# Some Contributions to High Dimensional Mixed Effects Logistic Regression Models

by

Jun Guo

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2018

Doctoral Committee:

Associate Professor Yves Atchadé, Chair
Associate Professor Ambuj Tewari
Assistant Professor Zhenke Wu
Assistant Professor Gongjun Xu
Professor Ji Zhu

Jun Guo

guojun@umich.edu

ORCID iD: 0000-0002-3702-2364

To my fiancée Ping Hou.

# Acknowledgement

Firstly, I would like to thank my Ph.D. advisor, Professor Yves Atchadé. Yves is a highly learned scholar who is very patient in guiding me through my thesis research. He is creative in selecting tailored problems for me to work on and always help me with my obstacles along the way. The knowledge and skills I have gained while working with him have prepared me well for my future. I would also like to thank many other professors for their meaningful and informative discussions with me throughout my Ph.D. career, which contribute incredibly in my academic and personal growth, among them Professors Ji Zhu, Ambuj Tewari, Elizaveta Levina and Vijayan N. Nair have left me with lasting impact. I greatly appreciate the many support I have received for dissertation writing while I am working at Wells Fargo.

During my Ph.D. years, I am greatly blessed to have shared my life with many Christian brothers and sisters. With them, I am learning to love God and love other people. In our Olive Tree Campus Church, I have met my fiancée Ping, with whom, in God, we see purpose, hope and strength in everything we experience together in life.

Last and above all, I give all my thanks to my Lord God Jesus Christ. He is with me till the end. *Soli Deo Gloria*.

# Table of Contents

# List of Figures

**Figure**

# List of Tables

**Table**

## Abstract

High dimensional mixed effects generalized linear models extend the generalized linear models (GLMs) by adding random effects to the linear predictors of the original high dimensional GLMs. The high dimensional mixed effect logistic regression is a typical example. These models are useful in analyzing categorical or discrete data with group structure. Inference for these models is challenging because of the intractable and generally non-convex negative log-likelihood function. In this dissertation, we propose and analyze four different algorithms to solve the high dimensional mixed-effects logistic regression model.

The first two algorithms we develop are stochastic proximal gradient and second order approximate algorithms, which are both proximal gradient based algorithms. As the gradient of the loss function is intractable, the stochastic proximal gradient algorithm uses a Markov chain Monte Carlo technique to approximate the gradient, while the second order approximate algorithm approximates the objective function based on Taylor expansion to the second order, and solves an approximate problem. We prove the convergence of the second order approximate algorithm using the Kurdyka-Lojasiewicz (K-L) property based techniques. To analyze convergence behavior of the stochastic proximal gradient algorithm, we expand this K-L based technique to incorporate stochastic perturbations in the algorithm updates. We show that the stochastic algorithm's limiting points are the stationary points of the orig-

inal objective function. We illustrate the good performance of our algorithms in several numerical examples. We also apply the two algorithms in a breast cancer data analysis.

The next algorithm we consider is based on a "fixed effect approximation" of the mixed effects models. Here we treat the random effects as unknown fixed effects coefficients, and estimate them without penalty. The approximation reduces the original problem to the usual high dimensional logistic regression with offset terms. Computational efficiency is a clear gain, non-convex problem is also replaced by a convex one. We have derived a non-asymptotic estimation error bound for its solution with respect to the true model parameters. In this effort, we have expanded the restricted eigenvalue (RE) condition to a stochastic setting, which holds with high probability in our problem. We have conducted extensive numerical study of this approximation scheme, and compared its performance with the previous two algorithms. The same breast cancer data is analyzed by this algorithm.

Our final algorithms are the iterated filtering algorithms. The core of this algorithm is a novel "pseudo proximal map" which computes the mean of a constructed log-likelihood function to approximate the optimum of the objective function. We explore its connections to the proximal and gradient descent algorithms and focus on its application in composite objective function optimization. We then devise the iterated filtering algorithm and its block coordinate update version to solve the high dimensional mixed effect logistic regression model. Under strong convexity assumption, we derive new convergence results for the algorithm sharper than previous results in the literature. We use numerical studies to demonstrate the effectiveness of our algorithm.

# Chapter I

# Introduction

In this dissertation, we try to solve the high dimensional mixed effects logistic regression model. This model is very useful for modeling discrete data with hidden group structures, for example in cross population genome wise association studies. However, it has received scarce literature attention, especially in the theoretical front. Estimation of this model is challenging because of the intractable and in general non-convex model log-likelihood function. We have developed and analyzed four different algorithms to tackle this problem.

Building on the popular proximal gradient algorithms (*Combettes and Wajs* (2005); *Parikh and Boyd* (2013b)), we have used a Polya-Gamma Markov chain Monte Carlo (MCMC) sampler (*Polson et al.* (2012)) and a Taylor series expansion to approximate the intractable gradient of the log-likelihood function, these lead to our stochastic proximal gradient and second order approximate algorithms. We have used Kurdyka-Łojasiewicz (K-L) property (*Kurdyka* (1998)) based technique (*Attouch and Bolte* (2009)) to analyze convergence of our algorithms in a non-convex and non-smooth setting. We have expanded this technique to incorporate the stochas-

tic perturbation in the updates of stochastic proximal gradient algorithm, and have shown that the limiting points of the updates are all stationary points of the objective function. Our analysis and numerical evidence of stochastic proximal gradient algorithm shows that proper Markov chain Monte Carlo techniques can be incorporated in the algorithm to *exactly* solve the maximum likelihood problem of high dimensional mixed effects logistic regression model, which is a meaningful addition to the previous literature results which use numerical techniques to *approximately* solve the problem (*Schelldorfer et al.* (2014) and references therein).

In pursuit of efficient solutions, we have proposed to approximate the model by taking the mixed effects as additional unknown fixed effects coefficients, this reduces the original non-convex and intractable problem to be convex and tractable, and leads to our "fixed effects approximate algorithm". An obvious gain of this algorithm is computational efficiency. In addition to computational gains, we have established a non-asymptotic estimation error bound to basically show that when the random effects noise level (standard deviation) is reasonably small, with other suitable conditions the solution of this approximation is close to the true model parameters with high probability and the solution is consistent. To establish the error bound, we have extended the restricted eigenvalue (RE) condition (*Bickel et al.* (2009); *Koltchinskii* (2009)) to a stochastic setting. This development shows that in some cases, the challenging high dimensional mixed effects logistic regression can be highly efficiently solved by treating the relatively weak random effects as unknown fixed effects coefficients, moreover, the solutions of this approximation are statistically sound under suitable conditions.

Finally, we explore and apply the novel iterated filtering algorithm recently

proposed in the statistical literature (*Ionides et al.* (2006, 2011, 2015)). In our development, we have removed a redundancy in the original iterated filtering algorithm (*Ionides et al.* (2006, 2011)) by leveraging a result showing that the "pseudo proximal map" - the core of iterated filtering algorithm, is close to the gradient map under reasonable conditions (*Doucet et al.* (2013)). We have also related this algorithm more closely to well-known stochastic gradient methods, and with strongly convexity assumption derived sharper convergence results than those of *Ionides et al.* (2011). By incorporating a usual importance sampling technique, we have successfully applied this algorithm to solve the high dimensional mixed effect logistic regression model. While we have not spelled out the details, it is not hard to see the convergence analysis done for stochastic proximal gradient algorithm can be adapted to show that in our case the limiting points of the iterated filtering updates are also stationary points of the objective function. This development assures us that simple importance sampling Monte Carlo, versus the advanced MCMC methods like Polya-Gamma sampler, can come in handy to exactly solve the high dimensional mixed effect logistic regression model.

In the following, we give a comprehensive introduction of the background, literature, and other details of our problem, algorithms, their analysis and application. We also give more details of our contributions and the organization of the dissertation toward the end of this chapter.

The high dimensional generalized linear models (GLMs) (*McCullagh and Nelder* (1989)), and linear mixed effects models (LMMs) (*Rosenberg* (1973) as one of the early references) are well known and widely applied in nearly every field of data analysis. While the GLMs assume independent observations, and LMMs are applicable

only to continuous observations, there are many cases in practice where the observations are correlated as well as discrete or categorical. Any common longitudinal study with binary response serves a simple example. The generalized linear mixed effects models (GLMMs) (*McCullagh and Nelder* (1989); *Breslow and Clayton* (1993); *McCulloch and Searle* (2005); *Molenberghs and Verbeke* (2005)), which add random effects to the linear predictors in the GLMs, are likely candidate models for these situations, the mixed effects logistic regression is a typical example. GLMMs have been widely applied in fields like genomics, genetics, biology, ecology, medicine, pharmaceutical science, just to name a few ( *Yu* (2006); *Jiang* (2007); *Atwell* (2010); *Zhang* (2010b); *Yang* (2011); *Zhou et al.* (2013); *Bühlmann et al.* (2014); *Aulchenko* (2007)).

The building blocks of GLMMs are fixed effects covariates with a corresponding $p$-dimensional parameter vector, and random effects with a $q$-dimensional random effect parameter vector. To define such a model in mathematical terms, suppose that condition on a random effects vector $\mathbf{u} \in \mathbb{R}^q$, the observations $y_1, \ldots, y_n$ are conditionally independent such that the conditional distribution of $y_i$ given $\mathbf{u}$ is a member of the exponential family with probability density function

$$f_i(y_i|\mathbf{u}) = \exp\left\{\frac{y_i\xi_i - b(\xi_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\} \tag{1.1}$$

where $b(\cdot), a_i(\cdot), c_i(\cdot,\cdot)$ are known functions, and $\phi$ is a dispersion parameter which may or may not be known. The quantity $\xi_i$ is associated with the conditional mean $\mu_i = \mathbb{E}(y_i|\mathbf{u})$, $\mu_i$ is in turn associated with a linear predictor $\eta_i = x_i'\beta + z_i'\mathbf{u}$ through a specified link function $g(\cdot)$ such that $g(\mu_i) = \eta_i$, where $x_i,\ z_i \in \mathbb{R}^n$ are known fixed effects data and random effects loading vectors. $\beta \in \mathbb{R}^p$ is a vector of fixed unknown parameters. Under canonical link functions (*McCullagh and Nelder* (1989), pp. 32),

we have $\xi_i = \eta_i$. Furthermore, it is assumed that random effect $\mathbf{u} \sim \mathsf{N}(0, \Sigma)$. From general frequentist model inference point of view, the unknown parameters in the model are $\beta$, $\Sigma$, and possibly an error variance term $\sigma_\epsilon^2$, for example in linear mixed effects models, which is a special case of GLMMs.

The exponential families in model 1.1 can be binomial, multinomial, negative-binomial, Poisson, Gaussian etc. The Gaussian and binomial cases lead to the Linear mixed effects models and mixed effect logistic models, which are two popular special cases of GLMMs. In linear mixed effects models, for all $i = 1, \ldots, n$, with $a_i(\phi) \equiv \sigma_\epsilon^2$, $b(\eta_i) = \eta_i^2/2$, $c_i(y_i, \phi) = -\log(2\pi\sigma^2)/2 - y_i^2/2\sigma^2$, and the mixed effects linear predictor being $\eta_i = x_i'\beta + z_i'\mathbf{u}$, (1.1) becomes,

$$f_i(y_i|\mathbf{u}) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(\frac{(y_i - x_i'\beta - z_i'\mathbf{u})^2}{2\sigma_\epsilon^2}\right) \tag{1.2}$$

While in mixed effect logistic regression models, for all $i = 1, \ldots, n$, with $y_i \in \{0, 1\}$, $a_i(\phi) \equiv 1$, $b(\eta_i) = \log(1 + \exp(\eta_i))$, $c_i(y_i, \phi) \equiv 0$, and $\eta_i = x_i'\beta + z_i'\mathbf{u}$, (1.1) becomes,

$$f_i(y_i|u) = \frac{\exp\left[y_i\left(x_i'\beta + z_i'\mathbf{u}\right)\right]}{1 + \exp(x_i'\beta + z_i'\mathbf{u})} \tag{1.3}$$

The problem of regularized maximum likelihood based estimation of the fixed effects coefficients $\beta$ in high dimensional mixed effects logistic regression model is of our interest. The LMM likelihood function of $\beta$ given $\Sigma$ and $\sigma_\epsilon$ is concave, thus estimation of the fixed effect coefficients of LMMs given the variance components remains a convex problem. However, the likelihood functions of other GLMMs, including the mixed effect logistic regression model, are in general non-convex in both the fixed effects parameter $\beta \in \mathbb{R}^p$ and variance-covariance matrix $\Sigma$. In this thesis, we will assume the random effects variance-covariance matrix $\Sigma$ is given as known.

In fact, in applied studies, $\Sigma$ is often estimated with data or *a-priori* knowledge of the random effects before model fitting and is treated as a given parameter subsequently. In our real data analysis in Chapters II and III, we will illustrate one way of estimating the random effect variance-covariance matrix when GLMMs are applied in breast cancer study with gene expression information.

Most of the existing literature dealing with mixed effect generalized linear models are low-dimensional and classical in nature (*Schelldorfer et al.* (2014)). Some works focus on regularization driven variable selection procedures in GLMMs with low dimensional data: *Ibrahim et al.* (2011); *Groll and Tutz* (2014). The high dimensional scenario typically refers to the case when number of fixed effects coefficients $p$ is large (larger than the sample size $N$), but the dimension of the random effects factor, for instance in terms of the rank $q$ of its variance-covariance matrix $\Sigma$, is small. In this setting, we can again apply various sparsity penalties to the log-likelihood function to obtain the regularized MLEs:(*Bühlmann et al.* (2014))

$$\widehat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ -\ell_n(\beta;\ Y) + g(\beta) \right\} \tag{1.4}$$

Where $g$ is the regularization function. We will focus on $\ell_1$ Lasso and elastic-net penalties in this thesis.

The difficulties of the estimator in (1.4) lie mainly in two aspects: first in the log-likelihood $-\ell_n(\beta;\ Y)$ being a non-convex function in the unknown parameters in general, second in the log-likelihood function being intractable and generally hard to access.

In the non-convex log-likelihood aspect, our problem is non-convex in the loss function, while well studied non-convex problem examples in statistical literature focused on cases where non-convex penalty functions sum with convex loss functions (*Fan and Li* (2001); *Zhang* (2010a); *Loh and Wainwright* (2017)). Recently a few literatures have focused on the non-convex loss functions. *Schelldorfer et al.* (2011) have devised algorithms to solve the high dimensional linear mixed effects models for both fixed effects parameters $\beta$ and the variance-covariance component $\Sigma + \sigma_\epsilon^2 I$, although it is non-convex solving for the the variance components, it remains a convex problem solving for $\beta$. There are some algorithms commonly used in solving non-convex objective functions, for example the proximal gradient algorithm, the EM algorithm, the alternating direction method of multipliers (ADMM), and the iterated filtering algorithms. We will consider developing our algorithms based on the proximal gradient algorithms in chapter II and the iterated filtering algorithms in chapter IV. Our initial numerical experiments showed that one ADMM we developed performs similar in estimation to our other algorithms, but converges much slower and takes significant longer time, so we will not pursue it further in this dissertation. The EM algorithm applied to our problem, which will be very different from gradient based algorithms we consider, can be an interesting independent research in the future. There is in general no guarantee of algorithm convergence for the proximal gradient and iterated filtering algorithms applied to solve non-convex problems. Recently, *Bolte et al.* (2006); *Attouch and Bolte* (2009); *Attouch et al.* (2010, 2013) have proposed an framework of analyzing proximal gradient algorithms solving for possibly non-convex optimization problems, based on a Kurdyka-Lojasiewicz (K-L) property of the loss function, which we will use and extend to a case involving stochastically perturbed gradient. We observe that this analysis can be adapted in iterated filtering algorithm to show similar convergence results as in proximal gradi-

ent based algorithms, the difference amounts to distinct stochastic approximations for the gradient step. We will also consider using a misspecified model to approximate the original model, and use a usual high dimensional logistic regression convex problem as a surrogate to the original non-convex problem, we elaborated this in chapter III.

In the intractable log-likelihood aspect, computation of the log-likelihood function and its gradient in GLMMs is a notorious challenge even in low dimensional cases (*Jiang* (2007)). Except only in the case of linear mixed effect models, the GLMMs' negative log-likelihood functions are in general intractable integrations. There were various numerical integration techniques applicable to this problem, for instance the Laplace method applied in *Schelldorfer et al.* (2014), on which their GLMMLasso algorithm was built. However, it is known that when the dimension of the intractable integration is high, numerical techniques like Laplace approximation typically break, and Monte Carlo methods are possibly unavoidable in these situations. *Atchadé et al.* (2017) has demonstrated in an numerical example using a Markov chain Monte Carlo based algorithm fits a logistic mixed effect model very well. Iterated filtering is another stochastic algorithm which in many cases uses only simple importance sampling Monte Carlo to effectively approximate the intractable integration we have. We will develop algorithms using different Monte Carlo techniques to build these stochastic algorithms in chapters II and IV. We will also develop deterministic approximate based benchmark algorithm in chapter II. In chapter III, we will use a convex and tractable problem as a surrogate of the original intractable problem, and use highly efficient algorithms like *glmnet* to solve the surrogate problem, this helps us to bypass the computation of intractable integrations. Exsisting softwares exist to fit the mixed effects logistic regression models, among them lme4 in R, NLMIXED

in SAS are applicable for low dimensional GLMMs, while in high dimensional cases, the glmmixedlasso *Schelldorfer et al.* (2014) available from R-Forge is a recent development.

Our contribution in this dissertation can be described in three aspects: methodology, theory, and application.

For methodology contribution, we have first proposed and analyzed the stochastic proximal gradient algorithm, which applies the Polya-Gamma MCMC sampler to approximate the loss gradient. In addition, we have developed and analyzed a deterministic approximate algorithm based on a second order Taylor approximation of the conditional log-likelihood, these two algorithms in chapters II are both solving non-convex and non-smooth optimization problems. The stochastic proximal gradient algorithm does exact likelihood based inference while the second order approximate algorithm is solving an approximate problem. We have also developed and analyzed a "fixed effect approximate" algorithm in chapter III, which solves a high dimensional logistic regression model as a convex surrogate to the original non-convex problem. In chapter IV, we have explored and applied a block coordinate update version of iterated filtering algorithm to also exactly solve the original model. we have demonstrated the performance of all the developed algorithms in numerical studies and have applied some to a real data analysis.

Theoretically, we have analyzed the convergence behavior of the stochastic proximal gradient algorithm in the non-convex and non-smooth setting for the first time to our knowledge. We have extended the Kurdyka-Łojasiewicz (K-L) property based technique to incorporate stochastic perturbations in the updates to analyze the con-

vergence behavior of the stochastic proximal gradient algorithm. The algorithm convergence analysis for deterministic approximation algorithm is done by adapting arguments of *Attouch and Bolte* (2009); *Attouch et al.* (2010). The fixed effect approximation algorithm is solving a convex surrogate of the original non-convex optimization problem. We ask the question of how close its solution is to the true fixed effects parameters of the original non-convex problem, and we answer it by deriving a high dimensional non-asymptotic estimation error bound between the solution and the truth with high probability. This is done for the first time in high dimensional mixed effect logistic regression model to our knowledge. For our block update version of iterated filtering algorithm in chapter IV, we have related this algorithm closer to well-known stochastic gradient methods like in those of *Atchadé et al.* (2017). These new connections allow us to derive sharper algorithm convergence results than those of *Ionides et al.* (2011), assuming strong convexity of objective function. We also point out, without detailed proof, that the same technique used in convergence analysis of stochastic proximal gradient algorithm can be applied to analyze the convergence behavior of iterated filtering algorithms applied to possibly non-convex problems in chapter IV, and reach the same convergence analysis conclusion.

For applications of our developed algorithms and corresponding theory, we have conducted numerous simulation studies of our algorithms to solve the regularized maximum likelihood estimation problem in high dimensional mixed effects logistic regression model. We have demonstrated the effectiveness of our algorithms in different numerical scenarios and designs with comparions among themselves. Futher, we have applied our algorithms to analyze a well known breast cancer data set (*van Vliet et al.* (2008); *van't Veer et al.* (2002); *Vijver et al.* (2002)) modeled by high

dimensional mixed effect logistic regression. In this study, the distant metastasis within five years event is the binary response, which is a widely used indicator of breast cancer survival. The collected gene expression information and a few other clinical variables are the fixed effects. The goal of the study is to find a gene set which is most predictive of distant metastases within 5 years. For the mixed effects model applied to this data, we have constructed the random effects which take the relatedness of the individual gene information to form their variance-covariance matrix. In the end, we have selected a few genes (among thousands of candidate genes) as prognosis predictors of the response. Our gene set has novelties in gene discovery compared with different published findings (which have very limited overlap among themselves). Our potential prognostic gene discoveries may be useful for the clinical scientists to consider for their future studies of breast cancer.

The rest of this dissertation is organized as follows. We will present the development and analysis of our stochastic and deterministic approximate algorithms in chapter II, numerical study of the high dimensional mixed effects logistic regression and its application in a breast cancer real data analysis will be presented in chapter II too. In chapter III, we will propose the fixed effects approximation to the mixed effects model and a corresponding algorithm. In this chapter, we derive the high dimensional statistical estimation error bound of the algorithm solution with respect to the fixed effects parameters of the true model. We conduct numerical simulation study of this algorithm, and compare its solution to those of the algorithms in chapter II. We also apply this algorithm to the breast cancer data analysis. Next, in chapter IV, we devise a different algorithm based on iterated filtering algorithms, which is related to the proximal gradient algorithms we have developed before. We give a broad presentation of the iterated filtering algorithm and relate it to other stochastic

gradient methods. We derive a property of iterated filtering algorithm concerning its closeness to the proximal map, and derive a convergence result of the algorithm solving for strongly convex objective function. We demonstrate its estimation performance in numerical studies, with comparison with those of other algorithms we have developed in previous chapters.

# Chapter II

# Proximal Gradient Algorithms for Logistic Mixed Effects Regression Models

## 2.1 Introduction

In this chapter, we will deal with a class of optimization problems and develop a stochastic proximal gradient algorithm with other benchmark algorithms to solve the problem. The objective functions of such problems are composite functions of a generally non-convex loss function and a non-smooth component acting as a regularization. Moreover, the non-convex loss function we consider could involve analytically intractable integration that are also numerically challenging to approximate.

The non-convex and non-smooth problem we consider and their corresponding algorithms we will develop are different from those non-convex problems we have reviewed in the general introduction, mainly in three aspects. Firstly, in our case, the loss function, instead of the regularization function, is the non-convex component of the composite objective function; secondly, the non-convexity we face is not well structured like in biconvex problems where alternating direction methods or other

methods are known to be readily useful; the third and the most distinctive difference in our problem is that we need to develop specific strategy in the algorithm to deal with a non-convex function that involves intractable integration, which is challenging to approximate in the first place, and not easy to close the approximation gap along the algorithm updates. Our contribution in these aspects will be discussed toward the end of the introduction.

A typical application of these optimization problems can be fitting the high dimensional mixed effect generalized linear models, except in the case of usual linear model with Gaussian errors, which degenerates to a convex high dimensional estimation problem. We will use the high dimensional mixed effect logistic regression model as an example of the problems we consider in this chapter. The wide and useful applications of generalized mixed effect models have already been mentioned in the general introduction.

To deal with composite objective functions with intractable and non-smooth components, various algorithms have been proposed. *Nemirovski et al.* (2008); *Duchi et al.* (2012); *Lan* (2012); *Juditsky and Nemirovski* (2012a,b) have focused on stochastic sub-gradient and mirror descent algorithms. Others like *Combettes and Wajs* (2005); *Hu et al.* (2009); *Xiao* (2010); *Juditsky and Nemirovski* (2012a,b) have developed algorithms based on proximal operators to exploit the smoothness of the loss function and properties of the penalty component in the objective function. However, these algorithms have been studied only in the case of convex composite objective functions.

The algorithms we consider are developed based on proximal gradient algo-

rithms, for which we refer to *Beck and Teboulle* (2010); *Combettes and Pesquet* (2015b) or others for literature review and additional references). We will introduce relevant concepts about the proximal map and proximal gradient algorithm later in this chapter. Based on the proximal gradient algorithms, we have developed a stochastic proximal gradient algorithm, and a second order deterministic approximate algorithm as well. While the stochastic algorithm aims to exactly solve the problem, the deterministic algorithm solves the problem approximately thus introduces bias, especially when the dimension of integration involved in the objective function is high. Our deterministic algorithm can be seen as an approximation of the Laplace's method of intractable integration. In this sense ours is of the same spirit to the "GLMMLasso" algorithm developed by *Schelldorfer et al.* (2014)

In terms of algorithm convergence analysis, *Combettes and Wajs* (2005); *Rosasco et al.* (2014); *Nitanda* (2014); *Xiao and Zhang* (2014) have analyzed the proximal algorithm with perturbations. These methods and analysis again only apply to convex or strongly convex objective functions. *Beck and Teboulle* (2010) has analyzed the case with non-convex objective functions, but only for exact proximal operator without stochastic perturbation. The closest development we have seen so far is *Atchadé et al.* (2017). They have proposed and analyzed a similar stochastic proximal gradient algorithm, however their convergence analysis is limited to convex composite objective functions.

Our technical contribution in this chapter is that we have proposed and analyzed both the stochastic proximal gradient algorithm and our deterministic approximation algorithm solving for problem with *non-convex, non-smooth* objective function, where the loss function involves an *intractable integration*. In the theoreti-

cal convergence analysis, we have extended the Kurdyka-Łojasiewicz (K-L) property based technique to incorporate stochastic perturbation to analyze convergence of the stochastic proximal gradient algorithm. The Kurdyka-Łojasiewicz (K-L) inequality is a geometric property of a function, which provides sufficient curvature for the function at its stationary points.

In addition to the technical contribution, we have applied our algorithms - the stochastic (MCMC) proximal gradient algorithm for exact likelihood inference, second order approximate algorithm for approximate inference for high dimensional mixed effect logistic regression model, with comparison to glmnet solutions ignoring the random effects. We have demonstrated that considering the random effects in the data leads to clearly better estimation performance, and that the MCMC based stochastic algorithm performs better than the deterministic algorithm in many cases, with comparable running time.

As an overview of what follows in this chapter, in section 2.2, we formulate the optimization problem we will solve in this chapter, and we will introduce the high dimensional mixed effect generalized linear model as a typical example for the optimization problem we aim to solve. Then in section 2.3 we will develop the stochastic proximal gradient algorithm and a deterministic approximate algorithm we use to solve our optimization problem. Next, we carry out our non-convex algorithm convergence analysis of both the stochastic and deterministic algorithms in section 2.4. Extensive simulation study will be carried out in section 2.5 to numerically demonstrate the performance of our algorithms and their theoretical properties, we will also use the high dimensional mixed effect logistic regression model as our numerical example and study it in detail.

## 2.2 The Optimization Problem

This chapter deals with the optimization problem

$$\textbf{(P)} \quad \min_{\beta \in \mathbb{R}^p} F(\beta) \quad \text{with } F(\beta) = f(\beta) + g(\beta) \tag{2.1}$$

Where $f$ is a Lipschitz continuously differentiable function, possibly non-convex, and $g$ is a possibly non-smooth convex function.

**Assumption II.1.** *The function $f : \mathbb{R}^p \to \mathbb{R}$ is continuously differentiable on $\mathbb{R}^p$ and there exists a finite non-negative constant $L$ such that, for all $\beta, \beta' \in \mathbb{R}^p$,*

$$\|\nabla f(\beta) - \nabla f(\beta')\| \leq L \|\beta - \beta'\|,$$

*where $\nabla f$ denotes the gradient of $f$. The function $g : \mathbb{R}^p \to [0, +\infty]$ is convex, not identically $\infty$ (proper), and lower semi-continuous.*

We denote by $\Theta$ the domain of $g : \Theta \stackrel{\text{def}}{=} \{\beta \in \mathbb{R}^p : g(\beta) < \infty\}$.

**Assumption II.2.** *The set $argmin_{\beta \in \Theta} F(\beta)$ is a non empty subset of $\Theta$, and $\inf_{\beta \in \Theta} F(\beta) > -\infty$. With out loss of generality, we take $\inf_{\beta \in \Theta} F(\beta) = 0$.*

Fitting high dimensional mixed effect generalized linear models gives a typical example of problem **(P)**. In this case, $f$ is the negative log-likelihood function of the mixed effect generalized linear models, as the following when $y_i \in \{0, 1\}$

$$f(\beta) = -\log \int_{\mathbb{R}^q} \exp\left\{ \sum_{i=1}^{n} \left( y_i(x_i'\beta + z_i'u_i) - \log(1 + x_i'\beta + z_i'u_i) \right) \right\} \pi(du) \tag{2.2}$$

where $\pi(u)$ is a $q$ dimensional Gaussian density of random effects $u$ in our problem, other quantities are the same as introduced in (1.3) in chapter I

In general, the loss function we consider in this chapter has the following form,

$$f(\beta) = -\log \int_{\mathbb{R}^q} \exp \ell(\beta, u)\pi(du) \tag{2.3}$$

where $\pi(u)$ is a $q$ dimensional Gaussian density, $\ell(\beta, u)$ can be thought of as the conditional log-likelihood functions of the mixed effects logistic regressions and its quadratic approximation later in this chapter. It is not hard to show that so long as gradient and Hessian of $\ell$ are uniformly bounded, then $f$ will satisfy Assumption II.1, and this is the case for mixed effect logistic regression, and its quadratic approximation we will use later.

$g$ is a regularization function that imposes structure to the solution, say sparsity when $p$ is larger than $n$. The elastic net function is widely used as a sparsity inducing regularization function.

$$g(\beta) = \lambda \left( \frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \tag{2.4}$$

## 2.3 Algorithms

To solve for problem **(P)** in (2.1), we propose and study two algorithms. Namely, the stochastic perturbed proximal gradient algorithm and proximal gradient algorithm for a deterministic approximation of the objective function.

### 2.3.1 Stochastic Proximal Gradient Algorithm

Proximal algorithms are well established optimization algorithms for dealing with non-smooth objective functions and composite optimization problems like problem **(P)** in (2.1). See *Beck and Teboulle* (2010); Parikh and Boyd (2013)*Parikh and Boyd* (2013a); *Juditsky and Nemirovski* (2012a,b). In this paper, we focuse on the proximal gradient algorithm(see also *Nesterov* (2004)) and its perturbed version first proposed by *Atchadé et al.* (2017): the gradient of $f(\beta)$ at the current estimate $\beta^k$ is replaced by a Monte Carlo approximation $H_{k+1}$. Besides the perturbed gradient, our objective function $f(\beta)+g(\beta)$ is possibly non-convex and non-smooth. We typically assumes the proximity operator of $g(\beta)$ can be easily computed.

**Algorithm 1** stochastic proximal gradient algorithm

For $k \geq 1$, given the current $\beta^{(k)}$, repeat until convergence:

1. Compute an approximation of $\nabla f(\beta^{(k)})$ as $H_{k+1}$;
2. Compute $\beta^{(k+1)} = \text{Prox}_{\gamma g}\left(\beta^{(k)} - \gamma H_{k+1}\right)$.

In general, $\text{Prox}_{\gamma g}\left(\beta - \gamma \nabla f(\beta)\right)$ denotes the proximal operator of a function $\gamma \cdot g(\cdot)$ applied to the vector $\beta - \gamma \nabla f(\beta)$, defined as:

$$\text{Prox}_{\gamma g}\left(\beta - \gamma \nabla f(\beta)\right) = \underset{\vartheta \in \text{Dom}(g)}{\arg\min} \left\{ \langle \nabla f(\beta), \vartheta - \beta \rangle + \frac{1}{2\gamma} \left\| \vartheta - \beta \right\|^2 + g(\vartheta) \right\} \tag{2.5}$$

For the elastic regularization function $g$ in (2.4),

$$\text{Prox}_{\gamma g}(\vartheta) = \begin{cases} \frac{\vartheta_j - \gamma\lambda\alpha}{1+\gamma\lambda(1-\alpha)}, & \text{if } \vartheta_j \geq \gamma\lambda\alpha, \\ \frac{\vartheta_j + \gamma\lambda\alpha}{1+\gamma\lambda(1-\alpha)}, & \text{if } \vartheta_j \leq -\gamma\lambda\alpha, \\ 0, & \text{if } \vartheta_j \in (-\gamma\lambda\alpha, \gamma\lambda\alpha) \end{cases}$$

We now derive the Markov chain Monte Carlo algorithm we have used to approximate the gradient $\nabla f(\beta)$. The algorithm has utilized the Polya-Gamma distribution, it is proposed in *Polson et al.* (2012) based on data augmentation strategy. We first describe the Gibbs sampler below:

### 2.3.2 Deterministic Approximate Algorithm

Instead of directly solving the original problem **(P)** in (2.1), one can solve an approximation problem

$$\widetilde{\mathbf{(P)}} \quad \min_{\beta \in \mathbb{R}^p} \widetilde{F}(\beta) \quad \text{with } \widetilde{F} = \widetilde{f} + g \tag{2.6}$$

where $\widetilde{f}$ is a deterministic approximation of the intractable loss function $f$ we will derive.

Depending on the functional form of $\widetilde{f}$, various algorithms can be applied to solve problem $\widetilde{\mathbf{(P)}}$ in (2.6), we apply the proximal gradient algorithm as it performs well and enables a comparable development of convergence theory with stochastic proximal gradient algorithm.

**Algorithm 2** Second Order Approximate Algorithm

For $k \geq 1$, given the current $\beta^{(k)}$, repeat until convergence:

1. Compute $\nabla \widetilde{f}(\beta^{(k)})$;
2. Compute $\beta^{(k+1)} = \text{Prox}_{\gamma g}\left(\beta^{(k)} - \gamma \nabla \widetilde{f}(\beta^{(k)})\right)$.

Next we investigate the convergence of the proposed algorithms for solving in-

tractable, possibly non convex and non smooth composite objective functions.

We assume Assumption (II.1) also applies to $\nabla \widetilde{f}(\beta)$, that is there exists a finite non-negative constant $\widetilde{L}$ such that, for all $\beta, \beta' \in \mathbb{R}^p$,

$$\left\| \nabla \widetilde{f}(\beta) - \nabla \widetilde{f}(\beta') \right\|_2 \leq \widetilde{L} \left\| \beta - \beta' \right\|_2 \tag{2.7}$$

This can be routinely verified to be true from our discussion of (2.3) in 2.2.

## 2.4 Algorithm Convergence

We first show that Algorithms (1) and (2) solving for respective problems (2.1) and (2.6) always find the stationary points of the objective functions so long as the objective functions satisfy Assumptions (II.1) and (II.2). To be precise, *the limiting points of iterative updates* $\{\beta_k\}_{k\in\mathbb{N}}$ *are all stationary points of* $F(\beta)$ *or* $\widetilde{F}(\beta)$ *respectively, such that if* $\beta_*$ *is a limiting point of* $\{\beta_k\}_{k\in\mathbb{N}}$*, then* $\beta_* \in \mathcal{L} = \{\beta \in \mathbb{R}^p : 0 \in \nabla f(\beta) + \partial g(\beta)\}$*, or* $\widetilde{\mathcal{L}} = \left\{\beta \in \mathbb{R}^p : 0 \in \nabla \widetilde{f}(\beta) + \partial g(\beta)\right\}$, along with this we show that $\lim_k \widetilde{F}(\beta_k) = \widetilde{F}(\beta_*)$ under mild assumptions.

For algorithm convergence, that is, concerning $\lim_k \beta_k$, we need to further characterize the stationary points of the objective functions. In the case of mixed effect Gaussian linear regression, proximal gradient algorithm (2) solving for **(P)** is enough since approximation of objective function is not necessary, objective function convexity at the stationary points is enough to guarantee convergence of Algorithm (2) (*Atchadé et al.* (2017)), while in other cases of mixed effect generalized linear regression, since the objective function is possibly non-convex, and the curvature at its stationary points is difficult to characterize due to its non-smooth part, we discuss in detail for different problems and algorithms. Convergence of Algorithm (2) ap-

plied to problem $(\widetilde{\mathbf{P}})$ always holds under assumptions (II.1) and (II.2), we propose to utilize an approach developed in *Attouch and Bolte* (2009), which is based on the *Kurdyka-Łojasiewicz property* of the possibly non-convex objective function.

### 2.4.1 Stochastic Proximal Gradient Algorithm Convergence Analysis

We will use a proposition (*Bauschke and Combettes* (2011)) concerning proximal operators for convex functions:

**Proposition II.3.** *For function $g$ assuming A2, with*

$$Prox_{\gamma g}(\beta) \triangleq \arg\min_{\vartheta \in \Theta} \left[ g(\vartheta) + \frac{1}{2\gamma} \|\vartheta - \beta\|_2^2 \right]$$

,

*(i) Define function*

$$g_\gamma(\beta) = \min_{\vartheta \in \Theta} \left[ g(\vartheta) + \frac{1}{2\gamma} \|\vartheta - \beta\|_2^2 \right]$$

*$g_\gamma(\beta)$ is differentiable everywhere and*

$$\nabla g_\gamma(\beta) = \frac{1}{\gamma} \left( \beta - Prox_\gamma^g(\beta) \right). \tag{2.8}$$

*Furthermore $\beta \mapsto \nabla g_\gamma(\beta)$ is Lipschitz with Lipschitz constant $\frac{1}{\gamma}$.*

*(ii) For $u \in \Theta$, $\nabla g_\gamma(u) \in \partial g(Prox_\gamma^g(u))$. This means that for all $\vartheta \in \Theta$,*

$$g(\vartheta) \geq g(Prox_\gamma^g(u)) + \left\langle \frac{1}{\gamma} \left( u - Prox_\gamma^g(u) \right), \vartheta - Prox_\gamma^g(u) \right\rangle \tag{2.9}$$

We will then establish several lemmas.

**Lemma II.4** (*Atchadé et al.* (2017)). *Let* $\{\nu_k, k \in \mathbb{N}\}$ *and* $\{\chi_k, k \in \mathbb{N}\}$ *be non-negative sequences and* $\{\xi_k, k \in \mathbb{N}\}$ *be such that* $\sum_k \xi_k$ *exists. If for any* $k \geq 0$,

$$\nu_{k+1} \leq \nu_k - \chi_{k+1} + \xi_{k+1}$$

*then* $\sum_k \chi_k < \infty$ *and* $\lim_k \nu_k$ *exists.*

**Lemma II.5.** *Suppose Assumption II.1 holds. Let* $\{\beta_k, k \in \mathbb{N}\}$ *be given by Algorithm (1) with non increasing step size* $\{\gamma_k, k \in \mathbb{N}\}$. *The sequence* $\{F(\beta_k)\}_{k\in\mathbb{N}}$ *satisfies*

$$F(\beta_k) - F(\beta_{k+1}) \geq \frac{1}{2\gamma_{k+1}} \|\beta_{k+1} - \beta_k\|^2 + \langle \beta_{k+1} - \beta_k, \eta_{k+1} \rangle \tag{2.10}$$

**Lemma II.6.** *Suppose Assumptions II.1 and II.2 hold. Let* $\eta_{k+1} = H_{k+1} - \nabla f(\beta_k), k \geq 1$ *denote the Monte Carlo gradient approximation error in (1). If approximation error* $\eta_k$ *satisfies*

$$\sum_k \langle \beta_{k+1} - \beta_k, \eta_{k+1} \rangle < \infty \quad a.s. \tag{2.11}$$

*Then for the same sequence* $\{\beta_k, k \in \mathbb{N}\}$ *in Lemma II.5, the following hold.*

*(i)* $\sum_k \|\beta_{k+1} - \beta_k\|^2 < \infty$ *and* $\lim_{k\to\infty} \|\beta_{k+1} - \beta_k\| = 0$ *a.s.*

*(ii)* $\lim_k \gamma_{k+1} F(\beta_{k+1})$ *exists.*

**Lemma II.7.** *Suppose Assumption II.1 holds. Let* $\{\beta_k, k \in \mathbb{N}\}$ *be given by Algorithm (1) with non increasing step size* $\{\gamma_k, k \in \mathbb{N}\}$. *The following results hold.*

*(i)* $A_k := \frac{1}{\gamma_k}(\beta_{k-1} - \beta_k) + \nabla f(\beta_k) - \nabla f(\beta_{k-1}) - \eta_k$. *Then* $A_k \in \partial F(\beta_k)$

*(ii)* $\|A_k\| \leq \frac{2}{\gamma_k} \|\beta_{k-1} - \beta_k\| + \|\eta_k\|$

Now we are ready to characterize the limit point set of the sequence $\{\beta_k, k \in \mathbb{N}\}$ produced by stochastic proximal gradient algorithm Algorithm(1). We show that

the limit point(s) of $\{\beta_k, k \in \mathbb{N}\}$ are stationary point(s) of the objective function $F(\beta) = f(\beta) + g(\beta)$ in problem **(P)**, along with other properties.

**Theorem II.8.** *Suppose that Assumptions II.1 and II.2 hold. Denote $\omega(\beta_0)$ as the limit point set of the sequence $\{\beta_k, k \in \mathbb{N}\}$ which is assumed to be bounded, be given by Algorithm with non increasing step size $\{\gamma_k, k \in \mathbb{N}\}$. The following assertions hold.*

*(i)* $\emptyset \neq \omega(\beta_0) \subset \mathcal{L}$, *defined as* $\mathcal{L} \triangleq \{\beta \in \Theta : 0 \in \nabla f(\beta) + \partial g(\beta)\}$, *the critical point set of F.*

*(ii) We have*

$$\lim_{k\to\infty} dist\left(\beta_k, \omega\left(\beta_0\right)\right) = 0 \tag{2.12}$$

*(iii)* $\omega\left(\beta_0\right)$ *is a nonempty, compact and connected set.*

Proofs for the above are presented in the proof section at the end of this chapter.

Now that our algorithms always find the stationary points of the objective functions, we aim to establish the convergence of $\{\beta_k\}_{k\in\mathbb{N}}$ and it converges to a local minimum for possibly non-convex objective functions. For this purpose, we explore in two aspects.

Firstly, both algorithms produce $\{\beta_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \|\beta_{k+1} - \beta_k\| = 0$. This implies that either the sequence $\{\beta_k\}_{k\in\mathbb{N}}$ converges to $\beta_* \in \mathcal{L}$, or the set of limit points of this sequence forms a continuum, and the sequence does not converge. It also implies that if $\{\beta_k\}_{k\in\mathbb{N}}$ has an isolated limit point $\beta_*$, then $\lim_k \beta_k = \beta_*$. In addition, given $\lim_{k\to\infty} F(\beta_k) = F(\beta_*)$ (Theorem(II.12) $(iv)$), it is not hard to see that for any given $c_* \in \mathbb{R}$, if $\{\beta_* \in \mathcal{L} : F(\beta_*) = c_*\}$ is a countable set we denote $\mathcal{A}_*$,

then $\{\beta_k\}_{k\in\mathbb{N}}$ will converge to one point in $\mathcal{A}_*$. Ofttimes, $\mathcal{A}_*$ is finite, even has only one point.

Secondly, what kind of stationary points do the algorithms find? Saddle points and local maximums are undesirable, we will check the conditions under which local minimums are discovered.

### 2.4.2 Second Order Approximate Algorithm Convergence

We begin with some lemmas. Similar but much simpler than Lemma(II.5), we have the following lemma:

**Lemma II.9.** *Suppose Assumptions II.1 and II.2 hold. Let $\{\beta_k, k\in\mathbb{N}\}$ be given by Algorithm(2) with non increasing positive step size $\{\gamma_k, k\in\mathbb{N}\}$. We have descent property for the objective function, that the sequence $\left\{\widetilde{F}(\beta_k)\right\}_{k\in\mathbb{N}}$ is decreasing and satisfies*

$$\widetilde{F}(\beta_k) - \widetilde{F}(\beta_{k+1}) \geq \frac{1}{2\gamma_{k+1}} \|\beta_{k+1} - \beta_k\|_2^2 . \tag{2.13}$$

**Lemma II.10.** *Suppose Assumptions II.1 and II.2 hold. Let $\{\beta_k, k\in\mathbb{N}\}$ be given by Algorithm(2). We have the square summable result, $\sum_k \|\beta_{k+1} - \beta_k\|_2^2 \leq \infty$ and $\lim_k \|\beta_{k+1} - \beta_k\|_2 = 0$*

**Lemma II.11.** *Suppose Assumptions II.1 and II.2 holds. Let $\{\beta_k, k\in\mathbb{N}\}$ be given by Algorithm(2). The following results hold.*

*(i) Let $\widetilde{A}_k := \frac{1}{\gamma_k}(\beta_{k-1} - \beta_k) + \nabla\widetilde{f}(\beta_k) - \nabla\widetilde{f}(\beta_{k-1})$. Then $\widetilde{A}_k \in \partial\widetilde{F}(\beta_k)$*

*(ii) $\left\|\widetilde{A}_k\right\|_2 \leq \frac{2}{\gamma_k}\|\beta_{k-1} - \beta_k\|_2$*

**Theorem II.12** (Properties of the limit point set of the sequence $\{\beta_k, k\in\mathbb{N}\}$)**.** *Suppose that Assumptions II.1 and II.2 holds. Denote $\widetilde{\omega}(\beta_0)$ as the set of accumu-*

*lation points of the sequence* $\{\beta_k, k \in \mathbb{N}\}$ *generated by Algorithm*(2) *and assumed to be bounded. We have the following assertions hold.*

*(i)* $\emptyset \neq \widetilde{\omega}(\beta_0) \subset \widetilde{\mathcal{L}}$, *defined as* $\widetilde{\mathcal{L}} \triangleq \left\{\beta \in \Theta : 0 \in \nabla \widetilde{f}(\beta) + \partial g(\beta)\right\}$, *the critical point set of* $\widetilde{F}$.

*(ii) We have*

$$\lim_{k\to\infty} dist\left(\beta_k, \widetilde{\omega}\left(\beta_0\right)\right) = 0 \tag{2.14}$$

*(iii)* $\widetilde{\omega}\left(\beta_0\right)$ *is a nonempty, compact and connected set.*

*(iv) The objective function* $\widetilde{F}$ *is finite and constant* $(:= \widetilde{F_-})$ *on* $\widetilde{\omega}\left(\beta_0\right)$, *and for all* $\bar{\beta} \in \widetilde{\omega}\left(\beta_0\right)$,

$$\lim_{k\to\infty} \widetilde{F}(\beta_k) = \widetilde{F}(\bar{\beta}) = \widetilde{F_-}. \tag{2.15}$$

We derive algorithm convergence in this case by extending an approach developed based on the Kurdyka-Łojasiewicz property (K-L-property) of the possibly non-convex objective function (*Attouch et al.* (2010)). We further derive the convergence rate in this case.

Since the geometric concept of *Kurdyka-Łojasiewicz property* is not a very widely known in statistics community, we will take a digression to introduce this concept in the following, and derive useful properties from this concept to pave the way of proving convergence of the deterministic approximate algorithm in Theorem 5.

#### 2.4.2.1 Kurdyka-Łojasiewicz (K-L) property

First we introduce the "KL property" characterizing the curvature of a possibly non-convex function. As an motivating example, let us consider a toy triangular

wave function as our objective.

$$z(t) = \left(t - 2\left\lfloor \frac{t}{2} + \frac{1}{2} \right\rfloor\right)(-1)^{\left\lfloor \frac{t}{2}+\frac{1}{2} \right\rfloor}, x \in \mathbb{R} \tag{2.16}$$

This function is non-convex and non-smooth, it is convex in a neighborhood of

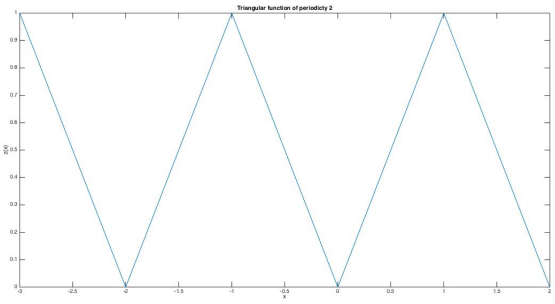


Figure 2.1: Triangular wave function $z(t)$

the local minimum points, but not strictly convex and thus not strongly convex. However, it is easy to see that for any $\bar{t} \in \mathbb{R}$, the function has a "KL-property" that

$$|z(t) - z(\bar{t})|^r \leq \frac{1}{4^r}, \text{ for all } r \in [0, 1) \text{ and } t \in \left\{t : |t - \bar{t}| < \frac{1}{4} \text{ and } 0 < z(t) - z(\bar{t}) < \frac{1}{4}\right\}$$

It can be shown that piecewise linear functions are all KL-functions, regardless of their non-convexity and non-smoothness.

Such property around its local minimum points, which in the above example are the even integer points, are of our interest, since such geometric property of the objective function together with algorithmic properties established in section (**??**) would imply algorithm convergence as we will prove.

To further motivate the role of KL-property in optimizing a function, consider a

hypothetical toy function $\zeta(x), x \in \mathbb{R}$, $\zeta(x)$ is twice continuously differentiable such that $\zeta^{''}(x) > 0$ for all $x$ in $O_* = \{x \in \mathbb{R} : \zeta'(x) = 0\}$.

Taylor expansion of $\zeta(x)$ around any $x \in O_*$ gives

$$\zeta(y) - \zeta(x) = \frac{1}{2}\left(\zeta^{''}(x) + o(1)\right)(y-x)^2 \tag{2.17}$$

with

$$\left(\zeta^{''}(x) + o(1)\right)^2 = \frac{\zeta'(y)^2}{(y-x)^2} \tag{2.18}$$

since $\zeta'(x) = 0$. Combine (2.17) we get

$$|\zeta(y) - \zeta(x)| = \frac{1}{2}\frac{\zeta'(y)^2}{|\zeta''(x) + o(1)|} \tag{2.19}$$

Then there exists some positive $\eta$, for all $y$ satisfying $|x-y| < \eta$, there exists some positive constant $C$ such that

$$|\zeta(y) - \zeta(x)|^{\frac{1}{2}} \leq C\,|\zeta'(y)| \tag{2.20}$$

Now we give a formal definition of Kurdyka-Łojasiewicz property with some remarks.

**Definition II.13.** *(*Kurdyka-Łojasiewicz property*) Let* $F : \mathbb{R}^d \to (-\infty, +\infty]$ *be proper and lower semi-continuous.*

*(i) The function* $F$ *is said to have the* Kurdyka-Łojasiewicz (KL) property *at* $\bar{x} \in \operatorname{dom} \partial F := \left\{x \in \mathbb{R}^d : \partial F(x) \neq \emptyset\right\}$ *if there exist* $\eta \in (0, \infty]$, $r \in [0, 1)$, $c > 0$, *a neighborhood* $\mathcal{B}$ *of* $\bar{x}$, *such that for all*

$$x \in \mathcal{B} \cap \{x : F(\bar{x}) < F(x) < F(\bar{x}) + \eta\},$$

*the following inequality holds for all* $x^* \in \partial F(x)$

$$|F(x) - F(\bar{x})|^r \leq c \|x^*\|_2 \tag{2.21}$$

*(ii) If* $F$ *satisfy the K-L property at each point of* dom $\partial F$ *then* $F$ *is called a* K-L function *(we will later say "a function is K-L" meaning a function is a K-L function).*

*Remark* II.14. (on Definition 1) *K-L property* characterize a geometrical feature of the function. As *Attouch et al.* (2013) pointed out *K-L property* does not pertain a function's convexity or smoothness. On another hand, while not all convex functions are *K-L functions*, strongly convex functions are *K-L*, and while not all $\mathcal{C}^\infty$ functions are *K-L*, smooth functions whose Taylor series converges to the function in some neighborhood for every point in its domain, or *real analytic functions* are all *K-L*. *K-L property* has useful consequences in the study of first-order descent methods.(see *Attouch et al.* (2013)).

We refer the definition of real analytic functions of several variables to *Kranz and Parks* (2002) *Definition 2.2.1*, and elementary properties to *Proposition 1.4.2* and *Proposition 2.2.2*.

**Proposition II.15.** *(*Proposition 1.4.2*) Let* $I$ *and* $J$ *be open intervals in* $\mathbb{R}$, $f : I \to J$ *and* $g : J \to \mathbb{R}$ *are both real analytic. Then* $g \circ f : I \to \mathbb{R}$ *is real analytic.*

**Proposition II.16.** *(*Proposition 2.2.2*) Let* $U, V \subseteq \mathbb{R}^m$ *be open. If* $f : U \to \mathbb{R}$ *and* $g : V \to \mathbb{R}$ *are real analytic, then* $f + g$, $f \cdot g$ *are real analytic on* $U \cap V$, *and* $f/g$ *is real analytic on* $U \cap V \cap \{x : g(x) \neq 0\}$.

We aim to show that the objective function in (2.6) is a *K-L* function. We show this in several steps.

**Proposition II.17.** $\widetilde{f}(\beta) = -\overset{\circ}{\ell}(\beta) - \frac{\sigma^2}{2} g(\beta)^T \left[I_q - \sigma^2 h(\beta)\right]^{-1} g(\beta) + \frac{1}{2} \log det\left(I_q - \sigma^2 h(\beta)\right)$ *in* (2.49) *is real analytic.*

*Proof.* Firstly, $-\overset{\circ}{\ell}(\beta) = \sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i \langle x_i, \beta\rangle\right)\right)$ is real analytic. This is because $\exp\left(-y_i \langle x_i, \beta\rangle\right)$ being a composition of elementary exponential and linear function, is real analytic, then $1 + \exp\left(-y_i \langle x_i, \beta\rangle\right)$ is real analytic. Since $\log(x)$ is real analytic on $x \in (0, +\infty)$, $\log\left(1 + \exp\left(-y_i \langle x_i, \beta\rangle\right)\right)$ is real analytic by composition proposition (II.15). Now $-\overset{\circ}{\ell}(\beta)$ is a real analytic function by proposition (II.16).

Secondly, since $s_i(\beta) = \frac{1}{1+\exp(-y_i \langle x_i, \beta\rangle)}$ in (2.51) is real analytic by proposition (II.16), each element in the vector $g(\beta)$ in (2.51) and matrix $h(\beta)$ in (2.52) is a real analytic function of $\beta \in \mathbb{R}^d$ again by proposition (II.16).

Next, $h(\beta)$ is a diagonal matrix, since $Z$ is an orthogonal matrix, so

$$\begin{aligned} det\left(I_q - \sigma^2 h(\beta)\right) &= \left|I_q - \sigma^2 W_\beta\right| \\ &= \prod_{k=1}^{q} \left[1 - \sigma^2 s_i(\beta)(1 - s_i(\beta))\right] \end{aligned}$$

so $det\left(I_q - \sigma^2 h(\beta)\right)$ is an analytic function of $\beta$ by proposition (II.16).

Similarly,

$$\left[I_q - \sigma^2 h(\beta)\right]^{-1} = Z^T \left[I_q - \sigma^2 W_\beta\right]^{-1} Z$$

where $\left[I_q - \sigma^2 W_\beta\right]^{-1}$ is a $n \times n$ diagonal matrix with $i$th diagonal entry being

$$1/\left[1 - \sigma^2 s_i(\beta)(1 - s_i(\beta))\right]$$

.

Assume that random effect noise $\sigma^2$ satisfies that $\sigma^2 s_i(\beta)(1 - s_i(\beta)) < 1, \forall i =$

$1, 2, \ldots, n$, then $\det\left(I_q - \sigma^2 h(\beta)\right) > 0$, $\log\det\left(I_q - \sigma^2 h(\beta)\right)$ is real analytic by proposition (II.15), and each entry in $\left[I_q - \sigma^2 h(\beta)\right]^{-1}$ is real analytic function in $\beta$ by proposition (II.16).

Now it is easy to see after matrix multiplication $\frac{\sigma^2}{2} g(\beta)^T \left[I_q - \sigma^2 h(\beta)\right]^{-1} g(\beta)$ is real analytic in $\beta$ by proposition (II.16).

Finally, as a sum of real analytic functions,

$$\widetilde{f}(\beta) = -\mathring{\ell}(\beta) - \frac{\sigma^2}{2} g(\beta)^T \left[I_q - \sigma^2 h(\beta)\right]^{-1} g(\beta) + \frac{1}{2}\log\det\left(I_q - \sigma^2 h(\beta)\right)$$

in (2.49) is real analytic. □

So $\widetilde{f}(\beta)$ is a *K-L* function by *Attouch and Bolte* (2009)

**Proposition II.18.** *The sum of two* K-L *functions is* K-L.

*Proof.* Suppose $F_1$ and $F_2$ are two *K-L* functions. Specifically for any $\bar{x} \in \mathrm{dom}\partial F_1 \cap \mathrm{dom}\partial F_2$, there exist $\eta \in (0, \infty]$, a neighborhood $\mathcal{B} \subseteq \mathrm{dom}F_1 \cap \mathrm{dom}F_2$ of $\bar{x}$, such that for all

$$x \in \mathcal{B} \cap \{x : F_i(\bar{x}) < F_i(x) < F_i(\bar{x}) + \eta/2, i = 1, 2\}$$

the following inequalities hold for all $x_i^* \in \partial F_i(x), i = 1, 2$

$$\left|F_i\left(x\right) - F_i\left(\bar{x}\right)\right| \leq c_i \left\|x_i^*\right\|_2^{1/r_i} \tag{2.22}$$

Adding the above inequalities and by triangle inequality we have:

$$\left|F_1(x) + F_2(x) - F_1(\bar{x}) - F_2(\bar{x})\right| \leq c_1 \left\|x_1^*\right\|_2^{1/r_1} + c_2 \left\|x_2^*\right\|_2^{1/r_2} \tag{2.23}$$

Let $c = c_1 \vee c_2$, suppose w.o.l.g. $\|x_1^*\|_2 \geq \|x_2^*\|_2$, then we get

$$|F_1(x) + F_2(x) - F_1(\bar{x}) - F_2(\bar{x})| \leq c\left(\|x_1^*\|_2^{1/r_1} + c_2 \|x_1^*\|_2^{1/r_2}\right) \tag{2.24}$$

Since $v(x) = a^x, a > 0$ is a convex function and $\|x_1^*\|_2 \geq 0$, let $r = 2/(1/r_1 + 1/r_2) \in [0, 1)$, by Jensen's inequality we get

$$|F_1(x) + F_2(x) - F_1(\bar{x}) - F_2(\bar{x})| \leq 2c \|x_1^*\|_2^{1/r} \tag{2.25}$$

Rearranging terms we get

$$|F_1(x) + F_2(x) - F_1(\bar{x}) - F_2(\bar{x})|^r \leq C \|x_1^*\|_2 \tag{2.26}$$

for some constant $C > 0$. So $F_1 + F_2$ is a *K-L* function. □

*Remark* II.19. By mathematical induction, the sum of finitely many *K-L* functions is *K-L*.

**Proposition II.20.** *The objective function* $\widetilde{F}(\beta) = \widetilde{f}(\beta) + g(\beta), \beta \in \mathbb{R}^d$ *in problem* (2.6) *is a* K-L function.

*Proof.* Proposition (II.17) shows that $\widetilde{f}$ is real analytic function and thus is K-L function. The elastic net function

$$g(\beta) = \lambda\left(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right)$$

is a sum of the polynomial(quadratic) function $\frac{1-\alpha}{2}\|\beta\|_2^2$, which is real analytic, and the $\ell_1$ norm function $\alpha\|\beta\|_1$, which is itself a sum of $p$ absolute value functions $|\beta_j|, j = 1, 2, \ldots, p$. The *K-L property* of absolute value function is essentially established in the *triangular wave function* example (2.16). By proposition (II.18), $g(\beta)$

is a K-L function, and again by proposition (II.18) we conclude that $\widetilde{F}(\beta)$ is a K-L function. □

Theorem(II.12) established that the limit point set $\omega(\beta_0)$ of the sequence $\{\beta_k, k \in \mathbb{N}\}$ generated by Algorithm(2) is a non-empty, compact and connected set, and that the objective function $\widetilde{F}$ is a constant on $\omega(\beta_0)$. As a consequence, we will derive a *uniformized K-L property* used in the proof of the main convergence theorem.

**Lemma II.21** (Uniformize K-L property)**.** *Let $\Omega$ be compact, in addition to assumption A, assume $F$ is constant on $\Omega$ and satisfies K-L property (see Definition 2) at each point of $\Omega$. Then there exists $\delta \in (0, \infty], \epsilon > 0$, and $c > 0, r \in [0, 1)$ such that for all $\bar{u}$ in $\Omega$ and all $u$ in the following set:*

$$\{u \in \Theta : dist(u, \Omega) < \epsilon\} \cap \{u \in \Theta : F(\bar{u}) < F(u) < F(\bar{u}) + \delta\} \tag{2.27}$$

*one has,*

$$|F(u) - F(\bar{u})|^r \leq c \left\|u^{\#}\right\|, \text{ for all } u^{\#} \in \partial F(u) \tag{2.28}$$

*Proof.* Denote $\mu$ as the constant value of $F$ on $\Omega$. The compact set $\Omega$ can be covered by a finite number of open balls $B(u_i, \epsilon_i)$ with $u_i \in \Omega$, $i = 1, \ldots, p$ on which the K-L property holds, that is, for any $u \in B(u_i, \epsilon_i) \cap \{u \in \Theta : 0 < F(u) - \mu < \delta_i\}$ we have $|F(u) - F(u_i)|^{r_i} = |F(u) - \mu|^{r_i} \leq c_i \left\|u^{\#}\right\| \; \forall u^{\#} \in \partial F(u)$.

Choose $\epsilon > 0$ sufficiently small, so that

$$\{u \in \Theta : dist(u, \Omega) < \epsilon\} \subset \cup_{i=1}^{p} B(u_i, \epsilon_i) \tag{2.29}$$

Set $\delta = \min\{\delta_i, i = 1, \ldots, p\} > 0$ and

$c = \max\{c_i, i = 1, \ldots, p\}$,

$r = \max\{r_i, i = 1, \ldots, p\}$

we get that for all $u \in \{u \in \Theta : dist(u, \Omega) < \epsilon\} \cap \{u \in \Theta : 0 < F(u) - F(\bar{u}) < \delta\}$,

$$|F(u) - F(\bar{u})|^r \leq c \left\|u^{\#}\right\|, \text{ for all } u^{\#} \in \partial F(u) \tag{2.30}$$

This completes the proof. □

Now we are ready to show the theorem of convergence of proximal gradient algorithm solving non-smooth and possibly non-convex problem $\widetilde{P}$ in (2.6).

**Theorem II.22** (almost sure finite path)**.** *Let $\gamma_{k+1} \in (0, 1/L]$ and $\{\beta_k, k \in \mathbb{N}\}$ be given by Algorithm (2) solving problem $\widetilde{P}$ in (2.6), the sequence is assumed to be bounded.*

*(i) The sequence $\{\beta_k\}_{k\in\mathbb{N}}$ has finite length,*

$$\sum_{k=1}^{\infty} \|\beta_{k+1} - \beta_k\|_2 < \infty \tag{2.31}$$

*(ii) The sequence $\{\beta_k\}_{k\in\mathbb{N}}$ converges to a stationary point*

$$\beta^* \in \mathcal{L} \triangleq \{\beta \in \Theta : 0 \in \nabla f(\beta) + \partial g(\beta)\}$$

.

*Proof.* Suppose $\bar{\beta} \in \omega(\beta_0)$ is any limit point of the sequence $\{\beta_k\}_{k\in\mathbb{N}}$. Since $\{F(\beta_k)\}_{k\geq 1}$ is a decreasing sequence and converges to $F(\bar{\beta}) = \inf_{k\in\mathbb{N}}\{F(\beta_k)\}$ (ref. Lemma II.9, theorem II.12), if there exists $k_0 \in \mathbb{N}$ for which $F(\beta_{k_0}) = F(\bar{\beta})$ then $F(\beta_k) = F(\beta_{k_0})$ and $\beta_k = \beta_{k_0}, \forall k > k_0$ (ref. lemma II.9), and induction shows (2.31) easily.

Otherwise, as $\{F(\beta_k)\}_{k\in\mathbb{N}}$ is decreasing, together with (2.15) we have for any $\delta > 0$, there exists a nonnegative integer $K_0$, such that $0 < F(\beta_k) - F(\bar{\beta}) < \delta$ for all $k > K_0$. Theorem II.12 (ii) established that $\lim_{k\to\infty} dist(\beta_k, \omega(\beta_0)) = 0$, thus for any $\epsilon > 0, \exists K_1 \in \mathbb{N}$, such that $dist(\beta_k, \omega(\beta_0)) < \epsilon$, for all $k > K_1$. Summing up these facts, we get that

$$\beta_k \in \{\beta \in \Omega : dist(\beta, \Omega) < \epsilon\} \cap \left\{\beta : 0 < F(\beta) - F(\bar{\beta}) < \delta\right\}, \text{ for all } k > l := K_0 \vee K_1$$

Let $\Omega = \omega(\beta_0)$, Theorem II.12 (ii) says $\Omega$ is compact and $F$ is constant on $\Omega$, then we can apply Lemma II.21 of uniformize K-L property to get for any $k > l$, there exists $c > 0, r \in (0, 1]$:

$$\left|F(\beta_k) - F(\bar{\beta})\right|^r \leq c \left\|A_k\right\|, \quad \forall A_k \in \partial F(\beta_k) \tag{2.32}$$

Consider the concave function $\phi(s) = s^{1-r}, s > 0$, by the concavity inequality $\phi(y) - \phi(x) \geq \langle y - x, \phi'(y)\rangle, \forall x, y \in (0, \delta]$ we have that

$$\left(F(\beta_k) - F(\bar{\beta})\right)^{1-r} - \left(F(\beta_{k+1}) - F(\bar{\beta})\right)^{1-r} \geq [F(\beta_k) - F(\beta_{k+1})] \cdot (1-r) \left|F(\beta_k) - F(\bar{\beta})\right|^{-r} \tag{2.33}$$

Summarizing (2.32) and (2.33), and let $c > 0$ denote generic constants, we get

$$\left(F(\beta_k) - F(\bar{\beta})\right)^{1-r} - \left(F(\beta_{k+1}) - F(\bar{\beta})\right)^{1-r} \geq \frac{F(\beta_k) - F(\beta_{k+1})}{c \left\|A_k\right\|}, \quad \forall A_k \in \partial F(\beta_k) \tag{2.34}$$

Denote

$$\Delta_{p,q} := \left(F(\beta_p) - F(\bar{\beta})\right)^{1-r} - \left(F(\beta_q) - F(\bar{\beta})\right)^{1-r} \tag{2.35}$$

Apply lemma II.11 on *subgradient growth bound* we get:

$$\Delta_{k,k+1} \geq \frac{F(\beta_k) - F(\beta_{k+1})}{\frac{2}{\gamma_k} \|\beta_{k-1} - \beta_k\|} \tag{2.36}$$

By Theorem II.12 (i) we get :

$$\Delta_{k,k+1} \geq \frac{\|\beta_{k+1} - \beta_k\|^2}{4 \|\beta_{k-1} - \beta_k\|} \geq 0 \tag{2.37}$$

that is

$$\|\beta_{k+1} - \beta_k\|^2 \leq 4\Delta_{k,k+1} \|\beta_{k-1} - \beta_k\| \tag{2.38}$$

Take squre root in both sides of the above inequality and use the fact that $2\sqrt{ab} \leq a + b, \forall a, b \geq 0$, we get:

$$\|\beta_{k+1} - \beta_k\| \leq \Delta_{k,k+1} + \|\beta_{k-1} - \beta_k\| \tag{2.39}$$

Summing up (2.39) $\|\beta_{i+1} - \beta_i\| \leq \Delta_{i,i+1} + \|\beta_{i-1} - \beta_i\|$ for $i = l+1, \ldots, k, \forall k > l \geq 1$ yields

$$2 \sum_{i=l+2}^{k+1} \|\beta_i - \beta_{i-1}\| \leq \sum_{i=l+2}^{k+1} \|\beta_i - \beta_{i-1}\| + \|\beta_{l+1} - \beta_l\| - \|\beta_{k+1} - \beta_k\| + 4 \sum_{i=l+1}^{k} \Delta_{i,i+1}$$

$$\sum_{i=l+1}^{k} \|\beta_{i+1} - \beta_i\| \leq \|\beta_{l+1} - \beta_l\| - \|\beta_{k+1} - \beta_k\|$$

$$+ 4 \left( \left(F(\beta_{l+1}) - F(\bar{\beta})\right)^{1-r} - \left(F(\beta_{k+1}) - F(\bar{\beta})\right)^{1-r} \right)$$

where in the last inequality one had $4 \sum_{i=l+1}^{k} \Delta_{i,i+1} = 4\Delta_{l+1,k+1}$ by the fact that $\Delta_{p,q} + \Delta_{q,r} = \Delta_{p,r}$ for all $p, q, r \in \mathbb{N}$. The limit of the right hand side of the above

inequality with $k, l \to \infty$ is zero, so by Cauchy's test of series we have

$$\sum_{k=0}^{\infty} \|\beta_{k+1} - \beta_k\| < \infty$$

For (ii), theorem II.12 (i) has shown that the limiting point set $\omega(\beta_0)$ of the sequence $\{\beta_k\}_{k\in\mathbb{N}}$ is a subset of $\mathcal{L}$, so we only need to show that the sequence $\left\{\beta_k \in \mathbb{R}^d\right\}_{k\in\mathbb{N}}$ is a Cauchy sequence and hence converges to a point in $\mathcal{L}$.

For any $q > p > l$ we have

$$\beta_q - \beta_p = \sum_{k=p}^{q-1} (\beta_{k+1} - \beta_k) \tag{2.40}$$

hence,

$$\|\beta_q - \beta_p\| = \left\| \sum_{k=p}^{q-1} (\beta_{k+1} - \beta_k) \right\| \leq \sum_{k=p}^{q-1} \|\beta_{k+1} - \beta_k\| \leq \sum_{k=p}^{\infty} \|\beta_{k+1} - \beta_k\| \overset{p\to\infty}{\to} 0,$$

it follows that the sequence $\{\beta_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}^d$ and hence is a convergence sequence, which converges to a stationary point

$$\beta^* \in \mathcal{L} \triangleq \{\beta \in \Theta : 0 \in \nabla f(\beta) + \partial g(\beta)\} \neq \emptyset$$

. □

## 2.5 Numerical Example: Mixed Effect Logistic Regression Model in High Dimensions

Here, we would like to use designed numerical examples to illustrate the algorithms developed in this chapter, demonstrate their convergence numerically, and in

the case of high dimensional examples, explore the sparsity properties of the algorithm solutions. We will also compare the solutions from stochastic exact algorithms with the solutions from deterministic approximate algorithm, and other benchmark algorithms.

We will focus on the mixed effect logistic regression model as an example in our numerical studies. It has non-convex negative loglikelihood function, and has non-convex and non-smooth objective function in high dimensional case when the number of unknown covariate coeffecients is much larger than the sample size. We will first introduce the model used in our numerical study below.

### 2.5.1 Model

A mixed effect logistic regression model models correlated binary responses using both fixed covariates and random effects. The *a priori* designed or estimated covariance structure of the random effect term in the model is used to model the correlation among the response.

We model the binary responses $\{y_i\}_{i=1}^N$, where $y_i \in \{-1, 1\}$, for all $i$, as conditionally independent realizations of the following Bernoulli model:

$$\mathbf{Y}_i|\mathbf{U}_\star \overset{ind.}{\sim} \begin{cases} 1, & \text{with probability (w.p.) } s\left(x_i'\beta + \sigma z_i'\mathbf{U}_\star\right) \\ 0, & \text{w.p. } 1 - s\left(x_i'\beta + \sigma z_i'\mathbf{U}_\star\right) \end{cases} \tag{2.41}$$

where $x_i \in \mathbb{R}^p$ is the vector of the $i$-th covariate, $z_i \in \mathbb{R}^q$ is the $i$-th loading vector for the random effect, which is known. The random effect $\mathbf{U}_\star$ is assumed to follow standard Gaussian distribution: $\mathbf{U}_\star \sim \mathcal{N}_q(0, I)$. We focus on estimating high dimensional covariate coefficients $\beta \in \mathbb{R}^p$ and we assume the random effect covariance

level parameter $\sigma > 0$ in (2.41) is given. In (2.41),

$$s(x) = \frac{e^x}{1+e^x}$$

denotes the cumulative distribution function of the standard logistic distribution.

We estimate $\beta$ via maximum likelihood approach, via solving the following optimization problem approximately and numerically:

$$\min_{\beta \in \mathbb{R}^p} -\ell(\beta) + \lambda P(\beta) \tag{2.42}$$

where $\ell(\beta)$ denotes the *log-likelihood function* of the model (2.41), $\phi(\mathbf{u})$ denotes the density function of $q$-dimensional standard Gaussian random variable $\mathbf{u} \in \mathbb{R}^q$:

$$\begin{aligned} \ell(\beta) &= \log \int_{\mathbf{u}} \prod_{i=1}^{N} \frac{1}{1+\exp\left(-y_i\left(\langle x_i, \beta\rangle + \sigma\langle z_i, u\rangle\right)\right)} \phi(\mathbf{u}) d\mathbf{u} \\ &= \log \int_{\mathbf{u}} \exp\left(-\sum_{i=1}^{N} \log\left[1+\exp\left(-y_i(\langle x_i, \beta\rangle + \sigma\langle z_i, \mathbf{u}\rangle)\right)\right]\right) \phi(\mathbf{u})\, d\mathbf{u} \end{aligned} \tag{2.43}$$

where

$$P(\beta) = \frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1$$

denotes the elastic net penalty, where $\|\beta\|_r = \left(\sum_{j=1}^{p} |\beta_j|^r\right)^{1/r}$, and $\alpha \in [0,1]$. $\alpha$ controls the trade-off between $\ell_1$ and $\ell_2$ errors. $P(\beta)$ encourage sparsity in the solution $\widehat{\beta}$ and controls multicolinearity in the fixed design matrix. Tuning parameter $\lambda > 0$ controls the level of regularization, larger $\lambda$ leads to more severe regularization and results in more parsimonious model.

### 2.5.2 Data

We generate synthetic response data $y_1, y_2, \ldots, y_N$ for our simulation study according to the Bernoulli model (2.41) introduced above.

For the high dimensional fixed effect specification, we let $X \in \mathbb{R}^{N\times p}$ denote the design matrix with row vectors $x_i \in \mathbb{R}^p$, and $\langle x_i, \beta\rangle$ denotes the fixed effect term, $i = 1, 2, \ldots, N$, $\beta = (\beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{R}^p$ is the unknown parameter vector. We generate the fixed design covariate matrix $X$ by drawing random $\mathbb{R}^N$-vectors from $\mathbf{N}(0, \Sigma_X)$ to form each column of the $N \times p$ sized design matrix $X$. In many simulation settings to follow, $\Sigma_X = I_N$, while in other settings, $\Sigma_X$ has an explicit structure, we will specify $\Sigma_X$ in each setting.

For the random effect speficication, we have the random effect loading matrix $Z \in \mathbb{R}^{N\times q}$, such that $Z^TZ$ is the convariance matrix of the $q$ dimensional random effect $U$. We assume $U$ follows Gaussian distribution $\mathsf{N}\left(0, \sigma Z^TZ\right)$. We have assumed a given as known $\sigma$ in most simulation studies we have carried out, however, in a few cases where we will specify, we have also estimated $\sigma$. $\langle z_i, \sigma U\rangle$ denotes the random effect term in the linear predictor, where $z_i$ is the $i$-th row of $Z$, $i = 1, 2, \ldots, N$, and $\sigma U \sim \mathsf{N}(0, \sigma^2 I_q)$ is a $q$-variate Gaussian random vector. We have specified two distinct structures of $Z$ in our different simulation study settings for different purpose. Let us specify the two ways below.

The first form of $Z$ we use in the study is such that $z_i = e_{\lceil iq/N\rceil}$ where $\{e_j, j \le q\}$ is the canonical basis of $\mathbb{R}^q$ and $\lceil\cdot\rceil$ denotes the upper integer part, in words, each $\mathbb{R}^q$-row vector of $Z$ is composed of $(q-1)$ zeros and an one indicating group label

for the $i$-th . The repeated measurement structure is usually used for longitudinal grouping structure in the data, which corresponds to $q$-group repeated measurement in the response $y_i, i = 1, 2, \ldots, N$.

The other form of $Z$ we use in the study is such that $Z^T Z \in \mathbb{R}^{q \times q}$ forms the low rank approximation to the (underlying) random effect covariance marix $\Sigma_U \in \mathbb{R}^{N \times N}$ based on *Eckart-Young-Mirsky* theorem (*Eckart and Young* (1936)). In this case the working model random effect vector $U \in \mathbb{R}^q$ is a low dimensional representation of the underlying true random effect $\mathbb{R}^N$ vector which follows Gaussian $\mathsf{N}(0, \sigma\Sigma_U)$ distribution, and the utilization of $Z \in \mathbb{R}^{N \times q}$ is such that the low dimensional working model random effect vector $U$ follows $\mathsf{N}\left(0, \sigma Z^T Z\right)$ distribution. We assume the random effect covariance matrix is given as $\Sigma_U = \frac{1}{p} X X^T$ which encodes the overall measurement similarity between all pairs of samples. Then we form $Z$ by the following Singular Value Decomposition (SVD) procedure: Let the SVD of $\Sigma_U$ be that $\Sigma_U = V D V^T$, where $V =: [V_q, V_{N-q}]$ is an orthornormal matrix, and $D = \begin{bmatrix} D_q & 0 \\ 0 & D_{N-q} \end{bmatrix}$ is a positive-semidefinite diagnal matrix, where $V_q \in \mathbb{R}^{N \times q}$ and $D_q \in \mathbb{R}^{q \times q}$. We then let $Z = V_q D_q^{1/2}$, and $Z^T Z = D_q$, thus the working model low dimensional random effect $U \sim \mathsf{N}(0, \sigma D_q)$.

### 2.5.3 Algorithms and Simulation Study Design

We will implement the stochastic exact algorithm, and the deterministic approximate algorithm we have developed in this chapter to solve for high dimensional mixed effect logistic regression model. Let us first derive for the details of our algorithms below.

#### 2.5.3.1 Stochastic Proximal Gradient Algorithm for Mixed Effect Logistic Regression Model

In the context of high dimensional mixed effect logistic regression model, for stochastic proximal gradient algorithm in (1), the objective function $F(\beta) = f(\beta) + g(\beta)$ in (2.1) has become the sum of the negative model loglikelihood function and the elastic net penalty.

In particular, from (2.43), $f(\beta)$ is

$$f(\beta) = -\ell(\beta) = -\log \int_{\mathbf{u}} \exp\left(-\sum_{i=1}^{N} \log\left[1 + \exp\left(-y_i(x_i'\beta + \sigma z_i'\mathbf{u})\right)\right]\right) \phi(\mathbf{u})\, d\mathbf{u} \tag{2.44}$$

where $\phi(\mathbf{u})$, and the elastic penalty $g(\beta)$ are defined in (2.43) too.

For implementation of the algorithm, it is useful to define the conditional loglikelihood of the observations $\mathbf{y} = \{y_1, \ldots, y_N\}$ given the random effect $\mathbf{U}_\star$:

$$\ell_c(\beta|\mathbf{U}_\star) = -\sum_{i=1}^{N} \log\left[1 + \exp\left(-y_i(x_i'\beta + \sigma z_i'\mathbf{u})\right)\right]$$

And the derivative $\nabla \ell_c(\beta|\mathbf{U}_\star)$:

$$\nabla \ell_c(\beta|\mathbf{U}_\star) = \sum_{i=1}^{N} s(-y_i(x_i'\beta + \sigma z_i'\mathbf{u})) \cdot (y_i x_i)$$

On another hand, the conditional distribution of the random effect $\mathbf{U}_\star$ given the observations $\mathbf{y}$ and the parameters $\beta$ is

$$\pi_\beta(\mathbf{U}_\star) = \exp\left(\ell_c(\beta|\mathbf{U}_\star) - \ell(\beta)\right) \phi(\mathbf{U}_\star) \tag{2.45}$$

The stochastic algorithm involves the gradient of $f(\beta)$, which can be routinely derived with Fisher's identity to be

$$\nabla f(\beta) = -\int \nabla \ell_c\left(\beta|\mathbf{U}_\star\right) \pi_\beta(\mathbf{U}_\star) d\mathbf{U}_\star = -\int \sum_{i=1}^{N} s(-y_i(x_i'\beta + \sigma z_i'\mathbf{u}))\cdot(y_i x_i)\, \pi_\beta(\mathbf{U}_\star) d\mathbf{U}_\star \tag{2.46}$$

The integration above is analytically intractable. To approximate $\nabla f(\beta)$ in the stochastic algorithm, we sample from the distribution $\pi_\beta$ using the MCMC sampler proposed in *Polson et al.* (2012) Polson et al. (2013) based on data-augmentation strategy.

To approximate $\nabla f(\beta) = -\int \nabla \ell_c\left(\beta|\mathbf{u}\right) \pi_\beta(\mathbf{u}) d\mathbf{u}$ via data-augmentation based MCMC, we write $\nabla f(\beta) = -\int H_\beta(\mathbf{u}) \widetilde{\pi}_\beta(\mathbf{u}, \mathbf{w}) d\mathbf{u} d\mathbf{w}$, where $\mathbf{u} := \mathbf{u}$, and $H_\beta(\mathbf{u}) := \nabla \ell_c\left(\beta|\mathbf{u}\right)$ notation-wise.

$\widetilde{\pi}_\beta(\mathbf{u}, \mathbf{w})$ is defined for $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{w} = (w_1, \ldots, w_N) \in \mathbb{R}^N$ by

$$\widetilde{\pi}_\beta(\mathbf{u}, \mathbf{w}) = \left( \prod_{i=1}^{N} \widetilde{\pi}_{\mathrm{PG}}\left(w_i; |x_i'\beta + \sigma z_i'\mathbf{u}|\right) \right) \pi_\beta\left(\mathbf{u}\right) \tag{2.47}$$

where $\widetilde{\pi}_{\mathrm{PG}}\left(\cdot; c\right)$ is the probability density of the Polya-Gamma distribution on the positive real line with parameter $c$. It has explicit function form as:

$$\widetilde{\pi}_{\mathrm{PG}}\left(w; c\right) = cosh(c/2) \exp(-wc^2/2) \rho(w) \mathbf{1}_{\{\mathbb{R}^+(w)\}},$$

where $\rho \propto \sum_{k\geq 0} (-1)^k (2k+1) \exp(-(2k+1)^2/(8w)) w^{-3/2}$ (*Biane et al.* (2001) Biane et al. 2001, Section 3.1). The target distribution $\widetilde{\pi}_\beta(\mathbf{u}, \mathbf{w})$ can be sampled by the Gibbs sampler below:

With the current value $(\mathbf{u}^t, \mathbf{w}^t)$ of the Markov chain, we sample the next point from the conditional distribution of $\mathbf{u}$ given $\mathbf{w}^t$, and the conditional distribution of $\mathbf{w}$ given $\mathbf{u}^{t+1}$:

$$\widetilde{\pi}_\beta\left(\mathbf{u}|\mathbf{w}\right) \equiv \mathsf{N}_q\left(\mu_\beta(\mathbf{w}); \Gamma_\beta(\mathbf{w})\right) \qquad \widetilde{\pi}_\beta\left(\mathbf{w}|\mathbf{u}\right) \equiv \prod_{i=1}^{N} \widetilde{\pi}_{\mathrm{PG}}\left(w_i; |x_i'\beta + \sigma z_i'\mathbf{u}|\right)$$

with

$$\Gamma_\beta(\mathbf{w}) = \left(I_q + \sigma^2 \sum_{i=1}^{N} w_i z_i z_i'\right)^{-1}, \qquad \mu_\beta(\mathbf{w}) = \sigma\Gamma_\beta(\mathbf{w}) \sum_{i=1}^{N} \left(Y_i/2 - w_i x_i'\beta\right) z_i.$$

The details of the above data augmentation Gibbs sampler derivation, based on nice properties of Polya-Gamma distribution , can be consulted in Section 3.1 of *Polson et al.* (2012).

With the Monte Carlo sample $\left\{\mathbf{u}^{(t)}\right\}_{t=1}^{T}$ drawn from the above Polya-Gamma sampler, we can proceed to give Monte Carlo approximation of the gradient $\nabla f(\beta)$:

$$\nabla f(\beta) \approx H(\beta) \triangleq -\frac{1}{T} \sum_{t=1}^{T} \nabla \ell_c(\beta|\mathbf{u}^{(t)}) \tag{2.48}$$

The MCMC approximation error $\eta = -\nabla\ell(\beta) - H(\beta)$ for $\nabla f(\beta)$ is seen as a stochastic perturbation of the gradient operator.

#### 2.5.3.2 Second Order Approximate Algorithm for Mixed Effect Logistic Regression Model

Through Taylor series expansion of

$$-\sum_{i=1}^{N}\log\left[1+\exp\left(-y_i(\langle x_i,\beta\rangle+\sigma\left\langle z_i,u\right\rangle)\right)\right]$$

to the second order at $\sigma u=0$ , we can approximate the loss function $f$ as

$$\widetilde{f}(\beta)=-\overset{\circ}{\ell}(\beta)-\frac{\sigma^2}{2}g(\beta)^T\left[I_q-\sigma^2h(\beta)\right]^{-1}g(\beta)+\frac{1}{2}\log\det\left[I_q-\sigma^2h(\beta)\right] \tag{2.49}$$

where $-\overset{\circ}{\ell}(\beta)$ is the negative log-likelihood function of logistic regression model

$$-\overset{\circ}{\ell}(\beta)=\sum_{i=1}^{N}\log\left(1+\exp\left(-y_i\left\langle x_i,\beta\right\rangle\right)\right) \tag{2.50}$$

$g(\beta)$ is the gradient of $-\sum_{i=1}^{N}\log\left[1+\exp\left(-y_i(\langle x_i,\beta\rangle+\sigma\left\langle z_i,u\right\rangle)\right)\right]$ with respect to $\sigma u$, evaluated at zero

$$g(\beta)=Z^T\left[y_i\left(1-s\left(y_i\left\langle x_i,\beta\right\rangle\right)\right)\right]_{i=1:N} \tag{2.51}$$

where we denote $s_i(\beta):=s\left(y_i\left\langle x_i,\beta\right\rangle\right)=\frac{1}{1+\exp(-y_i\langle x_i,\beta\rangle)}$.
$h(\beta)$ is the Hessian of $-\sum_{i=1}^{N}\log\left[1+\exp\left(-y_i(\langle x_i,\beta\rangle+\sigma\left\langle z_i,u\right\rangle)\right)\right]$ with respect to $\sigma u$, evaluated at zero

$$h(\beta)=-Z^TW_\beta Z \tag{2.52}$$

where $W_\beta=\mathrm{Diag}(w_i(\beta)_{i=1:N})$ and $w_i(\beta)=s_i(\beta)\left(1-s_i(\beta)\right)$.

We denote $A(\beta):=\left[I_q-\sigma^2h(\beta)\right]^{-1}$ and $B(\beta)=-A(\beta)$, the gradient of the

approximation function $\widetilde{f}$ in (2.49) can be derived to have closed form

$$\nabla \widetilde{f}(\beta) = -X^T \left[ y_i \left( 1 - s_i(\beta) \right) \right]_{i=1:N} - \frac{\sigma^2}{2} \left[ 2A(\beta) \frac{\partial g}{\partial \beta} + \left( \frac{\partial A}{\partial \beta_1} g(\beta), \ldots, \frac{\partial A}{\partial \beta_p} g(\beta) \right) \right]^T g(\beta) + \frac{1}{2} \left[ \operatorname{tr} \left( A \cdot \frac{\partial B}{\partial \beta_j} \right) \right]_{j=1:p} \quad (2.53)$$

The detailed derivation of the above second order approximations can be found in the end of the chapter.

#### 2.5.3.3 Simulation Study Design

We compare our developed algorithms with several benchmark algorithms to demonstrate the estimation performance and illustrate their computational time. We conduct the comparison in two steps. In the first step, we compare the *stochastic proximal gradient algorithm* with the *glmnet* algorithm to solve for the high dimensional mixed effect logistic regression model, with the *glm* algorithm ignoring the random effects in the model. We show that considering the random effects into the model clearly improves statistical estimation performance. Secondly, we compare the Monte Carlo and the deterministic Laplace approximation of the integration in the objective function, and show that the Markov chain Monte Carlo ((*Polson et al.*, 2012)) we leverage in *stochastic proximal gradient algorithm*, which accounts for the "stochastic" ingredient of our algorithm, leads to better estimation performance, while not increase too much on computational complexity.

We intended to compare the stochastic approximation of the integration in (2.43), versus the Laplace approximation, which is implemented in the *glmmLasso* algorithm in "glmmLasso" R package *Schelldorfer et al.* (2014). However, the "glmmLasso" implemented the Laplace approximation approximately, which is essentially

carrying out a second order Taylor approximation of the log-integrand at a particular fixed value of random effect $U$. In this regard, we deem comparing the stochastic algorithm with our second order Taylor approximation algorithm is sufficient for our purpose.

### 2.5.4 Model Selection Method

In simulation study, when the algorithm solves high dimensional problem with *elastic net* or *Lasso* penalty, we do model selection for a sequence of tuning parameters $\lambda$. The tuning parameter $\alpha$ in *elastic net* penalty is chosen subjectively as a common practice.

In simulation study, we can independently generate testing data set corresponding to each training data set that the algorithm trys to fit with a sequence of tuning parameters $\lambda$. We will select the fitted models associated with different $\lambda$'s based on their prediction performance on the testing data set. Precisely, we define the prediction error as the $\ell_2$ error of the predicted $\widehat{y}_i$ to the testing set $y_i$, that is $\frac{1}{N_t} \|\widehat{y}_i - y_i\|_2$, $i = 1, \ldots, N_t$, where $N_t$ is the testing sample size. $N_t$ typically equals $N/2$, half of the training sample size. We choose the model corresponding to a specific $\lambda$ that results in the smallest prediction error.

### 2.5.5 Results and Conclusion

In this section, we describe our simulation results on a synthetic data which resembles real world problem scale. We set sample size $N = 400$, fixed effect size or the problem dimension $p = 2000$. The random effect size is set to $q = 7$. We

randomly pick $10\beta_j$'s to be non-zero among the 2000 fixed effect coefficient $\beta$'s. The fixed and random effect design matrices $X$ and $Z$ respectively, and binary response $Y$ are all generated according to the fashion introduced in section 7.1. The random effects were also introduced as Gaussian variables in section 7.1. The parameters $\lambda$s were tuned based on prediction performance on the testing data, the details are in section 7.3.

The following plots presents the simulation results, based on 30 simulation runs on independent data sets. The $x$-axis codes each of the independent simulation run, while the $y$-axis denotes different performance metrics.

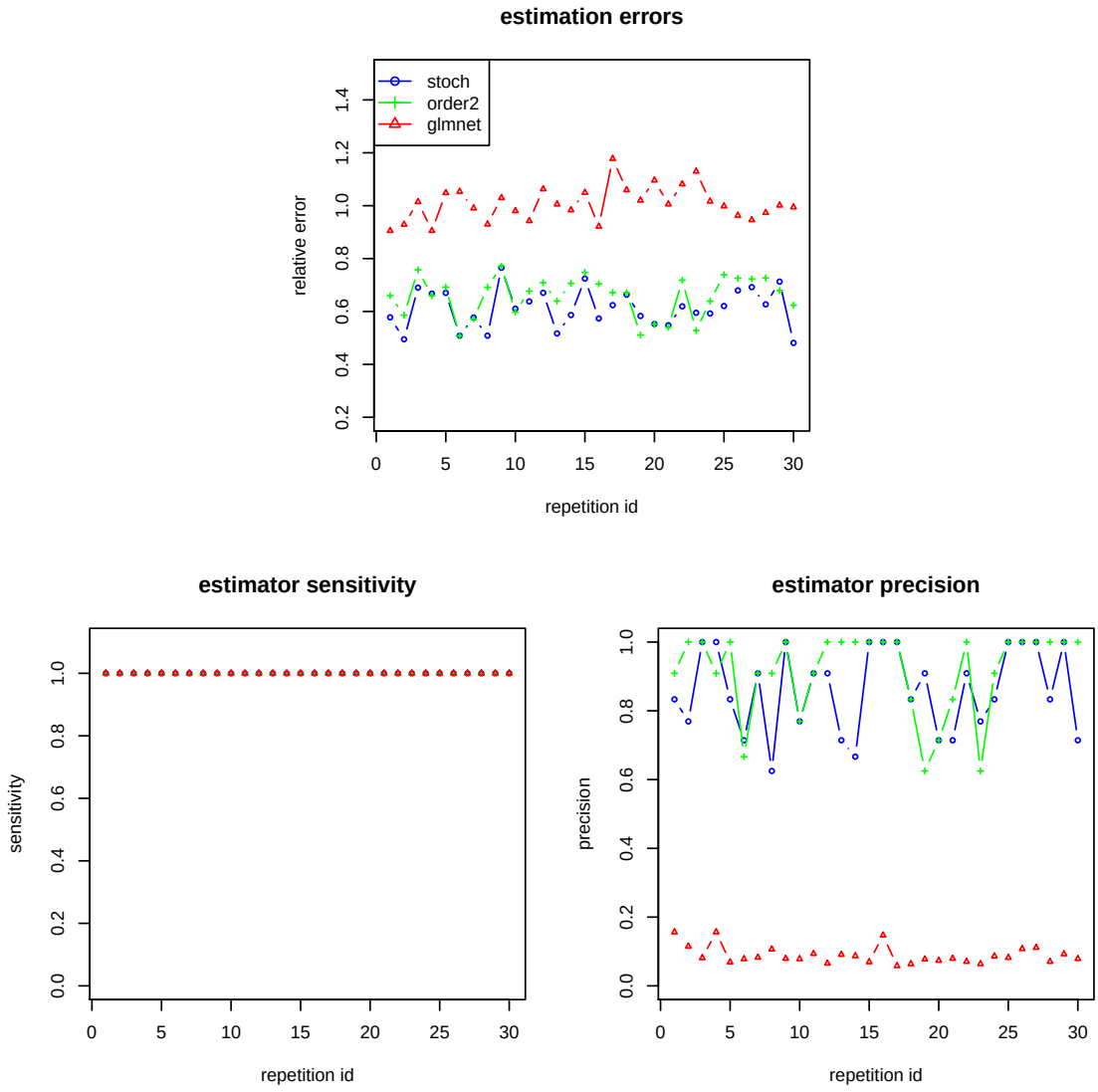


Figure 2.2: $N400p2000s10sig3$ step size $\gamma = 0.005$ stochastic proximal gradient descent, second order approximate, and glmnet algorithms.

In the above figure, *glmnet* results represents the algorithms which solve the problem ignoring the random effect in the data. We can tell in this case, *glmnet* algorithm cannot estimate the unknown fixed coefficients well, as the estimation $\ell_2$ norm percent error in the top left panel shows its errors are around 1.0, and its corresponding minimum prediction errors (after tuning for $\lambda$) around 0.1. Worst of

all, it has failed to recover the sparse set of the fixed effect coefficients. Indeed, the bottom graphs have shown that *glmnet* cannot tell the non-zero coefficients from the zero components, it basically estimate all the 2000 coefficients to be non zero.

The two effective algorithms here are the *stochastic proximal gradient algorithm* and *second order Taylor approximation algorithm*. Their estimation and prediction performance are similar in this setting. After 150 iteration steps, for both the algorithms, the estimation errors are around 0.6, and the minimum prediction errors for each of the independent are around $0.05 \sim 0.06$. In our setting here, on average the deterministic *second order Taylor approximate* algorithm outperforms the *stochstic proximal gradient descent algorithm* in terms of estimation precision, or sparsity recovery. Both algorithms have recovered all the non-zero components of fixed effect coefficients, while precision concerns overshooting, there are 14 out of 30 runs where *Taylor approximation* algorithm recovers more than 10 non-zero coefficients; whereas *stochastic proximal gradient descent* overestimates in 22 out of 30 runs, this is due to the stochasticity of the algorithm. However, the two algorithms both achieve above 0.6 of precision (recover about 16 non-zero variables, 6 more than the truth) in 90% of the cases, except in few cases where every algorithm recovers sparsity poorly.

## 2.6 Real Data Analysis

We use our stochastic proximal gradient and second order approximate algorithms to analyze a well known breast cancer data set.

Like the original Vijer study *Vijver et al.* (2002), we take the Distant Metastasis

within five years event, which is coded in $\{0, 1\}$ as the response vector, and level of expression of estrogen-receptors (ER), called ER status, diameter of the tumor, age, NIH score, St. Gallen score and lymph-node status (positive or negative) of each consecutively enrolled breast cancer patient as their clinical variables, as part of the fixed effects in our model. Our goal in this real data analysis is to identify certain genes to be potential prognosis predictors, or the prognostic signatures in breast cancer, of distant metastases within 5 years, which is a major clinical indicator of survival.

The data set includes 295 patients, gene intensity measurements of 24496 genes to begin with. Pre-processing is done by pruning the genes by individual gene T-test with clinical variables as the off-set terms for all the 24496 genes. We determine to use only the probes with p-value $< 0.01$, with the 70 gene set identified with Veer et al. 2002 added to the initial gene set, as some of the 70 genes were pruned out in the screening step. We end up with 295 patients and 1083 genes. For the screening procedure, we first enrolled the clinical variables "ESR1", "NIH", "StG", and "Posnodes" as the clinical characteristic for each patient, then we screen the genes by fitting logistic regression models to 5 year metastases event against all the clinical variables with each gene expression intensity measurement in one logistic model at a time. Then we conduct t-test for each fitted gene covariate coefficient, and pick the genes with corresponding testing p-value $< 0.01$. Out of the 24496 genes, the procedure screens 1024 genes with corresponding coefficient test p-value $< 0.01$. Then we check that there are only 16% or 11 genes in Vijer et al. 2002's 70 genes enrolled in these 1024 genes, we decide to include all the rest of the 70 genes into the pre-processed gene set, this gives us 1083 genes as an initial set for subsequent gene-selection in our model fittings.

From sample design perspective, the 295 patients are consecutively enrolled, and sample heterogeneity is very likely to present. We intend to apply mixed effect logistic model to the model the data, and select genes as potential prognostic signature for 5 year metastasis indicators. The advantage of using a mixed effects model also includes taking the small effect genes into proper account, so that the major effect genes could be more effectively discovered.

In order to apply our stochastic proximal gradient and second order approximate algorithms to identify the gene prognostic signatures, the first step is to construct the random effect variance - covariance matrix. We intend the random effect covariance matrix to code the genetic relationship among individuals, suppose $G \in \mathbb{R}^{n\times p}$ codes the prognostic gene expression signature, in which genes are depicted in the columns and samples in the rows. We use $K = \frac{1}{p}GG^T$, which captures the overall genetic similarity between all pairs of samples. As the random effect factor could only be low dimensional, we fix its dimension at $q \ll 295$, and do SVD for $K$ as $K = UDU^T$, and we take the top $q$ eigenvalues of $D$ to form the approximation of $K$ as $\Sigma_q = UD_qU^T$. Thus the random effect $U_\star \sim \mathsf{N}(0, \sigma^2\Sigma_q)$. Finally, before applying our algorithms to the screened gene data, we perform a standardization of the fixed effects design matrix $X$ and the random effects loading matrix $Z$ such that their columns are all of mean zero and unit standard deviation.

To describe the fitting and model selection schemes, both stochastic proximal gradient and second order approximate algorithms involve Lasso regularization, and model selection is done by solving a sequence of Lasso regularized optimization problems with different penalty amount $\lambda$'s, this is usually called "regularization path" in the literature (*Friedman et al.* (2010a)). For a given sequence of lambda, we run

the two algorithms for each lambda, and plot the solution path along the sequence of lambdas from largest to smallest. We will regard the genes that constantly stays in the solution path to be potential prognostic signatures.

In the following solution path plots, we have fixed the random effect factor dimension $q = 5$. We use Lasso regularization with a sequence of $\lambda = 30, 29.5, \ldots, 24$. Notice the x-axis is order reversed, so that log of lambda values decreases from left to right.

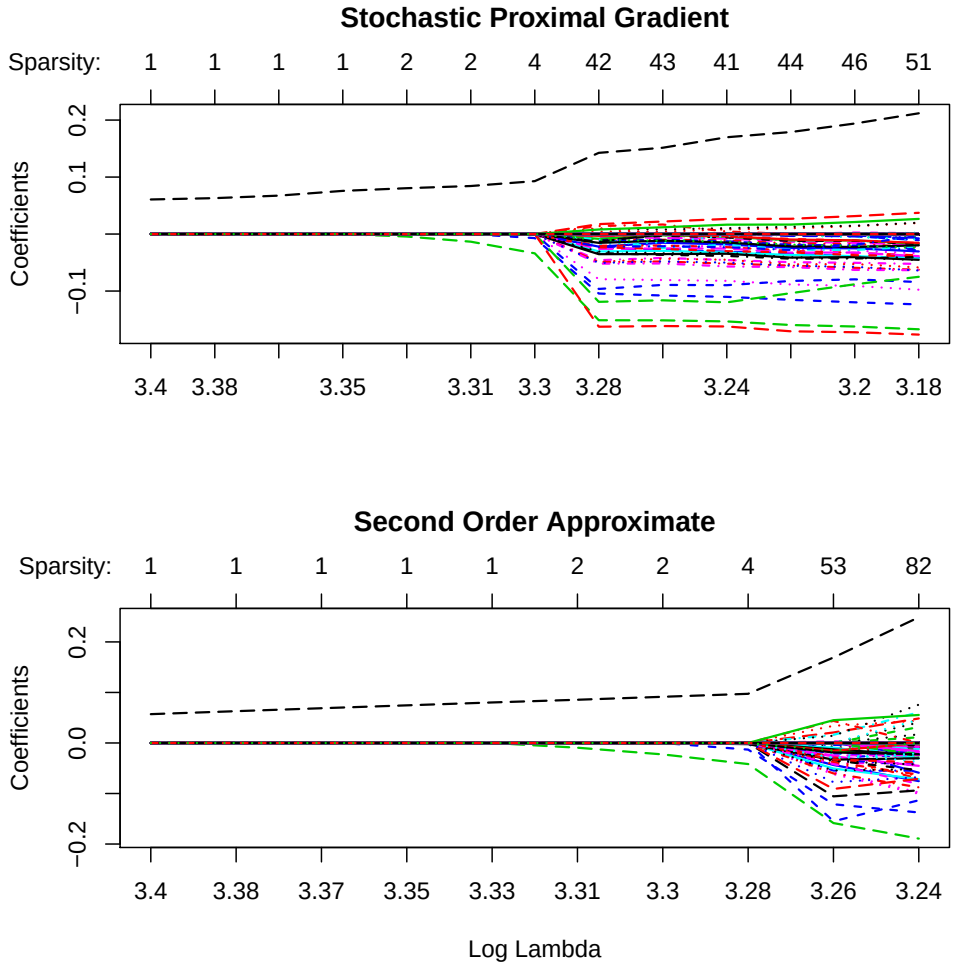


Figure 2.3: Solution paths for mixed effect logistic regression on breast cancer data, $q = 5$

We notice the solution paths of both algorithms have sudden jump ups with the number of selected genes as the regularization amount $\log(\lambda)$ decreases from 3.3 to 3.28 for the stochastic proximal gradient algorithm, and from 3.28 to 3.26 for the second order approximate algorithm. In the following, we first present the selected genes that stays along the solution paths of both algorithms, then we provide some explanation for the jump ups.

There are 4 genes selected by both the stochastic proximal gradient and the second order approximate algorithms which stay along the solution paths both before and after the jump ups. They are genes named "AF055033","NM˙006573", "NM˙002985" and "Contig44265˙RC" in the data set.

After the jump ups, there are 32 common genes selected and stays in the solution paths of both algorithms, before the number of selected genes grows above 100 as $\lambda$ continue to decrease toward zero. The commonly selected genes are "NM˙004120", "NM˙002727", "NM˙002985", "Contig54010˙RC", "Contig54425", "Contig60753˙RC", "NM˙005455", "Contig37281˙RC", "NM˙007019", "NM˙014395", "AB002304", "NM˙007204", "NM˙006573", "AF055033", "NM˙016009", "NM˙007358", "Contig26022˙RC", "AB028985", "Contig44265˙RC", "Contig47106˙RC", "AL049667", "AF049524", "NM˙001165", "NM˙000599", "NM˙020188", "NM˙003875", "Contig32185˙RC", "NM˙016577", "Contig51464˙RC", "NM˙005915", "NM˙001282", and "Contig20217˙RC".

For the jump ups in the number of selected genes, we have tried to use elastic-net penalty with different $\alpha$ values, but the results are similar. Judging from the solution paths, we see that many of the selected genes after the jump ups have coefficients close to zero, and the effect sizes are very close; on another hand, both

of our algorithm update all $p$-components of the gradient at a time in each iteration, instead of updating them in a coordinate-wise fashion. These factors very likely contribute to the jump ups in the number of non-zero coefficients. We will see in next chapter that with coordinate-wise updates, the solution path will evolve much more smoothly.

## 2.7 Proofs and Derivation

### 2.7.1 Proofs for Section 2.4.1

Proof for *Lemma* II.5:

*Proof.* By Assumption II.1 and consequently the descent lemma, we have

$$f(\beta_{k+1}) \leq f(\beta_k) + \langle \nabla f(\beta_k), \beta_{k+1} - \beta_k \rangle + \frac{1}{2\gamma_{k+1}} \|\beta_{k+1} - \beta_k\|_2^2, \quad \forall k \geq 0 \tag{2.54}$$

With the convexity of $g$ over $\Theta$, let $u = \beta_k - \gamma_{k+1} H_{k+1}$, and $\vartheta = \beta_k$ in proposition II.3 we get

$$g(\beta_{k+1}) \leq g(\beta_k) - \frac{1}{\gamma_{k+1}} \langle \beta_k - \gamma_{k+1} H_{k+1} - \beta_{k+1}, \beta_k - \beta_{k+1} \rangle \tag{2.55}$$

Summing up (2.54) and (2.55) we conclude:

$$F(\beta_k) - F(\beta_{k+1}) \geq \frac{1}{2\gamma_{k+1}} \|\beta_{k+1} - \beta_k\|^2 + \langle \beta_{k+1} - \beta_k, \eta_{k+1} \rangle \tag{2.56}$$

□

Proof for Lemma II.6:

*Proof.* By Lemma II.5 with the facts that $\gamma_k$ is positive non-increasing, and $\{F(\beta_k)\}_{k \in \mathbb{N}}$

is non-negative (Assumption II.2), we have

$$2\gamma_{k+1}F(\beta_{k+1}) \leq 2\gamma_k F(\beta_k) - \|\beta_{k+1} - \beta_k\|_2^2 + \langle \beta_{k+1} - \beta_k, \gamma_{k+1}\eta_{k+1}\rangle \tag{2.57}$$

In Lemma II.4, let $\nu_k = 2\gamma_k F(\beta_k)$, $\xi_{k+1} = \langle \beta_{k+1} - \beta_k, \gamma_{k+1}\eta_{k+1}\rangle$, $\chi_{k+1} = \|\beta_{k+1} - \beta_k\|_2^2$, for all $k \in \mathbb{N}$. Assume $\eta_k$ satisfies $\sum_{k\geq 0} \langle \beta_{k+1} - \beta_k, \gamma_{k+1}\eta_{k+1}\rangle < \infty$ a.s., then Lemma II.4 concludes that

$$\sum_{k\geq 0} \|\beta_{k+1} - \beta_k\|_2^2 < \infty \text{ and } \lim_{k\to\infty} \gamma_k F(\beta_k) \text{ exists.}$$

□

Proof for Lemma II.7:

*Proof.* From Algorithm (1) we know, for all $k \geq 1$, with $\gamma_k \in (0, \frac{1}{L}]$,

$$\beta_k := \arg\min_{\beta\in\Theta} \left\{ \langle \beta - \beta_{k-1}, H_k\rangle + \frac{1}{2\gamma_k} \|\beta - \beta_{k-1}\|^2 + g(\beta) \right\} \tag{2.58}$$

by the global optimization criterion of (2.72) we have

$$H_k + \frac{1}{\gamma}(\beta_k - \beta_{k-1}) + u_k = 0 \tag{2.59}$$

where $u_k \in \partial g(\beta_k)$, by additivity of subdifferential, that is $\nabla f(\beta_k) + \partial g(\beta_k) = \partial F(\beta_k)$, thus (i) is established:

$$A_k := \frac{1}{\gamma_k}(\beta_{k-1} - \beta_k) + \nabla f(\beta_k) - \nabla f(\beta_{k-1}) - \eta_k \in \partial F(\beta_k) \tag{2.60}$$

For (ii),

$$\begin{aligned}\|A_k\| &\leq \|\eta_k\| + \|\nabla f(\beta_k) - \nabla f(\beta_{k-1})\| + \frac{1}{\gamma_k}\|\beta_{k-1} - \beta_k\| \\ &\leq \|\eta_k\| + \left(L + \frac{1}{\gamma_k}\right)\|\beta_{k-1} - \beta_k\| \\ &\leq \|\eta_k\| + \frac{2}{\gamma_k}\|\beta_{k-1} - \beta_k\|, \text{ since } \gamma_k \leq \frac{1}{L}, \forall k \geq 1.\end{aligned}$$

□

Proof for Theorem II.8:

*Proof.* (i) Let $\beta^* \in \omega(\beta_0)$ be a limit point of $\{\beta_k\}_{k\in\mathbb{N}}$. To show that $\beta^* \in \mathcal{L}$, we need to show that for a sequence $\alpha_n \to \beta^*$ as $n \to \infty$, if $A_{\alpha_n} \in \partial F(\alpha_n)$ converges to 0, with $F(\alpha_n) \to F(\beta^*)$, then (by an elementary argument with the definition of subderivative) $0 \in \partial F(\beta^*)$. We will begin as the following.

$\{\beta_k\}_{k\in\mathbb{N}}$ is a bounded sequence, so there is a subsequence $\{\beta_{k_q}\}_{q\in\mathbb{N}}$ such that $\beta_{k_q} \overset{a.s.}{\to} \beta^*$ as $q \to \infty$. Since $g$ is lower semicontinous, we have

$$\liminf_{q\to\infty} g(\beta_{k_q}) \geq g(\beta^*) \tag{2.61}$$

From Algorithm (1), we have for all $k \in \mathbb{N}$

$$\beta_{k+1} \in \arg\min_{\beta\in\Theta}\left\{\langle \beta - \beta_k, H_{k+1}\rangle + \frac{1}{2\gamma_k}\|\beta - \beta_k\|^2 + g(\beta)\right\}$$

Thus letting $\beta = \beta^*$ in the above, we have

$$\begin{aligned}&\langle \beta_{k+1} - \beta_k, \eta_{k+1} + \nabla f(\beta_k)\rangle + \frac{1}{2\gamma_k}\|\beta_{k+1} - \beta_k\|^2 + g(\beta_{k+1}) \\ &\leq \langle \beta^* - \beta_k, \eta_{k+1} + \nabla f(\beta_k)\rangle + \frac{1}{2\gamma_k}\|\beta^* - \beta_k\|^2 + g(\beta^*)\end{aligned}$$

Choosing $k = k_q - 1$ in the above inequality and letting $q$ goes to $\infty$, we obtain

$$\begin{aligned}
&\limsup_{q\to\infty}\left\{\left\langle \beta_{k_q}-\beta_{k_q-1},\eta_{k_q}+\nabla f(\beta_{k_q-1})\right\rangle+\frac{1}{2\gamma_{k_q-1}}\left\|\beta_{k_q}-\beta_{k_q-1}\right\|^2+g(\beta_{k_q})\right\}\\
&\leq \limsup_{q\to\infty}\left\{\left\langle \beta^*-\beta_{k_q-1},\eta_{k_q}+\nabla f(\beta_{k_q-1})\right\rangle+\frac{1}{2\gamma_{k_q-1}}\left\|\beta^*-\beta_{k_q-1}\right\|^2+g(\beta^*)\right\}
\end{aligned} \tag{2.62}$$

We have from Lemma II.6 that almost surely,

$$\begin{cases}
\lim\limits_{k\to\infty}\|\beta_{k+1}-\beta_k\|=0\\
\lim\limits_{k\to\infty}\|\eta_k\|=0\\
\lim\limits_{q\to\infty}\nabla f(\beta_{k_q})=\nabla f(\beta^*) \text{ by continuity}
\end{cases}
\Rightarrow
\begin{cases}
\limsup\limits_{q\to\infty}\left\|\beta_{k_q}-\beta_{k_q-1}\right\|=0\\
\beta_{k_q-1}\to\beta^* \text{ as } q\to\infty\\
\limsup\limits_{q\to\infty}\left\|\eta_{k_q}\right\|=0\\
\limsup\limits_{q\to\infty}\nabla f(\beta_{k_q-1})=\nabla f(\beta^*)
\end{cases} \tag{2.63}$$

Combining the above (2.77) results with (2.76) we get

$$\limsup_{q\to\infty} g(\beta_{k_q})\leq g(\beta^*) \tag{2.64}$$

Recalling (2.75) one has

$$\lim_{q\to\infty} g(\beta_{k_q})=g(\beta^*) \tag{2.65}$$

Thus we finally obtain

$$\begin{aligned}
\lim_{q\to\infty}F(\beta_{k_q}) &= \lim_{q\to\infty}f(\beta_{k_q})+\lim_{q\to\infty}g(\beta_{k_q})\\
&= f(\beta^*)+g(\beta^*), f \text{ is continuously differentiable}\\
&= F(\beta^*)
\end{aligned} \tag{2.66}$$

From Lemma II.7 we know that

$$A_{k_q} := \frac{1}{\gamma_{k_q}}\left(\beta_{k_q-1} - \beta_{k_q}\right) + \nabla f(\beta_{k_q}) - \nabla f(\beta_{k_q-1}) - \eta_{k_q} \in \partial F(\beta_{k_q})$$

So

$$\lim_{q\to\infty} A_{k_q} = 0 \tag{2.67}$$

Now that $\beta_{k_q} \to \beta^*$, and in view of (2.80), (2.81) we get by definition of $\partial F$:

$$0 \in \partial F(\beta^*)$$

This shows $\beta^* \in \mathcal{L}$ and $\emptyset \neq \omega(\beta_0) \subset \mathcal{L}$.

(ii) By the definition of limiting points, this item follows as an elementary consequence.

(iii) Since the sequence $\{\beta_k\}_{k\in\mathbb{N}}$ is bounded, its closure $clo\,\{\beta_k\}_{k\in\mathbb{N}}$ is compact. By definition of limiting points, $\omega(\beta_0)$ is a closed subset of $clo\,\{\beta_k\}_{k\in\mathbb{N}}$, thus it is also compact.

It is a fact that a metric space is connected if and only if every continous $\{0,1\}$ valuded function defined on the space is a constant. (Apostol Theorem 4.36).

Suppose $f$ is an arbitrary $\{0,1\}$ valuded continuous function defined on the closure of the sequence $\{\beta_n\}_{n\in\mathbb{N}}$, in particular is such defined on $\omega\,(\beta_0)$. W.l.o.g., let $\beta^* \neq \beta'$ be any two limit points of the sequence $\{\beta_n\}_{n\geq 0}$, there are two subsequences converging to them respectively, $\beta_{n_p} \to \beta^*$ as $p \to \infty$ and $\beta_{n_q} \to \beta'$ as $q \to \infty$. Suppose $f(\beta^*) = 0$, by continuity of $f$, $\exists P_1 \in \mathbb{N}$, s.t. $\forall p > P_1, f(\beta_{n_p}) = 0$. On another hand, $\lim_{n\to\infty} \|\beta_{n+1} - \beta_n\| = 0$, $f$ is continuous on a compact set, thus is uniformly continuous, so $\lim_{n\to\infty} \|f(\beta_{n+1}) - f(\beta_n)\| = 0$,

in this $\{0,1\}$ valued case, $\exists N \in \mathbb{N}, \forall n > N, f(\beta_{n+1}) = f(\beta_n)$. To summarize, $\exists P_2 \in \mathbb{N}, \forall p > \max(P_1, P_2), s.t.\ n_p + m > N$ for all $m \geq 0$, thus $f(\beta_{n_p+m}) = f(\beta_{n_p}) = 0$. Now for any $\forall p > \max(P_1, P_2)$ there exists $Q \in \mathbb{N}$, s.t. $\forall q > Q, m_q = n_q - n_p \geq n_Q - n_p \geq 0$, and $f(\beta_{n_q}) = f(\beta_{n_p+m_q}) = f(\beta_{n_p}) = 0$, so by continuity of $f$, $f(\beta') = 0$. We conclude that $f \equiv 0$ is a constant on $\omega(\beta_0)$, which is now shown to be connected.

(iv) Since $F(\beta_k)$ is decreasing on $k$ and is assumed to be bounded from below, denote by $F_-$ the finite limit of $F(\beta_k)$ as $k \to \infty$. Take $\bar{\beta} \in \omega(\beta_0)$. There exists a subsequence $\beta_{k_q} \to \bar{\beta}$ as $q \to \infty, a.s.$ On one hand $\lim_{q\to\infty} F(\beta_{k_q}) = l, a.s.$, one the other hand as we proved in (i) $\lim_{q\to\infty} F(\beta_{k_q}) = F(\bar{\beta}), a.s.$, so $F(\bar{\beta}) = F_-, a.s.$, and $\lim_{k\to\infty} F(\beta_k) = F(\beta^*) = F_-, a.s.$.

□

Proof of Lemma II.9

*Proof.* By Assumption II.1 and consequently the descent lemma, we have

$$\widetilde{f}(\beta_{k+1}) \leq \widetilde{f}(\beta_k) + \left\langle \nabla \widetilde{f}(\beta_k), \beta_{k+1} - \beta_k \right\rangle + \frac{1}{2\gamma_{k+1}} \|\beta_{k+1} - \beta_k\|_2^2, \quad \forall k \geq 0 \tag{2.68}$$

With the convexity of $g$ over $\Theta$, let $u = \beta_k - \gamma \nabla \widetilde{f}(\beta_k)$, and $\vartheta = \beta_k$ in proposition II.3 we get

$$g(\beta_{k+1}) \leq g(\beta_k) - \frac{1}{\gamma_{k+1}} \left\langle \beta_k - \gamma_{k+1} \nabla \widetilde{f}(\beta_k) - \beta_{k+1}, \beta_k - \beta_{k+1} \right\rangle \tag{2.69}$$

Summing up (2.68) and (2.69) we conclude:

$$\widetilde{F}(\beta_k) - \widetilde{F}(\beta_{k+1}) \geq \frac{1}{2\gamma_{k+1}} \|\beta_{k+1} - \beta_k\|_2^2 \tag{2.70}$$

□

Proof of Lemma II.10

*Proof.* We sum up the inequality (2.70) for any fixed integer $l \geq 0$, $k = 0, \ldots, l$. Since $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_k \geq 0$, we have:

$$\begin{aligned} \sum_{k=0}^{l} \|\beta_{k+1} - \beta_k\|^2 &\leq 2\gamma_1 \sum_{k=0}^{l} \left[\widetilde{F}(\beta_k) - \widetilde{F}(\beta_{k+1})\right] \\ &= 2\gamma_1 \left(\widetilde{F}(\beta_0) - \widetilde{F}(\beta_{l+1})\right) \end{aligned} \tag{2.71}$$

From Assumption A we know $F(\beta) \geq 0, \forall \beta \in \Theta$. Taking limit of $l \to \infty$ on both sides of (2.71), we get the desired square summable result

$$\sum_{k=0}^{\infty} \|\beta_{k+1} - \beta_k\|_2^2 \leq \infty$$

and its elementary consequence

$$\lim_{k \to \infty} \|\beta_{k+1} - \beta_k\|_2 = 0$$

□

Proof of Lemma II.11:

*Proof.* From Algorithm 1 we know, for all $k \geq 1$, with $\gamma_k \in (0, 1/\tilde{L}]$,

$$\beta_k := \arg\min_{\beta \in \Theta} \left\{ \left\langle \beta - \beta_{k-1}, \nabla \widetilde{f}(\beta_{k-1}) \right\rangle + \frac{1}{2\gamma_k} \|\beta - \beta_{k-1}\|_2^2 + g(\beta) \right\} \tag{2.72}$$

by the global optimization criterion of (2.72) we have

$$\nabla \widetilde{f}(\beta_{k-1}) + \frac{1}{\gamma_k} (\beta_k - \beta_{k-1}) + u_k = 0 \tag{2.73}$$

where $u_k \in \partial g(\beta_k)$, by additivity of subdifferential we have $\nabla \widetilde{f}(\beta_k) + \partial g(\beta_k) = \partial \widetilde{F}(\beta_k)$, with these two results we get (i):

$$\widetilde{A}_k := \frac{1}{\gamma_k}(\beta_{k-1} - \beta_k) + \nabla \widetilde{f}(\beta_k) - \nabla \widetilde{f}(\beta_{k-1}) \in \partial \widetilde{F}(\beta_k) \tag{2.74}$$

For (ii),

$$\begin{aligned}
\left\|\widetilde{A}_k\right\|_2 &\leq \left\|\nabla \widetilde{f}(\beta_k) - \nabla \widetilde{f}(\beta_{k-1})\right\|_2 + \frac{1}{\gamma_k}\|\beta_{k-1} - \beta_k\|_2 \\
&\leq \left(\widetilde{L} + \frac{1}{\gamma_k}\right)\|\beta_{k-1} - \beta_k\|_2 \\
&\leq \frac{2}{\gamma_k}\|\beta_{k-1} - \beta_k\|_2, \text{ since } \gamma_k \leq \frac{1}{\widetilde{L}}
\end{aligned}$$

□

Proof of Theorem II.12:

*Proof.* (i) Let $\beta^* \in \widetilde{\omega}(\beta_0)$ be a limit point of $\{\beta_k\}_{k\in\mathbb{N}}$. To show that $\beta^* \in \widetilde{\mathcal{L}}$, we need to show that for a sequence $\alpha_n \to \beta^*$ as $n \to \infty$, if $\widetilde{A}_{\alpha_n} \in \partial \widetilde{F}(\alpha_n)$ converges to 0, with $\widetilde{F}(\alpha_n) \to \widetilde{F}(\beta^*)$, then (by an elementary argument with the definition of subderivative) $0 \in \partial \widetilde{F}(\beta^*)$. We will do this in the following.

$\{\beta_k\}_{k\in\mathbb{N}}$ is a bounded sequence, so there is a subsequence $\{\beta_{k_q}\}_{q\in\mathbb{N}}$ such that $\beta_{k_q} \xrightarrow{a.s.} \beta^*$ as $q \to \infty$. Since $g$ is lower semicontinous, we have

$$\liminf_{q\to\infty} g(\beta_{k_q}) \geq g(\beta^*) \tag{2.75}$$

From Algorithm (2), we have for all $k \in \mathbb{N}$

$$\beta_{k+1} \in \arg\min_{\beta\in\Theta}\left\{\left\langle \beta - \beta_k, \nabla \widetilde{f}(\beta_k)\right\rangle + \frac{1}{2\gamma_k}\|\beta - \beta_k\|_2^2 + g(\beta)\right\}$$

Thus letting $\beta = \beta^*$ in the above, we have

$$\begin{aligned}
&\left\langle \beta_{k+1} - \beta_k, \nabla \widetilde{f}(\beta_k) \right\rangle + \frac{1}{2\gamma_k} \|\beta_{k+1} - \beta_k\|_2^2 + g(\beta_{k+1}) \\
&\leq \left\langle \beta^* - \beta_k, \nabla \widetilde{f}(\beta_k) \right\rangle + \frac{1}{2\gamma_k} \|\beta^* - \beta_k\|_2^2 + g(\beta^*)
\end{aligned}$$

Choosing $k = k_q - 1$ in the above inequality and letting $q$ goes to $\infty$, we obtain

$$\begin{aligned}
&\limsup_{q\to\infty} \left\{ \left\langle \beta_{k_q} - \beta_{k_q-1}, \nabla \widetilde{f}(\beta_{k_q-1}) \right\rangle + \frac{1}{2\gamma_{k_q-1}} \left\|\beta_{k_q} - \beta_{k_q-1}\right\|^2 + g(\beta_{k_q}) \right\} \\
&\leq \limsup_{q\to\infty} \left\{ \left\langle \beta^* - \beta_{k_q-1}, \nabla \widetilde{f}(\beta_{k_q-1}) \right\rangle + \frac{1}{2\gamma_{k_q-1}} \left\|\beta^* - \beta_{k_q-1}\right\|^2 + g(\beta^*) \right\}
\end{aligned} \tag{2.76}$$

We have from Lemma II.6 we have,

$$\begin{cases} \lim\limits_{k\to\infty} \|\beta_{k+1} - \beta_k\| = 0 \\ \lim\limits_{q\to\infty} \nabla \widetilde{f}(\beta_{k_q}) = \nabla \widetilde{f}(\beta^*) \text{ by continuity} \end{cases} \Rightarrow \begin{cases} \limsup\limits_{q\to\infty} \left\|\beta_{k_q} - \beta_{k_q-1}\right\| = 0 \\ \beta_{k_q-1} \to \beta^* \text{ as } q \to \infty \\ \limsup\limits_{q\to\infty} \nabla \widetilde{f}(\beta_{k_q-1}) = \nabla \widetilde{f}(\beta^*) \end{cases} \tag{2.77}$$

Combining the above (2.77) results with (2.76) we get

$$\limsup_{q\to\infty} g(\beta_{k_q}) \leq g(\beta^*) \tag{2.78}$$

Recall (2.75), we have

$$\lim_{q\to\infty} g(\beta_{k_q}) = g(\beta^*) \tag{2.79}$$

Thus we finally obtain

$$\begin{aligned}\lim_{q\to\infty}\widetilde{F}(\beta_{k_q}) &= \lim_{q\to\infty}\widetilde{f}(\beta_{k_q}) + \lim_{q\to\infty} g(\beta_{k_q}) \\ &= \widetilde{f}(\beta^*) + g(\beta^*), \widetilde{f} \text{ is continuously differentiable} \\ &= \widetilde{F}(\beta^*)\end{aligned} \tag{2.80}$$

From Lemma II.7 we know that

$$\widetilde{A}_{k_q} := \frac{1}{\gamma_{k_q}}\left(\beta_{k_q-1} - \beta_{k_q}\right) + \nabla\widetilde{f}(\beta_{k_q}) - \nabla\widetilde{f}(\beta_{k_q-1}) \in \partial\widetilde{F}(\beta_{k_q})$$

So

$$\lim_{q\to\infty}\widetilde{A}_{k_q} = 0 \tag{2.81}$$

Now that $\beta_{k_q} \to \beta^*$, and in view of (2.80), (2.81) we get by definition of $\partial\widetilde{F}$:

$$0 \in \partial\widetilde{F}(\beta^*)$$

This shows $\beta^* \in \widetilde{\mathcal{L}}$ and $\emptyset \neq \widetilde{\omega}(\beta_0) \subset \widetilde{\mathcal{L}}$.

(ii) By the definition of limiting points, this item follows as an elementary consequence.

(iii) Since the sequence $\{\beta_k\}_{k\in\mathbb{N}}$ is bounded, its closure $clo\{\beta_k\}_{k\in\mathbb{N}}$ is compact. By definition of limiting points, $\widetilde{\omega}(\beta_0)$ is a closed subset of $clo\{\beta_k\}_{k\in\mathbb{N}}$, thus it is also compact.

It is a fact that a metric space is connected if and only if every continous $\{0,1\}$ valuded function defined on the space is a constant. (Apostol Theorem 4.36).

Suppose $f$ is an arbitrary $\{0,1\}$ valued continuous function defined on the closure of the sequence $\{\beta_n\}_{n\in\mathbb{N}}$, in particular on $\widetilde{\omega}(\beta_0)$. W.l.o.g., let $\beta^* \neq \beta'$

be any two limit points of the sequence $\{\beta_n\}_{n\geq 0}$, there are two subsequences converging to them respectively, $\beta_{n_p} \to \beta^*$ as $p \to \infty$ and $\beta_{n_q} \to \beta'$ as $q \to \infty$. Suppose $f(\beta^*) = 0$, by continuity of $f$, $\exists P_1 \in \mathbb{N}$, s.t. $\forall p > P_1, f(\beta_{n_p}) = 0$. On another hand, $\lim_{n\to\infty} \|\beta_{n+1} - \beta_n\| = 0$, $f$ is continuous on a compact set, thus is uniformly continuous, so $\lim_{n\to\infty} \|f(\beta_{n+1}) - f(\beta_n)\| = 0$, in this $\{0, 1\}$ valued case, $\exists N \in \mathbb{N}, \forall n > N, f(\beta_{n+1}) = f(\beta_n)$. To summarize, $\exists P_2 \in \mathbb{N}, \forall p > \max(P_1, P_2), s.t.\ n_p + m > N$ for all $m \geq 0$, thus $f(\beta_{n_p+m}) = f(\beta_{n_p}) = 0$. Now for any $\forall p > \max(P_1, P_2)$ there exists $Q \in \mathbb{N}$, s.t. $\forall q > Q, m_q = n_q - n_p \geq n_Q - n_p \geq 0$, and $f(\beta_{n_q}) = f(\beta_{n_p+m_q}) = f(\beta_{n_p}) = 0$, so by continuity of $f$, $f(\beta') = 0$. We conclude that $f \equiv 0$ is a constant on $\widetilde{\omega}(\beta_0)$, which is shown to be connected now.

(iv) Since $\widetilde{F}(\beta_k)$ is decreasing in $k$ and is assumed to be bounded from below, denote by $\widetilde{F_-}$ the finite limit of $\widetilde{F}(\beta_k)$ as $k \to \infty$. Take $\bar{\beta} \in \widetilde{\omega}(\beta_0)$. There exists a subsequence $\beta_{k_q} \to \bar{\beta}$ as $q \to \infty$. On one hand $\lim_{q\to\infty} \widetilde{F}(\beta_{k_q}) = l$, on the other hand we have proved in (i) that $\lim_{q\to\infty} \widetilde{F}(\beta_{k_q}) = \widetilde{F}(\bar{\beta})$, so $\widetilde{F}(\bar{\beta}) = \widetilde{F_-}$, and $\lim_{k\to\infty} \widetilde{F}(\beta_k) = \widetilde{F}(\beta^*) = \widetilde{F_-}$.

□

### 2.7.2 Derivation of the Second Order Approximation Gradient

We derive the deterministic approximation to the objective function (2.43), which we use to develop the second order approximation algorithm in the follow-

ing. We denote

$$
\begin{aligned}
F(\beta, \sigma u) &= \exp\left(f_\beta(\sigma u)\right) \\
&= \exp\left(-\sum_{i=1}^{N} \log\left[1+\exp\left(-y_i(\langle x_i, \beta\rangle + \sigma\left\langle z_i, u\right\rangle)\right)\right]\right) \\
&= \exp\left(\sum_{i=1}^{N} y_i(\langle x_i, \beta\rangle + \sigma\left\langle z_i, u\right\rangle) - \sum_{i=1}^{N} \log\left[1+\exp\left(y_i(\langle x_i, \beta\rangle + \sigma\left\langle z_i, u\right\rangle)\right)\right]\right),
\end{aligned}
\tag{2.82}
$$

In the following we expand $f_\beta(\sigma u)$ in $\log \mathbb{E}_u\left(\exp\left(f_\beta(\sigma u)\right)\right)$ at $\sigma u = 0$. One has:

$$
\begin{aligned}
\ell(\beta) &= \log \mathbb{E}_u\left(\exp\left(f_\beta(\sigma u)\right)\right) \\
&= \log \mathbb{E}_u\left(\exp\left(f_\beta(0) + \langle \nabla_{\sigma u} f_\beta(0), \sigma u\rangle + \frac{\sigma^2}{2} u^T \nabla^2_{\sigma u} f(\beta, \sigma\bar{u}) u\right)\right) \\
&= \log \mathbb{E}_u\left(F(\beta, 0) * \exp\left(\langle \nabla_{\sigma u} f_\beta(0), \sigma u\rangle + \frac{\sigma^2}{2} u^T \nabla^2_{\sigma u} f_\beta(\sigma\bar{u}) u\right)\right) \\
&= \log F(\beta, 0) + \log\left(\int_u \exp\left(\langle \nabla_{\sigma u} f_\beta(0), \sigma u\rangle + \frac{\sigma^2}{2} u^T \nabla^2_{\sigma u} f_\beta(\sigma\bar{u}) u\right) \phi(u) du\right) \\
&\approx \log \overset{\circ}{L}(\beta) + \\
&\log\left[\exp\left(\frac{\sigma^2}{2} \nabla f(0)^T \left[I_q - \sigma^2 \nabla^2 f(0)\right]^{-1} \nabla f(0)\right)\left(\sqrt{\det\left(I_q - \sigma^2 \nabla^2 f(0)\right)}\right)^{-1}\right] \\
&\approx \overset{\circ}{\ell}(\beta) + \frac{\sigma^2}{2} g(\beta)^T \left[I_q - \sigma^2 h(\beta)\right]^{-1} g(\beta) - \frac{1}{2}\log\left|I_q - \sigma^2 h(\beta)\right|
\end{aligned}
\tag{2.83}
$$

Where we assume $u \sim \mathsf{N}(0, I_q)$ so $\mathbb{E}\left(uu^T\right) = I_q$, $\bar{u}$ lies in between 0 and $u$;

Now we have an approximation of the loglikelihood function:

$$
\widetilde{\ell}(\beta) = \overset{\circ}{\ell}(\beta) + \frac{\sigma^2}{2} g(\beta)^T \left[I_q - \sigma^2 h(\beta)\right]^{-1} g(\beta) - \frac{1}{2}\log\left|I_q - \sigma^2 h(\beta)\right| \tag{2.84}
$$

We can derive that in the approximation function (2.84),

$\overset{\circ}{\ell}(\beta)$: the logistic regression model log-likelihood is

$$\begin{aligned}\overset{\circ}{\ell}(\beta) &= \log \Pi_{i=1}^{N} \frac{1}{1+\exp\left(-y_i \left\langle x_i, \beta \right\rangle\right)} \\ &= -\sum_{i=1}^{N} \log\left(1+\exp\left(-y_i \left\langle x_i, \beta \right\rangle\right)\right)\end{aligned} \tag{2.85}$$

$g(\beta)$: the gradient of $f_\beta(\sigma u)$ with respect to $\sigma u$ at zero is

$$\begin{aligned}g(\beta) &= \frac{\partial f}{\partial(\sigma u)}|_{\sigma u=0}(\beta) \\ &= Z^T\left[y_i\left(1-s\left(y_i\left\langle x_i, \beta\right\rangle\right)\right)\right]_{i=1:N}\end{aligned} \tag{2.86}$$

we henceforce denote $s_i(\beta) := s\left(y_i \left\langle x_i, \beta \right\rangle\right) = \frac{1}{1+\exp(-y_i\langle x_i,\beta\rangle)}$.

$h(\beta)$: the Hessian of $f_\beta(\sigma u)$ with respect to $\sigma u$ at zero is

$$\begin{aligned}h(\beta) &= \frac{\partial^2 f}{\partial(\sigma u)^2}|_{\sigma u=0}(\beta) \\ &= -Z^T W_\beta Z\end{aligned} \tag{2.87}$$

where $W_\beta = \mathrm{Diag}(w_i(\beta)_{i=1:N})$ and $w_i(\beta) = s_i(\beta)\left(1-s_i(\beta)\right)$.

Then we can derive the gradient for **the negative** of approximation function (2.84):

firstly

$$\begin{aligned}\frac{\partial \overset{\circ}{\ell}}{\partial \beta}(\beta) &= -\frac{\partial}{\partial \beta}\sum_{i=1}^{N}\log\left(1+\exp\left(-y_i\left\langle x_i,\beta\right\rangle\right)\right)\\ &= \sum_{i=1}^{N} y_i\left(1-\frac{1}{1+\exp(-y_i\left\langle x_i,\beta\right\rangle)}\right)x_i\\ &= \sum_{i=1}^{N} y_i\left(1-s_i(\beta)\right)x_i\\ &= X^T\left[y_i\left(1-s_i(\beta)\right)\right]_{i=1:N}\end{aligned} \tag{2.88}$$

and, denote $A(\beta) := \left[I_q - \sigma^2 h(\beta)\right]^{-1}$ in the following, we have

$$\frac{\partial}{\partial \beta}\left(\frac{\sigma^2}{2}g(\beta)^T\left[I_q-\sigma^2 h(\beta)\right]^{-1}g(\beta)\right) = \frac{\sigma^2}{2}\left[\left(\frac{\partial g}{\partial \beta}\right)^T A(\beta)g(\beta) + \left(\frac{\partial (A\cdot g)}{\partial \beta}\right)^T g(\beta)\right] \tag{2.89}$$

where,

$$\frac{\partial g}{\partial \beta} = -Z^T \mathrm{Diag}\left(s_i(\beta)(1-s_i(\beta))\right)X \tag{2.90}$$

and,

$$\begin{aligned}\left(\frac{\partial (A\cdot g)}{\partial \beta}\right)^T &= \left(\frac{\partial A}{\partial \beta}g(\beta) + A(\beta)\frac{\partial g}{\partial \beta}\right)^T\\ &= \left[\left(\frac{\partial A}{\partial \beta_1},\ldots,\frac{\partial A}{\partial \beta_p}\right)g(\beta) + A(\beta)\frac{\partial g}{\partial \beta}\right]^T\\ &= \left[\left(\frac{\partial A}{\partial \beta_1}g(\beta),\ldots,\frac{\partial A}{\partial \beta_p}g(\beta)\right) + A(\beta)\frac{\partial g}{\partial \beta}\right]^T\end{aligned} \tag{2.91}$$

Compute for $j = 1, \ldots, p$, denote $B = I_q - \sigma^2 h(\beta) = A^{-1}$ in the following,

$$\begin{aligned}
\frac{\partial A}{\partial \beta_j} &\triangleq \left[\frac{\partial A_{mn}}{\partial \beta_j}\right]_{m,n=1:q} \\
&= \frac{\partial}{\partial \beta_j}\left[\left(I_q - \sigma^2 h(\beta)\right)^{-1}\right] \\
&= \frac{\partial}{\partial \beta_j}(B^{-1}) \\
&= -B^{-1}\frac{\partial B}{\partial \beta_j}B^{-1} \\
&= -A\frac{\partial B}{\partial \beta_j}A
\end{aligned} \tag{2.92}$$

in the above,

$$\begin{aligned}
\frac{\partial B}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left[I_q - \sigma^2 h(\beta))\right] \\
&= \sigma^2 Z^T \mathrm{Diag}\left(\frac{\partial}{\partial \beta_j}\left(s_i(\beta)\left(1 - s_i(\beta)\right)\right)_{i=1:N}\right) Z \\
&= \sigma^2 Z^T \mathrm{Diag}\left(\left[y_i x_{ij} s_i(\beta)(1 - s_i(\beta))(1 - 2s_i(\beta))\right]_{i=1:N}\right) Z \\
&\in M(q \times q)
\end{aligned} \tag{2.93}$$

For the *logdet* term,

$$\begin{aligned}
\frac{\partial}{\partial \beta}\log\det(B) &= \left[\frac{\partial}{\partial \beta_j}\log\det(B)\right]_{j=1:p} \\
&= \left[\mathrm{tr}\left(A \cdot \frac{\partial B}{\partial \beta_j}\right)\right]_{j=1:p} \\
&\in \mathbb{R}^p
\end{aligned} \tag{2.94}$$

Now for the approximate log-likelihood function $\widetilde{\ell}(\beta)$,

$$\begin{aligned}\nabla\widetilde{\ell}(\beta) = & X^T \left[y_i \left(1 - s_i(\beta)\right)\right]_{i=1:N} \\ & + \frac{\sigma^2}{2}\left[2A(\beta)\frac{\partial g}{\partial \beta} + \left(\frac{\partial A}{\partial \beta_1}g(\beta), \ldots, \frac{\partial A}{\partial \beta_p}g(\beta)\right)\right]^T g(\beta) \\ & - \frac{1}{2}\left[\operatorname{tr}\left(A \cdot \frac{\partial B}{\partial \beta_j}\right)\right]_{j=1:p}\end{aligned} \tag{2.95}$$

Finally, the gradient of the negative log-likelihood function for which we minimize over $\beta$,

$$\begin{aligned}\nabla - \widetilde{\ell}(\beta) &= -X^T \left[y_i \left(1 - s_i(\beta)\right)\right]_{i=1:N} \\ &\quad - \frac{\sigma^2}{2}\left[2A(\beta)\frac{\partial g}{\partial \beta} + \left(\frac{\partial A}{\partial \beta_1}g(\beta), \ldots, \frac{\partial A}{\partial \beta_p}g(\beta)\right)\right]^T g(\beta) \\ &\quad + \frac{1}{2}\left[\operatorname{tr}\left(A \cdot \frac{\partial B}{\partial \beta_j}\right)\right]_{j=1:p} \\ &= -\nabla\widetilde{\ell}(\beta)\end{aligned} \tag{2.96}$$

# Chapter III

# A Fixed Effects Model Approximation to Mixed Effects Logistic Models

## 3.1 Introduction

In this chapter, we will devise another algorithm to solve for the high dimensional mixed effects logistic regression model, based on a very different approximation of the model. The approximation we propose in this chapter is to treat the random effects in the model as if they are fixed effects, we call this algorithm the *fixed effect approximate algorithm* ("FEAME"). Specifically, we will combine the true $p$-dimensional ($p \gg N$) fixed effects coefficients and the random effects $u \in \mathbb{R}^q$ to be a $p+q$ dimensional coefficient vector in model estimation, where the $q$ dimensional component $u$ will not be penalized. This approximation will reduce the high dimensional generalized mixed effects model to be a $p+q$ dimensional generalized linear model, which we already have highly efficient algorithms to solve, *glmnet* (*Friedman et al.* (2010a,b)) is a popular one.

The major goal of this chapter is to solve the fixed effect approximate problem

via Lasso regularized maximum likelihood inference, and to establish non-asymptotic estimation consistency results with high probability, for the *fixed effect approximate solution* $\widehat{\beta}$ from this algorithm, with respect to the true model generating fixed effects coefficients $\beta_\star$. In establishing this estimation error bound with high probability, we could show that under certain conditions of the design matrices, especially in terms of the problem dimension $p$, random effect dimension $q$ and the magnitude of the random effect noise level $\sigma$, one can actually get the fixed effect approximate solution $\widehat{\beta}$ reasonably close to the true generalized mixed effects model parameter $\beta_\star$.

This theoretical development takes its framework foundation in the estimation consistency theory of the high dimensional generalized linear models. For generalized linear models, the high dimensional point estimation theoretical development inherits largely from the work in high-dimensional linear models, the statistical properties derived in high dimensional linear models using Lasso hold analogously in generalized linear models regularized by Lasso *van de Geer* (2008), this is especially true when the distribution of dependent variable $Y|X = x$ (we treat X as fixed input information) is from the exponential family model. Our focus of statistical property of the *fixed effect approximate solution* $\widehat{\beta}$ is its estimation consistency with respect to the true generalized mixed effects model parameter $\beta_\star$, in terms of a non-asymptotic estimation error bound with high probability.

To estimate parameter $\beta_\star$, an identifiability assumption on the design matrix $X$ is needed *Bickel et al.* (2009); *Koltchinskii* (2009), due to the well known fact that in high dimensions, the design matrix $X_{n\times p}$ is column rank deficient when $p > n$, which leads to non-identifiable model parameters. One of such conditions is the *restricted eigenvalue* (RE) condition. First introduced by *Bickel et al.* (2009), RE is a

less restrictive condition than other compatibility conditions like restricted isometry property *Candés and Tao* (2005). The RE condition is frequently seen in literature for providing estimation error bounds for Lasso estimators, *Geer and Bühlmann* (2009) provides a comparison between RE and other related compatibility conditions for establishing Lasso error bound in high dimension.

Briefly speaking, the restricted eigenvalue condition tailors an affine vector subspace (usually a cone) in the $p$-dimensional vector space such that the loss function will be strongly convex in this subspace. In the case of linear models, that means the design matrix $\mathbf{X}$ is positive definite restricted to this affine subspace. The concept of restricted eigenvalue allows for establishing optimality in Lasso estimation, and development of estimation error bound such as the following in high dimensional linear model *Bühlmann and van de Geer* (2011):

$$\left\|\widehat{\beta} - \beta_\star\right\|_2 = O_p\left(s_\star^{1/2}\gamma\sqrt{\log(p)/n}\right) \tag{3.1}$$

where $s_\star$ denotes the number of non-zero coefficients in the true parameter vector $\beta_\star$, and $\gamma$ denotes a restricted eigenvalue of the design matrix $\mathbf{X}$ in the linear model. The corresponding error bound for generalized linear model is quite similar. The rate in (3.1) is optimal up to the $\log(p)$ factor and the restricted eigenvalue $\gamma$, in the context that the oracle least squares estimation would have an error rate $O_p(s/n)$ should we knew the non-zero true effects variables beforehand. Numerous elegant works are dedicated to dealing with the many facets of (3.1), see for example *Bunea et al.* (2007); *van de Geer* (2008); *Zhang and Huang* (2008); *Meinshausen and Yu* (2009); *Bickel et al.* (2009).

Our main technical contribution in establishing the high dimensional estimation error bound is that we have extended the restricted eigenvalue condition to a stochastic setting where both the fixed effect approximate solution $\widehat{\beta}(\lambda, \mathbf{U}_\star)$, and the true model log-likelihood function $\ell(\beta; data, \mathbf{U}_\star)$ are essentially random functions of the random effects $\mathbf{U}_\star$. We show that the extended restricted eigenvalue condition holds with high probability in this setting. We will describe the details en route our theoretical development.

We also contribute to the tool box of solving high dimensional generalized linear mixed effect models the fixed effects approximate algorithm, which reduces the problem to solving a high dimensional generalized linear model. Under suitable conditions, the approximate solution will be reasonably close to the true model parameters. From an algorithmic and computational point of view, fitting high dimensional generalized linear model with Lasso penalty are convex optimization problems. These models have convex negative log-likelihood function and convex $\ell_1$ penalty on the unknown coefficients to form a convex objective function (strongly convex if the model fisher information matrix is positive definite), which enables tractable computation, efficient optimization via many major algorithms. The recent very efficient coordinate gradient descent approach carried out in *glmnet* package *Friedman et al.* (2010b) is a favorable choices. It has been argued that the coordinate gradient descent approach is usually more efficient to solve $\ell_1$ penalized smooth convex optimization problems *Meier et al.* (2008); *Wu and Lange* (2008); *Friedman et al.* (2010a), we will use *glmnet* to solve the fixed effect approximate problem.

The rest of this chapter is organized as follows. In section 3.2 we will introduce the true and approximate models and corresponding optimization problems

and algorithms to fit the models, the fixed effect approximate algorithm will be also outlined. In section 3.3, we will develop our high dimensional statistical estimation error bound for the fixed effect approximate solution in detail. Finally in 3.4, we will present our comprehensive numerical simulation studies for the approximate algorithm, and numerical echos to the statistical property we have derived.

## 3.2 The Model and Problem

Recall that a mixed effect logistic regression model models correlated binary responses where the correlation among the response could be counted in the covariance structure of a random effect term introduced into the model. Specifically, we model the binary response or observation $y_1, y_2, \ldots, y_n$, for all $i$, $y_i \in \{0, 1\}$ as conditionally independent realizations as the following Bernoulli model:

$$\mathbf{Y}_i|\mathbf{U}_\star = u \overset{ind.}{\sim} \begin{cases} 1, & \text{with probability (w.p.) } s\left(x_i'\beta + \sigma z_{i,\cdot}'u\right) \\ 0, & \text{w.p. } 1 - s\left(x_i'\beta + \sigma z_{i,\cdot}'u\right) \end{cases} \tag{3.2}$$

where $x_i \in \mathbb{R}^p$ is the vector of the $i$-th covariate, $z_{i,\cdot} \in \mathbb{R}^q$ is the $i$-th loading vector for the random effect. The random effect $\mathbf{U}$ is assumed to follow standard Gaussian distribution: $\mathbf{U}_\star \sim \mathcal{N}_q(0, I)$. We focus on estimating high dimensional covariate coefficients $\beta$ and assume the random effect covariance level parameter $\sigma$ is given.

$$s(x) = 1/(1 + e^{-x})$$

denotes the cumulative distribution function of the standard logistic distribution.

We present and compare the exact and a fixed effect approximate regularized maximum likelihood inference problems relevant to fitting model (3.2). The Exact problem has been solved in Chapter 1. The current chapter will focus on the fixed effect approximate problem.

### 3.2.1 Exact Model and Problem

For the original probabilistic model in (3.2), when the dimension of the model covariate $p$ is greater than the sample size $n$, we adopt the well developed regularized maximum likelihood estimation framework to fit and infer about the model.

The regularized maximum likelihood estimation framework tries to maximize the model likelihood (or log-likelihood) function with respect to the unknown parameters to fit the model to the observed data, while it puts a contraint on the model parameter space to encourage certain desirable structure, control model complexity and avoid model overfitting.

For high dimensional mixed logistic regression model, the model likelihood function at $\beta \in \mathbb{R}^p$ given the observations $\{y_i\}_{i=1}^n$ is:

$$L(\beta) = \int_{\mathbb{R}^q} \exp\left(\ell_\beta(u)\right) \phi\left(u\right) du \tag{3.3}$$

where

$$\ell_\beta(u) = \sum_{i=1}^{n} \log\left[s(y_i(x_i'\beta + \sigma z_{i,\cdot}'u))\right] \tag{3.4}$$

is the log-likelihood function of the observations at $\beta$, conditioning on the random effect $\mathbf{U}_\star$ at $u \in \mathbb{R}^q$. And $\phi\left(u\right)$ is the density function of the standard Gaussian random effect $\mathbf{U}_\star$ evaluated at $u \in \mathbb{R}^q$.

The log-likelihood function of the model at $\beta \in \mathbb{R}^p$ given the observations is:

$$\begin{aligned}\ell(\beta) &= \log \int_{\mathbb{R}^q} \exp\left(\ell_\beta(u)\right) \phi\left(u\right) du \\ &= \log \int_{\mathbb{R}^q} \prod_{i=1}^{N} \frac{1}{\exp\left[-y_i\left(x_i'\beta + \sigma z_{i,\cdot}' u\right)\right] + 1} \phi(u) du;\end{aligned} \tag{3.5}$$

Again, $\ell(\beta)$ is a non-concave function, and it typically involves intractable $q$ dimensional integration.

The unknown parameter to be inferred is $\beta \in \mathbb{R}^p$ in our case. Apart from the model log-likelihood function in (3.5) to maximize, we apply the Lasso penalty on $\beta$ to encourage sparsity of the solution, which is necessary here as we are in $p > n$ regime; also this penalty is useful to counter the multi-colinearity problem in the high dimensional covariates. The model fitting problem is formulated in the following, as in chapter 1 problem

**(M1)**:

$$\min_{\beta \in \mathbb{R}^p} -\ell(\beta; y) + g(\beta) \tag{3.6}$$

where

$$g(\beta) = \lambda \left\|\beta\right\|_1 \tag{3.7}$$

is the Lasso penalty function applied to non-intercept coefficients of the covariates. $\lambda > 0$ is the regularization tuning parameter, $\left\|\beta\right\|_r = \left(\sum_{i=1}^{p} |\beta_i|^r\right)^{1/r}$.

Problem **M1** above is a nonconvex problem involving intractable $q$ dimensional integration. In Chapter 1 we have seen the stochastic proximal gradient algorithm solving problem **M1** exactly, which outperforms the other state-of-the-art algorithm

in usual cases. One possible practical concern for the stochastic proximal gradient algorithm, for now, is its relatively low computation efficiency. This is due to the fact that the algorithm involves solving the intractable high dimensional integration via Markov chain Monte Carlo techniques.

### 3.2.2 Approximate Model and Problem

Now that the original problem seems too challenging to solve in both an accurate and efficient manner, we apply the common wisdom to approximate the problem in a reasonable form, such that the approximate problem can be much more efficiently solved, while the approximate solution being reasonably close to its exact counterpart, and to the true data generating parameter value when sample size is large.

To motivate a simple way to approximate model (3.2), we have observed that in many simulation studies, when the presence of random effect in the high dimensional data is of moderate strength, in terms of the random effect dimension $q$ being much lower than the sample size $n$, and the covariance level parameter $\sigma$ being small, we can approximately solve the original model by solving a misspecified model which treats the random effect $\mathbf{U}_\star$ as an unknown fixed effect $u \in \mathbb{R}^q$. It turns out this approximate model can be highly efficiently solved, with solution being reasonablely close to the solution given by the stochastic proximal gradient algorithm solving the exact model (3.2) in chapter 1, we will demonstrate the performance comparison in the simulation studies.

Specifically, we approximate the exact solution by fitting a misspecified logistic regression model which models correlated binary observations $y_1, y_2, \cdots, y_n$ as

realizations of independent Bernoulli random variables $Y_1, Y_2, \cdots, Y_n$, such that for any $i = 1, 2, \ldots, n$, $\mathbb{P}(Y_i = 1) = s(x_i'\theta)$, and $\mathbb{P}(Y_i = 0) = 1 - \mathbb{P}(Y_i = 1)$. $\theta$ is the unknown covariate coefficient. We denote the augmented covariate matrix as $X_A = (X, Z) \in \mathbb{R}^{n\times(p+q)}$, where $X \in \mathbb{R}^{n\times p}$ is the original covariate marix, and $Z \in \mathbb{R}^q$ is the original random effect loading matrix presented in chapter 1. For all $i = 1, 2, \ldots, n$, $\widetilde{x}_i$ denotes the $i$-th row of the augmented covariate matrix. $\theta = (\beta, u)$ denotes the unknown model parameters. The misspecified logistic regression model is:

$$\mathbf{Y}_i \overset{ind.}{\sim} \begin{cases} 1, & \text{w.p. } s\left(\widetilde{x}_i'\theta\right) \\ 0, & \text{w.p. } 1 - s\left(\widetilde{x}_i'\theta\right) \end{cases} \tag{3.8}$$

The misspecified model has the negative log-likelihood function as the following:

$$-\widetilde{\ell}_n\left(\theta; y\right) = -\sum_{i=1}^{n}\left[y_i\left\langle\widetilde{x}_i, \theta\right\rangle - \log\left(1 + \exp\left(\left\langle\widetilde{x}_i, \theta\right\rangle\right)\right)\right] \tag{3.9}$$

It is routine to check that $-\widetilde{\ell}_n(\theta; y)$ is a convex function in $\theta = (\beta, u)$, where $\beta \in \mathbb{R}^p$ is the high dimensional component of the model parameter vector, while $u \in \mathbb{R}^q$ is its low dimensional component.

The approximate regularized maximum likelihood estimation problem is:

**problem M2**:

$$\min_{\theta\in\mathbb{R}^{p+q}} -\widetilde{\ell}_n\left(\theta; y\right) + g\left(\beta\right) \tag{3.10}$$

where $\widetilde{\ell}_n\left(\theta; y\right)$ is the log-likelihood function (3.9) of the misspecified model (3.8), and $g\left(\beta\right)$ is the Lasso function at $\beta \in \mathbb{R}^p$ specified in (3.7).

To solve **problem M2**, we treat the fixed effect and the random effect approx-

imated as fixed effect factors as one enlarged unknown fixed vector of parameters, we let the random effect factors be free of penalty, and fit the rest of the parameters as usual high dimensional logistic regression, via, say *glmnet*. The algorithm is summarized as the following: We outline the *Fixed effects approximate algorithm* below:

---

**Algorithm 3** Fixed effects approximate algorithm (FEAME)

---

1. Initialize $(\beta, u) = (\beta_0, u_0) \in \mathbb{R}^{p+q}$;
2. Solve the following problem via *glmnet* algorithm:

$$\left(\widehat{\beta}, \widehat{u}\right) = \underset{\beta \in \mathbb{R}^p, u \in \mathbb{R}^q}{\arg\min} -\widetilde{\ell}_n\left(\beta, u; y\right) + g\left(\beta\right)$$

   for $\widetilde{\ell}_n(\cdot)$ and $g(\cdot)$ in (3.10)

---

## 3.3 Statistical High Dimensional Estimation Theory

Consider the convex optimization problem **problem M2** defined in (3.10), we will show in the following that the solution of **problem M2** exsists and is well defined. The stochastic behavior of the solution stems from that of the random vectors $Y$ and $U_\star$. Henceforth we denote its solution as

$$\begin{aligned}\widehat{\theta}_\lambda :&= \widehat{\theta}_\lambda\left(Y, U_\star\right) \\ &= \underset{\theta \in \mathbb{R}^{p+q}}{\arg\min} \left\{-\widetilde{\ell}_n\left(\theta; Y, U_\star\right) + g\left(\theta_{[p]}\right)\right\}\end{aligned}$$

where $\theta_{[p]} = \beta \in \mathbb{R}^p$ denotes the subvector composed of the first $p$ elements of $\theta \in \mathbb{R}^{p+q}$.

One aspect of high dimensional statistical estimation theory concerns the con-

vergence behavior, especially the convergence rate of the M-estimator $\widehat{\theta}_\lambda$ to the true data generating model parameter $\theta_\star = (\beta_\star, U_\star) \in \mathbb{R}^{p+q}$. This convergence behavior in $\mathbb{R}^{p+q}$ is naturally expressed by the convergence behavior of various norms of $\widehat{\theta}_\lambda - \theta_\star$ in $\mathbb{R}$, when sample size $n$ and problem dimension $p$ goes to infinity.

In this chapter, we would like to establish the finite sample bound of the $\ell_2$-norm error of our estimation procedure, which in notation is to bound $\left\|\widehat{\theta}_\lambda\left(Y, U_\star\right) - \theta_\star\right\|_2^2$ with high probability. In so doing we would like to investigate the convergence behavior of our estimator.

In a nutshell , we point out the difference of our problem with other high dimensional convex statistical inference problems. In one hand, our true parameter value $\theta_\star$ is not a constant vector, but instead a degenerate $p+q$ dimensional Gaussian random vector composed of $p$ unknown sparse atoms and $q$ standard Gaussian variables: $\beta_\star$ is a unknown constant vector in $\mathbb{R}^p$ and $\mathbf{U}_\star \sim \mathcal{N}_q\left(0, I\right)$; On another hand, our M-estimator $\widehat{\theta}_\lambda$ has stochastic behavior stems not only from the random vector $Y$ with an observed sample $\{y_i\}_{i=1}^N$, but also from the unobserved random effect vector $U_\star$.

For the high dimensional estimation theory of our estimation procedure, we make the following basic assumption:

**A1**: $X_A = (X, Z) \in \mathbb{R}^{n\times(p+q)}$ is given as fixed. $Y = \{Y_i \in \mathbb{R}\}_{i=1:n}$ are conditionally independent given $U_\star$, and follows conditional *Bernoulli* distribution with $\mathbb{P}\left(Y_i = 1|U_\star\right) = \exp\left(X_A\theta_\star\right)/(1+\exp\left(X_A\theta_\star\right))$, where $\theta_\star = (\beta_\star, U_\star)$ and $U_\star \sim \mathsf{N}(0, \sigma^2 I_q)$, $\sigma > 0$ is assumed to be known.

Before we delve into the main theorem, let us first introduce the key quantities involved in the theory development. For the unknown $p$-dimensional parameter vector of interest $\beta_\star$, we denote its support as $S_\star = \{j \in \{1, 2, \ldots, p\} \, | \beta_{\star j} \neq 0\}$, and $s_\star := |S_\star|$ as the number of none-zero entries in $\beta_\star$. For the fixed design matrix $X_A = (X, Z)$, let $\|X_A\|_2 := \max_{j=1,\ldots,p+q} \|\widetilde{x}_{\cdot,j}\|_2^2$, and $\|\|X_A\|\|_\infty := \max_{i,j} |\widetilde{x}_{ij}|$, where $\widetilde{x}_{\cdot,j}$ denotes the $j$-th column, and $\widetilde{x}_{ij}$ the $ij$-th entry of the augmented design matrix $X_A$. We also let $\bar{\nu}_Z^2 := \max_{1 \leq i \leq n} \|z_{i,\cdot}\|_\infty^2 = \|\|Z\|\|_\infty^2$, where $z_{i,\cdot}$ denotes the $i$-th row of the random effect loading matrix $Z$.

#### 3.3.0.1 Restricted Eigenvalues for Mixed Effect GLM Regression

Analogue to high dimensional ($p >> n$) linear regression, the relevant constraint set $\mathcal{C}$ for restricted eigenvalues turns out to be a cone. Specifically, for appropriate choices of the regularization parameter $\lambda_N$, the lasso error $\widehat{\zeta} = \widehat{\beta} - \widehat{\beta}_\star$ satisfies a cone constraint of the form

$$\left\|\widehat{\zeta}_{S^c}\right\|_1 \leq \alpha \left\|\widehat{\zeta}_S\right\|_1 \tag{3.11}$$

for some constant $\alpha \geq 1$, where $\mathcal{S} := \{j \in \{1, \ldots, p\} : \beta_{\star j} \neq 0\}$ and $\zeta_\mathcal{S} \in \mathbb{R}^{|S|}$ denotes the subvector indexed by elements of $\mathcal{S}$, such that $(\zeta_\mathcal{S})_j = \zeta_j \cdot \mathbf{1}_{\{j \in \mathcal{S}\}}$. In fact, with appropriate choice of the regularization parameter $\lambda_N$, the lasso error in the mixed effect logistic regression model is also restricted to a cone we define in the following:

$$\mathcal{C} := \left\{\zeta \in \mathbb{R}^{p+q} : \|\zeta_{\mathcal{S}^C}\|_1 \leq 3 \|\zeta_\mathcal{S}\|_1\right\} \tag{3.12}$$

#### 3.3.0.2 Restricted Strong Convexity

In $\ell_2$ error bound theory development, in general one would desire that the objective function is sufficiently curved, so that a bound on the function difference translates into a bound on $\ell_2$ error.

To be specific, in our case, where $\widehat{\theta}$ is the lasso minimizer to the objective function $f_N(\theta)$, and $\theta_\star$ is the true parameter vector, it is desirable that a small difference in $\Delta f_N = \left| f_N(\widehat{\theta}) - f_N(\theta_\star) \right|$ would lead directly to a small difference in $\Delta\theta = \left\| \widehat{\theta} - \theta_\star \right\|_2$.

The notion of strong convexity specifies a desirable curvature of a function. To formalize, given a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$, $f$ is said to be *strongly covex* with parameter $\gamma > 0$ at $\theta \in \mathbb{R}^p$ if the inequality

$$f(\theta') - f(\theta) \geq \nabla f(\theta)^T (\theta' - \theta) + \frac{\gamma}{2} \|\theta' - \theta\|_2^2 \tag{3.13}$$

hold for all $\theta' \in \mathbb{R}^p$. When the function $f$ is twice continuously differentiable, an alternative characterization of strong convexity is expressed through the Hessian $\nabla^2 f$, such that, the function $f$ is strongly convex with parameter $\gamma > 0$ around $\theta_\star \in \mathbb{R}^p$ if and only if the minimum eigenvalue of the Hessian matrix $\nabla^2 f(\theta)$ is at least $\gamma$ for all vectors $\theta$ in a neighborhood of $\theta_\star$. In our particular statistical context, $f$ is the negative log-likelihood under the mixed effect logistic model parametrized by $\theta \in \mathbb{R}^{p+q}$, then $\nabla^2 f(\theta_\star)$ is the observed *Fisher information matrix*, so that strong convexity corresponds to a uniform lower bound on the Fisher information in *all directions*.

However, the above notion of strong convexity is not applicable in high dimensional linear regression, as well as mixed effect logistic regression, exactly because

the uniform lower bound of $\gamma$ needs to be applied in *all directions*.

Recall that in the high-dimensional setting, where the number of parameters, or the problem dimension $p$ is larger than sample size $N$, the objective function $\widetilde{\ell}_n(\theta; y) = \sum_{i=1}^{n} [y_i \langle \widetilde{x}_i, \theta \rangle - \log(1 + \exp(\langle \widetilde{x}_i, \theta \rangle))]$ in (3.9) is always convex for all $\theta$ in its domain. However, under what condition is it strongly convex? Notice the function $\widetilde{\ell}_N(\theta; y)$ is twice continuously differentiable and its Hessian matrix at $\theta \in \mathbb{R}^{p+q}$ is: $\nabla^2 \widetilde{\ell}_n(\theta; y) = \left(X_A^T W_\theta X_A\right)/N$. Thus, the logistic loss is strongly convex if and only if the eigenvalues of the positive semidefinite matrix $X_A^T W_\theta X_A$ are uniformly bounded away from zero. However, this matrix has rank at most $\min(N, p+q)$, thus it is always rank-deficient in high-dimensional setting where $p \geq N$, and hence not strongly convex. For this reason, we need a relaxed notion of strong convexity suitable for high dimensional analysis setting.

Let us note the difference between the notions of locally strongly convex and restricted strongly convex here. By literature convention, the notion of locally strongly convex refers to a function $f(\beta)$ being strongly convex in a neighborhood of a fixed $\beta \in \mathbb{R}^p$ in its domain, the definition applies to *all p directions* of any vector in a neighborhood of $\beta \in \mathbb{R}^p$, thus locally strongly convexity would not meet the challenge we face in our high dimensional problem. Whereas the notion of restricted strongly convex we need should at least not require strongly convex in all directions of the argument vector. It turns out in our theory development, one only needs to impose a strong convexity condition for some subset $\mathcal{C} \in \mathbb{R}^p$ of vectors $v \in \mathbb{R}^p$. In particular, we say that a function $f$ satisfies *restricted strong convexity* at $w^*$ with

respect to $\mathcal{C}$ if there is a constant $\gamma > 0$ such that

$$v^T \nabla^2 f(w) v \geq \gamma \text{ for all } v \in \mathcal{C} \text{ and } \|v\|_2 = 1, \tag{3.14}$$

and for all $w \in \mathbb{R}^p$ in a neighborhood of $w^*$.

Let us compare the case of linear regression with the case for our problem.

In the case of linear regression, this notion is equivalent to lower bounding the *restricted eigenvalues* of the model matrix, in particular, requiring that

$$\frac{1}{N} v^T X^T X v \geq \gamma \textit{ for all } v \in \mathcal{C} \text{ and } \|v\|_2 = 1 \tag{3.15}$$

While in the case of our problem of fixed effect approximation to random effect logistic regression model, it is equivalent to lower bounding the *restricted random eigenvalues* of the model matrix, which is requiring that

$$\frac{1}{N} \left( X_A^T W_{\theta_\star} X_A \right) \geq \gamma \textit{ for all } v \in \mathcal{C} \text{ and } \|v\|_2 = 1 \tag{3.16}$$

To explore the restricted strong convexity in the context of mixed effect logistic regression, we will inspect the following *restricted random eigenvalue* $\underline{\nu}_{\mathcal{C}}(U_\star)$:

$$\underline{\nu}_{\mathcal{C}}(U_\star) = \inf_{v \in \mathcal{C}, \|v\|_2 = 1} \left\{ v^T \left( X_A^T W_{\theta_\star} X_A \right) v \right\} / N \tag{3.17}$$

where $W_{\theta_\star}$ is a random $n \times n$ diagonal matrix with the $i$th random diagonal entry equal to $\frac{\exp(\langle \widetilde{x}_i, \theta_\star \rangle)}{(1+\exp(\langle \widetilde{x}_i, \theta_\star \rangle))^2}$, $\widetilde{x}_i$ is the $i$-th row of the augmented design matrix $X_A$. The randomness of $W_{\theta_\star}$ matrix is due to that $U_\star$ is a random $\mathbb{R}^q$ subvector in $\theta_\star = (\beta_\star, U_\star)$;

In an effort to lower bound the above *restricted random eigenvalue* (3.17), we define the corresponding usual *restricted eigenvalue* $\underline{\nu}_{\mathcal{C}}(0)$ in the following:

$$\underline{\nu}_{\mathcal{C}}(0) = \inf_{v \in \mathcal{C}, \|v\|_2 = 1} \left\{ v^T \left( X_A^T W_{\beta_\star} X_A \right) v \right\} / n, \tag{3.18}$$

where $W_{\beta_\star}$ is an $n \times n$ diagonal matrix with the $i$th diagonal entry equal to $\frac{\exp(\langle x_i, \beta_\star \rangle)}{(1+\exp(\langle x_i, \beta_\star \rangle))^2}$, where $x_i$ is the $i$-th row of the fixed effect design matrix $\mathbf{X}$;

We assume that the above defined *restricted eigenvalue* $\underline{\nu}_{\mathcal{C}}(0)$ is positive in our theory development. We note that there are design matrices $\mathbf{X}$ which can guarantee the positiveness of $\underline{\nu}_{\mathcal{C}}(0)$ in (3.18), for example the Gaussian and sub-Gaussian ensembles.

We denote $\underline{\nu}_{\mathcal{C}} := \underline{\nu}_{\mathcal{C}}(0)/2$. This deterministic constant $\underline{\nu}_{\mathcal{C}}$ will bound $\underline{\nu}_{\mathcal{C}}(0)$ from below.

As our theory develops, we define a constant $c := \|\|X_A \underline{\nu}_{\mathcal{C}}(0)\|\|_\infty^2 > 0$.

We present and prove our main theorem regarding convergence of our estimator $\widehat{\theta}_\lambda$ to the true parameter vector $\theta_\star$ in the following.

**Theorem III.1.** *Assume* $\sigma \leq \frac{2\underline{\nu}_{\mathcal{C}}}{c\|\|Z\|\|_\infty \sqrt{q \log(n)}}$. *Take the regularization parameter* $\lambda$ *such that* $\lambda/\sqrt{n} \geq 2\sqrt{2\|\|X_A\|\|_\infty \log(p+q)}$, *we have:*

*With probability at least* $\left(1 - \frac{2}{n} - \frac{2}{p+q}\right)$,

$$\left\| \widehat{\theta}_\lambda - \theta_\star \right\|_2 \leq \frac{48}{\underline{\nu}_{\mathcal{C}}} \sqrt{\frac{2\|\|X_A\|\|_\infty s_\star \log(p+q)}{n}} \tag{3.19}$$

*Given that sample size* $n$ *satisfies*

$$n \geq \frac{96}{\underline{\nu}_{\mathcal{C}}} \|\|X_A\|\|_\infty \sqrt{2\|\|X_A\|\|_\infty s_\star \log(p+q)}$$

In our main theorem above, $c$ and $\underline{\nu}_{\mathcal{C}}$ are positive constants that we have mentioned before, and will describe in detail as we develop the theorem later.

Before we prove our main result, let us first discuss the various factors in the above finite sample bound to put them into perspective. We further compare our result with non-asymptotic estimation error bounds in literature for high dimensional generalized linear models.

In most "with high probability" results, it is usually certain critical conditions that the design or objective functions need to satisfy with large probability. For our result specifically, we need $Y \in \{0,1\}^n$ and $U_\star \in \mathbb{R}^q$ satisfy the two conditions with high probability:

$$[\mathbf{C1}] : \left\|\nabla\widetilde{\ell}(\theta_\star, Y)\right\|_\infty \leq \frac{\lambda}{2}, \text{ and} \tag{3.20}$$

$$[\mathbf{C2}] : \underline{\nu}_{\mathcal{C}}(U_\star) \geq \underline{\nu}_{\mathcal{C}} \tag{3.21}$$

*Proof.* To prove the main theorem, we first state the conditions under which the conclusion of the theorem follows; we later show the high probability type of results guaranteeing the conditions hold with high probability when sample size and problem dimension are large.

[Condition **C2**]: $Y \in \{0,1\}^n$ and $U_\star \in \mathbb{R}^q$ satisfy that there $\underline{\nu}_{\mathcal{C}}(U_\star) \geq \underline{\nu}_{\mathcal{C}}$.

Before we systematically develop the main conclusion in Theorem 1, let us first introduce several basic conditions that we will use as intermediate tools en route the development.

[Condition **C1**]: $Y \in \{0,1\}^n$ and $U_\star \in \mathbb{R}^q$ satisfy that $\left\|\nabla\widetilde{\ell}(\theta_\star, Y)\right\|_\infty \leq \frac{\lambda}{2}$.

This basic condition **C1** says that the gradient of $\widetilde{\ell}(\theta, Y)$ at or around the true parameter value $\theta_\star$ should be small, which is often necessary for the M-estimator $\widehat{\theta}_\lambda$ derived from regularized maximization of $\widetilde{\ell}(\theta, Y)$ to be close to the truth $\theta_\star$.

We begin to develop our result with the following first lemma.

**Lemma III.2.** *Suppose for a fixed tuning parameter $\lambda > 0$, $Y \in \{0,1\}^n$ and $U_\star \in \mathbb{R}^q$ satisfy conditions **C1** and a restricted eigenvalue condition **C2** that we will introduce in proving this lemma, the solution $\widehat{\theta}_\lambda$ to (3.10) is well defined and satisfies:*

$$\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2 \leq \frac{24\lambda\sqrt{s_\star}}{n\underline{\nu}_{\mathcal{C}}} \tag{3.22}$$

We prove Lemma 1 below. We choose to introduce our *restricted eigenvalue condition* **C2** in the proof for Lemma 1 because the relevant derivation and notations necessary to present this condition are best developed while proving Lemma 1 for coherent presentation of ideas and logic. Furthermore, conditions **C1** and **C2** will be shown to hold with high probability in later Lemmas, all these lemmas will eventually bring us to our main conclusion in Theorem 1.

*Proof.* We begin with showing that the estimator $\widehat{\theta}_\lambda$ is well defined and has the property that $\widehat{\theta}_\lambda - \theta_\star$ lies in a cone $\mathcal{C} := \{\zeta \in \mathbb{R}^{p+q} : \|\zeta_{\mathcal{S}^C}\|_1 \leq 3\|\zeta_{\mathcal{S}}\|_1\}$, assuming condition [**C1**] holds.

For a given $\lambda > 0$, define

$$\mathcal{U}_n(\theta) := -\widetilde{\ell}_n(\theta; Y) + \lambda\|\theta\|_1$$

Let $\theta_\star \in \mathbb{R}^{p+q}$ be the true (random) parameter vector. By concavity of $\ell_n(\theta)$ we

have:

$$\begin{aligned}\mathcal{U}_n(\theta_\star) - \mathcal{U}_n(\theta) &= -\widetilde{\ell}_n(\theta_\star; Y) + \widetilde{\ell}_n(\theta; Y) + \lambda\left(\|\theta_\star\|_1 - \|\theta\|_1\right) \\ &\leq \left\langle \nabla\widetilde{\ell}_n(\theta_\star; Y), \theta - \theta_\star \right\rangle + \lambda\left(\|\theta_\star\|_1 - \|\theta\|_1\right)\end{aligned}$$

Apply Cauchy-Schwarz inequality we get

$$\mathcal{U}_n(\theta_\star) - \mathcal{U}_n(\theta) \leq \left\|\nabla\widetilde{\ell}_n(\theta_\star; Y)\right\|_\infty \cdot \|\theta - \theta_\star\|_1 + \lambda\left(\|\theta_\star\|_1 - \|\theta\|_1\right)$$

Apply condition **[C1]** to $\left\|\nabla\widetilde{\ell}_n(\theta_\star; Y)\right\|_\infty$ we get

$$\begin{aligned}\mathcal{U}_n(\theta_\star) - \mathcal{U}_n(\theta) &\leq \frac{\lambda}{2}\left(\|\theta - \theta_\star\|_1 - \|\theta\|_1\right) + \lambda\|\theta_\star\|_1 - \frac{\lambda}{2}\|\theta\|_1 \\ &\leq \frac{3\lambda}{2}\|\theta_\star\|_1 - \frac{\lambda}{2}\|\theta\|_1\end{aligned} \tag{3.23}$$

Thus $\mathcal{U}_n(\theta) > \mathcal{U}_n(\theta_\star)$ in the open set $\{\theta \in \mathbb{R}^{p+q} : \|\theta\|_1 > 3\|\theta_\star\|_1\}$. By continuity, $\mathcal{U}_n(\theta)$ has well defined global minimum in the compact set

$$\left\{\theta \in \mathbb{R}^{p+q} : \|\theta\|_1 \leq 3\|\theta_\star\|_1\right\}$$

.

That is $\widehat{\theta}_\lambda := \arg\min_{\theta\in\mathbb{R}^{p+q}} -\widetilde{\ell}_n(\theta) + \lambda\|\theta\|_1$ is well defined.

On another hand, let $\mathcal{S} = \{j \in \{1, \ldots, p+q\} : \theta_{\star j} \neq 0\}$, and $(\theta_{\mathcal{S}})_j = \theta_j \cdot \mathbf{1}_{\{j\in\mathcal{S}\}}$.

Let $\mathcal{S}^c$ denote the complement set of $\mathcal{S}$. We have,

$$\begin{aligned}\mathcal{U}_n(\theta_\star) - \mathcal{U}_n(\theta) &\leq \left\|\nabla\widetilde{\ell}_n(\theta_\star; Y)\right\|_\infty \cdot \|\theta - \theta_\star\|_1 + \lambda\left(\|\theta_\star\|_1 - \|\theta\|_1\right) \\ &\leq \frac{\lambda}{2}\|\theta_{\mathcal{S}} + \theta_{\mathcal{S}^c} - \theta_\star\|_1 + \lambda\left(\|\theta_\star - \theta_{\mathcal{S}} + \theta_{\mathcal{S}}\|_1 - \|\theta_{\mathcal{S}} + \theta_{\mathcal{S}^c}\|_1\right) \quad (3.24) \\ &\leq \frac{3\lambda}{2}\|\theta_{\mathcal{S}} - \theta_\star\|_1 - \frac{\lambda}{2}\|\theta_{\mathcal{S}^c}\|_1\end{aligned}$$

By the definition of $\mathcal{S}$ we see that $(\theta - \theta_\star)_{\mathcal{S}^c} = \theta_{\mathcal{S}^c}$ and $(\theta - \theta_\star)_{\mathcal{S}} = \theta_{\mathcal{S}} - \theta_\star$. Recall the cone $\mathcal{C} := \{\zeta \in \mathbb{R}^{p+q} : \|\zeta_{\mathcal{S}^c}\|_1 \leq 3\|\zeta_{\mathcal{S}}\|_1\}$, (3.24) above indicates that when $\theta - \theta_\star \notin \mathcal{C}$, $\mathcal{U}_n(\theta) > \mathcal{U}_n(\theta_\star)$; It is also clear that $\theta_\star \in \mathcal{C}$.

With the above two aspects, we conclude that $\widehat{\theta}_\lambda - \theta_\star$ lies in $\mathcal{C}$.

To further investigate $\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2$ by exploring the convexity of the negative loglikelihood function $\widetilde{\ell}_n(\theta; Y)$ around $\theta_\star$, we define,

$$\mathcal{L}_{n,\theta_\star}(\theta) = \widetilde{\ell}_n(\theta; Y) - \widetilde{\ell}_n(\theta_\star; Y) - \left\langle \nabla\widetilde{\ell}_n(\theta_\star), \theta - \theta_\star \right\rangle \quad (3.25)$$

Then for the difference of objective function at $\widehat{\theta}_\lambda$ and $\theta_\star$ we have

$$\begin{aligned}U_n(\theta_\star) - U_n(\widehat{\theta}_\lambda) &= \widetilde{\ell}_n(\widehat{\theta}_\lambda; Y) - \widetilde{\ell}_n(\theta_\star; Y) + \lambda\left(\|\theta_\star\|_1 - \left\|\widehat{\theta}_\lambda\right\|_1\right) \\ &= \mathcal{L}_{n,\theta_\star}(\widehat{\theta}_\lambda) + \left\langle \nabla\widetilde{\ell}_n(\theta_\star), \widehat{\theta}_\lambda - \theta_\star \right\rangle + \lambda\left(\|\theta_\star\|_1 - \left\|\widehat{\theta}_\lambda\right\|_1\right)\end{aligned}$$

By condition **C1** that $\left\|\nabla\widetilde{\ell}_n(\theta_\star)\right\|_\infty \leq \frac{\lambda}{2}$, we have:

$$\begin{aligned}\left|\left\langle \nabla\widetilde{\ell}_n(\theta_\star), \widehat{\theta}_\lambda - \theta_\star \right\rangle\right| + \left|\lambda\left(\|\theta_\star\|_1 - \left\|\widehat{\theta}_\lambda\right\|_1\right)\right| &\leq \left(\left\|\nabla\widetilde{\ell}_n(\theta_\star)\right\|_\infty + \lambda\right) \cdot \left\|\widehat{\theta}_\lambda - \theta_\star\right\|_1 \\ &\leq \frac{3\lambda}{2}\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_1\end{aligned}$$

Since $\widehat{\theta}_\lambda - \theta_\star$ lies in cone $\mathcal{C} = \{\zeta \in \mathbb{R}^{p+q} : \|\zeta_{\mathcal{S}^c}\|_1 \leq 3\|\zeta_{\mathcal{S}}\|_1\}$, we have:

$$\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_1 \leq 4\left\|(\widehat{\theta}_\lambda)_{\mathcal{S}} - \theta_\star\right\|_1 \leq 4\sqrt{s_\star}\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2$$

Since $U_n(\theta_\star) - U_n(\widehat{\theta}_\lambda) \geq 0$,

We have

$$U_n(\theta_\star) - U_n(\widehat{\theta}_\lambda) \leq \mathcal{L}_{n,\theta_\star}(\widehat{\theta}_\lambda) + 6\lambda\sqrt{s_\star}\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2 \tag{3.26}$$

And,

$$-\mathcal{L}_{n,\theta_\star}(\widehat{\theta}_\lambda) \leq 6\lambda\sqrt{s_\star}\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2 \tag{3.27}$$

Now, with (3.27) obtained and by convexity of $-\widetilde{\ell}_n(\theta; Y)$, if we are able to lower bound $-\mathcal{L}_{n,\theta_\star}(\widehat{\theta}_\lambda)$ by a positive quantity relating to $\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2$, we might be able to form an inequality in $\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2$ alone, and find the finite sample bound for the estimation error. Following this line, we need to explore the curvature of $\widetilde{\ell}_n(\theta; Y)$ at $\theta_\star$ and make use of the fact that $\widehat{\theta}_\lambda - \theta_\star$ lies in cone $\mathcal{C}$. We will do this in the following.

Define

$$\mathcal{L}_{i,\theta_\star}(\theta; Y_i) = \widetilde{\ell}_i(\theta; Y_i) - \widetilde{\ell}_i(\theta_\star; Y_i) - \left\langle \nabla\widetilde{\ell}_i(\theta_\star), \theta - \theta_\star \right\rangle$$

where for $Y_i$ in logistic model,

$$\widetilde{\ell}_i(\theta; Y_i) = Y_i \langle \widetilde{x}_i, \theta\rangle - \log\left(1 + \exp\left(\langle \widetilde{x}_i, \theta\rangle\right)\right), \quad \text{for all } i = 1, 2, \ldots, n;$$

Fix any $\alpha \in \mathbb{R}$, we define a univariate function $g_\alpha(h) : \mathbb{R} \to \mathbb{R}$, that

$$g_\alpha(h) = \log\left(1 + \exp(\alpha + h)\right), \text{ for all } h \in \mathbb{R} \tag{3.28}$$

Let $\alpha_i = \langle \widetilde{x}_i, \theta_\star \rangle$, $h = \langle \widetilde{x}_i, \theta - \theta_\star \rangle$ for all $i = 1, 2, \ldots, n$, we have:

$$-\mathcal{L}_{i,\theta_\star}(\theta; Y_i) = g_{\alpha_i}(h) - g_{\alpha_i}(0) - g'_{\alpha_i}(0)h, \quad \text{where } g'_{\alpha_i}(0) = s(\langle \widetilde{x}_i, \theta_\star \rangle) \tag{3.29}$$

where $s(\cdot)$ is defined in section 3.2.

From algebraic simplifications in the above (3.28) and (3.29), we observe that lower bounding $-\mathcal{L}_{i,\theta_\star}(\theta; Y_i)$ through the curvature information of $\widetilde{\ell}_i(\theta; Y_i)$ at $\theta_\star$ has been equivalently transformed into lower bounding the right hand side of (3.29) via the curvature information of the univariate function $g_\alpha(h)$ at 0. For this purpose, we have the following proposition, the proof of which can be found in the end of this chapter.

**Proposition III.3.** *For function $g_\alpha(h)$ defined in (3.28), we have $g_\alpha(h) - g_\alpha(0) - g'_\alpha(0)h \geq g''_\alpha(0)\frac{h^2}{|h|+2}$ , for all $\alpha, h \in \mathbb{R}$*

Apply Proposition (III.3) to (3.29), we get:

$$-\mathcal{L}_{i,\theta_\star}(\theta; Y_i) \geq g''_{\alpha_i}(0)\frac{(\theta - \theta_\star)^T \widetilde{x}_i \widetilde{x}_i^T (\theta - \theta_\star)}{2 + |\langle \widetilde{x}_i, \theta - \theta_\star \rangle|}$$

Apply Cauchy-Schwarz inequality to $\langle \widetilde{x}_i, \theta - \theta_\star \rangle$ in the above denominator we get:

$$-\mathcal{L}_{i,\theta_\star}(\theta; Y_i) \geq g''_{\alpha_i}(0)\frac{(\theta - \theta_\star)^T \widetilde{x}_i \widetilde{x}_i^T (\theta - \theta_\star)}{2 + \|\widetilde{x}_i\|_\infty \|\theta - \theta_\star\|_1}$$

To generalize the inequality for all $i = 1, 2, \ldots, n$, we replace $\|\widetilde{x}_i\|_\infty$ by its matrix counterpart $|||X_A|||_\infty$ and use Cauchy-Schwarz inequality to change $\ell_1$ norm to $\ell_2$ norm in the above. Then we have:

$$-\mathcal{L}_{i,\theta_\star}(\theta; Y_i) \geq g''_{\alpha_i}(0)\frac{(\theta - \theta_\star)^T \widetilde{x}_i \widetilde{x}_i^T (\theta - \theta_\star)}{2 + 4\sqrt{s_\star}|||X_A|||_\infty \|\theta - \theta_\star\|_2} \tag{3.30}$$

where $g''_{\alpha_i}(0) = s(\langle \widetilde{x}_i, \theta_\star \rangle)\,(1 - s(\langle \widetilde{x}_i, \theta_\star \rangle))$.

Now, $\mathcal{L}_{n,\theta_\star}(\theta)$ defined in (3.25) relates to $\mathcal{L}_{i,\theta_\star}(\theta; Y_i)$ in the above as:

$$\mathcal{L}_{n,\theta_\star}(\theta) = \sum_{i=1}^{n} \mathcal{L}_{i,\theta_\star}(\theta; Y_i)$$

So by summing up (3.30) for $i = 1, \ldots, n$ we get

$$-\mathcal{L}_{n,\theta_\star}(\theta) \geq \frac{1}{2} \cdot \frac{1}{1 + 2\sqrt{s_\star}\|\|X_A\|\|_\infty \|\theta - \theta_\star\|_2} (\theta - \theta_\star)^T X_A^T W_{\theta_\star} X_A (\theta - \theta_\star) \tag{3.31}$$

where $W_{\theta_\star}$ is an $n \times n$ diagonal matrix with the $i$th diagonal entry equal to $g''_{\alpha_i}(0)$.

From (3.31) we see that by controlling the minimum eigenvalue of the matrix $X_A^T W_{\theta_\star} X_A$ to be positive, we would be able to reach our goal of lower bounding $-\mathcal{L}_{n,\theta_\star}(\theta)$ by a positive quantity in $\|\theta - \theta_\star\|_2$.

Note that the matrix $X_A^T W_{\theta_\star} X_A$ is random in nature due to the randomness in $\theta_\star$, so its eigenvalues are naturally random. We define the following random quantity analogous to the smallest eigenvalue of a fixed matrix:

$$\begin{aligned} \underline{\nu}_{\mathcal{C}}(U_\star) &:= \inf_{v \in \mathcal{C}, \|v\|_2 = 1} \left\{ v^T \left( X^T W_{\theta_\star} X \right) v \right\} / n \\ &\equiv \inf_{v \in \mathcal{C} \setminus \{0\}} \left\{ \frac{v^T \left( X^T W_{\theta_\star} X \right) v}{n \|v\|_2} \right\} \end{aligned}$$

where $\theta_\star = (\beta_\star, U_\star)$, $U_\star \sim \mathsf{N}(0, \sigma^2 I)$, and thus $\underline{\nu}_{\mathcal{C}}(U_\star)$ is a random function, for which we assume the following condition holds

[**C2**] There exists a constant $\underline{\nu}_{\mathcal{C}} > 0$, such that $U_\star \in \mathbb{R}^q$ satisfies that $\underline{\nu}_{\mathcal{C}}(U_\star) \geq \underline{\nu}_{\mathcal{C}}$.

In fact, the above condition [**C2**] is our version of the *restricted eigenvalue condition* that we plan to discuss in detail in Lemma 3 later, where we prove that it holds with high probability as $n$ and $p$ grows large. For now, we make use of this condition and draw conclusion for Lemma 1 below.

Having shown that $\widehat{\theta}_\lambda - \theta_\star$ lies in the cone $\mathcal{C}$, together with (3.31) and [**C2**] applied to $X_A^T W_{\theta_\star} X_A$, we have,

$$-\mathcal{L}_{n,\theta_\star}(\widehat{\theta}_\lambda; Y) \geq \frac{1}{2} \cdot \frac{n\underline{\nu}_{\mathcal{C}}}{1 + 2\sqrt{s_\star}|||X_A|||_\infty \left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2} \left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2^2 \tag{3.32}$$

Now combine (3.27) with (3.32), we have

$$\frac{1}{2} \cdot \frac{n\underline{\nu}_{\mathcal{C}}}{1 + 2\sqrt{s_\star}|||X_A|||_\infty \left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2} \left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2^2 \leq 6\lambda\sqrt{s_\star} \left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2$$

By simple algebraic arrangement we get the conclusion of Lemma 1:

$$\left\|\widehat{\theta}_\lambda - \theta_\star\right\|_2 \leq \frac{24\lambda\sqrt{s_\star}}{n\underline{\nu}_{\mathcal{C}}}, \text{ assuming } n \geq \frac{48\lambda\sqrt{s_\star}|||X_A|||_\infty}{\underline{\nu}_{\mathcal{C}}}$$

□

Next, we show that when sample size $n$ and problem dimension $p$ go large, conditions **C1** and **C2** hold with high probability.

We define the following event:

$$\mathcal{E}_n(\lambda, \sigma) \stackrel{\text{def}}{=} \left\{ Y \in \{0,1\}^n, U_\star \in \mathbb{R}^q : \|\nabla \ell(\theta_\star, Y)\|_\infty \leq \frac{\lambda}{2}, \underline{\nu}_{\mathcal{C}}(U_\star) \geq \underline{\nu}_{\mathcal{C}} \right\} \tag{3.33}$$

For the high probability results, we aim to show that for certain choice of $\lambda$ and $\sigma$, we have:

$$\mathbb{P}_{Y,U_\star}\left(\mathcal{E}_n(\lambda,\sigma)\right) \to 1, \text{ as } n,p \to \infty.$$

For the above it suffices to show that

$$\mathbb{P}\left(\left\|\nabla \ell(\theta_\star, Y)\right\|_\infty \geq \frac{\lambda}{2}\right) \to 0 \tag{3.34}$$

and

$$\mathbb{P}\left(\underline{\nu}_{\mathcal{C}}(U_\star) \leq \underline{\nu}_{\mathcal{C}}\right) \to 0 \tag{3.35}$$

when $n,p \to \infty$, with suitable choice of $\lambda, \sigma$. It is understood that $\mathbb{P}$ denotes joint $(Y, U_\star)$ probability measure.

**Lemma III.4.** *For a fixed $\lambda > 0$, it holds that :*

$$\mathbb{P}_{Y,U_\star}\left(\left\|\nabla \widetilde{\ell}(\theta_\star, Y)\right\|_\infty \geq \frac{\lambda}{2}\right) \to 0 \tag{3.36}$$

*as $p \to \infty$.*

*Proof.* Recall the misspecified model log-likelihood function defined in (3.9), we have:

$$\nabla \widetilde{\ell}_n(\theta_\star; Y) = \sum_{i=1}^{n} \epsilon_{\star i} \widetilde{x}_i \tag{3.37}$$

where we denote the deviance $\epsilon_{\star i} := Y_i - \mu_{\star i}$, with $\mu_{\star i} = \mathbb{E}(Y_i|U_\star) = \frac{\exp(\langle \widetilde{x}_i, \theta_\star \rangle)}{1+\exp(\langle \widetilde{x}_i, \theta_\star \rangle)}$, and $\widetilde{x}_i$ denotes the $i$-th row of the augmented matrix $\mathbf{X}_A$.

From the above we have

$$\left\|\nabla\widetilde{\ell}_n(\theta_\star;Y)\right\|_\infty = \max_{j=1:(p+q)} |\langle \epsilon_\star, \widetilde{x}_{\cdot,j}\rangle| \tag{3.38}$$

where $\epsilon_\star = (\epsilon_{\star 1}, \ldots, \epsilon_{\star n})$, $\widetilde{x}_{\cdot,j}$ is the $j$th column of matrix $X_A = (X, Z)$.

By an equivalent transformation, for any $j$-th entry of $\nabla\widetilde{\ell}_n(\theta_\star; y)$, $j = 1, \ldots, p+q$, we have

$$\mathbb{P}_{(Y,U_\star)}\left(\sum_{i=1}^n \epsilon_{\star i}\widetilde{x}_{ij} > \frac{\lambda}{2}\right) = \inf_{t\geq 0} \mathbb{P}_{(Y,U_\star)}\left[\exp\left(t\sum_{i=1}^n \epsilon_{\star i}\widetilde{x}_{ij}\right) > \exp\left(\frac{t\lambda}{2}\right)\right]$$

Apply Markov's inequality to the above, we get

$$\mathbb{P}_{(Y,U_\star)}\left(\sum_{i=1}^n \epsilon_{\star i}\widetilde{x}_{ij} > \frac{\lambda}{2}\right) \leq \inf_{t\geq 0} \frac{\mathbb{E}_{Y,U_\star}\left[\prod_{i=1}^n e^{t\widetilde{x}_{ij}(Y_i-\mu_{\star i})}\right]}{e^{t\lambda/2}} \tag{3.39}$$

For the numerator in (3.39) above, we observe that condition on random effect $U_\star$, the random variables $(Y_i - \mu_{\star i}) \in [-1, 1], i = 1, \ldots, n$ are zero-mean, supported on interval $[-1, 1]$. So they are *sub-Gaussian* random variables satisfying $\mathbb{E}_{U_\star}\left[e^{t(Y_i-\mu_{\star i})}\right] \leq e^{t^2/2}$. Then we have:

$$\begin{aligned}
\mathbb{E}_{Y,U_\star}\left[\prod_{i=1}^n e^{t\widetilde{x}_{ij}(Y_i-\mu_{\star i})}\right] &= \mathbb{E}_{U_\star}\left[\prod_{i=1}^n \mathbb{E}_Y\left[e^{t\widetilde{x}_{ij}(Y_i-\mu_{\star i})}|U_\star\right]\right] \\
&\leq \mathbb{E}_{U_\star}\left[\prod_{i=1}^n e^{\frac{t^2\widetilde{x}_{ij}^2}{2}}\right] \\
&= \exp\left(\frac{t^2}{2}\left\|\widetilde{x}_{\cdot,j}\right\|_2^2\right)
\end{aligned} \tag{3.40}$$

Now back to (3.39) we have

$$\begin{aligned}\mathbb{P}_{Y,U_\star}\left(\sum_{i=1}^{n}\epsilon_{\star i}\widetilde{x}_{ij} > \frac{\lambda}{2}\right) &\leq \inf_{t\geq 0}\frac{\mathbb{E}_{Y,U_\star}\left[\prod_{i=1}^{n} e^{t\widetilde{x}_{ij}(Y_i-\mu_{\star i})}\right]}{e^{t\lambda/2}} \\ &\leq \inf_{t\geq 0}\exp\left(\frac{t^2}{2}\left\|\widetilde{x}_{\cdot,j}\right\|_2^2 - \frac{\lambda}{2}t\right) \\ &= \exp\left(-\frac{\lambda^2}{4\left\|\widetilde{x}_{\cdot,j}\right\|^2}\right)\end{aligned} \tag{3.41}$$

By symmetry we get the following

$$\begin{aligned}\mathbb{P}_{Y,U_\star}\left(\left|\sum_{i=1}^{n}\epsilon_{\star i}\widetilde{x}_{ij}\right| > \frac{\lambda}{2}\right) &\leq 2\exp\left(-\frac{\lambda^2}{4\left\|\widetilde{x}_{\cdot,j}\right\|_2^2}\right) \\ &\leq 2\exp\left(-\frac{\lambda^2}{4n|||X_A|||_\infty^2}\right)\end{aligned} \tag{3.42}$$

where $|||X_A|||_\infty = \max_{j=1,\ldots,p+q}\left\|\widetilde{x}_{\cdot,j}\right\|_\infty$ by definitions of $|||X_A|||_\infty$ and $\left\|\widetilde{x}_{\cdot,j}\right\|_\infty$.

To proceed with the above $j$-th component result, we apply a simple union bound argument and get the followng

$$\begin{aligned}\mathbb{P}_{Y,U_\star}\left(\left\|\nabla\widetilde{\ell}_n(\theta_\star;Y)\right\|_\infty > \frac{\lambda}{2}\right) &\leq 2(p+q)\exp\left(-\frac{\lambda^2}{4n|||X_A|||_\infty^2}\right) \\ &= 2\exp\left(\log(p+q) - \frac{\lambda^2}{4n|||X_A|||_\infty^2}\right)\end{aligned} \tag{3.43}$$

If we choose $\lambda$ such that $2\log(p+q) = \frac{\lambda^2}{4n|||X_A|||_\infty^2}$, we can get

$$\mathbb{P}_{Y,U_\star}\left(\left\|\nabla\widetilde{\ell}_n(\theta_\star;Y)\right\|_\infty > \frac{\lambda}{2}\right) \leq \frac{2}{p+q} \tag{3.44}$$

Thus as problem dimension $p \to \infty$, $\left\|\nabla\ell_n(\theta_\star;y)\right\|_\infty \leq \frac{\lambda}{2}$ holds with probability

converging to 1.

□

To show that for some positive constant $\underline{\nu_{\mathcal{C}}}$, $\underline{\nu_{\mathcal{C}}}(U_\star) \geq \underline{\nu_{\mathcal{C}}}$ with high probability, it is sensible to control the random effect noise level $\sigma$. We show this in the following lemma:

**Lemma III.5.** *For a fixed $\lambda > 0$ and $\sigma \leq \frac{2\underline{\nu_{\mathcal{C}}}}{c\bar{\nu_Z}\sqrt{\log(n)}}$, it holds that*

$$\mathbb{P}_{Y,U_\star}\left(\underline{\nu_{\mathcal{C}}}\left(U_\star\right) \leq \underline{\nu_{\mathcal{C}}}\right) \to 0$$

*as $n, p \to \infty$.*

*Proof.* Recall the cone $\mathcal{C} := \{v \in \mathbb{R}^{p+q} : \|v_{\mathcal{S}^C}\|_1 \leq 3\|v_{\mathcal{S}}\|_1\}$, $\mathcal{S}$ is the support of $\theta_\star$. Now we denote $\mathcal{C}_1 = \mathcal{C} \cap \{v : \|v\|_2 = 1\}$, and $\theta_{\star 1} = (\beta_\star, U_1)$, $\theta_{\star 2} = (\beta_\star, U_2)$, where $\beta_\star$ is the true covariate coefficient vector, $U_1, U_2$ are any $q$-dimensional standard Gaussian random vectors with a noise level $\sigma$ to be specified.

We have:

$$
\begin{aligned}
|\underline{\nu}_{\mathcal{C}}(U_1) - \underline{\nu}_{\mathcal{C}}(U_2)| &= \left| \inf_{v \in \mathcal{C}_1} \left\{ v^T \left( X^T W_{\theta_{\star 1}} X \right) v \right\} - \inf_{v \in \mathcal{C}_1} \left\{ v^T \left( X^T W_{\theta_{\star 2}} X \right) v \right\} \right| / n \\
&\leq \left| v_2^T X^T W_{\theta_{\star 1}} X v_2 - v_2^T X^T W_{\theta_{\star 2}} X v_2 \right| / n \\
&\left( \inf_{v \in \mathcal{C}_1} \left\{ v^T \left( X^T W_{\theta_{\star 2}} X \right) v \right\} \text{ is attainable at some } v_2 \in \mathcal{C}_1 \right) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left| (W_{\theta_{\star 1}})_{ii} - (W_{\theta_{\star 2}})_{ii} \right| \cdot \left[ (X v_2)_i \right]^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left| g_i'''(t_{mi}) \right| \left| \langle z_{i,\cdot}, U_1 - U_2 \rangle \right| \left[ (X v_2)_i \right]^2 \\
&\left( \text{there exists vector } t_{mi} \text{ lies between } \theta_{\star 1} \text{ and } \theta_{\star 2} \right) \\
&\leq \frac{\max_{1 \leq i \leq n} \left\{ \left[ (X v_2)_i \right]^2 \right\}}{n} \cdot \sum_{i=1}^{n} g_i''(t_{mi}) \left| \langle z_{i,\cdot}, U_1 - U_2 \rangle \right| \\
&\leq \frac{c}{4} \max_{1 \leq i \leq n} \left| \langle z_{i,\cdot}, U_1 - U_2 \rangle \right|
\end{aligned}
\tag{3.45}
$$

Where $c = \left\| \left| X v_2 \right| \right\|_\infty^2 > 0$ in the above.

Now, let $\theta_{\star 1} = \theta_\star = (\beta_\star, U_\star)$, which is the true parameter vector; and $\theta_{\star 2} = (\beta_\star, 0)$, a deterministic parameter vector which consists the true model fixed effect covariate coefficients $\beta_\star$.

Let $\underline{\nu}_{\mathcal{C}}(0) := \inf_{v \in \mathcal{C}, \|v\|_2 = 1} \left\{ v^T \left( X^T W_{(\beta_\star, 0)} X \right) v \right\} / n$, which is a deterministic quantity. It is known that if matrix $X \in \mathbb{R}^{n \times p}$ is formed by independently sampling each row $X_i \sim \mathsf{N}(0, \Sigma)$, which is referred to as the $\Sigma$-*Gaussian ensemble*, then with high probability we have $\underline{\nu}_{\mathcal{C}}(0) > 0$. (*Raskutti et al.* (2010); *Negahban et al.* (2012)). *Rudelson and Zhou* (2011) extends this result to the cases of sub-Gaussian designs, allowing substantial dependencies among the covariates, such that sub-Gaussian ensembles $X$ also has its corresponding $\underline{\nu}_{\mathcal{C}}(0) > 0$.

Let $t$ be any positive number, we inspect the following:

$$\begin{aligned}
\mathbb{P}\left[|\underline{\nu}_{\mathcal{C}}(U_\star) - \underline{\nu}_{\mathcal{C}}(0)| \leq t\right] &\leq \mathbb{P}\left[\max_{1\leq i\leq n} |\langle z_{i,\cdot}, U_\star\rangle| \leq \frac{4t}{c}\right] \\
&\leq \sum_{i=1}^{n} \mathbb{P}\left[|\langle z_{i,\cdot}, U_\star\rangle| \leq \frac{4t}{c}\right] \text{ (union bound)} \\
&\leq 2\sum_{i=1}^{n} \exp\left(-\frac{8t^2}{c^2\sigma^2 \left\|z_{i,\cdot}\right\|_2^2}\right) \text{ (Gaussian tail bound)} \\
&\leq 2\exp\left(\log(n) - \frac{8t^2}{c^2\sigma^2 q\bar{\nu}_Z^2}\right)
\end{aligned} \tag{3.46}$$

Where $\bar{\nu}_Z^2 = \left\|\left|Z\right|\right\|_\infty^2$ and $c = \left\|\left|X_A\underline{\nu}_{\mathcal{C}}(0)\right|\right\|_\infty^2 > 0$ in the above. So if one chooses the standard deviation $\sigma$ of the random effect variable $U_\star$ as $\sigma \leq 2t/(c\bar{\nu}_Z\sqrt{q\log(n)})$, and let $t = \underline{\nu}_{\mathcal{C}}(0)/2$, then we have:

$$\mathbb{P}\left[|\underline{\nu}_{\mathcal{C}}(U_\star) - \underline{\nu}_{\mathcal{C}}(0)| \leq \frac{\underline{\nu}_{\mathcal{C}}(0)}{2}\right] \leq \frac{2}{n} \tag{3.47}$$

That is, if one can choose the random effect variable $U_\star \sim \mathsf{N}(0, \sigma^2 I_q)$ such that $\sigma \leq \frac{\underline{\nu}_{\mathcal{C}}(0)}{c\bar{\nu}_Z\sqrt{q\log(n)}}$, then with probability at least $1 - \frac{2}{n}$, $\underline{\nu}_{\mathcal{C}}(U_\star) \geq \underline{\nu}_{\mathcal{C}}(0)/2 > 0$. So we find the positive constant to be $\underline{\nu}_{\mathcal{C}} = \underline{\nu}_{\mathcal{C}}(0)/2$, such that $\underline{\nu}_{\mathcal{C}}(U_\star) \geq \underline{\nu}_{\mathcal{C}} > 0$ with high probability. □

□

## 3.4 Numerical Simulation

In the simulation study for the *fixed effect approximation algorithm*, we have generated data $X, Z, Y$ and parameters $\beta_\star, U_\star$ according to section 7.1 in Chapter 1. In the first simple simulation study below, we set sample size $N = 200$, fixed effect

size (problem dimension) $p = 50$, in which randomly selected 5 $\beta_{\star j}$'s are non-zero. The number of non-zero singular values of the Gaussian random effects covariance matrix is $q = 2$.

The following plots presents the simulation results, based on 18 repeated runs on independently generated data sets. The $x$-axis denotes each of the independent run, while the $y$-axis denotes different performance metrics.

Compared with the two major algorithms in chapter 1, the *fixed effect approximation algorithm* performs similarly in terms of estimation errors and sparsity recovery. The average estimation errors of the fixed effect approximate solutions are around 0.65, while those of the stochastic proximal gradient and second order approximate solutions are around 0.75. Its sparsity recovery is comparable to those of the stochastic proximal gradient and second order approximate algorithms. As sensitivity of a solution captures"how much true effects (non-zero coefficients) does the algorithm find", there is only one instance out of 18 runs the fixed effect approximate solution missed one true effect, which is in general performing slightly better than the second order approximate and stochastic proximal gradient algorithms. The precision of a solution measures"among those non-zero coefficients in the solution, how much are true effects", and we can see that the fixed effect approximate solution is over covering the non-zero coefficients in a few cases, slightly more than that of the other two algorithms, but overall performs similarly.

Next, we will conduct a comprehensive numerical experiment to explore the performance of the fixed effects approximate algorithm with respect to different problem dimensions $p$, random effect rank $q$, and random effect noise level $\sigma$. As our estimation consistency theory points out that these three design quantities affect the fixed

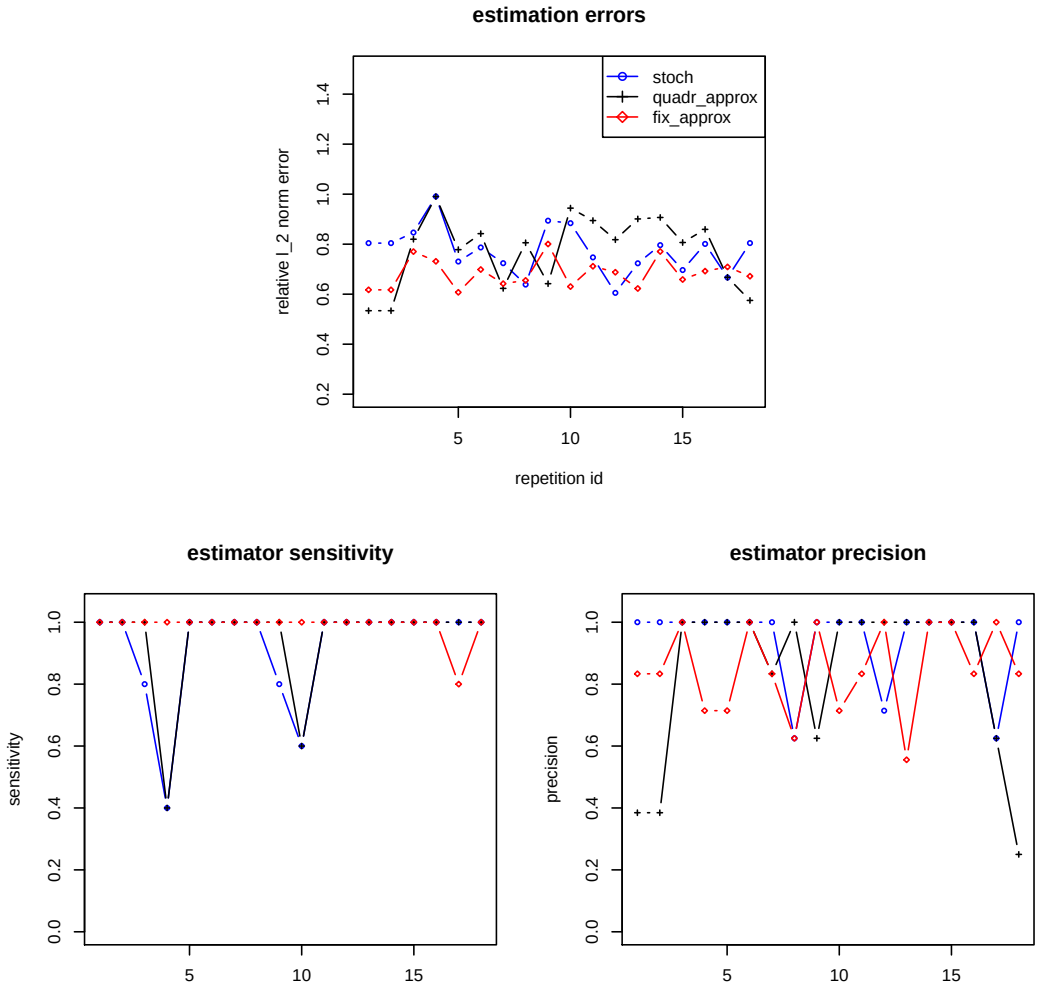


Figure 3.1: $N200p50s5sigma1.5$ step size $\gamma = 0.005$ Stochastic proximal gradient, second order (quadratic) approximate, and fixed effects approximate algorithms.

effects approximate solution performance the most.

We generate data as before, but let the training sample size equals 200, and testing sample size equals 100, 5 non-zero true fixed effects in all cases. We run one experiment on each of the following problem design: fixed effect dimensions

$p = 50, 100, 200, 250$, random effect ranks $q = 2, 5, 7, 10$, and random effect noise levels $\sigma = 0.1, 1.5, 2.5, 3.5$. So in total, for each of the three algorithms: fixed effects approximate (FEA), stochastic proximal gradient (SPG), second order approximate (SOA), we have $4 \times 4 \times 4 = 64$ different settings. The results are tabulated below.

Table 3.1: Relative estimation error of fixed effect approximate (FEA) vs. stochastic proximal gradient (SPG) and second order approximate (SOA) algorithms

| | $p = 50$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 0.1$ | | | $\sigma = 1.5$ | | | $\sigma = 2.5$ | | | $\sigma = 3.5$ | | |
| | FEA | SPG | SOA | FEA | SPG | SOA | FEA | SPG | SOA | FEA | SPG | SOA |
| $q = 2$ | 0.63 | 0.56 | 0.56 | 0.57 | 0.64 | 0.67 | 0.58 | 0.58 | 0.75 | 0.73 | 0.72 | 0.82 |
| 5 | 0.61 | 0.53 | 0.53 | 0.55 | 0.73 | 0.71 | 0.66 | 0.94 | 0.93 | 0.68 | 1.0 | 1.0 |
| 7 | 0.44 | 0.39 | 0.39 | 1.0 | 0.73 | 0.83 | 1.0 | 0.96 | 0.95 | 1.0 | 1.0 | 1.0 |
| 10 | 0.57 | 0.80 | 0.43 | 0.69 | 0.67 | 0.70 | 0.71 | 0.68 | 0.95 | 0.92 | 1.0 | 0.97 |
| | $p = 100$ | | | | | | | | | | | |
| $q = 2$ | 0.62 | 0.55 | 0.45 | 0.57 | 0.39 | 0.67 | 0.73 | 0.74 | 0.84 | 0.58 | 0.69 | 0.86 |
| 5 | 0.65 | 0.59 | 0.59 | 0.60 | 0.58 | 0.82 | 0.52 | 0.79 | 0.81 | 0.77 | 0.72 | 0.96 |
| 7 | 0.55 | 0.35 | 0.56 | 0.87 | 0.87 | 0.88 | 0.94 | 0.78 | 0.99 | 1.0 | 1.0 | 1.0 |
| 10 | 0.63 | 0.31 | 0.31 | 0.79 | 0.73 | 0.84 | 0.88 | 0.87 | 0.92 | 1.0 | 0.95 | 0.96 |
| | $p = 200$ | | | | | | | | | | | |
| $q = 2$ | 0.50 | 0.40 | 0.40 | 0.57 | 0.56 | 0.62 | 0.58 | 0.63 | 0.64 | 0.58 | 0.69 | 0.66 |
| 5 | 0.80 | 0.64 | 0.64 | 0.81 | 0.72 | 0.73 | 0.68 | 0.79 | 0.86 | 0.83 | 1.0 | 0.92 |
| 7 | 0.36 | 0.37 | 0.44 | 0.87 | 0.69 | 0.82 | 0.74 | 0.72 | 0.98 | 1.0 | 0.76 | 1.0 |
| 10 | 0.64 | 0.43 | 0.56 | 0.86 | 0.81 | 0.86 | 1.0 | 0.90 | 0.92 | 1.0 | 0.90 | 0.94 |
| | $p = 250$ | | | | | | | | | | | |
| $q = 2$ | 0.67 | 0.68 | 0.55 | 0.57 | 0.64 | 0.64 | 0.70 | 0.65 | 0.70 | 0.70 | 0.79 | 0.82 |
| 5 | 0.70 | 0.37 | 0.59 | 0.73 | 0.74 | 0.76 | 0.90 | 0.81 | 0.91 | 1.0 | 0.90 | 1.0 |
| 7 | 0.65 | 0.66 | 0.59 | 0.70 | 0.69 | 0.82 | 0.88 | 0.96 | 0.95 | 1.0 | 1.0 | 1.0 |
| 10 | 0.72 | 0.52 | 060 | 0.85 | 0.82 | 0.80 | 0.83 | 0.92 | 0.93 | 0.87 | 0.92 | 0.91 |

The $\ell_2$ norm relative estimation error is defined as $\left\|\widehat{\beta} - \beta_\star\right\|_2 / \|\beta_\star\|_2$, where $\beta_\star$ is the true parameter vector. Reading the above results corresponding to our estimation error bound theorem III.1, we recall that for generated design matrices $X$ and $Z$, $X_A = (X, Z)$, and $|||X_A|||_\infty = 3.81$ when $p = 50$, $|||X_A|||_\infty = 4.55$ when $p = 250$. The error bounds $\left\|\widehat{\beta}_\lambda - \beta_\star\right\|_2 \leq \frac{48}{\nu_{\mathcal{C}}} \sqrt{\frac{2|||X_A|||_\infty s_\star \log(p+q)}{n}}$ have been well satisfied. We also observe that for the fixed sample size $n = 200$ and true effect size $s_\star = 5$,

when the random effect factor dimension $q$ increases, smaller random effect noise $\sigma$ will in general have lower estimation errors, this is in line with the requirement that $\sigma < \frac{2\nu_c}{c\|Z\|_\infty \sqrt{q\log(n)}}$ for the estimation error bound to hold. Also, from the error bound we can say that when problem dimension $p$ increases, the upper bound will enlarge, so as the actual estimation errors show the increasing trend.

Table 3.2: Harmonic Mean of Sensitivity and Precision for FEA, SPG and SOA algorithms

| | $p = 50$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 0.1$ | | | $\sigma = 1.5$ | | | $\sigma = 2.5$ | | | $\sigma = 3.5$ | | |
| | FEA | SPG | SOA | FEA | SPG | SOA | FEA | SPG | SOA | FEA | SPG | SOA |
| $q = 2$ | 0.67 | 0.48 | 0.48 | 0.63 | 0.83 | 0.71 | 0.83 | 0.83 | 0.91 | 1.0 | 1.0 | 0.91 |
| 5 | 0.47 | 0.38 | 0.38 | 0.77 | 0.91 | 0.91 | 0.83 | 0.57 | 0.75 | 0.91 | NaN | NaN |
| 7 | 0.53 | 0.48 | 0.77 | NaN | 0.43 | 0.77 | NaN | 0.50 | 0.46 | NaN | NaN | NaN |
| 10 | 0.67 | 0.18 | 0.48 | 0.77 | 0.71 | 0.77 | 0.83 | 0.59 | 0.57 | 0.67 | NaN | 0.50 |
| | $p = 100$ | | | | | | | | | | | |
| $q = 2$ | 0.67 | 0.59 | 0.42 | 0.63 | 0.22 | 0.59 | 0.83 | 0.89 | 0.80 | 1.0 | 0.67 | 0.73 |
| 5 | 0.71 | 0.53 | 0.53 | 0.40 | 0.23 | 0.80 | 0.31 | 0.83 | 0.62 | 0.80 | 0.10 | 0.33 |
| 7 | 0.53 | 0.18 | 0.53 | 0.80 | 0.77 | 0.62 | 0.57 | 0.14 | 0.28 | NaN | NaN | NaN |
| 10 | 0.77 | 0.20 | 0.20 | 0.75 | 0.83 | 0.53 | 0.75 | 0.67 | 0.46 | NaN | 0.29 | 0.44 |
| | $p = 200$ | | | | | | | | | | | |
| $q = 2$ | 0.36 | 0.26 | 0.26 | 0.30 | 0.30 | 0.42 | 0.32 | 0.42 | 0.45 | 0.37 | 0.67 | 0.50 |
| 5 | 0.89 | 0.26 | 0.26 | 0.75 | 0.47 | 0.42 | 0.44 | 0.73 | 0.60 | 0.57 | NaN | 0.55 |
| 7 | 0.20 | 0.17 | 0.26 | 0.80 | 0.15 | 0.57 | 0.53 | 0.18 | 0.67 | NaN | 0.23 | 0.29 |
| 10 | 0.59 | 0.13 | 0.29 | 0.75 | 0.73 | 0.75 | NaN | 0.50 | 0.57 | NaN | 0.55 | 0.50 |
| | $p = 250$ | | | | | | | | | | | |
| $q = 2$ | 0.53 | 0.47 | 0.22 | 0.34 | 0.45 | 0.48 | 0.91 | 0.56 | 0.91 | 0.67 | 0.91 | 1.0 |
| 5 | 0.77 | 0.15 | 0.37 | 0.50 | 0.42 | 0.43 | 0.67 | 0.38 | 0.50 | NaN | 0.50 | NaN |
| 7 | 0.71 | 0.63 | 0.36 | 0.50 | 0.28 | 0.59 | 0.62 | 0.04 | 0.50 | NaN | NaN | NaN |
| 10 | 0.83 | 0.18 | 0.30 | 0.91 | 0.67 | 0.71 | 0.53 | 0.71 | 0.60 | 0.62 | 0.73 | 0.62 |

The harmonic mean of sensitivity and precision is defined as

$$\frac{2}{(1/sensitivity + 1/precision)} \in (0, 1]$$

, it is a measure of the trade off in recovering the true non-zero effects while main-

taining the model to be sparse. A value of the harmonic mean closer to 1 indicates the sparsity pattern is closer to the true parameter vector. If this value attains 1, then the solution exactly recovers the non-zero true effects. From the above results, we can conclude that when random effect noise level $\sigma$ is small, the sparsity recovery is in general better for fixed effects approximate solutions.

From other numerical experiments we have carried out for the three algorithms, we have found that the stochastic proximal gradient and second order approximate algorithms generally perform better than fixed effect approximate model based algorithm when the problem dimension $p$ is large, say $p = 2000$, especially when random effects dimension $q$ are large to be around $30 \sim 50$.

## 3.5 Real Data Analysis

We apply the fixed effects approximate algorithm to the same breast cancer data we have analyzed in chapter II, 2.6.

Recall that the original data set (*Vijver et al.* (2002)) includes 295 patients consecutively enrolled. There are 24496 gene expression intensity measurements to start with. Our pre-processing screening has selected 1083 genes's expression measurements as the fixed effects, on top of a few clinical variables "ESR1", "NIH", "StG", and "Posnodes" as the clinical characteristic for each patient. To model the 5-year distant metastasis event, which is coded $\{0, 1\}$ by the mixed effects logistic regression model, the Gaussian random effects variance-covariance matrix $Z^T Z$ is similarly generated as in chapter II.

To describe the fitting and model selection schemes, the fixed effects approximate algorithm involves Lasso regularization, and model selection is done by solving a sequence of Lasso regularized optimization problems with different penalty amount $\lambda$'s, this is usually called "regularization path" in the literature (*Friedman et al.* (2010a)). For a given sequence of lambda, we run the algorithm for each lambda, and plot the solution path along the sequence of lambdas from largest to smallest. We will regard the genes that constantly stays in the solution path to be potential prognostic signatures.

In the following solution path plots, we have fixed the random effect factor dimension $q = 2$. We use Lasso regularization with a sequence of $\lambda = 38, 36, 34, \ldots, 12$. Notice the x-axis is order reversed, so that log of lambda values decreases from left to right.

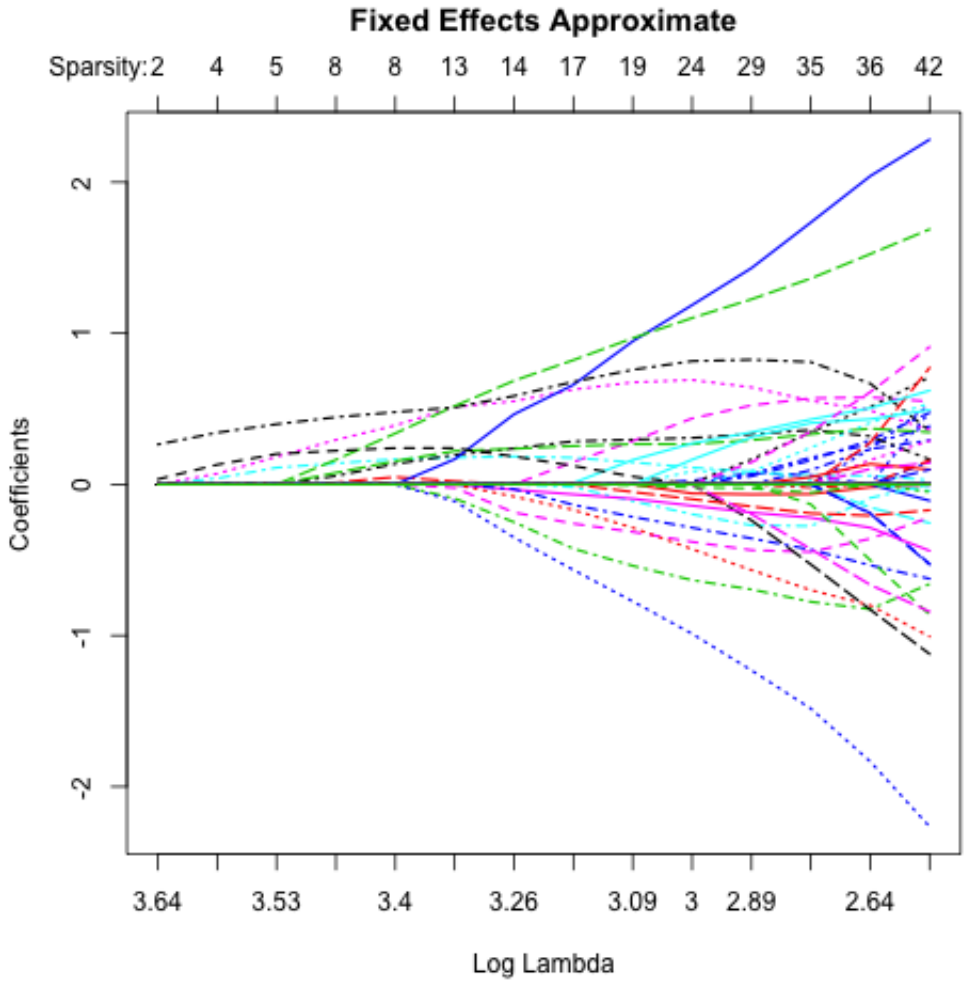


Figure 3.2: Solution paths for mixed effect logistic regression on breast cancer data, $q = 2$

We observe that there are 4 genes selected by FEAME which stay along the solution paths, they are named "NM˙003258", "NM˙003662", "NM˙003981" and "Contig41977˙RC" in the data set; "Contig57584˙RC" appeared in the beginning of the path but subsided later when more gene expressions are selected, while "Contig41977˙RC" pops up when $\lambda$ gets a bit smaller and stays in the path. This result has "Contig41977˙RC" overlapped with those in stochastic proximal gradient and

second order approximate algorithm's solution path. Except the limited overlap of result compared with previous algorithms, we have also noticed that the number of genes selected in the fixed effects approximate algorithm solution path tend to increase more continuously than those of stochastic proximal gradient and second order approximate algorithms, partly because we have treated the random effects as fixed now, which reduces much of the noise ($\sigma = 1.0$ in our experiments) from the random effects in the model. This matters because the gene expression measurements are of a small scale in this data set, and is more prone to noise in the model.

We have also run the algorithm when we set the random effects dimension $q = 5$ and get the following result:

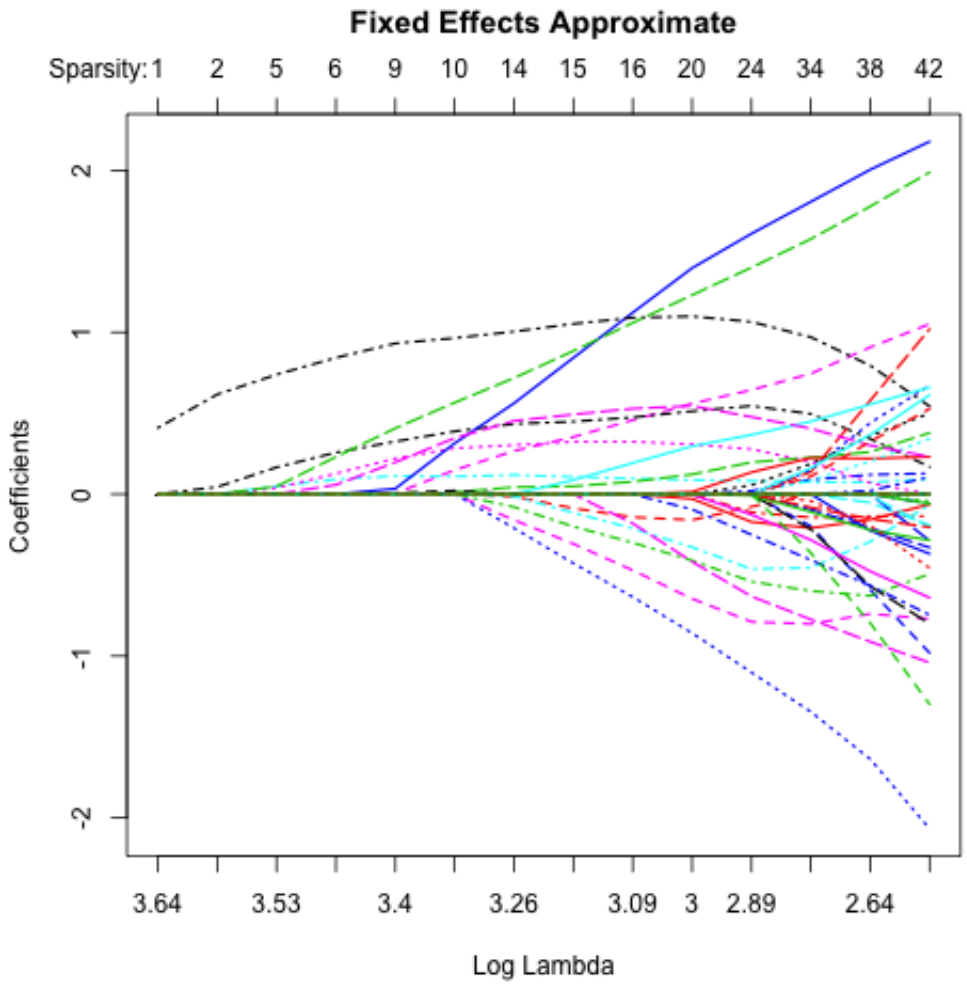


Figure 3.3: Solution paths for mixed effect logistic regression on breast cancer data, $q = 5$

This time the genes selected from the solution path are "NM˙003258", "NM˙003662", "NM˙003981", "NM000903", "M94096", "AF055033" and "Contig41977˙RC" which appears in a later stage when more genes are selected. The overlap with $q = 2$ result are NM˙003258", "NM˙003662", "NM˙003981" and "Contig41977˙RC". Again we only provide these results as contenders for future clinical study, however their validity still needs to be proved by scientific means.

## 3.6 Proofs

We give the proof of Proposition III.3 below:

**Proposition III.6.** *For the function $g_\alpha(h)$ defined in (3.28), we have:*

$$g_\alpha(h) - g_\alpha(0) - g'_\alpha(0)h \geq g''_\alpha(0)\frac{h^2}{|h|+2} \text{ , for all } \alpha, h \in \mathbb{R}$$

*Proof.* Fix any $\alpha \in \mathbb{R}$. For all $h \in \mathbb{R}$, we have

$$\begin{aligned} g'_\alpha(h) &= \frac{\exp(\alpha+h)}{1+\exp(\alpha+h)} \\ g''_\alpha(h) &= g'_\alpha(h)(1-g'_\alpha(h)) \\ g'''_\alpha(h) &= g''_\alpha(h)(1-2g'_\alpha(h)) \end{aligned}$$

Observe that $g'_\alpha(h) \in (0,1)$, $g''_\alpha(h) > 0$. And, $1 - 2g'_\alpha(h) \in (-1,1)$, so $|g'''_\alpha(h)| \leq g''_\alpha(h), \forall h \in \mathbb{R}$. We repeatedly apply the theorem of calculus for the center quantity in $h$ in the following, for all $h > 0$:

$$\begin{aligned} -1 \leq& \left(\log(g''_\alpha(h))\right)' \leq 1 \\ -h \leq& \log\left(\frac{g''_\alpha(h)}{g''_\alpha(0)}\right) \leq h \\ g''_\alpha(0)e^{-h} \leq& g''_\alpha(h) \leq g''_\alpha(0)e^h \\ g''_\alpha(0)(1-e^{-h}) \leq& g'_\alpha(h) - g'_\alpha(0) \leq g''_\alpha(0)(e^h-1) \\ g''_\alpha(0)(e^{-h}+h-1) \leq& g_\alpha(h) - g_\alpha(0) - g'_\alpha(0)h \leq g''_\alpha(0)(e^h-h-1) \end{aligned} \tag{3.48}$$

On another hand, for all $h \leq 0$, similar to the above procedure we have

$$g''_\alpha(0)(e^h-h-1) \leq g_\alpha(h) - g_\alpha(0) - g'_\alpha(0)h \leq g''_\alpha(0)(e^{-h}+h-1) \tag{3.49}$$

So for all $h \in \mathbb{R}$ we have

$$g''_\alpha(0)(e^{-|h|} + |h| - 1) \leq g_\alpha(h) - g_\alpha(0) - g'_\alpha(0)h \leq g''_\alpha(0)(e^{|h|} - |h| - 1) \tag{3.50}$$

In addition, we verify that the following holds for all $x \geq 0$:

$$e^{-x} + x - 1 \geq x^2/(x+2)\,. \tag{3.51}$$

Denote

$$f(x) = e^{-x} + x - 1 - \frac{x^2}{x+2}$$

Note that $f(0) = 0$, and for $x > 0$, we have

$$\begin{aligned} f'(x) &= \frac{4}{(x+2)^2} - e^{-x} \\ &= \frac{4e^x - x^2 - 4x - 4}{e^x(x+2)^2} \\ &= \frac{4\sum_{i=0}^{\infty} x^i/i! - x^2 - 4x - 4}{e^x(x+2)^2} \\ &> \frac{x^2}{e^x(x+2)^2} \\ &> 0 \end{aligned}$$

So for all $h \in \mathbb{R}$ we have,

$$g(h) - g(0) - g'(0)h \geq g''(0)\frac{h^2}{|h|+2} \tag{3.52}$$

□

# Chapter IV

# Iterated Filtering Algorithms Revisited

## 4.1 Introduction

Iterated filtering algorithms are a class of stochastic algorithms recently proposed in the statistical literature (*Ionides et al.* (2006, 2011, 2015)) to address some uniquely challenging optimization problems that arise when dealing with state space models. A state space model is comprised of a latent (un-observed) state $X_{1:T} = (X_1, \ldots, X_T) \in \mathcal{X}^T$ with distribution $f_\beta$, and an observation variable $Y_{1:T} = (Y_1, \ldots, Y_T) \in \mathcal{Y}^T$ with conditional distribution $q_\beta(\cdot|x)$ given $X_{1:T} = x$. The parameter $\beta \in \Theta \subseteq \mathbb{R}^p$ is unknown and the problem at hand is the estimation of $\beta$ from a realization $y_{1:T}$ of $Y_{1:T}$. Since the state variable is not observed, the log-likelihood function of the model is

$$\ell(\beta) \stackrel{\text{def}}{=} \log \int_{\mathcal{X}^T} q_\beta(y_{1:T}|x_{1:T}) f_\beta(x_{1:T}) \mathrm{d}x_{1:T}. \tag{4.1}$$

State space models are widely used in science and engineering (*Cappé et al.* (2005); *Anderson and Collins* (2007); *Fernandez-Villaverde and Rubio-Ramirez* (2007); *Ergun et al.* (2007); *Newman et al.* (2008)), and the problem of maximizing the log-

likelihood function (4.1) is very common. We should add that the function $\ell$ is not concave in general, so local modes and stationary points are typically the best one can hope for.

The problem of maximizing (4.1) is particularly difficult when dealing with state space models for which the density of the state model $f_\beta$ is intractable (not easily computable). This is the case for instance when the state variable $(X_1, \ldots, X_T)$ is obtained from a diffusion process observed at some discrete time $\{t_1, \ldots, t_T\}$. This type of state variable models are commonly used in the applications (see for instance *Ionides et al.* (2011) and the references therein).

Notice that the integral in (4.1) is intractable in general, so direct access to the function $\ell$ is rarely available. One of the simplest methods available to approach this optimization problem is to approximate $e^\ell$ (the likelihood function) by Monte Carlo (importance sampling) estimate:

$$\widetilde{L_N(\beta)} = \frac{1}{N} \sum_{i=1}^{N} \frac{q_\beta(y_{1:T}|X_{1:T}^{(i)}) f_\beta(X_{1:T}^{(i)})}{p(X_{1:T}^{(i)})}, \quad \text{where} \quad X_{1:T}^{(i)} \overset{i.i.d.}{\sim} p.$$

One can then proceed to maximize $\widetilde{L_N}$ using standard optimization tools. *Fearnhead* (2008) reviews several examples where this approach was successful. However that success hinges on the choice of the proposal density $q$: the method produces terribly large variance unless $p$ is carefully chosen. Sequential Monte Carlo algorithms (instead of importance sampling) typically produce better estimates of $\widetilde{L_N(\beta)}$. But these estimates are typically discontinuous functions of $\beta$. Another issue is the well-known fact that approximating and maximizing the likelihood function $e^\ell$ itself is typically not a numerically stable problem (it is more susceptible to over/under-flow).

Another well-established strategies for maximizing the function $\ell$ is the expectation maximization (EM) algorithm and Stochastic approximation (SA). These al-

gorithms are described at length in *Cappé et al.* (2005). The $Q$ function for the EM algorithm is

$$Q(\beta, \beta') = \int_{\mathcal{X}^T} \log \left[ q_{\beta'}(y_{1:T}|x_{1:T}) f_{\beta'}(x_{1:T}) \right] \pi_\beta(x_{1:T}|y_{1:T}) \mathrm{d}x_{1:T}, \tag{4.2}$$

where $\pi_\beta(x_{1:T}|y_{1:T})$ is the conditional distribution of $X_{1:T}$ given $Y_{1:T} = y_{1:T}$. Similarly the gradient of the log-likelihood function $\ell$ is

$$\nabla \ell(\beta) = \int_{\mathcal{X}^T} \nabla \log \left[ q_\beta(y_{1:T}|x_{1:T}) f_\beta(x_{1:T}) \right] \pi_\beta(x_{1:T}|y_{1:T}) \mathrm{d}x_{1:T}. \tag{4.3}$$

The EM algorithm is based on (4.2); whereas SA uses (4.3). Due to their integral form, neither of these functions is readily available, but Monte Carlo approximation can be obtained by sampling from the filtering distribution $\pi_\beta(x_{1:T}|Y_{1:T})$. This can be done by Markov Chain Monte Carlo (MCMC) or sequential Monte Carlo (SMC). There is a large literature on MCMC/SMC driven EM and SA algorithms for computing stationary points of $\ell$ (*Cappé et al.* (2005)). One important limitation of the EM and SA algorithms is that they cannot be easily applied when dealing with state space models for which the density of the state is intractable.

One clever strategy devised in the finance literature to dealing with the case of discretely observed diffusion is data-augmentation (*Eraker* (2001); *Ola et al.* (2001); *Roberts and Stramer* (2001); *Beskos et al.* (2006)). If $t_1 < \ldots < t_T$ denote the time points of the latent observations $X_{1:T}$, the basic idea is to add more time points to get $t'_1 < \ldots < t'_K$, such that $\{t_1, \ldots, t_T\} \subset \{t'_1, \ldots, t'_K\}$, and such that the Euler scheme approximation of the diffusion process based on $(X_{t'_1}, X_{t'_1}, \ldots, X_{t'_K})$ is reasonably accurate. The EM and SA strategies can then be adapted to the augmented model. The approach has limitation though: the mixing of the resulting algorithm deteriorates with the amount of additional data imputation, and the posterior for

the volatility parameter becomes severely degenerate as the number of augmented variable increases (*Roberts and Stramer* (2001); *Beskos et al.* (2006)).

Iterated filtering algorithms give a simple, yet effective strategy to deal with state space in general. The method is particularly effective in dealing with state space models where the density of the state is intractable. The goal of this work is to give a broad presentation of iterated filtering algorithms, and relate more closely these algorithms to well-known stochastic gradient methods. These new connections will allow us to derive new convergence results for iterated filtering algorithms that are sharper than those of *Ionides et al.* (2011). Although iterated filtering algorithms are commonly used to address nonconvex optimization problems, the theoretical results established here assume strong convexity. The general convex case and the nonconvex case are left as possible future research. By and large the convergence analysis of stochastic optimization algorithms in nonconvex setting remains an open problem.

The rest of the manuscript is organized as follows. In Section 4.2 we introduce iterated algorithms and explores its connection with gradient and proximal algorithms. We focus on the problem of minimizing composite objective functions as commonly seen in high-dimensional statistics, and the main iterated filtering algorithm that we propose is Algorithm 5, as well as its block coordinate version described in Algorithm 6. We illustrate the behavior of the algorithm in Section 4.4, using a mixed effects logistic regression model. In Section 4.3 we establish the convergence of Algorithm 6 under a strong convexity and boundedness assumption. Technical details are gathered in Section 4.5.

## 4.2 Iterated Filtering Algorithms

We consider the problem of minimizing a function $F : \mathbb{R}^p \to (-\infty, +\infty]$, that we think of as a negative log-likelihood function, or a penalized negative log-likelihood function. For $\sigma > 0$, and $u \in \mathbb{R}^p$, let $\mathsf{K}_\sigma(u, \cdot)$ denote the density on $\mathbb{R}^p$ of the normal density $\mathbf{N}(u, \sigma^2 I_p)$. Given $\sigma > 0$, and $\beta \in \mathbb{R}^p$, we define

$$\mathsf{B}^F_{\sigma,\beta}(z) \stackrel{\text{def}}{=} \frac{e^{-F(z)}\mathsf{K}_\sigma(\beta, z)}{\int_{\mathbb{R}^p} e^{-F(u)}\mathsf{K}_\sigma(\beta, u)\mathrm{d}u}, \quad z \in \mathbb{R}^p.$$

And we introduce the map $\Pi^F_\sigma : \mathbb{R}^p \to \mathbb{R}^p$ by

$$\Pi^F_\sigma(\beta) \stackrel{\text{def}}{=} \int z\mathsf{B}^F_{\sigma,\beta}(z)\mathrm{d}z.$$

The map $\Pi^F_\sigma$ is closely related to the proximal map of $F$ defined as

$$\mathrm{Prox}^F_\sigma(\beta) \stackrel{\text{def}}{=} \mathsf{Argmin}_{u\in\mathbb{R}^p} \left[F(u) + \frac{1}{2\sigma^2}\|u - \beta\|^2\right] = \mathsf{Argmin}_{u\in\mathbb{R}^p} \mathsf{B}_{\sigma,\beta}(u).$$

In other words $\Pi^F_\sigma(\beta)$ is the mean of $\mathsf{B}^F_{\sigma,\beta}$, whereas $\mathrm{Prox}^F_\sigma(\beta)$ is its mode. Therefore, we shall sometimes refer to the map $\Pi^F_\sigma$ as the pseudo-proximal map of $F$. It is well known that one can approximate the minimizer of $F$ by iterating the proximal map $\mathrm{Prox}^F_\sigma$ (*Parikh and Boyd* (2013b)). Such iterations schemes are also known as implicit gradient schemes. When $F$ is differentiable, its minimizers $F$ can also be found by iterating the gradient map

$$G^F_\sigma(\beta) \stackrel{\text{def}}{=} \beta - \sigma^2\nabla F(\beta), \quad \beta \in \mathbb{R}^p, \tag{4.4}$$

where $\nabla F$ denotes the gradient of $F$. Such schemes are also known as explicit gradient schemes. We introduce here the pseudo-proximal $\Pi^F_\sigma$ as an alternative to

the proximal and gradient maps. The following result initially due to *Ionides et al.* (2011) and improved by *Doucet et al.* (2013) show that $\Pi_\sigma^F$ is closely related to the gradient map $G_\sigma^F$.

**Proposition IV.1.** *Suppose that $F$ is four times continuously differentiable. Then for any compact set $\mathcal{C} \subset \mathbb{R}^p$, we can find $\sigma_0 > 0$ $c > 0$ such that*

$$\sup_{0<\sigma\leq\sigma_0} \sup_{\beta\in\mathcal{C}} \|\Pi_\sigma^F(\beta) - G_\sigma^F(\beta)\|_2 \leq c\sigma^4.$$

*Proof.* See Theorem 1 of *Doucet et al.* (2013). □

The next result shows that when $\sigma$ is small, the pseudo-proximal map $\Pi_\sigma^F$ and the proximal map $\mathrm{Prox}_\sigma^F$ are also close.

**Proposition IV.2.** *Suppose that $F$ is differentiable and its gradient is Lipschitz with constant $L$. Then for all $\beta \in \Theta$, and all $\sigma > 0$ such that $\sigma^2 L \leq 1$,*

$$\|\Pi_\sigma(\beta) - \mathrm{Prox}_\sigma(\beta)\| \leq \sigma\sqrt{p}\left(1 + L\sigma^2\right)^{p/4}. \tag{4.5}$$

*Proof.* See Section 4.5.1. □

There are several classes of problems – in state space modeling and more generally in modeling with latent variables – where the map $\Pi_\sigma^F$ proves much easier to approximate by Monte Carlo. Indeed, one can easily approximate $\Pi_\sigma^F$ by the importance sampling estimate

$$H_{\sigma,N}^F(\beta) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^N \vartheta_i e^{-F(\vartheta_i)}}{\sum_{i=1}^N e^{-F(\vartheta_i)}}, \quad \text{where} \quad \vartheta_{1:N} \overset{i.i.d.}{\sim} \mathsf{K}_\sigma(\beta, \cdot). \tag{4.6}$$

In the last display the notation $U_{1:K}$ is a short for the vector $(U_1, \ldots, U_K)$. The performance of iterated filtering algorithms hinges on the fact $H_{\sigma,N}^F(\beta)$ is a good

approximation for $\Pi_\sigma^F(\beta)$, for large $N$. This is summarized in the next result.

**Lemma IV.3.** *Let $\mathcal{C} \subset \mathbb{R}^p$ be a compact set. Then there exists $\sigma_0 > 0$, and a finite constant $c_0$ such that*

$$\sup_{0<\sigma<\sigma_0} \sup_{\beta\in\mathcal{C}} \left|\mathbb{E}\left[H_{\sigma,N}^F(\beta) - \Pi_\sigma^F(\beta)\right]\right| + \mathbb{E}\left[\left(H_{\sigma,N}^F(\beta) - \Pi_\sigma^F(\beta)\right)^2\right] \leq \frac{c_0}{N}.$$

*Proof.* This follows from Theorem 7 of *Ionides et al.* (2011). □

This leads to the following stochastic algorithm to minimize $F$. Let $\{\sigma_k,\ k \geq 0\}$ be a sequence of positive numbers, and $\{N_k,\ k \geq 0\}$ a sequence of integers.

**Algorithm 4** Iterated Filtering Algorithm I

Given $\beta^{(k)}$ generate $\vartheta_{1:N_k} \overset{i.i.d.}{\sim} \mathsf{K}_{\sigma_k}(\beta^{(k)}, \cdot)$, and compute

$$\beta^{(k+1)} = H_{\sigma_k,N_k}^F(\beta^{(k)}).$$

*Remark* IV.4. Variants of this algorithm can be easily constructed depending on the application. For example, in the case of the state space model discussed in the introduction with log-likelihood function given in (4.1), the map $\Pi_\sigma^F$ takes the form

$$\frac{\int_\Theta \int_{\mathcal{X}^T} \vartheta q_\vartheta(y_{1:T}|x_{1:T}) \mathsf{K}_\sigma(\beta,\vartheta) f_\vartheta(x_{1:T}) \mathrm{d}x_{1:T} \mathrm{d}\vartheta}{\int_\Theta \int_{\mathcal{X}^T} q_\vartheta(y_{1:T}|x_{1:T}) \mathsf{K}_\sigma(\beta,\vartheta) f_\vartheta(x_{1:T}) \mathrm{d}x_{1:T} \mathrm{d}\vartheta},$$

which can be approximation for instance by importance sampling by drawing $\vartheta_j \sim \mathsf{K}_\sigma(\beta, \cdot)$, and $(X_{1:T})_j|\vartheta_j \sim f_{\vartheta_j}(\cdot)$ for $j = 1, \ldots, N$, and taking

$$\frac{\sum_{j=1}^N \vartheta^{(j)} q_{\vartheta^{(j)}}(y_{1:T}|X_{1:T}^{(j)})}{\sum_{j=1}^N q_{\vartheta^{(j)}}(y_{1:T}|X_{1:T}^{(j)})}.$$

Note that this estimator does not require the computation of the density of the latent variable $X_{1:T}$, it requires only the ability to sample from it. The temporal dynamics

of the state space can be further exploited to construct more robust sequential Monte Carlo sampler approximation of $\Pi_\sigma^F(\beta)$. We refer the reader to *Ionides et al.* (2011) for more details.

*Remark* IV.5. It is worth pointing out that Algorithm 4 differs slightly from the original iterated filtering algorithm of *Ionides et al.* (2006, 2011). Indeed, in these works Proposition IV.1 is used to approximate the gradient $\nabla F(\beta)$ by

$$\widehat{\nabla F}(\beta) \stackrel{\text{def}}{=} \frac{1}{\sigma^2}\left(\beta - H_{\sigma,N}^F(\beta)\right),$$

which is then used in a standard gradient update with step-size $\gamma > 0$:

$$\beta^{(k+1)} = \beta^{(k)} - \gamma\widehat{\nabla F}(\beta^{(k)}).$$

For $\gamma = \sigma^2$, one recovers the same iteration as in Algorithm 4, however Proposition IV.1 shows that this strategy is redundant. Furthermore, the computation of $\widehat{\nabla F}(\beta)$ can be unstable when $\sigma$ is small.

### 4.2.1 The Case of Composite Function

In many problems of interest the function $F$ takes the form

$$F = f + g,$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is a smooth function, and $g : \mathbb{R}^p \to (-\infty, +\infty]$ is non-smooth but is simple enough for its proximal – denoted $\text{Prox}_\sigma^g$ – to be easily computed. In this setting the proximal map of $F$ itself is typically intractable. The hugely successful forward-backward splitting algorithm (*Beck and Teboulle* (2010); *Combettes and Pesquet* (2015b); *Parikh and Boyd* (2013b)) comes to the rescue, and leads to the

iterations

$$\beta^{(k)} = \text{Prox}_{\sigma}^{g}\left(\beta^{(k-1)} - \sigma^2 \nabla f(\beta^{(k-1)})\right).$$

In many statistical problems involving latent variables, the gradient $\nabla f(\beta)$ is typically intractable and is approximated by Monte Carlo or Markov Chain Monte Carlo simulation. The resulting stochastic optimization algorithm has been investigated by several authors in recent year (*Rosasco et al.* (2014); *Combettes and Pesquet* (2015a); *Atchadé et al.* (2017)). However this strategy can prove difficult in latent variable models where the density of the latent variable is intractable. We propose an iterated filtering algorithms whereby we replace the gradient map update $G_{\sigma}^{f}(\beta)$ by the pseudo-proximal update $\Pi_{\sigma}^{f}(\beta)$, leading to the deterministic iteration

$$\beta^{(k)} = \text{Prox}_{\sigma}^{g}\left(\Pi_{\sigma}^{f}(\beta^{(k-1)})\right).$$

If we approximate the pseudo-proximal map $\Pi_{\sigma}^{f}(\beta)$ by its Monte Carlo estimate as in Algorithm 4, we obtain the following stochastic algorithm.

**Algorithm 5** Iterated Filtering Algorithm II : Composite Objective Function

Given $\beta^{(k)}$, generate $\vartheta_{1:N_k} \overset{i.i.d.}{\sim} \mathsf{K}_{\sigma_k}(\beta^{(k)}, \cdot)$, set $H^{(k+1)} = \frac{\sum_{i=1}^{N_k} \vartheta_i e^{-f(\vartheta_i)}}{\sum_{i=1}^{N_k} e^{-f(\vartheta_i)}}$, and compute

$$\beta^{(k+1)} = \text{Prox}_{\sigma_k}^{g}\left(H^{(k+1)}\right).$$

### 4.2.2 Bloc Update Implementation

For large scale problems, it may be advantageous to use a block update strategy to minimize $F$. We consider again the case where $F$ is a composite function $F = f+g$, and $f(\beta) = f(\beta_1, \beta_2)$, and $g(\beta_1, \beta_2) = g_1(\beta_1) + g_2(\beta_2)$. We focus on two blocks, but the idea can be readily extended to any finite number of blocks. Suppose that the dimension of $\beta_i$ is $p_i$. For $\sigma > 0$, and $u_i \in \mathbb{R}^{p_i}$, let $\mathsf{K}_{i,\sigma}(u_i, \cdot)$ denote the density on

$\mathbb{R}^{p_i}$ of the normal density $\mathbf{N}(u_i, \sigma^2 I_{p_i})$. Given $\beta_i \in \mathbb{R}^{p_i}$, we define

$$H^{(1,f)}_{\sigma,N}(\beta_1,\beta_2) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^N \vartheta_i e^{\ell(\vartheta_i^{(1)},\beta_2)}}{\sum_{i=1}^N e^{\ell(\vartheta_i^{(1)},\beta_2)}}, \quad \text{and} \quad H^{(2,f)}_{\sigma,N}(\beta_1,\beta_2) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^N \vartheta_i e^{\ell(\beta_1,\vartheta_i^{(2)})}}{\sum_{i=1}^N e^{\ell(\beta_1,\vartheta_i^{(2)})}},$$

where $\vartheta^{(i)}_{1:N} \overset{iid}{\sim} \mathsf{K}_{i,\sigma}(\beta_i, \cdot)$, $i = 1, 2$.

**Algorithm 6** Block Update Iterated Filtering Algorithm: Composite Objective Function

Given $\beta^{(k)} = (\beta_1^{(k)}, \beta_2^{(k)})$:

1. generate $\vartheta^{(1)}_{1:N_k} \overset{i.i.d.}{\sim} \mathsf{K}_{1,\sigma_k}(\beta_1^{(k)}, \cdot)$, and compute $H_1^{(k+1)} \stackrel{\text{def}}{=} H^{(1,f)}_{\sigma_k,N_k}(\beta_1^{(k)}, \beta_2^{(k)})$,

$$\beta_1^{(k+1)} = \text{Prox}^{g_1}_{\sigma_k}\left(H_1^{(k+1)}\right).$$

2. generate $\vartheta^{(2)}_{1:N} \overset{i.i.d.}{\sim} \mathsf{K}_{2,\sigma_k}(\beta_2^{(k)}, \cdot)$, and compute $H_2^{(k+1)} \stackrel{\text{def}}{=} H^{(2,f)}_{\sigma_k,N_k}(\beta_1^{(k+1)}, \beta_2^{(k)})$

$$\beta_2^{(k+1)} = \text{Prox}^{g_2}_{\sigma_k}\left(H_2^{(k+1)}\right).$$

## 4.3 Some Theory

We study here the convergence of Algorithm 6. Block coordinate descent algorithms have attracted a lot of attention in recent years due due their ability to deal with very large problems. The analysis of these algorithms has been considered by several authors (*Saha and Tewari* (2013); *Beck and Tetruashvili* (2013); *Bolte et al.* (2014)) for convex and nonconvex problems. However stochastic version of these algorithms have received comparatively little attention[1]. We study Algorithm 6 by adapting ideas from *Atchadé et al.* (2017). We make the simplifying assumption

[1]By stochastic we mean that the gradient update is stochastic, as opposed to stochastic block coordinate descent algorithms where the randomness comes from a random selection of the blocks. This latter class of algorithms have also been extensively studied in recent year (see for instance *Richtárik and Takáč* (2014) and the references therein). These two types of stochastic block coordinate descent algorithms lead to very different challenges

that the function $g$ is convex, and $f$ is strongly convex, even though this assumption does not hold in general with latent variable models. Convergence analysis under convexity assumption can still be useful in nonconvex settings to understand the local behavior of the algorithm around local modes. We should also note that the ideas developed in the first part of the thesis can also be used here to show that in the nonconvex case limit points of the optimization sequences are stationary points. However we shall not pursue this here.

We assume the function $\ell$ satisfies the following.

**Assumption IV.6.** *The function $g_i : \mathbb{R}^{p_i} \to (-\infty, +\infty]$ is convex not identically $+\infty$, and lower semi-continuous. The function $\ell$ is four times continuously differentiable on $\mathbb{R}^p$ and there exist finite constant $0 < \mu \leq L$ such that for all $\beta \in \mathbb{R}^p$,*

$$\mu I_p \preceq \nabla^{(2)} f(\beta) \preceq L I_p,$$

*where $I_p$ is the identity matrix of $\mathbb{R}^p$, $\nabla^{(2)} f$ denotes the Hessian matrix of $f$ evaluated at $\beta$, and $A \preceq B$ means that $B - A$ is symmetric positive semi-definite.*

**Theorem IV.7.** *Assume AIV.6 and $\sigma_k^2 L \leq 1$ for all $k \geq 1$. Suppose also that the sequence $\{\beta^{(k)},\ k \geq 0\}$ produced by Algorithm 6 remains in a compact set $\mathcal{C}$ that contains $\beta_\star \stackrel{\text{def}}{=} \mathsf{Argmin}_{u \in \mathbb{R}^p} F(\beta)$. Then there exists a finite constant $C_0$ such that*

$$\mathbb{E}\left[\|\beta^{(k)} - \beta_\star\|_2^2\right] \leq \left(1 - \frac{\mu}{4}\right)^k \mathbb{E}\left[\|\beta^{(0)} - \beta_\star\|_2^2\right] + C_0 \left(\frac{1}{N_k} + \sigma_k^4\right)$$

*Proof.* See Section 4.5.2. □

## 4.4 Numerical Experiments

### 4.4.1 Toy Example: Comparing Algorithm 4 and the Iterated Filtering of *Ionides et al.* (2011)

In this section we use a toy proble to illustrate the behavior of Algorithm 4 (that we refer below as PROX) can perform a comparison with the initial iterated filtering of *Ionides et al.* (2011) (that we refer below as IF1). We consider a simple bivariate discrete time Gaussian autoregressive process, with Gaussian measurement error. We chose this model so that the Monte Carlo calculations can be verified using a Kalman filter. The model is given by the state space forms:

$$\begin{aligned} X_n|X_{n-1} = x_{n-1} &\sim \mathcal{N}(\alpha x_{n-1}, \sigma^\top \sigma), \\ Y_n|X_n = x_n &\sim \mathcal{N}(x_n, I_2). \end{aligned}$$

where $\alpha$, $\sigma$ are $2 \times 2$ matrices and $I_2$ is $2 \times 2$ identity matrix. We simulate the data set with the following parameters:

$$\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} = \begin{bmatrix} 0.8 & -0.5 \\ 0.3 & 0.9 \end{bmatrix}, \ \sigma = \begin{bmatrix} 3 & 0 \\ -0.5 & 2 \end{bmatrix}.$$

We set the number of time points $N = 100$ and initial starting point $X_0 = (-3, 4)$. We estimate parameters $\alpha_2$ and $\alpha_3$ for this model using both PROX and IF1. We run our experiment with 25 iterations ($M = 25$) and with 1000 particles ($J = 1000$) on a Linux computer with 12 cores 3.07GHz processors. As seen from Fig. 4.1, while the maximum likelihood (ML) value obtained from both algorithms appear to be fairly close to the true ML value – vertical broken line – the distribution of the estimate produced by Algorithm 4 appear to smaller bias and a smaller variance,

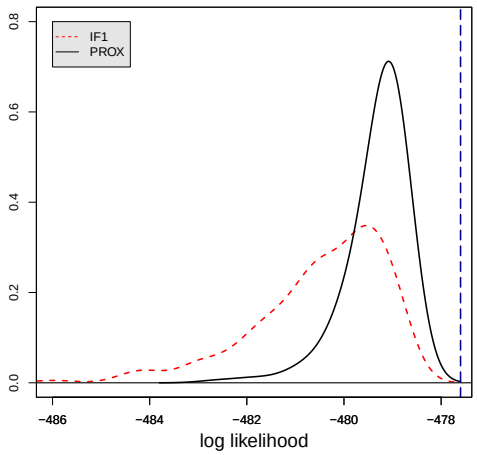


Figure 4.1: Comparison of estimators for the linear, Gaussian toy example, showing the densities of the MLEs estimated by the PROX and IF1 methods. The parameters $\alpha_2$ and $\alpha_3$ were estimated, started from 200 randomly uniform initial values over a large rectangular region $[-1, 1] \times [-1, 1]$.

implying better convergence rate in this case. In addition, Algorithm 4 seems to be more robust to the initialization of the algorithm, since we start at random values uniformly in a large rectangle. Furthermore as shown in Table 4.4.1 PROX has similar computational costs as IF1.

For this toy example, Fig. 4.2 shows the results of 40 Monte Carlo replications so that we can see the clustering of the MLE estimates around the true MLE. For PROX, most of the replications clustered near the true MLE while none of them stays in a lower likelihood region. Fig. 1, can be viewed as a statistical summary of Fig. 4.2, with 200 Monte Carlo replications. These results indicate that PROX is clearly the better of the investigated methods for this test compared to IF1.

We also checked how the methods compared when given additional computational resources, setting $M = 100$ iterations and $J = 10000$ particles, with the random walk standard deviation decreasing geometrically from 0.02 down to 0.0018 for both methods. In this situation, PROX is better than IF1. Both methods have

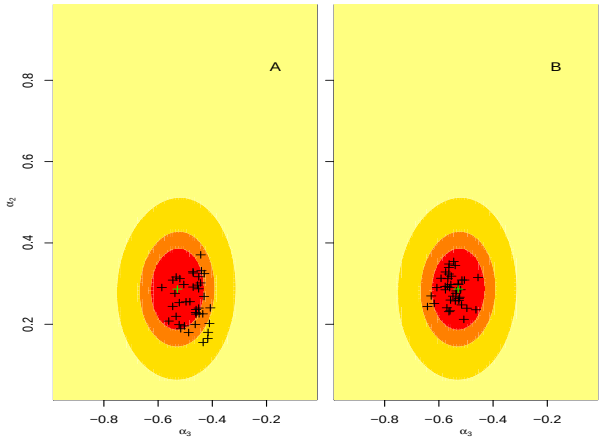


Figure 4.2: Comparison of different estimators. The likelihood surface for the linear, Gaussian model, with likelihood within 2 log units of the maximum shown in red, within 4 log units in orange, within 10 log units in yellow, and lower in light yellow. The location of the MLE is marked with a green cross. The black crosses show final points from 40 Monte Carlo replications of the estimators: (A) IF1 method; (B) PROX method; Each method, was started uniformly over the rectangle shown, with $M = 25$ iterations, $N = 1000$ particles, and a random walk standard deviation decreasing from 0.02 geometrically to 0.011 for both $\alpha_2$ and $\alpha_3$.

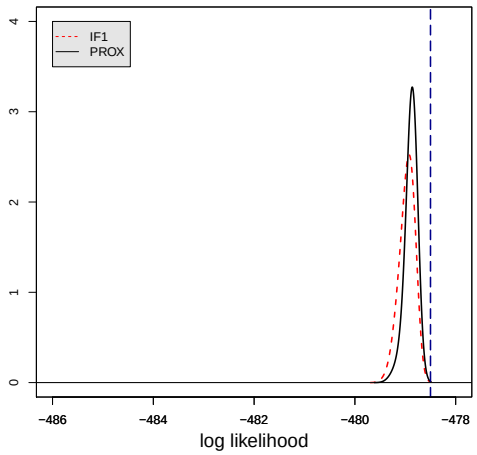


Figure 4.3: The distributions of likelihoods corresponding to Monte Carlo MLE approximations estimated by IF1 and PROX methods for toy model. The MLE is shown as a dashed vertical line (dark blue in electronic version). The optimizations were started from 200 randomly uniform initial values over a rectangle.

Table 4.1: Computation times, in seconds, for the toy example.

| | $J = 200$ | $J = 1000$ | $J = 10000$ |
|---|---|---|---|
| IF1 | 1.332 | 3.653 | 36.564 |
| PROX | 1.329 | 3.640 | 36.594 |

comparable computational demands for given $M$ and $J$.

In addition, average computational time of ten independent runs of each approach is given in Table4.4.1. Additional overheads for estimating score make the computation time of IF1 a bit larger compared to computational time of PROX. However, with complex models and large enough number of particles, these overheads become negligible and computational time of IF1 and PROX are similar.

### 4.4.2 High-Dimensional Mixed Effects Logistic Regression Models

Although iterated filtering algorithms were developed specifically in the context of state space models, we show here that they can also be employed to fit random effects models. We focus on the high-dimensional logistic regression case.

Let $X \in \mathbb{R}^{n\times p}$, $Z \in \mathbb{R}^{n\times q}$. The $i$-th row of $X$ is $x_i$, and the $i$-column of $Z$ is $z_i$. For a regularization parameter $\lambda > 0$, $p_\lambda : \mathbb{R}^p \to [0,\infty)$ is a convex penalty. The random effect logistic regression model leads to the problem of maximizing $F(\beta) = f(\beta) + p_\lambda(\beta)$, where $f$ is the negative penalized negative log-likelihood function given by

$$f(\beta) = -\log \int_{\mathbb{R}^q} \exp\left[\sum_{i=1}^{n} y_i \left(\langle x_i, \beta\rangle + \kappa \langle z_i, u\rangle\right) - \log\left(1 + e^{\langle x_i,\beta\rangle + \kappa\langle z_i,u\rangle}\right)\right] G(u)\mathrm{d}u,$$

where $G$ is the density of $\mathbf{N}(0, I_q)$ on $\mathbb{R}^q$, and $\kappa > 0$ is a noise parameter that we assume known. In the sequel we take $p_\lambda(\beta) = \lambda\|\beta\|_1$. This problem falls squarely in the framework developed above and Algorithm 5 applied to this problem becomes.

**Algorithm 7** Iterated Filtering Algorithm solving mixed effects logistic regression

Given $\beta^{(k)}$:

1. Generate $\vartheta_{1:N_k} \overset{i.i.d}{\sim} \mathbf{N}(\beta^{(k)}, \sigma_k I_p)$, and $U_{1:N_k} \overset{i.i.d.}{\sim} \mathbf{N}(0, I_q)$.
2. For each $1 \le i \le N_k$, compute

$$w_j = \exp\left[\sum_{i=1}^{n} y_i \left(\langle x_i, \vartheta_j\rangle + \kappa \langle z_i, U_j\rangle\right) - \log\left(1 + e^{\langle x_i,\vartheta_j\rangle + \kappa\langle z_i,U_j\rangle}\right)\right],$$

3. Compute

$$\beta^{(k+1)} = \mathrm{Prox}_{\sigma_k}^{p_\lambda}\left(-\frac{\sum_{j=1}^{N_k} w_j\vartheta_j}{\sum_{j=1}^{N_k} w_j}\right).$$

When the dimension $p$ is large, this joint update strategy is likely to perform

poorly. The block update version (Algorithm 6) is straightforward to design, provides a better alternative. The resulting algorithm is as follows.

**Algorithm 8** Block Iterated Filtering Algorithm solving mixed effects logistic regression

Given $\beta^{(k)}$:

1. Set $\bar{\beta} = \beta^{(k)}$, and $s = 1$.

   (a) Generate $U_{1:N_k} \overset{i.i.d.}{\sim} \mathbf{N}(0, I_q)$. For $j = 1, \ldots, N_k$, set $\vartheta_{j,\ell} = \bar{\beta}_\ell$, if $\ell \neq s$, and draw $\vartheta_{j,s} \sim \mathbf{N}(\bar{\beta}_s, \sigma_k)$. Compute

   $$w_j = \exp\left[\sum_{i=1}^{n} y_i \left(\langle x_i, \vartheta_j \rangle + \kappa \langle z_i, U_j \rangle\right) - \log\left(1 + e^{\langle x_i, \vartheta_j \rangle + \kappa \langle z_i, U_j \rangle}\right)\right], \quad 1 \leq j \leq N_k$$

   (b) Set

   $$\bar{\beta}_s = \mathrm{Prox}_{\sigma_k}^{p_\lambda}\left(-\frac{\sum_{j=1}^{N_k} w_j \vartheta_j}{\sum_{j=1}^{N_k} w_j}\right).$$

   (c) If $s < p$, set $s = s + 1$, and go back to (a).

2. Set $\beta^{(k+1)} = \bar{\beta}$.

#### 4.4.2.1 Numerical examples

We have carried out a comprehensive numerical study for the iterated filtering algorithm compared with the stochastic proximal gradient and second order approximate algorithms below. For the simulation data settings, we have kept the training sample size to be 200, true non-zero fixed effects size to be 5, and testing sample size to be 100 for model selection with respect to the regularization parameter $\lambda$'s. Data generation is done according to section 2.5 of chapter II.

Then we run one experiment on each of the following problem design:s fixed effect dimensions $p = 50, 100, 200, 250$, random effect ranks $q = 2, 5, 7, 10$, and random effect noise levels $\sigma = 0.1, 1.5, 2.5, 3.5$. In total, for each of the three algorithms:

iterated filtering (IF), stochastic proximal gradient (SPG), second order approximate (SOA), we have $4 \times 4 \times 4 = 64$ different settings. In each of these settings, we have used $N_k = 250$ Monte Carlo particles in algorithm (8) above. The results are tabulated below.

Table 4.2: Relative estimation error for Iterated Filtering (IF), Stochastic Proximal Gradient (SPG), and Second Order Approximate (SOA) algorithms

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p = 50$ | | | | | | | | | | | |
| | $\sigma = 0.1$ | | | $\sigma = 1.5$ | | | $\sigma = 2.5$ | | | $\sigma = 3.5$ | | |
| | IF | SPG | SOA | IF | SPG | SOA | IF | SPG | SOA | IF | SPG | SOA |
| $q = 2$ | 0.39 | 0.73 | 0.47 | 0.64 | 0.69 | 0.71 | 0.62 | 0.46 | 0.72 | 1.0 | 0.78 | 0.80 |
| 5 | 0.35 | 0.36 | 0.46 | 1.01 | 0.60 | 0.82 | 1.0 | 0.95 | 0.99 | 1.0 | 0.82 | 1.0 |
| 7 | 0.59 | 0.49 | 0.49 | 1.0 | 0.55 | 0.74 | 1.0 | 0.46 | 0.92 | 1.0 | 0.78 | 0.92 |
| 10 | 0.36 | 0.73 | 0.46 | 1.0 | 0.55 | 0.81 | 1.0 | 0.78 | 0.92 | 1.0 | 0.79 | 0.98 |
| | $p = 100$ | | | | | | | | | | | |
| $q = 2$ | 0.54 | 0.45 | 0.45 | 0.69 | 0.53 | 0.73 | 0.71 | 0.61 | 0.83 | 0.92 | 0.65 | 0.91 |
| 5 | 0.36 | 0.43 | 0.43 | 0.93 | 0.66 | 0.86 | 1.0 | 0.82 | 0.97 | 1.0 | 0.77 | 0.98 |
| 7 | 0.41 | 0.50 | 0.67 | 0.99 | 0.82 | 0.87 | 1.01 | 0.80 | 0.95 | 1.0 | 0.86 | 1.0 |
| 10 | 0.57 | 0.66 | 0.66 | 1.0 | 0.69 | 0.81 | 1.0 | 0.66 | 0.89 | 1.0 | 1.0 | 1.0 |
| | $p = 200$ | | | | | | | | | | | |
| $q = 2$ | 0.39 | 0.41 | 0.41 | 0.43 | 0.75 | 0.47 | 0.61 | 0.71 | 0.66 | 0.66 | 0.66 | 0.68 |
| 5 | 0.59 | 0.50 | 0.66 | 1.0 | 0.58 | 0.72 | 1.0 | 0.66 | 0.85 | 1.0 | 1.0 | 0.98 |
| 7 | 0.74 | 0.49 | 0.65 | 1.0 | 0.80 | 0.73 | 1.0 | 0.86 | 0.88 | 1.0 | 0.93 | 0.98 |
| 10 | 0.51 | 0.41 | 0.61 | 1.0 | 0.75 | 0.78 | 1.0 | 0.66 | 0.86 | 1.0 | 0.92 | 0.93 |
| | $p = 250$ | | | | | | | | | | | |
| $q = 2$ | 0.42 | 0.55 | 0.55 | 0.59 | 0.48 | 0.67 | 0.62 | 0.67 | 0.67 | 0.74 | 0.77 | 0.77 |
| 5 | 0.65 | 0.49 | 0.58 | 1.0 | 0.64 | 0.78 | 1.0 | 0.63 | 0.77 | 1.0 | 0.59 | 0.85 |
| 7 | 0.44 | 0.46 | 0.70 | 1.0 | 0.87 | 0.88 | 1.0 | 0.91 | 0.88 | 1.0 | 0.84 | 0.98 |
| 10 | 0.37 | 0.41 | 0.53 | 0.99 | 0.60 | 0.80 | 1.0 | 0.64 | 0.82 | 1.0 | 0.76 | 0.89 |

For estimation performance from the above results, problem dimensions $p$ within our experimental range seem not be a major factor affecting the performance. While the random effect factor dimension $q$ and noise level $\sigma$ play clearer role in solution performance. In general, the iterative filtering algorithm performs well, in some cases better than the stochastic proximal gradient second order approximate algorithms when $q = 2$, or when $\sigma = 0.1$, which is relatively small compared with other $\sigma$ values.

When $\sigma$ becomes larger than 1.5, it would only perform relatively well when $q = 2$, and its performance deteriorates when $q$ increases.

Table 4.3: Harmonic Mean of Sensitivity and Precision for IF, SPG and SOA algorithms

| | $p = 50$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 0.1$ | | | $\sigma = 1.5$ | | | $\sigma = 2.5$ | | | $\sigma = 3.5$ | | |
| | IF | SPG | SOA | IF | SPG | SOA | IF | SPG | SOA | IF | SPG | SOA |
| $q = 2$ | 0.22 | 0.77 | 0.33 | 0.55 | 0.91 | 0.91 | 0.33 | 0.63 | 0.83 | NaN | 0.83 | 0.73 |
| 5 | 0.22 | 0.23 | 0.37 | 0.24 | 0.43 | 0.77 | NaN | 0.57 | 0.29 | NaN | 0.67 | NaN |
| 7 | 0.52 | 0.42 | 0.37 | NaN | 0.19 | 1.0 | NaN | 0.28 | 0.75 | 0.09 | 0.57 | 0.40 |
| 10 | 0.20 | 0.19 | 0.43 | NaN | 0.23 | 0.67 | NaN | 0.63 | 0.57 | NaN | 0.28 | 0.33 |
| | $p = 100$ | | | | | | | | | | | |
| $q = 2$ | 0.59 | 0.42 | 0.42 | 0.19 | 0.36 | 0.91 | 0.13 | 0.10 | 0.91 | 0.16 | 0.10 | 0.67 |
| 5 | 0.14 | 0.38 | 0.38 | 0.11 | 0.12 | 0.83 | NaN | 0.10 | 0.5 | 0.04 | 0.13 | 0.44 |
| 7 | 0.14 | 0.29 | 0.67 | 0.11 | 0.60 | 0.60 | 0.10 | 0.14 | 0.62 | NaN | 0.19 | NaN |
| 10 | 0.45 | 0.91 | 0.91 | 0.09 | 0.56 | 0.75 | NaN | 0.18 | 0.75 | NaN | NaN | NaN |
| | $p = 200$ | | | | | | | | | | | |
| $q = 2$ | 0.09 | 0.26 | 0.26 | 0.06 | 0.06 | 0.24 | 0.12 | 0.73 | 0.72 | 0.10 | 0.50 | 0.57 |
| 5 | 0.34 | 0.24 | 0.56 | NaN | 0.24 | 0.77 | NaN | 0.13 | 0.50 | NaN | 0.29 | 0.20 |
| 7 | 0.77 | 0.21 | 0.53 | NaN | 0.57 | 0.42 | 0.03 | 0.60 | 0.33 | NaN | 0.36 | 0.22 |
| 10 | 0.34 | 0.11 | 0.71 | NaN | 0.62 | 0.57 | NaN | 0.09 | 0.43 | NaN | 0.57 | 0.57 |
| | $p = 250$ | | | | | | | | | | | |
| $q = 2$ | 0.07 | 0.22 | 0.22 | 0.18 | 0.25 | 0.67 | 0.13 | 0.56 | 0.56 | 0.08 | 0.91 | 0.91 |
| 5 | 0.21 | 0.08 | 0.19 | NaN | 0.21 | 0.59 | NaN | 0.24 | 0.63 | NaN | 0.12 | 0.67 |
| 7 | 0.08 | 0.10 | 0.48 | NaN | 0.62 | 0.67 | NaN | 0.75 | 0.59 | NaN | 0.36 | 0.22 |
| 10 | 0.07 | 0.09 | 0.18 | 0.04 | 0.07 | 0.83 | NaN | 0.04 | 0.62 | NaN | 0.04 | 0.62 |

The sparsity recovery performance of iterated filtering algorithm is comparable to the stochastic proximal gradient and second order approximate algorithm when sample size $N$, problem dimension $p$ are relatively small and the random effects are relatively weak in the models. However, iterated filtering algorithm would perform poorly in problems with larger sizes, especially with large $q$ or $\sigma$'s. In a number of large $q$ or $\sigma$ settings, the iterated filtering algorithm could be unstable, such that it estimates the fixed effects coefficients to be all zero, which leads to poor precision. Thus in close to real scale problems, we recommend at least using the stochastic

proximal gradient or second order approximate algorithms to check the results of the iterated filtering algorithms.

## 4.5 Proofs

### 4.5.1 Proof of Proposition IV.2

*Proof.* Let $q$ denote the density of $\mathbf{N}(0, \sigma^2 I_p)$. Write

$$\Pi_\sigma(\beta) = \frac{\int q(z) e^{\ell(\beta+\sigma z)}(\beta+\sigma z)\mathrm{d}z}{\int q(z) e^{\ell(\beta+\sigma z)}\mathrm{d}z} = \frac{\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))}(\beta+\sigma z)\mathrm{d}z}{\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))}\mathrm{d}z}.$$

Hence

$$\Pi_\sigma(\beta) - \mathrm{Prox}_\sigma(\beta) = \frac{\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))}\left(\beta+\sigma z - \mathrm{Prox}_\sigma(\beta)\right)\mathrm{d}z}{\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))}\mathrm{d}z}.$$

We note that for all $x \in \mathbb{R}$, $e^x = 1 + x + x^2 \int_0^1 (1-t) e^{tx} \mathrm{d}t$. Hence

$$\begin{aligned}\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))}\left(\beta+\sigma z - \mathrm{Prox}_\sigma(\beta)\right)\mathrm{d}z &= \beta - \mathrm{Prox}_\sigma(\beta) \\ &+ \int \left(\beta+\sigma z - \mathrm{Prox}_\sigma(\beta)\right)\left(\ell(\beta+\sigma z) - \ell(\mathrm{Prox}_\sigma(\beta))\right) q(z)\mathrm{d}z \\ + \int_0^1 (1-t)\left[\int \left(\beta+\sigma z - \mathrm{Prox}_\sigma(\beta)\right)\left(\ell(\beta+\sigma z) - \ell(\mathrm{Prox}_\sigma(\beta))\right)^2 e^{t(\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta)))} q(z)\mathrm{d}z\right]\mathrm{d}t.\end{aligned}$$

Using the assumption that $\nabla\ell$ is $M$ Lipschitz, and setting $p = \mathrm{Prox}_\sigma(\beta)$, we get

$$\int \left(\beta + \sigma z - \mathrm{Prox}_\sigma(\beta)\right)\left(\ell(\beta+\sigma z) - \ell(\mathrm{Prox}_\sigma(\beta))\right) e^{t(\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta)))} q(z)\mathrm{d}z$$

Hence, Jensen's inequality gives

$$\|\Pi_\sigma(\beta) - \mathrm{Prox}_\sigma(\beta)\| \leq \sigma \left[ \frac{\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))} \|z - \sigma^{-1}(\mathrm{Prox}_\sigma(\beta) - \beta))\|^2 \mathrm{d}z}{\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))} \mathrm{d}z} \right]^{1/2}.$$

By the optimality condition in the maximization that defines $\mathrm{Prox}_\sigma(\beta)$, we have:

$$\nabla \ell \left(\mathrm{Prox}_\sigma(\beta)\right) = \frac{1}{\sigma^2} \left(\mathrm{Prox}_\sigma(\beta) - \beta\right). \tag{4.7}$$

Using this AIV.6 and a straightforward Taylor expansion we obtain

$$q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))} \geq \left(\frac{1}{1+M\sigma^2}\right)^{p/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathrm{Prox}_\sigma(\beta) - \beta\|^2\right) \left(\frac{1+M\sigma^2}{2\pi}\right)^{p/2} \times \exp\left(-\frac{1+M\sigma^2}{2}\|z - \sigma^{-1}(\mathrm{Prox}_\sigma(\beta) - \beta)\|^2\right).$$

Hence

$$\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))} \mathrm{d}z \geq \left(\frac{1}{1+M\sigma^2}\right)^{p/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathrm{Prox}_\sigma(\beta) - \beta\|^2\right).$$

Similar calculations for the numerator gives

$$q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))} \leq \left(\frac{1}{1+m\sigma^2}\right)^{p/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathrm{Prox}_\sigma(\beta) - \beta\|^2\right) \left(\frac{1+m\sigma^2}{2\pi}\right)^{p/2} \times \exp\left(-\frac{1+m\sigma^2}{2}\|z - \sigma^{-1}(\mathrm{Prox}_\sigma(\beta) - \beta)\|^2\right),$$

So that

$$\int q(z) e^{\ell(\beta+\sigma z)-\ell(\mathrm{Prox}_\sigma(\beta))} \|z - \sigma^{-1}(\mathrm{Prox}_\sigma(\beta) - \beta))\|^2 \mathrm{d}z \leq \left(\frac{1}{1+m\sigma^2}\right)^{p/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathrm{Prox}_\sigma(\beta) - \beta\|^2\right) \frac{p}{1+m\sigma^2}.$$

We conclude that

$$\|\Pi_\sigma(\beta) - \mathrm{Prox}_\sigma(\beta)\| \leq \sigma \left(\frac{1+M\sigma^2}{1+m\sigma^2}\right)^{p/4} \left(\frac{p}{1+m\sigma^2}\right)^{1/2},$$

as claimed. □

### 4.5.2 Proof of Theorem IV.7

We then denote by $\Theta_i \in \mathbb{R}^{p_i}$ the domain of $g_i$, That is $\Theta_i = \{u \in \mathbb{R}^{p_i} : \ g_i(u) < \infty\}$. We introduce the function

$$F_1(u|\beta_2) \stackrel{\text{def}}{=} f(u, \beta_2) + g_1(u), \quad F_2(v|\beta_1) = f(\beta_1, v) + g_2(v),$$

where $u, \beta_1 \in \mathbb{R}^{p_1}$, and $v, \beta_2 \in \mathbb{R}^{p_2}$. We then write $\nabla_1 f(u, v)$ (resp. $\nabla_2 f(u, v)$) to denote the partial derivative of $f$ with respect to $u$ (resp. $v$) and evaluated at $(u, v)$. We will need the following well-known result.

**Lemma IV.8.** *Assume that $g : \mathbb{R}^p \to (-\infty, +\infty]$ is a convex lower semi-continuous function with domain $\Theta$. For $\beta, \beta' \in \Theta$ and $\gamma > 0$*

$$g\Big(\mathrm{Prox}_\gamma^g(\beta)\Big) - g(\beta') \leq -\frac{1}{\gamma}\left\langle \mathrm{Prox}_\gamma^g(\beta) - \beta', \mathrm{Prox}_\gamma^g(\beta) - \beta \right\rangle. \tag{4.8}$$

*For any $\gamma > 0$ and for any $\beta, \beta' \in \Theta$,*

$$\|\mathrm{Prox}_\gamma^g(\beta) - \mathrm{Prox}_\gamma^g(\beta')\|^2 + \|(\mathrm{Prox}_\gamma^g(\beta) - \beta) - (\mathrm{Prox}_\gamma^g(\beta') - \beta')\|^2 \leq \|\beta - \beta'\|^2. \tag{4.9}$$

*Proof.* See (*Bauschke and Combettes*, 2011, Propositions 4.2., 12.26 and 12.27). □

We will also need the following result taken from *Atchadé et al.* (2017).

**Lemma IV.9.** *Assume IV.6 and take $\sigma > 0$ such that $\sigma^2 L \leq 1$.*

1. *For all $u, u', x \in \Theta_1$, and $\beta_2 \in \mathbb{R}^{p_2}$, we have*

$$2\sigma^2 \left(F_1\left(\mathrm{Prox}_\sigma^{g_1}(u)|\beta_2\right) - F_1(x|\beta_2)\right) + \|\mathrm{Prox}_\sigma^{g_1}(u) - x\|_2^2 - \left(1 - \frac{\mu}{2}\right)\|u' - x\|_2^2 \\ \leq 2\left\langle u - \left(u' - \sigma^2 \nabla_1 f(u', \beta_2)\right), \mathrm{Prox}_\sigma^{g_1}(u) - x\right\rangle. \quad (4.10)$$

2. *for all $v, v', y \in \Theta_2$, and $\beta_1 \in \mathbb{R}^{p_1}$,*

$$2\sigma^2 \left(F_2\left(\mathrm{Prox}_\sigma^{g_2}(v)|\beta_1\right) - F_2(y|\beta_1)\right) + \|\mathrm{Prox}_\sigma^{g_2}(v) - y\|_2^2 - \left(1 - \frac{\mu}{2}\right)\|v' - y\|_2^2 \\ \leq 2\left\langle v - \left(v' - \sigma^2 \nabla_2 f(v', \beta_1)\right), \mathrm{Prox}_\sigma^{g_2}(v) - y\right\rangle. \quad (4.11)$$

*Proof.* We prove (1), (2) is similar. The $L$-Lipschitz property of $f_1$ which follows from AIV.6 give:

$$f\left(\mathrm{Prox}_\sigma^{g_1}(u), \beta_2\right) \leq f(u', \beta_2) + \langle \nabla_1 f(u', \beta_2), \mathrm{Prox}_\sigma^{g_1}(u) - u'\rangle + \frac{L}{2}\|\mathrm{Prox}_\sigma^{g_1}(u) - u'\|_2^2.$$

Hence

$$f\left(\mathrm{Prox}_\sigma^{g_1}(u), \beta_2\right) - f(x; \beta_2) \leq \left[f(u', \beta_2) + \langle \nabla_1 f(u', \beta_2), x - u'\rangle - f(x, \beta_2)\right] \\ + \langle \nabla_1 f(u', \beta_2), \mathrm{Prox}_\sigma^{g_1}(u) - x\rangle + \frac{L}{2}\|\mathrm{Prox}_\sigma^{g_1}(u) - u'\|_2^2.$$

Then we use the strong convexity of $f$ to conclude that

$$f\left(\mathrm{Prox}_\sigma^{g_1}(u), \beta_2\right) - f(x; \beta_2) \leq -\frac{\mu}{2}\|x - u'\|_2^2 + \langle \nabla_1 f(u', \beta_2), \mathrm{Prox}_\sigma^{g_1}(u) - x\rangle \\ + \frac{L}{2}\|\mathrm{Prox}_\sigma^{g_1}(u) - u'\|_2^2. \quad (4.12)$$

On the other hand Lemma IV.8 gives

$$g_1(x) \geq g_1(\mathrm{Prox}_\sigma^{g_1}(u)) + \frac{1}{\sigma^2}\left\langle u - \mathrm{Prox}_\sigma^{g_1}(u), x - \mathrm{Prox}_\sigma^{g_1}(u)\right\rangle.$$

We combine this with (4.12) to get

$$\begin{aligned}
F_1\left(\mathrm{Prox}_\sigma^{g_1}(u)|\beta_2\right) - F_1(x|\beta_2) &\leq -\frac{\mu}{2}\|x-u'\|_2^2 \\
&+ \frac{1}{\sigma^2}\left\langle u + \sigma^2\nabla_1 f(u',\beta_2) - \mathrm{Prox}_\sigma^{g_1}(u), \mathrm{Prox}_\sigma^{g_1}(u) - x\right\rangle + \frac{L}{2}\|\mathrm{Prox}_\sigma^{g_1}(u) - u'\|_2^2 \\
&\leq -\frac{\mu}{2}\|x-u'\|_2^2 + \frac{1}{\sigma^2}\left\langle u - \left(u' - \sigma^2\nabla_1 f(u',\beta_2)\right), \mathrm{Prox}_\sigma^{g_1}(u) - x\right\rangle \\
&\quad + \frac{1}{\sigma^2}\left\langle u' - \mathrm{Prox}_\sigma^{g_1}(u), \mathrm{Prox}_\sigma^{g_1}(u) - x\right\rangle + \frac{1}{2\sigma^2}\|\mathrm{Prox}_\sigma^{g_1}(u) - u'\|_2^2,
\end{aligned}$$

where the last inequality also uses the assumption that $\sigma^2 L \leq 1$. The result follows noticing that for all $\beta, \beta_0, \bar{\beta} \in \mathbb{R}^q$ for some $q \geq 1$, we have

$$\begin{aligned}
\frac{1}{2}\|\bar{\beta}-\beta\|^2 + \left\langle \bar{\beta}-\beta, \beta_0 - \bar{\beta}\right\rangle &= \frac{1}{2}\left\langle \bar{\beta}-\beta, \bar{\beta}-\beta\right\rangle + \left\langle \bar{\beta}-\beta, \beta_0-\bar{\beta}\right\rangle \\
&= \frac{1}{2}\left\langle \bar{\beta}-\beta, \bar{\beta}-\beta+2\beta_0-2\bar{\beta}\right\rangle = \frac{1}{2}\left\langle \bar{\beta}-\beta, 2\beta_0-\beta-\bar{\beta}\right\rangle \\
&= \frac{1}{2}\left[\left\langle \bar{\beta}-\beta_0, \beta_0-\beta+\beta_0-\bar{\beta}\right\rangle + \left\langle \beta_0-\beta, \beta_0-\beta+\beta_0-\bar{\beta}\right\rangle\right] \\
&= \frac{1}{2}\left[\|\beta-\beta_0\|^2 - \|\bar{\beta}-\beta_0\|^2\right].
\end{aligned}$$

□

We apply Lemma IV.9-(4.10) with $u = H_1^{(k+1)}$, $u' = \beta_1^{(k)}$, $x = \beta_{\star,1}$, $\beta_2 = \beta_{\star,2}$ to get

$$\begin{aligned}
2\sigma_k^2\left[F_1(\beta_1^{(k+1)}|\beta_{\star,2}) - F_1(\beta_{\star,1}|\beta_{\star,2})\right] + \|\beta_1^{(k+1)} - \beta_{\star,1}\|_2^2 - \left(1-\frac{\mu}{2}\right)\|\beta_1^{(k)} - \beta_{\star,1}\|_2^2 \\
\leq 2\left\langle H_1^{(k+1)} - \left(\beta_1^{(k)} - \sigma_k^2\nabla_1 f(\beta_1^{(k)}, \beta_{\star,2})\right), \beta_1^{(k+1)} - \beta_{\star,1}\right\rangle. \quad (4.13)
\end{aligned}$$

Then we apply Lemma IV.9-(4.11) with $v = H_2^{(k+1)}$, $\beta_1 = \beta_1^{(k+1)}$, $v' = \beta_2^{(k)}$, and $y = \beta_{\star,2}$, and we get

$$2\sigma_k^2 \left[F_2(\beta_2^{(k+1)}|\beta_1^{(k+1)}) - F_2(\beta_{\star,2}|\beta_1^{(k+1)})\right] + \|\beta_2^{(k+1)} - \beta_{\star,2}\|_2^2 - \left(1 - \frac{\mu}{2}\right)\|\beta_2^{(k)} - \beta_{\star,2}\|_2^2 \\ \leq 2\left\langle H_2^{(k+1)} - \left(\beta_2^{(k)} - \sigma_k^2 \nabla_2 f(\beta_1^{(k+1)}, \beta_2^{(k)})\right), \beta_2^{(k+1)} - \beta_{\star,2}\right\rangle. \quad (4.14)$$

We then add (4.13) and (4.14) to get

$$2\sigma_k^2 \left[F(\beta^{(k+1)}) - F(\beta_\star)\right] + \|\beta^{(k+1)} - \beta_\star\|_2^2 - \left(1 - \frac{\mu}{2}\right)\|\beta^{(k)} - \beta_\star\|_2^2 \\ \leq 2\left\langle \eta_1^{(k+1)}, \beta_1^{(k+1)} - \beta_{\star,1}\right\rangle + 2\left\langle \eta_2^{(k+1)}, \beta_2^{(k+1)} - \beta_{\star,2}\right\rangle \\ + 2\sigma_k^2 \left\langle \nabla_1 f(\beta_1^{(k)}, \beta_{\star,2}) - \nabla_1 f(\beta_1^{(k)}, \beta_2^{(k)}), \beta_1^{(k+1)} - \beta_{\star,1}\right\rangle, \quad (4.15)$$

where

$$\eta_1^{(k+1)} \stackrel{\text{def}}{=} H_1^{(k+1)} - \left(\beta_1^{(k)} - \sigma_k^2 \nabla_1 f(\beta_1^{(k)}, \beta_2^{(k)})\right), \\ \text{and} \quad \eta_2^{(k+1)} \stackrel{\text{def}}{=} H_2^{(k+1)} - \left(\beta_2^{(k)} - \sigma_k^2 \nabla_2 f(\beta_1^{(k+1)}, \beta_2^{(k)})\right).$$

Since the gradient $\nabla f$ is Lipschitz as assumed in HIV.6, we have

$$+2\left|\left\langle \nabla_1 f(\beta_1^{(k)}, \beta_{\star,2}) - \nabla_1 f(\beta_1^{(k)}, \beta_2^{(k)}), \beta_1^{(k+1)} - \beta_{\star,1}\right\rangle\right| \leq 2L\|\beta_2^{(k)} - \beta_{\star,2}\|_2 \|\beta_1^{(k+1)} - \beta_{\star,1}\|_2 \\ \leq 2L\|\beta^{(k)} - \beta_\star\|_2 \|\beta^{(k+1)} - \beta_\star\|_2 \leq L\|\beta^{(k)} - \beta_\star\|_2^2 + L\|\beta^{(k+1)} - \beta_\star\|_2^2,$$

where the last inequality uses the fact that $2ab \leq a^2 + b^2$. Using this together with (4.15) and the choice $\sigma_k^2 \leq \mu/(4L)$, we conclude that

$$2\sigma_k^2 \left[F(\beta^{(k+1)}) - F(\beta_\star)\right] + \|\beta^{(k+1)} - \beta_\star\|_2^2 \leq \left(1 - \frac{\mu}{4}\right)\|\beta^{(k)} - \beta_\star\|_2^2$$

$$+ 2 \left\langle \eta^{(k+1)}, \beta^{(k+1)} - \beta_\star \right\rangle . \quad (4.16)$$

Iterating this inequality we obtain,

$$\|\beta^{(k)} - \beta_\star\|_2^2 \leq \left(1 - \frac{\mu}{4}\right)^k \|\beta^{(0)} - \beta_\star\|_2^2 + 2 \sum_{j=1}^{k} \left(1 - \frac{\mu}{4}\right)^{k-j} \left\langle \eta^{(j)}, \beta^{(j)} - \beta_\star \right\rangle . \quad (4.17)$$

Define $\bar{\beta}^{(k+1)} = (\bar{\beta}_1^{(k+1)}, \bar{\beta}_2^{(k+1)})$, where

$$\bar{\beta}_1^{(k+1)} \stackrel{\text{def}}{=} \text{Prox}_{\sigma_k}^{g_1} \left( \beta_1^{(k)} - \sigma_k^2 \nabla_1 f(\beta_1^{(k)}, \beta_2^{(k)}) \right),$$
$$\text{and} \quad \bar{\beta}_2^{(k+1)} \stackrel{\text{def}}{=} \text{Prox}_{\sigma_k}^{g_2} \left( \beta_2^{(k)} - \sigma_k^2 \nabla_2 f(\beta_1^{(k+2)}, \beta_2^{(k)}) \right).$$

We then write $\beta^{(k)} - \beta_\star = \beta^{(k)} - \bar{\beta}^{(k)} + \bar{\beta}^{(k)} - \beta_\star$. Then by the Lipschitz property of the proximal map (4.9),

$$\left\langle \eta^{(j)}, \beta^{(j)} - \beta_\star \right\rangle = \left\langle \eta^{(j)}, \beta^{(j)} - \bar{\beta}^{(j)} \right\rangle + \left\langle \eta^{(j)}, \bar{\beta}^{(j)} - \beta_\star \right\rangle$$
$$\leq \|\eta^{(j)}\|_2^2 + \left\langle \eta^{(j)}, \bar{\beta}^{(j)} - \beta_\star \right\rangle$$

Hence, taking the expectation on both side of (4.17 yields,

$$\mathbb{E}\left[\|\beta^{(k)} - \beta_\star\|_2^2\right] \leq \left(1 - \frac{\mu}{4}\right)^k \mathbb{E}\left[\|\beta^{(0)} - \beta_\star\|_2^2\right]$$
$$+ 2 \sum_{j=1}^{k} \left(1 - \frac{\mu}{4}\right)^{k-j} \mathbb{E}\left[\|\eta^{(j)}\|_2^2 + \left\langle \eta^{(j)}, \bar{\beta}^{(j)} - \beta_\star \right\rangle\right].$$

We apply Lemma IV.3 to conclude that

$$\mathbb{E}\left[\|\beta^{(k)} - \beta_\star\|_2^2\right] \leq \left(1 - \frac{\mu}{4}\right)^k \mathbb{E}\left[\|\beta^{(0)} - \beta_\star\|_2^2\right]$$

$$+ 2\sum_{j=1}^{k}\left(1-\frac{\mu}{4}\right)^{k-j}\mathbb{E}\left[\frac{1}{N_j}+\sigma_k^2\right] \leq \left(1-\frac{\mu}{4}\right)^{k}\mathbb{E}\left[\|\beta^{(0)}-\beta_\star\|_2^2\right] + C\left(\frac{1}{N_k}+\sigma_k^4\right).$$

This completes the proof.

# Bibliography


Anderson, J. L., and N. Collins (2007), Scalable implementations of ensemble filter algorithms for data assimilation, *Journal of Atmospheric and Oceanic Technology*, *24*(8), 1452–1463.

Atchadé, Y. F., G. Fort, and E. Moulines (2017), On perturbed proximal gradient algorithms, *J. Mach. Learn. Res.*, *18*(1), 310–342.

Attouch, H., and J. Bolte (2009), On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Math. Program.*, *116*, 5–16.

Attouch, H., J. Bolte, P. Redont, and A. Soubeyran (2010), Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-łojasiewicz inequality, *Math. Oper. Res.*, *35*(2), 438–457.

Attouch, H., J. Bolte, and B. F. Svaiter (2013), Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, *Math. Program.*, *137*(1-2, Ser. A), 91–129.

Atwell, S. e. a. (2010), Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines, *Nature*, *465*, 627–631.

Aulchenko, Y. S. e. a. (2007), Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci, *Genetics*, *177*, 577–585.

Bauschke, H., and P. Combettes (2011), *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, xvi+468 pp., Springer, New York, with a foreword by Hédy Attouch.

Beck, A., and M. Teboulle (2010), Gradient-based algorithms with applications to signal-recovery problems, in *Convex optimization in signal processing and communications*, pp. 42–88, Cambridge Univ. Press, Cambridge.

Beck, A., and L. Tetruashvili (2013), On the convergence of block coordinate descent type methods, *SIAM Journal on Optimization*, *23*(4), 2037–2060.

Beskos, A., O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead (2006), Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(3), 333–382.

Biane, P., J. Pitman, and M. Yor (2001), Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions, *Bull. Amer. Math. Soc. (N.S.)*, *38*(4), 435–465 (electronic).

Bickel, P. J., Y. Ritove, and A. B. Tsybakov (2009), Simultaneous analysis of lasso and dantzig selector, *The Annals of Statistics*, *37*, 1705–1732.

Bolte, J., A. Daniilidis, and A. Lewis (2006), The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM J. Optim.*, *17*(4), 1205–1223 (electronic).

Bolte, J., S. Sabach, and M. Teboulle (2014), Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Math. Program.*, *146*(1-2, Ser. A), 459–494.

Breslow, N. E., and D. G. Clayton (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, *88*, 9–25, doi:doi:10.2307/2290687.

Bühlmann, P., and S. van de Geer (2011), *Statistics for high-dimensional data*, Springer Series in Statistics, xviii+556 pp., Springer, Heidelberg, methods, theory and applications.

Bühlmann, P., M. Kalisch, and L. Meier (2014), High-dimensional statistics with a view toward applications in biology, *Annual Review of Statistics and Its Application*, *1*, 255–278.

Bunea, F., A. Tsybakov, and M. Wegkamp (2007), Sparsity oracle inequalities for the lasso, *Electronic Journal of Statistics*, *1*, 169–194.

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

Cappé, O., E. Moulines, and T. Ryden (2005), *Inference in Hidden Markov Models (Springer Series in Statistics)*, Springer-Verlag, Berlin, Heidelberg.

Combettes, P., and J. Pesquet (2015a), Stochastic Quasi-Fejer block-coordinate fixed point iterations with random sweeping, *SIAM J. Optim.*, *25*(2), 1221–1248.

Combettes, P., and J. Pesquet (2015b), Stochastic Approximations and Perturbations in Forward-Backward Splitting for Monotone Operators, *Tech. rep.*, arXiv:1507.07095v1.

Combettes, P., and V. Wajs (2005), Signal recovery by proximal forward-backward splitting, *Multiscale Modeling and Simulation*, *4*(4), 1168–1200.

Doucet, A., P. E. Jacob, and S. Rubenthaler (2013), Derivative-Free Estimation of the Score Vector and Observed Information Matrix with Application to State-Space Models, *ArXiv e-prints*.

Duchi, J. C., P. L. Bartlett, and M. J. Wainwright (2012), Randomized smoothing for stochastic optimization, *SIAM J. Optim.*, *22*(2), 674–701.

Eckart, G., and G. Young (1936), The approximation of one matrix by another of lower rank, *Psychometrika*, *1*, 211–218.

Eraker, B. (2001), Mcmc analysis of diffusion models with application to finance, *Journal of Business & Economic Statistics*, *19*(2), 177–191.

Ergun, A., R. Barbieri, U. T. Eden, M. A. Wilson, and E. N. Brown (2007), Construction of point process adaptive filter algorithms for neural systems using sequential monte carlo methods, *54*, 419 – 428.

Fan, J., and R. Li (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, *96*(456), 1348–1360, doi:10.1198/016214501753382273.

Fearnhead, P. (2008), Computational methods for complex stochastic systems: a review of some alternatives to mcmc, *Statistics and Computing*, *18*(2), 151–171.

Fernandez-Villaverde, J., and J. F. Rubio-Ramirez (2007), Estimating macroeconomic models: A likelihood approach, *The Review of Economic Studies*, *74*(4), 1059–1087.

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

Friedman, J., T. Hastie, R. Tibshirani, N. Simon, B. Narasimhan, and J. Y. Qian (2010b), glmnet: Lasso and elastic-net regularized generalized linear models.

Geer, S. A. v. d., and P. Bühlmann (2009), On the conditions used to prove oracle results for the lasso, *Electronic Journal of Statistics*, *3*, 1360–1392.

Groll, A., and G. Tutz (2014), Variable selection for generalized linear mixed models by l1-penalized estimation, *Journal Statistics and Computing*, *24*, 137–154, doi: 10.1007/s11222-012-9359-z.

Hu, C., W. Pan, and J. T. Kwok (2009), Accelerated gradient methods for stochastic optimization and online learning, in *Advances in Neural Information Processing Systems*, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, , and A. Culotta, pp. 781–789.

Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011), Fixed and random effects selection in mixed effects models, *Biometrics*, *67*, 495–503, doi:10.1111/j.1541-0420.2010.01463.x.

Ionides, E. L., C. Bretó, and A. A. King (2006), Inference for nonlinear dynamical systems, *Proceedings of the National Academy of Sciences*, *103*(49), 18,438–18,443.

Ionides, E. L., A. Bhadra, Y. Atchadé, and A. King (2011), Iterated filtering, *Ann. Statist.*, *39*(3), 1776–1802.

Ionides, E. L., D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King (2015), Inference for dynamic and latent variable models via iterated, perturbed bayes maps, *Proceedings of the National Academy of Sciences*, *112*(3), 719–724.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics, xiv+257 pp., Springer-Verlag, New York.

Juditsky, A., and A. Nemirovski (2012a), First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods, in *Oxford Handbook of Innovation*, edited by S. Sra, S. Nowozin, and S. Wright, pp. 121–146, MIT Press, Boston.

Juditsky, A., and A. Nemirovski (2012b), First-order methods for nonsmooth convex large-scale optimization, ii: Utilizing problem's structure, in *Oxford Handbook of Innovation*, edited by S. Sra, S. Nowozin, and S. Wright, pp. 149–181, MIT Press, Boston.

Koltchinskii, V. (2009), The dantzig selector and sparsity oracle inequalities, *Bernoulli*, *15*, 799–828, doi:10.3150/09/09-BEJ187.

Kranz, S. G., and H. R. Parks (2002), *A Primer of Real Analytic Functions*, 2 ed., XIII, 209 pp., Birkhäuser Basel, doi:10.1007/978-0-8176-8134-0.

Kurdyka, K. (1998), On gradients of functions definable in o-minimal structures, *Annales de l'institut Fourier*, *48*, 769–783.

Lan, G. (2012), An optimal method for stochastic composite optimization, *Math. Program.*, *133*(1-2, Ser. A), 365–397.

Loh, P., and M. J. Wainwright (2017), Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, *45*, 2455–2482.

McCullagh, P., and A. Nelder (1989), *Generalized Linear Models, Second Edition*, CRC, London.

McCulloch, C. E., and S. R. Searle (2005), *Generalized, Linear, and Mixed Models (GLMMs)*, Wiley, USA.

Meier, L., S. van de Geer, and P. Bühlmann (2008), The group lasso for logistic regression, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *36*, 53–71.

Meinshausen, N., and B. Yu (2009), Lasso-type recovery of sparse representations for high-dimensional data, *Annals of Statistics*, *37*, 246–270.

Molenberghs, G., and G. Verbeke (2005), *Models for Discrete Longitudinal Data*, Springer, New York.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and Y. Bin (2012), Nearly unbiased variable selection under minimax concave penalty, *Statistical Science*, *27*(4), 538–557.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2008), Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.*, *19*(4), 1574–1609, doi:10.1137/070704277.

Nesterov, Y. (2004), *Introductory Lectures on Convex Optimization, A basic course*, Kluwer Academic Publishers.

Newman, K., C. Fernandez, L. Thomas, and S. T Buckland (2008), Monte carlo inference for state-space models of wild animal populations, *65*, 572–83.

Nitanda, A. (2014), Stochastic proximal gradient descent with acceleration techniques, in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, pp. 1574–1582, Curran Associates, Inc.

Ola, E., C. Siddhartha, and S. Neil (2001), Likelihood inference for discretely observed nonlinear diffusions, *Econometrica*, *69*(4), 959–993.

Parikh, N., and S. Boyd (2013a), Proximal algorithms, *Foundations and Trends in Optimization*, *1*(3), 123–231.

Parikh, N., and S. Boyd (2013b), Proximal algorithms, *Foundations and Trends in Optimization*, *1*(3), 123–231.

Polson, N. G., J. G. Scott, and J. Windle (2012), Bayesian inference for logistic models using Polya-Gamma latent variables, *ArXiv e-prints*.

Raskutti, G., M. J. Wainwright, and B. Yu (2010), Restricted eigenvalue properties for correlated gaussian designs, *Journal of Machine Learning Research*, *11*, 2241–2259.

Richtárik, P., and M. Takáč (2014), Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function, *Mathematical Programming*, *144*(1), 1–38.

Roberts, G. O., and O. Stramer (2001), On inference for partially observed nonlinear diffusion models using the metropolis-hastings algorithm, *Biometrika*, *88*(3), 603–621.

Rosasco, L., S. Villa, and B. Vu (2014), Convergence of a Stochastic Proximal Gradient Algorithm, *Tech. rep.*, arXiv:1403.5075v3.

Rosenberg, B. (1973), Linear regression with randomly dispersed parameters, *Biometrika*, *60*, 65–72.

Rudelson, M., and S. Zhou (2011), Reconstruction from anisotropic random measurements, *Tech. rep.*, https://arxiv.org/abs/1106.1151.

Saha, A., and A. Tewari (2013), On the nonasymptotic convergence of cyclic coordinate descent methods, *SIAM Journal on Optimization*, *23*(1), 576–601.

Schelldorfer, J., P. Bühlmann, and S. van de Geer (2011), Estimation for high-dimensional linear mixed-effects models using l˙1 penalization, *Scandinavian Journal of Statistics*, *38*, 197–214, doi:10.1111/j.1467-9469.2011.00740.x.

Schelldorfer, J., L. Meier, and P. Bühlmann (2014), Glmmlasso: An algorithm for high-dimensional generalized linear mixed models using 1-penalization, *23*(2), 460–477, doi:10.1080/10618600.2013.773239.

van de Geer, S. (2008), High-dimensional generalized linear models and the lasso, *Annals of Statistics*, *36*(2), 614–645, doi:10.1214/009053607000000929.

van Vliet, M. H., F. Reyal, H. M. Horlings, M. van de Vijver, M. J. Reinders, and L. F. Wessels (2008), Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability, *BMC Genomics*, *9*, 375, doi:10.1186/1471-2164/9/375.

van't Veer, L., et al. (2002), Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, *415*, 530–535.

Vijver, M. v. d., et al. (2002), A gene-expression signature as a predictor of survival in breast cancer, *The New England Journal of Medicine*, *347*(25), 1999–2009, doi:10.1186/1471-2164/9/375.

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

Xiao, L. (2010), Dual averaging methods for regularized stochastic learning and online optimization, *J. Mach. Learn. Res.*, *11*, 2543–2596.

Xiao, L., and T. Zhang (2014), A proximal stochastic gradient method with progressive variance reduction, *Tech. rep.*, arXiv:1403.4699.

Yang, J. e. a. (2011), Genomic inflation factors under polygenic inheritance, *European Journal of Human Genetics*, *19*, 807–812.

Yu, J. M. e. a. (2006), A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nature Genetics*, *38*, 203–208.

Zhang, C. H. (2010a), Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, *38*, 894–942.

Zhang, G. H., and J. Huang (2008), The sparsity and bias of the lasso selection in high-dimensional linear regression, *Annals of Statistics*, *36*, 1567–1594.

Zhang, Z. W. e. a. (2010b), Mixed linear model approach adapted for genome-wide association studies, *Nature Genetics*, *42*, 355–360.

Zhou, X., P. Carbonetto, and M. Stephens (2013), Polygenic modeling with bayesian sparse linear mixed models, *PLOS Genetics*, *9*.