Indian Institute of Technology, Kanpur

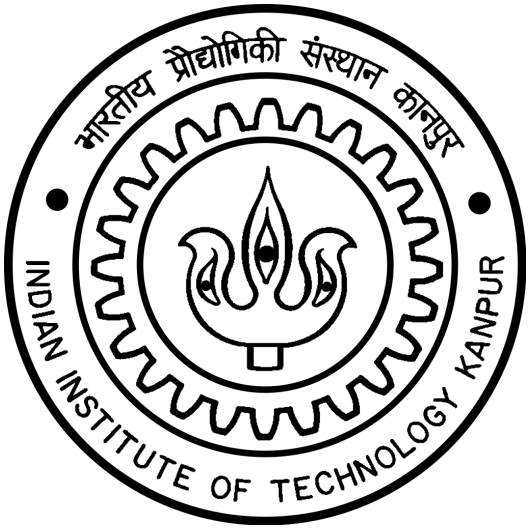
MTH697A Report

C-Optimal Design of Experiments for High-Dimensional Linear Models: Theory

Author
Keshav Ranjan

Supervisor
Prof. Satya Prakash Singh

31st October 2024



CERTIFICATE

This is to certify that the project entitled "C-Optimal Design of Experiments for High-Dimensional Linear Models: Theory" submitted by Keshav Ranjan (Roll no.: 208170508) as a part of MTH697A course offered by the Indian Institute of Technology, Kanpur, is a Bonafide record of the work done by him under my guidance and supervision from 1^{st} August 2024 to 31^{st} October 2024.

Dr. Satya Prakash Singh
Assistant Professor,
Dept. of Mathematics & Statistics
Indian Institute of Technology Kanpur

Annexure-II

DECLARATION

I/We hereby declare that the work presented in the project report entitled C-Optimal Design of Experiments for High-Dimensional Linear Models: Theory is written by me in my own words and contains my own or borrowed ideas. At places, where ideas and words are borrowed from other sources, proper references and acknowledgements, as applicable, have been provided. To the best of my knowledge this work does not emanate from or resemble work created by person(s) other than those mentioned and acknowledged herein.

Name and Signature Keshav Ranjan Keshav Ranjan

Date: 30/10/2024

Acknowledgement

I would like to express my heartfelt gratitude to the Indian Institute of Technology Kanpur and my project supervisor, **Prof. Satya Prakash Singh**, for the opportunity to pursue my MS research through the **MTH697A** course. The knowledge and experience I gained from this project have been invaluable for my future endeavors. I am particularly grateful to **Prof. Satya Prakash Singh** for his mentorship, guidance, and the constructive feedback he provided on my technical reports. His support has significantly enriched my research journey, and I truly appreciate the time and effort he dedicated to my development.

Abstract

This research addresses the design of c-optimal experiments in the context of high-dimensional linear models, where the dimension of covariates exceeds the sample size. In such scenarios, traditional inference methods which focus on low-dimension settings, face significant challenges, particularly when estimating linear combinations of parameters in sparse regression models. Our study focuses on the use of debiased lasso estimators, which are known for their strong theoretical properties but can exhibit bias when the dimensionality is high.

To address this, we propose a novel framework utilizing randomized designs to minimize the asymptotic variance of the debiased lasso estimator when a large pool of unlabeled data is available, but measuring the corresponding responses is costly. By framing the optimal sampling distribution as a solution to a semidefinite program (SDP), we derive constrained c-optimal designs that effectively control bias while optimizing variance.

The results of this work provide a practical and theoretically sound method for designing high-dimensional experiments, bridging the gap between classical optimal design theory and modern high-dimensional statistical inference. Our findings highlight the utility of constrained c-optimal designs in sparse regression problems, offering both computational and statistical advantages.

Keywords: c-optimal design; high-dimensional; sparsity; lasso; semidefinite programming.

Contents

Chapter 1

INTRODUCTION

1.1 Motivation

Optimal experimental design is a crucial framework in statistics, especially when resources are limited or costly. In many modern applications, such as genomics or medical imaging, the number of the covariates p often exceeds the available fixed sample size n. This high-dimensional setting presents challenges for traditional regression methods, making consistent estimation and inference difficult without structural assumptions like sparsity. Additionally, measuring the responses (y_i's) for all covariate data (x_i's) can be expensive, motivating the need for designs that minimize the cost while maximizing statistical efficiency. This is especially relevant when only a limited number of responses can be measured, and the goal is to derive accurate estimates from these limited observations.

In classical framework, the c-optimal design is the design (i.e. a distribution on the covariate data $(x_i)_i$) that minimizes the variance of an ordinary least squares (OLS) estimate of $\langle c, \beta^* \rangle$, where $\beta^* \in \mathbb{R}^p$ is the true regression parameter and c is a pre-specified vector. However, in high-dimensional scenarios where the number of covariates p significantly exceeds the sample size n (i.e., $p \gg n$), the ordinary least squares (OLS) method becomes impractical. In such cases, traditional regression techniques often fail to produce reliable estimates due to overfitting and high variance. Alternative methods, like the Lasso, introduce regularization to combat these issues, but this can lead to biases in the estimates.

A key assumption in high-dimensional settings is the sparsity of the regression parameter β^*, meaning that only a small subset of covariates has a significant effect on the response. Recent developments, particularly the debiased Lasso, leverage this sparsity by providing estimates that are both consistent and asymptotically normal. However, the challenge of designing optimal experiments that effectively utilize these advanced estimation techniques, while accounting for the inherent sparsity, remains largely unaddressed in the existing literature. This gap highlights the necessity for innovative design strategies specifically tailored for high-dimensional models, ensuring that we can harness the advantages of modern statistical methods while mitigating issues related to bias and variance.

1.2 Problem Statement

In the context of high-dimensional linear models, we consider a scenario where the number of covariates p exceeds the number of observations n. Let us denote the covariate data as $X = \{x_i \in \mathbb{R}^p\}_{i=1}^N$, where N is the total number of available design points. The objective is to observe the responses $Y = \{y_i\}_{i=1}^n$ corresponding to a selected subset of these covariates.

The relationship between the response y and covariates x is modeled linearly as:

$$y = \langle x, \beta^* \rangle + \epsilon, \tag{1.1}$$

where $\beta^* \in \mathbb{R}^p$ is the true regression parameter, and ϵ is a mean-zero noise term, often assumed to be sub-Gaussian. The parameter of interest for inference is typically a linear combination of the regression coefficients, denoted as $\langle c, \beta^* \rangle$, where $c \in \mathbb{R}^p$ is a predefined contrast vector.

We assume that $N \geq p$ throughout our analysis, as this condition is essential for ensuring that the population covariance matrix is non-singular. A non-singular covariance matrix is necessary to guarantee the restricted eigenvalue condition for sample design matrices. In scenarios where $N < p$, although the restricted eigenvalue condition may hold for either the population or sample covariance matrices, it is important to note that verifying

these conditions is *NP-hard* (Dobriban and Fan, 2016). Consequently, enforcing such conditions on design matrices also becomes a complex challenge.

In high-dimensional settings where $p > n$, a critical assumption is the sparsity of the true regression parameter β^*. Specifically, we assume that only a small number of components of β^* are non-zero, reflecting a common structure in many high-dimensional datasets. This sparsity assumption allows for the effective use of regularization techniques such as the Lasso, which selects relevant covariates while mitigating overfitting. However, the presence of sparsity also complicates the design of experiments, as the optimal choice of covariate points must account for the underlying structure of the parameter vector.

While the Lasso provides sparse solutions, it can introduce bias in the estimates. To address this bias, we utilize the debiased Lasso estimator that offers asymptotically unbiased estimates of the regression coefficients, but minimizing its variance remains a challenge. The c-optimality criterion in classical settings focuses on minimizing the variance of $\langle c, \beta^* \rangle$ for some fixed vector c, but this criterion needs to be adapted for high-dimensional models, where the estimator bias cannot be ignored.

In this report, we address the following questions:

1. How can we design experiments that minimize the asymptotic variance of a debiased Lasso estimator in high-dimensional linear models?

2. How can we ensure that the resulting designs account for the potential bias of the estimator, which may not vanish in finite samples?

To solve these problems, we formulate the design task as a semidefinite programming (SDP) problem, imposing constraints that ensure both the variance and bias of the estimator are controlled.

1.3 Objective

The main objective of this report is to develop a framework for c-optimal experimental design in high-dimensional linear models. Specifically, we aim to:

1. Extend the classical c-optimal design criterion to the high-dimensional setting by introducing constraints that control the bias of the debiased Lasso estimator.

2. Formulate the design problem as a semidefinite program (SDP) that minimizes the variance of the estimator while ensuring the bias remains controlled.

By bridging the gap between classical optimal design theory and modern high-dimensional inference methods, we provide a comprehensive solution to the problem of designing experiments in high-dimensional settings.

Chapter 2

Literature Review

The topic of c-optimal experimental designs has been extensively studied in the context of low-dimensional regression models, where various optimality criteria have been developed to improve the efficiency of parameter estimation. The seminal work of Elfving [1] provided an elegant geometric solution for the (approximate) c-optimal design problem, which can be formulated as a linear program admitting efficient solutions, as detailed in the work by Harman and Jurík [2]. This linear programming approach allows researchers to minimize the variance of the estimates effectively while satisfying design constraints. Chernoff [3] later extended these results to include local and asymptotic notions of optimality, enriching the theoretical landscape of optimal designs. Pukelsheim [4] further contributed to this body of work by systematically organizing optimal design theory with a focus on practical implementation across various regression models.

In high-dimensional settings (where the number of covariates exceeds the number of observations, $p > n$), classical methods face challenges related to inefficiency and inconsistency. The advent of penalized regression techniques such as the Lasso [5], which introduced sparsity-inducing regularization, provided a solution to handle high-dimensional covariates. However, Lasso's bias, induced by regularization, led to the development of debiasing techniques for obtaining consistent parameter estimates. Zhang et al. [6] and van de Geer et al. [7] contributed significantly by proposing methods to achieve $\sqrt{n}$-consistent estimators from the Lasso under certain sparsity conditions, forming the foundation for high-dimensional inference and extending the scope of optimal experimental design to bias correction.

Despite these advances, the literature on optimal experimental design for high-dimensional linear models remains scarce. Seeger [8] explored sequential design strategies for maximizing information in sparse models, while Ravi et al. [9] introduced D-optimal designs tailored for sparse models. Huang et al. [10] expanded the design criteria to include the covariance matrix of the debiased Lasso, further highlighting the increasing complexity of balancing estimation accuracy with computational efficiency in high-dimensional settings.

This report extends these discussions by proposing constrained c-optimal designs for high-dimensional linear models that account for both the bias and variance of the debiased Lasso estimator. This is achieved through a semidefinite programming approach, which allows for more efficient designs that control the bias without sacrificing precision. The application of these designs to sparse MRI reconstruction demonstrates the practical importance of reducing measurement costs while preserving the accuracy of inference, further emphasizing the need for targeted experimental designs in real-world applications.

Chapter 3

Notations & Definitions

This section outlines the key notations and definitions used in this report:

1. **Set Definition**: For any natural number q: $[q] := \{0, 1, \ldots, q\}$
2. **Parameters**:
   - p: Dimension of covariates.
   - N: Total number of available design points.
   - n: Sample size for observed responses, satisfying $n \leq p \leq N$.
3. **Covariate Data**: $x_i \in \mathbb{R}^p$: Potential covariate data; experimental domain defined as $(x_i)_{i=1}^N$.
4. **Sample and Responses**:
   - Sample $(X_i)_{i=1}^n$ from $(x_i)_{i=1}^N$ with responses $(y_i)_{i=1}^n$.
   - $\mathbb{X}$: The $n \times p$ matrix with rows X_i^T.
   - $Y = (y_1, \ldots, y_n)^T$: Vector of responses.
5. **Inner Product and Norms**:
   - **Inner Product**: $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^p$.
   - ℓ_q**-Norm**: $\|a\|_q = \left(\sum_i |a_i|^q\right)^{1/q}$ for $q \geq 1$.
   - ℓ_0**-Norm**: $\|a\|_0 = |supp(a)|$.
   - **Standard Basis**: e_j for the j-th element in $\mathbb{R}^p$.
6. **Random Variables**: For a random variable or vector Z:
   - **Expected Value**: $\mathbb{E}[Z]$.
   - $\mathbb{E}_x[X] : \mathbb{E}[X \mid x]$ (conditional expectation of X given x).
   - $\mathrm{Var}_x(X) : \mathrm{Var}(X \mid x)$ (conditional variance of X given x).
   - **Probability of event E**: $\mathbb{P}(E)$.
   - **Sub-Gaussian Norm**: $\|Z\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(|Z|^2/t^2)] < 2\}$.
   - **Sub-Exponential Norm**: $\|Z\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[\exp(|Z|/t)] < 2\}$.
7. **Matrices**: For positive semidefinite matrix Σ:
   - Eigenvalues: $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denotes its smallest and largest eigenvalues.
   - Matrix relation: $A \preccurlyeq B$ means $B - A$ is positive semidefinite.
   - Pseudo-inverse: Σ^+.
8. **Sequences**: For sequences a_n and b_n:
   - $a_n \lesssim b_n$ implies $a_n \leq C b_n$ for some constant $C > 0$.
   - Reverse inequality: $a_n \gtrsim b_n$.
   - Both inequalities: $a_n \asymp b_n$ implies $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Chapter 4

C-optimality in low-dimensions

4.1 Introduction

In this section, we provide a thorough mathematical analysis of C-optimality for low-dimensional settings, exploring the relevant theorems and optimization techniques. This serves as the foundation for extending the analysis to high-dimensional cases.

4.2 Linear Regression Framework

Suppose that $n > p$, and we are given a fixed sample $\{(X_i, y_i)\}_{i=1}^n$, and let $\mathbb{X} \in \mathbb{R}^{n\times p}$ be the matrix with X_i^T in its i-th row. We consider the standard linear regression model:

$$Y = \mathbb{X}\beta^\star + \epsilon \tag{4.1}$$

where:

- $Y \in \mathbb{R}^n$ is the response vector,
- $\beta^\star \in \mathbb{R}^p$ is the vector of regression parameters, and
- $\epsilon \in \mathbb{R}^n$ is the error vector, assumed to be independent with mean zero and variance $\sigma^2 I_n$.

The goal is to estimate a specific linear combination $c^T\beta^\star$, where $c \in \mathbb{R}^p$ is a fixed vector. In C-optimality, we aim to minimize the variance of this estimator.

4.3 Variance of the OLS Estimator

Assuming that $c \in \text{span}(X_1, \ldots, X_n)$, by the Gauss-Markov theorem, the best linear unbiased estimate of $c^T\beta^\star$ is $c^T\hat{\beta}_{\text{OLS}}$, where the ordinary least squares (OLS) estimate of $\beta^\star$ is defined by:

$$\hat{\beta}_{OLS} = (\mathbb{X}^T\mathbb{X})^+\mathbb{X}^T Y \tag{4.2}$$

The variance of $c^T\hat{\beta}_{OLS}$ is given by:

$$\text{Var}(c^T\hat{\beta}_{OLS}) = \sigma^2 c^T(\mathbb{X}^T\mathbb{X})^+ c \tag{4.3}$$

The objective of C-optimality is to minimize this variance. This corresponds to finding the optimal design points (rows of $\mathbb{X}$) that minimize $c^T(\mathbb{X}^T\mathbb{X})^+c$.

4.4 Exact C-Optimal Design Problem

Let N_i represent the number of times design point x_i is selected in the experiment. The exact C-optimal design problem can be formulated as:

$$\min_{N_1,\ldots,N_N} \; c^T\Sigma^+ c \tag{4.4}$$

subject to:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{N} N_i x_i x_i^T, \quad \sum_{i=1}^{N} N_i = n, \quad N_i \in \{0, 1, \ldots, n\} \tag{4.5}$$

This problem is NP-complete and hence, computationally hard to solve due to the integer constraints on N_i[11]. Therefore, a more computationally feasible solution is needed.

4.5 Approximate C-Optimality Problem

To simplify the problem, the integer constraints on N_i are relaxed by allowing them to be non-negative real values, denoted as weights w_i. The approximate C-optimal design problem becomes:

$$\mathbf{P1}: \min_{w_1,\ldots,w_N} c^T \Sigma^+ c \tag{4.6}$$

subject to:

$$\Sigma = \sum_{i=1}^{N} w_i x_i x_i^T, \quad \sum_{i=1}^{N} w_i = 1, \quad w_i \geq 0, \quad c \in col(\Sigma) \tag{4.7}$$

This transforms the problem into a convex optimization problem, which is computationally solvable. It is important to note that when dealing with an approximate optimal design $(w_i^*)_{i=1}^N$, the values nw_i^* may not be whole numbers. Consequently, rounding methods may be necessary to convert the approximate optimal design into an exact design [12]. An alternative approach involves randomization, where the set $\{X_1, \ldots, X_n\}$ is treated as an independent and identically distributed (iid) sample drawn from $(x_i)_{i=1}^N$ with the corresponding probabilities $(w_i^*)_{i=1}^N$. In our work, we adopt the latter method, as randomization facilitates the establishment of high-probability bounds on the bias of the de-biased lasso by ensuring that a restricted eigenvalue condition holds with high probability.

4.6 Elfving's Theorem

Elfving's theorem [1] provides an elegant geometric solution to the approximate C-optimal design problem. It states:

- Let $x_1, x_2, \ldots, x_N \in \mathbb{R}^p$ represent the design points, and let $c \in \mathbb{R}^p$ be the target vector.
- The **Elfving set** $\mathcal{E}$ is defined as the convex hull formed by the set of points $(\pm x_i)_{i=1}^N$:

$$\mathcal{E} := \text{conv}\left(\{x_i : 1 \leq i \leq N\} \cup \{-x_i : 1 \leq i \leq N\}\right) \tag{4.8}$$

- Let x_c be the point on the boundary of the **Elfving set** $\mathcal{E}$ that intersects with the half-line extending from the origin in the direction of c:

$$x_c = \partial\mathcal{E} \cap \{tc : t \geq 0\} \tag{4.9}$$

- Assuming that we can express x_c as a linear combination of the design points i.e. Let $x_c = \sum_{i=1}^N v_i x_i$, then the optimal weights for the approximate design are given by:

$$w_i^* = \frac{|v_i|}{\sum_{i=1}^N |v_i|} \tag{4.10}$$

4.7 Linear Programming Formulation

Using Elfving's theorem, the approximate C-optimality problem is transformed into a linear programming problem:

$$\mathbf{P2}: \min_{b \in \mathbb{R}^N} \|b\|_1 \tag{4.11}$$

subject to:

$$c = \sum_{i=1}^{N} b_i x_i \tag{4.12}$$

If b^* is the optimal solution for **P2**, then the optimal weights for **P1** is given by $w_i^* = \frac{b_i^*}{\sum_{i=1}^N b_i^*}$.

4.8 Conclusion & Remark

In low-dimensional settings where $n > p$, it is generally feasible to express a target linear combination $\mathbf{c}$ as a combination of fewer than n design points. This allows for the use of rounding techniques to convert approximate designs, represented by weights w_i, into deterministic exact designs where the number of times each design point $\mathbf{x}_i$ is selected is an integer value close to nw_i. These techniques are supported by C-optimality, which minimizes the variance of a desired linear combination of parameters and simplifies the problem via Elfving's theorem. This approach is computationally efficient, transforming the design problem into a convex optimization task.

However, in high-dimensional scenarios where $p \gg n$, expressing $\mathbf{c}$ as a linear combination of n design points becomes infeasible. In such cases, it is possible to encounter many non-zero weights w_i, often with values close to zero, making traditional rounding techniques impractical. This motivates the adoption of randomized designs. Randomization not only bypasses the limitations of rounding but also provides probabilistic guarantees, such as the restricted eigenvalue condition, which is crucial for the analysis of the lasso estimator. These methods are particularly relevant in high-dimensional contexts, where standard deterministic design approaches may fail to be effective.

Chapter 5

Theory for the Lasso

In high-dimensional settings, where the dimension of covariates p exceeds the number of observations n, traditional regression methods like ordinary least squares (OLS) often fail due to overfitting and non-uniqueness in the solution. The Least Absolute Shrinkage and Selection Operator (LASSO) addresses these issues by introducing an ℓ_1-norm regularization, which encourages sparsity in the coefficient estimates.

For the linear model (model 4.1), The LASSO estimator, $\hat{\beta} := \hat{\beta}(\lambda)$, is defined as the solution to the following optimization problem:

$$\hat{\beta}(\lambda) \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{1}$$

where $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and $\lambda > 0$ is the regularization parameter controlling the level of shrinkage.

Some Useful Notations:

- Let β_0 be the truth for model 4.1
- The true active set, denoted by S_0, is the support of the subset selection solution ($S_0 = \text{supp}(\beta_0)$) and is defined as:
$$S_0 = \{j; \beta_{0_j} \neq 0\}.$$
- Let s_0 be the cardinality of S_0, i.e., $s_0 = |S_0|$, hence $s_0 < p$.

5.1 Some Useful Bounds

In this section, we derive important bounds which are required for proving oracle bounds for the Lasso.

Lemma 1 (Basic Inequality)**.** :

$$\frac{1}{n} \|X(\hat{\beta} - \beta_0)\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{2}{n} \epsilon^T X(\hat{\beta} - \beta_0) + \lambda \|\beta_0\|_1. \tag{5.1}$$

Proof. The Lasso estimator $\hat{\beta}$ minimizes the objective function:

$$\frac{1}{n} \|Y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1, \tag{5.2}$$

so for any β_0, we have the inequality:

$$\frac{1}{n} \|Y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{n} \|Y - X\beta_0\|_2^2 + \lambda \|\beta_0\|_1. \tag{5.3}$$

Using $Y = X\beta_0 + \epsilon$, where ϵ is the noise term, we can express the left-hand side of (5.3) as:

$$\frac{1}{n} \|Y - X\hat{\beta}\|_2^2 = \frac{1}{n} \|X(\beta_0 - \hat{\beta}) + \epsilon\|_2^2. \tag{5.4}$$

Expanding (5.4) gives:

$$\frac{1}{n}\|X(\beta_0-\hat{\beta})\|_2^2+\frac{2}{n}\epsilon^T X(\beta_0-\hat{\beta})+\frac{1}{n}\|\epsilon\|_2^2. \tag{5.5}$$

Substituting (5.5) into (5.3), we obtain:

$$\frac{1}{n}\|X(\beta_0-\hat{\beta})\|_2^2+\frac{2}{n}\epsilon^T X(\beta_0-\hat{\beta})+\frac{1}{n}\|\epsilon\|_2^2+\lambda\|\hat{\beta}\|_1\leq\frac{1}{n}\|\epsilon\|_2^2+\lambda\|\beta_0\|_1. \tag{5.6}$$

Cancel the common term $\frac{1}{n}\|\epsilon\|_2^2$ from both sides of (5.6):

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2+\lambda\|\hat{\beta}\|_1\leq\frac{2}{n}\epsilon^T X(\hat{\beta}-\beta_0)+\lambda\|\beta_0\|_1. \tag{5.7}$$

□

The term $2\epsilon^T X(\hat{\beta}-\beta_0)/n$ is called the stochastic process part, and it can be bounded by the ℓ_1-norm of the Lasso error. Applying Hölder's inequality (14), we get the following:

$$2|\epsilon^T X(\hat{\beta}-\beta_0)|\leq\left(\max_{1\leq j\leq p}2|\epsilon^T X_j|\right)\|\hat{\beta}-\beta_0\|_1.$$

Now let's introduce an event:

$$\mathscr{T}:=\left\{\max_{1\leq j\leq p}\frac{2|\epsilon^T X_j|}{n}\leq\lambda_0\right\}.$$

where we assume (quite arbitrarily) that $\lambda\geq 2\lambda_0$ to make sure that, on $\mathscr{T}$, we can get rid of the random part of the problem and thus we can write:

$$2\epsilon^T X(\hat{\beta}-\beta_0)/n\leq\frac{\lambda}{2}\|\hat{\beta}-\beta_0\|_1. \tag{5.8}$$

We need to show that event $\mathscr{T}$ has high probability for suitable value of λ_0. This is answered by the following lemma.

Lemma 2. Suppose that $\hat{\sigma}_j=\hat{\Sigma}_{j,j}=1$ where $\hat{\Sigma}=\frac{X^TX}{n}$ for all j. Then we have for all $t>0$, and for

$$\lambda_0=2\sigma\sqrt{\frac{t^2+2\log p}{n}},$$

$$\mathbb{P}(\mathscr{T})\geq 1-2\exp\left(-\frac{t^2}{2}\right).$$

Proof. Recall that $\mathscr{T}=\left\{\max_{1\leq j\leq p}\frac{2|\epsilon^T X_j|}{n}\leq\lambda_0\right\}$. We have

$$1-\mathbb{P}(\mathscr{T})=\mathbb{P}\left\{\max_{1\leq j\leq p}\frac{2|\epsilon^T X_j|}{n}>\lambda_0\right\}=\mathbb{P}\left\{\max_{1\leq j\leq p}\frac{2|\epsilon^T X_j|}{\sqrt{n}\sigma^2}>2\sqrt{t^2+2\log p}\right\}$$

$$\leq p\mathbb{P}\left\{\frac{2|\epsilon^T X_j|}{\sqrt{n}\sigma^2}>2\sqrt{t^2+2\log p}\right\}=2p\mathbb{P}\left\{\frac{2\epsilon^T X_j}{\sqrt{n}\sigma^2}>2\sqrt{t^2+2\log p}\right\}$$

$$\leq 2p\exp\left\{-\frac{t^2+2\log p}{2}\right\}=2\exp\left(-\frac{t^2}{2}\right).$$

□

5.2 Oracle Inequality for Lasso

In this section, we study various conditions required on design matrix X to establish oracle results for the Lasso. Let us write, for an index set $S \subset \{1, \ldots, p\}$,

$$\beta_{j,S} := \beta_j \mathbb{I}\{j \in S\},$$

and (hence)

$$\beta_{j,S^c} := \beta_j \mathbb{I}\{j \notin S\}.$$

Thus, β_S has zeroes outside the set S, and the elements of β_{S^c} can only be non-zero in the complement S^c of S. Clearly,

$$\beta = \beta_S + \beta_{S^c}.$$

Definition 1. (Restricted Eigenvalue Condition$(RE(L, s_0, X))$ For a set $S \subset \{1, \ldots, p\}$ and constant $L > 0$, the (L, s_0, X)-restricted eigenvalue for a matrix X is

$$\phi^2_{RE}(L, s_0, X) := \min \left\{ \frac{\|X\Delta\|_2^2}{\|\Delta_S\|_2^2} : \|\Delta_{S^c}\|_1 \leq L\|\Delta_S\|_1 \neq 0 \right\}.$$

The restricted eigenvalue condition is said to be met if $\phi_{RE}(L, s_0, X) > 0$ for all subsets S of size less than equal to s_0 i.e. $|S| \leq s_0$.

RE-condition has been used to derive oracle results for the estimation and prediction. A slightly stronger version of restricted eigenvalue condition as given by :

Definition 2. ((Strong) Restricted Eigenvalue Condition) For a set $S \subset \{1, \ldots, p\}$ and constant $L > 0$, the (L, s_0, X)-strong restricted eigenvalue for a matrix X is

$$\phi^2_{\text{str}}(L, s_0, X) := \min \left\{ \frac{\|X\Delta\|_2^2}{\|\Delta\|_2^2} : \|\Delta_{S^c}\|_1 \leq L\|\Delta_S\|_1 \neq 0 \right\}.$$

The strong restricted eigenvalue condition is said to be met if $\phi_{\text{str}}(L, s_0, X) > 0$ for all subsets S of size less than equal to s_0 i.e. $|S| \leq s_0$.

Definition 3. (Compatibility Condition) For a fixed active set S_0 with cardinality $s_0 = |S_0|$ and constant $L > 0$, the (L, s_0, X)-restricted ℓ_1 eigenvalue is

$$\phi^2_{\text{comp}}((L, s_0, X)) := \min \left\{ \frac{\|X\Delta\|_2^2 s_0}{\|\Delta_S\|_1^2} : \|\Delta_{S^c}\|_1 \leq L\|\Delta_S\|_1 \neq 0 \right\}.$$

The (L, s_0, X) compatibility condition is said to be satisfied for the set S, if $\phi_{\text{comp}}(L, s_0, X) > 0$.

Note: Throughout the report we will assume L=3.
We note that the RE-condition implies that the compatibility condition holds for all subsets S of size s. The compatibility condition depends on the set S, whereas the RE condition depends only on the cardinality $s = |S|$. It follows that the RE condition is stronger than the compatibility condition.

Now, let's get back to our proof for oracle bound,

Lemma 3. We have on $\mathscr{T}$, with $\lambda \geq 2\lambda_0$,

$$2\|X(\hat{\beta} - \beta_0)\|_2^2/n + \lambda\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\beta}_{S_0} - \beta_{0_{S_0}}\|_1.$$

Proof. On $\mathscr{T}$, by the Basic Inequality, and using $2\lambda_0 \leq \lambda$,

$$\frac{\|X(\hat{\beta} - \beta_0)\|_2^2}{n} + \lambda\|\hat{\beta}\|_1 \leq \frac{2\epsilon^T X(\hat{\beta} - \beta_0)}{n} + \lambda\|\beta_0\|_1.$$

Using (5.8),

$$\frac{\|X(\hat{\beta} - \beta_0)\|_2^2}{n} + \lambda\|\hat{\beta}\|_1 \leq \frac{\lambda}{2}\|\hat{\beta} - \beta_0\|_1 + \lambda\|\beta_0\|_1,$$

we have

$$2\frac{\|X(\hat{\beta}-\beta_0)\|_2^2}{n}+2\lambda\|\hat{\beta}\|_1\leq\lambda\|\hat{\beta}-\beta_0\|_1+2\lambda\|\beta_0\|_1.$$

But on the left-hand side, using the triangle inequality,

$$\|\hat{\beta}\|_1=\|\hat{\beta}_{S_0}\|_1+\|\hat{\beta}_{S_0^c}\|_1\geq\|\beta_{0_{S_0}}\|_1-\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1+\|\hat{\beta}_{S_0^c}\|_1,$$

whereas on the right-hand side, we can invoke

$$\|\hat{\beta}-\beta_0\|_1=\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1+\|\hat{\beta}_{S_0^c}\|_1.$$

Hence, By simplifying the inequalities above, we obtain the required result. □

In Lemma 3 above, a term involving the ℓ_1-norm $\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1$ occurs on the right-hand side. To eliminate it, we want to incorporate this term into the left-hand side's $2\|X(\hat{\beta}-\beta_0)\|_2^2/n$.

Clearly, by the Cauchy-Schwarz inequality, we can replace the ℓ_1-norm with the ℓ_2-norm given by:

$$\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1\leq\sqrt{s_0}\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_2.$$

Theorem 1. (Consistency of Lasso)

(A) (Slow rate of Convergence) An optimal solution $\hat{\beta}$ of the Lasso problem satisfies the following bound on $\mathscr{T}$, for $\lambda\geq 2\lambda_0$:

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2\leq\frac{3\lambda}{2}\|\beta_0\|_1\lesssim\cdot\sigma\|\beta_0\|_1\sqrt{\frac{\log p}{n}}.$$

(B) (Fast Rate of Convergence) If the design matrix satisfies the $\phi_{\text{comp}}(3,S_0)$ compatibility condition over S_0, then an optimal solution $\hat{\beta}$ satisfies the following bound:

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2\leq\frac{9\lambda^2 s_0}{4\phi_{\text{comp}}^2}\lesssim\cdot\frac{\sigma^2\log p}{n}\cdot\frac{s_0}{\phi_{\text{comp}}^2}.$$

Proof. **(A):** From the basic inequality and inequality (5.8), we derive the following:

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2\leq\frac{\lambda}{2}\|\hat{\beta}-\beta_0\|_1+\lambda\left(\|\beta_0\|_1-\|\hat{\beta}\|_1\right)$$

which leads to

$$0\leq\frac{\lambda}{2}\|\hat{\beta}-\beta_0\|_1+\lambda\left(\|\beta_0\|_1-\|\hat{\beta}\|_1\right).$$

Applying the inequality $\|\hat{\beta}-\beta_0\|_1\leq\|\hat{\beta}\|_1+\|\beta_0\|_1$, we obtain:

$$0\leq\frac{\lambda}{2}\left(\|\hat{\beta}\|_1+\|\beta_0\|_1\right)+\lambda\left(\|\beta_0\|_1-\|\hat{\beta}\|_1\right).$$

This simplifies to:

$$\implies\|\hat{\beta}\|_1\leq 3\|\beta_0\|_1.$$

Considering the basic inequality again:

$$\begin{aligned}\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2&\leq\frac{\lambda}{2}\left(\|\hat{\beta}\|_1+\|\beta_0\|_1\right)+\lambda\left(\|\beta_0\|_1-\|\hat{\beta}\|_1\right)\\&\leq\frac{3\lambda}{2}\|\beta_0\|_1-\frac{\lambda}{2}\|\hat{\beta}\|_1.\end{aligned}$$

We can rearrange this as:

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2+\frac{\lambda}{2}\|\hat{\beta}\|_1\leq\frac{3\lambda}{2}\|\beta_0\|_1.$$

Thus,

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2\leq\frac{3\lambda}{2}\|\beta_0\|_1\lesssim\cdot\sigma\|\beta_0\|_1\sqrt{\frac{\log p}{n}}\quad(\because\lambda=O\left(\sigma\sqrt{\frac{\log p}{n}}\right)).$$

We note that consistency for the prediction can be achieved only if $\|\beta_0\|_1 \ll \sqrt{\frac{n}{\log p}}$.

(B): Further simplifying lemma 3 we get,

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 \leq \frac{3}{2}\lambda\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1.$$

Substituting $\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1 \leq \frac{\|X(\hat{\beta}-\beta_0)\|_2\sqrt{s_0}}{\sqrt{n}\phi_{\text{comp}}(3,s_0,X)}$ (Compatibility Condition):

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 \leq \frac{3\lambda}{2}\|X(\hat{\beta}-\beta_0)\|_2\frac{\sqrt{s_0}}{\sqrt{n}\phi_{\text{comp}}(3,s_0,X)}.$$

Dividing both sides by $\sqrt{n}$:

$$\frac{1}{\sqrt{n}}\|X(\hat{\beta}-\beta_0)\|_2 \leq \frac{3\lambda\sqrt{s_0}}{2\phi_{\text{comp}}(3,s_0,X)}.$$

Thus, we have:

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 \leq \frac{9\lambda^2 s_0}{4\phi_{\text{comp}}^2(3,s_0,X)} \lesssim \frac{\lambda^2 s_0}{\phi_{\text{comp}}^2(3,s_0,X)}.$$

□

Theorem 2. (Oracle Inequality) Suppose the compatibility condition holds for S_0. Then on $\mathscr{T}$, for $\lambda \geq 2\lambda_0$, we have

$$\frac{\|X(\hat{\beta}-\beta_0)\|_2^2}{n} + \lambda\|\hat{\beta}-\beta_0\|_1 \leq \frac{4\lambda^2 s_0}{\phi_{\text{comp}}^2(3,s_0,X)}.$$

Proof. We continue with Lemma 3 and add $\lambda\|\hat{\beta}-\beta_0\|_1$ on both side which gives,

$$\frac{2}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 + \lambda\|\hat{\beta}-\beta_0\|_1 \leq 3\lambda\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1 - \lambda\|\hat{\beta}_{S_0^c}\|_1 + \lambda\left(\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1 + \|\hat{\beta}_{S_0^c}-\beta_{0_{S_0^c}}\|_1\right)$$

$$\leq 4\lambda\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1.$$

Substituting $\|\hat{\beta}_{S_0}-\beta_{0_{S_0}}\|_1 \leq \frac{\|X(\hat{\beta}-\beta_0)\|_2\sqrt{s_0}}{\sqrt{n}\phi_{\text{comp}}(3,s_0,X)}$ (Compatibility Condition):

$$\frac{2}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 + \lambda\|\hat{\beta}-\beta_0\|_1 \leq 4\lambda\frac{\|X(\hat{\beta}-\beta_0)\|_2\sqrt{s_0}}{\sqrt{n}\phi_{\text{comp}}(3,s_0,X)}.$$

Applying the inequality $4ab \leq a^2 + 4b^2$:

$$\frac{2}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 + \lambda\|\hat{\beta}-\beta_0\|_1 \leq \frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 + 4s_0\frac{\lambda^2}{\phi_{\text{comp}}^2(3,s_0,X)}.$$

Thus, we have:

$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 + \lambda\|\hat{\beta}-\beta_0\|_1 \leq 4s_0\frac{\lambda^2}{\phi_{\text{comp}}^2(3,s_0,X)}.$$

□

The above inequality provides two bounds:

1. The prediction error bound:
$$\frac{1}{n}\|X(\hat{\beta}-\beta_0)\|_2^2 \leq \frac{4s\lambda^2}{\phi_{\text{comp}}^2}.$$
2. The ℓ_1-estimation error bound:
$$\lambda\|\hat{\beta}-\beta_0\|_1 \leq \frac{4s\lambda^2}{\phi_{\text{comp}}^2}.$$

   Hence, we conclude that:
$$\|\hat{\beta}-\beta_0\|_1 \leq \frac{4s\lambda}{\phi_{\text{comp}}^2} \lesssim \frac{s\lambda}{\phi_{\text{comp}}^2}.$$

Chapter 6

Debiased Lasso in High-Dimensional Inference

In high-dimensional settings, the Lasso is a commonly used estimator for parameter estimation due to its ability to handle sparsity. However, the Lasso estimator is inherently biased, which presents challenges for inference, particularly for linear functionals such as $\gamma := \langle c, \beta \rangle$. To overcome this, a debiasing procedure is applied to correct the bias and allow valid inference. Javanmard and Lee [13] developed a flexible framework for hypothesis testing in high-dimensional settings, addressing general null hypotheses of the form $\beta \in \Omega_0$ for arbitrary Ω_0. In parallel, Cai, Cai, and Guo [14] focused on the inference of individualized treatment effects $\langle c, \beta_1 - \beta_2 \rangle$ in two-sample problems, and linear functionals $\langle c, \beta \rangle$ in one-sample settings. A shared methodology across these works is the application of debiasing techniques, where a biased estimator, such as Lasso, is corrected using projection methods. We outline a variant of the method proposed by Cai, Cai, and Guo [14] in the following.

A variant of the Lasso(or Weighted Lasso) estimator was used to obtain an initial estimate of β^* and is defined by:

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} \hat{W}_j |\beta_j| \right\},$$

where $\hat{W}_j$ is a normalization factor to account for the variability in the covariates i.e. $\hat{W}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2}$, and λ is the regularization parameter, typically chosen as $\lambda \asymp \tilde{\sigma}_\varepsilon \sqrt{\log p / n}$, with $\tilde{\sigma}_\varepsilon^2$ representing the noise variance.

To address the bias in the Lasso estimator $\langle c, \hat{\beta} \rangle$, a debiasing step is introduced. Cai, Cai, and Guo [4] propose to use the following debiased estimator:

$$\hat{\gamma} = \langle c, \hat{\beta} \rangle + \frac{1}{n} \hat{u}^T \mathbb{X}^T (Y - \mathbb{X}\hat{\beta}),$$

where

$$\hat{u} := \arg\min_{u} u^T \hat{\Sigma} u$$

$$\text{s.t.} \quad \|\hat{\Sigma}u - c\|_\infty \le \|c\|_2 \lambda, \quad |c^T \hat{\Sigma} u - \|c\|_2^2| \le \|c\|_2^2 \lambda.$$

The above minimization problem effectively estimates u $= \Sigma^{-1} c$ in the typical setting where Σ is unknown. However, in scenarios where Σ is known, such as in our case, we can directly use $u = \Sigma^{-1} c$ for debiasing procedure. We will use the debiased estimator

$$\hat{\gamma} := \langle c, \hat{\beta} \rangle + \frac{1}{n} u^T \mathbb{X}^T (Y - \mathbb{X}\hat{\beta}) \tag{6.1}$$

with $u = \Sigma^{-1} c$.

6.1 Poisson Sampling Scheme

Before presenting our main theorem on the consistency and asymptotic normality of $\hat{\gamma}$, we first describe and motivate the Poisson sampling scheme employed in our analysis.

6.1.1 Poisson Sampling

Let $\mathbf{w} = (w_i)_{i=1}^N$ be a probability distribution over the set $(x_i)_{i=1}^N$. Given an i.i.d. sample $X_1, \ldots, X_n$ drawn according to the probabilities specified by $\mathbf{w}$, we define $\tilde{N}_i$ to represent the count of how many times x_i appears in the sample:

$$\tilde{N}_i := |\{k \in \{1, \ldots, n\} : X_k = x_i\}|.$$

Random quantities such as $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ that arise in our theoretical framework can be expressed as functions of the multinomial random variables $\tilde{N}_i$:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^N \tilde{N}_i x_i x_i^T.$$

This formulation shifts the focus from high-dimensional random vectors $(X_i)_{i=1}^n$ to the random variables $(\tilde{N}_i)_{i=1}^N$, which are more tractable for theoretical analysis. Consequently, studying the concentration properties of $\hat{\Sigma}$ requires an examination of $\tilde{N}_i$.

However, the dependence among the $\tilde{N}_i$ introduces challenges for concentration arguments, which typically rely on independence. Note that under i.i.d. sampling, the distribution of $(\tilde{N}_i)_{i=1}^N$ follows a multinomial distribution with probabilities $(w_i)_{i=1}^N$ and a total sum of n.

To circumvent these issues, we can utilize a Poisson sampling approach, which effectively breaks the dependence among the $\tilde{N}_i$. Specifically, we define independent random variables N_i where $N_i \sim \text{Poisson}(nw_i)$. We then set $K = \sum_{i=1}^N N_i$, creating a sample $X_1, \ldots, X_K$ such that each x_j appears N_j times in this sample.

6.1.2 Properties of Poisson Sampling

1. **Total Number of Samples:** The total count of samples drawn, K, converges in probability to n as n approaches infinity:

$$\frac{K}{n} \xrightarrow{p} 1 \quad \text{as } n \to \infty.$$

Proof. Since $N_i \sim \text{Poisson}(nw_i)$, the expectation of N_i is:

$$\mathbb{E}[N_i] = nw_i.$$

Therefore, the expected total number of samples is:

$$\mathbb{E}[K] = \mathbb{E}\left[\sum_{i=1}^N N_i\right] = \sum_{i=1}^N \mathbb{E}[N_i] = n\sum_{i=1}^N w_i = n.$$

To show that $\frac{K}{n}$ converges in probability to 1, we can use the law of large numbers. The Poisson distribution's variance is equal to its mean:

$$\text{Var}(N_i) = nw_i.$$

Thus, the variance of K can be computed as:

$$\text{Var}(K) = \sum_{i=1}^N \text{Var}(N_i) = \sum_{i=1}^N nw_i = n.$$

By Chebyshev's inequality:

$$\mathbb{P}\left(|K - n| \geq \epsilon n\right) \leq \frac{\text{Var}(K)}{\epsilon^2 n^2} = \frac{1}{\epsilon^2 n} \to 0 \quad \text{as } n \to \infty.$$

Therefore, we conclude that K converges in probability to n:

$$\frac{K}{n} \xrightarrow{p} 1 \quad \text{as } n \to \infty.$$

□

2. **Conditioned Distribution:** Given $K = k$, the distribution of $(N_i)_{i=1}^N$ follows a multinomial distribution with parameters $\big(k, (w_i)_{i=1}^N\big)$:

Proof. Let N_i be independent random variables such that $N_i \sim \text{Poisson}(nw_i)$ for $i = 1, 2, \ldots, N$. We denote the total count of samples as $K = \sum_{i=1}^N N_i$. Since N_i are independent Poisson random variables, their sum K is also Poisson distributed:

$$K \sim \text{Poisson}\left(n \sum_{i=1}^N w_i\right) = \text{Poisson}(n),$$

because $\sum_{i=1}^N w_i = 1$.

The joint probability mass function (PMF) for $(N_1, N_2, \ldots, N_N, K)$ can be expressed as:

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_N = n_N, K = k) = \prod_{i=1}^N P(N_i = n_i),$$

where $n_1 + n_2 + \cdots + n_N = k$. The PMF for each N_i is given by:

$$P(N_i = n_i) = \frac{(nw_i)^{n_i} e^{-nw_i}}{n_i!}.$$

The probability mass function $P(K = k)$ is given by:

$$P(K = k) = \frac{n^k e^{-n}}{k!},$$

since $K \sim \text{Poisson}(n)$.

We can express the conditional distribution as:

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_N | K = k) = \frac{P(N_1 = n_1, N_2 = n_2, \ldots, N_N, K = k)}{P(K = k)}.$$

Substituting the joint distribution gives:

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_N | K = k) = \frac{\prod_{i=1}^N \frac{(nw_i)^{n_i} e^{-nw_i}}{n_i!}}{\frac{n^k e^{-n}}{k!}}.$$

Simplifying the fraction, we have:

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_N | K = k) = \frac{k!}{n_1! n_2! \cdots n_N!} \cdot \frac{(nw_1)^{n_1} (nw_2)^{n_2} \cdots (nw_N)^{n_N} e^{-n \sum_{i=1}^N w_i}}{n^k e^{-n}}.$$

Noticing that $n^k = n^{n_1 + n_2 + \cdots + n_N}$ and since $\sum_{i=1}^N w_i = 1$ we can factor out e^{-n}.

This can be rearranged as follows:

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_N | K = k) = \frac{k!}{n_1! n_2! \cdots n_N!} \cdot w_1^{n_1} w_2^{n_2} \cdots w_N^{n_N},$$

Therefore, we conclude that conditioned on $K = k$, the distribution of $(N_i)_{i=1}^N$ is multinomial:

$$(N_1, N_2, \ldots, N_N | K = k) \sim \text{Multinomial}(k, (w_1, w_2, \ldots, w_N)).$$

□

As a result, the samples obtained through Poisson sampling are effectively analogous to sampling with replacement according to the distribution w. This method eliminates the necessity for assuming sub-Gaussianity of the vectors X_i, making the analysis more robust.

6.2 Theorem for Asymptotic Distribution of debiased estimator

The following theorem describes the asymptotic distribution of the debiased estimator $\hat{\gamma}$ derived from data obtained through Poisson sampling. This theorem is conceptually aligned with the results presented by Cai, Cai, and Guo [14], as well as those by Javanmard and Lee [13]. However, a significant distinction lies in its lack of assumptions regarding the covariates being sub-Gaussian vectors. This aspect is particularly relevant in the design setting, as enforcing the sub-Gaussianity property can be challenging when striving for optimal design. To address this issue, we utilize Poisson sampling in conjunction with an innovative proof technique that does not depend on the sub-Gaussianity of the design.

The asymptotic framework presented in the theorem follows a triangular array structure, wherein all parameters (such as N_n, p_n, s_n, M_n, Σ_n, etc.) are allowed to vary with n as it approaches infinity. Furthermore, in line with conventional practices in high-dimensional statistics, we simplify the notation by omitting the dependence on n, using N, p, s, M, and Σ in place of N_n, p_n, s_n, M_n, and Σ_n.

Before presenting our main theorem, we will first outline several relevant theorems and definitions necessary for its proof. We begin by noting the following theorem by Rudelson and Zhou [15], which establishes a connection between the restricted eigenvalues of random matrices and the corresponding (restricted) eigenvalues of the population covariance matrices. In this theorem, the smallest k-sparse eigenvalue of X is defined as

$$\rho_{\min}(k, X) = \min_{\substack{\|t\|_0 \le k \\ t \neq 0}} \frac{\|Xt\|_2}{\|t\|_2}.$$

Theorem 3. (Rudelson & Zhou[15]) Let $0 < \delta < 1$ and $0 < s_0 < p$. Let $X \in \mathbb{R}^p$ be a random vector such that $\|X\|_\infty \le M$ almost surely, and denote $\Sigma = \mathbb{E}XX^T$. Let $\mathbb{X}$ be an $n \times p$ matrix whose rows $X_1, X_2, \ldots, X_n$ are independent copies of X. Let Σ satisfy the restricted eigenvalue condition $\mathrm{RE}(3L, s_0, \Sigma^{1/2})$ as in Definition 1. Define

$$d = s_0 \left(1 + \max_j \|\Sigma^{1/2} e_j\|_2^2 \frac{16(3k_0)^2(3k_0+1)}{\delta^2 \cdot \phi_{\mathrm{RE}}^2(L, s_0, \Sigma^{1/2})}\right).$$

Assume that $d \le p$ and $\rho = \rho_{\min}(d, \Sigma^{1/2}) > 0$. Assume that the sample size n satisfies

$$n \ge n_0 := C_{\mathrm{RZ}} M^2 d \cdot \frac{\log p}{\rho^2 \delta^2} \cdot \log^3\left(C_{\mathrm{RZ}} M^2 d \cdot \frac{\log p}{\rho^2 \delta^2}\right),$$

for an absolute constant C_{RZ}. Then with probability at least $1 - \exp\left(\frac{-\delta\rho^2 n}{6M^2 d}\right)$, the restricted eigenvalue condition $\mathrm{RE}(L, s_0, \mathbb{X}/\sqrt{n})$ holds for matrix $\mathbb{X}/\sqrt{n}$ with

$$\phi_{\mathrm{RE}}(L, s_0, \mathbb{X}/\sqrt{n}) \ge (1 - \delta) \cdot \phi_{\mathrm{RE}}(L, s_0, \Sigma^{1/2}).$$

Note: We take L=3 for Lasso.

Proposition 1. Suppose that $K_1, \ldots, K_N$ are independent Poisson random variables with $K_j \sim \mathrm{Poisson}(nw_j)$ and $\sum_{j=1}^N w_j = 1$, so that $K := \sum_{j=1}^N K_j \sim \mathrm{Poisson}(n)$. Let $\mathbb{X}$ be a $K \times p$ matrix where x_j^T is repeated in the rows of $\mathbb{X}$ exactly K_j times. Suppose that

- The population covariance matrix $\Sigma = \sum_{j=1}^N w_j x_j x_j^T$ satisfies
$$\lambda_\star \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le \lambda^\star.$$
- The expected sample size n satisfies
$$n \ge \frac{5}{4}\tilde{n}_0 \quad \text{where} \quad \tilde{n}_0 = \frac{\tilde{C} M^2 \lambda^\star s_0 \log p}{\lambda_\star^2} \log^3\left(\frac{\tilde{C} M^2 \lambda^\star s_0 \log p}{\lambda_\star^2}\right),$$
and $\tilde{C} = 4 \times 51841 C_{\mathrm{RZ}}$.

Then with probability at least $1 - e^{-\tilde{n}_0/4} - e^{-n_0\lambda_\star/(12M^2 d)}$ we have

$$\phi_{\mathrm{RE}}(3, s_0, \mathbb{X}/\sqrt{K}) \ge \frac{1}{2}\phi_{\mathrm{RE}}(3, s_0, \Sigma^{1/2}).$$

Proof. see [16]. □

Theorem 4. (Conditional Central Limit Theorem) [17] Let $\{U_{n,k} : k = 1, \ldots, k_n \text{ and } n \in \mathbb{N}\}$ be an array of random variables, which are $\mathcal{A}_n$-independent (i.e., independent given $\mathcal{A}_n$) in each row (for some σ-algebra $\mathcal{A}_n \subseteq \mathcal{F}$, where $n \in \mathbb{N}$), and $\mathrm{Var}_{\mathcal{A}_n}(U_{n,k}) < \infty$ (almost surely) for $k = 1, \ldots, k_n, n \in \mathbb{N}$. Define $S_n := \sum_{i=1}^{N} U_{n,k}$. Assume that $\sigma^2_{n_{\mathcal{A}_n}} := \mathrm{Var}_{\mathcal{A}_n}(S_n) > 0$ (almost surely) for all large n. Then the two relations

$$\max_{k=1,\ldots,k_n} \frac{\mathrm{Var}_{\mathcal{A}_n}(U_{n,k})}{\sigma^2_{n_{\mathcal{A}_n}}} \xrightarrow{p} 0$$

and

$$\mathbb{E}_{\mathcal{A}_n}\left[\exp\left(it\frac{S_n - \mathbb{E}_{\mathcal{A}_n} S_n}{\sigma_{n_{\mathcal{A}_n}}}\right)\right] \xrightarrow{p} \exp\left(-\frac{t^2}{2}\right), \quad n \to \infty,$$

hold if and only if the $\mathcal{A}_n$-Lindeberg condition is satisfied in a weak form: for any $t > 0$,

$$T_n := \frac{1}{\sigma^2_{n_{\mathcal{A}_n}}} \sum_{i=1}^{k_n} \mathbb{E}_{\mathcal{A}_n}\left[(U_k - \mathbb{E}_{\mathcal{A}_n} U_k)^2 \mathbb{I}\{|U_k - \mathbb{E}_{\mathcal{A}_n} U_k| > t\sigma_{n_{\mathcal{A}_n}}\}\right] \xrightarrow{p} 0. \tag{A}$$

Furthermore, if the above $\mathcal{A}_n$-Lindeberg condition holds, then we have

$$\frac{S_n - \mathbb{E}_{\mathcal{A}_n} S_n}{\sigma_{n_{\mathcal{A}_n}}} \to Z \sim N(0,1), \quad \text{as } n \to \infty.$$

Theorem 5. (Main Theorem) [18] Suppose that $\mathbb{X} = (X_i^T)_{i=1}^K$ is a Poisson sample according to a distribution w on $(x_i)_{i=1}^N$ with $\mathbb{E}K = n$ and that $(y_i, X_i)_{i=1}^K$ follow **(Model 1.1)** with a s-sparse regression parameter β^*. Also assume that the following conditions are satisfied:

1. $\max_{1 \le i \le N} \|x_i\|_\infty \le M = o\left(\sqrt{\frac{n}{\log(p)}}\right)$.
2. $\lambda_\star I \preccurlyeq \Sigma_w := \sum_{i=1}^N w_i x_i x_i^T \preccurlyeq \lambda^\star I$ where $0 < \lambda_\star, \lambda^\star < \infty$ do not depend on n.
3. The noise terms ε_j are i.i.d. mean-zero sub-Gaussian random variables with $\|\varepsilon_j\|_{\psi_2} \le \sigma_\varepsilon$ and a variance $\mathbb{E}\varepsilon_n^2$ that is bounded away from 0 and ∞.
4. $s \log^{\frac{3}{2}}(p) = o(\sqrt{n})$.

Then the debiased estimate (6.1) satisfies:

$$\sqrt{n}(\hat{\gamma} - \gamma) = \frac{1}{\sqrt{n}} c^T \Sigma_w^{-1} \mathbb{X}^T \varepsilon + b_n,$$

where:

- **(Bias Bound)** As $n \to \infty$, with probability $1 - o(1)$, we have:

$$|b_n| \lesssim \frac{M\sigma_\varepsilon \sqrt{c^T \Sigma_w^{-1} c} \cdot s \log^{\frac{3}{2}}(p)}{\lambda_\star \sqrt{n}}.$$

- **(Asymptotic Normality)** Let $v^2 := (c^T \Sigma_w^{-1} \hat{\Sigma} \Sigma_w^{-1} c)\mathbb{E}\varepsilon_n^2$, where $\hat{\Sigma} = \frac{\sum_{i=1}^K X_i X_i^T}{n}$. Then:

$$\frac{1}{v\sqrt{n}} c^T \Sigma_w^{-1} \mathbb{X}^T \varepsilon \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty.$$

- **(Variance Approximation)** The variance of the noise term can be approximated (asymptotically) by $c^T \Sigma_w^{-1} c$:

$$\frac{c^T \Sigma_w^{-1} \hat{\Sigma} \Sigma_w^{-1} c}{c^T \Sigma_w^{-1} c} \xrightarrow{p} 1 \quad \text{as} \quad n \to \infty.$$

Proof. Since the size of the Poisson sample $K = \sum_{i=1}^{N} N_i$ is itself a Poisson random variable with mean n, it follows that

$$P(K = 0) = e^{-n} \to 0.$$

Therefore, in the subsequent analysis, we will implicitly assume that $K > 0$. More formally, the analysis is restricted to the event $[K > 0]$, which occurs with probability $1 - e^{-n}$.

Furthermore, we note that conditionally on $K = k$, the random variables $X_1, \ldots, X_k$ form an i.i.d. sample drawn from $(x_i)_{i=1}^N$ with weights $(w_i)_{i=1}^N$. This result follows from the well-known fact that, given $\sum_{i=1}^N N_i = k$, the distribution of $(N_i)_{i=1}^N$ is multinomial with parameters k and $(w_i)_{i=1}^N$.

We present the proof in three parts:

Part 1: Bias Bound

To begin, we observe that using weighted Lasso with weights

$$\hat{W}_j = \sqrt{\frac{1}{K}\sum_{i=1}^{K} X_{ij}^2} \quad \text{for } 1 \le j \le p$$

is equivalent to normalizing the columns of X before applying the Lasso method. This equivalence is crucial for establishing the performance of the Lasso estimator. Using the property 6.1.2(2), we first compute the expected value of $\hat{W}_j^2$ conditioned on $K = k$:

$$\mathbb{E}[\hat{W}_j^2 \mid K = k] = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} X_{ij}^2 \mid K = k\right] = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}X_{ij}^2 = \mathbb{E}X_{1j}^2 \quad (\because X_1, X_2, \ldots, X_k \text{ are i.i.d.}).$$

and The X_{ij} are drawn from a weighted distribution, so we need to compute the expectation of X_{ij}^2 under this distribution:

$$\mathbb{E}X_{1j}^2 = \sum_{i=1}^{N} w_i x_{ij}^2.$$

which is equivalent to the (j, j)-th element of the covariance matrix Σ_w:

$$\Sigma_{w,jj} = \sum_{i=1}^{N} w_i x_{ij}^2.$$

Thus,

$$\mathbb{E}[\hat{W}_j^2 \mid K = k] = \Sigma_{w,jj}$$

Using the law of total expectation, we have:

$$\mathbb{E}[\hat{W}_j^2] = \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{K} X_{ij}^2}{K} \mid K\right]\right] = \Sigma_{w,jj},$$

where $\Sigma_{w,jj}$ is the jj-th element of the population covariance matrix.

Since each $|X_{ij}|$ is bounded by M by assumption, we can apply Hoeffding's concentration inequality. For each j and every $t > 0$, we obtain:

$$\mathbb{P}\left(|\hat{W}_j^2 - \Sigma_{w,jj}| \ge t \mid K\right) \le 2\exp\left(-\frac{2Kt^2}{M^2}\right).$$

Setting $t = M\sqrt{\frac{\log(p)}{K}}$ in the above inequality gives:

$$\mathbb{P}\left(|\hat{W}_j^2 - \Sigma_{w,jj}| \ge M\sqrt{\frac{\log(p)}{K}} \mid K\right) \le 2\exp\left(-\frac{2K \cdot (M\sqrt{\frac{\log(p)}{K}})^2}{M^2}\right).$$

This simplifies to:

$$\mathbb{P}\left(|\hat{W}_j^2 - \Sigma_{w,jj}| \geq M\sqrt{\frac{\log(p)}{K}} \mid K\right) \leq 2p^{-2}.$$

Taking expectations on both sides with respect to the distribution of K:

$$\mathbb{P}\left(|\hat{W}_j^2 - \Sigma_{w,jj}| \geq M\sqrt{\frac{\log(p)}{K}}\right) \leq 2p^{-2}.$$

Now, applying the union bound over $j = 1, \ldots, p$ and noting that $\Sigma_{w,jj} \geq \lambda_\star > 0$ (by assumption), we obtain:

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{\hat{W}_j^2}{\Sigma_{w,jj}} - 1\right| \geq \frac{M}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\right) \leq 2p^{-1} \to 0.$$

Using property 6.1.2(1), we note that $\frac{K}{n} \to_p 1$ since $\text{Var}\left[\frac{K}{n}\right] = n^{-1} \to 0$. By assumption, we have $M\sqrt{\log(p)} = o(\sqrt{n})$. Thus:

$$M\sqrt{\frac{\log(p)}{K}} \to 0.$$

Given that $\lambda_\star$ is bounded away from zero by assumption, we conclude that:

$$\max_{1\leq j\leq p}\left|\frac{\hat{W}_j^2}{\Sigma_{w,jj}} - 1\right| \to_p 0 \text{ as } n \to \infty.$$

Finally, we have established that, with high probability, the weights $\hat{W}_j$ are bounded away from 0 and ∞. This ensures that the standard (unweighted) Lasso guarantees can be applied, as detailed in [19].

Let $\hat{w}_i = \frac{N_i}{n}$, where N_i are obtained using Poisson sampling. Then the debiased lasso estimator can be written as

$$\hat{\gamma} = \langle c, \hat{\beta}\rangle + u^T\hat{\Sigma}(\beta - \hat{\beta}) + \frac{1}{n}u^T\mathbb{X}^T\varepsilon$$

$$= \gamma + c^T\left(\Sigma_w^{-1}\hat{\Sigma} - I\right)(\beta - \hat{\beta}) + \frac{1}{n}c^T\Sigma_w^{-1}\mathbb{X}^T\varepsilon.$$

Subtracting γ from both sides and multiplying by $\sqrt{n}$, we obtain

$$\sqrt{n}(\hat{\gamma} - \gamma) = \sqrt{n}c^T\left(\Sigma_w^{-1}\hat{\Sigma} - I\right)(\beta - \hat{\beta}) + \frac{1}{\sqrt{n}}c^T\Sigma_w^{-1}\mathbb{X}^T\varepsilon.$$

we will show that the $1st$ quanitity of above is o(1) in probability.
Using Holder's inequality,

$$\sqrt{n}\left|c^T\left(\Sigma_w^{-1}\hat{\Sigma} - I\right)(\beta - \hat{\beta})\right| \leq \sqrt{n}\|c^T\left(\Sigma_w^{-1}\hat{\Sigma} - I\right)\|_\infty \cdot \|\hat{\beta} - \beta\|_1.$$

Now, we will find an upper bound for these two norms. So, we observe that the eigenvalues of the population covariance matrix Σ are bounded, as $0 < \lambda_\star \leq \lambda_{\min}(\Sigma_w) \leq \lambda_{\max}(\Sigma_w) \leq \lambda^\star < \infty$. Additionally, since it is assumed that $\sqrt{n} \gg s\log^{3/2}(p)$, Proposition 1 ensures that the scaled design matrix $\mathbb{X}/\sqrt{K}$ satisfies the Restricted Eigenvalue (RE) condition with a restricted eigenvalue greater than $\sqrt{\lambda_\star}/2$ with probability approaching 1, where K is the number of samples drawn through Poisson sampling.

Let G denote the event that the RE condition holds, i.e.,

$$\phi_{\text{RE}}(3, s, \mathbb{X}/\sqrt{K}) \geq \frac{\sqrt{\lambda_\star}}{2}.$$

Then, we have $\mathbb{P}(G) = 1 - o(1)$, meaning that with high probability, the matrix $\mathbb{X}/\sqrt{K}$ will meet the required RE condition. Given the event G, we have

$$\mathbb{P}\left(\|\hat{\beta}-\beta\|_1 \lesssim \frac{\sigma_\epsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\,\middle|\, G\right) = 1-o(1).$$

using a similar proof of Theorem 1 part (B). It follows that

$$\mathbb{P}\left(\|\hat{\beta}-\beta\|_1 \lesssim \frac{\sigma_\epsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\right) \geq \mathbb{P}\left(\|\hat{\beta}-\beta\|_1 \lesssim \frac{\sigma_\epsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\,\middle|\, G\right)\cdot P(G)$$

$$= (1-o(1))\cdot(1-o(1)) = 1-o(1).$$

Next, since $K/n \xrightarrow{p} 1$ using property 6.1.2(1), we can substitute n for K in the above bound and write

$$\mathbb{P}\left(\|\hat{\beta}-\beta\|_1 \lesssim \frac{\sigma_\epsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{n}}\right) \xrightarrow{p} 1. \tag{1}$$

Let $\hat{w}_i = \frac{N_i}{n}$ and note that

$$c^T\left(\Sigma_w^{-1}\hat{\Sigma}-I\right) = \sum_{i=1}^N (\hat{w}_i - w_i)\, c^T\Sigma_w^{-1}x_i x_i^T.$$

Recall that N_i is a Poisson R.V. with mean nw_i so, MGF of N_i, $M_{N_i}(t) = \mathbb{E}e^{tN_i} = \exp(nw_i(e^t-1))$ and hence

$$\mathbb{E}\left[\exp\left(t(N_i-nw_i)\right)\right] = \exp\left(nw_i\left(e^t-1-t\right)\right) \leq \exp\left(nw_i e^{|t|}\frac{t^2}{2}\right) \leq \exp\left(nw_i t^2\right) \quad \text{for } |t|\leq\frac{1}{2}$$

So, Using Proposition 2.7.1 and Definition 2.7.5 of [20] we can say that $N_i - nw_i$ is sub-exponential and thus $\|N_i-nw_i\|_{\psi_1} \lesssim \sqrt{nw_i}$, and therefore $\|\hat{w}_i - w_i\|_{\psi_1} \lesssim \sqrt{w_i/n}$.

Define $V_{ij} = (\hat{w}_i - w_i)c^T\Sigma_w^{-1}x_i x_{ij}$. Using Bernstein's inequality for sub-exponential random variables [20, Theorem 2.8.1], for some absolute constant $b>0$ and all $j=1,\ldots,p$, we have:

$$\mathbb{P}\left(\left|\sum_{i=1}^N V_{ij}\right| > t\right) \leq 2\exp\left(-b\min\left\{\frac{t^2}{\sum_{i=1}^N\|V_{ij}\|_{\psi_1}^2},\ \frac{t}{\max_i\|V_{ij}\|_{\psi_1}}\right\}\right). \tag{2}$$

Using the bound $\max_{i,j}|x_{ij}| \leq M$, we have:

$$\sum_i \|V_{ij}\|_{\psi_1}^2 \leq \frac{M^2}{n}\sum_i w_i(c^T\Sigma_w^{-1}x_i)^2 = \frac{M^2}{n}c^T\Sigma_w^{-1}\left(\sum_i w_i x_i x_i^T\right)\Sigma_w^{-1}c = \frac{M^2}{n}c^T\Sigma_w^{-1}c.$$

Similarly,

$$\max_i\|V_{ij}\|_{\psi_1} \leq M\cdot\max_i\sqrt{\frac{w_i}{n}}|c^T\Sigma_w^{-1}x_i| \leq M\sqrt{\frac{c^T\Sigma_w^{-1}c}{n}}.$$

Using these bounds and for $t=\sqrt{\frac{2\log(p)}{nb}}$, the Bernstein bound (2) implies:

$$\mathbb{P}\left(\left|\sum_{i=1}^N V_{ij}\right| > \frac{2M\log(p)}{b}\sqrt{\frac{c^T\Sigma_w^{-1}c}{n}}\right) \leq 2\exp\left(-\min\left\{\frac{4\log^2(p)}{b},\ 2\log(p)\right\}\right).$$

For $p > e^{b/2}$, the exponential tail prevails, and we obtain:

$$\mathbb{P}\left(\left|\sum_{i=1}^N V_{ij}\right| > \frac{2M\log(p)}{b}\sqrt{\frac{c^T\Sigma_w^{-1}c}{n}}\right) \leq 2p^{-2}, \quad \text{for all } j=1,\ldots,p.$$

A union bound over all $j=1,\ldots,p$ now yields:

$$\mathbb{P}\left(\max_{1\le j\le p}\left|\sum_{i=1}^{N} V_{ij}\right| > \frac{2M\log(p)}{b}\sqrt{\frac{c^T\Sigma_w^{-1}c}{n}}\right) \le 2p^{-1}, \quad \text{for } p > e^{b/2}. \tag{3}$$

Using the upper bound (3) and the error rate of the lasso estimate (1), with probability $1 - o(1)$, we have:

$$\sqrt{n}\left|c^T\left(\Sigma_w^{-1}\hat{\Sigma} - I\right)(\beta - \hat{\beta})\right| \le \sqrt{n}\|c^T\left(\Sigma_w^{-1}\hat{\Sigma} - I\right)\|_\infty \cdot \|\hat{\beta} - \beta\|_1.$$

$$\lesssim \sqrt{n}\cdot\left(M\log(p)\sqrt{\frac{c^T\Sigma_w^{-1}c}{n}}\right)\cdot\frac{\sigma_\epsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{n}}$$

$$\lesssim \frac{M\sigma_\epsilon\sqrt{c^T\Sigma_w^{-1}c}}{\lambda_\star}\cdot\frac{s\log^{3/2}(p)}{\sqrt{n}}.$$

Part 2: Variance approximation

Next, we first prove the third part of the theorem, as the argument used here will be useful in the proof of asymptotic normality [18]. The conditional variance of the noise term is given by:

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}c^T\Sigma_w^{-1}\mathbb{X}^T\varepsilon \,|\, \mathbb{X}\right) = c^T\Sigma_w^{-1}\hat{\Sigma}\Sigma_w^{-1}c.$$

We will show that this variance can be approximated by $c^T\Sigma_w^{-1}c$, i.e.,

$$\frac{c^T\Sigma_w^{-1}\hat{\Sigma}\Sigma_w^{-1}c}{c^T\Sigma_w^{-1}c} \xrightarrow{p} 1.$$

Assume $N = |\{i : w_i \neq 0\}|$; otherwise, we can discard the zero weights. We want to demonstrate:

$$A := \frac{\sum_i w_i(c^T\Sigma_w^{-1}x_i)^4}{(c^T\Sigma_w^{-1}c)^2} \le \frac{1}{n}.$$

Let $d_i = \Sigma_w^{-1/2}x_i$ and $v = \Sigma_w^{-1/2}c$. Then A can be rewritten as:

$$A = \sum_i w_i(d_i^T v)^4\frac{1}{n\|v\|_2^4}.$$

Step 1. First, suppose that $N = p$. We have:

$$\sum_i w_i d_i d_i^T = \Sigma_w^{-1/2}\left(\sum_i w_i x_i x_i^T\right)\Sigma_w^{-1/2} = I_p.$$

Let d_j^* be the projection of d_j on the orthogonal complement of the span of $\{d_i \,|\, i \neq j\}$. Multiplying both sides by d_j^* yields:

$$w_j(d_j^T d_j^*)d_j = d_j^*.$$

Note that $d_j^T d_j^* \neq 0$ for all $j = 1, \ldots, p$, as $d_1, \ldots, d_p$ form a basis for $\mathbb{R}^p$ (since Σ_w is non-singular by construction). From this equation, it follows that $d_1, \ldots, d_p$ are orthogonal, and multiplying both sides by d_j results in:

$$\|d_j\|_2^2 = w_j^{-1}.$$

Since A does not depend on $\|v\|_2$, we have:

$$A \le \max_{u\neq 0}\sum_i w_i(d_i^T u)^4\frac{1}{n\|u\|_2^4} = \frac{1}{n}\max_{\|u\|_2^2=1}\sum_i w_i(d_i^T u)^4.$$

The Lagrangian for the last maximization problem is:

$$L(u,\lambda) = \sum_i w_i (d_i^T u)^4 - 2\lambda(u^T u - 1).$$

Taking the derivative with respect to u and setting it to zero yields:

$$\sum_i w_i (d_i^T \hat{u})^3 d_i = \lambda \hat{u}.$$

Changing variables to $\tilde{u} = \sqrt{1/\lambda}\hat{u}$, we can rewrite this as:

$$\sum_i w_i (d_i^T \tilde{u})^3 d_i = \tilde{u}.$$

Multiplying on the left once by $\tilde{u}^T$ and once by d_j^T gives:

$$\sum_i w_i (d_i^T \tilde{u})^4 = \tilde{u}^T \tilde{u} \quad \text{and} \quad (d_j^T \tilde{u})^2 = 1.$$

Note that $\sum_i w_i (d_i^T u)^4 / \|u\|_2^4$ does not depend on the norm of u, so any nonzero multiple of $\hat{u}$, and in particular $\tilde{u}$, is a maximizer. Plugging $\tilde{u}$ into this expression and using the previous result gives:

$$A \leq \frac{\sum_i w_i (d_i^T \tilde{u})^4}{n(\tilde{u}^T \tilde{u})^2} \leq \frac{1}{n}.$$

Step 2. Now consider the $N > p$ case. The idea is to reduce this case to the $N = p$ case by appropriately extending the length of $d_i, v \in \mathbb{R}^p$ to N.

Once again, below identity

$$\sum_i w_i d_i d_i^T = \Sigma_w^{-1/2} \left(\sum_i w_i x_i x_i^T \right) \Sigma_w^{-1/2} = I_p.$$

will hold. We define the vectors $\tilde{d}_i \in \mathbb{R}^N$ by a tuple $\tilde{d}_i^T = (d_i^T, f_i^T)$ for some vectors $f_i \in \mathbb{R}^{N-p}$ such that

$$\sum_i w_i \tilde{d}_i \tilde{d}_i^T = I_N.$$

A construction of these f_i's is provided in the Appendix .1. For any $u \in \mathbb{R}^N$, we used a similar decomposition $u^T = (u_1^T, u_2^T)$ with $u_1 \in \mathbb{R}^p$ and $u_2 \in \mathbb{R}^{N-p}$. Then:

$$n \cdot A \leq \max_{\|v\|_2^2 = 1} \sum_i w_i (d_i^T v)^4 = \max_{\|u\|_2^2 = 1, u_2 = 0} \sum_i w_i (\tilde{d}_i^T u)^4 \leq \max_{\|u\|_2^2 = 1} \sum_i w_i (\tilde{d}_i^T u)^4 \leq 1,$$

where the last inequality follows from the argument in Step 1 (since now $\tilde{p} := \dim(\tilde{d}_i) = N$).

Part 3. (Asymptotic Normality)

Let $\mathcal{A}$ be the σ-algebra generated by $(N_i)_{i=1}^N$ (note that we are suppressing the dependence of $\mathcal{A} = \mathcal{A}_n$ on n to lighten notation). As in Theorem 4 and [16] we define

$$U_{n,k} := \begin{cases} \frac{1}{\sqrt{n}} c^T \Sigma_w^{-1} x_k (\varepsilon_1^{(k)} + \cdots + \varepsilon_{N_k}^{(k)}), & N_k > 0 \\ 0, & N_k = 0. \end{cases}$$

Let

$$S_n := \sum_{i=1}^N U_{n,k}, \quad \sigma_{n_{\mathcal{A}}}^2 := \mathbb{E}_{\mathcal{A}} (S_n - \mathbb{E}_{\mathcal{A}} S_n)^2,$$

where $\varepsilon_i^{(j)} \sim \varepsilon_n$ are i.i.d. mean-zero random variables with variance equal to $E[\varepsilon_n^2] = \sigma^2$ and sub-Gaussian norm $\|\varepsilon_i^{(j)}\|_{\psi_2} \leq \sigma_\varepsilon$. In general, we have $\sigma \leq \sigma_\varepsilon$. Since $\mathbb{E}_{\mathcal{A}}[S_n] = 0$, the conditional variance of S_n can be written as:

$$(\sigma^2_{n_{\mathcal{A}}} = \mathbb{E}_{\mathcal{A}}\left[S_n^2\right] = \frac{1}{n}\sum_{k=1}^{N}\left(c^T\Sigma_w^{-1}x_k\right)^2 \mathbb{E}\left[\left(\sum_{i=1}^{N_k}\varepsilon_i^{(k)}\right)^2 \middle| N_k\right].$$

This simplifies to:

$$= \frac{1}{n}\sum_{k=1}^{N}\left(c^T\Sigma_w^{-1}x_k\right)^2 N_k\sigma^2$$

and thus we can express it as:

$$= \sigma^2 \cdot c^T\Sigma_w^{-1}\hat{\Sigma}\Sigma_w^{-1}c.$$

Thus, we want to show that:

$$\frac{c^T\Sigma_w^{-1}\mathbb{X}^T\varepsilon}{\sqrt{n\sigma^2 \cdot c^T\Sigma_w^{-1}\hat{\Sigma}\Sigma_w^{-1}c}} = \frac{S_n - \mathbb{E}_{\mathcal{A}}[S_n]}{\sigma_{n_{\mathcal{A}}}} \to Z \sim N(0,1), \quad \text{as } n \to \infty.$$

using Theorem 4, we just need to prove that the conditional Lindeberg condition A is satisfied. So, we will find appropriate upper bounds on the terms in the Lindeberg condition. Additionally, since the noise variance σ^2 is bounded away from 0 and ∞, we can, without loss of generality, set $\sigma^2 = 1$ for subsequent analysis. Using the Cauchy-Schwarz inequality:

$$\mathbb{E}_{\mathcal{A}}[U_{n,k}^2\mathbb{I}\{|U_{n,k}| > t\sigma_{n_{\mathcal{A}}}\}] \le \left(\mathbb{E}_{\mathcal{A}}U_{n,k}^4\mathbb{E}_{\mathcal{A}}\mathbb{I}\{|U_{n,k}| > t\sigma_{n_{\mathcal{A}}}\}\right)^{\frac{1}{2}} \le \left(\frac{\mathbb{E}_{\mathcal{A}}U_{n,k}^4 \cdot \mathbb{E}_{\mathcal{A}}U_{n,k}^2}{t^2\sigma^2_{n_{\mathcal{A}}}}\right)^{\frac{1}{2}}. \tag{4}$$

The fourth moment of $U_{n,k}$ can be bounded as follows:

$$\mathbb{E}_{\mathcal{A}}U_{n,k}^4 = \frac{1}{n^2}(c^T\Sigma_w^{-1}x_k)^4\mathbb{E}_{\mathcal{A}}\left(\varepsilon_1^{(k)} + \cdots + \varepsilon_{N_k}^{(k)}\right)^4 = \frac{1}{n^2}(c^T\Sigma_w^{-1}x_k)^4 \cdot \left(N_k\mathbb{E}\varepsilon^4 + N_k(N_k-1)(\mathbb{E}\varepsilon^2)^2\right)$$

$$\le \frac{C}{n^2}(c^T\Sigma_w^{-1}x_k)^4 \cdot N_k^2\sigma_\epsilon^4,$$

where the last inequality follows since $N_k \le N_k^2$, and by the sub-Gaussianity of ε, we have $\mathbb{E}\varepsilon^4 \le C \cdot \sigma_\varepsilon^4$ for an absolute constant $C > 0$.

Also, $\mathbb{E}_{\mathcal{A}}U_{n,k}^2 = \frac{N_k}{n}(c^T\Sigma_w^{-1}x_k)^2\sigma_\varepsilon^2$ and we have assumed $\sigma^2 = 1$, thus the upper bound for (4) is given by:

$$\mathbb{E}_{\mathcal{A}}[U_{n,k}^2\mathbb{I}\{|U_{n,k}| > t\sigma_{n_{\mathcal{A}}}\}] \le \left(\frac{C}{t^2\sigma^2_{n_{\mathcal{A}}}}\left(\frac{N_k}{n}\right)^3 \cdot (c^T\Sigma_w^{-1}x_k)^6\right)^{\frac{1}{2}}$$

$$\le \frac{1}{\sigma_{n_{\mathcal{A}}}}\left(\frac{C}{t^2}\sum_{i=1}^{N}\frac{N_i}{n}\right)^{\frac{1}{2}}\left(\frac{N_k}{n}\right)\left|c^T\Sigma_w^{-1}x_k\right|^3.$$

Therefore, the expression in the Lindeberg condition A has the following upper bound:

$$T_n \le \left(\frac{C}{t^2}\sum_{i=1}^{N}\frac{N_i}{n}\right)^{\frac{1}{2}}\left(\frac{(c^T\Sigma_w^{-1}c)}{\sigma^2_{n_{\mathcal{A}}}}\right)^{\frac{3}{2}} \cdot \frac{1}{(c^T\Sigma_w^{-1}c)^{3/2}}\sum_{k=1}^{N}\left|c^T\Sigma_w^{-1}x_k\right|^3\hat{w}_k.$$

Since, as we have shown before in 6.1.2 and in 6.2, $\sum_i N_i/n \xrightarrow{p} 1$ and $\sigma^2_{n_{\mathcal{A}}}/(c^T\Sigma_w^{-1}c) \xrightarrow{p} 1$, it suffices to show that:

$$\frac{1}{(c^T\Sigma_w^{-1}c)^{3/2}}\sum_{k=1}^{N}\left|c^T\Sigma_w^{-1}x_k\right|^3\hat{w}_k \xrightarrow{p} 0.$$

We will show the variance of this term converges to zero:

$$\frac{1}{n \cdot (c^T \Sigma_w^{-1} c)^3} \sum_{k=1}^{N} \left(c^T \Sigma_w^{-1} x_k\right)^6 w_k \xrightarrow{p} 0.$$

A similar technique as in 6.2 has been used here. Let $d_k = \Sigma_w^{-1/2} x_k$ and $v = \Sigma_w^{-1/2} c$. The variance can be rewritten as:

$$\frac{1}{n\|v\|_2^6} \sum_k w_k (d_k^T v)^6 \leq \frac{1}{n} \cdot \max_{\|u\|_2=1} \left\{ \sum_k w_k (d_k^T u)^6 \right\} \leq \frac{1}{n} \xrightarrow{p} 0, \text{ as } n \to \infty. \text{16}$$

this completes the proof of the theorem. □

Chapter 7

Discussion

7.1 Assumptions on the Analysis

The assumptions made in our analysis are integral to ensuring that our methods for inference and optimization in high-dimensional settings are robust, computationally feasible, and statistically accurate. These assumptions address both the statistical properties of the data and the computational complexity of the optimization problem, particularly in the context of Lasso regression and debiased estimators.

7.1.1 Uniform Bound on Design Points

The first assumption imposes a uniform bound on the coordinates of all design points. This assumption is crucial because, in high-dimensional settings, the behavior of the covariate vectors can significantly affect the convergence properties of estimators like Lasso.

Specifically, if the design points $\{x_i\}_{i=1}^N$ are sampled from a population where the covariate coordinates x_{ij} are sub-Gaussian, with a sub-Gaussian norm $\|x_{ij}\|_{\psi_2} = O(1)$, it is well-known that the maximum of the absolute values $\max_{i,j} |x_{ij}|$ behaves as $\log(Np)$ with high probability. In this case, we can set $M \asymp \log(Np)$ and adjust the sparsity constraint accordingly. The modified sparsity condition becomes

$$s \log^{3/2}(p) \sqrt{\log(Np)} = o(\sqrt{n}),$$

which ensures that the high-dimensional inference remains valid under these relaxed conditions.

7.1.2 Restricted Eigenvalue Condition

The second assumption ensures that the restricted eigenvalue (RE) condition holds with high probability for the sample design matrix. The RE condition is crucial for establishing the fast rate of convergence for the Lasso estimator, as it guarantees that the sample design matrix behaves sufficiently well to recover the sparse signal.

Furthermore, the constraint in this assumption is semi-definite in nature, which significantly simplifies the computational aspect of the optimization. Computing restricted eigenvalues directly is known to be NP-hard in general, but by ensuring the population covariance matrix satisfies the RE condition indirectly, we avoid this computational bottleneck. Additionally, this assumption implies that $c \in \text{span}(\{x_i : w_i > 0\})$, which is a necessary condition for unbiasedness in low-dimensional settings. Without this assumption, the population covariance matrix Σ could have a vanishing restricted eigenvalue, resulting in an inconsistent debiased estimator.

7.1.3 Sub-Gaussianity of the Noise

The third assumption, which assumes that the noise term is sub-Gaussian, is standard in high-dimensional linear models. This assumption allows us to apply concentration inequalities, which are fundamental for deriving accurate confidence intervals and hypothesis tests in high-dimensional settings. Without the sub-Gaussianity assumption, the noise distribution could exhibit heavy tails, complicating the inference and potentially leading to suboptimal statistical properties.

7.1.4 Sparsity of the Regression Parameter

The final assumption imposes a constraint on the sparsity of the regression parameter β. This is a stronger condition than the typical "ultra-sparsity" conditions commonly assumed in high-dimensional inference literature. The additional factor of $\log(p)$ in this assumption is the trade-off for relaxing the usual assumption of sub-Gaussian covariates. Here, we only assume uniformly bounded entries, which is more general. This stronger sparsity assumption compensates for the relaxation in the covariate structure and ensures that the Lasso estimator retains good performance even when the covariates are not sub-Gaussian.

7.2 Computation Challenges

We will explore a natural definition of c-optimality for deterministic experiments. Let $N_i \in \{0, 1, \ldots, n\}$ represent the frequency of the x_i in the sample. Then, the variance of the debiased Lasso estimator can be approximated as

$$c^T \Sigma_{rel} \hat{\Sigma} \Sigma_{rel} c,$$

where $\hat{\Sigma} = n^{-1} \sum_{i=1}^{N} N_i x_i x_i^T$, and Σ_{rel} serves as a relaxed inverse of $\hat{\Sigma}$. This relaxed inverse typically needs to satisfy the condition

$$\|c^T \Sigma_{rel} \hat{\Sigma} - c^T\|_\infty \lesssim \sqrt{\frac{\log(p)}{n}}$$

to ensure that the bias of the debiased Lasso estimator remains $o\left(\frac{1}{\sqrt{n}}\right)$. Consequently, it becomes natural to minimize the variance described above while adhering to this constraint:

$$\text{P3:} \quad \min_{(N_i)_{i=1}^N \in [n]^N} \quad c^T \Sigma_{rel} \hat{\Sigma} \Sigma_{rel} c \quad \text{s.t. } \|c^T \Sigma_{rel} \hat{\Sigma} - c^T\|_\infty \lesssim \sqrt{\frac{\log(p)}{n}},$$

with

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{N} N_i x_i x_i^T \quad \text{and} \quad \sum_{i=1}^{N} N_i \leq n.$$

7.2.1 Computationally infeasibility

The above optimization problem, which aims to minimize the variance of the debiased Lasso estimator, is a non-convex problem set in a high-dimensional space. This non-convexity significantly complicates the search for global optima. Specifically, the design problem entails selecting the weights N_i (the number of times each design point x_i is included) in such a way that the estimator's variance is minimized. The dimensionality of the problem is $O(N+p^2)$, where N is the number of design points and p is the dimension of covariates. Even when the integer constraints on N_i are relaxed, addressing such a large-scale non-convex problem becomes computationally intractable for most practical applications.

7.2.2 Statistical Inaccuracy

Even if a locally optimal solution to the design problem is found, it is not guaranteed that the solution will satisfy the RE condition. Ensuring that the RE condition holds for a deterministic design is NP-hard. This makes it difficult to guarantee the statistical accuracy of the resulting design.

7.3 Advantages of Randomization

To overcome these challenges, randomization offers a practical solution. By randomizing the design, we can approximate the non-convex objective with a convex relaxation. Additionally, randomization helps ensure that the RE condition holds with high probability, thus preserving the statistical properties of the estimator.

7.3.1 Convex Relaxtion

As we have shown in Theorem 5 variance approximation, Randomization allows us to approximate the objective function

$$c^T \Sigma_{rel} \hat{\Sigma} \Sigma_{rel} c$$

with the simpler expression

$$c^T \Sigma_{rel} c,$$

where $\Sigma_{rel} := \Sigma_w^{-1}$. This approximation reduces the computational burden by transforming the original problem into a convex optimization problem given by:

$$\textbf{P4:} \quad \min_{w_1,\ldots,w_N} \quad c^T \Sigma_w^{-1} c$$

subject to

$$\Sigma_w = \sum_{i=1}^{N} w_i x_i x_i^T, \quad \lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star I, \quad \sum_{i=1}^{N} w_i = 1, \quad w_i \geq 0.$$

7.3.2 Why the LP Formulation is Challenging

Although P4 is technically a linear program (LP), it involves a matrix inversion term Σ_w^{-1}, complicating the optimization process. Additionally, while the problem aims to control the eigenvalues of Σ_w, solving a linear program with eigenvalue constraints is not straightforward.

Inversion of Covariance Matrix

The term $c^T \Sigma_w^{-1} c$ requires the inversion of Σ_w, which can be computationally expensive and numerically unstable in high dimensions, especially if Σ_w has eigenvalues close to zero.

Eigenvalue Constraints

Directly managing eigenvalue bounds in an LP setup is challenging. Ensuring that the population covariance matrix Σ_w satisfies the eigenvalue constraints $\lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star I$ while keeping the problem tractable is particularly difficult in high dimensions.

7.3.3 Transformation of P4 to a SDP P5

To address these challenges, the original LP problem is recast as a semidefinite program (SDP) in P5. SDPs facilitate handling eigenvalue constraints in a more tractable manner by employing matrix inequalities, which are standard in the SDP framework.

The transformed problem P5 is as follows:

$$\textbf{P5:} \quad \min_{t \in \mathbb{R}, w_1,\ldots,w_N} \quad t$$

subject to

$$\Sigma_w = \sum_{i=1}^{N} w_i x_i x_i^T, \quad \lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star I,$$

$$\begin{bmatrix} t & c^T \\ c & \Sigma_w \end{bmatrix} \succcurlyeq 0, \quad \sum_{i=1}^{N} w_i = 1, \quad w_i \geq 0.$$

This SDP formulation includes:

Matrix Inequalities

The eigenvalue constraints $\lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star I$ are enforced through linear matrix inequalities (LMIs). These constraints ensure that the eigenvalues of Σ_w remain bounded between $\lambda_\star$ and $\lambda^\star$, thereby preventing the covariance matrix from becoming ill-conditioned.

Convexity

The SDP formulation transforms the objective into a convex problem. By introducing a scalar variable t and the LMI

$$\begin{bmatrix} t & c^T \\ c & \Sigma_w \end{bmatrix} \succcurlyeq 0,$$

we avoid the direct inversion of Σ_w. This matrix inequality ensures that $t \geq c^T \Sigma_w^{-1} c$, providing a convex relaxation of the original objective function.

7.3.4 Why the SDP Formulation is Preferable

Handling Eigenvalue Constraints

SDPs are particularly suitable for problems that involve eigenvalue constraints, as they can enforce these constraints through linear matrix inequalities (LMIs). In high dimensions, controlling the spectrum of Σ is essential to ensure that the design matrix behaves well, and the eigenvalue bounds $\lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star I$ are naturally expressed as LMIs in the SDP framework.

Computational Feasibility

Transitioning to an SDP avoids the computational difficulties associated with inverting large matrices. Instead, the inversion is implicitly managed through the LMI

$$\begin{bmatrix} t & c^T \\ c & \Sigma_w \end{bmatrix} \succcurlyeq 0,$$

which ensures that the matrix Σ_w has well-behaved eigenvalues without requiring explicit inversion.

Convex Optimization

SDPs are convex optimization problems, and efficient solvers are available to find global optima for SDPs. This represents a significant advantage over the original LP formulation, which, due to matrix inversion and eigenvalue constraints, could be numerically unstable or difficult to solve in high-dimensional settings.

Statistical Guarantees

The SDP formulation helps ensure that the sample covariance matrix Σ_w satisfies the restricted eigenvalue (RE) condition with high probability. This is important because, as noted earlier, the RE condition is necessary for the consistency and accuracy of the Lasso estimator. In high-dimensional settings, randomization combined with the SDP formulation guarantees that the RE condition holds with high probability, whereas verifying this condition for a deterministic design would be NP-hard.

Bibliography

[1] Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23 (2):255–262, 1952.

[2] R. Harman and T. Jurík. Computing c-optimal experimental designs using the simplex method of linear programming. *Computational Statistics & Data Analysis*, 53:247–254, 2008. doi: 10.1016/j.csda.2008.06.023.

[3] Herman Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24(4):586–602, 1953.

[4] Friedrich Pukelsheim. *Optimal Design of Experiments*. SIAM, 2006.

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

[6] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[7] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On the asymptotic theory for high-dimensional models with penalized estimators: uniform control of approximate sparsity. *The Annals of Statistics*, 42(2):1166–1202, 2014.

[8] Matthias Seeger. Bayesian inference and optimal design for the sparse linear model. In *Proceedings of the 25th international conference on Machine learning*, pages 912–919. ACM, 2008.

[9] Ramya Ravi, Katya Scheinberg, and Shashanka Ubaru. Sparse d-optimal designs. *Advances in Neural Information Processing Systems*, 29, 2016.

[10] Wei Huang, John Lafferty, and Larry Wasserman. Optimal experimental designs for debiased lasso estimators. *Electronic Journal of Statistics*, 14(1):1207–1242, 2020.

[11] Michal Černý and Milan Hladík. Two complexity results on c-optimality in experimental design. *Computational Optimization and Applications*, 51(3):1397–1408, 2012.

[12] Friedrich Pukelsheim and Sabine Rieder. Efficient rounding of approximate designs. *Biometrika*, 79(4): 763–770, 1992.

[13] Adel Javanmard and Jason D Lee. A flexible framework for hypothesis testing in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):685–718, 2020.

[14] Tianxi Cai, T. Tony Cai, and Zijian Guo. Optimal Statistical Inference for Individualized Treatment Effects in High-Dimensional Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4): 669–719, 08 2021. ISSN 1369-7412. doi: 10.1111/rssb.12426. URL https://doi.org/10.1111/rssb.12426.

[15] Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1, 2012.

[16] Hamid Eftekhari, Moulinath Banerjee, and Ya'acov Ritov. Supplement to "design of c-optimal experiments for high-dimensional linear models", 2023.

[17] A. Bulinski. Conditional central limit theorem. *Theory of Probability and Its Applications*, 61:613–631, 2017.

[18] Hamid Eftekhari, Moulinath Banerjee, and Ya'acov Ritov. Design of c-optimal experiments for high-dimensional linear models. *Bernoulli*, 29(1):652–668, February 2023.

[19] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, Berlin, Heidelberg, 2011. ISBN 978-3-642-20192-9.

[20] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

.1 Construction of f_i's:

Step 1: Orthogonal Decomposition

Consider the vectors $d_1, d_2, \ldots, d_p \in \mathbb{R}^p$, which span a subspace $\mathcal{D}$ of dimension p within $\mathbb{R}^N$. Since $N > p$, there exists an orthogonal complement $\mathcal{D}^\perp \subset \mathbb{R}^N$, which is a subspace of dimension $N - p$.

Our goal is to extend the orthonormal basis $\{d_1, d_2, \ldots, d_p\}$ of $\mathcal{D}$ to an orthonormal basis of the entire space $\mathbb{R}^N$. To achieve this, we must find vectors $f_{p+1}, f_{p+2}, \ldots, f_N$ that span $\mathcal{D}^\perp$ and satisfy the required orthonormality condition.

Step 2: Projection onto $\mathcal{D}^\perp$

Given that $d_1, d_2, \ldots, d_p$ form an orthonormal set with respect to the weighted inner product $\langle \cdot, \cdot \rangle_w$, we can project any vector $v \in \mathbb{R}^N$ onto the subspace $\mathcal{D}^\perp$ by removing its components along $d_1, \ldots, d_p$.

Let $P_{\mathcal{D}}$ denote the projection operator onto $\mathcal{D}$. The projection of any vector $v \in \mathbb{R}^N$ onto $\mathcal{D}^\perp$ is given by:

$$v_{\mathcal{D}^\perp} = v - P_{\mathcal{D}}(v).$$

Step 3: Constructing the f_i's

We now construct the vectors $f_{p+1}, \ldots, f_N$ by selecting arbitrary linearly independent vectors in $\mathbb{R}^N$ and projecting them onto $\mathcal{D}^\perp$. Specifically, for each $i \in \{p+1, \ldots, N\}$, choose a vector $v_i \in \mathbb{R}^N$, and define:

$$f_i = v_{i,\mathcal{D}^\perp} = v_i - \sum_{j=1}^{p} \langle v_i, d_j \rangle_w d_j.$$

The resulting vectors $f_{p+1}, \ldots, f_N$ are automatically orthogonal to the d_i's and to each other due to the nature of projection onto the orthogonal complement $\mathcal{D}^\perp$. Hence, they form an orthonormal set under the weighted inner product $\langle \cdot, \cdot \rangle_w$.

Step 4: Verifying the Orthonormality

We now verify that the vectors $\tilde{d}_1, \ldots, \tilde{d}_N$, where $\tilde{d}_i = (d_i^T, f_i^T)^T$, satisfy the orthonormality condition. For $1 \leq i, j \leq p$, we already have:

$$\langle \tilde{d}_i, \tilde{d}_j \rangle_w = \delta_{ij}.$$

For $p + 1 \leq i, j \leq N$, the orthogonality of the vectors f_i's ensures:

$$\langle f_i, f_j \rangle_w = \delta_{ij}.$$

Finally, for any $i \in \{1, \ldots, p\}$ and $j \in \{p+1, \ldots, N\}$, the orthogonality of d_i and f_j ensures:

$$\langle d_i, f_j \rangle_w = 0.$$

Thus, we have constructed an orthonormal set $\{\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N\}$, where $\tilde{d}_i = (d_i^T, f_i^T)^T$, that satisfies the desired properties.

Hence, the matrix formed by the weighted sum $\sum_k w_k \tilde{d}_k \tilde{d}_k^T$ is the identity matrix. Specifically, for any $i, j \in \{1, \ldots, N\}$, we have:

$$\sum_k w_k \tilde{d}_{ki} \tilde{d}_{kj} = \delta_{ij}.$$

Thus, $\sum_k w_k \tilde{d}_k \tilde{d}_k^T = I_N$.