



University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

1-1-2016

Essays in Problems in Sequential Decisions and Large-Scale Randomized Algorithms

Peichao Peng

University of Pennsylvania, ppeichao@wharton.upenn.edu

Follow this and additional works at: <http://repository.upenn.edu/edissertations>

 Part of the [Statistics and Probability Commons](#)

Recommended Citation

Peng, Peichao, "Essays in Problems in Sequential Decisions and Large-Scale Randomized Algorithms" (2016). *Publicly Accessible Penn Dissertations*. 1941.

<http://repository.upenn.edu/edissertations/1941>

This paper is posted at ScholarlyCommons. <http://repository.upenn.edu/edissertations/1941>

For more information, please contact libraryrepository@pobox.upenn.edu.

Essays in Problems in Sequential Decisions and Large-Scale Randomized Algorithms

Abstract

In the first part of this dissertation, we consider two problems in sequential decision making.

The first problem we consider is sequential selection of a monotone subsequence from a random permutation. We find a two term asymptotic expansion for the optimal expected value of a sequentially selected monotone subsequence from a random permutation of length n . The second problem we consider deals with the multiplicative relaxation or constriction of the classical problem of the number of records in a sequence of n independent and identically distributed observations. In the relaxed case, we find a central limit theorem (CLT) with a different normalization than Renyi's classical CLT, and in the constricted case we find convergence in distribution to an unbounded random variable.

In the second part of this dissertation, we put forward two large-scale randomized algorithms.

We propose a two-step sensing scheme for the low-rank matrix recovery problem which requires far less storage space and has much lower computational complexity than other state-of-art methods based on nuclear norm minimization. We introduce a fast iterative reweighted least squares algorithm, `\textit{Guluru}`, based on subsampled randomized Hadamard transform, to solve a wide class of generalized linear models.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Michael Steele

Second Advisor

Dean Foster

Subject Categories

Statistics and Probability

ESSAYS IN PROBLEMS IN SEQUENTIAL DECISIONS AND LARGE-SCALE
RANDOMIZED ALGORITHMS

Peichao Peng

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

Co-Supervisor of Dissertation

J. Michael Steele, C.F. Koo Professor
Professor of Statistics

Dean Foster, Marie and Joseph Melone
Professor Emeritus of Statistics

Graduate Group Chairperson

Eric Bradlow, K.P. Chao Professor
Professor of Marketing, Statistics, and Education

Dissertation Committee

J. Michael Steele, C.F. Koo Professor; Professor of Statistics
Dean Foster, Marie and Joseph Melone Professor Emeritus of Statistics
Linda Zhao, Professor of Statistics

ESSAYS IN PROBLEMS IN SEQUENTIAL DECISIONS AND LARGE-SCALE
RANDOMIZED ALGORITHMS

©

2016

Peichao Peng

ACKNOWLEDGEMENTS

Completing a Ph.D. is a formidable journey that might not be possible without the steady support from advisors, friends and families.

First and foremost, I would like to express my utmost gratitude towards my advisors Prof. J. Michael Steele and Prof. Dean Foster for their guidance, patience and encouragement throughout my Ph.D. life. They provide me with precious opportunities to pursue research projects that fascinate me. They show me what constitutes good scientific research and always lead me to the right directions. More important are the subtle details that are difficult to describe precisely, but that do make a gigantic impact on me. I learn from them not only about probability, statistics and computer science, but also a variety of other things that are priceless in both academic and professional life.

I also receive continuous support from Prof. Linda Zhao. I especially appreciate her invitations to the Thanksgiving dinner at her house for the past two years. It is a genuine pleasure to have a mentor-like friend like her.

I would like to thank all my friends I made here at UPenn, especially to Zijian Guo, Shaokun Li, Zhuang Ma, Xin Lu Tan, Qinwen Wang, Haotian Xiang, Linjun Zhang, Yicheng Zhu. We laugh together through the past four years. I enjoy every meal and coffee with them. They really add lots of color to my Ph.D. life.

Thanks also goes to Hongyi Chen, Moren Gao, Junkai Jiang, Ziyang Gao, Simeng Kuang, Ya Le, Xincheng Lei, Junchi Li, Xi Lin, Junyang Qian, Xiangyu Wang, Wenzhe Wei, Lei Xu and Zeyu Zheng, who are my dearest friends dated back to high school and college. Sometimes I recall the old happy memories of the days we spent together. I am really fortunate that I can make friends with such a group of fantastic people.

Finally, I thank my parents, Shitao Peng and Liqun Yang. They show me love and care for the past 24 years without one day interruption. I am also grateful to my girlfriend, Dongfang, for being with me always. She sacrifices a lot for the convenience of me. Without their encouragement and tolerance, many beautiful moments in my life would not exist. Therefore I would like to dedicate this work to my parents and my girlfriend.

Peichao Peng
Philadelphia, PA
February, 2016

ABSTRACT

ESSAYS IN PROBLEMS IN SEQUENTIAL DECISIONS AND LARGE-SCALE RANDOMIZED ALGORITHMS

Peichao Peng

J. Michael Steele

Dean Foster

In the first part of this dissertation, we consider two problems in sequential decision making. The first problem we consider is sequential selection of a monotone subsequence from a random permutation. We find a two term asymptotic expansion for the optimal expected value of a sequentially selected monotone subsequence from a random permutation of length n . The second problem we consider deals with the multiplicative relaxation or constriction of the classical problem of the number of records in a sequence of n independent and identically distributed observations. In the relaxed case, we find a central limit theorem (CLT) with a different normalization than Renyi's classical CLT, and in the constricted case we find convergence in distribution to an unbounded random variable.

In the second part of this dissertation, we put forward two large-scale randomized algorithms. We propose a two-step sensing scheme for the low-rank matrix recovery problem which requires far less storage space and has much lower computational complexity than other state-of-art methods based on nuclear norm minimization. We introduce a fast iterative reweighted least squares algorithm, *Guluru*, based on subsampled randomized Hadamard transform, to solve a wide class of generalized linear models.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF TABLES	viii
LIST OF ILLUSTRATIONS	ix
CHAPTER 1 : Introduction	1
CHAPTER 2 : Sequential Selection of a Monotone Subsequence	
From a Random Permutation	4
2.1 Sequential Subsequence Problems	4
2.2 Recurrence Relations	7
2.3 Comparison Principles	8
2.4 An Approximation Solution	10
2.5 Proof of Theorem 2	14
2.6 Further Developments and Considerations	16
CHAPTER 3 : Relative Records: Relaxed or Constrained	19
3.1 Relaxed or Constrained Sequential Selection Processes	19
3.2 Representation as a Markov Additive Functional	22
3.3 The Dobrushin Coefficient and Its Consequences	23
3.4 When $\rho < 1$: Proof of Theorem 3	25
3.5 Proof of Theorem 4	31
3.6 The Stationary Measure and the Pantograph Equation	34
3.7 When $\rho > 1$: The Proof of Theorem 5	36
3.8 Complements to Classical Record Theory	37

3.9	More Records: Relaxed or Constrained	42
CHAPTER 4 : Low Rank Matrix Recovery via Sensing the Range Space		44
4.1	The Low Rank Matrix Recovery Problem	44
4.2	Sensing Scheme	46
4.3	Theoretical Guarantee of Recovery	48
4.4	Experiments	51
4.5	Details of Proof	56
CHAPTER 5 : Large-scale Estimation of Generalized Linear Model		66
5.1	Background	66
5.2	The Guluru Algorithm	68
5.3	Convergence Analysis	71
5.4	Experiments	73
5.5	Discussion	77
5.6	Details of Proof	78
BIBLIOGRAPHY		88

LIST OF TABLES

TABLE 1 :	Comparison Between Three Sensing Schemes	48
TABLE 2 :	Prediction Accuracy	77

LIST OF ILLUSTRATIONS

FIGURE 1 :	Comparison Between Proposed Scheme and Gaussian Ensemble	53
FIGURE 2 :	Recovery of Lena with Different Sampling Parameters	55
FIGURE 3 :	Results for Simulation Studies	75
FIGURE 4 :	Results for Real Data Studies	76

CHAPTER 1 : Introduction

In the first part of this dissertation we study two problems in sequential decision making, where the decision maker faces uncertain outcomes and has to make decisions throughout a discrete time horizon. We are most interested in the asymptotic performance of certain strategies carried out by the decision maker.

In Chapter 2, we consider the sequential monotone subsequence selection problem, where the decision maker faces the values $\pi[1], \pi[2], \dots$ from a random permutation $\pi : [1 : n] \rightarrow [1 : n]$ one by one and, when shown the value $\pi[i]$ at time i , must decide (once and for all) either to accept or reject $\pi[i]$ as element of the selected increasing subsequence. One can easily relate this to a similar problem, where we consider sequential selection from a sequence of n independently uniformly distributed random variables X_1, X_2, \dots, X_n instead of a random permutation. It was established by Samuels and Steele (1981) that

$$s(n) \sim \hat{s}(n) \sim \sqrt{2n} \tag{1.1}$$

where $s(n)$ and $\hat{s}(n)$ denote the expected value of optimal selection from a random permutation and n independent and identically distributed samples respectively. Nevertheless, there is a flurry of literature on characterizing $\hat{s}(n)$ while few have focused on analyzing $s(n)$. To the best of our knowledge, there is no finer analysis of $s(n)$ than (1.1). Given the similarities that lie between these two problems, one might hope there is a definite relationship between $s(n)$ and $\hat{s}(n)$, but this is far from intuitive. Our contributions are two-fold: in the first place we proved that $s(n) \geq \hat{s}(n)$, which implies that the decision maker is better off in the permutation problem; secondly, we managed to quantify the extent to which it is better. Specifically, $s(n)$ is larger by at least $(1/6) \log n + O(1)$.

When the decision maker adopts the greedy strategy, the selected values correspond to the record values. The number of records in a sequence of n independent and identically

distributed samples is well studied by Rényi (1962). In Chapter 3 we generalize classical records to relative records and consider the multiplicative relaxations and constrictions of the number of records, which have not been studied previously and lead to novel phenomena. First, the number of relative records is no longer independent of the distribution function. Moreover, the asymptotic behaviour is unlike the behavior that one finds for the classical record process. In the relaxed case, we find a central limit theorem (CLT) with a different normalization than Rényi's classical CLT, and in the constricted case we find convergence in distribution to an unbounded random variable.

The big data era has posed tremendous challenge to traditional non-scalable and computationally inefficient algorithms. Random projection and randomized subsampling are two powerful tools that have found their applications in a variety of problems. Based on these two ideas, in the second part of this dissertation we put forward two randomized algorithms addressing the low rank matrix recovery problem and large-scale estimation of generalized linear model respectively.

Exploration of low-rank structure is of great significance and interest in a wide range of applications. In Chapter 4, we propose a randomized two-step sensing scheme for the low-rank matrix recovery problem, which requires far less storage space and has much lower computational complexity compared with other state-of-art methods based on nuclear norm minimization. Besides exact recovery in the ideal low-rank and noiseless case, the proposed procedure is applicable to cases where the underlying matrix has full rank with decaying singular values and where the measurements suffer from noise. Expectation and concentration error bounds for both the spectral norm and the Frobenius norm are established. Finally, numerical experiments are given to support the theory.

In Chapter 5, we propose a fast iterative reweighted least squares algorithm, *Guluru*, based on subsampled randomized Hadamard transform, to solve a wide class of generalized linear models under the large n , large p and $n \gg p$ setting, where, as usual, n denotes the number of observations and p is the number of features. In each iteration, the proposed

algorithm reduces the computational complexity from $O(np^2)$ to $O(np)$. We provide theoretical guarantees that the log-likelihood achieved by Guluru upon convergence is at most $O(p^2/nr^2)$ away from the maximum log-likelihood, where r is the subsampling ratio. We also prove that the final estimator of Guluru is only $O(p/nr)$ away from the maximum likelihood estimator. Extensive empirical studies demonstrate the competitive performance of the proposed algorithm in both computational speed and prediction accuracy.

CHAPTER 2 : Sequential Selection of a Monotone Subsequence
From a Random Permutation

2.1. Sequential Subsequence Problems

In the classical monotone subsequence problem, one chooses a random permutation $\pi : [1 : n] \rightarrow [1 : n]$, and one considers the length of its longest increasing subsequence,

$$L_n = \max\{k : \pi[i_1] < \pi[i_2] < \cdots < \pi[i_k] \text{ where } 1 \leq i_1 < i_2 < \cdots < i_k \leq n\}.$$

On the other hand, in the *sequential* monotone subsequence problem one views the values $\pi[1], \pi[2], \dots$ as though they were presented over time to a decision maker who, when shown the value $\pi[i]$ at time i , must decide (once and for all) either to accept or reject $\pi[i]$ as element of the selected increasing subsequence.

The decision to accept or reject $\pi[i]$ at time i is based on just the knowledge of the time horizon n and the observed values $\pi[1], \pi[2], \dots, \pi[i]$. Thus, in slightly more formal language, the sequential selection problems amounts to the consideration of random variables of the form

$$L_n^\tau = \max\{k : \pi[\tau_1] < \pi[\tau_2] < \cdots < \pi[\tau_k] \text{ where } 1 \leq \tau_1 < \tau_2 < \cdots < \tau_k \leq n\}, \quad (2.1)$$

where the indices $\tau_i, i = 1, 2, \dots$ are stopping times with respect to the increasing sequence of σ -fields $\mathcal{F}_k = \sigma\{\pi[1], \pi[2], \dots, \pi[k]\}, 1 \leq k \leq n$. We call a sequence of such stopping times a *feasible selection strategy*, and, if we use τ as a shorthand for such a strategy, then the quantity of central interest here can be written as

$$s(n) = \sup_{\tau} \mathbb{E}[L_n^\tau], \quad (2.2)$$

where one takes the supremum over all feasible selection strategies.

It was conjectured in Baer and Brock (1968) that

$$s(n) \sim \sqrt{2n} \quad \text{as } n \rightarrow \infty, \quad (2.3)$$

and a proof of this relation was first given in Samuels and Steele (1981). A much simpler proof of (2.3) was later given by Gneden (2000) who made use of a recursion that had been used for numerical computations by Baer and Brock (1968). The main purpose of this note is to show how by a more sustained investigation of that recursion one can obtain a two term expansion.

Theorem 1 (Sequential Selection from a Random Permutation). *For $n \rightarrow \infty$ one has the asymptotic relation*

$$s(n) = \sqrt{2n} + \frac{1}{6} \log n + O(1). \quad (2.4)$$

Given what is known for some closely related problems, the explicit second order term $(\log n)/6$ gives us an unanticipated bonus. For comparison, suppose we consider sequential selection from a sequence of n independently uniformly distributed random variables X_1, X_2, \dots, X_n . In this problem a feasible selection strategy τ is again expressed by an increasing sequence of stopping times $\tau_j, j = 1, 2, \dots$, but now the stopping times are adapted to the increasing σ -fields $\widehat{\mathcal{F}}_j = \sigma\{X_1, X_2, \dots, X_j\}$. The analog of (2.1) is then

$$\widehat{L}_n^\tau = \max\{k : X_{\tau_1} < X_{\tau_2} < \dots < X_{\tau_k} \quad \text{where } 1 \leq \tau_1 < \tau_2 < \dots < \tau_k \leq n\}, \quad (2.5)$$

and the analog of (2.2) is given by

$$\widehat{s}(n) = \sup_{\tau} \mathbb{E}[\widehat{L}_n^\tau].$$

Bruss and Robertson (1991) found that for $\widehat{s}(n)$ one has a uniform upper bound

$$\widehat{s}(n) \leq \sqrt{2n} \quad \text{for all } n \geq 1, \tag{2.6}$$

so, by comparison with (2.4), we see there is a sense in which sequential selection of a monotone subsequence from a permutation is *easier* than sequential selection from an independent sequence. In part, this is intuitive; each successive observation from a permutation gives useful information about the subsequent values that can be observed. By (2.4) one quantifies how much this information helps, and, so far, we have only an analytical understanding of the source of $(1/6) \log n$. A genuinely probabilistic understanding of this term remains elusive.

Since (2.6) holds for all n and since (2.4) is only asymptotic, it also seems natural to ask if there is a relation between $\widehat{s}(n)$ and $s(n)$ that is valid for all n . There is such a relation if one gives up the logarithmic gap.

Theorem 2 (Selection for Random Permutations vs Random Sequences). *One has for all $n = 1, 2, \dots$ that*

$$\widehat{s}(n) \leq s(n).$$

Here we should also note that much more is known about $\widehat{s}(n)$ than just (2.6); in particular, there are several further connections between $s(n)$ and $\widehat{s}(n)$. These are taken up in a later section, but first it will be useful to give the proofs of Theorems 1 and 2.

The larger context for the problems studied here is the theory of Markov decision processes (or MDPs) which is closely tied to the theory of optimal stopping and the theory of on-line algorithms (cf. Puterman (1994), Shiryaev (2008), and Flat and Woeginger (1998)). The traditional heart of the theory of MDPs is the optimality equation (or Bellman equation) which presents itself here as the identity (2.7). One of our main motivations has been the expectation that (2.7) gives one an appropriate path for examining how one can extract delicate asymptotic information from max-type non-linear recursions of the kind that oc-

cur in the theory of MDPs. In this respect, it seems hopeful that tools that parallel the comparison principles of Section 2.3 and the approximate solutions of Section 2.4 may be broadly applicable, although the details will necessarily vary from problem to problem.

The proof of Theorem 1 takes most of our effort, and it is given over the next few sections. Section 2.2 develops the basic recurrence relations, and Section 2.3 develops stability relations for these recursions. In Section 2.4 we then do the calculations that support a candidate for the asymptotic approximation of $s(n)$, and we complete the proof of Theorem 1. Our arguments conclude in Section 2.5 with the brief — and almost computation free — proof of Theorem 2. Finally, in Section 2.6 we discuss further relations between $s(n)$, $\widehat{s}(n)$, and some other closely related quantities that motivate consideration of two open problems.

2.2. Recurrence Relations

One can get a recurrence relation for $s(n)$ by first step analysis. Specifically, we take a random permutation $\pi : [1 : n + 1] \rightarrow [1 : n + 1]$, and we consider its initial value $\pi[1] = k$. If we reject $\pi[1]$ as an element of our subsequence, we are faced with the problem of sequential selection from the reduced random permutation π' on an n -element set. Alternatively, if we choose $\pi[1] = k$ as an element of our subsequence, we are then faced with the problem of sequential selection for a reduced random permutation π'' of the set $\{k + 1, k + 2, \dots, n + 1\}$ that has $n + 1 - k$ elements. By taking the better of these two possibilities, we get from the uniform distribution of $\pi[1]$ that

$$s(n + 1) = \frac{1}{n + 1} \sum_{k=1}^{n+1} \max\{s(n), 1 + s(n + 1 - k)\}. \quad (2.7)$$

From the definition (2.2) of $s(n)$ one has $s(1) = 1$, so subsequent values can then be computed by (2.7). For illustration and for later discussion, we note that one has the approximate values:

n	1	2	3	4	5	6	7	8	9	10
$s(n)$	1	1.5	2	2.375	2.725	3.046	3.333	3.601	3.857	4.098
$\sqrt{2n}$	1.414	2	2.449	2.828	3.162	3.464	3.742	4	4.243	4.472.

Here we observe that for the 10 values in the table one has $s(n) \leq \sqrt{2n}$, and, in fact, this relation persists for all $1 \leq n \leq 174$. Nevertheless, for $n = 175$ one has $\sqrt{2n} < s(n)$, just as (2.4) requires for all sufficiently large values of n .

We also know from (2.2) that the map $n \mapsto s(n)$ is strictly monotone increasing, and, as a consequence, the recursion (2.7) can be written a bit more simply as

$$\begin{aligned} s(n+1) &= \frac{1}{n+1} \max_{1 \leq k \leq n} \left\{ (n-k+1)s(n) + \sum_{i=n-k+1}^n \{s(i) + 1\} \right\} \\ &= \frac{1}{n+1} \max_{1 \leq k \leq n} \left\{ (n-k+1)s(n) + k + \sum_{i=n-k+1}^n s(i) \right\}. \end{aligned} \quad (2.8)$$

In essence, this recursion goes back to Baer and Brock (1968, p. 408), and it is the basis of most of our analysis.

2.3. Comparison Principles

Given a map $g : \mathbb{N} \rightarrow \mathbb{R}$ and $1 \leq k \leq n$, it will be convenient to set

$$H(n, k, g) = k + (n - k + 1)g(n) + \sum_{i=n-k+1}^n g(i), \quad (2.9)$$

so the optimality recursion (2.8) can be written more succinctly as

$$s(n+1) = \frac{1}{n+1} \max_{1 \leq k \leq n} H(n, k, s). \quad (2.10)$$

The next two lemmas make rigorous the idea that if g is almost a solution of (2.10) for all n , then $g(n)$ is close to $s(n)$ for all n .

Lemma 2.1 (Upper Comparison). *If $\delta : \mathbb{N} \rightarrow \mathbb{R}^+$, $1 \leq g(1) + \delta(1)$, and*

$$\frac{1}{n+1} \max_{1 \leq k \leq n} H(n, k, g) \leq g(n+1) + \delta(n+1) \quad \text{for all } n \geq 1, \quad (2.11)$$

then one has

$$s(n) \leq g(n) + \sum_{i=1}^n \delta(i) \quad \text{for all } n \geq 1. \quad (2.12)$$

Proof. We set $\Delta(i) = \delta(1) + \delta(2) + \cdots + \delta(i)$, and we argue by induction. Specifically, using (2.12) for $1 \leq i \leq n$ we have

$$\begin{aligned} H(n, k, s) &= k + (n - k + 1)s(n) + \sum_{i=n-k+1}^n s(i) \\ &\leq k + (n - k + 1)(g(n) + \Delta(n)) + \sum_{i=n-k+1}^n \{g(i) + \Delta(i)\} \end{aligned}$$

so by monotonicity of $\Delta(\cdot)$ we have

$$\frac{1}{n+1} H(n, k, s) \leq \frac{1}{n+1} H(n, k, g) + \Delta(n).$$

Now, when we take the maximum over $k \in [1 : n]$, the recursion (2.8) and the induction condition (2.11), give us

$$\begin{aligned} s(n+1) &\leq \frac{1}{n+1} \max_{1 \leq k \leq n} H(n, k, g) + \Delta(n) \\ &\leq g(n+1) + \delta(n+1) + \Delta(n) = g(n+1) + \Delta(n+1), \end{aligned}$$

so induction establishes (2.12) for all $n \geq 1$. □

Naturally, there is a lower bound comparison principle that parallels Lemma 2.1. The statement has several moving parts, so we frame it as a separate lemma even though its proof can be safely omitted.

Lemma 2.2 (Lower Comparison). *If $\delta : \mathbb{N} \rightarrow \mathbb{R}^+$, $g(1) - \delta(1) \leq 1$, and*

$$g(n+1) - \delta(n+1) \leq \frac{1}{n+1} \max_{1 \leq k \leq n} H(n, k, g) \quad \text{for all } n \geq 0,$$

then one has

$$g(n) - \sum_{i=1}^n \delta(i) \leq s(n) \quad \text{for all } n \geq 1.$$

2.4. An Approximation Solution

We now argue that the function $f : \mathbb{N} \rightarrow \mathbb{R}$ defined by

$$f(n) = \sqrt{2n} + \frac{1}{6} \log n, \tag{2.13}$$

gives one an approximate solution of the recurrence equation (2.8) for $n \mapsto s(n)$.

Proposition 2.3. *There is a constant $0 < B < \infty$ such that for all $n \geq 1$, one has*

$$-Bn^{-3/2} \leq \frac{1}{n+1} \left\{ \max_{1 \leq k \leq n} H(n, k, f) \right\} - f(n+1) \leq Bn^{-3/2}. \tag{2.14}$$

FIRST STEP: LOCALIZATION OF THE MAXIMUM

To deal with the maximum in (2.14), we first estimate

$$k^*(n) = \text{locmax}_k H(n, k, f).$$

From the definition (2.9) of $H(n, k, f)$ we find

$$H(n, k+1, f) - H(n, k, f) = 1 - f(n) + f(n-k),$$

and, from the definition (2.13) of f , we see this difference is monotone decreasing function

of k ; accordingly, we also have the representation

$$k^*(n) = 1 + \max\{k : 0 \leq 1 - f(n) + f(n - k)\}. \quad (2.15)$$

Now, for each $n = 1, 2, \dots$ we then consider the function $D_n : [0, n] \rightarrow \mathbb{R}$ defined by setting

$$D_n(x) = 1 - f(n) + f(n - x) = 1 - \{\sqrt{2n} - \sqrt{2(n - x)}\} - \frac{1}{6}\{\log n - \log(n - x)\}.$$

This function is strictly decreasing with $D_n(0) = 1$ and $D_n(n) = -\infty$, so there is a unique solution of the equation $D_n(x) = 0$. For $x \in [0, n)$ we also have the easy bound

$$D_n(x) = 1 - \frac{1}{2} \int_{2(n-x)}^{2n} \frac{1}{\sqrt{u}} du - \frac{1}{6} \log(n/(n-x)) \leq 1 - \frac{x}{\sqrt{2n}}.$$

This gives us $D_n(\sqrt{2n}) \leq 0$, so by monotonicity we have $x_n \leq \sqrt{2n}$.

To refine this bound to an asymptotic estimate, we start with the equation $D_n(x_n) = 0$ and we apply Taylor expansions to get

$$\begin{aligned} 1 &= \sqrt{2n} \left\{ 1 - (1 - x_n/n)^{1/2} \right\} - \frac{1}{6} \log(1 - x_n/n) \\ &= \sqrt{2n} \left\{ \frac{x_n}{2n} + O(x_n^2/n^2) \right\} + O(x_n/n). \end{aligned}$$

By simplification, we then get

$$\sqrt{2n} = x_n + O(x_n^2/n) + O(x_n/n^{1/2}) = x_n + O(1), \quad (2.16)$$

where in the last step we used our first bound $x_n \leq \sqrt{2n}$.

Finally, by (2.16) and the characterization (2.15), we immediately find the estimate that we need for $k^*(n)$.

Lemma 2.4. *There is a constant $A > 0$ such that for all $n \geq 1$, we have*

$$\sqrt{2n} - A \leq k^*(n) \leq \sqrt{2n} + A. \quad (2.17)$$

Remark 2.5. The relations (2.16) and (2.17) can be sharpened. Specifically, if we use a two-term Taylor series with integral remainders, then one can show $\sqrt{2n} - 2 \leq x_n$. Since we already know that $x_n \leq \sqrt{2n}$, we then see from the characterization (2.15) and integrality of $k^*(n)$ that we can take $A = 2$ in Lemma 2.4. This refinement does not lead to a meaningful improvement in Theorem 1, so we omit the details of the expansions with remainders.

COMPLETION OF PROOF OF PROPOSITION 2.3

To prove Proposition 2.3, we first note that the definition (2.9) of $H(n, k, f)$ one has for all $1 \leq k \leq n$ that

$$\frac{1}{n+1}H(n, k, f) = f(n) + \frac{1}{n+1} \left\{ k - \sum_{i=1}^{k-1} (f(n) - f(n-i)) \right\} \quad (2.18)$$

The task is to estimate the right-hand side of (2.18) when $k = k^*(n)$ and $k^*(n)$ is given by (2.15).

For the moment, we assume that one has $k \leq D\sqrt{n}$ where $D > 0$ is constant. With this assumption, we find that after making Taylor expansions we get from explicit summations that

$$\begin{aligned} \sum_{i=1}^{k-1} (f(n) - f(n-i)) &= \sum_{i=1}^{k-1} \left(\sqrt{2n} - \sqrt{2(n-i)} \right) + \sum_{i=1}^{k-1} \left(\frac{\log n}{6} - \frac{\log(n-i)}{6} \right) \\ &= \sum_{i=1}^{k-1} \left(\frac{i}{\sqrt{2n}} + \frac{i^2}{4n\sqrt{2n}} + O\left(\frac{i^3}{n^{5/2}}\right) \right) + \sum_{i=1}^{k-1} \left(\frac{i}{6n} + O\left(\frac{i^2}{n^2}\right) \right) \\ &= \frac{(k-1)k}{2\sqrt{2n}} + \frac{(k-1)k(2k-1)}{24n\sqrt{2n}} + \frac{(k-1)k}{12n} + O(n^{-1/2}), \end{aligned} \quad (2.19)$$

where the implied constant of the remainder term depends only on D .

We now define $r(n)$ by the relation $k^*(n) = \sqrt{2n} + r(n)$, and we note by (2.17) that $|r(n)| \leq A$. Direct algebraic expansions then give us the elementary estimates

$$\frac{(k^*(n) - 1)k^*(n)}{12n} = \frac{1}{6} + O(n^{-1/2})$$

and

$$\frac{(k^*(n) - 1)k^*(n)(2k^*(n) - 1)}{24n\sqrt{2n}} = \frac{1}{6} + O(n^{-1/2}),$$

where in each case the implied constant depends only on A .

Estimation of the first summand of (2.19) is slightly more delicate than this since we need to account for the dependence of this term on $r(n)$; specifically we have

$$\begin{aligned} \frac{(k^*(n) - 1)k^*(n)}{2\sqrt{2n}} &= \frac{(\sqrt{2n} + r(n) - 1)(\sqrt{2n} + r(n))}{2\sqrt{2n}} \\ &= \sqrt{n/2} + r(n) - \frac{1}{2} + O(n^{-1/2}). \end{aligned}$$

Now, for a pleasing surprise, we note from the last estimate and from the definition of $k^*(n)$ and $r(n)$ that we have cancelation of $r(n)$ when we then compute the critical sum; thus, one has simply

$$k^*(n) - \sum_{i=1}^{k^*(n)-1} (f(n) - f(n-i)) = \sqrt{n/2} + \frac{1}{6} + O(n^{-1/2}). \quad (2.20)$$

Finally, from the formula (2.13) for $f(\cdot)$, we have the Taylor expansion

$$f(n+1) - f(n) = \frac{1}{\sqrt{2n}} + \frac{1}{6n} + O(n^{-3/2}), \quad (2.21)$$

so, when we return to the identity (2.18), we see that the estimates (2.20) and (2.21) give

us the estimate

$$\begin{aligned} & \frac{1}{n+1} \left\{ \max_{1 \leq k \leq n} H(n, k, f) \right\} - f(n+1) \\ &= \frac{1}{n+1} \left(\sqrt{n/2} + \frac{1}{6} + O(n^{-1/2}) \right) + f(n) - f(n+1) = O(n^{-3/2}). \end{aligned}$$

Here the implied constant is absolute, and the proof of Proposition 2.3 is complete.

COMPLETION OF PROOF OF THEOREM 1

Lemmas 2.1 and 2.2 combine with Proposition 2.3 to tell us that by summing the sequence $n^{-3/2}$, $n = 1, 2, \dots$ and by writing $\zeta(z) = 1 + 2^{-z} + 3^{-z} + \dots$ one has

$$|s(n) - f(n)| \leq \zeta(3/2)B \leq (2.62)B \quad \text{for all } n \geq 1.$$

This is slightly more than one needs to complete the proof of Theorem 1.

2.5. Proof of Theorem 2

The sequential monotone selection problem is a finite horizon Markov decision problem with bounded rewards and finite action space, and for such problems it is known one cannot improve upon an optimal deterministic strategy by the use of strategies that incorporate randomization, Bertsekas and Shreve (1978, Corollary 8.5.1)cf.. The proof of Theorem 2 exploits this observation by constructing a randomized algorithm for the sequential selection of a monotone subsequence from a random permutation.

We first recall that if e_i , $i = 1, 2, \dots, n+1$ are independent exponentially distributed random variables with mean 1 and if one sets

$$Y_i = \frac{e_1 + e_2 + \dots + e_i}{e_1 + e_2 + \dots + e_{n+1}},$$

then the vector (Y_1, Y_2, \dots, Y_n) has the same distribution as the vector of order statistics

$(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ of an i.i.d. sample of size n from the uniform distribution (see e.g. Feller (1971), p. 77). Next we let \mathcal{A} denote an optimal algorithm for sequential selection of an increasing subsequence from an independent sample X_1, X_2, \dots, X_n from the uniform distribution, and we let $\tau(\mathcal{A})$ denote the associated sequence of stopping times. If $\widehat{L}_n^{\tau(\mathcal{A})}$ denotes the length of the subsequence that is chosen from X_1, X_2, \dots, X_n when one follows the strategy $\tau(\mathcal{A})$ determined by \mathcal{A} , then by optimality of \mathcal{A} for selection from X_1, X_2, \dots, X_n we have

$$\widehat{s}(n) = \sup_{\tau} \mathbb{E}[\widehat{L}_n^{\tau}] = \mathbb{E}[\widehat{L}_n^{\tau(\mathcal{A})}].$$

We use the algorithm \mathcal{A} to construct a new randomized algorithm \mathcal{A}' for sequential selection of an increasing sequence from a random permutation $\pi : [n] \mapsto [n]$. First, the decision maker generates independent exponential random variables $e_i, i = 1, 2, \dots, n + 1$ as above. This is done off-line, and this step can be viewed as an internal randomization.

Now, for $i = 1, 2, \dots, n$, when we are presented with $\pi[i]$ at time i , we compute $X_i = Y_{\pi[i]}$. Finally, if at time i the value X_i would be accepted by the algorithm \mathcal{A} , then the algorithm \mathcal{A}' accepts $\pi[i]$. Otherwise the newly observed value $\pi[i]$ is rejected. By our construction we have

$$\mathbb{E}[L_n^{\tau(\mathcal{A}')}] = \mathbb{E}[\widehat{L}_n^{\tau(\mathcal{A})}] = \widehat{s}(n). \quad (2.22)$$

Moreover, \mathcal{A}' is a randomized algorithm for construction an increasing subsequence of a random permutation π . By definition, $s(n)$ is the expected length of a monotone subsequence selected from a random permutation by an optimal deterministic algorithm, and by our earlier observation, the randomized algorithm \mathcal{A}' cannot do better. Thus, from (2.22) one has $\widehat{s}(n) \leq s(n)$, and the proof of Theorem 1 is complete.

2.6. Further Developments and Considerations

The uniform upper bound (2.6) was obtained by Bruss and Robertson (1991) as a consequence of a bound on the expected value of the random variable

$$N(s) = \max \left\{ |A| : \sum_{i \in A} X_i \leq s \right\}$$

where the observations $\{X_i : i \in [1 : n]\}$ have a common continuous distribution with support in $[0, \infty)$. This bound was extended in Steele (2016) to accommodate non-identically distributed random variables, and, as a consequence, one finds some new bounds for the sequential knapsack problems.

On the other hand, this extension does not help one to refine or generalize (2.6), since, as Coffman et al. (1987) first observed, the sequential knapsack problem and the sequential increasing subsequence problem are equivalent only when the observations are uniformly and identically distributed. Certainly, one may consider the possibility of analogs of (2.6) for non-identically distributed random variables, but, as even deterministic examples show, the formulation of such analogs is problematical.

Here we should also note that Gnedin (1999) gave a much different proof of (2.6), and, moreover, he generalized the bound in a way that accommodates random samples with random sizes. More recently, Arlotto et al. (2015a) obtained yet another proof (2.6) as a corollary to bounds on the *quickest selection problem*, which is an informal dual to the traditional selection problem.

Since the bound (2.6) is now well understood from several points of view, it is reasonable to ask about the possibility of some corresponding uniform bound on $s(n)$. The numerical values that we noted after the recursion (2.6) and the relation

$$s(n) = \sqrt{2n} + \frac{1}{6} \log n + O(1) \tag{2.23}$$

from Theorem 1 both tell us that one cannot expect a uniform bound for $s(n)$ that is as simple as that for $\widehat{s}(n)$ given by (2.6). Nevertheless, numerical evidence suggests that the $O(1)$ term in (2.23) is always negative. The tools used here cannot confirm this conjecture, but the multiple perspectives available for (2.6) give one hope.

A closely related issue arises for $\widehat{s}(n)$ when one considers lower bounds. Here the first steps were taken by Bruss and Delbaen (2001) who considered i.i.d. samples of size N_ν where N_ν is an independent random variable with the Poisson distribution with mean ν . If now we write $\widehat{s}(\nu)$ for the expected value of the length of the subsequence selected by an optimal algorithm in the Bruss-Delbaen framework, then they proved that there is a constant $c > 0$ such that

$$\sqrt{2\nu} - c \log \nu \leq \widehat{s}(\nu);$$

moreover, Bruss and Delbaen (2004) subsequently proved that for the optimal feasible strategy $\tau_* = (\tau_1, \tau_2, \dots)$ the random variable

$$\widehat{L}_{N_\nu}^{\tau_*} = \max\{k : X_{\tau_1} < X_{\tau_2} < \dots < X_{\tau_k} \quad \text{where } 1 \leq \tau_1 < \tau_2 < \dots < \tau_k \leq N_\nu\},$$

also satisfies a central limit theorem. Arlotto et al. (2015b) considered the de-Poissonization of these results, and it was found that one has the corresponding CLT for $\widehat{L}_n^{\tau_*}$ where the sample size n is deterministic. In particular, one has the bounds

$$\sqrt{2n} - c \log n \leq \widehat{s}(n) \leq \sqrt{2n}.$$

Now, by analogy with (2.23), one strongly expects that there is a constant $c > 0$ such that

$$\widehat{s}(n) = \sqrt{2n} - c \log n + O(1). \tag{2.24}$$

Still, a proof this conjecture is reasonably remote, since, for the moment, there is not even a compelling candidate for the value of c .

For a second point of comparison, one can recall the *non-sequential* selection problem where one studies

$$\ell(n) = \mathbb{E}[\max\{k : X_{i_1} < X_{i_2} < \dots < X_{i_k}, 1 \leq i_1 < i_2 < \dots < i_k \leq n\}].$$

Through a long sequence of investigations culminating with Baik et al. (1999), it is now known that one has

$$\ell(n) = 2\sqrt{n} - \alpha n^{1/6} + o(n^{1/6}), \tag{2.25}$$

where the constant $\alpha = 1.77108\dots$ is determined numerically in terms of solutions of a Painlevé equation of type II. Romik (2014) gives an elegant account of the extensive technology behind (2.25), and there are interesting analogies between $\ell(n)$ and $s(n)$. Nevertheless, a proof of the conjecture (2.24) seems much more likely to come from direct methods like those used here to prove (2.23).

Finally, one should note that the asymptotic formulas for $n \mapsto \ell(n)$, $n \mapsto s(n)$, and $n \mapsto \widehat{s}(n)$ all suggest that these maps are concave, but so far only $n \mapsto \widehat{s}(n)$ has been proved to be concave (cf. Arlotto et al. (2015b, p. 3604)).

CHAPTER 3 : Relative Records: Relaxed or Constrained

3.1. Relaxed or Constrained Sequential Selection Processes

Let $X_i, i = 1, 2, \dots$ be a sequence of independent random variables with a common continuous distribution F with support in $[0, \infty)$, and let ρ denote a non-negative constant. Next, we set $\tau_1 = 1$, and we define a sequence of stopping times by taking

$$\tau_k = \min\{j : X_j \geq \rho X_{\tau_{k-1}}\} \quad \text{for } k > 1. \quad (3.1)$$

The random variables of main interest here are then given by

$$R_n(\rho) = \max\{k : \tau_k \leq n\}. \quad (3.2)$$

When $\rho = 1$ the times $\tau_k, k = 1, 2, \dots$ are precisely the times at which new record values are observed, and $R_n(1)$ is the total number of records that are observed in the time interval $[1 : n] = \{1, 2, \dots, n\}$.

The random variable $R_n(1)$ has been well understood for a long time. In particular, Rényi (1962) found among other things that $\mathbb{E}[R_n(1)] \sim \log n$ and $\text{Var}[R_n(1)] \sim \log n$; moreover, he found that one has

$$\frac{R_n(1) - \log n}{\sqrt{\log n}} \Rightarrow N(0, 1), \quad (3.3)$$

where, as usual, the symbol \Rightarrow denotes convergence in distribution and $N(0, 1)$ denotes the standard Gaussian distribution.

The cases $\rho \in (0, 1)$ and $\rho \in (1, \infty)$ have not been considered previously, and they lead to some novel phenomena. First, the distribution of $R_n(\rho)$ is no longer independent of F . Moreover, one finds that the asymptotic behavior is unlike the behavior that one finds for the classical record process $\{R_n(1) : n \geq 1\}$.

The most interesting case is when $\rho \in (0, 1)$ where, in comparison to the classical record process, one has *relaxed* the condition that is imposed on the condition for sequential selections. In this case one again has a central limit theorem, but it differs substantially from Renyi's. In particular, for $\rho \in (0, 1)$ the mean and variance grow linearly and the summands are not independent.

For the sake of brevity, we say a distribution function is in the *selection class* \mathcal{S}_L if there is an $L \in (0, \infty)$ such that $F(0) = 0$, $F(L) = 1$, and F is continuous and strictly monotone on $(0, L)$. For example, the uniform distribution on $[0, 1]$ is in \mathcal{S}_1 , and for any $L > 0$ the truncated exponential distribution $F(x) = (1 - e^{-x})/(1 - e^{-L})$ is in \mathcal{S}_L . For these examples, one has a density, but, in general, a distribution in \mathcal{S}_L need not have a density.

Theorem 3 (Mean, Variance, and CLT when $0 < \rho < 1$). *If $X_i, i = 1, 2, \dots$ are independent and if $F \in \mathcal{S}_L$, then there are constants $\mu_\rho(F) > 0$ and $\sigma_\rho(F) > 0$ such that*

$$\mathbb{E}[R_n(\rho)] \sim n\mu_\rho(F) \quad \text{and} \quad \text{Var}[R_n(\rho)] \sim n\sigma_\rho^2(F), \quad (3.4)$$

and one has a central limit theorem

$$\frac{R_n(\rho) - n\mu_\rho(F)}{\sigma_\rho(F)\sqrt{n}} \Rightarrow N(0, 1). \quad (3.5)$$

After we develop some useful connections with the theory of Markov chains in Sections 3.2 and 3.3, the proof of Theorem 3 is given in Section 3.4. The main issues are the proofs of the relations (3.4) and the proof of $\sigma_\rho^2(F) > 0$. Once these facts are in hand, the convergence (3.5) then follows from general theory; for example, one can obtain (3.5) directly from Arlotto and Steele (2016) Theorem 1, Corollary 2. Alternatively, with a page or two of extra work, one can obtain (3.5) by first generalizing other known central limit theorems for additive functionals of Markov processes. In either case, the proof that $\sigma_\rho^2(F) > 0$ is a make-or-break step.

For a general $F \in \mathcal{S}_L$, the task of determining the constants $\mu_\rho(F)$ and $\sigma_\rho(F)$ seems

intractable. Nevertheless, in the case of most interest when F is the uniform distribution U on $[0, 1]$, there is at least an explicit series formula for the mean.

Theorem 4 (Moments for Uniform Distribution). *If $\rho \in (0, 1)$ and if the random variables $X_i, i = 1, 2, \dots$ have the uniform distribution U on $[0, 1]$, then we have*

$$\mu_\rho(U) = 1 - \frac{\rho}{2} - \frac{\rho}{3} \left(\rho - \frac{\rho^2}{2} \right) - \frac{\rho}{4} \left(\rho - \frac{\rho^2}{2} \right) \left(\rho - \frac{\rho^3}{3} \right) - \frac{\rho}{5} \left(\rho - \frac{\rho^2}{2} \right) \left(\rho - \frac{\rho^3}{3} \right) \left(\rho - \frac{\rho^4}{4} \right) \dots \quad (3.6)$$

The proof of Theorem 4 is then given in Section 3.5 where we also develop an equation of the pantograph type for the stationary distribution of the driving Markov chain. We do not solve this equation, but we use it to derive the Mellin transform for the stationary distribution. This we use in turn to get the required explicit formula (3.6) for $\mu_\rho(U)$.

While there is little hope of finding a correspondingly explicit representation for $\sigma_\rho^2(F)$ even for $F = U$, we do find in Section 3.4 that there is a useful series representation (3.26) for $\sigma_\rho^2(F)$.

When $\rho > 1$, one no longer has a central limit theorem. Instead one has almost sure convergence to an unbounded random variable with a well-behaved moment generating function.

Theorem 5 (Distributional Limit when $\rho > 1$). *If $F \in \mathcal{S}_L$ and if $F(x) = O(x)$ in the neighbourhood of 0, then for each $\rho > 1$ there exists an unbounded random variable N_ρ with moment generating function*

$$\mathbb{E}[\exp(sN_\rho)] < \infty \quad \text{for } |s| < \log \rho, \quad (3.7)$$

such that with probability one $R_n(\rho) \nearrow N_\rho$ as $n \rightarrow \infty$.

The case $\rho > 1$ of the sequential selection process can be viewed as a “degenerate case” where one no longer has a central limit theory. This is less interesting than the cases $\rho = 1$ or $\rho \in (0, 1)$ where one has central limit theorems of differing kinds. Still, for completeness,

we give a brief — but complete — analysis of this case in Section 3.7.

Section 3.8 then gives refinements of several kinds of Renyi’s classic formula for the expected number of records. For example, consider the number $R_n^x(1)$ of records that are larger than x . When F is the uniform distribution on $[0, 1]$, we find

$$\mathbb{E}[R_n^x(1)] = H_n - \sum_{k=1}^n \frac{x^k}{k}. \quad (3.8)$$

This formula recaptures Renyi’s classic harmonic sum when we set $x = 0$, yet its proof shares nothing in common with classic argument of Rényi (1962). Moreover, the methods that lead one to (3.8) yield further generalizations for the quantities $\mathbb{E}[R_n^x(\rho)]$ and $\lim_n \{\mathbb{E}[R_n^x(\rho)] - \mathbb{E}[R_n^y(\rho)]\}$ that are defined more fully in Section 3.8.

In Section 3.9 we make more explicit the senses in which the values chosen by the selection process (3.1) can be viewed as relaxed or constrained records. We also show how in the relaxed case $\rho \in (0, 1)$, the selected values can differ greatly from any notion of approximate record, even though our selection process and various approximate record processes may both contain the record process as limiting cases.

3.2. Representation as a Markov Additive Functional

For $k = 1, 2, \dots$ we take Y_k to be the last value that has been accepted by the selection process during the time interval $[1 : k]$; that is, we set

$$Y_k = X_{\tau_j} \quad \text{where } j = \max\{m : \tau_m \leq k\}.$$

The values Y_k , $k = 1, 2, \dots$ determine a Markov chain where if one is in state x then one stays in state x with probability $F(\rho x)$ and with probability $1 - F(\rho x)$ one moves to a point y in the set $[\rho x, 1] \setminus \{x\}$ that is chosen according to the probability measure $dF(y)/(1 - F(\rho x))$. In other words, if we also set $Y_0 = 0$ then the process $\{Y_k : k \in [0 : \infty)\}$, has the transition

kernel

$$K_{\rho,F}(x, A) = F(\rho x) \mathbb{1}(x \in A) + \int_{\rho x}^L \mathbb{1}(y \in A) dF(y). \quad (3.9)$$

Now, in terms of the Markov chain $\{Y_k : 0 = 1, 2, \dots\}$ we have the representation

$$R_n(\rho) = \sum_{k=1}^n \mathbb{1}[Y_k \neq Y_{k-1}], \quad (3.10)$$

since we accept a new value precisely at the times when the state of the chain $\{Y_n\}$ changes. Most of Theorem 3 follows from this representation after we establish a few analytic properties of the Markov chain $\{Y_n\}$.

Remark 3.1. Here one should note that by definition $R_n(\rho)$ is a function of the independent random variables $\{X_1, X_2, \dots, X_n\}$, and we simply write $\mathbb{E}[R_n(\rho)]$ and $\text{Var}[R_n(\rho)]$ when $R_n(\rho)$ is viewed in this way. On the other hand, by the representation (3.10), we can also view $R_n(\rho)$ as a function of $\{Y_1, Y_2, \dots, Y_n\}$ and the distribution of this sequence depends on the initial distribution of the Markov chain. When we take the second point of view it is natural (and necessary) to write $\mathbb{E}_\mu[R_n(\rho)]$ and $\text{Var}_\mu[R_n(\rho)]$ whenever Y_0 has the distribution μ . By construction, we always have $\text{Var}[R_n(\rho)] = \text{Var}_0[R_n(\rho)]$ and $\mathbb{E}[R_n(\rho)] = \mathbb{E}_0[R_n(\rho)]$.

3.3. The Dobrushin Coefficient and Its Consequences

There are several ways one can investigate the Markov chain defined by (3.9), but here it is especially efficient to first estimate the Dobrushin coefficient.

Definition 3.2 (Dobrushin Coefficient). If K is a Markov transition function on a Borel state space \mathcal{X} and if $\mathcal{B}(\mathcal{X})$ denotes the collection of Borel subsets of \mathcal{X} , then the *Dobrushin coefficient* $\delta(K)$ of the kernel K is defined by

$$\delta(K) = \sup_{x_1, x_2 \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})} |K(x_1, A) - K(x_2, A)|.$$

Lemma 3.3 (Dobrushin Coefficient for $K_{\rho,F}$). *If $F \in \mathcal{S}_L$ and if $K_{\rho,F}$ is the transition*

kernel given by (3.9), then one has

$$\delta(K_{\rho,F}) \leq F(\rho L) < 1. \quad (3.11)$$

Proof. If we assume $x_1 < x_2$, then for any Borel set $A \subset [0, L]$ we have from (3.9) that

$$\begin{aligned} \Delta &\stackrel{\text{def}}{=} K_{\rho,F}(x_1, A) - K_{\rho,F}(x_2, A) \\ &= F(\rho x_1) \mathbb{1}(x_1 \in A) + \int_{\rho x_1}^1 \mathbb{1}(y \in A) dF(y) - F(\rho x_2) \mathbb{1}(x_2 \in A) - \int_{\rho x_2}^1 \mathbb{1}(y \in A) dF(y) \\ &= F(\rho x_1) \mathbb{1}(x_1 \in A) - F(\rho x_2) \mathbb{1}(x_2 \in A) + \int_{\rho x_1}^{\rho x_2} \mathbb{1}(y \in A) dF(y). \end{aligned}$$

After majorizing the positive terms, we see from monotonicity of F that

$$\Delta \leq F(\rho x_1) + \{F(\rho x_2) - F(\rho x_1)\} = F(\rho x_2) \leq F(\rho L).$$

On the other hand, if we keep just the one negative terms, then we have

$$\Delta \geq -F(\rho x_2) \geq -F(\rho L),$$

and these two bounds on Δ complete the proof of (3.11). \square

Nagaev (2015) proved that for any Markov chain with kernel K and Dobrushin coefficient $\delta(K) < 1$, there is a probability measure ν on the state space \mathcal{X} that is stationary under K , and, most notably, if $K^{(n)}$ denotes the n step transition kernel, then one has the total variation bound

$$|K^{(n)}(x, A) - \nu(A)| \leq 2[\delta(K)]^n \quad \text{for all } x \in \mathcal{X} \text{ and } A \in \mathcal{B}(\mathcal{X}). \quad (3.12)$$

Now, we let $\{Z_n : n = 0, 1, 2, \dots\}$ be the Markov chain associated with the kernel K , and we write \mathbb{E}_x and \mathbb{E}_ν for the corresponding expectation operators where accordingly as

$Z_0 = x \in \mathcal{X}$ or $Z_0 \sim \nu$. By the total variation bound (3.12) and approximation by step functions, one can then check that for any bounded measurable $g : D = \mathcal{X}^4 \rightarrow \mathbb{R}$, one has for any fixed $0 \leq i \leq j \leq k$ and $n \rightarrow \infty$ that

$$\mathbb{E}_x[g(Z_n, Z_{n+i}, Z_{n+j}, Z_{n+k})] - \mathbb{E}_\nu[g(Z_0, Z_i, Z_j, Z_k)] = O(\|g\|_\infty [\delta(K)]^n), \quad (3.13)$$

where here we set $\|g\|_\infty = \sup_{v \in D} |g(v)|$.

The implied constant in (3.13) is absolute; in fact, it can be taken to be 4. Naturally, we also have analogous relations for functions of fewer than four variables or more than four variables. Here we only need (3.13) and its analog for functions of two variables.

3.4. When $\rho < 1$: Proof of Theorem 3

We now restrict attention to the Markov chain with transition kernel $K_{\rho, F}(\cdot, \cdot)$ given by (3.9). By the bound (3.11) we have $\delta \equiv \delta(K_{\rho, F}) < 1$, so the stationary distribution exists. We consider two initial distributions: in the first case we take $Y_0 \equiv 0$; and in the second case we assume that Y_0 has the stationary distribution ν . By the two variable analog of (3.13) for $g(Y_n, Y_{n+1}) = \mathbb{1}(Y_n \neq Y_{n+1})$ we have

$$\mathbb{E}_0[\mathbb{1}(Y_{k-1} \neq Y_k)] = \mathbb{E}_\nu[\mathbb{1}(Y_0 \neq Y_1)] + O(\delta^k). \quad (3.14)$$

From (3.14) and the representation (3.10), we see by geometric summation that

$$\begin{aligned} \mathbb{E}_0[R_n(\rho)] &= \sum_{k=1}^n \mathbb{E}_0[\mathbb{1}(Y_{k-1} \neq Y_k)] = n \mathbb{E}_\nu[\mathbb{1}(Y_0 \neq Y_1)] + O(1) \\ &= n \int_{\mathcal{X}} \int_{\mathcal{X}} \mathbb{1}[s \neq t] K_{\rho, F}(s, dt) d\nu(s) + O(1), \end{aligned} \quad (3.15)$$

where the double integral is just $\mathbb{E}_\nu[\mathbb{1}(Y_0 \neq Y_1)]$ written in longhand. This gives us the first assertion (3.4) of Theorem 3 in a form that is a bit more explicit; in particular, (3.15) tells

us that in (3.4) we have

$$\mu_\rho(F) = \int_{\mathcal{X}} \int_{\mathcal{X}} \mathbb{1}[s \neq t] K_{\rho,F}(s, dt) d\nu(s). \quad (3.16)$$

To find the asymptotic variance of $R_n(\rho)$, we introduce two sequences of random variables:

$$U_k = \mathbb{1}[Y_{k-1} \neq Y_k] - \mathbb{E}_0(\mathbb{1}[Y_{k-1} \neq Y_k]) \text{ and } V_k = \mathbb{1}[Y_{k-1} \neq Y_k] - \mathbb{E}_\nu(\mathbb{1}[Y_{k-1} \neq Y_k]).$$

Both U_k and V_k are functions of the Markov process $\{Y_k : k = 0, 1, \dots\}$, so in particular, both $\{U_n\}$ and $\{V_n\}$ depend on the initial value Y_0 . For clarity one should note that U_k has mean zero when $Y_0 \equiv 0$ and V_k has mean zero when Y_0 follows the stationary distribution ν .

The random variables U_k and V_k differ by a *constant* that depends on k , and by (3.14) the constant is not larger than $O(\delta^k)$. Thus, by the representation (3.10), we have

$$\text{Var}[R_n(\rho)] = \text{Var}_0[R_n(\rho)] = \mathbb{E}_0 \left[\left(\sum_{k=1}^n U_k \right)^2 \right] = \mathbb{E}_0 \left[\left(\sum_{k=1}^n V_k \right)^2 \right] + O(1). \quad (3.17)$$

Now, when we expand the second sum in (3.17) and write

$$\mathbb{E}_0 \left[\left(\sum_{k=1}^n V_k \right)^2 \right] = \sum_{k=1}^n \mathbb{E}_0[V_k^2] + 2 \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} \mathbb{E}_0[V_i V_{i+j}] \stackrel{\text{def}}{=} A_n + B_n. \quad (3.18)$$

To estimate the first sum A_n of (3.18), we apply (3.13) just as we did in the derivation of (3.14), and this time we find

$$\mathbb{E}_0[V_k^2] = \mathbb{E}_\nu[V_k^2] + O(\delta^k) = \mathbb{E}_\nu[V_1^2] + O(\delta^k).$$

Summation then gives us

$$A_n = n \mathbb{E}_\nu[V_1^2] + O(1). \quad (3.19)$$

To deal with the double sum B_n , we first need a lemma to help us estimate the summands of B_n .

Lemma 3.4. *For any initial distribution μ one has*

$$\mathbb{E}_\mu[V_i V_{i+j}] = O(\delta^j) \quad \text{for all } i, j \geq 0. \quad (3.20)$$

Proof. To exploit the Markov property for the chain $\{Y_n : n = 0, 1, \dots\}$, we first condition on Y_{i-1} and Y_i and note that

$$\mathbb{E}_\mu[V_i V_{i+j}] = \mathbb{E}_\mu[V_i \mathbb{E}_\mu[V_{i+j} | Y_{i-1}, Y_i]] = \mathbb{E}_\mu[V_i \mathbb{E}_\mu[V_{i+j} | Y_i]] = \mathbb{E}_\mu[V_i \mathbb{E}_{Y_i}[V_j]]. \quad (3.21)$$

If we use (3.13) as before, then we see that for all $x \in \mathcal{X}$ we have

$$\mathbb{E}_x[V_n] = \mathbb{E}_\nu[V_n] + O(\delta^n),$$

and the implied constant does not depend on x . When we insert this in (3.21) and recall that the definition of V_n gives us $\mathbb{E}_\nu[V_n] = 0$, the proof of the lemma is complete. \square

Lemma 3.4 helps us deal with cross terms with large j , but we also need a relation that deals with arbitrary j . Here, we again use (3.13) to get for all $j \geq 0$ that

$$\mathbb{E}_0[V_i V_{i+j}] = \mathbb{E}_\nu[V_i V_{i+j}] + O(\delta^i) = \mathbb{E}_\nu[V_1 V_{1+j}] + O(\delta^i). \quad (3.22)$$

Now, to calculate B_n , we first apply Lemma 3.4 to the cross terms $\mathbb{E}_0[V_i V_{i+j}]$ where $i \leq j$

and then apply (3.22) to the rest to obtain

$$B_n = 2 \sum_{j=1}^{n-1} \sum_{i=1}^j O(\delta^j) + 2 \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} \sum_{i=j+1}^{n-1} \left\{ \mathbb{E}_\nu[V_1 V_{1+j}] + O(\delta^i) \right\}. \quad (3.23)$$

We have the sums

$$\sum_{j=1}^{n-1} \sum_{i=1}^j O(\delta^j) = \sum_{j=1}^{n-1} O(j\delta^j) = O(1), \quad \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} \sum_{i=j+1}^{n-1} O(\delta^i) = \sum_{j=2}^{n-1} O(j\delta^j) = O(1)$$

and we have the sum

$$\sum_{j=1}^{\lfloor (n-1)/2 \rfloor} \sum_{i=j+1}^{n-1} \mathbb{E}_\nu[V_1 V_{1+j}] = \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} (n-j-1) \mathbb{E}_\nu[V_1 V_{1+j}],$$

so (3.23) becomes

$$B_n = 2 \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} (n-j-1) \mathbb{E}_\nu[V_1 V_{1+j}] + O(1). \quad (3.24)$$

In summary, (3.18), (3.19) and (3.24) give us the key relation

$$\frac{1}{n} \text{Var}[R_n(\rho)] = \mathbb{E}_\nu[V_1^2] + 2 \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} \left(1 - \frac{j+1}{n} \right) \mathbb{E}_\nu[V_1 V_{1+j}] + O(1/n), \quad (3.25)$$

and by Lemma 3.4 the summands are absolutely convergent, so we can take the limit in (3.25) to get

$$\lim_{n \rightarrow \infty} \frac{1}{n} \text{Var}[R_n(\rho)] = \mathbb{E}_\nu[V_1^2] + 2 \sum_{j=1}^{\infty} \mathbb{E}_\nu[V_1 V_{1+j}] \stackrel{\text{def}}{=} \sigma_\rho^2(F). \quad (3.26)$$

This completes the proof of the asymptotic relations for the mean and variance of $R_n(\rho)$. When these relations are coupled with the bound (3.11) on the Dobrushin coefficient, the central limit theorem part of Theorem 3 is almost automatic. Specifically, after one shows that the constant $\sigma_\rho^2(F)$ defined by (3.26) is strictly positive, then, without any further

work, we get (3.5) from Arlotto and Steele (2016), Theorem 1 Corollary 2.

Now, to work toward a lower bound for $\sigma_\rho^2(F)$, we let \mathcal{F}_e be the σ -field generated by the evenly indexed terms Y_0, Y_2, Y_4, \dots , and to facilitate calculations that are conditional on the “even σ -field” \mathcal{F}_e we write

$$R_{2n}(\rho) = \sum_{j=0}^{n-1} W_j \quad \text{where } W_j = \mathbb{1}(Y_{2j+1} \neq Y_{2j}) + \mathbb{1}(Y_{2j+2} \neq Y_{2j+1}).$$

We already know by (3.26) that $\text{Var}[R_n(\rho)] = \text{Var}_0[R_n(\rho)] \sim n\sigma_\rho^2(F)$, and we have also shown that $\text{Var}_0[R_n(\rho)] \sim \text{Var}_\nu[R_n(\rho)]$. Thus, to show $\sigma_\rho^2(F) > 0$, it suffices to show that there is a constant $\alpha > 0$ such that $\text{Var}_\nu[R_{2n}(\rho)] \geq n\alpha$ for all $n \geq 1$. We begin by studying the conditional variances of the individual summands of $R_{2n}(\rho)$.

For specificity, we should also note that for each j the distribution of W_j given \mathcal{F}_e does not depend on the initial distribution; accordingly we simply write $\text{Var}[W_j|\mathcal{F}_e]$ for the corresponding conditional variance. On the other hand, the distribution of the random variable $\text{Var}[W_j|\mathcal{F}_e]$ depends on the distribution of Y_0 , so, for its expectation when $Y_0 \sim \nu$, we need to write $\mathbb{E}_\nu[\text{Var}[W_j|\mathcal{F}_e]]$.

Lemma 3.5. *For all $\rho \in (0, 1)$ and $F \in \mathfrak{S}_L$, there exists a constant $\alpha_F(\rho) > 0$ for which one has*

$$\mathbb{E}_\nu[\text{Var}[W_j|\mathcal{F}_e]] = \mathbb{E}_\nu[\text{Var}[W_j|Y_{2j}, Y_{2j+2}]] \geq \alpha_F(\rho) \quad \text{for all } j \geq 0.$$

Proof. When we condition on $\mathcal{F}_e = \sigma\{Y_0, Y_2, \dots\}$, the distribution of W_j requires the consideration of two cases. First, if we have $Y_{2j} = Y_{2j+2}$, then with probability one we have $Y_{2j} = Y_{2j+1} = Y_{2j+2}$ and hence $W_j = 0$. Second, given \mathcal{F}_e with $Y_{2j} \neq Y_{2j+2}$, then we have

$$W_j = \begin{cases} 0, & \text{with probability } 0, \\ 1, & \text{with probability } F(\rho Y_{2j}), \\ 2, & \text{with probability } 1 - F(\rho Y_{2j}). \end{cases} \quad (3.27)$$

From the representation (3.27) and the strict monotonicity of $F \in \mathfrak{S}_L$, we see there is a

constant $C_F(\rho) > 0$ such that for all $j \geq 0$

$$\text{Var}[W_j | Y_{2j}, Y_{2j+2}] \geq C_F(\rho) \mathbb{1}[Y_{2j} \neq Y_{2j+2}, \rho L \leq Y_{2j}, Y_{2j+2} \leq L]. \quad (3.28)$$

If we set $Z = \mathbb{1}[Y_{2j} \neq Y_{2j+2}, \rho L \leq Y_{2j}, Y_{2j+2} \leq L]$ and

$$A = \{Y_{2j} \in [\rho L, L]\}, B = \{Y_{2j+1} \in [\rho L, L], Y_{2j+1} \neq Y_{2j}\}, C = \{Y_{2j+2} \in [\rho L, L]\}.$$

Then $Z \geq \mathbb{1}(A \cap B \cap C)$ and

$$\mathbb{E}_\nu[Z] \geq P_\nu(A \cap B \cap C) = P_\nu(A)P_\nu(B|A)P_\nu(C|A, B).$$

Each term on the right hand side is at least $1 - F(\rho L)$ because any upcoming observation that falls within $[\rho L, L]$ will be accepted. This gives us

$$\mathbb{E}_\nu[Z] \geq (1 - F(\rho L))^3,$$

so by (3.28) one can take $\alpha_F(\rho) \equiv C_F(\rho)(1 - F(\rho L))^3 > 0$ to complete the proof of the lemma. \square

This is last of the tools we need to get a non-trivial lower bound for $\sigma_\rho^2(F)$. By the law of total variance and by Lemma 3.5, we have

$$\begin{aligned} \text{Var}[R_{2n}(\rho)] &= \mathbb{E}[\text{Var}[R_{2n}(\rho) | \mathcal{F}_e]] + \text{Var}[\mathbb{E}[R_{2n}(\rho) | \mathcal{F}_e]] \\ &\geq \mathbb{E}[\text{Var}[R_{2n}(\rho) | \mathcal{F}_e]] = \mathbb{E} \left[\sum_{j=1}^n \text{Var}[W_j | \mathcal{F}_e] \right] \geq n\alpha_F(\rho) \end{aligned} \quad (3.29)$$

where the last equality is due to the independence between W_i and W_j given \mathcal{F}_e when $i \neq j$.

Finally, given Lemma 3.5 and our earlier observations, the proof of Theorem 3 is complete.

3.5. Proof of Theorem 4

Before we take up the proof of Theorem 4 in earnest, it will be useful to know that when F is the uniform distribution we can work with the density of the stationary distribution of $K_{\rho,F}$. To get the required absolute continuity we begin with a general inequality.

Proposition 3.6. *If $F \in \mathfrak{S}_L$ and if ν is the stationary measure for the transition kernel $K_{\rho,F}$ given by (3.9), then for all Borel $A \subset \mathcal{X}$ one has*

$$\nu(A) \leq \frac{1}{1 - F(\rho L)} \int_0^L \mathbb{1}(y \in A) F(dy). \quad (3.30)$$

Proof. Stationarity of ν and the definition of $K_{\rho,F}$ give us

$$\begin{aligned} \nu(A) &= \int_{\mathcal{X}} K_{\rho,F}(x, A) \nu(dx) \\ &= \int_{\mathcal{X}} \mathbb{1}(x \in A) F(\rho x) \nu(dx) + \int_0^L \int_{\mathcal{X}} \mathbb{1}(y \in A) \mathbb{1}(\rho x \leq y \leq L) \nu(dx) F(dy) \\ &\leq \nu(A) F(\rho L) + \int_0^L \mathbb{1}(y \in A) F(dy), \end{aligned}$$

from which we get (3.30). □

From (3.30) we see that ν is always absolutely continuous with respect to F . Consequently, if F is absolutely continuous with respect to Lebesgue measure dx , then both ν and F have densities with respect to dx .

Now we take F to be the uniform distribution on $[0, 1]$, and we simply write K_{ρ} , M_{ρ} and m_{ρ} for the corresponding transition kernel, stationary distribution function and density function. The definition of K_{ρ} and equation of stationarity now tell us

$$m_{\rho}(y) = \int_0^1 m_{\rho}(x) K_{\rho}(x, y) dx = \rho y m_{\rho}(y) + \int_0^1 m_{\rho}(x) \mathbb{1}(\rho x \leq y) dx,$$

or, in other words,

$$m_\rho(y) - \rho y m_\rho(y) = M_\rho(y/\rho) \quad \text{for all } y \in [0, 1]. \quad (3.31)$$

Perhaps the quickest way to extract what we need from this key identity is to first introduce the Mellin transform of $m(\cdot)$:

$$\phi(s, \rho) \stackrel{\text{def}}{=} \int_0^1 x^s m_\rho(x) dx.$$

From (3.31) and the fact that $M_\rho(x) = 1$ for $x \geq 1$ we then find

$$\phi(s, \rho) - \rho \phi(s+1, \rho) = \int_0^1 x^s M_\rho(x/\rho) dx = \int_0^\rho x^s M_\rho(x/\rho) dx + \frac{1 - \rho^{s+1}}{s+1}. \quad (3.32)$$

A change of variables and integration by parts give us

$$\int_0^\rho x^s M_\rho(x/\rho) dx = \rho^{s+1} \int_0^1 u^s M_\rho(u) du = \rho^{s+1} \frac{1 - \phi(s+1, \rho)}{s+1},$$

so (3.32) becomes

$$\phi(s, \rho) - \rho \phi(s+1, \rho) = \frac{1 - \rho^{s+1} \phi(s+1, \rho)}{s+1}, \quad (3.33)$$

which we can rewrite as a recursion,

$$\phi(s, \rho) = \frac{1}{1+s} + \left(\rho - \frac{\rho^{s+1}}{s+1} \right) \phi(s+1, \rho). \quad (3.34)$$

Proposition 3.7 (Mellin Transform of the Stationary Density). *We have*

$$\phi(s, \rho) = \sum_{k=0}^{\infty} a_k(s) \quad \text{where} \quad a_0(s) = \frac{1}{1+s} \quad \text{and} \quad (3.35)$$

$$a_k(s) = \frac{1}{s+k+1} \prod_{i=1}^k \left(\rho - \frac{\rho^{s+i}}{s+i} \right) \quad \text{for } k \geq 1.$$

Proof. We just need to check that (3.35) satisfies the recursion (3.34). In fact we have

$$\begin{aligned} \left(\rho - \frac{\rho^{s+1}}{s+1}\right) a_k(s+1) &= \frac{1}{s+k+2} \left(\rho - \frac{\rho^{s+1}}{s+1}\right) \prod_{i=1}^k \left(\rho - \frac{\rho^{s+1+i}}{s+1+i}\right) \\ &= \frac{1}{s+k+2} \prod_{i=1}^{k+1} \left(\rho - \frac{\rho^{s+i}}{s+i}\right) = a_{k+1}(s), \end{aligned}$$

so summing from $k = 0$ to ∞ gives us

$$\left(\rho - \frac{\rho^{s+1}}{s+1}\right) \phi(s+1, \rho) = \sum_{k=0}^{\infty} a_{k+1}(s) = \sum_{k=1}^{\infty} a_k(s).$$

Since $a_0(s) = 1/(1+s)$, we have proved

$$\left(\rho - \frac{\rho^{s+1}}{s+1}\right) \phi(s+1, \rho) = \sum_{k=1}^{\infty} a_k(s) = \phi(s, \rho) - \frac{1}{1+s},$$

giving us the required recursion (3.34). □

For the first moment of $m_\rho(\cdot)$ we therefore find

$$\int_0^1 x m_\rho(x) dx = \phi(1, \rho) = \frac{1}{2} + \frac{1}{3} \left(\rho - \frac{\rho^2}{2}\right) + \frac{1}{4} \left(\rho - \frac{\rho^2}{2}\right) \left(\rho - \frac{\rho^3}{3}\right) + \dots,$$

and this is just what we need to complete the calculation of $\mu_\rho(U)$. Specifically, if we specialize the general formula (3.16) for $\mu_\rho(F)$ to the uniform distribution function U , we get some substantial simplification. Specifically, we have

$$\begin{aligned} \mu_\rho(U) &= \mathbb{E}_\nu[\mathbb{1}(Y_0 \neq Y_1)] = \int_0^1 \int_0^1 K_\rho(x, y) \mathbb{1}(x \neq y) m_\rho(x) dx dy \\ &= \int_0^1 \int_0^1 \mathbb{1}(\rho x \leq y) m_\rho(x) dx dy = 1 - \rho \phi(1, \rho), \end{aligned}$$

and, together with the expansion for $\phi(1, \rho)$, this completes the proof of the first assertion (3.6) of Theorem 4.

We will see in the next section that (3.31) is from a class of equations with a rich theory.

Nevertheless, there are situations where one can make use of (3.31) without knowing its solution and without appealing to wider theory.

3.6. The Stationary Measure and the Pantograph Equation

The first-order non-autonomous *pantograph equation* for $\lambda \in (0, \infty)$ is the functional differential equation

$$H'(t) = a(t)H(t) + b(t)H(\lambda t) \quad t \geq 0. \quad (3.36)$$

The connection to the problems considered here is that for $0 < \rho < 1$ the equation (3.31) for the distribution function M_ρ of the stationary measure of the transition kernel $K_{\rho,U}(\cdot, \cdot) \equiv K_\rho(\cdot, \cdot)$ can be written as

$$M'_\rho(t) = \frac{1}{1 - \rho t} M_\rho(t/\rho) \quad \text{for } 0 \leq t < 1. \quad (3.37)$$

Thus, on the interval $[0, 1]$, the distribution function M_ρ satisfies the pantograph equation (3.36) with $a(t) = 0$, $b(t) = 1/(1 - \rho t)$, and $\lambda = 1/\rho > 1$.

The pantograph equation occurs in many contexts, perhaps the earliest of which was a number of theoretic investigations of Mahler (1940) that exploited the equation $H'(t) = bH(\lambda t)$, $H(0) = 1$ and its solution

$$H(t) = \sum_{j=0}^{\infty} \frac{1}{j!} \lambda^{j(j-1)/2} b^j t^j, \quad (3.38)$$

which is an elegant — and useful — generalization of the exponential function.

The two-term pantograph equation (3.36) has mostly commonly occurred in the autonomous case where $a(t)$ and $b(t)$ are constant, and the equation got its name from Fox et al. (1971) where the autonomous equation was used to model the collection of current by the pantograph (or flat pan connection head) of a tram. The subsequent investigation of Kato and McLeod (1971) showed the full richness of the equation, and, ever since, the pantograph equation has been regularly studied and applied, see e.g. Iserles (1993), Derfel and Iser-

les (1997), Guglielmi and Zennaro (2003), Saadatmandi and Dehghan (2009), Yusufoglu (2010), and Hsiao (2015), all of which contain many references.

In the non-autonomous case, essentially all work on (3.36) has been asymptotic or numerical. Moreover, all of the recent work focuses on the case when $\lambda \in (0, 1)$, and there is a sound scientific reason for this. Specifically, for (3.36) to be useful in an engineering or scientific context, it seems natural to assume that it is a *causal* equation; that is, the current rate of change $H'(t)$ is required to be determined by information that is available at time t .

A noteworthy feature of the stationarity equation (3.37) is that it is *not* a causal equation; one has $\lambda = 1/\rho > 1$. The other interesting feature of (3.37) is that it was essentially solved in Section 3.5, at least in the sense that Proposition 3.7 gives explicit series expansion of its Mellin transform.

Mellin transforms have rarely been used in the theory of the pantograph equation; we know of only one other case. Specifically, van Brunt and Wake (2011) used Mellin transforms to study a second order non-autonomous pantograph equation. Intriguingly, their equation was also acausal, and it also had a probabilistic origin. Specifically, it arose as the Fokker-Plank equation in a diffusion model for a population of cells, and the acausal parameter came from a splitting constant for cell division.

For the moment, we do not make further use the pantograph equation. Nevertheless, given the richness of the theory of the pantograph equation, the connection may prove fruitful over time. The benefits may even flow both ways. For example, calculations like those of Section 3.5 provide explicit Mellin transforms for the solutions of some other pantograph equations in addition to (3.37). Such explicit solutions seem worth pursuing, even though it would be a distraction for us to do so here.

3.7. When $\rho > 1$: The Proof of Theorem 5

We now consider an infinite sequence X_1, X_2, \dots of independent random variables with distribution $F \in \mathfrak{S}_L$. We then fix $\rho > 1$, and we again use the recursive definition (3.1) to specify the set of selection times $\{\tau_k : k = 1, 2, \dots\}$. If we then set

$$N_\rho = \min\{k : X_{\tau_k} \in (L/\rho, L]\} \text{ and } M_\rho = \min\{\tau_k : X_{\tau_k} \in (L/\rho, L]\}$$

then the number of selections one makes from $\{X_1, X_2, \dots, X_n\}$ is simply given by $R_n(\rho) = R_{\min(n, M_\rho)}(\rho)$, since after we have made a selection larger than L/ρ no further selections are possible. Also, for each $\omega \in \{\omega : M_\rho(\omega) < \infty\}$, we have

$$R_n(\rho) = R_{\min(n, M_\rho)}(\rho) \nearrow R_{M_\rho}(\rho) = N_\rho \quad \text{as } n \rightarrow \infty,$$

so the main task is to prove the moment generating function bound (3.7).

Since each value accepted by the selection process with $\rho > 1$ must be at least a factor of ρ greater than the preceding selection we have the bounds

$$N_\rho \leq \max\{k : \rho^{k-1} X_1 \leq L\} \leq 1 + \log L / \log \rho - \log X_1 / \log \rho,$$

so for the moment generating function we find

$$\mathbb{E}[\exp(sN_\rho)] \leq \exp(s)L^{s/\log \rho} \mathbb{E}[X_1^{-s/\log \rho}] = \exp(s)L^{s/\log \rho} \int_0^L x^{-s/\log \rho} dF.$$

We know the integral is finite when $F(x) = O(x)$ near 0 and $|s| < \log \rho$, and this gives us (3.7).

To show N_ρ is unbounded, we first fix an integer $M > 1$, and we consider the disjoint

subintervals $\{I_1, I_2, \dots, I_M\}$ of $[0, L]$ that are defined by setting

$$I_k = [a_k, b_k] = \left[\frac{(\rho - 1)L}{\rho^M - 1} \sum_{i=1}^{k-1} \rho^i, \frac{(\rho - 1)L}{\rho^M - 1} \sum_{i=0}^{k-1} \rho^i \right], \quad 1 \leq k \leq M.$$

The main feature here is that one has $a_{k+1}/b_k = \rho > 1$ for all $1 \leq k < M$. If we have $X_i \in I_i$ for $i = 1, 2, \dots, M$, then all of the observations X_1, X_2, \dots, X_M are selected, so we always have the inequality

$$\prod_{k=1}^M \mathbb{1}(X_k \in I_k) \leq \mathbb{1}(N_\rho \geq M).$$

Finally, by the independence of the variables X_k , $1 \leq k \leq M$ and the strict monotonicity of F , we see that the expectation of the product is strictly positive. This gives us $P(N_\rho \geq M) > 0$ for all $M \geq 1$. Since M was arbitrary, we see that N_ρ is unbounded, and the proof of the theorem is complete.

3.8. Complements to Classical Record Theory

Here we consider the calculation of the expected number of selections where we assume that there was a selection made at “time zero” that had value $x \in [0, 1]$. Formally, we modify the definition (3.2) by first setting $\tau_1 = \min\{j : X_j \geq \rho x\}$. Next, for $k \geq 2$ we define τ_k as before by setting $\tau_k = \min\{j : X_j \geq \rho X_{\tau_{k-1}}\}$, and finally we set

$$R_n^x(\rho) = \max\{k : \tau_k \leq n\}. \quad (3.39)$$

In this notation, Renyi’s classical formula for the expected number of records is

$$\mathbb{E}[R_n^0(1)] = \sum_{k=1}^n \frac{1}{k} \stackrel{\text{def}}{=} H_n, \quad (3.40)$$

and the main goal of this section is to generalize this result in two ways. The immediate goal is to show that

$$\mathbb{E}[R_n^x(1)] = H_n - \sum_{k=1}^n \frac{x^k}{k}, \quad (3.41)$$

and then in Theorem 6 we will get a closely related formula for $\mathbb{E}[R_n^x(\rho)]$.

We begin by using first step analysis to get a useful recursion for the quantities

$$g_{n,\rho}(x) \stackrel{\text{def}}{=} \mathbb{E}[R_n^x(\rho)] \quad \text{and} \quad g_n(x) \stackrel{\text{def}}{=} g_{n,1}(x).$$

Specifically, if we consider the first observation $y = X_1$, then X_1 is not accepted if $y \leq \rho x$, and this happens with probability ρx . On the other hand if $y = X_1 \in [\rho x, 1]$ we do accept X_1 , and accordingly we find the basic recurrence relation

$$g_{n+1,\rho}(x) = \rho x g_{n,\rho}(x) + \int_{\rho x}^1 [1 + g_{n,\rho}(y)] dy. \quad (3.42)$$

For general $\rho \in (0, 1)$, this equation offers considerable resistance; in essence, it is a linearized non-autonomous pantograph equation in integrated form. Nevertheless, one can use (3.42) to extract some interesting information, including refinements of some classical facts.

For example, if we take $\rho = 1$ in (3.42), then we can make some quick progress. Specifically, if we write $g_n(x)$ for $g_{n,1}(x)$ then differentiation and a nice cancellation give us

$$g'_{n+1}(x) = x g'_n(x) - 1. \quad (3.43)$$

We have $g_1(x) = 1 - x$, so $g'_1(x) = -1$ and repeated applications of (3.43) give us

$$g'_2(x) = -x - 1, \quad g'_3(x) = -x^2 - x - 1, \quad \text{and} \quad g'_4(x) = -x^3 - x^2 - x - 1.$$

In general, one has

$$g'_n(x) = -x^{n-1} - x^{n-2} - \dots - 1 = -\frac{1-x^n}{1-x}, \quad (3.44)$$

so integration over $[0, x]$ gives us

$$g_n(x) = g_n(0) - x - \frac{x^2}{2} - \dots - \frac{x^n}{n}. \quad (3.45)$$

Now if we use the basic recursion (3.42) with $x = 0$ and $\rho = 1$ we have from (3.45) that

$$\begin{aligned} g_{n+1}(0) &= 1 + \int_0^1 g_n(y) dy = g_n(0) + 1 - \frac{1}{1 \cdot 2} - \frac{1}{2 \cdot 3} - \dots - \frac{1}{n \cdot (n+1)} \\ &= g_n(0) + \frac{1}{n+1}. \end{aligned}$$

By telescoping we then recover Renyi's formula $g_n(0) = H_n$, but from (3.45) we now also find our refinement of Renyi's formula (and its approximation):

$$\mathbb{E}[R_n^x(1)] = H_n - \sum_{k=1}^n \frac{x^k}{k} = \log n - \sum_{k=1}^n \frac{x^k}{k} + \gamma + \frac{1}{2n} + O(1/n^2), \quad (3.46)$$

where $\gamma = 0.577 \dots$ is Euler's constant.

For any $0 < \rho < 1$, one can derive a representation of $\mathbb{E}[R_n^x(\rho)]$ that is only a little less explicit than (3.41). The correcting term is again a truncated power series, but in this case principal term $g_{n,\rho}(0)$ is no longer a well-known quantity.

Theorem 6. *For all $0 < \rho \leq 1$ and $0 \leq x \leq 1$ we have*

$$g_{n,\rho}(x) = g_{n,\rho}(0) - \sum_{i=1}^n a_i x^i, \quad (3.47)$$

where $a_1 = \rho$, $a_2 = \rho(\rho - \rho^2/2)$, and $a_i = (\rho - \rho^i/i)a_{i-1}$ for all $i \geq 2$.

Proof. To argue by induction, we first recall that $g_{1,\rho}(x) = \rho x$ for all $0 \leq x \leq 1$, and this

gives us by direct evaluation that (3.47) holds for $n = 1$. Next, from the basic recursion (3.42) we have

$$g_{n+1,\rho}(0) = \int_0^1 [1 + g_{n,\rho}(y)] dy \quad \text{and} \quad g_{n+1,\rho}(x) = \rho x g_{n,\rho}(x) + \int_{\rho x}^1 [1 + g_{n,\rho}(y)] dy,$$

so taking the difference gives us

$$g_{k+1,\rho}(0) - g_{k+1,\rho}(x) = \rho x + \int_0^{\rho x} [g_{n,\rho}(y) - g_{n,\rho}(x)] dy. \quad (3.48)$$

By the induction hypothesis we can expand the last integrand as

$$g_{n,\rho}(y) - g_{n,\rho}(x) = \sum_{i=1}^n a_i (x^i - y^i), \quad (3.49)$$

so from (3.48) and the defining relation $a_{i+1} = (\rho - \rho^{i+1}/(i+1))a_i$ we have

$$\int_0^{\rho x} \sum_{i=1}^n a_i (x^i - y^i) dy = \sum_{i=1}^n a_i \left(\rho - \frac{\rho^{i+1}}{i+1} \right) x^{i+1} = \sum_{i=1}^n a_{i+1} x^{i+1}. \quad (3.50)$$

Finally, from (3.48) and (3.50) we then get

$$g_{n+1,\rho}(0) - g_{n+1,\rho}(x) = \sum_{i=1}^{n+1} a_i x^i,$$

which completes the induction step. □

Since $0 < a_i \leq \rho^i$, the identity (3.47) has an immediate corollary that underscores an informative difference between the case when $\rho \in (0, 1)$ and the case $\rho = 1$. Specifically, for $\rho = 1$ we see from (3.41) that the influence of x is unbounded, while the next corollary tells us that for $0 < \rho < 1$ the influence of the initial value x has only a *bounded influence*.

Corollary 3.8 (Insensitivity of the Initial Constraint). *For all $\rho \in (0, 1)$, $n \geq 0$, and all $0 \leq x \leq y \leq 1$, one has*

$$0 \leq g_{n,\rho}(x) - g_{n,\rho}(y) \leq \frac{\rho}{1 - \rho}. \quad (3.51)$$

The bounds (3.51) suggest that we should take limits, and from the geometric convergence in (3.49), we can define a continuous anti-symmetric function $B : [0, 1]^2 \rightarrow \mathbb{R}$ by setting

$$\sum_{i=1}^{\infty} a_i(y^i - x^i) = \lim_{n \rightarrow \infty} \{g_{n,\rho}(x) - g_{n,\rho}(y)\} \stackrel{\text{def}}{=} B(x, y). \quad (3.52)$$

A useful feature of this function is that it leads to an alternative characterization of $\mu_\rho(U)$, and it gives second proof of its series representation (3.6).

To derive the characterization, we subtract $g_{n,\rho}(x)$ from both sides of the basic recursion (3.42), and we simplify to get the identity

$$g_{n+1,\rho}(x) - g_{n,\rho}(x) = (1 - \rho x) + \int_{\rho x}^1 \{g_{n,\rho}(y) - g_{n,\rho}(x)\} dy.$$

Now, if we set $x = 0$ in the defining relation (3.52) and apply antisymmetry of $B(\cdot, \cdot)$, then we see that as $n \rightarrow \infty$ one has

$$g_{n+1,\rho}(0) - g_{n,\rho}(0) = 1 + \int_0^1 B_\rho(y, 0) dy + o(1) = 1 - \int_0^1 B_\rho(0, y) dy + o(1).$$

We now sum over $n \in [0 : N]$. By telescoping, division by $N + 1$, and taking limits we get a new formula for the mean $\mu_\rho(U)$ given by Theorem 4:

$$\mu_\rho(U) = 1 - \int_0^1 B_\rho(0, y) dy. \quad (3.53)$$

Finally, if we substitute the series expansion (3.52) for $B_\rho(0, y)$ into (3.53), we see that term-by-term integration that (3.53) gives us a second derivation of the original formula (3.6) for $\mu_\rho(U)$. In a sense, this integration also explains the presence of the harmonic factors $1/2, 1/3, 1/4, \dots$ in (3.6).

3.9. More Records: Relaxed or Constrained

For any $\rho \in (0, \infty)$ and any $F \in \mathfrak{S}_L$, we can consider the set of selected values $\mathcal{A}(\rho) = \{X_{\tau_1}, X_{\tau_2}, \dots\}$; these are formally defined by the stopping time recursion (3.1). The set $\mathcal{A}(1)$ is exactly the set of record values, and more generally we have the relations

$$\rho \in (0, 1) \Rightarrow \mathcal{A}(1) \subset \mathcal{A}(\rho) \quad \text{and} \quad \rho \in (1, \infty) \Rightarrow \mathcal{A}(\rho) \subset \mathcal{A}(1), \quad (3.54)$$

which give us a more explicit sense in which $\rho \in (0, 1)$ *relaxes* the record condition and $\rho \in (1, \infty)$ further *constrains* the record condition.

The first relation of (3.54) is obvious since whenever X_k is record, then we have $X_k \geq X_{\tau_i} \geq \rho X_{\tau_i}$ for all $\tau_i < k$. It is rather less obvious that for $\rho > 1$ one has the complementary relation $\mathcal{A}(\rho) \subset \mathcal{A}(1)$. To prove this by induction, we first note that $X_{\tau_1} = X_1 \in \mathcal{A}(\rho)$, and by definition X_1 is a record. Now we suppose by induction that the first n elements $\{X_{\tau_1}, X_{\tau_2}, \dots, X_{\tau_n}\}$ of $\mathcal{A}(\rho)$ are also all records.

There are two cases to consider. First, if $\tau_{n+1} < \infty$ and $X_{\tau_n} = x$, then we have $\tau_{n+1} = \min\{k : X_k \geq \rho x\}$. This tells us that τ_{n+1} is the first entrance time of the process X_1, X_2, \dots into the interval $[\rho x, L]$. Since all such first entrance times are also record times, we see that $X_{\tau_{n+1}}$ is a record, and induction gives us $\mathcal{A}(\rho) \subset \mathcal{A}(1)$. In the second case we have $\tau_{n+1} = \infty$, and $\mathcal{A}(\rho) = \{X_{\tau_1}, X_{\tau_2}, \dots, X_{\tau_n}\}$. We already have from our induction hypothesis that $\{X_{\tau_1}, X_{\tau_2}, \dots, X_{\tau_n}\} \subset \mathcal{A}(1)$, so again we get $\mathcal{A}(\rho) \subset \mathcal{A}(1)$.

Despite the first relation of (3.54), it is generally inappropriate to think of the values $\mathcal{A}(\rho) = \{X_{\tau_j} : j = 1, 2, \dots\}$ are anything like “approximate records” when $0 < \rho < 1$. To make this distinction explicit, fix $0 < \epsilon < 1$ and consider the events

$$A_k = \{X_k \text{ is selected and } X_k \leq \epsilon \max\{X_i : i \leq k\}\}, \quad (3.55)$$

where the random variables X_i , $i = 1, 2, \dots$ are independent and uniformly distributed on

$[0, 1]$.

When A_k occurs, the selected value X_k is only a small fraction of the current maximum, so it is not an approximate record (or a near-record) in any reasonable sense. Nevertheless, with probability one, infinitely many of the events A_1, A_2, \dots will occur, so infinitely often the selected values are quite unlike records.

To see this, we first note that for any $\epsilon > 0$ both of the sets $[0, \epsilon/2]$ and $[1/2, 1]$ have positive probability under the stationary measure ν for the associated Markov chain $\{Y_n : n = 1, 2, \dots\}$ of Section 3.2. Thus, they are also both recurrent sets for the chain. Now, if at time k the chain enters $[0, \epsilon/2]$ after having entered $[1/2, 1]$ at some time previous to k , then the event A_k also occurs. The positive recurrence of the respective sets then tells us that infinitely many of the events $\{A_k : k = 1, 2, \dots\}$ will occur with probability one.

This construction shows that there is a disconnection between the theory of the selection process with $0 < \rho < 1$ and the theory of the near records such as studied in Balakrishnan et al. (2005), Gouet et al. (2007) or Gouet et al. (2012), but this construction does not tell the whole story. In Section 3.8 we saw several instances where the technology of selection processes could inform us about the classical record process. Still, it is reasonable to expect that one has at least some analogous carry forward to the theory of near-records, but here we cannot pursue that point except to acknowledge the possibility.

4.1. The Low Rank Matrix Recovery Problem

Low-rank structure has been studied extensively in a wide range of applications including image compression (Andrews and Patterson III, 1976), subspace segmentation (Liu et al., 2013), face recognition (Basri and Jacobs, 2003; Candès et al., 2011), Netflix problems (Candès and Recht, 2009) and quantum state tomography (Gross et al., 2010; Liu, 2011). In this chapter, we focus on the low-rank matrix recovery problem, which refers to finding an unknown matrix X of dimension $m \times n$ with rank $r \ll \min\{m, n\}$ based on l linear measurements $\langle A_1, X \rangle, \dots, \langle A_l, X \rangle$. Here $\{A_1, \dots, A_l\}$ is called the set of measurement matrices and the inner product is defined as $\langle X, Y \rangle = \sum_{i,j} X_{ij} Y_{ij} = \text{Tr}(X^\top Y)$. When the matrix is constrained to be diagonal, the problem is reduced to sparse vector recovery and usually termed as Compressed Sensing or Compressed Sampling (Donoho, 2006; Candès et al., 2006; Candès et al., 2006). Intuitively, a rank- r matrix of dimension $m \times n$ has $(m + n - r)r$ degrees of freedom, which implies that $l = O((m + n)r)$ is the minimal number of measurements needed for exact recovery. The optimal rate and exact recovery were achieved with overwhelming probability by using Gaussian ensemble or whatever matrix that satisfies the Restricted Isometry Property (RIP) as the measurement matrix and solving a convex optimization problem (Recht et al., 2010; Candès and Plan, 2011) as follows

$$\begin{aligned} \min \quad & \|M\|_*, \\ \text{subject to} \quad & \mathcal{A}(M) = y, \end{aligned} \tag{4.1}$$

where $\|\cdot\|_*$ denotes the nuclear norm, the sum of all the singular values, $\mathcal{A}(M) = (\text{Tr}(A_1^\top M), \dots, \text{Tr}(A_l^\top M))$ and $y = (\text{Tr}(A_1^\top X), \dots, \text{Tr}(A_l^\top X))$. In practice, X is usually only approximately low-rank. Under this situation, the goal becomes finding the rank- r matrix which best approximates X using $O((m + n)r)$ measurements. Furthermore, it is realistic to consider noisy measurements $y = (\text{Tr}(A_1^\top X) + z_1, \dots, \text{Tr}(A_l^\top X) + z_l)$ with z

being the noise vector (Candes and Plan, 2011). The problem can be solved by the so-called matrix Dantzig Selector,

$$\begin{aligned} \min \quad & \|M\|_* \\ \text{subject to} \quad & \mathcal{A}(M) - y \in \mathcal{Z}, \end{aligned} \tag{4.2}$$

where \mathcal{Z} is a region determined by the noise structure.

However, the Gaussian ensemble design suffers from the storage problem. Specifically, it requires $O(mnl)$ bytes of space to store the measurement matrices (A_1, \dots, A_l) . As pointed out by Cai and Zhang (2013), accurate reconstruction of $10,000 \times 10,000$ matrix of rank 10 requires 45 TB storage. To address this problem, they put forward the Rank One Projection approach which replaced the Gaussian random matrices A_i with rank one matrices $\beta_i \gamma_i^\top$, where β_i, γ_i are independent Gaussian random vectors. This reduces the required storage to $O((m+n)l)$, which is significantly smaller. It is worth mentioning that RIP does not hold for rank one matrices $\beta_i \gamma_i^\top$. Furthermore, a semi-definite programming (SDP) is required as the final recovery step for both choices of measurement matrices, which is not scalable. Standard SDP solver for nuclear norm minimization is only efficient for matrices of size up to dozens by dozens. It becomes intractable for larger matrices. So far, several fast algorithms have been put forward (Lee and Bresler, 2010; Fornasier et al., 2011; Goldfarb and Ma, 2011; Mohan and Fazel, 2012). However, Fornasier et al. (2011); Mohan and Fazel (2012); Goldfarb and Ma (2011) are only applicable to noiseless measurements. Besides, all the algorithms rely on RIP of the measurement matrices. Therefore, $O(mnl)$ bytes of storage is inevitable.

Our contributions are two fold. First, instead of following the mainstream research based on choosing measurement matrices with RIP followed by solving the nuclear norm minimization problem, we put forward a two-step sensing scheme that uses minimal number of measurements $O((m+n)r)$ and far less storage $O(nr)$, and has lower computational complexity $O(mnr)$. Our method involves only QR decomposition and matrix multiplication. Second, besides exact recovery for the ideal low-rank and noiseless case, the proposed

sensing scheme is applicable when the underlying matrix is only approximately low-rank or when the measurements suffer from Gaussian noise. We provide theoretical guarantees under various settings. Error bounds for both spectral norm and Frobenius norm are established, while for most of the literature only Frobenius norm is discussed (Lee and Bresler, 2010; Candes and Plan, 2011; Cai and Zhang, 2013).

The rest of this chapter is organized as follows. In Section 4.2, we introduce the proposed sensing scheme and compare it with Gaussian ensemble design and Rank One Projection. In Section 4.3, we provide error bounds for our sensing scheme. We present simulation and real data analysis results in Section 4.4. All the details of proofs are put in Section 4.5.

Notations: Throughout this chapter, we use $\|\cdot\|_*$, $\|\cdot\|_2$, $\|\cdot\|_F$ for nuclear norm, spectral norm and Frobenius norm respectively. Let $\text{vec}(X)$ be the long vector obtained by stacking the columns of matrix X . Without loss of generality, we assume $m \geq n$ and as usual, let the singular value decomposition of $X \in \mathbb{R}^{m \times n}$ be $U\Sigma V^\top$, where $U \in \mathbb{R}^{m \times n}$, $\Sigma \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{n \times n}$. Also let $X_r = U\Sigma_r V^\top$ denote the best rank- r approximation of X .

4.2. Sensing Scheme

Let $X = U\Sigma V^\top$ be the singular value decomposition (SVD) of X and suppose $\text{rank}(X) = r$. The proposed sensing scheme is based on the observation that $UU^\top X = X$. Actually, if the columns of Q constitute an orthonormal basis of the range space of X , then $QQ^\top X = X$. Let $Q = (Q_1, \dots, Q_r)$. Notice that $(Q^\top X)_{ij} = Q_i^\top X e_j = \text{Tr}(Q_i^\top X e_j) = \text{Tr}(e_j Q_i^\top X) = \langle Q_i e_j^\top, X \rangle$. Therefore, if we can sense the range space of X , we can recover X using measurements $\{Q_i e_j^\top, 1 \leq i \leq r, 1 \leq j \leq n\}$. It turns out that the range space can be approximated by sensing $X\Omega$, where Ω is a random matrix. This idea is widely used in the literature of randomized SVD algorithms for huge matrices (Sarlos, 2006; Woolfe et al., 2008; Halko et al., 2011), where random projection is first used to identify the subspace that contains most information of the low rank matrix. Fazel et al. (2008) put forward two schemes based on sensing the row and column space, which has the same idea as ours.

However, their idea is carried out in a different way and the error bounds in the paper are rough. Also, there is no theoretical analysis is given for the noisy case.

4.2.1. Algorithm

In practice, the rank of X is usually unknown. Let the target rank be r , that is to say, we aim to find the matrix with rank- r that best approximates X using minimal number of measurements $O((m+n)r)$. To be specific, we are trying to find $X_r = U\Sigma_r V^\top$. The details of the proposed algorithm is described in Algorithm 1.

Algorithm 1 Computationally Efficient Low Rank Matrix Recovery

- 1: **Target Rank:** r
 - 2: **Sampling Parameter:** $k = r + p$, p is the oversampling parameter
 - 3: **Preparation:** Generate standard Gaussian matrix Ω with size $n \times k$
 - 4: **Step 1** Sense the range space using measurements $\mathcal{A}_1 = \{e_i \Omega_j, 1 \leq i \leq m, 1 \leq j \leq k\}$:
 $S_1 = X\Omega$
 - 5: **Intermediate Processing Step:** QR decomposition: $Q \leftarrow S_1 = QR$
 - 6: **Step 2** Sense QX using measurements $\mathcal{A}_2 = \{Q_i e_j^\top, 1 \leq i \leq r, 1 \leq j \leq n\}$: $S_2 = Q^\top X$
 - 7: **Recovery** $\hat{X} = QS_2$
-

The oversampling parameter p is beneficial and necessary. As we will see in the Theorems below, the failure probabilities decrease exponentially with p . In practice, setting $p = 5$ or 10 is adequate. The total number of measurements for the proposed scheme is $(m+n)k$. If we maintain $p = O(r)$, the number of measurements used is $O((m+n)r)$, which is rate optimal.

To match the terms in model (4.1), we have:

- Measurements: $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$,
- Observation: $y = (y_1, y_2)$, where $y_1 = \mathcal{A}_1(X) = \text{vec}(S_1)$ and $y_2 = \mathcal{A}_2(X) = \text{vec}(S_2)$,
- Recovery Step: $\hat{X} = QS_2$, where Q is obtained from the QR decomposition of S_1 .

Different from previous recovery models where the measurement matrices are independent, the proposed sensing scheme adopts a two-step measurement procedure and \mathcal{A}_2 is completely

	Number of Measurements	Storage	Recovery Complexity
Proposed Scheme	$O((m+n)r)$	$O(nr)$	$O(mnr)$
Gaussian ensemble	$O((m+n)r)$	$O(mn(m+n)r)$	SDP
Rank One Projection	$O((m+n)r)$	$O((m+n)^2r)$	SDP

Table 1: Comparison Between Three Sensing Schemes

determined by $\mathcal{A}_1(X)$ through a QR decomposition. Notice that S_1 is of dimension $n \times k$. Since $k = O(r) \ll \min\{m, n\}$, the QR decomposition only has complexity $O(nr^2)$, which is efficient. The recovery step just involves matrix multiplication. It has computational complexity $O(mnr)$ and is much faster than nuclear norm minimization and existing fast algorithms. Also notice that we do not need to store the measurement matrices \mathcal{A}_1 and \mathcal{A}_2 directly because \mathcal{A}_1 is determined by Ω and \mathcal{A}_2 is determined by $\mathcal{A}_1(X)$. Therefore, we only need to store Ω and the storage complexity is simply $O(nr)$, significantly smaller than any other schemes.

4.3. Theoretical Guarantee of Recovery

4.3.1. Noiseless Case

Theorem 7. *If $\text{rank}(X) \leq k$, with probability 1, our algorithm recovers X exactly.*

Proof. If $\text{rank}(X) \leq k$, then with probability 1, $\text{rank}(X\Omega) = r$, which implies the column space of $A\Omega$ is exactly the range space of X . Let $X = U\Sigma V^\top$ be the singular value decomposition. Then there exists orthonormal matrix P such that $Q = UP$. Therefore, $\hat{X} = QQ^\top X = UPP^\top U^\top U\Sigma V^\top = X$. \square

For the case when X is not exactly low rank. The approximation error $\|X - \hat{X}\| = \|X - QQ^\top X\|$ is well studied in the literature of randomized SVD algorithms. We cite the error bounds from Halko et al. (2011).

Theorem 8 (Theorem 10.5 and 10.7 of Halko et al. (2011)). *Let $k = r + p$, p is the degree of oversampling. If $p \geq 2$, then*

$$\mathbb{E} \|\hat{X} - X\|_F \leq \left(1 + \frac{r}{p-1}\right)^{\frac{1}{2}} \left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}}.$$

If $p \geq 4$, for all $u, t \geq 1$,

$$\|\hat{X} - X\|_F \leq \left(1 + t\sqrt{12r/p}\right) \left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}} + ut \frac{e\sqrt{r+p}}{p+1} \sigma_{r+1}$$

with probability at least $1 - 5t^{-p} - 2e^{-u^2/2}$.

Theorem 9 (Theorem 10.6 and 10.8 of Halko et al. (2011)). *Let $k = r + p$, p is the degree of oversampling. If $p \geq 2$, then*

$$\mathbb{E} \|\hat{X} - X\|_2 \leq \left(1 + \sqrt{\frac{r}{p-1}}\right) \sigma_{r+1} + \frac{e\sqrt{r+p}}{p} \left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}}.$$

If $p \geq 4$, for all $u, t \geq 1$,

$$\|\hat{X} - X\|_2 \leq \left(1 + t\sqrt{\frac{12r}{p}} + ut \frac{e\sqrt{r+p}}{p+1}\right) \sigma_{r+1} + t \frac{e\sqrt{r+p}}{p+1} \left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}}$$

with probability at least $1 - 5t^{-p} - e^{-u^2/2}$.

For Frobenius norm, the oracle risk is $\left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}}$ which is obtained by the best rank r approximation X_r . Therefore, if we choose the oversampling parameter $p = O(r)$, the risk of the proposed procedure is within constant factor of the oracle risk. For spectral norm, it suffers an extra term $\left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}}$, which becomes negligible if the singular values have a quick decay.

4.3.2. Gaussian Noise Case

In this section, we assume the measurements suffer from Gaussian noise, that is to say

$$S_1 = X\Omega + E_1 = QR,$$

$$S_2 = Q^\top X + E_2,$$

where E_1 and E_2 are independent Gaussian random matrices with variance σ^2 . With noisy measurement results S_1 and S_2 , we recover the unknown matrix X by

$$\hat{X} = QS_2 = QQ^\top X + QE_2.$$

Compared with the noiseless case, now there are two extra sources of errors. One is involved with sensing the range space in the sense that Q is computed from a noisy observation of $X\Omega$. The other comes from the additive part in the recovery step, QE_2 . We are going to establish expectation and concentration error bounds for both the Frobenius norm and the spectral norm of the residual matrix.

Remark 4.1. In this additive noise model, it is worth noting that increasing the signal of the measurement matrices is equivalent to decreasing the noise variance. Recall that $\mathcal{A}_1 = \{e_i\Omega_j, 1 \leq i \leq m, 1 \leq j \leq k\}$. If replacing measurement matrices $e_i\Omega_j^\top$ with $Ce_i\Omega_j^\top$ where C is a positive constant, the observation becomes $Ce_i^\top X\Omega_j + z = C(e_i^\top X\Omega_j + z/C)$, which is equivalent to using measurement matrices $e_i\Omega_j^\top$ and additive noise z' with variance σ^2/C^2 . However, in practice, the noise of the measurements should be proportional to the signal of the measurement matrices. Therefore, for the purpose of analysis, it is valid and beneficial to fix both the signal and noise at a certain level. The error bounds below are derived based on rescaling the measurement matrices such that the expectation of the Frobenius norm is equal to mn .

Theorem 10. Let $k = r + p$ where p is the degree of oversampling. If $p \geq 2$, then

$$\mathbb{E} \|\hat{X} - X\|_F \leq \sqrt{1 + \frac{r}{p-1}} \left(\sqrt{\sum_{j \geq r+1} \sigma_j^2} + \sigma \right). \quad (4.3)$$

If $p \geq 4$, for all $u, t \geq 1$,

$$\|\hat{X} - X\|_F \leq \left(1 + \sqrt{\frac{12r}{p}}t\right) \sqrt{\sum_{j \geq r+1} \sigma_j^2} + ut \frac{e\sqrt{r+p}}{p+1} (\sigma_{r+1} + \sigma) + \sqrt{\frac{12r}{p}}t\sigma \quad (4.4)$$

with probability at least $1 - 5t^{-p} - 3e^{-u^2/2}$.

Theorem 11. Let $k = r + p$ where p is the degree of oversampling. If $p \geq 2$, then

$$\mathbb{E} \|\hat{X} - X\|_2 \leq \left(1 + \sqrt{\frac{r}{p-1}}\right) \sigma_{r+1} + \frac{e\sqrt{r+p}}{p} \left(\sqrt{\sum_{j \geq r+1} \sigma_j^2} + 3\sigma \right). \quad (4.5)$$

If $p \geq 4$, for all $u, t \geq 1$,

$$\|\hat{X} - X\|_2 \leq \left(1 + \sqrt{\frac{12r}{p}}t + \frac{e\sqrt{r+p}}{p+1}ut\right) \sigma_{r+1} + \frac{e\sqrt{r+p}}{p+1}t \sqrt{\sum_{j \geq r+1} \sigma_j^2} + 4\frac{e\sqrt{r+p}}{p+1}ut\sigma \quad (4.6)$$

with probability at least $1 - 5t^{-p} - 3e^{-u^2/2}$.

The error bounds derived above are tight in the sense that only extra noise terms are added to the original noiseless bounds which are within a constant factor of oracle risk. Also, different from most previous literature only focusing on Frobenius norms, we establish error bounds for spectral norms as well.

4.4. Experiments

So far, we have shown that the proposed procedure is much better in terms of required storage space and computational complexity. From theoretical point of view, the error bounds of the proposed procedure, Gaussian ensemble and Rank One Projection are comparable with each other and nearly of the same order. In this section, we carry out numerical ex-

periments to compare the error of recovery of the proposed scheme and Gaussian ensemble design.

4.4.1. Simulation Results

Gaussian ensemble design with nuclear norm minimization can be solved by the following semi-definite programming

$$\begin{aligned} & \min \text{Tr}(B_1) + \text{Tr}(B_2), \\ \text{subject to } & \begin{pmatrix} B_1 & X \\ X^\top & B_2 \end{pmatrix} \succeq 0 \text{ and } \|\mathcal{A}^*(y - \mathcal{A}(X))\|_2 \leq \lambda, \end{aligned} \tag{4.7}$$

where B_1, B_2, X are to be optimized and λ is a constant depending on the noise level. We use CVX package (Grant and Boyd, 2008, 2014) to solve all the involved SDP. Specifically, we investigate four scenarios: 1. low-rank matrix and noiseless measurements; 2. low-rank matrix and measurements with Gaussian noise; 3. full-rank matrix with decaying singular values and noiseless measurements; 4. full-rank matrix with decaying singular values and measurements with Gaussian noise. A summary of the results is presented in Figure 1.

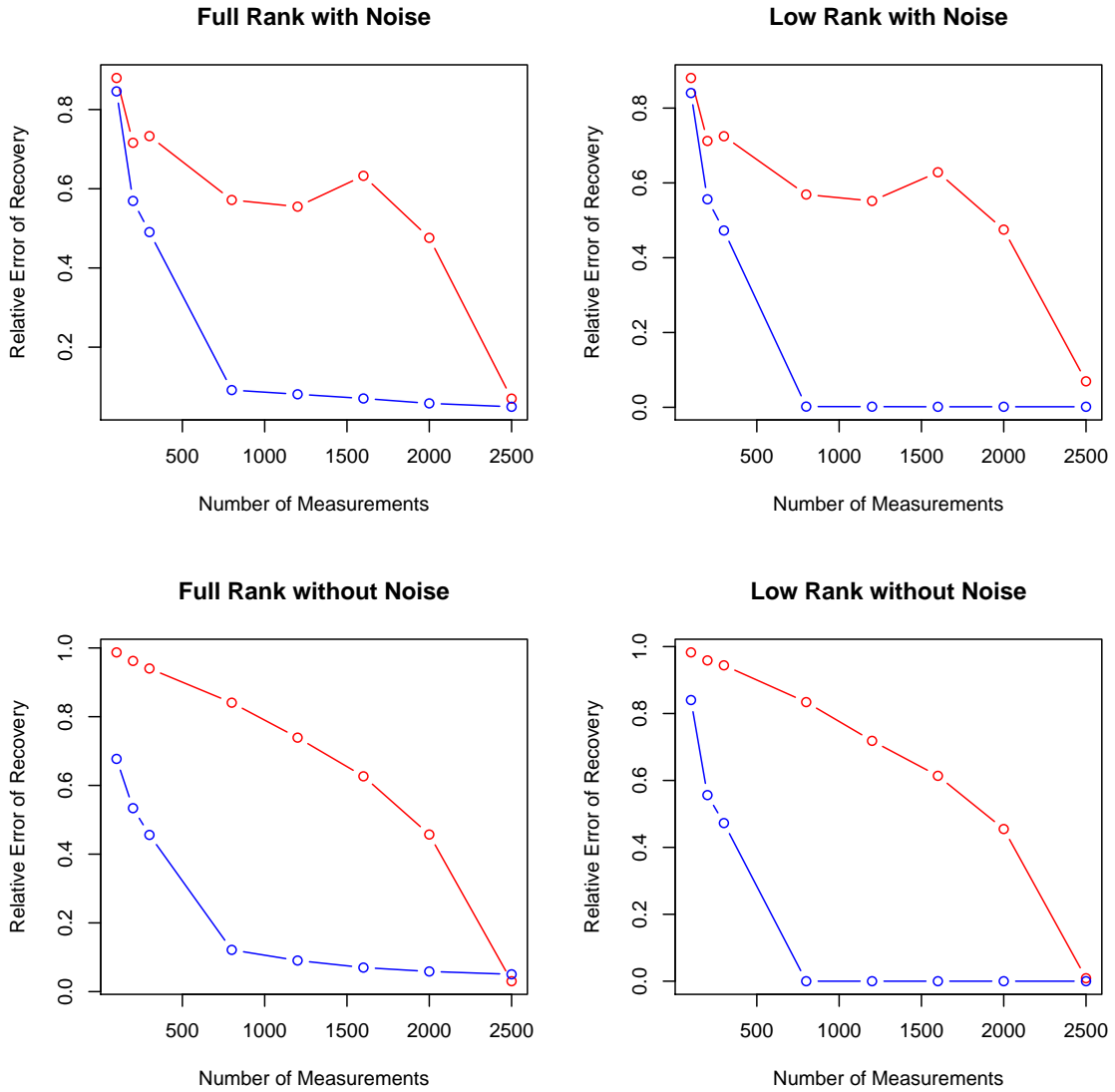


Figure 1: Comparison Between Proposed Scheme and Gaussian Ensemble

Figure 1 shows relative error $\|\hat{X} - X\|_F / \|X\|_F$ for the proposed scheme (blue line) and Gaussian ensemble design with nuclear norm minimization (red line). We use a 50×50 matrix with rank 4. For full rank case, we add small rank perturbations and for noisy measurements, we add Gaussian noise such that the signal-to-noise ratio is equal to 2.

As we can see from the plots, the proposed sensing scheme nearly dominates the Gaussian ensemble design with nuclear norm minimization. It is reasonable because the proposed

scheme essentially mimics the performance of SVD. Empirical study has shown that methods based on nuclear norm minimization have a phase transition curve, that is to say, recovery typically fails when the number of measurements is smaller than a certain threshold and succeeds otherwise (Donoho et al., 2013), which explains why the red curve drops more slowly at first. For our proposed sensing scheme, the error decreases most at the beginning because the scheme first captures the leading singular vector space.

4.4.2. Image Compression

Images are natural matrices and one approach for image compression is to use its low rank approximation. Here, we test the proposed sensing scheme through image recovery. We choose the famous image of Lena, which is a 512×512 grayscale image of full rank with decaying singular values.

From the results in Figure 2, we can see that the image is nearly the same as the original one when the sampling parameter is set at 120.

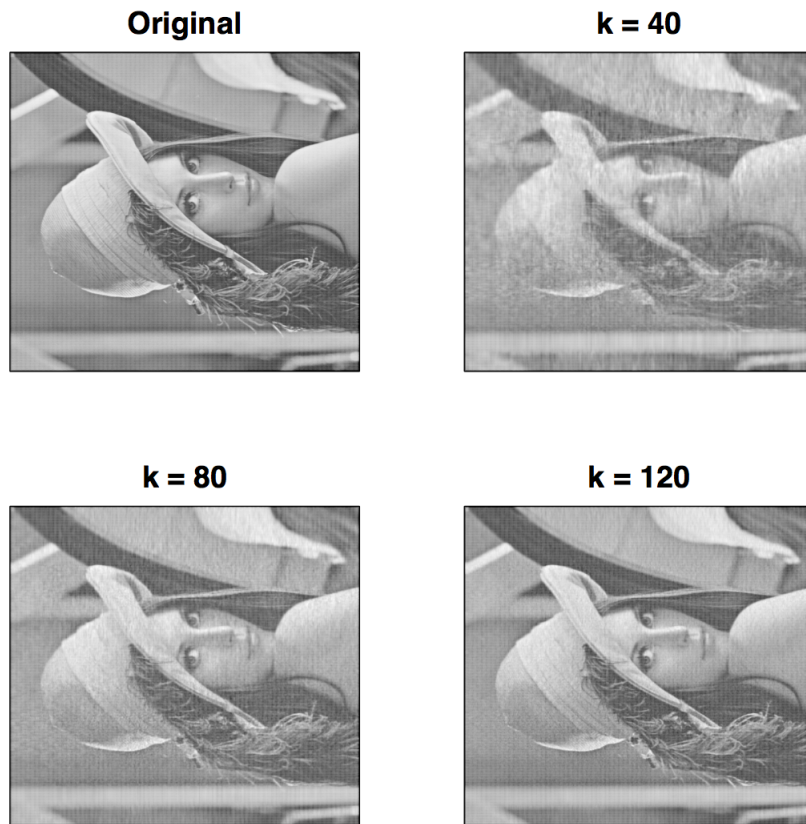


Figure 2: Recovery of Lena with Different Sampling Parameters

4.5. Details of Proof

Notations: We use G^\dagger for the pseudoinverse of matrix G , $\|\cdot\|_F$ for Frobenius norm, $\|\cdot\|_2$ for spectral norm. When we use $\|\cdot\|$, it means the statement is valid for both Frobenius norm and Spectral norm. $X =_d Y$ means that random variables X and Y have the same distribution and $X \leq_d Y$ means $P(X \geq t) \leq P(Y \geq t), \forall t \in \mathbb{R}$.

4.5.1. Technical Lemmas

We first present several lemmas. They are all from the extensive survey paper Halko et al. (2011).

On Matrix Algebra

Lemma 4.2. *Suppose $M \geq 0$, then*

$$I - (I + M)^{-1} \leq M.$$

Lemma 4.3. *Suppose $M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \geq 0$, then $\|M\|_2 \leq \|A\|_2 + \|C\|_2$.*

On Orthogonal Projectors

Lemma 4.4. *Suppose U is unitary, then $U^\top P_M U = P_{U^\top M}$.*

Lemma 4.5. *Suppose $\text{range}(N) \subset \text{range}(M)$, then for each matrix X , one has $\|P_N X\| \leq \|P_M X\|$ and $\|(I - P_M)X\| \leq \|(I - P_N)X\|$.*

On Gaussian Matrices

Lemma 4.6. *For real matrices S, T , and a standard Gaussian random matrix G , one has*

$$(\mathbb{E} \|SGT\|_F^2)^{\frac{1}{2}} = \|S\|_F \|T\|_F.$$

Lemma 4.7. For real matrices S, T , and a standard Gaussian random matrix G , one has

$$\mathbb{E} \|SGT\| \leq \|S\|_2 \|T\|_F + \|S\|_F \|T\|_2.$$

Lemma 4.8 (Frobenius Norm of Pseudoinverse). Let G be an $m \times n$ standard Gaussian matrix with $n - m \geq 2$, then

$$\mathbb{E} \|G^\dagger\|_F^2 = \frac{m}{n - m - 1}.$$

Lemma 4.9 (Spectral Norm of Pseudoinverse). Let G be an $m \times n$ standard Gaussian matrix with $n - m \geq 1$ and $m \geq 2$, then

$$\mathbb{E} \|G^\dagger\|_2 \leq \frac{e\sqrt{n}}{n - m}.$$

Lemma 4.10 (Concentration Norm Bounds of Pseudoinverse). Let G be a $k \times (k + p)$ standard Gaussian matrix where $p \geq 4$. For all $t \geq 1$, one has

$$\begin{aligned} P \left\{ \|G^\dagger\|_F \geq \sqrt{\frac{12k}{p}} t \right\} &\leq 4t^{-p} \\ P \left\{ \|G^\dagger\|_2 \geq \frac{e\sqrt{k+p}}{p+1} t \right\} &\leq t^{-p+1}. \end{aligned}$$

Lemma 4.11 (Concentration for Functions of a Gaussian Matrix). Suppose $h(\cdot)$ is a Lipschitz function on matrices, which means that for any X, Y

$$|h(X) - h(Y)| \leq L \|X - Y\|_F$$

holds, and G is a standard Gaussian matrix, then

$$P \{h(G) \geq \mathbb{E} h(G) + Lt\} \leq e^{-\frac{t^2}{2}}.$$

4.5.2. Proof of Theorem 10 and Theorem 11

As mentioned in Remark 4.1, we scale the measurement matrices such that $\mathbb{E} \|\cdot\|_F^2 = mn$. Recall that $\mathcal{A}_1 = \{e_i \Omega_j, 1 \leq i \leq m, 1 \leq j \leq k\}$ and $\mathcal{A}_2 = \{Q_i e_j^T, 1 \leq i \leq r, 1 \leq j \leq n\}$. Therefore, the rescale parameter should be \sqrt{m} and \sqrt{mn} respectively. Following the arguments in Remark 4.1, this is equivalent to assuming $\sigma_{E_1} = \frac{\sigma}{\sqrt{m}}$ and $\sigma_{E_2} = \frac{\sigma}{\sqrt{mn}}$.

$$\begin{aligned} \|\hat{X} - X\| &\leq \|QQ^\top X - X\| + \|QE_2\| \\ &= \|(I - P_{X\Omega+E_1})X\| + \|QE_2\|. \end{aligned}$$

Without loss of generality, we can assume $m \geq n$. Let $X = U\Sigma V^\top$ be the singular value decomposition, where U is $m \times n$, Σ and V are $n \times n$.

Step 1. U plays no essential role in the argument and X can be reduced to \tilde{X} , where

$$\tilde{X} = \begin{matrix} & & n \\ n & & \left(\begin{array}{c} \Sigma V^\top \\ 0 \end{array} \right) \\ m-n & & \end{matrix}. \quad (4.8)$$

Let $\tilde{U} = (U, U^\perp)$ be an orthogonal matrix. Then

$$\|(I - P_{X\Omega+E_1})X\| = \|\tilde{U}^\top (I - P_{X\Omega+E_1}) \tilde{U} \tilde{U}^\top X\| \quad (4.9)$$

$$= \|(I - P_{\tilde{U}^\top X \Omega + \tilde{U}^\top E_1}) \tilde{U}^\top X\|. \quad (4.10)$$

The first equality uses the fact that unitary transformation preserves both Frobenius and Spectral norm. The second equality is due to Lemma 4.4. Step 1 follows by noticing that $\tilde{U}^\top X = \tilde{X}$ and $\tilde{U}^\top E_1 =_d E_1$.

Step 2. The error term $\|(I - P_{X\Omega+E_1})X\|$ can be divided into three parts.

Divide the matrix into blocks

$$X\Omega = \begin{matrix} & r & n-r \\ & \left(\begin{array}{cc} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{array} \right) & \\ \begin{matrix} r \\ n-r \\ m-n \end{matrix} & & \left(\begin{array}{c} n \\ V_1^\top \\ V_2^\top \end{array} \right) \end{matrix} \Omega^{n \times k}. \quad (4.11)$$

Let $\Omega_1 = V_1^\top \Omega$ and $\Omega_2 = V_2^\top \Omega$. Then

$$S_1 = X\Omega + E_1 = \begin{matrix} & k \\ & \left(\begin{array}{c} \Sigma_1 \Omega_1 + E_{11} \\ \Sigma_2 \Omega_2 + E_{12} \\ E_{13} \end{array} \right) \\ \begin{matrix} r \\ n-r \\ m-n \end{matrix} & & \end{matrix}. \quad (4.12)$$

Without loss of generality, we assume all elements in Σ_1 are strictly positive (or else the error bounds will be smaller than the bound we will prove later).

$$Z = S_1(\Sigma_1 \Omega_1 + E_{11})^\dagger = \begin{pmatrix} I \\ F \end{pmatrix} \text{ where } F = \begin{pmatrix} \Sigma_2 \Omega_2 + E_{12} \\ E_{13} \end{pmatrix} (\Omega_1 + \Sigma_1^{-1} E_{11})^\dagger \Sigma_1^{-1}. \quad (4.13)$$

Since $\text{range}(Z) \subset \text{range}(S_1)$, by Lemma 4.5 we have

$$\|(I - P_{X\Omega + E_1})X\| \leq \|(I - P_Z)X\|. \quad (4.14)$$

Let $\tilde{\Sigma} = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{pmatrix}$, for spectral norm

$$\|(I - P_{X\Omega + E_1})X\|_2^2 \leq \|(I - P_Z)X\|_2^2 = \|X^\top (I - P_Z)X\|_2 = \|\tilde{\Sigma}^\top (I - P_Z)\tilde{\Sigma}\|_2. \quad (4.15)$$

Observe that

$$\begin{aligned}
P_Z &= Z(Z^\top Z)^{-1}Z^\top = \begin{pmatrix} I \\ F \end{pmatrix} (I + F^\top F)^{-1} \begin{pmatrix} I & F^\top \end{pmatrix} \\
I - P_Z &= \begin{pmatrix} I - (I + F^\top F)^{-1} & -(I + F^\top F)^{-1}F^\top \\ -F(I + F^\top F)^{-1} & I - F(I + F^\top F)^{-1}F^\top \end{pmatrix}.
\end{aligned} \tag{4.16}$$

By Lemma 4.2, $I - (I + F^\top F)^{-1} \leq F^\top F$. Also notice that $I - F(I + F^\top F)^{-1}F^\top \leq I$. Then

$$I - P_Z \leq \begin{pmatrix} F^\top F & B \\ B^\top & I \end{pmatrix}, \tag{4.17}$$

which implies

$$\tilde{\Sigma}^\top (I - P_Z) \tilde{\Sigma} \leq \begin{pmatrix} \Sigma_1^\top F^\top F \Sigma_1 & \Sigma_1^\top B \Sigma_2 \\ \Sigma_2^\top B^\top \Sigma_1 & \Sigma_2^\top \Sigma_2 \end{pmatrix}. \tag{4.18}$$

By Lemma 4.3,

$$\|\tilde{\Sigma}^\top (I - P_Z) \tilde{\Sigma}\|_2 \leq \|F \Sigma_1\|_2^2 + \|\Sigma_2\|_2^2. \tag{4.19}$$

Recall that $F = \begin{pmatrix} \Sigma_2 \Omega_2 + E_{12} \\ E_{13} \end{pmatrix} (\Omega_1 + \Sigma_1^{-1} E_{11})^\dagger \Sigma_1^{-1}$, we have

$$\|\tilde{\Sigma}^\top (I - P_Z) \tilde{\Sigma}\|_2 \leq \left\| \begin{pmatrix} \Sigma_2 \Omega_2 + E_{12} \\ E_{13} \end{pmatrix} (\Omega_1 + \Sigma_1^{-1} E_{11})^\dagger \right\|_2^2 + \|\Sigma_2\|_2^2. \tag{4.20}$$

Let $\tilde{E} = (E_{12}, E_{13})^\top$, then

$$\|\tilde{\Sigma}^\top (I - P_Z) \tilde{\Sigma}\|_2 \leq \|\Sigma_2 \Omega_2 (\Omega_1 + \Sigma_1^{-1} E_{11})^\dagger\|_2^2 + \|\tilde{E} (\Omega_1 + \Sigma_1^{-1} E_{11})^\dagger\|_2^2 + \|\Sigma_2\|_2^2. \tag{4.21}$$

Let G be a $r \times k$ standard Gaussian matrix, then $(\Omega_1 + \Sigma_1^{-1}E_{11})^\dagger =_d ((I + \sigma_{E_1}\Sigma_1^{-1})G)^\dagger$, and $\|(\Omega_1 + \Sigma_1^{-1}E_{11})^\dagger\|_2 \leq_d \|G^\dagger(I + \sigma_{E_1}\Sigma_1^{-1})\|_2 \leq \|G^\dagger\|_2$. Thus

$$\|\tilde{\Sigma}^\top(I - P_Z)\tilde{\Sigma}\|_2 \leq_d \|\Sigma_2\Omega_2G^\dagger\|_2^2 + \|\tilde{E}G^\dagger\|_2^2 + \|\Sigma_2\|_2^2. \quad (4.22)$$

Similarly, for Frobenius norm, we simply change (4.15) by

$$\|(I - P_{X\Omega+E_1})X\|_F^2 \leq \|(I - P_Z)X\|_F^2 = \text{Tr}(X^\top(I - P_Z)X) = \text{Tr}(\tilde{\Sigma}^\top(I - P_Z)\tilde{\Sigma}) \quad (4.23)$$

$$\leq \text{Tr} \begin{pmatrix} \Sigma_1^\top F^\top F \Sigma_1 & \Sigma_1^\top B \Sigma_2 \\ \Sigma_2^\top B^\top \Sigma_1 & \Sigma_2^\top \Sigma_2 \end{pmatrix} \quad (4.24)$$

$$= \text{Tr}(\Sigma_1^\top F^\top F \Sigma_1) + \text{Tr}(\Sigma_2^\top \Sigma_2) \quad (4.25)$$

$$= \|F\Sigma_1\|_F^2 + \|\Sigma_2\|_F^2. \quad (4.26)$$

Then it follows that

$$\|(I - P_{X\Omega+E_1})X\|_F^2 \leq_d \|\Sigma_2\Omega_2G^\dagger\|_F^2 + \|\tilde{E}G^\dagger\|_F^2 + \|\Sigma_2\|_F^2 \quad (4.27)$$

Combine (4.22) and (4.27), we have

$$\|(I - P_{X\Omega+E_1})X\|_2^2 \leq_d \|\Sigma_2\Omega_2G^\dagger\|_2^2 + \|\tilde{E}G^\dagger\|_2^2 + \|\Sigma_2\|_2^2. \quad (4.28)$$

Step 3. Error bound for Frobenius norm.

By Lemma 4.6 and Lemma 4.8

$$\begin{aligned} \mathbb{E} \|\Sigma_2\Omega_2G^\dagger\|_F^2 &= \|\Sigma_2\|_F^2 \mathbb{E} \|G^\dagger\|_F^2 = \frac{r}{p-1} \sum_{j \geq r+1} \sigma_j^2 \\ \mathbb{E} \|\tilde{E}G^\dagger\|_2^2 &= \|I_{m-r}\|_F^2 \mathbb{E} \|G^\dagger\|_F^2 = (m-r) \frac{r}{p-1} \sigma_{E_1}^2 \\ \mathbb{E} \|QE_2\|_F^2 &= \|Q\|_F^2 \|I_n\|_F^2 = (r+p)n\sigma_{E_2}^2. \end{aligned} \quad (4.29)$$

Then

$$\mathbb{E} \|\hat{X} - X\|_F \leq \mathbb{E} \left(\|\Sigma_2 \Omega_2 G^\dagger\|_F^2 + \|\tilde{E} G^\dagger\|_F^2 + \|\Sigma_2\|_F^2 + \|QE_2\|_F^2 \right)^{\frac{1}{2}} \quad (4.30)$$

$$\leq \mathbb{E} \left(\|\Sigma_2 \Omega_2 G^\dagger\|_F^2 + \|\Sigma_2\|_F^2 \right)^{\frac{1}{2}} + \mathbb{E} \left(\|\tilde{E} G^\dagger\|_F^2 + \|QE_2\|_F^2 \right)^{\frac{1}{2}} \quad (4.31)$$

$$\leq \sqrt{\left(\frac{r}{p-1} + 1\right) \sum_{j \geq r+1} \sigma_j^2} + \sqrt{(m-r) \frac{r}{p-1} \sigma_{E_1}^2 + (r+p) n \sigma_{E_2}^2} \quad (4.32)$$

$$\leq \sqrt{\left(\frac{r}{p-1} + 1\right) \sum_{j \geq r+1} \sigma_j^2} + \sqrt{\frac{r}{p-1} + 1} \sigma \quad (4.33)$$

$$= \sqrt{1 + \frac{r}{p-1}} \left(\sqrt{\sum_{j \geq r+1} \sigma_j^2} + \sigma \right). \quad (4.34)$$

Step 4. Error bound for Spectral norm.

$$\mathbb{E} \|(I - P_{X\Omega+E})X\|_2 \leq \mathbb{E} \left(\|\Sigma_2 \Omega_2 G^\dagger\|_2^2 + \|\tilde{E} G^\dagger\|_2^2 + \|\Sigma_2\|_2^2 \right)^{\frac{1}{2}} \quad (4.35)$$

$$\leq \mathbb{E} \|\Sigma_2 \Omega_2 G^\dagger\|_2 + \mathbb{E} \|\tilde{E} G^\dagger\|_2 + \|\Sigma_2\|_2 \quad (4.36)$$

$$\leq \|\Sigma_2\|_2 \mathbb{E} \|G^\dagger\|_F + \|\Sigma_2\|_F \mathbb{E} \|G^\dagger\|_2 \quad (4.37)$$

$$+ \sigma_{E_1} (\|I_{m-r}\|_2 \mathbb{E} \|G^\dagger\|_F + \|I_{m-r}\|_F \mathbb{E} \|G^\dagger\|_2) + \sigma_{r+1} \quad (4.38)$$

$$\leq \sigma_{r+1} \sqrt{\frac{r}{p-1}} + \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \frac{e\sqrt{r+p}}{p} \quad (4.39)$$

$$+ \sigma_{E_1} \left(\sqrt{\frac{r}{p-1}} + \sqrt{m-r} \frac{e\sqrt{r+p}}{p} \right) + \sigma_{r+1} \quad (4.40)$$

$$= \sigma_{r+1} \left(\sqrt{\frac{r}{p-1}} + 1 \right) + \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \frac{e\sqrt{r+p}}{p} \quad (4.41)$$

$$+ \sigma_{E_1} \left(\sqrt{\frac{r}{p-1}} + \sqrt{m-r} \frac{e\sqrt{r+p}}{p} \right) \quad (4.42)$$

$$\mathbb{E} \|QE_2\|_2 \leq \|I_n\|_F \|Q\|_2 + \|Q\|_F \|I_n\|_2 \leq (\sqrt{n} + \sqrt{r+p}) \sigma_{E_2} \quad (4.43)$$

Thus,

$$\begin{aligned}
\mathbb{E} \|\hat{X} - X\|_2 &\leq \mathbb{E} \|(I - P_{X\Omega+E})X\|_2 + \mathbb{E} \|QE_2\|_2 \\
&\leq \sigma_{r+1} \left(1 + \sqrt{\frac{r}{p-1}}\right) + \left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}} \frac{e\sqrt{r+p}}{p} \\
&\quad + \left(\sqrt{\frac{r}{p-1}} + \sqrt{m-r} \frac{e\sqrt{r+p}}{p}\right) \sigma_{E_1} + (\sqrt{n} + \sqrt{r+p}) \sigma_{E_2}.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E} \|\hat{X} - X\|_2 &\leq \mathbb{E} \|(I - P_{X\Omega+E})X\|_2 + \mathbb{E} \|QE_2\|_2 \\
&\leq \sigma_{r+1} \left(1 + \sqrt{\frac{r}{p-1}}\right) + \left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}} \frac{e\sqrt{r+p}}{p} + \frac{2e\sqrt{r+p}}{p} \sigma + \frac{1}{\sqrt{m}} \sigma \\
&\leq \left(1 + \sqrt{\frac{r}{p-1}}\right) \sigma_{r+1} + \frac{e\sqrt{r+p}}{p} \left(\sum_{j \geq r+1} \sigma_j^2\right)^{\frac{1}{2}} + \frac{3e\sqrt{r+p}}{p} \sigma.
\end{aligned}$$

Step 5. Concentration Bound for Frobenius Norm

For any $t \geq 1$, let

$$E_t = \left\{ G : \|G^\dagger\|_2 \leq \frac{e\sqrt{r+p}}{p+1} t \text{ and } \|G^\dagger\|_F \leq \sqrt{\frac{12r}{p}} t \right\}. \quad (4.44)$$

By Lemma 4.10, $P(E_t) \leq t^{-(p+1)} + 4t^{-p} \leq 5t^{-p}$. Let $h(X) = \|\Sigma_2 X G^\dagger\|_F$, then

$$|h(X) - h(Y)| \leq \|\Sigma_2(X - Y)G^\dagger\|_F \leq \|\Sigma_2\|_2 \|X - Y\|_F \|G^\dagger\|_2. \quad (4.45)$$

By Lemma 4.11, with $L \leq \|\Sigma_2\|_2 \|G^\dagger\|_2$ and $\mathbb{E}[h(\Omega_2)|G] \leq \|\Sigma_2\|_F \|G^\dagger\|_F$

$$P \left\{ \|\Sigma_2 \Omega_2 G^\dagger\|_F \geq \|\Sigma_2\|_F \|G^\dagger\|_F + \|\Sigma_2\|_2 \|G^\dagger\|_2 u \mid E_t \right\} \leq e^{-\frac{u^2}{2}}, \quad (4.46)$$

which implies

$$P \left\{ \|\Sigma_2 \Omega_2 G^\dagger\|_F \geq \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \sqrt{\frac{12r}{p}} t + \sigma_{r+1} \frac{e\sqrt{r+p}}{p+1} ut | E_t \right\} \leq e^{-\frac{u^2}{2}}. \quad (4.47)$$

Similar arguments can be applied to $\|\tilde{E}G^\dagger\|_F$, with $L \leq \|I_{n-r}\|_2 \|G^\dagger\|_2 = \|G^\dagger\|_2$. Then

$$P \left\{ \|\tilde{E}G^\dagger\|_F \geq \sigma_{E_1} \left(\sqrt{m-r} \sqrt{\frac{12r}{p}} t + \frac{e\sqrt{r+p}}{p+1} ut \right) | E_t \right\} \leq e^{-\frac{u^2}{2}} \quad (4.48)$$

$$P \left\{ \|QE_2\|_F \geq \sigma_{E_2} (\sqrt{(r+p)n} + u) \right\} \leq e^{-\frac{u^2}{2}} \quad (4.49)$$

$$\begin{aligned} P \{ \|X - \hat{X}\|_F \geq & \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \left(\sqrt{\frac{12r}{p}} t + 1 \right) + \sigma_{r+1} \frac{e\sqrt{r+p}}{p+1} ut \\ & + \sigma_{E_1} \left(\sqrt{m-r} \sqrt{\frac{12r}{p}} t + \frac{e\sqrt{r+p}}{p+1} ut \right) + \sigma_{E_2} (\sqrt{(r+p)n} + u) \} \\ & \leq 3e^{-\frac{u^2}{2}} + 5t^{-p} \end{aligned} \quad (4.50)$$

$$\begin{aligned} P \{ \|X - \hat{X}\|_F \geq & \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \left(\sqrt{\frac{12r}{p}} t + 1 \right) + \sigma_{r+1} \frac{e\sqrt{r+p}}{p+1} ut + \\ & \sigma \left(\sqrt{\frac{12r}{p}} t + \frac{e\sqrt{r+p}}{p+1} ut \right) + \sigma \left(1 + \frac{1}{\sqrt{mn}} u \right) \} \leq 3e^{-\frac{u^2}{2}} + 5t^{-p} \end{aligned} \quad (4.51)$$

$$P \{ \|X - \hat{X}\|_F \geq \left(\left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} + \sigma \right) \left(\sqrt{\frac{12r}{p}} t + 1 \right) + \frac{e\sqrt{r+p}}{p+1} ut (\sigma_{r+1} + 2\sigma) \} \leq 3e^{-\frac{u^2}{2}} + 5t^{-p}. \quad (4.52)$$

Step 6. Concentration Bound for Spectral Norm

For any $t \geq 1$, let

$$E_t = \left\{ G : \|G^\dagger\|_2 \leq \frac{e\sqrt{r+p}}{p+1} t \text{ and } \|G^\dagger\|_F \leq \sqrt{\frac{12r}{p}} t \right\}. \quad (4.53)$$

By Lemma 4.10, $P(E_t) \leq t^{-(p+1)} + 4t^{-p} \leq 5t^{-p}$. Let $g(X) = \|\Sigma_2 X G^\dagger\|$, then

$$|h(X) - h(Y)| \leq \|\Sigma_2(X - Y)G^\dagger\|_F \leq \|\Sigma_2\|_2 \|X - Y\|_F \|G^\dagger\|_2. \quad (4.54)$$

By Lemma 4.11, with $L \leq \|\Sigma_2\|_2 \|G^\dagger\|_2$ and $\mathbb{E}[h(\Omega_2)|G] \leq \|\Sigma_2\|_2 \|G^\dagger\|_F + \|\Sigma_2\|_F \|G^\dagger\|_2$,

$$P\{\|\Sigma_2 \Omega_2 G^\dagger\|_2 \geq \|\Sigma_2\|_2 \|G^\dagger\|_F + \|\Sigma_2\|_F \|G^\dagger\|_2 + \|\Sigma_2\|_2 \|G^\dagger\|_2 u \mid E_t\} \leq e^{-\frac{u^2}{2}}, \quad (4.55)$$

which implies

$$P\{\|\Sigma_2 \Omega_2 G^\dagger\|_2 \geq \sigma_{r+1} \sqrt{\frac{12r}{p}} t + \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \frac{e\sqrt{r+p}}{p+1} t + \sigma_{r+1} \frac{e\sqrt{r+p}}{p+1} ut \mid E_t\} \leq e^{-\frac{u^2}{2}}. \quad (4.56)$$

Similar arguments can be applied to $\|\tilde{E}G^\dagger\|_F$, with $L \leq \|I_{n-r}\|_2 \|G^\dagger\|_2 = \|G^\dagger\|_2$. Then

$$P\{\|\tilde{E}G^\dagger\|_2 \geq \sigma_{E_1} \left(\sqrt{\frac{12r}{p}} t + (\sqrt{m-r} \frac{e\sqrt{r+p}}{p+1} t + \frac{e\sqrt{r+p}}{p+1} ut) \mid E_t \right) \leq e^{-\frac{u^2}{2}} \quad (4.57)$$

$$P\{\|QE_2\|_2 \geq \sigma_{E_2} (\sqrt{n} + \sqrt{r+p} + u)\} \leq e^{-\frac{u^2}{2}}. \quad (4.58)$$

Put them together, we have

$$\begin{aligned} P\{\|X - \hat{X}\|_2 \geq \sigma_{r+1} \sqrt{\frac{12r}{p}} t + \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \frac{e\sqrt{r+p}}{p+1} t + \\ \sigma_{r+1} \frac{e\sqrt{r+p}}{p+1} ut + \sigma_{E_1} \left(\sqrt{\frac{12r}{p}} t + \sqrt{m-r} \frac{e\sqrt{r+p}}{p+1} t + \frac{e\sqrt{r+p}}{p+1} ut \right) \\ + \sigma_{E_2} (\sqrt{n} + \sqrt{r+p} + u) + \sigma_{r+1} \} \leq 5t^{-p} + e^{-\frac{u^2}{2}}. \quad (4.59) \end{aligned}$$

By simplifying this we get

$$\begin{aligned} P\{\|X - \hat{X}\|_2 \geq \sigma_{r+1} \left(\sqrt{\frac{12r}{p}} t + 1 \right) + \left(\sum_{j \geq r+1} \sigma_j^2 \right)^{\frac{1}{2}} \frac{e\sqrt{r+p}}{p+1} t + (\sigma_{r+1} + 4\sigma) \frac{e\sqrt{r+p}}{p+1} ut \} \\ \leq 5t^{-p} + e^{-\frac{u^2}{2}}. \quad (4.60) \end{aligned}$$

CHAPTER 5 : Large-scale Estimation of Generalized Linear Model

5.1. Background

Generalized linear model (McCullagh, 1984) is one of the most important statistical models that are widely used for prediction and classification. It achieves its success in a great many areas including machine learning (Chang et al., 2008), natural language processing (Genkin et al., 2007), data mining (Komarek and Moore, 2005), computational biology (Wu et al., 2009), epidemiology (Zou, 2004) and many others. Novel theory and applications about generalized linear model are still being put forward (Cleophas and Zwinderman, 2014; Claassen, 2014), but relatively few of them are focusing on computational aspects. The solution to the generalized linear models is typically sought by the maximum likelihood estimator, which is further computed by iterative reweighted least squares (IRLS) (Green, 1984). IRLS is essentially the Newton-Raphson method, where the parameters are updated by the solution of some weighted least squares in each step and the weights are updated by the new parameters iteratively until convergence.

The surge of big data in the past years has posed great challenges to traditional algorithms. There is recently a flurry of research on developing computationally efficient algorithms to handle the challenges across all kinds of areas (Tsang et al., 2005; Xu et al., 2006; Rokhlin and Tygert, 2008; Drineas et al., 2011; Zhang et al., 2012; Dhillon et al., 2013; Lu et al., 2013; Lu and Foster, 2014), including well-studied problems such as least squares. There are considerable amount of works contributed to speeding up the computation of least squares (Rokhlin and Tygert, 2008; Drineas et al., 2011; Zhang et al., 2012; Lu et al., 2013; Dhillon et al., 2013), most of which rely on the idea of random projection, subsampling or parallel computing. Least squares is most commonly used as the solution to linear models. As for generalized linear models, there have been sporadic works focusing on computational aspects, but most of them are restricted to some limited class of models (Keerthi et al.,

2005; Krishnapuram et al., 2005; Ifrim et al., 2008). To the best of our knowledge, no algorithm has been put forward that is capable of estimating any kind of generalized linear model efficiently.

We focus on the problem of how to efficiently estimate the parameters of generalized linear models under the large n , large p and $n \gg p$ setting, where, as usual, n denotes the number of observations and p denotes the number of features. Typical applications could include spam filter, email classification or ad click prediction where sample size could easily accumulate to the millions or billions level while the number of features ranges from several dozens to thousands.

The main contributions of this study are two-fold. Firstly, we propose the Guluru algorithm which solves a wide class of generalized linear models in a unified framework and greatly outperforms the usual IRLS in terms of speed. Specifically, in each iteration we reduce the computational cost from $O(np^2)$ to $O(np)$. Secondly, we provide theoretical justifications that the final log-likelihood achieved by the Guluru algorithm is at most $O(p^2/nr^2)$ away from the maximum log-likelihood under certain conditions, where r is the ratio of subsampling. We also prove that the final estimator of Guluru is only $O(p/nr)$ away from the maximum likelihood estimator, implying that our Guluru algorithm performs asymptotically as well as the IRLS algorithm. We evaluate our approach through extensive synthetic data and real world data studies.

The chapter is organized as follows: Section 5.2 introduces the basic notations and concepts of generalized linear models as well as IRLS, followed by the details of the Guluru algorithm; Section 5.3 provides theoretical guarantees on the performance of Guluru; Section 5.4 presents the results of simulation as well as real case studies; Section 5.5 summarizes the chapter with discussions on possible directions for future work. All the technical proofs are given in Section 5.6.

5.2. The Guluru Algorithm

5.2.1. Notations

We first introduce some notations for generalized linear models which will be used throughout the rest of the chapter. X is the predictor matrix with size $n \times p$. y is the response vector with size $n \times 1$. A typical generalized linear model which links X to y consists of three parts:

- The response y_i which obeys the canonical exponential distribution:

$$f(y_i; \theta_i) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y, \phi_i) \right\}.$$

Here θ_i is the canonical parameter and ϕ is the dispersion parameter.

- The linear predictor associated with each observation:

$$\eta_i = X_i \beta, \quad i = 1, 2, \dots, n$$

Here X_i is the i -th observation and β is the parameter to be estimated.

- The link function $g(\cdot)$ which connects η_i to μ_i , defined as $\mathbb{E}[y_i]$:

$$g(\mu_i) = \eta_i, \quad i = 1, 2, \dots, n.$$

The goal is to estimate β based on y and X . For notational convenience, we focus only on canonical link and suppose there is no dispersion for the exponential family throughout the rest of the chapter, so we have

$$\theta_i = \eta_i = X_i \beta$$

and

$$a(\phi_i) = 1.$$

As a special case, Logistic regression is a class of generalized linear model where $f(y; \theta)$ is binomial distribution with parameter θ and $g(\mu) = \ln(\mu/(1 - \mu))$. Other familiar examples include linear regression, softmax regression and Poisson regression.

5.2.2. Iterative Reweighted Least Squares

The generalized linear model is solved by finding maximizers of its log-likelihood function, which is computed by the Newton-Raphson method.

Excluding the constant terms, the log-likelihood can be written as

$$L(X, y) = \sum_{i=1}^n y_i X_i \beta - b(X_i \beta).$$

The derivative is

$$\frac{\partial L}{\partial \beta} = X^\top (y - \mu),$$

and the second order Hessian matrix is

$$\frac{\partial^2 L}{\partial \beta^2} = -X^\top W X,$$

where $\mu = \mathbb{E}[y]$ and $W = \text{diag}(b''(X_1^\top \beta), b''(X_2^\top \beta), \dots, b''(X_n^\top \beta))$. The Newton-Raphson iteration is usually written as

$$\beta^{new} = \left(X^\top W X \right)^{-1} X^\top W \left(X \beta^{old} + W^{-1} (y - \mu) \right), \quad (5.1)$$

and is the weighted least squares solution of $X \beta^{old} + W^{-1} (y - \mu)$ regressed on X with weight W . To find the maximizer, β^{new} are kept being updated until convergence.

5.2.3. Guleru Algorithm

We can first rewrite formula (5.1) in another form as

$$\beta^{new} = \beta^{old} + \left(X^\top W X \right)^{-1} X^\top (y - \mu). \quad (5.2)$$

Define $\tilde{X} = W^{1/2}X$ and $\tilde{y} = W^{-1/2}(y - \mu)$, then formula (5.2) is reduced to

$$\beta^{new} = \beta^{old} + \left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top \tilde{y}. \quad (5.3)$$

In formula (5.3) $(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{y}$ is the least squares solution of \tilde{y} regressed on \tilde{X} , which we will denote as $\tilde{\beta}$.

It is possible to apply fast algorithms for the least squares to get an approximation for $\tilde{\beta}$, denoted as $\tilde{\beta} + \Delta$ and Δ is the error as a result of the approximation. We first put forward a two-step fast least squares algorithm to compute $\tilde{\beta}$ based on subsampling, which is a slight modification of the Uluru algorithm in (Dhillon et al., 2013) in the sense that we use whole sample here instead of the remaining sample in their paper. The details are described in Algorithm 2. We use $n_s = nr$ to denote the number of observations in the subsample.

Algorithm 2 Faster Least Squares

- 1: **Input:** Response variable y , design matrix X and subsampling ratio r .
 - 2: **Output:** Approximation to the least squares $(X^\top X)^{-1} X^\top y$.
 - 3: Randomly subsample X_s and corresponding y_s from X and y with proportion r .
 - 4: Compute the subsample least squares $\gamma_s \leftarrow (X_s^\top X_s)^{-1} X_s^\top y_s$.
 - 5: Compute the correction $\gamma_* \leftarrow n_s/n(X_s^\top X_s)^{-1} X^\top (y - X\beta_s)$.
 - 6: **Return** $\gamma_s + \gamma_*$.
-

When r is chosen to have order $O(1/p)$, as suggested by (Dhillon et al., 2013), Algorithm 2 runs in $O(np)$ FLOPS in contrast to $O(np^2)$ for the usual least squares. The theoretical properties are provided in Section 5.3.

Formula (5.3) shows that for each iteration in IRLS a least squares solution is to be computed. We put forward our second algorithm, Guluru, to efficiently compute the solution of generalized linear models. The details are sketched below.

Algorithm 3 can be viewed as a variant of the approximate Newton-Raphson method. If we regard the number of iterations required upon convergence as constants, Algorithm 3 runs in $O(np)$ FLOPS in contrast with the $O(np^2)$ for the usual IRLS.

Algorithm 3 Guluru

- 1: **Input:** Response y , design matrix X , log-likelihood function L , subsampling ratio r and predefined tolerance constant ε_0 .
 - 2: **Output:** Approximation to the maximum likelihood estimator $\hat{\beta}$.
 - 3: While $L(\beta^{new}) - L(\beta^{old}) > \varepsilon_0$
 - 4: $\beta^{old} \leftarrow \beta^{new}$.
 - 5: Update \tilde{X} and \tilde{y} based on β^{new} .
 - 6: Compute $\tilde{\beta}$ using \tilde{X} , \tilde{y} and r by Algorithm 2.
 - 7: $\beta^{new} \leftarrow \beta^{old} + \tilde{\beta}$.
 - 8: Update $L(\beta^{new})$ and $L(\beta^{old})$.
 - 9: **Return** β^{new} .
-

5.3. Convergence Analysis

In this section we provide theoretical guarantees for the algorithms put forward in section 5.2. To this aim we need to make some assumptions. First we assume that X_i is i.i.d. subgaussian and the second order moment matrix of X , denoted as Σ , satisfies

$$\lambda_1 I_p \succeq \Sigma \succeq \lambda_2 I_p, \quad (5.4)$$

where I_p is the identity matrix with size $p \times p$ and $\lambda_1 \geq \lambda_2$ are positive constants. We also assume that W is bounded, i.e., there exist positive constants M_1 and M_2 such that $M_1 \geq M_2$ and

$$M_1 I_n \succeq W \succeq M_2 I_n. \quad (5.5)$$

Sometimes the subgaussian assumption can be ideal. For fixed design, it is possible to get similar bounds with same order but higher failure rate by similar techniques shown in the last section. Preconditioning, which usually refers to the randomized Hadamard transform (Tropp, 2011; Lu et al., 2013), is required for fixed design cases, which helps possibly to uniformize data and eliminate high leverage points (Lu et al., 2013).

Assumption (5.5) is natural since we are just assuming that there exists bounded variance for each μ_i . Take Logistic regression for instance, we have $W_{ii} = \hat{y}_i(1 - \hat{y}_i)$ where \hat{y}_i is the predicted value for i -th observation given the parameters. W_{ii} is bounded below means that

\hat{y}_i is prevented from getting too close to 0 or 1.

Please see section 5.6 for a detailed description and of these assumptions.

For Algorithm 2, we have the bound below.

Theorem 12. *Suppose $n_s \gg p \geq \log(2/\delta)$, we have*

$$\|\Delta\| \leq C_1 \frac{p}{n_s} \tag{5.6}$$

with probability at least $1 - 13\delta$. Here C_1 is a constant depending only on λ_1 , λ_2 , M_1 , M_2 and the structure of the subgaussian distribution. So are C_2 and C_3 in the next two theorems.

The bound here proved for Algorithm 2 is a novel bound different from the ones proved in Dhillon et al. (2013). Here we do not assume linear relationship between \tilde{y} and \tilde{X} and we are comparing our estimator with the least squares estimator instead of the true unknown parameters. When we assume the usual linear relationship between \tilde{y} and \tilde{X} , $\tilde{\beta}$ has become the maximum likelihood estimator. Theorem 12 implies that the estimator proposed by Algorithm 2 is only $O(p/nr)$ away from the maximum likelihood estimator, which complements the fact that the estimator is $O(\sqrt{p/n})$ away from the true unknown parameters as proved in Dhillon et al. (2013).

The convergence of generalized linear model is usually diagnosed by the convergence of log-likelihood. Denote the maximum likelihood estimator of the model by β_{MLE} and denote the final estimator of Guluru by β^* . For Algorithm 3 we have

Theorem 13. *Suppose $n_s \gg p \geq \log(2/\delta)$, the final log-likelihood achieved by Guluru are at most*

$$C_2 \frac{p^2}{nr^2} \tag{5.7}$$

less than the maximum log-likelihood with probability at least $1 - \delta$.

Under the big data setting when $n \gg p$ and $r = O(1/p)$, (5.7) has essentially become

$O(p^4/n)$. Here the log-likelihood is the sum for n observations. When it comes to comparing the log-likelihood per sample, we divide (5.7) further by n , resulting in an $O(p^4/n^2)$ bound on average for each sample.

Now we are ready to state our final theorem which gives a bound on the distance between Guluru estimator and the maximum likelihood estimator. We can conclude that

Theorem 14. *Suppose $n_s \gg p \geq \log(2/\delta)$, we have*

$$\|\beta^* - \beta_{MLE}\| \leq C_3 \frac{p}{nr}$$

with probability at least $1 - \delta$.

Theorem 12 illustrates that for each iteration, Guluru is within an $O(p/nr)$ distance to the exact Newton-Raphson step. Theorem 14 guarantees that the final estimator is also within the same order of distance to the maximum likelihood estimator.

All the proof of the theorems are presented in Section 5.6.

5.4. Experiments

In this section we assess the performance of the Guluru algorithm with several examples.

5.4.1. Measure of Performance

We use the number of theoretical FLOPS to compare the speed of different algorithms, which is independent of the CPU settings. Given the number of theoretical FLOPS, we use the gap between current log-likelihood and the maximum log-likelihood to judge the performance of the algorithms. The maximum log-likelihood is computed using the *glm* function in R.

5.4.2. Simulation Studies

First we test Guluru on some synthetic datasets. For the synthetic data, the elements in the design matrix X are i.i.d. drawn from standard normal distribution. The elements of the coefficient β are generated according to

$$\beta_i \sim (-1)^{\mathcal{B}(1,0.5)} \times (0.1 + |\mathcal{N}(0,1)|), \quad i = 1, 2, \dots, p,$$

where \mathcal{B} is the Bernoulli distribution and \mathcal{N} is the normal distribution. The response vector y is generated according to the Logistic model:

$$y_i \sim \mathcal{B}\left(1, \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)}\right), \quad i = 1, 2, \dots, n.$$

To accommodate to the large n , large p and $n \gg p$ setting, we fix sample size to be $n = 1,000,000$ and let p vary from 25 to 100 with increments of 25. The subsampling ratio is fixed to at 0.02. We compare our algorithm with the usual IRLS, gradient descent (GD) and stochastic gradient descent (SGD). For stochastic gradient descent, we get almost identical results to the gradient descent for various settings of batch proportions. In order to avoid confusion resulting from overlapping curves, we choose not to present the results of stochastic gradient descent in the following plots.

The simulation results are plotted in Figure 3. The top row shows the result for $p = 25$ and 50. The bottom row shows the result for $p = 75$ and $p = 100$. Red line is the result for Guluru. Blue line is the result for IRLS and green line is the result for GD. The X-axis denotes the log (base 10) of theoretical FLOPS divided by np and the Y-axis denotes the gap between current log-likelihood and the maximum log-likelihood.

From Figure 3 we can see that overall Guluru is significantly faster than all the others for all choices of p . Upon convergence, our algorithm requires much less FLOPS than the others. The gaps in speed between Guluru and other algorithms widen as p goes larger. As

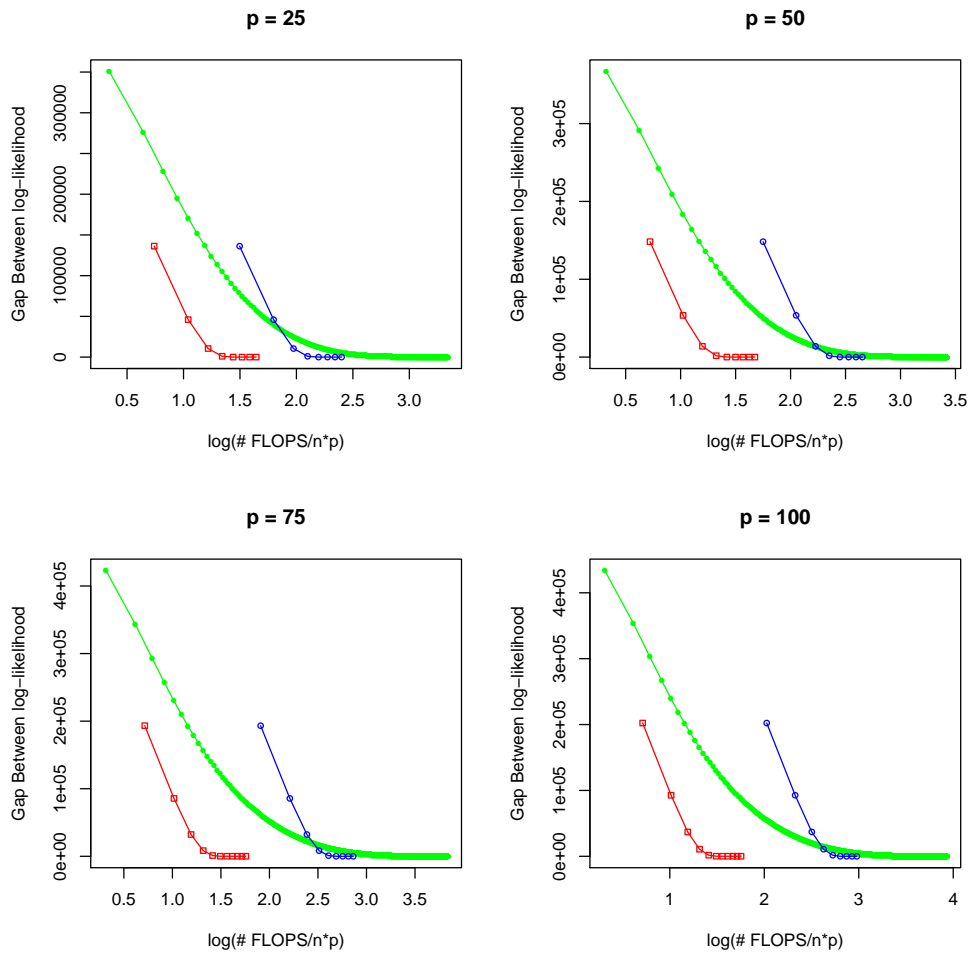


Figure 3: Results for Simulation Studies

for accuracy, all the algorithms finally converge to the maximum log-likelihood¹. Gradient descent is strictly dominated by Guluru. It is also almost dominated by IRLS except for the first few steps.

It is also worth mentioning that although Guluru relies on subsampling, it has almost the same convergence rate as IRLS. After the same number of iterations, they have almost equal log-likelihood, which shows that Guluru is a very good approximation to IRLS.

¹There is a tiny gap between the final log-likelihood returned by Guluru and the maximum log-likelihood in the simulations, which is expected. The gap is negligible compared to the amount of maximum log-likelihood. Specifically, the ratio of the gap to the maximum log-likelihood is less than 10^{-5} .

5.4.3. Real Case Studies

We also try to fit Logistic regression on two UCI datasets: banknote authentication dataset ($n = 1,372, p = 4$) and SUSY dataset ($n = 5,000,000, p = 18$). The previous one is a toy dataset and the latter one is larger. For real data, we compare the algorithms from two aspects: one is the speed of convergence as in the simulation studies; the other is the prediction accuracy. Since neither of these two datasets comes with separate training and test data, we randomly divide each of datasets into training and test data with a ratio of 3 to 1.

For speed of convergence, we present a similar plot as in the simulation studies. The results are presented in Figure 4. We can see that Guluru is slightly faster than the IRLS on the toy dataset and significantly faster than IRLS on the bigger dataset. These two algorithms still dominate the gradient descent in terms of speed. All the algorithms converge to the maximum log-likelihood finally.

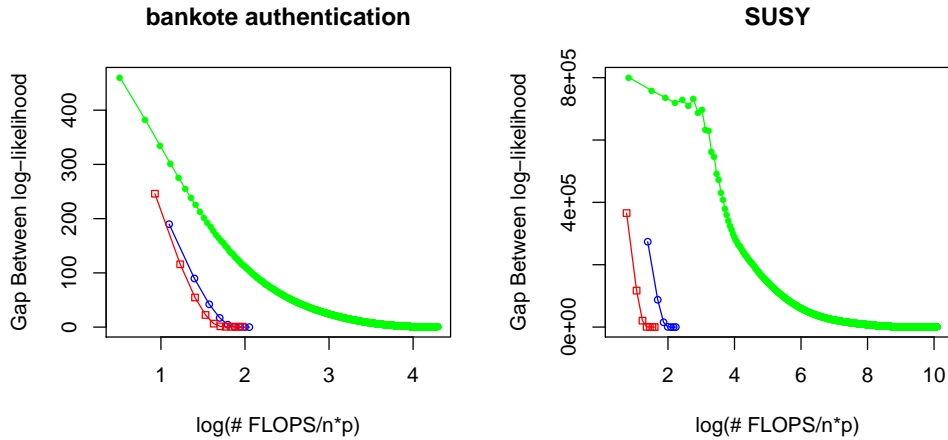


Figure 4: Results for Real Data Studies

As for prediction accuracy, the results are summarized in Table 2.

From Table 2 we can see that there is no significant difference between the prediction accuracy for each method. This is mainly because that all the algorithms are supposed to

	Guluru	IRLS	GD
banknote	322/343 (93.88%)	329/343 (95.91%)	329/343 (95.91%)
SUSY	958,957/1,250,000 (76.71%)	982,593/1,250,000 (78.61%)	982,603/1,250,000 (78.61%)

Table 2: Prediction Accuracy

converge to the maximum likelihood and thus result in very similar estimates for β . We also note that the results produced by IRLS and GD are almost the same. This is expected because they are both supposed to converge to the exact maximum log-likelihood while Guluru is an approximation to the maximum likelihood.

5.5. Discussion

In this paper we proposed the Guluru algorithm for solving generalized linear models in the large n , large p and $n \gg p$ setting. In addition to being significantly faster than the usual IRLS, Guluru leads to a final log-likelihood within an $O(p^2/nr^2)$ neighborhood of the maximum log-likelihood and an estimator within $O(p/nr)$ distance to the maximum likelihood estimator. Experiments on both simulated and real world datasets show that Guluru indeed achieves high speed with almost no loss of optimality compared to other optimization algorithms.

The techniques we use in the proof are quite general. The derivation of error bounds for Algorithm 2 and derivation of optimality for final log-likelihood are independent of each other. It is possible to use other fast least squares algorithm and adapt the framework of our proof to give similar bounds for the final likelihood. We believe as long as the error $\|\Delta\|$ arising from the approximation has order $o(\sqrt{1/n})$, the final likelihood is going to converge to the maximum likelihood.

Our proof is established on the subgaussian assumption. In practice this is sometimes not

satisfied, but we believe that similar theory can be established for the fixed design case after the randomized Hadamard transform (Tropp, 2011), possibly with higher failure rate. Also, there should be room for improvement of the gap between the maximum log-likelihood and log-likelihood achieved by our algorithm. Finally, in Gurlu the subsampling ratio is kept as constant throughout the algorithm, it is of interest to develop an adaptive version of the algorithm to achieve faster speed as well as higher accuracy.

5.6. Details of Proof

5.6.1. Assumptions

We adopt the same notations as in the main article. First we assume that X is i.i.d. subgaussian and the second order moment matrix of X , denoted as Σ , satisfies

$$\lambda_1 I_p \succeq \Sigma \succeq \lambda_2 I_p, \quad (5.8)$$

where I_p is the identity matrix with size $p \times p$ and $\lambda_1 \geq \lambda_2$ are positive constants. Furthermore we assume that W is also bounded, i.e., there exists positive constants M_1 and M_2 such that $M_1 \geq M_2$ and

$$M_1 I_n \succeq W \succeq M_2 I_n. \quad (5.9)$$

Finally we assume that X is still subgaussian after scaled by $y - \mu$, i.e., $\text{diag}(y - \mu)X$ is subgaussian.

The assumption that X is i.i.d. subgaussian is a common assumption in most regression literature. This may be ideal, but for fixed design case, after the randomized Hadamard transform the theory is almost the same, as is shown in Dhillon et al. (2013).

Formula (5.9) is a natural assumption since in some sense we are just assuming that there exists bounded variance for each μ_i . Take Logistic regression for instance, we have $W_{ii} = \hat{y}_i (1 - \hat{y}_i)$ where \hat{y}_i is the predicted value for i -th observation given the parameters. W_{ii} is bounded below means that \hat{y}_i is prevented from getting too close to 0 or 1. The last

assumption is also obviously true for Logistic regression and softmax regression.

5.6.2. Proof of Theorem 1

As in the main paper, each iteration can be written as

$$\beta^{new} = \beta^{old} + \tilde{\beta}$$

where

$$\tilde{\beta} = \left(\tilde{X}^\top \tilde{X} \right)^{-1} \tilde{X}^\top \tilde{y}. \quad (5.10)$$

Algorithm 1 is proposed to compute $\tilde{\beta}$ more efficiently with some error Δ and

$$\Delta = \frac{n_s}{n} \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \tilde{X}^\top \left(\tilde{y} - \tilde{X} \tilde{\beta}_s \right) + \tilde{\beta}_s - \tilde{\beta}.$$

In this section, we try to give a bound on $\|\Delta\|$.

Lemma 5.1. *We have*

$$\lambda_1 + C_1 \sqrt{\frac{p}{n}} + C_2 \sqrt{\frac{\log(2/\delta)}{n}} \geq \left\| \frac{X^\top X}{n} \right\| \geq \lambda_2 - C_1 \sqrt{\frac{p}{n}} - C_2 \sqrt{\frac{\log(2/\delta)}{n}}$$

with probability at least $1 - \delta$.

Proof. From (5.25) in Vershynin (2010) we know that there exist constant C_1 and C_2 which only depend on the structure of the subgaussian distribution such that

$$\left\| \frac{X^\top X}{n} - \Sigma \right\| \leq C_1 \sqrt{\frac{p}{n}} + C_2 \sqrt{\frac{\log(2/\delta)}{n}} \quad (5.11)$$

with probability at least $1 - \delta$. Then the result follows from (5.11) and assumption (5.8). \square

Lemma 5.2. *Conditioned on current β^{old} , we have*

$$\left\| \frac{\tilde{X}_s^\top \tilde{X}_s}{n_s} - \frac{\tilde{X}^\top \tilde{X}}{n} \right\| \leq C_3 \sqrt{\frac{p}{n_s}} + C_4 \sqrt{\frac{\log(2/\delta)}{n_s}}$$

with probability at least $1 - 2\delta$.

Proof. Conditioned on β^{old} , we know the rows of \tilde{X} are still i.i.d. Thus still from (5.25) in Vershynin (2010) we know that there exist C'_3 and C'_4 which only depend on the structure of the subgaussian distribution such that each of

$$\left\| \frac{\tilde{X}_s^\top \tilde{X}_s}{n_s} - \tilde{\Sigma} \right\| \leq C'_3 \sqrt{\frac{p}{n_s}} + C'_4 \sqrt{\frac{\log(2/\delta)}{n_s}}$$

and

$$\left\| \frac{\tilde{X}^\top \tilde{X}}{n} - \tilde{\Sigma} \right\| \leq C'_3 \sqrt{\frac{p}{n}} + C'_4 \sqrt{\frac{\log(2/\delta)}{n}}$$

holds with probability at least $1 - \delta$. Here $\tilde{\Sigma}$ is the second moment matrix of \tilde{X} . The result follows from the above two inequalities. \square

Lemma 5.3. *We have*

$$M_1 \left(\lambda_1 + C_1 \sqrt{\frac{p}{n}} + C_2 \sqrt{\frac{\log(2/\delta)}{n}} \right) \succeq \left\| \frac{\tilde{X}^\top \tilde{X}}{n} \right\| \succeq M_2 \left(\lambda_2 - C_1 \sqrt{\frac{p}{n}} - C_2 \sqrt{\frac{\log(2/\delta)}{n}} \right)$$

with probability at least $1 - \delta$.

Proof. This is an immediate observation from Lemma (5.1) and (5.9). \square

Lemma 5.4. *Suppose A is non-singular and let $\tilde{A} = A + E$, then*

$$\frac{\|A^{-1} - \tilde{A}^{-1}\|}{\|\tilde{A}^{-1}\|} \leq \|A^{-1}E\|.$$

Proof. This comes directly from Theorem 2.5 in Stewart (1990). \square

Lemma 5.5. *Conditioned on current β^{old} , we have*

$$\left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right\| \leq \frac{C_3 \sqrt{\frac{p}{n_s}} + C_4 \sqrt{\frac{\log(2/\delta)}{n_s}}}{M_2^2 \left(\lambda_2 - C_1 \sqrt{\frac{p}{n_s}} - C_2 \sqrt{\frac{\log(2/\delta)}{n_s}} \right)^2}$$

with probability at least $1 - 4\delta$.

Proof. From Lemma 5.2, Lemma 5.3 and Lemma 5.4 we have

$$\begin{aligned}
& \left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right\| \\
& \leq \left\| n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right\| \left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \right\| \left\| \frac{\tilde{X}_s^\top \tilde{X}_s}{n_s} - \frac{\tilde{X}^\top \tilde{X}}{n} \right\| \\
& \leq \frac{C_3 \sqrt{\frac{p}{n_s}} + C_4 \sqrt{\frac{\log(2/\delta)}{n_s}}}{M_2^2 \left(\lambda_2 - C_1 \sqrt{\frac{p}{n}} - C_2 \sqrt{\frac{\log(2/\delta)}{n}} \right) \left(\lambda_2 - C_1 \sqrt{\frac{p}{n_s}} - C_2 \sqrt{\frac{\log(2/\delta)}{n_s}} \right)} \\
& \leq \frac{C_3 \sqrt{\frac{p}{n_s}} + C_4 \sqrt{\frac{\log(2/\delta)}{n_s}}}{M_2^2 \left(\lambda_2 - C_1 \sqrt{\frac{p}{n_s}} - C_2 \sqrt{\frac{\log(2/\delta)}{n_s}} \right)^2}
\end{aligned}$$

with probability at least $1 - 4\delta$. □

Remark 5.6. The bound in Lemma 5.5 simply reduces to

$$\left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right\| \leq C_5 \sqrt{\frac{p}{n_s}}$$

with probability at least $1 - 4\delta$ when $n_s \gg p \geq \log(2/\delta)$.

Lemma 5.7. *Conditioned on current β^{old} , we have*

$$\left\| \tilde{\beta}_s - \tilde{\beta} \right\| \leq C_6 \sqrt{\frac{p \log(2/\delta)}{n_s}}$$

with probability at least $1 - 8\delta$ when $n_s \gg p \geq \log(2/\delta)$.

Proof. We have

$$\begin{aligned}
\|\tilde{\beta}_s - \tilde{\beta}\| &= \left\| \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \tilde{X}_s^\top \tilde{Y}_s - \left(\tilde{X}^\top \tilde{X} \right)^{-1} \tilde{X}^\top \tilde{Y} \right\| \\
&= \left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \frac{\tilde{X}^\top \tilde{Y}}{n} \right\| \\
&\leq \left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} \right\| \\
&\quad + \left\| n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \frac{\tilde{X}^\top \tilde{Y}}{n} \right\| \\
&\leq \left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right\| \left\| \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} \right\| \\
&\quad + \left\| \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} - \frac{\tilde{X}^\top \tilde{Y}}{n} \right\| \left\| n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right\|.
\end{aligned}$$

From the last assumption we made and Corollary 5.17 in Vershynin (2010) we have

$$\left\| \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} - \mathbb{E} \left[\frac{\tilde{X}^\top \tilde{Y}}{n} \right] \right\| \leq C'_6 \sqrt{\frac{p \log(2/\delta)}{n_s}}$$

and

$$\left\| \frac{\tilde{X}^\top \tilde{Y}}{n} - \mathbb{E} \left[\frac{\tilde{X}^\top \tilde{Y}}{n} \right] \right\| \leq C'_6 \sqrt{\frac{p \log(2/\delta)}{n}} \tag{5.12}$$

respectively with probability at least $1 - \delta$.

Thus have

$$\left\| \frac{\tilde{X}_s^\top \tilde{Y}_s}{n_s} - \frac{\tilde{X}^\top \tilde{Y}}{n} \right\| \leq 2C'_6 \sqrt{\frac{p \log(2/\delta)}{n_s}} \tag{5.13}$$

with probability at least $1 - 2\delta$. From Formula (5.12) we know $\|\tilde{X}_s^\top \tilde{Y}_s/n_s\|$ is bounded by a constant with probability at least $1 - \delta$. The result follows from formula (5.13), Lemma 5.3 and Lemma 5.5. \square

Theorem 15. *Conditioned on current β^{old} , we have*

$$\|\Delta\| \leq C_7 \frac{p}{n_s} \sqrt{\log(2/\delta)}$$

with probability at least $1 - 13\delta$ when $n_s \gg p \geq \log(2/\delta)$.

Proof. The difference can be written as

$$\begin{aligned} \|\Delta\| &= \left\| \frac{n_s}{n} \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \tilde{X}^\top \left(\tilde{y} - \tilde{X} \tilde{\beta}_s \right) + \tilde{\beta}_s - \tilde{\beta}_f \right\| \\ &= \left\| \frac{n_s}{n} \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \tilde{X}^\top \tilde{X} \left(\tilde{\beta}_f - \tilde{\beta}_s \right) + \tilde{\beta}_s - \tilde{\beta}_f \right\| \\ &= \left\| \left(\frac{n_s}{n} \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} \tilde{X}^\top \tilde{X} - I \right) \left(\tilde{\beta}_f - \tilde{\beta}_s \right) \right\| \\ &= \left\| \left(n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right) \frac{\tilde{X}^\top \tilde{X}}{n} \left(\tilde{\beta}_f - \tilde{\beta}_s \right) \right\| \\ &\leq \left\| n_s \left(\tilde{X}_s^\top \tilde{X}_s \right)^{-1} - n \left(\tilde{X}^\top \tilde{X} \right)^{-1} \right\| \left\| \frac{\tilde{X}^\top \tilde{X}}{n} \right\| \left\| \tilde{\beta}_f - \tilde{\beta}_s \right\|. \end{aligned}$$

Then the theorem is true from Lemma 5.5, Remark 5.6 and Lemma 5.7. \square

5.6.3. Proof of Theorem 2

Lemma 5.8. *The Hessian of the log-likelihood is negative semi-definite and bounded both from sides*

$$-nM_2 \left(\lambda_2 - C_1 \sqrt{\frac{p}{n}} - C_2 \sqrt{\frac{\log(2/\delta)}{n}} \right) \succeq -X^\top W X \succeq -nM_1 \left(\lambda_1 + C_1 \sqrt{\frac{p}{n}} + C_2 \sqrt{\frac{\log(2/\delta)}{n}} \right) \quad (5.14)$$

with probability at least $1 - \delta$.

Proof. This is an obvious result from Lemma 5.3 \square

Suppose after several iterations we are now at β^{old} . The exact Newton-Raphson step drives

β^{old} to β^{new} where

$$\beta^{new} = \beta^{old} + \tilde{\beta} = \beta^{old} - \left[\nabla^2 L(\beta^{old}) \right]^{-1} \nabla L(\beta^{old})$$

while Guluru drives β^{old} to $\beta^{new} + \Delta$ and

$$\beta^{new} + \Delta = \beta^{old} + \tilde{\beta} + \Delta = \beta^{old} - \left[\nabla^2 L(\beta^{old}) \right]^{-1} \nabla L(\beta^{old}) + \Delta.$$

For notational convenience, denote

$$M'_1 = M_1 \left(\lambda_1 + C_1 \sqrt{\frac{p}{n}} + C_2 \sqrt{\frac{\log(2/\delta)}{n}} \right)$$

and

$$M'_2 = M_2 \left(\lambda_2 - C_1 \sqrt{\frac{p}{n}} - C_2 \sqrt{\frac{\log(2/\delta)}{n}} \right),$$

then formula (5.14) can be rewritten as

$$-nM'_2 \succeq -X^\top W X \succeq -nM'_1.$$

Lemma 5.9. *Suppose overshooting does not happen in the Newton-Raphson step, then we have*

$$L(\beta^{new}) - L(\beta^{old}) \geq \frac{M_2'^2}{2nM_1'^3} \left\| \nabla L(\beta^{old}) \right\|^2$$

with probability at least $1 - \delta$.

Proof. We consider function $g(t) = L(\beta^{old} + t\tilde{\beta})$ restricted on the line segment $0 \leq t \leq 1$. No overshooting implies that $g(t)$ is increasing within this domain. Thus it suffices to prove that we can find t_0 such that

$$L(\beta^{old} + t_0\tilde{\beta}) - L(\beta^{old}) \geq \frac{M_2'^2}{2nM_1'^3} \left\| \nabla L(\beta^{old}) \right\|^2.$$

To this aim, we do a second order Taylor expansion and for any given t and have

$$\begin{aligned}
L(\beta^{old} + t\tilde{\beta}) - L(\beta^{old}) &\geq t\nabla L(\beta^{old})^\top \tilde{\beta} - \frac{nM_1't^2}{2} \|\tilde{\beta}\|^2 \\
&= -t\nabla L(\beta^{old})^\top \left[\nabla^2 L(\beta^{old})^{-1} \right] \nabla L(\beta^{old}) \\
&\quad - \frac{nM_1't^2}{2} \nabla L(\beta^{old})^\top \left[\nabla^2 L(\beta^{old})^{-2} \right] \nabla L(\beta^{old}) \\
&\geq \frac{t}{nM_1'} \|\nabla L(\beta^{old})\|^2 - \frac{M_1't^2}{2nM_2'^2} \|\nabla L(\beta^{old})\|^2,
\end{aligned}$$

where for the inequalities we have used either upper or lower bound for $\nabla^2 L$ (Lemma 5.8) and for the equality we have used the fact that

$$\tilde{\beta} = - \left[\nabla^2 L(\beta^{old})^{-1} \right] \nabla L(\beta^{old}).$$

As this is true for any t , we choose $t_0 = M_2'^2/M_1'^2 < 1$ and have

$$L(\beta^{old} + t_0\tilde{\beta}) - L(\beta^{old}) \geq \frac{M_2'^2}{2nM_1'^3} \|\nabla L(\beta^{old})\|^2.$$

□

Lemma 5.10. *Under the same assumptions as Lemma 5.9, when*

$$\|\nabla L(\beta^{old})\| \geq \frac{nM_1'^2 \left(M_1' + \sqrt{M_1'^2 + M_2'^2} \right)}{M_2'^2} \|\Delta\|,$$

our approximate Newton-Raphson step from Algorithm 1 will increase the likelihood with probability at least $1 - \delta$.

Proof. With probability at least $1 - \delta$ we have

$$\begin{aligned}
L(\beta^{new*}) - L(\beta^{old}) &= L(\beta^{new*}) - L(\beta^{new}) + L(\beta^{new}) - L(\beta^{old}) \\
&\geq L(\beta^{old} + \tilde{\beta} + \Delta) - L(\beta^{old} + \tilde{\beta}) + \frac{M_2'^2}{2nM_1'^3} \left\| \nabla L(\beta^{old}) \right\|^2 \\
&\geq \nabla L(\beta^{old} + \tilde{\beta})^\top \Delta - \frac{nM_1'}{2} \|\Delta\|^2 + \frac{M_2'^2}{2nM_1'^3} \left\| \nabla L(\beta^{old}) \right\|^2 \\
&\geq - \left\| \nabla L(\beta^{old} + \tilde{\beta}) \right\| \|\Delta\| - \frac{nM_1'}{2} \|\Delta\|^2 + \frac{M_2'^2}{2nM_1'^3} \left\| \nabla L(\beta^{old}) \right\|^2 \\
&\geq - \left\| \nabla L(\beta^{old}) \right\| \|\Delta\| - \frac{nM_1'}{2} \|\Delta\|^2 + \frac{M_2'^2}{2nM_1'^3} \left\| \nabla L(\beta^{old}) \right\|^2.
\end{aligned}$$

For the last inequality we have used the fact that Newton-Raphson step is going to decrease $\left\| \nabla L(\beta^{old}) \right\|$, thus

$$\left\| \nabla L(\beta^{old}) \right\| \geq \left\| \nabla L(\beta^{old} + \tilde{\beta}) \right\|.$$

The right hand side is a quadratic function with respect to $\left\| \nabla L(\beta^{old}) \right\|$. It has two roots with different signs and is greater than 0 when $\left\| \nabla L(\beta^{old}) \right\| > nM_1'^2(M_1' + \sqrt{M_1'^2 + M_2'^2})\|\Delta\|/M_2'^2$, which finishes the proof. \square

Lemma 5.11. *Suppose the true unknown maximizer of the log-likelihood function $L(\beta)$ is β_{MLE} , then for any given β other than β_{MLE} we have*

$$L(\beta_{MLE}) \leq L(\beta) + \frac{1}{2nM_2'} \left\| \nabla L(\beta) \right\|^2$$

with probability at least $1 - \delta$.

Proof. From our assumption we know that given β_1 and β_2 , we have

$$\begin{aligned}
L(\beta_1) &\leq L(\beta_2) + \nabla L(\beta_2)^\top (\beta_1 - \beta_2) - \frac{nM_2'}{2} \|\beta_1 - \beta_2\|^2 \\
&\leq L(\beta_2) + \frac{1}{2nM_2'} \left\| \nabla L(\beta_2) \right\|^2
\end{aligned}$$

with probability at least $1 - \delta$. Take $\beta_1 = \beta_{MLE}$ and $\beta_2 = \beta$ we have the desired inequality.

□

From Lemma 5.10 we know each iteration in Algorithm 2 increases likelihood with probability at least $1 - \delta$ as long as the norm of derivative is greater than $nM_1'^2(M_1' + \sqrt{M_1'^2 + M_2'^2})\|\Delta\|/M_2'^2$. Combining the result of Theorem 15 we have, conditioned on β^{old} , each iteration in Algorithm 2 increases log-likelihood with probability at least $1 - 14\delta$ when

$$\|\nabla L(\beta^{old})\| \geq \frac{C_8 p}{r} \sqrt{\log(2/\delta)} \quad (5.15)$$

and $n_s \gg p \geq \log(2/\delta)$.

Theorem 16. *The final log-likelihood is at most*

$$\frac{C_9 p^2}{nr^2}$$

less than the maximized log-likelihood with probability at least $1 - \delta$ when $n_s \gg p \geq \log(2/\delta)$.

Proof. This is obvious from Lemma 5.11 and formula (5.15). □

5.6.4. Proof of Theorem 3

The proof is straightforward. On the one hand we know from Theorem 16 that

$$L(\beta_{MLE}) - L(\beta^*) \leq \frac{C_9 p^2}{nr^2}$$

and on the other hand we have the inequality

$$L(\beta^*) - L(\beta_{MLE}) \leq \nabla L(\beta_{MLE})^\top (\beta^* - \beta_{MLE}) - nM_2' \|\beta^* - \beta_{MLE}\|^2 = -nM_2' \|\beta^* - \beta_{MLE}\|^2$$

each with probability at least $1 - \delta$. Combining the above two inequalities we know with probability at least $1 - 2\delta$.

$$\|\beta^* - \beta_{MLE}\| \leq \frac{C_{10} p}{nr}.$$

BIBLIOGRAPHY

- H. C. Andrews and C. Patterson III. Singular value decomposition (svd) image coding. *Communications, IEEE Transactions on*, 24(4):425–432, 1976.
- A. Arlotto and J. M. Steele. A central limit theorem for temporally non-homogenous markov chains with applications to dynamic programming. to appear in *Mathematics of Operations Research (eprint arXiv:1505.00749)*, 2016.
- A. Arlotto, E. Mossel, and J. M. Steele. Quickest online selection of an increasing subsequence of specified size. to appear in *Random Structures and Algorithms (eprint arXiv:1412.7985)*, 2015a.
- A. Arlotto, V. V. Nguyen, and J. M. Steele. Optimal online selection of a monotone subsequence: a central limit theorem. *Stochastic Process. Appl.*, 125(9):3596–3622, 2015b. ISSN 0304-4149. doi: doi:10.1016/j.spa.2015.03.009.
- R. M. Baer and P. Brock. Natural sorting over permutation spaces. *Math. Comp.*, 22: 385–410, 1968. ISSN 0025-5718.
- J. Baik, P. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999. ISSN 0894-0347. doi: 10.1090/S0894-0347-99-00307-0.
- N. Balakrishnan, A. G. Pakes, and A. Stepanov. On the number and sum of near-record observations. *Adv. in Appl. Probab.*, 37(3):765–780, 2005. ISSN 0001-8678. doi: 10.1239/aap/1127483746.
- R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):218–233, 2003.
- D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control: the discrete time case*, volume 139 of *Mathematics in Science and Engineering*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, NY, 1978. ISBN 0-12-093260-1.
- F. T. Bruss and F. Delbaen. Optimal rules for the sequential selection of monotone subsequences of maximum expected length. *Stochastic Process. Appl.*, 96(2):313–342, 2001. ISSN 0304-4149.
- F. T. Bruss and F. Delbaen. A central limit theorem for the optimal selection process for monotone subsequences of maximum expected length. *Stochastic Process. Appl.*, 114(2): 287–311, 2004. ISSN 0304-4149. doi: 10.1016/j.spa.2004.09.002.
- F. T. Bruss and J. B. Robertson. “Wald’s lemma” for sums of order statistics of i.i.d. random variables. *Adv. in Appl. Probab.*, 23(3):612–623, 1991. ISSN 0001-8678. doi: 10.2307/1427625.

- T. T. Cai and A. Zhang. Rop: Matrix recovery via rank-one projections. *arXiv preprint arXiv:1310.5791*, 2013.
- E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.
- E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
- E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.
- E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.
- E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
- M.-w. Chang, W.-t. Yih, and C. Meek. Partitioned logistic regression for spam filtering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 97–105. ACM, 2008.
- E. A. Claassen. *A Reduced Bias Method of Estimating Variance Components in Generalized Linear Mixed Models*, 2014.
- T. J. Cleophas and A. H. Zwinderman. Generalized linear models for predicting event-rates (50 patients). In *Machine Learning in Medicine-Cookbook*, pages 37–41. Springer, 2014.
- E. G. Coffman, Jr., L. Flatto, and R. R. Weber. Optimal selection of stochastic intervals under a sum constraint. *Adv. in Appl. Probab.*, 19(2):454–473, 1987. ISSN 0001-8678. doi: 10.2307/1427427.
- G. Derfel and A. Iserles. The pantograph equation in the complex plane. *J. Math. Anal. Appl.*, 213(1):117–132, 1997. ISSN 0022-247X. doi: 10.1006/jmaa.1997.5483. URL <http://dx.doi.org/10.1006/jmaa.1997.5483>.
- P. Dhillon, Y. Lu, D. P. Foster, and L. Ungar. New subsampling algorithms for fast least squares regression. In *Advances in Neural Information Processing Systems*, pages 360–368, 2013.
- D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.
- D. L. Donoho, M. Gavish, and A. Montanari. The phase transition of matrix recovery from

- gaussian measurements matches the minimax mse of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, 2013.
- P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.
- M. Fazel, E. Candes, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1043–1047. IEEE, 2008.
- W. Feller. *An introduction to probability theory and its applications. Vol. II*. Second edition. John Wiley & Sons, Inc., New York-London-Sydney, 1971.
- A. Flat and G. J. Woeginger. *Online Algorithms: The State of the Art*. Lecture Notes in Computer Science. Springer, New York, 1998. ISBN 10: 3540649174.
- M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.
- L. Fox, D. F. Mayers, J. R. Ockendon, and A. B. Tayler. On a functional differential equation. *J. Inst. Math. Appl.*, 8:271–307, 1971. ISSN 0020-2932.
- A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.
- A. V. Gnedin. Sequential selection of an increasing subsequence from a sample of random size. *J. Appl. Probab.*, 36(4):1074–1085, 1999. ISSN 0021-9002.
- A. V. Gnedin. A note on sequential selection from permutations. *Combin. Probab. Comput.*, 9(1):13–17, 2000. ISSN 0963-5483.
- D. Goldfarb and S. Ma. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210, 2011.
- R. Gouet, F. J. López, and G. Sanz. Asymptotic normality for the counting process of weak records and δ -records in discrete models. *Bernoulli*, 13(3):754–781, 2007. ISSN 1350-7265. doi: 10.3150/07-BEJ6027. URL <http://dx.doi.org/10.3150/07-BEJ6027>.
- R. Gouet, F. J. López, and G. Sanz. Central limit theorem for the number of near-records. *Comm. Statist. Theory Methods*, 41(2):309–324, 2012. ISSN 0361-0926. doi: 10.1080/03610926.2010.522753.
- M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.

- M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. <http://cvxr.com/cvx>, Mar. 2014.
- P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society B*, 46(2): 149–192, 1984.
- D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.
- N. Guglielmi and M. Zennaro. Stability of one-leg Θ -methods for the variable coefficient pantograph equation on the quasi-geometric mesh. *IMA J. Numer. Anal.*, 23(3):421–438, 2003. ISSN 0272-4979. doi: 10.1093/imanum/23.3.421. URL <http://dx.doi.org/10.1093/imanum/23.3.421>.
- N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- C.-H. Hsiao. A Haar wavelets method of solving differential equations characterizing the dynamics of a current collection system for an electric locomotive. *Appl. Math. Comput.*, 265:928–935, 2015. ISSN 0096-3003. doi: 10.1016/j.amc.2015.06.007. URL <http://dx.doi.org/10.1016/j.amc.2015.06.007>.
- G. Ifrim, G. Bakir, and G. Weikum. Fast logistic regression for text categorization with variable-length n-grams. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 354–362. ACM, 2008.
- A. Iserles. On the generalized pantograph functional-differential equation. *European J. Appl. Math.*, 4(1):1–38, 1993. ISSN 0956-7925. doi: 10.1017/S0956792500000966. URL <http://dx.doi.org/10.1017/S0956792500000966>.
- T. Kato and J. B. McLeod. The functional-differential equation $y'(x) = ay(\lambda x) + by(x)$. *Bull. Amer. Math. Soc.*, 77:891–937, 1971. ISSN 0002-9904.
- S. S. Keerthi, K. Duan, S. K. Shevade, and A. N. Poo. A fast dual algorithm for kernel logistic regression. *Machine learning*, 61(1-3):151–165, 2005.
- P. Komarek and A. W. Moore. Making logistic regression a core data mining tool with tr-irls. In *Data Mining, Fifth IEEE International Conference on*, pages 4–pp. IEEE, 2005.
- B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):957–968, 2005.
- K. Lee and Y. Bresler. Admira: Atomic decomposition for minimum rank approximation. *Information Theory, IEEE Transactions on*, 56(9):4402–4416, 2010.

- G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171–184, 2013.
- Y.-K. Liu. Universal low-rank matrix recovery from pauli measurements. In *NIPS*, pages 1638–1646, 2011.
- Y. Lu and D. P. Foster. Fast ridge regression with randomized principal component analysis and gradient descent. *arXiv preprint arXiv:1405.3952*, 2014.
- Y. Lu, P. Dhillon, D. P. Foster, and L. Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems*, pages 369–377, 2013.
- K. Mahler. On a special functional equation. *J. London Math. Soc.*, 15:115–123, 1940. ISSN 0024-6107.
- P. McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.
- K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473, 2012.
- S. V. Nagaev. The spectral method and ergodic theorems for general Markov chains. *Izv. Ross. Akad. Nauk Ser. Mat.*, 79(2):101–136, 2015. ISSN 0373-2436. doi: 10.4213/im8198. URL <http://dx.doi.org/10.4213/im8198>.
- M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1994. ISBN 0-471-61977-9. A Wiley-Interscience Publication.
- B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- A. Rényi. Théorie des éléments saillants d’une suite d’observations. *Ann. Fac. Sci. Univ. Clermont-Ferrand No.*, 8:7–13, 1962.
- V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.
- D. Romik. *The Surprising Mathematics of Longest Increasing Subsequences*. Cambridge University Press, Cambridge, 2014. ISBN 1107075831; 978-1107428829.
- A. Saadatmandi and M. Dehghan. Variational iteration method for solving a generalized pantograph equation. *Comput. Math. Appl.*, 58(11-12):2190–2196, 2009. ISSN 0898-1221.

- doi: 10.1016/j.camwa.2009.03.017. URL <http://dx.doi.org/10.1016/j.camwa.2009.03.017>.
- S. M. Samuels and J. M. Steele. Optimal sequential selection of a monotone sequence from a random sample. *Ann. Probab.*, 9(6):937–947, 1981. ISSN 0091-1798.
- T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.
- A. N. Shiryaev. *Optimal stopping rules*, volume 8 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2008. ISBN 978-3-540-74010-0. Translated from the 1976 Russian second edition by A. B. Aries, Reprint of the 1978 translation.
- J. M. Steele. The Bruss-Robertson inequality: Elaborations, extensions, and applications. to appear in *Mathematica Applicanda (Annales Societatis Mathematicae Polonae Series III) Volume 44, 2016 (eprint arXiv:1510.00843)*, 2016.
- G. W. Stewart. *Matrix perturbation theory*, 1990.
- J. A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.
- I. W. Tsang, J. T. Kwok, P.-M. Cheung, and N. Cristianini. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6(4), 2005.
- B. van Brunt and G. C. Wake. A Mellin transform solution to a second-order pantograph equation with linear dispersion arising in a cell growth model. *European J. Appl. Math.*, 22(2):151–168, 2011. ISSN 0956-7925. doi: 10.1017/S0956792510000367. URL <http://dx.doi.org/10.1017/S0956792510000367>.
- R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.
- T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.
- Y. Xu, D. Zhang, Z. Jin, M. Li, and J.-Y. Yang. A fast kernel-based nonlinear discriminant analysis for multi-class problems. *Pattern Recognition*, 39(6):1026–1033, 2006.
- E. Yusufoglu. An efficient algorithm for solving generalized pantograph equations with linear functional argument. *Appl. Math. Comput.*, 217(7):3591–3595, 2010. ISSN 0096-3003. doi: 10.1016/j.amc.2010.09.005. URL <http://dx.doi.org/10.1016/j.amc.2010.09.005>.

- Y. Zhang, M. J. Wainwright, and J. C. Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.
- G. Zou. A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, 159(7):702–706, 2004.