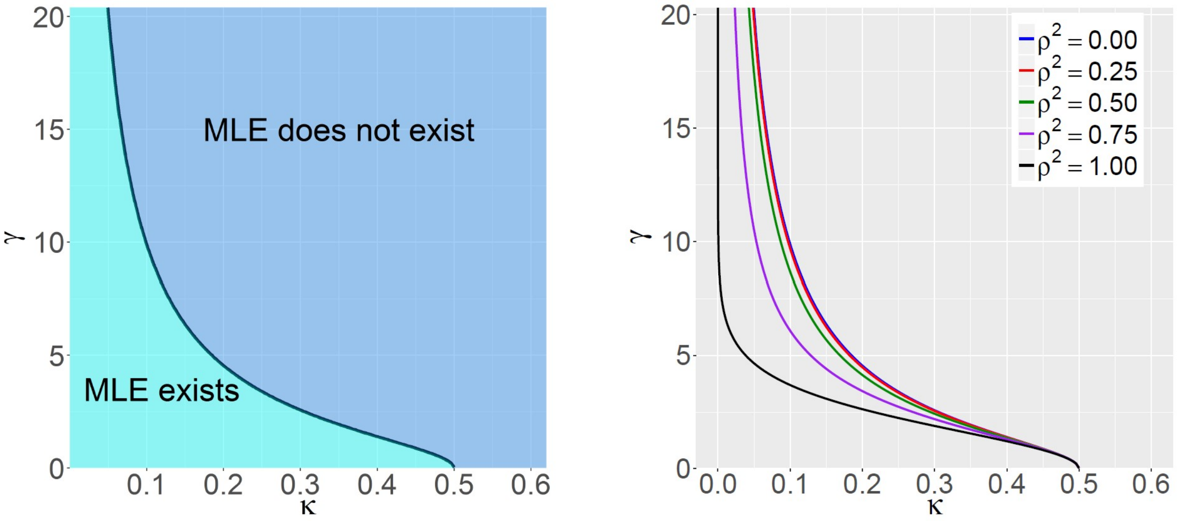
4.2 Types of Phase Transition Studies

In GLMs, there are several ways that phase transitions studies can be conducted depending on different parameter estimation methods or model regimes (sparse verses dense signal regimes). And talking about parameter estimation methods, there are various methods that can be used for GLMs estimation, such as maximum likelihood estimation (MLE), Bayesian estimation, or regularized estimation procedures like the Lasso or Ridge regression. Making a decision regarding the estimation technique, can have important consequences on the model's behavior and performance. The types of phase transitions seen here usually occur when the model assumptions or the data characteristics change.

There are numerous studies of phase transition for different parameter estimation methods. For instance, in the case of maximum likelihood estimation, it is a well-known phenomenon which has sparked several interesting investigations that the existence of the MLE is not guaranteed in all situations, even when the dimension p of the convariates is much smaller than the sample size n. An early study in this direction is the work of Silvapulle (1981), where the author put forward the necessary and sufficient conditions for the existence of the MLE, utilizing a geometric characterization that involves convex cones. This was then quickly followed by the work of Albert and Anderson (1984) on the existence of MLE for multinomial logistic regression models. The authors proved the existence theorems by considering the possible conditions of data geometry, which fell into three mutually exclusive and exhaustive categories: complete separation, quasicomplete separation and overlap (reader is refered to Albert and Anderson (1984) for details). They proved that the MLE exists if and only if the data points overlap. The work of Albert and Anderson (1984), was a major breakthrough that spurred several other studies including linear programming approach for the detection of separation (see e.e., Silvapulle and Burridge (1986), Lesaffre and Albert (1989), Kolassa (1997), Konis (2007)). Finally, Christmann and Rousseeuw (2001) applied the notion of regression depth as a data-analytic tool to measure the amount of overlap in datasets.

The above named results based on geometric characterizations, though beautiful, do not provide a practical guide for a data analyst to be able to tell when to expect the MLE to exist or not a priori given some random sample of data from some distribution. Fortunately, the early work of Cover (1964, 1965) provides an exception to this case. Cover's main result, provides that for the logistic regression, in the asymptotic regime where $n, p \to \infty$ with $p/n \to k$, and the covariates X_i are drawn from some distribution D, obeying certain conditions, and independent of the class labels $\{y_i\}_i$, having equal marginal probabilities, for $k < 1/2$, the data points tend to asymptotically overlap with near certainty, and for $k > 1/2$ the data points are separated also with the same degree of certainty. In the case where the MSE exists, Candes and Sur (2018) complemeted Cover's work and calculated the limiting distribution of MLE for Gaussian covariates.

Very recently, Candes and Sur (2018) have further investigated the existence of MLE in high-dimensional logistic regression models with Gaussian covariates. The authors established a phase boundary curve, h_{MLE}, which determines whether the MLE exists or not. They showed that if the problem is sufficiently high dimensional, meaning that the dimensionality ratio $p/n = k$ is greater than h_{MLE}, then the MLE does not exist with probability one. On the



(a) (b)

Figure 4.1: (a) Boundary curve $\gamma \mapsto g_{MLE}^{-1}(0,\gamma)$ separating the regions where the MLE asymptotically exists and where it does not (in this case $\beta_0 = 0$). (b) Boundary curves $\gamma \mapsto g_{MLE}^{-1}(\rho\gamma, \sqrt{1-\rho^2}\gamma)$ for various values of ρ.

other hand, if the dimensionality ratio is smaller than h_{MLE}, the MLE asymptotically exists with probability one. Specifically, they proved the following theorem:

Theorem 4.1 (Theorem 1 in Candes and Sur (2018)). *Let Z be a standard normal variable with density $\phi(t)$ and V be an independent continuous random variable with density $2\rho'(\gamma t)\phi(t)$. With $x_+ = max\{x, 0\}$, set*

$$g_{MLE}^{-1}(\gamma) = \min_{t\in\mathbb{R}} \left\{E(Z - tV)_+^2\right\} \tag{4.1}$$

which is a decreasing function of γ. Then in the setting described above,

$$\gamma > g_{MLE}(\kappa) \longrightarrow \lim_{n,p} \to \infty P\{MLEexists\} = 0, \tag{4.2}$$

$$\gamma < g_{MLE}(\kappa) \longrightarrow \lim_{n,p} \to \infty P\{MLEexists\} = 1, \tag{4.3}$$

The phase curve from Theorem 4.1 above is ploted in Fig 4.1.

It is essential to note that all the results from the above survey are applicable only in the regime where the dimension p of the data is smaller than the sample size n. For high dimensional problems in general, as was discussed in Chapter 2, the method of ℓ_1 minimization is a well established approached for handling it. But in this setting, the paradigm for phase transition immediately shifts from merely existence and uniqueness. An estimate is admitted as solution only if it lies within a certain threshold of the performance metric, the most popular of which is the mean squared error. Here, phase transition is considered in relation to the model regime. In the sparse signal regime, the model has a small number of non-zero parameters, meaning that only a few of the attributes of the predictor variables are important for characterizing the general behaviour of the response variable. On the other hand, in the dense signal regime, a larger number of predictors attributes is needed to effectively characterize the response variable. This phase transition in this setting is usually observed when the sparsity parameter, under-sampling parameter, or the regularization strength is/are varied. Typically, interest is usually in finding the best combination of sparsity verses under-sampling. By focusing on sparse solutions, which have a smaller number of non-zero coefficients, researchers can reduce the computational complexity and improve the interpretability of the results. Additionally, sparsity under-sampling helps to identify the critical threshold at which the phase transition occurs, providing insights into the behavior of the estimator in high-dimensional settings.

The work of Donoho et al. (2009) on compressed sensing introduced a modified iterative thresholding algorithms (this was refered to as AMP algorithms in earlier chapters) which while being far less expensive in application, achieves an equivalent sparsity–under-sampling tradeoff as the convex optimization method which was the best known technique at the time. In their study, they examined a parameter space that measures sparsity and under-sampling. In the limit of large dimensions p and n, they demonstrated a phase transition where the parameter space is divided into two partitions. In one phase, the AMP approach is able to accurately reconstruct the signal, while in the other phase, it is unsuccessful. Previous studies have identified regions of success and failure for LP-based recovery. Surprisingly, They have found that these two partitions of the sparsity-under-sampling parameter space are actually identical. Both reconstruction approaches succeed or fail in the same regions. Inpired by the findings in

Donoho et al. (2009) were base on numerical simulations and heuristic arguments, Bayati and Montanari (2011) went on and provided a rigorous theoretical support.

Several variations and extensions of AMP algorithms including the GAMP have been proposed in the literature by different authors for various types of M-estimation problem in high dimensions, and phase transition analysis is a constant step each time as it helps in understanding the performance and limitations of the algorithm and can provide valuable insights for signal recovery and estimation problems.

Some examples are the following. Bayati and Montanari (2012) considered a sequences of matrices with increasing dimensions and independent Gaussian entries. They proved that the normalized risk of the LASSO converges to a limit, and provided an explicit expression for this limit. Their result was the first rigorous derivation of an explicit formula for the asymptotic mean square error of the LASSO for random instances. The proof technique used in their study is also based on the analysis of AMP. Huang (2022) extended the work of Bayati and Montanari (2012) and studied the LASSO phase transition under arbitrary covariance dependence. The authors considerd a matrix X consisting of i.i.d. Gaussian rows with a general covariance matrix Σ. They presented explicit formulas that precisely characterize the trade-off between sparsity and under-sampling for arbitrary Σ.

Huang (2020) discussed the derivation of the asymptotic mean square error (MSE) of ℓ_1-penalized robust estimators in the context of high-dimensional regression models. The paper focused on the ℓ_1-penalized least absolute deviation and ℓ_1-penalized Huber's regressions. The authors analyzed the appearance of a sharp phase transition in the two-dimensional sparsity-under-sampling phase space and derive the explicit formula of the phase boundary. They find that the phase boundary is identical to the phase transition curve of LASSO and the Donoho-Tanner phase transition for sparse recovery. The derivation is based on the asymptotic analysis of the GAMP algorithm. They establish the asymptotic MSE of the ℓ_1-penalized robust estimator by connecting it to the asymptotic MSE of the corresponding GAMP estimator. Their results provide theoretical insights into high-dimensional regression methods, and computational experiments validate the correctness of the analytic results.

4.3 Phase Transition for ℓ_1 Penalized Logistic Regression

In this section, a simulation study of the phase transition properties of the ℓ_1-penalized logistic regression estimator based on the asymptotic results of the previous chapter is presented. This is done by following the same approach used in the paper Huang (2020) on asymptotic analysis.

Under the same setting as in Section 3.2, consider the sparsity/under-sampling phase space $(\epsilon, \delta) \in [0, 1]^2$ i.e., different combination of sparsity ϵ, and under-sampling δ. As can be easily deduced from the literature survey from the previous sections, the problem of reconstruction of sparse signal for under-determined systems is subject to sparsity ϵ and under-sampling δ trade-offs. There is a function $\delta(\epsilon)$, commonly referred to as phase curve that splits the space $(\epsilon, \delta) \in [0, 1]^2$ into two regions, a "success" region where exact reconstruction is attained, and a "failure" region where it is not, and our goal is to derive the formula of the curve, or at the very least provide numerical proof of its existence and location in the phase space.

At the point of this writing, the analytical formulation for the theoretical phase curve remains to be completed, but a robust empirical evidence is presented. First, a grid of 31 ϵ values is fixed in [0.05, 0.95]. For each ϵ, the following sequence of 20 δ values $\{0.05, 0.1, ..., 1\}$ is then considered. For each $\epsilon - \delta$ combination, 20 instances of $(\boldsymbol{X}, \beta)$ with dimension $p = 1000$, is generated, and $y = (y_1, ..., y_n)$ with $y_i \in \{0, 1\}$, is such that $P(y_i = 1|X_i) = \rho'(X_i^T\beta)$. For the ith iteration, an estimate β_i is obtain by using the *glmnet* function in R, on the ith simulated data. A success indicator variable is then defined as

$$S_i = 1 \text{ if } \frac{||\hat{\beta}_i - \beta||_2}{||\beta||_2} \leq 8.5 \times 10^{-1}$$

, and $S_i = 0$ otherwise. Finally, at each (ϵ, δ) combination, a variable S is defined by $S = \sum_{i=1}^{20} S_i$.

Next, a matrix with dimensions 20 by 31 is created. The rows of the matrix correspond to the values of the delta sequence, and the columns correspond to the values of the epsilon sequence. The matrix is then filled with the corresponding success rate $(S/20)$ of signal recovery for each combination of delta and epsilon. Fig. 4.2 is the heat map of the success rate

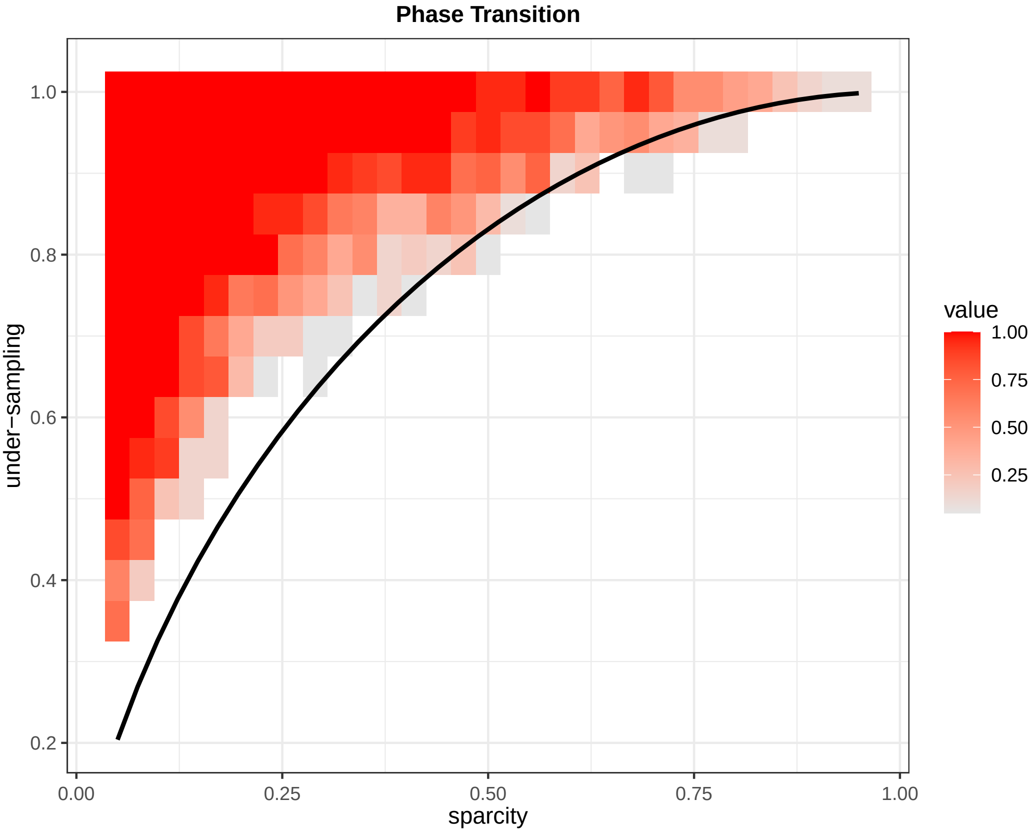


Figure 4.2: Heat Map of the success rate matrix of ℓ_1-penalized logistic regression. The solid line is the plot of the curve given in Conjecture 1.

matrix, and shows clear evidence and the location of a phase transition for ℓ_1-penalized logistic regression.

Conjecture 1. Consider the sparse class

$$\mathcal{S}_\epsilon := \{\pi_\beta : \pi_\beta \text{ is a probability measure with } \pi_\beta \geq 1-\epsilon\}. \tag{4.4}$$

Then, the phase space $0 \leq \delta, \epsilon \leq 1$ can be partitioned into two regions separated by a curve $\delta_* = \delta(\epsilon)$. Above this curve, the ℓ_1-penalized logistic regression estimator perfectly recovers the sparse signal β with high probability, after carefully choosing the tuning parameter λ. Below this curve, the estimator fail in the recovery task with high probability. We conjecture that the phase transition curve is determined by

$$\delta_* = \frac{2[\phi(\alpha) - \alpha\Phi(-\alpha)]}{\alpha + 2[\phi(\alpha) - \alpha\Phi(-\alpha)]} \tag{4.5}$$

where α is determined by

$$\epsilon = \frac{2\phi(\alpha)}{\alpha + 2[\phi(\alpha) - \alpha\Phi(-\alpha)]}, \tag{4.6}$$

with $\alpha \in [0, \infty)$.

Figure 4.2 shows that overall, the conjectured curve aligns fairly well with empirical data. Given this agreement, it is reasonable to consider pursuing a proof in the direction of the conjecture. The close correspondence between the predicted outcomes of the curve and the observed results strongly suggests that there may be a deeper underlying relationship at play. By rigorously establishing the validity of the conjecture or some variant of it through a formal proof, we can gain a deeper understanding of the phenomenon and potentially unlock new insights and applications.

Chapter 5

Conclusion and Future Work

The logistic regression is one of, if not the most popular statistical model used for binary classification problems. One of the key advantages of logistic regression is its interpretability. The model provides estimates of the coefficients for each independent variable, which can be interpreted as the effect of that variable on the probability of the binary outcome. This makes logistic regression a valuable tool for understanding the factors that influence a particular outcome.

Given that interpretablitly is one of the major attractions of the logistic regression, several estimation techniques have been proposed in the literature for the logistic regression, including maximum likelihood estimation (MLE), Bayesian estimation, and penalized estimation methods such as ridge regression and lasso. Introduce a penalty term to the likelihood function to encourage sparsity in the estimated coefficients. These techniques have been widely used in logistic regression to improve model performance and handle high-dimensional data.

Sur and Candès (2019), outlined the limitations of classical maximum likelihood theory in the context of logistic regression when the number of features and sample size are large and comparable. The authors highlighted that classical results, such as the unbiasedness of the maximum likelihood estimate (MLE) and the Chi-Squared distribution of the log-likelihood ratio (LLR) statistic, are inaccurate in this setting.

Inspired by the work of Sur and Candès, and noting the ubiquity of high dimensional problems with $p > n$ in modern applications where the ML theory fails, our interest was naturally drawn to investigate alternate procedures that would remain applicable under these settings. Specifically, we explored the problem of ℓ_1-penalized logistic regression given by (1.6), and our results extend even to the *large-n-large-p* regime.

The main result of this work contains an explicit characterization of the high dimensional limit of the ℓ_1 penalized logistic regression estimator. Using this characterization, the formula of asymptotic mean square error and the asymptotic selection error rate, to name a few were derived, and in both cases, were backed up by results from extensive numerical experiments. Further numerical experimentation revealed the existence and location of a phase transition in the two-dimensional sparsity-undersampling phase space. The formalism underpinning the approach used here is based on the asymptotic analysis of the GAMP algorithm. The results provide theoretical insights into high-dimensional regression methods. For instance, it can be used to tune the regularization parameter since it gives an exact formula for the asymptotic MSE. The phase transition result is also new and will now serve as a guide for when logistic regression estimates based on the ℓ_1 regularization technique are reliable.

5.1 Future Work

An immediate future work here is the completion the theoretical derivation of an explicit formula of the phase transition curve for which there has been provided a numerical evidence here.

It would also be interesting to try to recover the results established here for the logit model in the case of the probit model and complementary log-log model. This is because the logit model, probit model, and complementary log-log model are all commonly used statistical models for binary classification, but they have different underlying assumptions and estimation methods. By comparing the results between the two models, we can gain a better understanding of the similarities and differences in their performance and applicability.

Furthermore, in addition to the lasso, there have been several other types of regularization methods that have been studied for GLMs. A close relative of the lasso is the ridge regression, which adds an ℓ_2 penalty term to the likelihood function to control the complexity of the model. Also, there is the Elastic net regularization which uses a combination of ridge and lasso regression, and allows for both variable selection and shrinkage. Another method is the group lasso, which encourages sparsity at the group level rather than the individual variable level. There are also methods such as adaptive lasso, which adaptively weights the penalty term based on the estimated coefficients. Overall, these regularization methods have been studied and applied in various GLM settings to improve model performance and interpretability. We will like to explore all these different types of regularization methods with the logistic regression and other related GLMs.

Considering multiple types of regularization techniques on a GLM can have several benefits. it allows for a more flexible and robust modeling approach. Different regularization techniques have different strengths and weaknesses, so by considering multiple techniques, we can potentially capture a wider range of patterns and relationships in the data. This can lead to improved model performance and better predictive accuracy.

To summarize, there are various new paths that are open for further exploration from here. These new directions can succinctly be represented by the following optimization problem:

$$\hat{\beta}_h = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \ell(\boldsymbol{\beta}) + h(\boldsymbol{\beta}) \right\} \tag{5.1}$$

for different combinations of loss functions $\ell(\cdot)$ and convex penalty functions $h(\cdot)$. In the case of ℓ_2 regularization, the penalty is taken to be $h(\beta) = ||\beta||^2$, and for the probit model, $\ell(\cdot)$ would be standard normal cdf.

References

A., E. M. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, 191–203.

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.

Ali, A. and R. J. Tibshirani (2018). The generalized lasso problem and uniqueness.

Amelunxen, D., M. Lotz, M. B. McCoy, and J. A. Tropp (2013). Living on the edge: Phase transitions in convex programs with random data.

Bai, Z. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices*, Volume 20. Springer.

Bayati, M. and A. Montanari (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory 57*(2), 764–785.

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

Bean, D., P. J. Bickel, N. E. Karoui, and B. Yu (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences 110*(36), 14563–14568.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705 – 1732.

Boyd, S., S. Boyd, L. Vandenberghe, and C. U. Press (2004). *Convex Optimization*. Number pt. 1 in Berichte über verteilte messysteme. Cambridge University Press.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg.

Cai, T. T., Z. Guo, and R. Ma (2021). Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association 0*(0), 1–14.

Candes, E. and T. Tao (2005). Decoding by linear programming. *IEEE Transactions on Information Theory 51*(12), 4203–4215.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313 – 2351.

Candes, E. J. and P. Sur (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression.

Candès, E. J. and P. Sur (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics 48*(1), 27 – 42.

Chen, S. S., D. L. Donoho, and M. A. Saunders (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing 20*(1), 33–61.

Christmann, A. and P. J. Rousseeuw (2001). Measuring overlap in binary regression. *Computational Statistics Data Analysis 37*(1), 65–75.

Cohen, A., W. Dahmen, and R. Devore (2009, January). Compressed sensing and best k -term approximation. *Journal of the American Mathematical Society 22*(1), 211–231.

Cover, T. M. (1964). *Geometrical and statistical properties of linear threshold devices*. Ph. D. thesis, Stanford University.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers EC-14*(3), 326–334.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological) 20*(2), 215–242.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics 6*(4).

Deshpande, Y., E. Abbe, and A. Montanari (2016, 12). Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA 6*(2), 125–170.

Donoho, D., M. Elad, and V. Temlyakov (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory 52*(1), 6–18.

Donoho, D. and X. Huo (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory 47*(7), 2845–2862.

Donoho, D. and A. Montanari (2013). High dimensional robust m-estimation: Asymptotic variance via approximate message passing.

Donoho, D. L., A. Javanmard, and A. Montanari (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory 59*(11), 7434–7464.

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

Donoho, D. L., A. Maleki, and A. Montanari (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory 57*(10), 6920–6941.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407 – 499.

Fan, J. and R. Li (2011). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*(3), 928 – 961.

Feng, O. Y., R. Venkataramanan, C. Rush, and R. J. Samworth (2021). A unifying tutorial on approximate message passing. *arXiv preprint arXiv:2105.02180*.

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

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302 – 332.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Fu, W. and K. Knight (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356 – 1378.

Guo, Z., P. Rakshit, D. S. Herman, and J. Chen (2021). Inference for the case probability in high-dimensional logistic regression. *Journal of Machine Learning Research 22*(254), 1–54.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

He, X. and Q.-M. Shao (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis 73*(1), 120–135.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics 32*(1), 1–49.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Huang, H. (2020). Asymptotic risk and phase transition of l_1-penalized robust estimator. *The Annals of Statistics 48*(5), 3090–3111.

Huang, H. (2022). LASSO risk and phase transition under dependence. *Electronic Journal of Statistics 16*(2), 6512 – 6552.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics 1*(5), 799–821.

Janková, J., R. D. Shah, P. Bühlmann, and R. J. Samworth (2020). Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(3), 773–795.

Javanmard, A. and A. Montanari (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA 2*(2), 115–144.

Javanmard, A. and A. Montanari (2014a, jan). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res. 15*(1), 2869–2909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Karoui, E. and Noureddine (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields 170*(none).

Karoui, N. E., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences 110*(36), 14557–14562.

Kolassa, J. E. (1997). Infinite parameter estimates in logistic regression, with application to approximate conditional inference. *Scandinavian Journal of Statistics 24*(4), 523–530.

Konis, K. P. (2007). Linear programming algorithms for detecting separated data in binary logistic regression models.

Krzakala, F., M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová (2012, May). Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X 2*, 021005.

Lange, K., E. C. Chi, and H. Zhou (2014). A brief survey of modern optimization for statisticians. *International Statistical Review / Revue Internationale de Statistique 82*(1), 46–70.

Lehmann, E. L. and J. P. Romano (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lesaffre, E. and A. Albert (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society. Series B (Methodological) 51*(1), 109–116.

Ma, R., T. T. Cai, and H. Li (2021). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association 116*(534), 984–998. PMID: 34421157.

Mammen, E. (1989). Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap. *The Annals of Statistics 17*(1), 382 – 400.

Matsushita, R. and T. Tanaka (2013). Low-rank matrix reconstruction and clustering via approximate message passing. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

McCullagh, P. and J. Nelder (1983). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Springer US.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics 34*(3), 1436 – 1462.

Mondelli, M., C. Thrampoulidis, and R. Venkataramanan (2021, August). Optimal combination of linear and spectral estimators for generalized linear models. *Foundations of Computational Mathematics* (1615-3383).

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics 45*(1), 158–195.

Parikh, N. and S. Boyd (2014, jan). Proximal algorithms. *Found. Trends Optim. 1*(3), 127–239.

Park, M. Y. and T. Hastie (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 659–677.

Portnoy, S. (1984). Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *The Annals of Statistics 12*(4), 1298 – 1309.

Portnoy, S. (1985). Asymptotic behavior of m estimators of p regression parameters when p2 / n is large; ii. normal approximation. *The Annals of Statistics 13*(4), 1403–1417.

Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity. *The Annals of Statistics 16*(1), 356 – 366.

Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 2168–2172. IEEE.

Reinsel, D., J. Gantz, and J. Rydning (2017). Data age 2025: The evolution of data to life-critical. *Don't Focus on Big Data 2*.

Rosset, S. and J. Zhu (2007). Piecewise linear regularized solution paths. *The Annals of Statistics 35*(3), 1012 – 1030.

Schniter, P. and S. Rangan (2012). Compressive phase retrieval via generalized approximate message passing. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 815–822.

Shi, C., R. Song, Z. Chen, and R. Li (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics 47*(5), 2671 – 2703.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological) 43*(3), 310–313.

Silvapulle, M. J. and J. Burridge (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(1), 100–106.

Spokoiny, V. (2012). Penalized maximum likelihood estimation and effective dimension.

Sur, P. and E. J. Candès (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences 116*(29), 14516–14525.

Sur, P., Y. Chen, and E. J. Candès (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square.

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

Tibshirani, R. J. and J. Taylor (2011). The solution path of the generalized lasso. *The Annals of Statistics 39*(3), 1335 – 1371.

Tropp, J. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory 50*(10), 2231–2242.

van de Geer, S. (2007). The deterministic lasso. Report, Zürich.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166 – 1202.

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*(none), 1360 – 1392.

Wainwright, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity.

Wu, T. T. and K. Lange (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics 2*(1), 224 – 244.

Xia, L., B. Nan, and Y. Li (2020). A revisit to de-biased lasso for generalized linear models.

Zhang, C.-H. and S. S. Zhang (2011). Confidence intervals for low-dimensional parameters in high-dimensional linear models.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*(90), 2541–2563.

Zhao, T., H. Liu, and T. Zhang (2014). Pathwise coordinate optimization for sparse learning: Algorithm and theory.

Zhu, Y., X. Shen, and W. Pan (2020). On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association 115*(529), 217–230. PMID: 32788818.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics 37*(4), 1733 – 1751.