# STAT 578: Topics in High Dimensional Statistics

Jingbo Liu

This version: August 2, 2021

# Contents

# Chapter 0

# Course Information

Welcome! This is a graduate level topics course on high dimensional statistics. Traditional statistics (parametric statistics), which concerns methods and the analysis of settings with fixed number of parameters and increasing number of samples, is by now well-understood. Over the past decades, however, it becomes increasingly important to consider models in which the number of parameters (a.k.a., features, predictors, covariates) grows with the number of samples. The reasons are probably that with higher computation powers, it is possible to handle larger number of parameters and sample sizes; and for larger sample sizes, better prediction can be achieved using more model parameters.

Foundational work on high dimensional statistics and the related nonparametric statistics were laid by the Russian school since the seventies and by Donoho and Johnstone in the early nineties. The simplest example problem in high dimensional statistics is sparse linear regression, which can be solved by the Lasso algorithm for reasonable data sizes. We will discuss Lasso, its analysis, as well as other similar algorithms. Moreover, while sparse recovery is certainly well-studied, recent years have seen growing interests in matrix or tensor recovery, which share some common ideas and techniques.

Below is a tentative list of topics by weeks.

1. Introduction; Gaussian sequence model

2. Lasso; restricted eigenvalue condition; fast rate

3. Nullspace condition

4. Statistical dimension

5. Graphical Lasso

6. Matrix estimation

7. Robust PCA

8. Information-theoretic technique for lower bounds

9. Empirical distribution of error: leave-one-out

10. Empirical distribution of error: replica

11. Iterative soft/hard thresholding

12. Approximate message passing

1-5 are mostly sparse linear regression or the related. 2-4 introduce some tools for analyzing Lasso and provide guarantees on the risk. 8 is about general ideas of showing lower bounds on the risk. 9, 10 are techniques for more refined analysis that even provides the empirical distribution of the errors, and 11,12 additional algorithms for sparse regression. 9, 10, 12 are conceptually more challenging than the others in the list, but are also of greater interest in recent research (in my opinion). Papers on these topics are generally challenging to read, which, in my opinion, is partly because their primary aims are to present original results to get credits. In this course, however, my goal is tutorial, so I will "deconstruct" their technique by working on the simplest model possible. Nevertheless, you are not required to understand deeply all the topics (see grading below).

**Required background.** Basic courses in probability and statistics are assumed. You should also know basic notions of linear algebra, such as eigenvalue decomposition or singular value decomposition.

**Grading.**

- Homework (50%) I will leave a few problems as homework (see last chapter of this document). There is no TA for this course, so the homework will not be graded weekly. I recommend submitting once by the midterm and once by the end of the term. Also, you are not required to learn all the topics, and hence not required to solve all the problems. Number of credits will be given next to each problem, and you only need to solve enough so that the sum $\geq 50$.

- Midterm (25%) Take home exam, more or less a homework but within a fixed time (perhaps 1 week).

- Final project (25%) Read any of the papers below (or some other you find interesting; you may drop me an email to discuss). Write a 10-page report. Also recommend presenting the summary to the class (or upload a video), if convenient. Discussions on potential research topics are welcome.

**Readings Project Examples** [updated periodically]

1. D. Amelunxen, M. Lotz, M. McCoy, J. Tropp, Living on the edge: Phase transitions in convex programs with random data, Information and Inference, 2014.

2. D. Donoho and J. Tanner, Neighborliness of randomly projected simplices in high dimensions, PNAS, 2005.

3. T. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE trans. on electronic computers, 1965.

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

5. Narayana P. Santhanam, Martin J. Wainwright, Information-Theoretic Limits of Selecting Binary Graphical Models in High Dimensions `https://people.eecs.berkeley.edu/~wainwrig/Papers/SanWai12.pdf`

6. I. Daubechies M. Defrise C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint `https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20042`

7. Benjamin Recht, Weiyu Xu, Babak Hassibi, Null Space Conditions and Thresholds for Rank Minimization `https://people.eecs.berkeley.edu/~brecht/papers/10.RecXuHas.Thresholds.pdf`

8. Yuchen Zhang, Distributed machine learning with communication constraints `https://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-47.pdf`

9. Noureddine El Karoui, Derek Bean, Peter Bickel, Chingway Lim and Bin Yu, On robust regression with high-dimensional predictors `https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.294.5920&rep=rep1&type=pdf`

10. Noureddine El Karoui, On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators, Probab. Theory Relat. Fields (2018) 170:95-175

11. Mohsen Bayati, Andrea Montanari, The dynamics of message passing on dense graphs, with applications to compressed sensing `https://arxiv.org/pdf/1001.3448.pdf`

12. Yash Deshpande, Andrea Montanari, Information-theoretically Optimal Sparse PCA `https://arxiv.org/pdf/1402.2238.pdf`

13. The replica trick for the analysis of random matrices `https://meisong541.github.io/jekyll/update/2019/08/04/Replica_method_1.html#ref1`

# Chapter 1

# Linear Regression

*Regression* is a fundamental problem in statistics. Given paired observations $(X_1, Y_1)$, $(X_2, Y_2)$,..., $(X_n, Y_n)$, the statistician wants to find a relationship $f$ between the *predictors* $X_i$'s and the *responses* $Y_i$'s. The response for a fresh input $X$ in future can then be predicted as $f(X)$.

Often, each $Y_i$ is a real number, whereas each $X_i$ is a vector in $\mathbb{R}^d$. Making the ansatz that $f(X) = X^\top \theta$ is linear in the unknown coefficient $\theta$, we may reduce the problem to *linear regression*, i.e. to finding the unknown coefficient vector $\theta$.

To evaluate algorithms for regression, it is useful to analyze simple generative models for the data. The *Gaussian linear model* is given by

$$Y_i = X_i^\top \theta + \xi_i, \quad i = 1, \ldots, n \tag{1.1}$$

where $\xi_1, \dots, \xi_n$ are i.i.d. Gaussian; or, in the matrix form:

$$Y = \mathbb{X}\theta + \xi. \tag{1.2}$$

Regarding the genesis of the matrix $\mathbb{X}$ (equivalently, its row vectors $X_1$,...,$X_n$), there are two types of models that are commonly studied:

- *Fixed design models:* where $\mathbb{X}$ is a deterministic matrix chosen by the statistician.
- *Random design models:* where $X_1$,...,$X_n$ are i.i.d. random vectors drawn from some distribution on $\mathbb{R}^d$.

## 1.1 Gaussian Sequence Model

The model is

$$Y_i = \theta_i + \xi_i, \quad i = 1, \dots, d \tag{1.3}$$

where $\xi_1$,...,$\xi_d$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. This corresponds to a Gaussian linear model with identity fixed design matrix $\mathbb{X} = I_d$. The Gaussian sequence model is perhaps the simplest since the observations of the coordinates of the coefficient vector are decoupled. Yet, the model sheds light on many key ideas that will carry over into general linear regression models or beyond. One key idea, which we are going to explore in the next, is that we can do better than simply returning $\hat{\theta} = Y$, by using an idea called *shrinkage*.

### 1.1.1 Sparsity adaptive thresholding

Let us explore how shrinkage helps in the case of *sparse* signals. We shall use the following notations: for $\theta \in \mathbb{R}^d$,

$$\|\theta\|_0 := \sum_{i=1}^{d} 1\{\theta_i \neq 0\}. \tag{1.4}$$

The set of $k$-sparse vectors is denoted as

$$\mathcal{B}_0(k) := \{\theta\colon \|\theta\|_0 = k\}. \tag{1.5}$$

Note that the naive estimator $\hat{\theta}(Y) = Y$ has squared error

$$\begin{align}
\mathbb{E}\|\hat{\theta}(Y) - \theta\|_2^2 &= \mathbb{E}\|\hat{\theta}(\theta + \xi) - \theta\|_2^2 \tag{1.6}\\
&= \mathbb{E}\|\xi\|_2^2 \tag{1.7}\\
&= d\sigma^2. \tag{1.8}
\end{align}$$

In contrast, if the estimator has the oracle information of $\mathrm{supp}(\theta) := \{i\colon \theta_i \neq 0\}$, then the statistician can use the estimator

$$\hat{\theta}_i = Y_i 1\{i \in \mathrm{supp}(\theta)\}, \tag{1.9}$$

and the squared error should be $k\sigma^2$ instead, by a similar calculation as above. If $k$ scales linearly with $d$, the naive estimator achieves the performance of the oracle estimator up to a constant factor of $d/k$. However, more interesting is the case where $k$ is much smaller than $d$ asymptotically (as a model example, think of $k$ as a polynomial of $d$, say $k = d^{-0.5}$). This is the setting where the following estimator shows power:

**Hard thresholding estimator:**

- Input: vector $\theta \in \mathbb{R}^d$ and parameter $\tau \in (0, \infty)$.
- Output: $\hat{\theta}$ where

$$\hat{\theta}_i := Y_i 1\{|Y_i| > \tau\}. \tag{1.10}$$

How should we pick the threshold $\tau$? The idea to pick it just large enough so that there is no type-I error, with constant probability. Using the union bound, we can see that

$$\max_{1 \le i \le d} |\xi_i| \le \sigma \sqrt{2 \log(2d/\delta)} \tag{1.11}$$

with probability at least $1 - \delta$.

**Theorem 1.** *Consider the Gaussian sequence model and the hard thresholding estimator with*

$$\tau = 2\sigma \sqrt{2 \log(2d/\delta)}. \tag{1.12}$$

*For any $\theta \in \mathcal{B}_0(k)$, we have*

$$\|\hat{\theta} - \theta\|_2^2 \lesssim \sigma^2 k \log \frac{2d}{\delta} \tag{1.13}$$

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

*Proof.* As mentioned, with probability at least $1-\delta$ there is

$$\max_{1\le i\le d}|\xi_i| \le \tau/2 \tag{1.14}$$

which we call the event $\mathcal{A}$. Now under $\mathcal{A}$, we have

$$\begin{aligned}
|\hat{\theta}_i-\theta_i| &= |Y_i-\theta_i|1\{|Y_i|>\tau\}+|\theta_i|1\{|Y_i|\le\tau\} && (1.15)\\
&= |\xi_i|1\{|\theta_i+\xi_i|>\tau\}+|\theta_i|1\{|Y_i|\le\tau\} && (1.16)\\
&\le \frac{\tau}{2}1\{|\theta_i+\xi_i|>\tau\}+|\theta_i|1\{|Y_i|\le\tau\} && (1.17)\\
&\le \frac{\tau}{2}1\{|\theta_i|>\frac{\tau}{2}\}+|\theta_i|1\{|\theta_i|\le\frac{3}{2}\tau\} && (1.18)\\
&\lesssim \min\{|\theta_i|,\tau\} && (1.19)
\end{aligned}$$

for each $i=1,\dots,d$. This yields

$$\|\hat{\theta}-\theta\|_2^2 \lesssim \sum_{i=1}^{d}\min\{|\theta_i|^2,\tau^2\} \lesssim \|\theta\|_0\tau^2. \tag{1.20}$$

□

Remarkably, within a factor of $\log\frac{2d}{\delta}$, the hard thresholding estimator achieved the performance of the oracle estimator. Moreover, the estimator is *adaptive* in the sense that $\tau$ is selected without reference to the sparsity level $k$.

The following estimator shares similar desirable properties (achieving $O(\sigma^2 k\log\frac{2d}{\delta})$ error with probability at least $1-\delta$, and adaptivity), while having the advantage of being continuous:

**Soft thresholding estimator:**

- Input: vector $\theta \in \mathbb{R}^d$ and parameter $\tau \in (0, \infty)$.
- Output: $\hat{\theta}$ where

$$\hat{\theta}_i := (Y_i - \tau) 1\{Y_i > \tau\} + (Y_i + \tau) 1\{Y_i < -\tau\}. \tag{1.21}$$

### 1.1.2 Stein's paradox and the James-Stein estimator

In this section we do not impose any sparsity constraint on ground truth $\theta$. By translation invariance (Is it really true? See Exercise 3), it then seems that nothing more can be done than the naive estimator

$$\hat{\theta}_{\mathsf{naive}} = Y. \tag{1.22}$$

It is therefore paradoxical that the following actually holds:

**Theorem 2.** *The estimator $\hat{\theta}_{\mathsf{naive}}$ is* **inadmissible***, in the sense that there exists another estimator $\hat{\theta}$ such that*

$$\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \le \mathbb{E}\|\hat{\theta}_{\mathsf{naive}} - \theta\|_2^2 \tag{1.23}$$

*for any ground truth $\theta \in \mathbb{R}^d$, and moreover, there exists at least one $\theta$ such that the inequality is strict.*

The James-Stein estimator is one such estimator $\hat{\theta}$. It is defined as

$$\hat{\theta}_{JS} := \left(1 - \frac{\sigma^2(d-2)}{\|Y\|^2}\right) Y. \tag{1.24}$$

Clearly, the idea is still based on shrinkage. In retrospect, the paradox is resolved since the problem is not completely translation invariant: the noise vector $\xi$ is centered (see Exercise 3). Intuitively, a larger $\|Y\|$ suggests a larger "signal-to-noise ratio", and that the estimator should shrink less.

Stein considered estimators of the following form, which clearly encapsulates the estimator of (1.24).

$$\hat{\theta} = g(Y)Y, \tag{1.25}$$

where $g\colon \mathbb{R}^d \to \mathbb{R}$ is a function to be chosen.

*Stein's unbiased risk estimator (SURE)* gives an estimate of the mean squared error (risk) for (1.25):

**Lemma 3.** *Assume that $g$ is a function satisfying regularity conditions*[1]*. Suppose that $\hat{\theta}$ is given by* (1.25)*. Then an unbiased estimator of the risk $\mathbb{E}[\|\hat{\theta} - \theta\|^2]$ is given by*

$$\mathrm{SURE} = \sigma^2 d(2g(Y) - 1) + 2\sigma^2 \sum_{i=1}^{d} Y_i \frac{\partial g}{\partial y_i}(Y) + \|Y\|_2^2 (1 - g(Y))^2. \tag{1.26}$$

[1] See [Tsy08].

Note that SURE is a function only of $Y$ (not of the unknown $\theta$), and it will be shown that its expectation equals $\mathbb{E}[\|\hat{\theta}-\theta\|^2]$; this is the meaning of "unbiased estimator of the risk". Let us remark that estimators of the risk is broadly useful in statistics: for example in nonparametric statistics the nuisance parameters (such as the bandwidth) can be tuned by minimizing an estimator of the risk which can be computed from the observations. Another common technique for risk estimation is cross-validation.

*Proof of Lemma 3.* We begin with the expansion

$$\mathbb{E}\|\hat{\theta}-\theta\|_2^2 = \sum_{i=1}^{d} \mathbb{E}(g(Y)Y_i - \theta_i)^2 \tag{1.27}$$

$$= \sum_{i=1}^{d} \left\{ \mathbb{E}(Y_i-\theta_i)^2 + 2\mathbb{E}[(\theta_i - Y_i)(1-g(Y))Y_i] + \mathbb{E}[Y_i^2(1-g(Y))^2] \right\}. \tag{1.28}$$

Clearly $\mathbb{E}(Y_i-\theta_i)^2 = \sigma^2$. We now apply the Gaussian integration by parts (Exercise 4) to find

$$\mathbb{E}[(\theta_i - Y_i)(1-g(Y))Y_i] = -\sigma^2\mathbb{E}\left[1-g(Y)-Y_i\frac{\partial g}{\partial y_i}(Y)\right] \tag{1.29}$$

which get rids of the dependence on $\theta$, and the claim follows. □

Applying Lemma 3 to the James-Stein estimator, we obtain the following

**Theorem 4.** *Let $d \geq 3$. For any $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\|\hat{\theta}_{JS} - \theta\|^2 = d\sigma^2 - \mathbb{E}\left[\frac{\sigma^4(d-2)^2}{\|Y\|^2}\right] \tag{1.30}$$

*which is strictly smaller than $\mathbb{E}\|Y - \theta\|^2$. In particular, $\hat{\theta} = Y$ is inadmissible.*

*Proof.* From the expression in SURE we see that

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = d\sigma^2 + \mathbb{E}[W(Y)], \tag{1.31}$$

where we have defined the function

$$W(y) := -2\sigma^2 d(1 - g(y)) + 2\sigma^2 \sum_{i=1}^{d} y_i \frac{\partial g}{\partial y_i}(y) + \|y\|_2^2(1 - g(y))^2. \tag{1.32}$$

If $g(y) = 1 - \frac{c}{\|y\|^2}$ for some $c > 0$, then we can explicitly compute

$$W(y) := \frac{1}{\|y\|^2}\left(-2dc\sigma^2 + 4\sigma^2 c + c^2\right). \tag{1.33}$$

Minimizing $W(y)$ over $c > 0$ yields

$$c_{opt} = \sigma^2(d-2), \tag{1.34}$$

and the claim follows by plugging in $c_{opt}$ into SURE. □

## 1.2 Gaussian Linear Model and the Least Squares Regression

The next a few sections concern the Gaussian linear model (1.2):

$$Y = \mathbb{X}\theta + \xi \tag{1.35}$$

where $\mathbb{X} \in \mathbb{R}^{n\times d}$. First, let us review the basic least square estimator, which will reveal the typical behavior of the risk. Solving

$$\hat{\theta}^{LS} \in \operatorname{argmin}_{\theta\in\mathbb{R}^d} \|Y - \mathbb{X}\theta\|_2^2 \tag{1.36}$$

gives

$$\hat{\theta}^{LS} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y \tag{1.37}$$

when $\mathbb{X}$ has full column rank. In general, $(\mathbb{X}^\top\mathbb{X})^{-1}$ is replaced by the Moore-Penrose pseudoinverse of $\mathbb{X}^\top\mathbb{X}$.

**Theorem 5.** *The mean square error of the least square estimator is*

$$\mathbb{E}\|\mathbb{X}\hat{\theta}^{LS} - \mathbb{X}\theta\|_2^2 = d\sigma^2. \tag{1.38}$$

*Proof.* Assuming $\sigma^2 = 1$, we have

$$\begin{aligned}
\mathbb{E}\|\mathbb{X}\hat{\theta}^{LS} - \mathbb{X}\theta\|_2^2 &= \mathbb{E}\|\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y - \mathbb{X}\theta\|_2^2 && (1.39)\\
&= \mathbb{E}\|\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top \xi\|_2^2 && (1.40)\\
&= \operatorname{tr}(\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top) && (1.41)\\
&= \operatorname{tr}((\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{X}) && (1.42)\\
&= \operatorname{tr}(I_d) && (1.43)\\
&= d. && (1.44)
\end{aligned}$$

□

*Remark* 1. We see that the scaling of the mean square risk is the same as in the Gaussian sequence model (1.8), which is the degree of freedom multiplied by the noise variance. Moreover, the result does not depend on (the full rank) $\mathbb{X}$ at all!

*Remark* 2. In general if $\mathbb{X}$ has rank $r$, we can show that $\mathbb{E}\|\mathbb{X}\hat{\theta}^{LS} - \mathbb{X}\theta\|_2^2 = r\sigma^2$. Moreover, if the noise $\xi_i$ is *subgaussian* instead of Gaussian, a similar result of $\mathbb{E}\|\mathbb{X}\hat{\theta}^{LS} - \mathbb{X}\theta\|_2^2 \lesssim r\sigma^2$ holds, using a different derivation via the maximal inequality [Rig15, Theorem 2.2].

## 1.3 $\ell_0$ and $\ell_1$ Regularized Regressions

The least squares estimator does not apply to the case of $d > n$. Moreover, as we saw in the case of Gaussian sequence model (Theorem 1), when $\theta$ is $k$-sparse we should expect the mean square error to scale as $k\sigma^2$ (up to logarithmic factors) instead of $d\sigma^2$. Motivated by the soft and hard thresholding estimators in the Gaussian sequence model, it is natural to consider the Bayes Information Criterion (BIC) estimator and the Lasso estimator:

$$\hat{\theta}^{BIC} \in \mathrm{argmin}_{\theta\in\mathbb{R}^d}\left\{\|Y-\mathbb{X}\theta\|_2^2+\tau^2\|\theta\|_0\right\}; \tag{1.45}$$

$$\hat{\theta}^{L} \in \mathrm{argmin}_{\theta\in\mathbb{R}^d}\left\{\|Y-\mathbb{X}\theta\|_2^2+2\tau\|\theta\|_1\right\} \tag{1.46}$$

where we recall that

$$\|\theta\|_0 := \sum_{i=1}^d 1\{\theta_i\neq 0\}; \tag{1.47}$$

$$\|\theta\|_1 := \sum_{i=1}^d |\theta_i|. \tag{1.48}$$

Indeed, we can check that when $n=d$ and $\mathbb{X}=I_d$, BIC and Lasso are reduced to the hard-thresholding and the soft thresholding estimators (Exercise 5).

Let us remark that the contents of this and the following sections are mostly from [Rig15], where $\mathbb{X}$ has spectral norm scaling as $\sqrt{n}$. However, in this note we rescale $\mathbb{X}$ so that the spectral norm is order 1, so that the correspondence to the Gaussian sequence model ($\mathbb{X}=I_d$ case) is cleaner. Correspondingly, there is no extra $\frac{1}{n}$ factor in front of the 2-norms in (1.45)-(1.46).

### Computation issues.

Exactly solving BIC is known to be NP-hard, since it requires enumerating all possible sparsity patterns $\operatorname{supp}(\theta) := \{i \colon \theta_i \neq 0\}$. In contrast, the Lasso estimator is a convex optimization and can be solved efficiently by a number of algorithms. One of the most popular methods is coordinate gradient descent, which has been implemented in the glmnet package in R.

### Fast rate for the Lasso.

Under a *restricted eigenvalue condition* for $\mathbb{X}$, we will show that Lasso can achieve mean square error of the same $k\sigma^2 \log(\dots)$ order as the soft/hard-thresholding estimator for the Gaussian sequence model. This is called the fast rate for the Lasso, as opposed to the slow rate which holds for a more relaxed class of $\mathbb{X}$ (see [Rig15, Section 2.4]).

**Definition 6.** We say that $\mathbb{X}$ satisfies the $(k, \kappa)$-restricted eigenvalue condition (RE) if

$$\inf_{|\mathcal{S}| \leq k} \inf_{\theta \in \mathcal{C}_S} \frac{\|\mathbb{X}\theta\|_2^2}{\|\theta_S\|_2^2} \geq \kappa \tag{1.49}$$

where $\mathcal{C}_S = \{\theta \colon \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1\}$.

*Remark* 3. Our definition (1.49) is slighted different from some literature by a factor of $n$ and also the denominator has $\theta_S$ instead of $\theta$; in [RWY10] the restricted eigenvalue condition is defined as $\frac{1}{n}\inf_{|\mathcal{S}|\le k}\inf_{\theta\in\mathcal{C}_S}\frac{\|\mathbb{X}\theta\|_2^2}{\|\theta\|_2^2}\ge\kappa$ (and as a consequence other parts in the theory are rescaled).

*Remark* 4. A "typical behavior" for random matrices is that RE is satisfied when $n \geq k\log d$; see Corollary 1 and Section 3.2 in [RWY10] for more precise statements.

The following result is taken from [Rig15, Theorem 2.18]

**Theorem 7.** *Consider the Gaussian linear model* (1.2)*. Let* $n \geq 2$*. Assume that the ground truth* $\|\theta^*\|_0 \leq k$*, and* $\mathbb{X}$ *satisfies the* $(k, 1/2)$*-restricted eigenvalue condition and that each column satisfies* $\|\mathbb{X}_j\|_2^2 \leq 2$*. Then the Lasso estimator* $\hat{\theta}^L$ *with regularization parameter satisfying*

$$2\tau = 8\sigma\left(\sqrt{\log(2d)}+\sqrt{\log\frac{1}{\delta}}\right) \tag{1.50}$$

*satisfies*

$$\|\mathbb{X}\hat{\theta}^L - \mathbb{X}\theta^*\|_2^2 \lesssim k\sigma^2\log(2d/\delta) \tag{1.51}$$

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

*Proof.* First, we write out the optimality condition for the Lasso estimator:

$$\|Y-\mathbb{X}\hat{\theta}^L\|_2^2 \leq \|Y-\mathbb{X}\theta^*\|_2^2 + 2\tau\|\theta^*\|_1 - 2\tau\|\hat{\theta}^L\|_1. \tag{1.52}$$

Comparing with our goal of bounding $\|\mathbb{X}\hat{\theta}^L - \mathbb{X}\theta^*\|_2$, we see that we should use $Y = \mathbb{X}\theta^* + \xi$ to open up the squares in the above inequality. This yields

$$\|\mathbb{X}\hat{\theta}^L - \mathbb{X}\theta^*\|_2^2 \leq 2\xi^\top\mathbb{X}(\hat{\theta}^L - \theta^*) + 2\tau\|\theta^*\|_1 - 2\tau\|\hat{\theta}^L\|_1. \tag{1.53}$$

The hard part now is to control $\xi^\top\mathbb{X}(\hat{\theta}^L - \theta^*)$. The difficulty lies in the fact that $\hat{\theta}^L$ and $\xi$ are correlated. A few intuitive ideas for handling similar situations in high-dimensional statistics include:

- "sup-out" $\hat{\theta}$, i.e., use $\frac{\xi^\top\mathbb{X}(\hat{\theta}^L-\theta^*)}{\|\mathbb{X}(\hat{\theta}^L-\theta^*)\|_2} \leq \sup_{x:\, \|x\|_2\leq 1} \xi^\top x \lesssim \sqrt{n}$ (with high probability), so that $\xi^\top\mathbb{X}(\hat{\theta}^L - \theta^*) \lesssim \sqrt{n}\|\mathbb{X}(\hat{\theta}^L - \theta^*)\|_2$. An example of this argument can be found in [Rig15, Theorem 2.2] about the mean square error of the least squares estimator.

- The $\ell_1$-regularization forces $\theta^L$ to be "low-complexity", so that it cannot be too correlated with $\xi$. (In the extreme case where $\theta^L$ is a constant, we obviously have $\mathbb{E}[\xi^\top\mathbb{X}(\hat{\theta}^L - \theta^*) = 0]$.)

The derivation here will be essentially fleshing out the second idea. With probability $\geq 1 - \delta$ we have

$$\begin{aligned} \xi^\top\mathbb{X}(\hat{\theta}^L - \theta^*) &\leq \|\xi^\top\mathbb{X}\|_\infty\|\hat{\theta}^L - \theta^*\|_1 && (1.54)\\ &\leq \frac{\tau}{2}\|\hat{\theta}^L - \theta^*\|_1 && (1.55) \end{aligned}$$

where the first step follows by Hölder's inequality and the second follows by the assumption $\|\mathbb{X}_j\|_2^2 \leq 2$ for each column, and the bound on the Gaussian max in Exercise 1. Combining (1.55) and (1.53), we obtain

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^L - \mathbb{X}\theta^*\|_2^2 &\leq \tau\|\hat{\theta}^L - \theta^*\|_1 + 2\tau\|\theta^*\|_1 - 2\tau\|\hat{\theta}^L\|_1 && (1.56)\\
&= \tau\|\hat{\theta}^L_S - \theta^*_S\|_1 - \tau\|\hat{\theta}^L_{S^c}\|_1 + 2\tau\|\theta^*_S\|_1 - 2\tau\|\hat{\theta}^L_S\|_1 && (1.57)\\
&\leq 3\tau\|\hat{\theta}^L_S - \theta^*\|_1 - \tau\|\hat{\theta}^L_{S^c}\|_1 && (1.58)
\end{aligned}$$

where we have used $\theta^*_S = \theta^*$ and the triangle inequality. Thus, $3\|\hat{\theta}^L_S - \theta^*\|_1 \geq \|\hat{\theta}^L_{S^c}\|_1$ holds, and by Cauchy-Schwarz and the restricted eigenvalue condition,

$$\|\hat{\theta}^L_S - \theta^*\|_1 \leq \sqrt{k}\|\hat{\theta}^L_S - \theta^*\|_2 \leq \sqrt{2k}\|\mathbb{X}(\hat{\theta}^L - \theta^*)\|_2. \qquad (1.59)$$

Plugging this into (1.58), we find $\|\mathbb{X}\hat{\theta}^L - \mathbb{X}\theta^*\|_2^2 \leq 3\tau\sqrt{2k}\|\mathbb{X}(\hat{\theta}^L - \theta^*)\|_2$ or equivalently

$$\|\mathbb{X}\hat{\theta}^L - \mathbb{X}\theta^*\|_2 \leq 3\tau\sqrt{2k}. \qquad (1.60)$$

□

## 1.4 Basis Pursuit and the Null Space Condition

In this section we introduce the *Basis Pursuit* (BP) algorithm and its exact recovery property in the noiseless setting under a *null space condition*. Recall that the Lasso algorithm (1.46) has inputs $Y$ and $\mathbb{X}$, and selects $\hat{\theta}$ by solving the following:

$$\hat{\theta}^L \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left\{ \|Y - \mathbb{X}\theta\|_2^2 + 2\tau \|\theta\|_1 \right\}. \tag{1.61}$$

Now if $\mathbb{X}$ has full column rank and $\tau \downarrow 0$ (meaning that $\|Y - \mathbb{X}\theta\|_2^2$ has a much higher weight than $\|\theta\|_1$), we see that $\hat{\theta}^L$ converges to the solution of the following *basis pursuit linear program* (introduced by [CHE98]):

$$\hat{\theta}^{BP} \in \operatorname{argmin}_{\theta \colon Y = \mathbb{X}\theta} \|\theta\|_1. \tag{1.62}$$

**Definition 8.** Fix $\mathbb{X} \in \mathbb{R}^{n \times d}$ and $k \in \{1, 2, \dots\}$. We say $\mathbb{X}$ satisfies the *exact recovery property of order* $k$ if for any $k$-sparse vector $\theta^* \in \mathbb{R}^d$ and

$$y := \mathbb{X}\theta^*, \tag{1.63}$$

the BP algorithm with inputs $\mathbb{X}$ and $y$ returns the solution $\hat{\theta}^{BP} = \theta^*$.

Note that there is no additive noise in (1.63), hence the generative model is deterministic and we used the lowercase letter $y$.

**Definition 9.** Fix $\mathbb{X} \in \mathbb{R}^{n\times d}$ and $k \in \{1, 2, \dots\}$. We say $\mathbb{X}$ satisfies the *null space condition of order $k$* if

$$\|z_S\|_1 < \|z_{S^c}\|_1, \quad \forall z \in \mathcal{N}(\mathbb{X}) \setminus \{0\},\ S\colon |S| \le k \tag{1.64}$$

where $\mathcal{N}(\mathbb{X})$ denotes the null space of $\mathbb{X}$.

The definition of the null space condition appeared in earlier work of Donoho and Huo [DH06]; Feuer and Nemirovski [FN03], and was further explored by Cohen et al. [CDD09]. The significance of the null space condition is seen in its equivalence to the exact recovery by BP in the noiseless setting:

**Theorem 10.** *For given $\mathbb{X} \in \mathbb{R}^{n\times d}$ and $k \in \{1, 2, \dots\}$, the null space condition and the exact recovery condition of order $k$ are equivalent.*

The equivalence of the null space condition and exact recovery has been discussed in a number of papers (Cohen et al.[CDD09]; Donoho and Huo [DH06]; Elad and Bruckstein [EB02]; Feuer and Nemirovski [FN03]) for more discussion of restricted nullspaces and equivalence to exact recovery of basis pursuit.

*Proof of Theorem 10.* First, assume that the null space condition is satisfied. Assume that $\theta^*$ is a $k$-sparse vector with support $S$, and $y := \mathbb{X}\theta^*$. To show the exact recovery property, we need to show that for any $\theta' \in \mathbb{R}^d$, $\theta' \neq \theta^*$ satisfying $\mathbb{X}\theta' = \mathbb{X}\theta^*$, there is

$$\|\theta'\|_1 > \|\theta^*\|_1. \tag{1.65}$$

Define $z := \theta' - \theta^* \neq 0$, we see that $z \in \mathcal{N}(\mathbb{X})$, and hence $\|z_S\|_1 < \|z_{S^c}\|_1$ by the null space property. Therefore,

$$\begin{aligned}
\|\theta'\|_1 &= \|\theta'_S\|_1 + \|\theta'_{S^c}\|_1 && (1.66)\\
&= \|\theta'_S\|_1 + \|z_{S^c}\|_1 && (1.67)\\
&> \|\theta'_S\|_1 + \|z_S\|_1 && (1.68)\\
&\geq \|\theta'_S - z_S\|_1 && (1.69)\\
&= \|\theta^*_S\|_1 && (1.70)\\
&= \|\theta^*\|_1 && (1.71)
\end{aligned}$$

as desired, hence the exact recovery property holds.

Conversely, suppose that the exact recovery property holds. Now pick any $z \in \mathcal{N}(\mathbb{X})$, $z \neq 0$. Define $\theta^* := z_S$ (where we recall that $z_S$ is defined by $z_S(i) := z(i)1_{i \in S}$, $i - 1, \dots, d$) and $\theta' := -z_{S^c}$. We have $\mathbb{X}\theta^* = \mathbb{X}\theta'$ and therefore the exact recovery property implies $\|\theta^*\|_1 < \|\theta\|_1$. This is equivalent to $\|z_S\|_1 < \|z_{S^c}\|_1$ and hence the null space property holds. □

The equivalence between exact recovery and the null space condition can be extended to certain nonconvex optimization problems (Exercise 6). Furthermore, a version of the null space condition can be shown to be equivalent to the *robust recovery property* under the noisy observation model $y = \mathbb{X}\theta^* + \xi$, that is,

$$\frac{\|\hat{\theta}^{BP} - \theta^*\|_2}{\|\xi\|_2} \leq c \qquad (1.72)$$

for some constant $c$ independent of $\xi$ and $\theta^*$; see [LJG15] for details.

**Comparison with RE.**

**Proposition 11.** *The restricted eigenvalue condition (RE) in Definition 6 is stronger than the null space property.*

*Proof.* Suppose that $(k, \kappa)$-RE holds for $\mathbb{X}$ but the null space property of order $k$ fails. Then there exists $z \in \mathcal{N}(\mathbb{X})$, $z \neq 0$ such that $\|z_S\|_1 \geq \|z_{S^c}\|_1$ where $|S| = k$. This implies $z_S \neq 0$ and $z \in \mathcal{C}_S$ where $\mathcal{C}_S$ is as defined in Definition 6. The restricted eigenvalue condition implies

$$\|\mathbb{X}z\|_2^2 \geq \kappa \|z_S\|_2^2 > 0 \tag{1.73}$$

where we assumed $\kappa > 0$, This contradicts $z \in \mathcal{N}(\mathbb{X})$, hence the null space condition must hold. □

## 1.5 NSC for Random Designs

In this section we show that the null space conditions holds with high probability under the random design where $\mathbb{X} \in \mathbb{R}^{n \times d}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries and $k \lesssim n/\log d$. Up to the logarithmic term, this is the best we can expect. Our proof is based on Gordon's *escape through the mesh* theorem.

**Definition 12.** The Gaussian width of a set $\mathcal{K} \subseteq \mathbb{R}^d$ is defined as

$$w(\mathcal{K}) := \mathbb{E} \sup_{x \in \mathcal{K}} G^\top x \tag{1.74}$$

where $G \sim \mathcal{N}(0, I_d)$.

**Theorem 13** (Gordon)**.** *Let $\mathcal{K}$ be a subset of the unit Euclidean sphere $S^{d-1}$ in $\mathbb{R}^d$. Let $\nu$ be a uniformly distributed[2] random $(d-n)$-dimensional subspace of $\mathbb{R}^d$. Assume that*

$$w(\mathcal{K}) < \sqrt{n}. \tag{1.75}$$

*Then $\nu \cap \mathcal{K} = \emptyset$ with probability at least*

$$1 - 2.5 \exp\left( -\frac{(n/\sqrt{n+1} - w(\mathcal{K}))^2}{18} \right). \tag{1.76}$$

The proof of Theorem 13 in [Gor88] is based on the Gaussian comparison inequalities, which is a type of results establishing inequalities between the (expectations of the) extremal values of two Gaussian processes with related covariance structures. This is usually a topic of a high dimensional *probability* course which we shall not get into.

With the help of Gordon's theorem we can establish the null space property under the i.i.d. Gaussian random designs when $n \gtrsim k \log d$ [RV08], which is essentially the best we can hope for.

[2] That is, the distribution of the subspace is rotation-invariant; or, the distribution is the Haar measure on the Grassmannian.

**Theorem 14.** *Let $d \geq 10$, $k \geq 1$, and $n \geq 400k \log d$. Let $\mathbb{X} \in \mathbb{R}^{n \times d}$ be a random matrix where each entry is i.i.d. $\mathcal{N}(0, 1/n)$. Then $\mathbb{X}$ satisfies the null space condition of order $k$ with probability at least $1 - 2.5d^{-k/18}$.*

*Proof.* For any $S \subseteq \{1, 2, \ldots, d\}$, define

$$\mathcal{K}_S := \{z \in S^{d-1} \colon \|z_{S^c}\|_1 \leq \|z_S\|_1\} \tag{1.77}$$

and define

$$\mathcal{K} := \bigcup_{S \colon |S|=k} \mathcal{K}_S. \tag{1.78}$$

We now wish to upper bound $w(\mathcal{K})$ by controlling the upper tail of $\sup_{x \in \mathcal{K}} G^\top x$, which, in turn, is based controlling the upper tail fo $\sup_{x \in \mathcal{K}_S} G^\top x$ and applying the union bound over $S$. Note that for any $x \in \mathcal{K}_S$, we have $G^\top x = G_S^\top x_S + G_{S^c}^\top x_{S^c}$. Moreover,

$$G_S^\top x_S \leq \|G_S^\top\|_2 \tag{1.79}$$

and

$$\begin{aligned}
G_{S^c}^\top x_{S^c} &\leq \|G_{S^c}^\top\|_\infty \|x_{S^c}\|_1 && (1.80)\\
&\leq \|G\|_\infty \|x_{S^c}\|_1 && (1.81)\\
&\leq \|G\|_\infty \|x_S\|_1 && (1.82)\\
&\leq \|G\|_\infty \sqrt{k} \|x_S\|_2 && (1.83)\\
&\leq \sqrt{k} \|G\|_\infty && (1.84)
\end{aligned}$$

Therefore $\sup_{x\in\mathcal{K}} G^\top x \le \sup_{S\colon |S|=k} \|G_S\|_2 + \sqrt{k}\|G\|_\infty$. We have $\mathbb{E}[\sqrt{k}\|G\|_\infty] \le \sqrt{k}(\sqrt{2\log 2d}+1)$ (this can be shown from the bound on the Gaussian max Exercise 1), whereas

$$\begin{aligned}
&\mathbb{E}[\sup_{S\colon |S|=k} \|G_S\|_2] \\
&= \int_0^\infty \mathbb{P}[\sup_{S\colon |S|=k} \|G_S\|_2 > \lambda] d\lambda && (1.85)\\
&\le \sqrt{k} + \sqrt{2k\log d} + \int_{t=k\log d}^\infty \mathbb{P}[\sup_{S\colon |S|=k} \|G_S\|_2 > \sqrt{k} + \sqrt{2t}] d(\sqrt{2t}) && (1.86)\\
&\le 2\sqrt{2k\log d} + 2d^k \int_{k\log d}^\infty e^{-t} d(\sqrt{2t}) && (1.87)\\
&= 2\sqrt{2k\log d} + 2d^k \int_{\sqrt{2k\log d}}^\infty e^{-s^2/2} ds && (1.88)\\
&\le 2\sqrt{2k\log d} + 2 && (1.89)
\end{aligned}$$

where (1.87) follows by the tail bound on the chi-square distribution (Exercise (7)) together with the union bound which gives an additional $\binom{d}{k} \le d^k$ factor. Thus we have shown that

$$\begin{aligned}
w(k) &\le \sqrt{k}(\sqrt{2\log 2d} + 1) + 2\sqrt{2k\log d} + 2 && (1.90)\\
&\le 10\sqrt{k\log d} && (1.91)
\end{aligned}$$

which means that the condition (1.75) in Gordon's theorem is satisfied. Applying Gordon's theorem with $\nu := \mathcal{N}(\mathbb{X})$ shows that

$\mathcal{N}(\mathbb{X}) \cap \mathcal{K} = \emptyset$ with probability at least $1 - 2.5d^{-k/18}$. □

*Remark* 5. For the restricted eigenvalue condition (which is stronger than the null space condition according to Proposition 11), it is also true that $n \gtrsim k \log d$ suffices; see Corollary 1 in [RWY10] for the i.i.d. random designs. In fact, the bound in Corollary 1 in [RWY10] holds for the more general class of random designs where the rows are arbitrary Gaussian vectors with possibly correlated entries. The argument of [RWY10] is more lengthy than our Theorem 14 but the ideas are rather standard by today's standards: the goal is essentially to bound the tail probability of a Rayleigh quotient. This is based on two components 1) the expectation of the Rayleigh quotient is bounded using a comparison theorem for the Gaussian process (similar to Gordon's theorem). 2) The deviation from the expectation is controlled using *concentration inequalities*. Again, these are common topics in a high dimensional *probability* course, which we will not touch deeply in our high dimensional *statistics* course.

## 1.6 More on Convex Geometry

In the previous section we have seen a proof of the null space condition using Gordon's theorem, which says that a set on the sphere with small Gaussian width will miss a random hyperplane with high probability. As mentioned, the Gaussian width is essentially the mean width from convex geometry. In this section, we discuss a different approach for establishing the null space condition (among other results in high dimensional regression) by Amelunxen, Lotz, McCoy and Tropp [ALMT14]. While Gordon's theorem [Gor88] relied on the Gaussian comparison theorem, the approach of [ALMT14] has a more geometric flavor and also yields more general results.

### 1.6.1 A General Result on the Intersection of Convex Cones

**First definition of the statistical dimension.**

A set $\mathcal{K}$ in $\mathbb{R}^d$ is said to be a convex cone if it is a cone (i.e., $x \in \mathcal{K}$ implies $\lambda x \in \mathcal{K}$ for all $\lambda \geq 0$) and is convex. We define the statistical dimension of a closed convex cone $\mathcal{K}$ as

$$\delta(\mathcal{K}) := \mathbb{E}[\|P_{\mathcal{K}}(G)\|_2^2] \tag{1.92}$$

where $G \sim \mathcal{N}(0, I_d)$ and $P_{\mathcal{K}}(G) := \operatorname{argmin}_{z \in \mathcal{K}} \|G - z\|_2$.

An equivalent definition of the statistical dimension (which applies to general sets in $\mathbb{R}^d$) will be given in Section 1.6.3.

The statistical dimension of $\mathcal{K}$ equals the square of Gaussian width of $\mathcal{K} \cap S^{d-1}$ up to an additive constant; see Exercise 9.

The following result is found in Theorem I in [ALMT14].

**Theorem 15.** *Fix $\eta \in (0, 1)$. Let $\mathcal{C}$ and $\mathcal{K}$ be closed convex cones in $\mathbb{R}^d$, and let $Q \in \mathbb{R}^{d \times d}$ be a random orthogonal matrix (with* *rotationally invariant distribution). Then*

$$\delta(\mathcal{C}) + \delta(\mathcal{K}) \le d - a_\eta\sqrt{d} \Longrightarrow \mathbb{P}[\mathcal{C} \cap Q\mathcal{K} \neq \{0\}] \le \eta \tag{1.93}$$

$$\delta(\mathcal{C}) + \delta(\mathcal{K}) \ge d + a_\eta\sqrt{d} \Longrightarrow \mathbb{P}[\mathcal{C} \cap Q\mathcal{K} \neq \{0\}] \ge 1 - \eta \tag{1.94}$$

*where* $a_\eta := \sqrt{8\log(4/\eta)}$.

If $\mathcal{K}$ is a $(d-n)$-dimensional subspace, it is easy to see from the definition that $\delta(\mathcal{K}) = d - n$. Then (1.93) implies that if $\delta(\mathcal{C}) \le n - a_\eta\sqrt{d}$ then $\mathbb{P}[\mathcal{C} \cap Q\mathcal{K} \neq \{0\}] \le \eta$. This is already implied by Gordon's theorem (Theorem 13) since $w^2(\mathcal{C} \cap S^{d-1}) \le \delta(\mathcal{C}) \le w^2(\mathcal{C} \cap S^{d-1}) + 1$ (Exercise 9). On the other hand, Gordon's theorem did not provide the converse part (1.94). Note, however, a converse of Gordon's theorem for specific choices of $\mathcal{C}$ are easy to check (Exercise 8), and for the application in null space condition it is easy to see the near sharpness of the $n \gtrsim k\log d$ bound. Nevertheless, Theorem 15 is surprising since it shows that the statistical dimension/Gaussian width along is enough to determine the intersection probability, and that a *phase transition* exists for all convex cones (with a window size of $\sim \sqrt{d}$ for the statistical dimension).

### 1.6.2 Implication for Basis Pursuit

Given $\theta \in \mathbb{R}^d$, define the *descent cone*

$$\mathcal{D}(\theta) := \{z \in \mathbb{R}^d \colon \exists\epsilon > 0,\ \|\theta + \epsilon z\|_1 \le \|\theta\|_1\}, \tag{1.95}$$

that is, the set of directions at which the objective of the basis pursuit is decreased. Using a similar argument as in the proof of the equivalence between NSC and the exact recovery (Section 1.4), it is easy to see that

$$\text{Given } \mathbb{X},\, \theta,\, \text{BP exactly recovers } \theta \Longleftrightarrow \mathcal{D}(\theta) \cap \mathcal{N}(\mathbb{X}) = \{0\}. \tag{1.96}$$

The difference between this claim and Theorem 10 is that the former is about a condition of $\mathbb{X}$ and $\theta$ while the latter is about a condition of $\mathbb{X}$ and $k$.

If $\theta \in \mathbb{R}^d$ is $k$-sparse, using an argument similar to Theorem 14 we can show that $\delta(\mathcal{D}(\theta)) = \Theta(k \log d)$ (i.e. equals $k \log d$ up to constant factors) when $d$ is much larger than $k$; see Exercise 10. Combining this observation with Theorem 15, we obtain:

**Proposition 16.** *Suppose that $\mathbb{X} \in \mathbb{R}^{n \times d}$ and its null space is a (uniformly) random $(d-n)$-subspace, and $\theta \in \mathbb{R}^d$ is $k$-sparse where $d$ is much larger than $k$ ($d \geq k^\tau$ for some $\tau > 1$). Then there there exists universal constants $c, C > 0$ such that*

$$n \geq C\left(k \log d + \sqrt{d \log \frac{1}{\eta}}\right) \Longrightarrow \mathbb{P}[\mathit{BP\ exactly\ recovers\ } \theta] \geq 1 - \eta; \tag{1.97}$$

$$n \leq c\left(k \log d - \sqrt{d \log \frac{1}{\eta}}\right) \Longrightarrow \mathbb{P}[\mathit{BP\ exactly\ recovers\ } \theta] \leq \eta. \tag{1.98}$$

The part (1.97) is already implied by Theorem 14. Moreover, Theorem 14 is stronger since it provides a lower bound on the probability that BP recovers *any* $k$-sparse vector. Let us also remark that the factor $\log d$ in front of $k$ is not removed in (1.97) even though we are looking at a specific instance of $\theta$; indeed, while the $\log d$ factor can be removed in (1.89) when we bound $\mathbb{E}[\|G_S\|_2]$ instead, the $\log d$ factor in $\mathbb{E}[\sqrt{k}\|G\|_\infty]$ still remains.

Finally, let us remark that Theorem 15 has many applications in linear inverse problems beyond the analysis of BP, including demixing and low rank matrix recovery. Please refer to [ALMT14] for more details.

### 1.6.3 Proof Sketch

In this section we mention some ingredients for the proof of Theorem 15 in [ALMT14]. The materials require more background and are only intended for interested readers.

**Intrinsic volumes.**

A central idea of the proof is to define the so-called *intrinsic volume random variable* [MT14, LMN$^+$20], show its concentration property, and its connection to the statistical dimension. The intrinsic volume random variable has a distribution on $\{1, 2, \ldots, d\}$, where the probabilities are called the *(normalized) intrinsic volumes*. There are two equivalent definitions:

- One definition given in Definition 5.1 in [ALMT14] is as following: First, if $\mathcal{K}$ is a polyhedral cone (i.e., an intersection finitely many half spaces), then $\tilde{V}_j(\mathcal{K})$ is defined as the probability that $P_{\mathcal{K}}(G)$ lies in the relative interior of a $j$-dimensional face of $\mathcal{K}$. As before, $P_{\mathcal{K}}(G)$ denotes the projection of a standard Gaussian vector onto $\mathcal{K}$. Then the definition can be extended to a general closed cone $\mathcal{K}$ via approximation (in the conic Hausdorff metric) by polyhedral cones. It is clear from the definition that
$$\sum_{j=0}^{d} \tilde{V}_j(\mathcal{K}) = 1. \tag{1.99}$$

- Another definition is through the *spherical Steiner formula*, for which we refer to [MT14]. Here for simplicity, let us describe instead the intrinsic volume for convex bodies (compact convex sets with nonempty interior) via the standard (i.e. not spherical) Steiner formula. Note, though, that convex bodies are not cones and the corresponding intrinsic volumes are different albeit related concepts. The intrinsic volumes for convex bodies are closer to our intuitions about volumes in the Euclidean space.

  The following definition through the *mixed volumes* is taken from Definition 1.6 in [LMN$^+$20]. Let $\mathcal{K}$ be a convex body (compact and having nonempty interior). Let $B$ be the unit ball in $\mathbb{R}^d$. The Steiner formula states that for any $t \geq 0$,

$$\operatorname{vol}(\mathcal{K} + tB) = \sum_{j=0}^{d} \binom{d}{j} W_j^{(d)}(\mathcal{K}) t^j \tag{1.100}$$

where $W_j^{(d)}(\mathcal{K})$, $j = 0, \dots, d$ are nonnegative numbers, called the *quermassintegrals*. In fact, in general, $\operatorname{vol}(\mathcal{K} + t\mathcal{C})$ is a polynomial in $t$ for any convex sets $\mathcal{K}$ and $\mathcal{C}$, which can be shown by expressing the volume using the support function of the convex sets (see e.g. [SVH19]) with coefficients being the *mixed volumes*. The intrinsic volumes are defined by

$$V_{d-j}(\mathcal{K}) = \frac{1}{\kappa_j} \binom{d}{j} W_j^{(d)}(\mathcal{K}) \tag{1.101}$$

for $j = 0, \dots, d$, where $\kappa_j$ denotes the volume of the unit ball in $\mathbb{R}^j$. As a matter of fact, the intrinsic volumes are intrinsic: if $\mathcal{K}$ is embedded in a higher dimensional space, intrinsic volumes defined by (1.101) do not change. Now the normalized intrinsic volumed are defined by

$$\tilde{V}_j(\mathcal{K}) := \frac{V_j(\mathcal{K})}{W(\mathcal{K})}, \quad j = 0, \dots, d \tag{1.102}$$

where

$$W(\mathcal{K}) := \sum_{j=0}^{d} V_j(\mathcal{K}). \tag{1.103}$$

Finally, let us comment that the distribution of the (conic) intrinsic volume random variable is invariant under scaling, since if $\mathcal{K}$ is a closed convex cone, then $t\mathcal{K} = \mathcal{K}$ for any $t > 0$. In contrast, the distribution of the intrinsic volume random variable of a convex body is not invariant under scaling. Indeed, $V_j$ is $j$-homogenous; In fact, $\lim_{t\to\infty} V_j(t\mathcal{K}) = 1\{j = d\}$ for convex body $\mathcal{K}$.

**Second definition of the statistical dimension.**

The statistical dimension defined in (1.92) can be alternatively defined as the expectation of the intrinsic volume random variable. The proof of equivalence is shown in [MT14].

**Conic kinematic formula.**

The proof of Theorem 15 in [ALMT14] relied on the following *conic kinematic formula*:

**Theorem 17.** *Let $\mathcal{C}$ and $\mathcal{K}$ be closed convex cones in $\mathbb{R}^d$, one of which not a subspace*[3]*, and $Q$ be a (uniformly) random orthogo-*

[3] The formula does not hold when both $\mathcal{C}$ and $\mathcal{K}$ are linear subspaces, in which case $\tilde{V}$ becomes the indicator of the dimension. One might think of approximating of a subspace by convex cones which are not subspaces and applying a continuity argument for $\tilde{V}$; however, approximation of a subspace is not so easy due to the convex cone constraint.

*nal matrix. Then*

$$\mathbb{P}[\mathcal{C} \cap Q\mathcal{K} \neq \{0\}] = \sum_{i=0}^{d} \sum_{j=0}^{i-1} (1 + (-1)^{i-j+1}) \tilde{V}_i(\mathcal{C}) \tilde{V}_{d-j}(\mathcal{K}). \tag{1.104}$$

We refer this result to p260 in [SW08]. In the well-cited paper [ALMT14], this result was presented in Fact 2.1 as the following formula:

$$\mathbb{P}[\mathcal{C} \cap Q\mathcal{K} \neq \{0\}] = \sum_{i=0}^{d} (1 + (-1)^{i+1}) \sum_{j=i}^{d} \tilde{V}_i(\mathcal{C}) \tilde{V}_{d+i-j}(\mathcal{K}) \tag{1.105}$$

which appears to be incorrect (the formula is not symmetric in the roles of $\mathcal{C}$ and $\mathcal{K}$, and it fails when we consider the example where the intrinsic volume random variables for $\mathcal{C}$ and $\mathcal{K}$ are concentrated around some $i_0$ and $j_0$).

In fact, (5.8) stated an equivalent but more compact formula:

$$\mathbb{P}[\mathcal{C} \cap Q\mathcal{K} \neq \{0\}] = \sum_{k=d+1}^{2d} (1 + (-1)^{k-d+1}) \tilde{V}_k(\mathcal{C} \times \mathcal{K}) \tag{1.106}$$

where $\mathcal{C} \times \mathcal{K}$ denotes the product set in $\mathbb{R}^{2d}$. Using the fact that the conic intrinsic volume of a product set can be computed by convolution (Corollary 5.1 in [MT14]),

$$\tilde{V}_k(\mathcal{C} \times \mathcal{K}) = \sum_{i+j=k} \tilde{V}_i(\mathcal{C}) \tilde{V}_j(\mathcal{K}), \tag{1.107}$$

we obtain (1.104).

**Concentration of the intrinsic volume random variable.**

The last key technical ingredient for the proof of Theorem 15 is the concentration property of the intrinsic volume random variable. This was shown in Theorem 6.1 in [ALMT14] and improved in [MT14] for convex cones and in Theorem 1.11 in [LMN$^+$20] for convex bodies (via a beautiful information theoretic argument). Concentration states that the intrinsic volume random variable is close to its mean with high probability. For example, a basic (though not the strongest possible) result from Theorem 1.11 in [LMN$^+$20] states that

$$\mathrm{Var}(Z_{\mathcal{K}}) \leq 4d \tag{1.108}$$

where $Z_{\mathcal{K}}$ denotes the intrinsic random variable associated with the convex body $\mathcal{K}$. (This is nontrivial since a random variable over $\{1, 2, \ldots, d\}$ may have a variance as large as $n^2/4$.) Analogous variance bound for the conic intrinsic volume random variable can be found in Theorem 4.5 in [LMN$^+$20]. High probability bounds can then be deduced from the variance bound via the Chebyshev inequality.

Using Theorem 17 and the concentration of the intrinsic volume random variables $Z_{\mathcal{K}}$ and $Z_{\mathcal{C}}$, Theorem 15 can be proved using about one page (see [ALMT14]). Here we give a short heuristic proof just to illustrate the idea. Note that the double sum in (1.104) is over

$i$ and $j$ such that $i-j$ is odd. An interlacing property [ALMT14] shows that we can approximate by alleviating this parity constraint while reducing a factor of 2. Thus

$$\begin{aligned} \mathbb{P}[\mathcal{C} \cap Q\mathcal{K} \neq \{0\}] &\approx \sum_{i=0}^{d} \sum_{j=0}^{i-1} \tilde{V}_i(\mathcal{C}) \tilde{V}_{d-j}(\mathcal{K}) && (1.109) \\ &= \sum_{i=0}^{d} \sum_{l=0}^{d} \tilde{V}_i(\mathcal{C}) \tilde{V}_l(\mathcal{K}) 1_{i+l \geq d+1} && (1.110) \\ &= \mathbb{E}[1\{Z_{\mathcal{C}} + Z_{\mathcal{K}} \geq d+1\}] && (1.111) \\ &\approx 1\{\mathbb{E}[Z_{\mathcal{C}} + Z_{\mathcal{C}}] \geq d+1\} && (1.112) \end{aligned}$$

where the last approximation follows since $Z_{\mathcal{C}}$ and $Z_{\mathcal{K}}$ are close to their expectations with high probability (concentration). Theorem 15 then follows since $\mathbb{E}[Z_{\mathcal{C}}] = \delta(\mathcal{C})$.

# Chapter 2

# Graphical Lasso

This chapter is about graphical lasso, which is an algorithm for learning the covariance matrix under a sparsity assumption on the precision matrix (inverse of the covariance matrix).

## 2.1 Gaussian Graphical Model

Consider a random vector $X \in \mathbb{R}^p$ following the normal distribution $\mathcal{N}(0, \Sigma)$. The *precision matrix* is defined as

$$\Theta = \Sigma^{-1}. \tag{2.1}$$

We can construct an undirected graph $(\mathcal{V}, \mathcal{E})$ to encode the conditional independence structure of the coordinates of $X$: let $\mathcal{V} = \{1, \ldots, p\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ where

$$(u, v) \notin \mathcal{E} \Longleftrightarrow X_u \perp X_v | X_{\mathcal{V} \setminus \{u,v\}} \tag{2.2}$$

where $X_u \perp X_v | X_{\mathcal{V} \setminus \{u,v\}}$ denotes the conditional independence (that is, there is a Markov chain $X_u - X_{\mathcal{V} \setminus \{u,v\}} - X_v$). By the Hammersley-Clifford theorem[1], for any $u, v \in \{1, \dots, p\}$,

$$X_u \perp X_v | X_{\mathcal{V} \setminus \{u,v\}} \iff \Theta_{u,v} = 0. \tag{2.3}$$

In fact, for Gaussian distributions, we can directly prove (2.3) by computing the conditional covariance (Exercise 11). By (2.3), we see that in many applications, $\Theta$ is a sparse matrix, since the underlying graphs in many graphical models are sparse.

Suppose that there are $n$ i.i.d. samples $x^{(1)}$,. . . ,$x^{(n)}$ from the distribution $\mathcal{N}(0, \Theta^{-1})$, how can we estimate the covariance? The first idea is maximum likelihood estimation. Note that up to additive constants, the log-likelihood function is

$$\begin{aligned} \ell(\Theta) &= \sum_{i=1}^{n} \log \left( \det{}^{1/2}(\Theta) \exp \left( -\frac{1}{2} x^{(i)\top} \Theta x^{(i)} \right) \right) && (2.4) \\ &= \frac{n}{2} \log \det(\Theta) - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}(x^{(i)} x^{(i)\top} \Theta) && (2.5) \\ &= \frac{n}{2} \log \det(\Theta) - \frac{n}{2} \operatorname{tr}(S\Theta) && (2.6) \end{aligned}$$

where $S = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)\top}$ denotes the sample covariance matrix.

[1] The theorem states that for a general Markov network, the Markov and factorization properties are equivalent, if the probability distribution has a strictly positive mass or density (e.g. [LW13]). The latter is always fulfilled in the Gaussian case.

The classical theory tells that for fixed $p$, the maximum likelihood estimator (MLE) converges to the truth as $n \to \infty$. However, in the regime of $p > n$, MLE does not even exist (Exercise 12).

The *graphical lasso* [FHT08] addresses this issue by leveraging the sparsity of $\Theta$ and adding an $\ell_1$ penalty to the objective function:

$$\text{minimize}_{\Theta \succeq 0} \quad -\log\det\Theta + \operatorname{tr}(S\Theta) + \lambda\|\Theta\|_1. \tag{2.7}$$

Here, $A \succeq B$ means $A - B$ is a positive-semidefinite matrix; $\|\Theta\|_1$ denotes the sum of the absolute values of entries of $\Theta$, and $\lambda > 0$ is a tuning parameter. This is a convex optimization problem (Exercise 13) with $\sim p^2$ parameters to optimize.

## 2.2 Dual Formulation

The problem (2.7) is convex since each summand is convex in $\Theta$. The *dual* convex problem is the following:

$$\text{maximize}_{\Gamma\colon \|\Gamma\|_\infty \le \lambda} \quad \log\det(S + \Gamma) + p. \tag{2.8}$$

The equivalence between (2.7) and (2.8) can be shown by using the KKT condition to obtain properties of the optimizers [MH12]. Here, however, we introduce a systematic way of writing out the dual of any convex optimization:

**Theorem 18.** *Let $\Lambda_j \colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be convex functions, $j = 1, \ldots, k$ for some positive integer $k$. Suppose that there exist* *some $u_1, \dots, u_k$ such that $u_1 + \cdots + u_k = 0$ and $\Lambda_j(u_j) < \infty$, $j = 1, \dots, k$, and $\Lambda_j$ is upper semicontinuous at $u_j$ for some $j$. Then*

$$-\inf_{v \in \mathbb{R}^p} \sum_{j=1}^{k} \Lambda_j^*(v) = \inf_{u_1,\dots,u_k \in \mathbb{R}^p :\, u_1+\cdots+u_k=0} \sum_{j=1}^{k} \Lambda_j(u_j) \tag{2.9}$$

*where $\Lambda_j^*$ denotes the convex conjugate of $\Lambda_j$.*

Theorem 18 is a generalization of the Fenchel's duality theorem in convex analysis, the latter being the special case of $k = 2$ and also presented in a slightly different form. In fact, Theorem 18 holds in a very general setting where $\mathbb{R}^p$ is replaced by a general topological vector space; see Theorem 4 in [LCCV18].

Equipped with Theorem 18, we can easily show:

**Theorem 19.** (2.7) *and* (2.8) *are equivalent.*

*Proof.* We would like to have

$$\Lambda_1^*(\Theta) = -\log\det(\Theta); \tag{2.10}$$
$$\Lambda_2^*(\Theta) = \mathrm{tr}(S\Theta); \tag{2.11}$$
$$\Lambda_3^*(\Theta) = \lambda\|\Theta\|_1 \tag{2.12}$$

for $\Theta \in \mathbb{R}^{p^2}$; this is a $p^2$-dimensional space with inner product $\mathrm{tr}(\cdot)$. If $\Theta$ is not positive definite then it is understood that $\Lambda_1^*(\Theta) = +\infty$.

Assuming that the double-conjugates are the functions themselves, we can compute

$$\begin{aligned} \Lambda_1(\Gamma) &= \sup_{\Theta \in \mathbb{R}^{p^2}} \{\mathrm{tr}(\Gamma\Theta) + \log\det(\Theta)\} && (2.13) \\ &= \sup_{\Theta \text{ is psd}} \{\mathrm{tr}(\Gamma\Theta) + \log\det(\Theta)\} && (2.14) \\ &= -p - \log\det(-\Gamma); && (2.15) \end{aligned}$$

$$\begin{aligned} \Lambda_2(\Gamma) &= \sup_{\Theta \in \mathbb{R}^{p^2}} \{\mathrm{tr}(\Gamma\Theta) - \mathrm{tr}(S\Theta)\} && (2.16) \\ &= \begin{cases} 0 & \Gamma = S \\ \infty & \text{otherwise;} \end{cases} && (2.17) \end{aligned}$$

$$\begin{aligned} \Lambda_3(\Gamma) &= \sup_{\Theta \in \mathbb{R}^{p^2}} \{\mathrm{tr}(\Gamma\Theta) - \lambda\|\Theta\|_1\} && (2.18) \\ &= \begin{cases} 0 & \|\Gamma\|_\infty \le \lambda \\ \infty & \text{otherwise.} \end{cases} && (2.19) \end{aligned}$$

Indeed, after obtaining these formulae we can directly check that $\Lambda_j^*(\Theta) = \sup_\Gamma\{\mathrm{tr}(\Gamma\Theta) - \Lambda_j(\Gamma)\}$ is true. Now

$$\begin{aligned} &\sup_{\Gamma_1+\Gamma_2+\Gamma_3=0} \{-\Lambda_1(\Gamma_1) - \Lambda_2(\Gamma_2) - \Lambda_3(\Gamma_3)\} \\ &= \sup_{\|\Gamma_3\|_\infty \le \lambda, \Gamma_2 = S} \{p + \log\det(\Gamma_2 + \Gamma_3)\} && (2.20) \\ &= \sup_{\|\Gamma_3\|_\infty \le \lambda} \{p + \log\det(S + \Gamma_3)\} && (2.21) \end{aligned}$$

which is the formula for the dual of the graphical lasso. The equivalence to the graphical lasso is seen from Theorem 18. □

## 2.3 Blockwise Coordinate Descent

The problem (1.46) can be solved very quickly using the coordinate descent algorithm [WL$^+$08] whereby the coordinates are updated iteratively via line search. The graphical lasso problem (2.7) can also be solved by a blockwise coordinate descent algorithm (Friedman, Hastie, and Tibshirani [FHT08]), which can easily handle problem size of $p = 1000$.

The blockwise coordinate descent algorithm actually uses the lasso solver as a subroutine. The latter, of course, is convex optimization and can be solved in the primal or the dual form. Moreover, the problem (1.46) can also be solved in either the primal or the dual form. Thus there are at least four versions of the algorithm from these combinations [MH12]. Here, we briefly describe the one which is perhaps the simplest - the so called primal-glasso (P-lasso) in [MH12].

By setting the gradient in (2.7) to zero, we see that the optimizer $\Theta$ must satisfy

$$-\Theta^{-1} + S + \lambda \operatorname{sign}(\Theta) = 0. \tag{2.22}$$

The algorithm proceeds by iteratively updating $\Theta$, leveraging (2.22). Suppose that in $i$-th iteration, the matrix obtained previously is $\Theta^{(i)}$, and we pick an index $l \in \{1, \ldots, p\}$, and in $i$-th iteration we make updates on the $l$-th row and column of $\Theta$. The index $l$ is selected cyclically through the iterations; for simplicity of the subsequent discussions, we suppose that $l = p$ is the last index. Then the matrix $\Theta^{(i)}$ can then be viewed as a $2 \times 2$ block matrix, which can be represented as

$$\begin{pmatrix} \Theta_{11}^{(i)} & \Theta_{12}^{(i)} \\ \Theta_{21}^{(i)} & \Theta_{22}^{(i)} \end{pmatrix}. \tag{2.23}$$

Hereafter, note that notations such as $\Theta_{12}^{(i)}$ denote the submatrix in the block form (rather than a coordinate in the original matrix).

Checking on the condition (2.22) for the upper right block, we get

$$\frac{1}{\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}}\Theta_{11}^{-1}\Theta_{12} + S_{12} + \lambda \mathrm{sign}(\Theta_{12}) = 0 \tag{2.24}$$

where we have invoked the Schur complement theorem for the block matrix inverse. While (2.24) is a complicated formula for $\Theta$, the idea of iterative algorithms is to replace the complicated parts by the values of the previous round. We posit that

$$A\Theta_{12}^{(i)} + S_{12} + \lambda \mathrm{sign}(\Theta_{12}^{(i)}) = 0 \tag{2.25}$$

where

$$A := \frac{1}{\Theta_{22}^{(i-1)} - \Theta_{21}^{(i-1)}(\Theta_{11}^{(i-1)})^{-1}\Theta_{12}^{(i-1)}}(\Theta_{11}^{(i-1)})^{-1}. \tag{2.26}$$

However, (2.25) indicates that we can find $\Theta_{12}^{(i)}$ by solving the following lasso problem:

$$\text{minimize}_{\beta\in\mathbb{R}^{p-1}} \quad \frac{1}{2}\beta^\top A\beta + S_{12}^\top\beta + \lambda\|\beta\|_1. \tag{2.27}$$

Once $\Theta_{12}^{(i)}$ is obtained, we compute $\Theta_{22}^{(i)}$: Checking on the condition (2.22) for the lower right block, we get

$$(\Theta^{-1})_{22} = s_{22} + \lambda. \tag{2.28}$$

This suggests that we can compute

$$\Theta_{22}^{(i)} = \frac{1}{(\Theta^{-1})_{22}} + \Theta_{21}^{(i)}(\Theta_{11}^{(i-1)})^{-1}\Theta_{12}^{(i)} \tag{2.29}$$

according to the Schur complement theorem, where $(\Theta^{-1})_{22}$ is computed from (2.28). Finally, after $\Theta_{12}^{(i)}$ (and consequently, $\Theta_{21}^{(i)}$) and $\Theta_{22}^{(i)}$ are computed, the $i$-th iteration is completed. The iterates continues until convergence.

# Chapter 3

# Matrix Estimation

The problem of estimating a sparse vector discussed in Chapter 1 has been well studied, and many of its basic properties are now understood. In the recent years, matrices and tensors are becoming increasingly popular in high-dimensional statistics. Some argue that many practical problems are matrix recovery in disguise; one famous application is collaborative filtering (as popularized by the Netflix prize). For a recent video tutorial on the subject, see [YS18]. While matrices and tensors can certainly be thought of as extensions of vectors, not all the properties and results immediately carry over, and many basic statistical and computational questions about matrix and tensor estimation are under very active research today.

In this chapter we get a glimpse of matrix estimation through a basic problem of prediction in an additive noise model with linear measurements. There are other common variants of the problem, such as estimating a matrix from incomplete measurements of its entries, which we will not touch on. However, a common theme in these problems is that the low-rankness of the matrix is taken advantage of, just as sparsity is employed in vector estimation.

## 3.1 Multivariate Regression: Setup

We consider the following multivariate linear regression model:

$$\mathbb{Y} = \mathbb{X}\Theta^* + E \tag{3.1}$$

where $\mathbb{X} \in \mathbb{R}^{n\times d}$ is the design matrix, $\Theta^* \in \mathbb{R}^{d\times T}$ is the matrix of unknown parameters, $E$ is Gaussian noise matrix with independent $\mathcal{N}(0, \sigma^2)$ components, and $\mathbb{Y} \in \mathbb{R}^{n\times T}$ are the observations. The goal is prediction, i.e., to estimate $\mathbb{X}\Theta^*$ given $\mathbb{X}$ and $\mathbb{Y}$.

The following is an example scenario of application: suppose that there are $n$ movies, $d$ features of the movies (e.g., duration, music quality, plot quality...), and $T$ persons. The matrix $\mathbb{X}$ contains values quantifying the qualities of each movie regarding each feature. The matrix $\Theta^*$ contains values quantifying the preference of each person for each feature. Thus, it makes sense that $\mathbb{Y}$ quantifies the level of preference of each person for each movie.

Note that each row of $\Theta^*$ indicates how a particular feature is preferred by all the persons. It is possible that a certain feature (e.g. duration of the movie) is not quite relevant for the rating, in which case the corresponding row in $\Theta^*$ is zero. In general, we might expect that there are many zero rows in $\Theta^*$; in other words, the columns of $\Theta^*$ *share the same sparsity pattern.*

If we assume that there are at most $k$ nonzero rows in $\Theta^*$, then naively we can just apply Lasso for each column of $\Theta^*$ to obtain $\hat{\Theta}$. By Theorem 7, the mean square error is then

$$\mathbb{E}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \sigma^2 kT \log d \tag{3.2}$$

with constant probability. Here, $\|A\|_F := \sqrt{\sum_{ij} A_{ij}^2}$ denotes the Frobenius norm.

In the next section, we will show that we can do better than (3.2) by taking into account of the interactions of the columns of $\Theta^*$. First, the sparsity level $k$ can be reduced to the rank of $\Theta^*$. Second, the factor $\log d$, which is the price to pay for not knowing the sparsity pattern, can be get rid of.

## 3.2 Penalization by Rank

The low-rankness of $\Theta^*$ prompts the consideration of the following estimator:

$$\hat{\Theta}^{RK} := \operatorname{argmin}_{\Theta \in \mathbb{R}^{d\times T}} \left\{ \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\Theta\|_F^2 + 2\tau^2 \operatorname{rank}(\Theta) \right\}. \tag{3.3}$$

We call this *estimator by rank penalization with regularity parameter $\tau^2$.* At first sight, this looks similar to the BIC estimator introduced in Chapter 1. However, unlike BIC, $\hat{\Theta}^{RK}$ can be computed efficiently (Exercise 14).

Now, let us also show that $\hat{\Theta}^{RK}$ enjoys good statistical property[1]:

**Theorem 20.** *Let*

$$\tau := 4\sigma\sqrt{\frac{\log(12)\max\{d,T\}}{n}} + 2\sigma\sqrt{\frac{2\log(1/\delta)}{n}}.$$

*Then with probability* $1-\delta$,

$$\|\mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^*\|_F^2 \le 8n\,\mathrm{rank}(\Theta^*)\tau^2 \lesssim \sigma^2\,\mathrm{rank}(\Theta^*)(\max\{d,T\} + \log\frac{1}{\delta}). \tag{3.4}$$

It is interesting to note that $\mathbb{X}$ does not enter the bound.

*Proof.* As usual, the optimality condition implies that

$$\|\mathbb{Y} - \mathbb{X}\hat{\Theta}^{RK}\|_F^2 + 2n\tau^2\,\mathrm{rank}(\hat{\Theta}^{RK}) \le \|\mathbb{Y} - \mathbb{X}\Theta^*\|_F^2 + 2n\tau^2\,\mathrm{rank}(\Theta^*) \tag{3.5}$$

which is equivalent to

$$\begin{aligned}\|\mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^*\|_F^2 \le 2\left\langle E, \mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^*\right\rangle \\ - 2n\tau^2\,\mathrm{rank}(\hat{\Theta}^{RK}) + 2n\tau^2\,\mathrm{rank}(\hat{\Theta}^*).\end{aligned} \tag{3.6}$$

As usual, the inner product term is the main item we need to control; intuitively we need to leverage the "low complexity" of $\hat{\Theta}^{RK}$

[1] Adapted from Theorem 5.5 in [Rig15]

to show that the two terms in the inner product are approximately uncorrelated. By Young's inequality,

$$2\left\langle E, \mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^* \right\rangle \leq 2\left\langle E, U \right\rangle^2 + \frac{1}{2}\|\mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^*\|_F^2 \tag{3.7}$$

where

$$U := \frac{\mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^*}{\|\mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^*\|_F} \tag{3.8}$$

is a unit direction. We then "sup-out" the unit direction to decouple the noise $E$ and the optimizer $\hat{\Theta}^{RK}$. For notation simplicity, let us assume below that $\mathbb{X}$ is full column rank (if not, we can consider all inner products in the column space of $\mathbb{X}$ instead; this can be done as an exercise, or see Theorem 5.5 [Rig15]). Then

$$\begin{aligned}
\left\langle E, U \right\rangle^2 &\leq \|E\|_{\mathsf{op}}^2 \|U\|_{\mathsf{s1}}^2 && (3.9)\\
&\leq \|E\|_{\mathsf{op}}^2 \operatorname{rank}(U) && (3.10)\\
&= \|E\|_{\mathsf{op}}^2 \operatorname{rank}(\hat{\Theta}^{RK} - \Theta^*) && (3.11)\\
&\leq \|E\|_{\mathsf{op}}^2 (\operatorname{rank}(\hat{\Theta}^{RK}) + \operatorname{rank}(\Theta^*)) && (3.12)
\end{aligned}$$

where $\|\cdot\|_{\mathsf{s}p}$ denotes the Schatten $p$-norm; $\|\cdot\|_{\mathsf{op}} = \|\cdot\|_{\mathsf{s}\infty}$ denotes the operator norm of a matrix, i.e., the largest singular value. The first inequality above is the matrix Hölder inequality; the inequality follows since by Cauchy-Schwarz, $\|U\|_{\mathsf{s1}}^2 = \frac{\|U\|_{\mathsf{s1}}^2}{\|U\|_F^2} \leq \operatorname{rank}(U)$. Finally, random matrix theory provides us the estimate

$$\|E\|_{\mathsf{op}}^2 \leq n\tau^2 \tag{3.13}$$

for the top singular value with probability $1-\delta$, which gives

$$\langle E, U\rangle^2 \leq n\tau^2(\mathrm{rank}(\hat{\Theta}^{RK}) + \mathrm{rank}(\Theta^*)). \tag{3.14}$$

But (3.6) and (3.7) tell us

$$\|\mathbb{X}\hat{\Theta}^{RK} - \mathbb{X}\Theta^*\|_F^2 \leq 4\langle E, U\rangle^2 - 4n\tau^2\,\mathrm{rank}(\hat{\Theta}^{RK}) + 4n\tau^2\,\mathrm{rank}(\hat{\Theta}^*) \tag{3.15}$$

which, together with (3.14), gives the desired result. □

*Remark* 6. More general, the result of Theorem 20 holds in the more general setting where the error matrix $E$ is subgaussian, i.e., has gaussian-like tail when projected to any direction; see [Rig15].

*Remark* 7. While the rank-penalized estimator can be computed efficiently, it is also worth considering the "$\ell_1$ counterpart", that is,

$$\mathrm{argmin}_{\Theta\in\mathbb{R}^{d\times T}} \left\{ \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\Theta\|_F^2 + \tau\|\Theta\|_{\mathsf{s}1} \right\}. \tag{3.16}$$

Its error property is similar to (3.4) but with an additional multiplicative factor of the condition number of $\mathbb{X}$. Its analysis, however, is quite involved, rather than a simple adaptation of the proof for the error estimates in Lasso. See [KLT$^+$11].

## 3.3 Matrix Completion

Let $\Theta^* \in \mathbb{R}^{d\times T}$ be an unknown matrix. Let $Y$ be a set of incomplete observations of the entries of $\Theta^*$; we may think of $Y$ as a matrix obtained by replacing some entries of $\Theta^*$ by a question mark.

More generally, we may formulate a problem where $Y = \mathcal{A}(\Theta^*)$ is a set of linear measurements of $\Theta^*$:

$$Y_1 := \mathcal{A}_1(\Theta^*) = \langle A_1, \Theta^* \rangle \tag{3.17}$$

$$\dots \tag{3.18}$$

$$Y_n := \mathcal{A}_n(\Theta^*) = \langle A_n, \Theta^* \rangle \tag{3.19}$$

where $A_1, \dots, A_n \in \mathbb{R}^{d \times T}$.

If $\Theta^*$ is low-rank, we may recover it from $Y$ (and the known $A_1$,...,$A_n$) by the following

$$\text{argmin}_{\Theta\colon Y=\mathcal{A}(\Theta^*)} \text{rank}(\Theta). \tag{3.20}$$

Unfortunately, unlike the multivariate regression problem where the linear observation is of the specialized form $\mathcal{A}(\Theta^*) = \mathbb{X}\Theta^*$, we cannot solve (3.20) efficiently. Nevertheless, we can consider the convex relaxation

$$\text{argmin}_{\Theta\colon Y=\mathcal{A}(\Theta^*)} \|\Theta\|_{\mathsf{s}1} \tag{3.21}$$

which can also be recast as a semidefinite programming problem; see [CR09].

# Chapter 4

# Robust PCA

## 4.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a common dimension reduction technique. Given $n$ samples $\mathbb{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{n \times d}$ that are centered, PCA finds $r$ directions that best explain the most variance of the data:

$$\text{minimize}_{L:\ \text{rank}(L)=r} \|\mathbb{X} - L\|_F. \tag{4.1}$$

In other words, we find the best rank-$r$ approximation of $\mathbb{X}$ under the Frobenius norm. The optimization is easily solved by keeping the $r$ largest singular values, thanks to the rotation invariance of the Frobenius norm (Exercise 14).

## 4.2 Robust PCA

The optimization (4.1) fails when there are corrupted samples or outliers. In that case, the (additive) error can be thought of as a sparse matrix with possibly very large nonzero entries. The problem can therefore be formulated as disentangling sparse and low-rank matrices: Suppose we are given a matrix

$$M = L + S \in \mathbb{R}^{m,n} \tag{4.2}$$

where $L$ is low-rank and $S$ is sparse, can we recover both $L$ and $S$ from $M$? There are a few examples of applications:

**Example 21.** *Clustering/community recovery. Suppose that there are $n$ persons from $r$ communities, and the connectivity matrix $L$ is defined by*

$$L_{i,j} := 1\{i \textit{ and } j \textit{ are from the same community}\} \tag{4.3}$$

*for $i, j \in \{1, \dots, n\}$. Then $L$ has rank at most $r$ (why?). Suppose that we observe $M$ which is the connectivity structure but with a few flips. Then $S := M - L$ is sparse. Disentangling $L$ and $S$ will allow us to reconstruct the community structure.*

**Example 22.** *Videos surveillance. Suppose that the video data is stored in a matrix $M \in \mathbb{R}^{m\times n}$ where $m$ is the number of pixels in each frame and $n$ is the number of frames. The background* *does not change much across the frames, and hence should correspond to a low-rank matrix L (why?). On the other hand, the foreground (such as a person walking past) should correspond to a sparse matrix S since it is supported on a short space/time interval. Thus, disentangling S and L may allow us to extract useful information from the video.*

**Example 23.** *Graphical model with latent factors. Recall* (2.3) *that the zero pattern of the precision matrix* $\Theta$ *encodes the connectivity structure in the Gaussian graphical model. Now suppose that the Gaussian vector* $[X_{\mathcal{S}}, X_{\mathcal{L}}]$ *is from a large Gaussian graphical model, and* $\mathcal{S}$ *and* $\mathcal{L}$ *are some sets of indices. If this large graphical model is sparse, the precision matrix may not be sparse* $X_{\mathcal{S}}$ *(why?); however, if we further have that* $\mathcal{L}$ *is a small set, then*

$$\mathrm{Cov}^{-1}(X_{\mathcal{S}}) = \Theta_{\mathcal{S}} - \Theta_{\mathcal{S},\mathcal{L}}\Theta_{\mathcal{L}}^{-1}\Theta_{\mathcal{L},\mathcal{S}} \tag{4.4}$$

*is the sum of a sparse matrix and a low-rank matrix.*

### 4.2.1 Incoherence

Disentangling $L$ and $S$ is not always possible, since a matrix might be simultaneously low-rank and sparse. First, consider a matrix

$$M = uv^{\top} \tag{4.5}$$

where $u = v = [1, 0, 0, \dots, 0]$. Since $M$ is low-rank and sparse, either $L = M$, $S = 0$ or $L = 0$, $S = M$ is plausible. On the other hand, if $M = uv^\top$ still but $u = v = [1/\sqrt{n}, \dots, 1/\sqrt{n}]$, then $M$ is rank-1 but not sparse, hence we should expect $L = M$ and $S = 0$.

In general, the issue is that $u$ and $v$ should be incoherence with the basis used to measure sparsity. In the rest of the chapter, we shall assume $m = n$ in (4.2) for simplicity, though the results can be generalized, *mutatis mutandis*.

**Definition 24.** Let $M = U\Sigma V^\top$ be singular value decomposition, and $r = \operatorname{rank}(M)$. Define the coherence parameter $\mu_1$ as the smallest number such that

$$\max_i \|U^\top e_i\|_2^2 \leq \frac{\mu_1 r}{n}, \tag{4.6}$$

$$\max_i \|V^\top e_i\|_2^2 \leq \frac{\mu_1 r}{n}, \tag{4.7}$$

where $e_1, \dots, e_n$ form the standard basis.

**Definition 25.** Given $M = U\Sigma V^\top$, the joint coherence parameter $\mu_2$ is the smallest number such that

$$\|UV^\top\|_\infty \leq \sqrt{\frac{\mu_2 r}{n^2}}. \tag{4.8}$$

The parameters $\mu_1$ and $\mu_2$ are related: in general

$$\mu_1 \leq \mu_2 \leq \mu_1^2 r; \tag{4.9}$$

see Exercise 16.

Our goal is to show that $L$ and $S$ can be successfully disentangled when $\mu_1$ and $\mu_2$ are small. Naturally, we can solve this by a convex optimization, with 1-Schatten norm promoting low-rankness and the $L_1$-norm promoting sparsity. This leads to

$$\text{minimize}_{L,\,S}\|L\|_{\mathsf{s}1} + \lambda\|S\|_1 \text{ s.t. } M = L + S \tag{4.10}$$

where $\lambda > 0$ is a regularity parameter balancing the two terms. In [CLMW11], it is shown that the simple convex optimization (4.10) is, in some sense, nearly optimal:

**Theorem 26.** *Suppose that* $\operatorname{rank}(L) \lesssim \frac{n}{\max\{\mu_1,\mu_2\}\log^2 n}$*; the nonzero entries of $S$ are randomly located, and $\|S\|_0 \le \rho_s n^2$ for some constant $\rho_s > 0$. Let $M = L + S$. Then* (4.10) *recovers $L$ and $S$ with high probability.*

*Remark* 8. While $\mu_1$ and $\mu_2$ are related (4.15), the gap might be large. Is it possible to improve the rank constraint on $L$ in Theorem 26 to $\frac{n}{\mu_1 \text{polylog}(n)}$? This scaling is achievable for the related matrix completion problem (Section 3.3); see the result in [Che15]. However, it is not achievable for the robust PCA problem in (4.10), under the hardness assumption of the planted clique problem[1]; see the proof in [Che15]. More generally, this is along the line of the

[1] Karp [Kar77] conjectured in 1976 that there is no efficient algorithm for finding a planted clique of size $(1+\epsilon)\log_2 n$ in an Erdos-Renyi graph of size $n$ with edge connection probability 1/2. Nowadays, many people believe that in fact, there is no polynomial algorithm for finding

body of recent works about the average-case computational hardness of statistical problems under the hardness of the planted clique, which was initiated by [BR$^+$13].

Note that the rank of $L$ can be almost the largest possible (which is $n$ up to logarithmic factors), and the support size of the sparsity component $S$ is also almost the largest possible (which is order $n^2$). Also, the sparse component $S$ can have arbitrary signs and maginitudes.

The proof of Theorem 26 relies on the constructing dual certificate [CLMW11]. To get a flavor of why convex duality can be used in certifying the optimality of solution of convex optimization, see Exercise 15. We shall not include here the full proof of Theorem 26, which is a bit lengthy.

### 4.2.2 Guarantee via Convex Geometry

We shall demonstrate how the convex geometry tools in Section 1.6 can be used to bound the probability of successfully disentangling the sparse and the low-rank components. The setting here is taken from [ALMT14, Example 2.11], called rank-sparsity decomposition [CSPW11], which is a bit different from the setting in the preceding sections.

a planted clique of size $O(n^c)$ with $c < 1/2$, and have used this as an assumption in numerous proofs of average-case hardness results. However, there seems to be no good reason for this belief except that strong people have failed in finding such an algorithm. If you can find one, you will be instantly famous and find a very good job!

Suppose that we observe $M_0 = L_0 + \mathcal{U}(S_0)$ where $L_0$ is a low-rank matrix, $S_0$ is sparse, and $\mathcal{U}$ is a known orthogonal transformation on the linear space of matrices. Then we can attempt to disentangle $L_0$ and $S_0$ by solving the following convex optimization:

$$\text{minimize } \|L\|_{\mathsf{s1}} \text{ s.t. } \|S\|_1 \le \|S_0\|_1,\ M_0 = L + \mathcal{U}(S). \tag{4.11}$$

Note the differences from the setting in (4.10): here $\mathcal{U}$ is *known*, and it is an orthogonal transformation on the space of matrices (viewed as vectors in $\mathbb{R}^{n^2}$), not necessarily left and right multiplying orthogonal matrices as in SVD. In particular, $\mathcal{U}(S_0)$ may not be low-rank! Also, the sparsity level of the unknown $S_0$ is used as a parameter in (4.11).

Due to the convexity of (4.11), $L_0$ and $S_0$ are guaranteed to be the solution if there is no nonzero $\Delta \in \mathbb{R}^{n^2}$ such that the following holds: there exists $\epsilon > 0$ such that

$$\|L_0 + \epsilon\Delta\|_{\mathsf{s1}} \le \|L_0\|_{\mathsf{s1}}; \tag{4.12}$$

$$\|S_0 - \epsilon\mathcal{U}^{-1}(\Delta)\|_1 \le \|S_0\|_1. \tag{4.13}$$

Using the notation for the descent cone in Section 1.6, the above is equivalent to

$$\Delta \in \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}}); \tag{4.14}$$

$$-\mathcal{U}^{-1}(\Delta) \in \mathcal{D}(S_0, \|\cdot\|_1). \tag{4.15}$$

Since $\mathcal{U}$ is a random rotation, the condition is equivalent to saying that $\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})$ and a randomly rotated $\mathcal{D}(S_0, \|\cdot\|_1)$ have no nontrivial intersection. We can thus bound this probability using the tools in Section 1.6.

In [ALMT14], it is shown that if $r = \operatorname{rank}(L_0)$ scales linearly in $n$, then the statistical dimension of $\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s}1})$ is order $n^2$. The formula for the prefactor of $n^2$ is also obtained, but rather complicated, and the analysis is intricate. In this note, we give a simple derivation of a slighter cheaper version of the result. Using an argument similar to Theorem 14, we show that the statistical dimension of $\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s}1})$is $O(rn)$ (with a possibly worse prefactor).

**Theorem 27.** *Let $L_0 \in \mathbb{R}^{n^2}$ be a rank-r matrix. Then we can bound the Gaussian width*

$$w(\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s}1}) \cap B_2) = O(\sqrt{rn}) \tag{4.16}$$

*where $B_2$ denotes the unit ball (under the Frobenius norm) in the space of matrices. By Exercise 9, we can also bound the statistical dimension as $O(rn)$.*

*Proof.* The proof is similar to the proof of Theorem 14. First, by applying SVD, we can assume without loss of generality that

$$L_0 = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \tag{4.17}$$

where $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix with positive diagonal entries. The subgradient of $\|\cdot\|_{\mathsf{s}1}$ at $L_0$ consists of all matrices of the following form:

$$\begin{pmatrix} I_{r\times r} & 0 \\ 0 & A \end{pmatrix} \tag{4.18}$$

where $A \in \mathbb{R}^{(n-r)\times(n-r)}$ is an arbitrary matrix with operator norm bounded in $[-1, 1]$ (Exercise 17). Thus $\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s}1})$ consists all matrices of the form

$$\Delta = \begin{pmatrix} \Delta_1 & \Delta_2 \\ \Delta_3 & \Delta_4 \end{pmatrix} \tag{4.19}$$

where

$$\mathrm{tr}(\Delta_1) + \|\Delta_4\|_{\mathsf{s}1} \le 0. \tag{4.20}$$

Now let $G \in \mathbb{R}^{n^2}$ be a random matrix with i.i.d. $\mathcal{N}(0, 1)$ entries. Suppose that in the block form,

$$G = \begin{pmatrix} G_1 & G_2 \\ G_3 & G_4 \end{pmatrix}. \tag{4.21}$$

We have

$$\sup_{\Delta \in B_2 \cap \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s}1})} \langle G_1, \Delta_1 \rangle \le \|G_1\|_F. \tag{4.22}$$

Similarly,

$$\sup_{\Delta \in B_2 \cap \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s}1})} \langle G_2, \Delta_2 \rangle \le \|G_2\|_F; \tag{4.23}$$

$$\sup_{\Delta \in B_2 \cap \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s}1})} \langle G_3, \Delta_3 \rangle \le \|G_3\|_F. \tag{4.24}$$

Moreover,

$$\begin{aligned}
\sup_{\Delta \in B_2 \cap \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})} \langle G_4, \Delta_4 \rangle &\leq \|G_4\|_{\mathsf{s}\infty} \sup_{\Delta \in B_2 \cap \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})} \|\Delta_4\|_{\mathsf{s1}} && (4.25)\\
&\leq \|G_4\|_{\mathsf{s}\infty} \sup_{\Delta \in B_2 \cap \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})} \operatorname{tr}(-\Delta_1) && (4.26)\\
&\leq \|G_4\|_{\mathsf{s}\infty} \sqrt{r}. && (4.27)
\end{aligned}$$

Thus,

$$\begin{aligned}
&w(\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}}) \cap B_2)\\
&= \mathbb{E}[\sup_{\Delta \in B_2 \cap \mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})} \langle G, \Delta \rangle] && (4.28)\\
&\leq \mathbb{E}[\|G_1\|_F] + \mathbb{E}[\|G_2\|_F] + \mathbb{E}[\|G_3\|_F] + \sqrt{r}\mathbb{E}[\|G_4\|_{\mathsf{s}\infty}]. && (4.29)\\
&\leq \sqrt{nr} + \sqrt{r} \cdot O(\sqrt{n-r}) && (4.30)\\
&\leq O(\sqrt{nr}) && (4.31)
\end{aligned}$$

The step (4.30) used results about the top singular value in a random matrix, which follows from a standard $\epsilon$-net argument (e.g. [vH14]).

□

**Theorem 28.** *Suppose that we observe $M_0 = L_0 + \mathcal{U}(S_0) \in \mathbb{R}^{n^2}$, where*

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{rank}(L_0) = 0; \qquad (4.32)$$

$$\lim_{n\to\infty} \frac{\log n}{n^2} \|S_0\|_1 = 0, \qquad (4.33)$$

*and $\mathcal{U}$ is a known orthogonal transformation on the linear space of matrices. Then* (4.11) *exactly recovers $L_0$ and $S_0$ with probability tending to 1 as $n \to \infty$.*

*Proof.* In Theorem 14 and Theorem 27, we have shown the bounds on statistical dimensions:

$$\delta(\mathcal{D}(S_0, \|\cdot\|_1)) = O(\|S_0\|_1 \log n); \tag{4.34}$$

$$\delta(\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})) = O(\mathrm{rank}(L_0)n). \tag{4.35}$$

Under our assumptions on the rank and sparsity, we have

$$\delta(\mathcal{D}(S_0, \|\cdot\|_1)) + \delta(\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})) = o(n^2) \tag{4.36}$$

which implies that $-\mathcal{U}(\mathcal{D}(S_0, \|\cdot\|_1))$ and $\mathcal{D}(L_0, \|\cdot\|_{\mathsf{s1}})$ have only trivial intersection with probability tending to 1 (Theorem 15). These in turn imply the exact recovery of $L_0$ and $S_0$ according to the analysis around (4.14)-(4.15). □

# Chapter 5

# Lower Bounds

Lower bounds on statistical risks generally rely on information-theoretic techniques. Consider, for example, the Gaussian sequence model in Chapter 1. For an arbitrary $\theta \in \mathbb{R}^d$ and i.i.d. $\mathcal{N}(0, \sigma^2)$ noise, the naive estimator $\hat{\theta} = Y$ or the James-Stein estimator give

$$\mathbb{E}\|\hat{\theta} - \theta\|_2^2 = O(d\sigma^2). \tag{5.1}$$

Moreover, for $\theta \in \mathcal{B}_0(k)$, we can use a hard thresholding estimator with threshold $\tau = \Theta(\sigma\sqrt{\log d})$ to achieve

$$\|\hat{\theta} - \theta\|_2^2 = O(\sigma^2 k \log d). \tag{5.2}$$

with constant (say 0.9) probability; see Theorem 1. From the proofs of (5.1) and (5.2) we can see that they are in fact the true scalings of the risks of these estimator (assuming $k$ is much smaller than $d$). But can there be *other* estimators that achieve better scalings?

In this chapter, we will introduce general tools for lower bounds which, in particular, show that no estimators can perform strictly better than the scalings in (5.1) and (5.2). Three methods for lower bounds are widely known to statisticians: Le Cam's, Assouad's and Fano's [Yu97]. These methods are based on inequalities connecting information-theoretic measures (e.g. KL divergence) and operational quantities (estimation or testing errors). Among them, Fano's is simple enough, and, in some sense, stronger than Le Cam's and Assouad's (see [Yu97, p428], which is attributed to Birgé). We will only cover Fano's method. As we will see, the heart of the game is to construct an appropriate testing problem corresponding to the original estimation problem.

## 5.1 From Estimation to Testing

In this section we consider a general estimation problem where the parameter space $\Theta$ is a metric space equipped with metric $d$. The observed random variable $Y$ follows a known probability distribution $P_\theta$ once the parameter $\theta \in \Theta$ is specified. The idea of lower bounding the estimation error is find a discrete subset of $\Theta$, called *packing*, such that distinct parameters $\theta$ and $\theta'$ in the packing are well-separated under the metric distance. Then, a good estimator can be easily converted to a good $M$-ary testing scheme where the hypotheses are elements in the packing.

**Definition 29.** A subset $\mathcal{A}$ of $\Theta$ is called an *$\epsilon$-packing*, if for any $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$, we have

$$d(\theta_1, \theta_2) > \epsilon. \tag{5.3}$$

**Theorem 30.** *Suppose that $\mathcal{A} = \{\theta_1, \ldots, \theta_M\}$ is a $2D$-packing ($D > 0$), and that there exists an estimator $\hat{\theta} = \hat{\theta}(Y)$ such that*

$$\inf_{\theta \in \Theta} \mathbb{P}[d(\hat{\theta}, \theta) \leq D] \geq 1 - \delta \tag{5.4}$$

*for some $\delta \in [0, 1]$. Then there exists a map $\psi \colon \mathcal{Y} \to \{1, \ldots, M\}$ such that*

$$\min_{j \in \{1, \ldots, M\}} P_{\theta_j}[\psi(Y) = j] \geq 1 - \delta. \tag{5.5}$$

*Proof.* We simply choose

$$\psi(Y) := \operatorname{argmin}_{j=1,\ldots,M} d(\hat{\theta}(Y), \theta_j) \tag{5.6}$$

and break ties arbitrary (if exist). By assumption, for each $j \in \{1, \ldots, M\}$ and with probability $1 - \delta$ we have $d(\hat{\theta}(Y), \theta_j) \leq D$. Since $\mathcal{A}$ is a $2D$-packing, by the triangle inequality of the metric we see that

$$d(\hat{\theta}(Y), \theta_i) > D \tag{5.7}$$

for any $i \neq j$. Therefore $\psi(Y) = j$, and the claim follows. □

## 5.2 Fano's Inequality

The next ingredient for the lower bound is to show the impossibility of good $M$-ary hypothesis testers. This relies on the Fano inequality in information theory. First, let us introduce a few notations in information theory.

Given two probability measures $P$ and $Q$ on the same measurable space $\mathcal{X}$, define the Kullback-Leibler (KL) divergence

$$D(P\|Q) := \int \log \frac{dP}{dQ}(x) dP(x) \tag{5.8}$$

where $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative, if $P$ is absolutely continuous with respect to $Q$, and $+\infty$ otherwise. By Jensen's inequality, it is easy to see that $D(P\|Q)$ is always nonnegative. Moreover, $D(P\|Q) = 0$ if and only if $P = Q$.

Given random variables $(W, Y)$ on the measurable space $\mathcal{W} \times \mathcal{Y}$ with distribution $P_{WY}$, define the mutual information

$$I(W; Y) = D(P_{WY}\|P_W \times P_Y) \tag{5.9}$$

where $P_W$ and $P_Y$ denote the marginal distributions. If $W$ is discrete, we also have[1]

$$I(W; Y) = \sum_w D(P_{Y|W=w}\|P_Y)P_W(w); \tag{5.10}$$

[1] If $W$ is not discrete, we may replace the sum in (5.10) by an integral. There is a slight issue, though, that $D(P_{Y|W=w}\|P_Y)$ is not necessarily a measurable function in $w$ (in some artificially designed but practically uncommon examples), in which case the Lebesgue integral is not well-defined.

see Exercise 18.

**Theorem 31** (Fano's inequality)**.** *Suppose that $(W, Y)$ is a pair of random variables, $W$ is equiprobable on $\{1, \dots, M\}$, and $\psi \colon \mathcal{Y} \to \{1, \dots, M\}$ is map. Let $\delta := \mathbb{P}[\psi(Y) \neq W]$. Then*

$$I(W; Y) \geq (1 - \delta) \log M - h(\delta) \tag{5.11}$$

*where $h(\delta) := \delta \log \frac{1}{\delta} + (1 - \delta) \log \frac{1}{1-\delta}$ denotes the binary entropy function.*

*Proof.* Let $(\bar{W}, \bar{Y})$ be a pair of random variables with the same marginal distributions as $(W, Y)$, but $\bar{W}$ and $\bar{Y}$ are independent. In other words, $P_{\bar{W}\bar{Y}} = P_W P_Y$. Moreover, define the indicators

$$\begin{align} E &:= 1\{W \neq \psi(Y)\}; \tag{5.12} \\ \bar{E} &:= 1\{\bar{W} \neq \psi(\bar{Y})\}. \tag{5.13} \end{align}$$

$$\begin{align} I(W; Y) &= D(P_{WY} \| P_{\bar{W}\bar{Y}}) \tag{5.14} \\ &\geq D(P_E \| P_{\bar{E}}) \tag{5.15} \\ &= \delta \log \frac{\delta}{1 - \frac{1}{M}} + (1 - \delta) \log \frac{1 - \delta}{1/M} \tag{5.16} \\ &= -h(\delta) + (1 - \delta) \log M \tag{5.17} \end{align}$$

where (5.15) follows by the data processing inequality of the KL divergence, since $E$ and $\bar{E}$ are the same functions applied to $(W, Y)$ and $(\bar{W}, \bar{Y})$ (why?). □

Roughly speaking, Fano's inequality tells us that if $M$-ary hypothesis testing with constant $\delta \in (0,1)$ average error probability is possible, then $\log M$ cannot exceed the mutual information (up to a constant factor).

In statistical applications, it is often convenient to weaken Fano's inequality by bounding the mutual information with some KL divergence, which may be easier to compute:

**Corollary 32.** *Consider the setting of Theorem 31, and suppose that $Q$ is an arbitrary distribution on $\mathcal{Y}$. We have*

$$\frac{1}{M}\sum_{j=1}^{M} D(P_j\|Q) \ge (1-\delta)\log M - h(\delta). \tag{5.18}$$

*where $P_j = P_{Y|W=j}$.*

*Proof.*

$$I(W;Y) = D(P_{WY}\|P_W \times P_Y) \tag{5.19}$$
$$\le D(P_{WY}\|P_W \times Q_Y) \tag{5.20}$$

which is the left side of (5.18) (why?). □

*Remark* 9. Unfortunately, in most of the current statistics literature, Fano's inequality is stated in a weaker form involving pairwise KL-divergences $\frac{1}{M^2}\sum_{i,j=1}^{M} D(P_i\|P_j)$ (see e.g. [Rig15]). Although this form appears more symmetrical than Theorem 31 and is useful in many applications, it is not strong enough in some applications (e.g. [TLR21]). In contrast, (32) appears to be more convenient in that the reference measure $Q$ can be chosen freely, and in particular, one such that $D(P_j \| Q)$ is easy to compute.

*Remark* 10. Alternatively, we could have directly proved Corollary 32 by letting $\bar{Y} \sim Q$ in the proof of Theorem 31.

## 5.3 Construction of $M$-ary Tests

The last technical ingredient needed is a method of constructing the "right" packing $\mathcal{A}$ of $\Theta$. Note that by Theorem 30, we want $\mathcal{A}$ to be such that distinct elements in $\mathcal{A}$ are well-separated in the metric distance. On the other hand, by Corollary 32, we also want the elements in $\mathcal{A}$ to be close in the sense that $D(P_\theta \| Q)$ is small for some $Q$. We must keep in mind this tension when constructing $\mathcal{A}$.

Constructing a packing is the name of game in coding theory. The following basic result, called Varshamov-Gilbert bound, is useful in many statistical applications.

**Lemma 33** (Varshamov-Gilbert)**.** *There exist positive constants $C_1$ an $C_2$ such that the following holds for any integers $k$ and $d$ such that $1 \leq k \leq d/8$. There exist vectors $\omega_1, \ldots, \omega_M \in \{0, 1\}^d$ such that*

- *The Hamming distance*

$$\rho(\omega_i, \omega_j) := \sum_{l=1}^{d} 1\{\omega_{il} \neq \omega_{jl}\} \tag{5.21}$$

*is at least* $\frac{k}{2}$ *for all* $i \neq j$*;*

- $\log M \geq \frac{k}{8} \log\left(1 + \frac{d}{2k}\right)$*;*
- $|\omega_j|_0 = k$ *for all* $j$*.*

It is easy to see the following relation of the Hamming distance and the $\ell_2$-distance (the metric in the parameter space):

$$\rho(\omega_i, \omega_j) = \|\omega_i - \omega_j\|_2^2. \tag{5.22}$$

Therefore, we see that Lemma 33 is useful for our construction of packing under the metric distance.

Lemma 33 shows us we can find many binary vectors in $\mathbb{R}^d$ with Hamming weight $k$ such that the pairwise distance is large. How good is this lemma? To fix ideas, imagine if we had chosen $\mathcal{B}$ to be the set of all weight $k$ binary vectors instead. Then the cardinality is (assuming $k \leq d/8$)

$$\log |\mathcal{B}| = dh(k/d) + o(\log d) = \Theta(k \log \frac{d}{k}) \tag{5.23}$$

where we recall that $h(\cdot)$ denotes the binary entropy function (Exercise 21). Thus, the size of the packing $M$ in Lemma 33 is already the largest possible, up to multiplicative factors. However, for the naive choice $\mathcal{B}$ the pairwise Hamming distance of distinct elements can be as small as 1, whereas in Lemma 33 it is $k/2$, which is (up to a factor of 4) the largest possible for $k$-sparse binary vectors! This is the content of Lemma 33.

*Proof sketch of Lemma 33.* We choose $\omega_1, \ldots, \omega_M$ by the following simple algorithm: For $i = 1, \ldots, M$, choose $\omega_i$ as any element in

$$\mathcal{B} \setminus \bigcup_{j=1}^{i-1} B(\omega_j, k/2) \tag{5.24}$$

where $\mathcal{B}$ is the set of all $k$-sparse binary vectors, and $B(\omega_j, k/2)$ denotes the Hamming ball centered at $\omega_j$ with radius $k/2$. It remains to show that (5.24) is nonempty for $i = 1, \ldots, M$, which follows by a simple *volume argument*: As mentioned, the total number of $k$-sparse binary vectors is $|\mathcal{B}| = \exp(dh(k/d) + O(\log d))$, but

$$\bigcup_{j=1}^{i-1} B(\omega_j, k/2) \leq M \exp\left(dh(\frac{k}{2d}) + O(\log d)\right). \tag{5.25}$$

Therefore, (5.24) is nonempty as long as

$$\begin{aligned} M &\leq \exp\left(dh(\frac{k}{d}) - dh(\frac{k}{2d}) + O(\log d)\right) && (5.26)\\ &= \exp\left(k \log\frac{d}{k} - \frac{k}{2}\log\frac{2d}{k} + O(\log d)\right) && (5.27)\end{aligned}$$

$$= \exp\left(k \log \frac{d}{k} + O(\log d)\right). \tag{5.28}$$

□

## 5.4 Lower Bound for $\Theta = \mathbb{R}^d$

In this section we show the tightness of (5.1) in the Gaussian sequence model. Let $\{\omega_1 \dots \omega_M\}$ be as in Lemma 33 with $k = d/8$. Then Lemma 33 guarantees that

$$\|\omega_i - \omega_j\|_2^2 \geq \frac{k}{2} = \frac{d}{16}, \quad i \neq j; \tag{5.29}$$

$$\log M \geq \frac{k}{8} \log(1 + \frac{d}{2k}) = \frac{d}{64} \log 5. \tag{5.30}$$

We then choose

$$\theta_j := \beta\sigma\omega_j, \quad j = 1, \dots, M \tag{5.31}$$

where $\beta > 0$ is a constant to be chosen later.

Then we let $W$ be equiprobable on $\{1, \dots, M\}$, and let $Y \sim P_{\theta_j} = \mathcal{N}(\theta_j, \sigma^2 I_d)$ conditioned on $W = j$. By Corollary 32, any test $\psi\colon \{1, \dots, M\}$ must satisfy

$$\mathbb{P}[\psi(Y) = W] \log M - 1 \leq \inf_Q \frac{1}{M} \sum_{j=1}^{M} D(P_{\theta_j} \| Q). \tag{5.32}$$

While the optimal $Q$ in (5.32) is $\frac{1}{M}\sum_{j=1}^{M} P_{\theta_j}$, we would rather choose $Q = P_0 = \mathcal{N}(0, \sigma^2 I_d)$, from which we can simply compute

$$\frac{1}{M}\sum_{j=1}^{M} D(P_{\theta_j}\|Q) = D(P_{\theta_1}\|P_0) = \frac{d\beta^2}{16}\log e. \tag{5.33}$$

Thus

$$\mathbb{P}[\psi(Y) = W] \le 4\beta^2 \log_5 e + O(1/d). \tag{5.34}$$

Then invoking Theorem 30 with $D = \frac{\sigma\beta\sqrt{d}}{8}$, we see that any estimator $\hat{\theta}$ must satisfy

$$\|\hat{\theta} - \theta\|_2 \ge \frac{\sigma\beta\sqrt{d}}{8} \tag{5.35}$$

with probability at least $1 - 4\beta^2 \log_5 e + O(1/d)$. To sum up, we have:

**Theorem 34.** *In the Gaussian sequence model, for any $\delta \in (0,1)$, there exists $c > 0$ such that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{P}[\|\hat{\theta} - \theta\|_2 \ge c\sigma\sqrt{d}] \ge 1 - \delta \tag{5.36}$$

*where $\inf_{\hat{\theta}}$ means the infimum over all estimators $\hat{\theta}\colon \mathcal{Y} \to \Theta$.*

## 5.5 Lower Bound for $\Theta = \mathcal{B}_0(k)$

In this section we show the tightness of (5.2) in the Gaussian sequence model. The proof is similar to the preceding section, but we choose $\{\omega_1 \dots \omega_M\}$ as in Lemma 33 with arbitrary $k \leq d/8$. Most of the previous computations carry over. We have

$$\begin{aligned}
\mathbb{P}[\psi(Y) = W] \log M - 1 &\leq \inf_Q \frac{1}{M} \sum_{j=1}^{M} D(P_{\theta_j} \| Q). && (5.37) \\
&= D(P_{\theta_j} \| P_0) && (5.38) \\
&= \frac{k\beta^2}{2} \log e. && (5.39)
\end{aligned}$$

Thus

$$\mathbb{P}[\psi(Y) = W] \leq \frac{\frac{k\beta^2}{2} \log e}{\frac{k}{8} \log(1 + \frac{d}{2k})} + O(1/d). \tag{5.40}$$

Invoking Theorem 30 with $D = \beta\sigma\sqrt{\frac{k}{8}}$, we see that any estimator $\hat{\theta}$ must satisfy

$$\|\hat{\theta} - \theta\|_2 \geq \beta\sigma\sqrt{\frac{k}{8}} \tag{5.41}$$

with probability at least $1 - \frac{\frac{k\beta^2}{2} \log e}{\frac{k}{8} \log(1+\frac{d}{2k})} + O(1/d)$. By choosing $\beta$ to be order $\sqrt{\log \frac{d}{k}}$, we have:

**Theorem 35.** *In the Gaussian sequence model, for any* $\delta \in (0,1)$*, there exists* $c > 0$ *such that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathcal{B}_0(k)} \mathbb{P}\left[\|\hat{\theta} - \theta\|_2 \geq c\sigma\sqrt{k \log \frac{d}{k}}\right] \geq 1 - \delta. \tag{5.42}$$

*Remark* 11. In the regime of $k = d^{1-\epsilon}$ where $\epsilon > 0$, we have $\log \frac{d}{k} = \Theta(\log d)$, therefore Theorem 35 matches (5.2).

# Chapter 6

# Leave-one-out

The first a few chapters have discussed techniques for bounding the magnitude of errors in regression. Starting from this chapter, we will introduce a few common methods for establishing the *asymptotic empirical distribution* of the recovered signal $\hat{\theta}$. This can be viewed as more refined results, since asymptotic empirical distribution implies the asymptotic magnitude of error. On the other hand, the knowledge of the asymptotic empirical distribution is very useful for some statistical applications, such as constructing confidence intervals [JM14][CMW20] or false discovery rate control [LR19].

This chapter introduces the *leave-one-out* technique. It is more or less similar to a few other concepts in different contexts, such as cavity method, Thouless-Anderson-Palmer (TAP), coordinate descent, or predecessor comparison. To put it very simply, it is similar to how we solve a fixed-point equation in order to determine the limit of a sequence of numbers defined by an inductive formula.

## 6.1 A Pedagogical Example

Let us consider the problem of asymptotic empirical distribution in least squares. The setting is the following:

**(1)** $\mathbf{A}$: $n \times p$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries.

**(2)** $n/p = \delta > 1$ is fixed, and $n, p \to \infty$.

**(3)** $\mathbf{w} \in \mathbb{R}^n$: coordinates are i.i.d. $\mathcal{N}(0, n)$.

**(4)** $\hat{\theta} := \operatorname{argmin}_{\theta \in \mathbb{R}^p} \|\mathbf{w} - \mathbf{A}\theta\|_2 = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{w}$.

In other words, $\hat{\theta}$ is the solution to least squares regression when the ground truth $\theta = 0$. This is without loss of generality, since for general $\theta$ we have

$$(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top (\mathbf{A}\theta + \mathbf{w}) = \theta + (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{w}, \tag{6.1}$$

so the general result is obtained simply by a translation.

Also comment that while the scaling $w_i \sim \mathcal{N}(0, n)$ may appear strange at first sight, it is the right scaling ensuring that $(\hat{\theta}_i)_{i=1}^p$ has a nontrivial asymptotic empirical distribution.

Goal: we will show that the empirical distribution of $(\hat{\theta}_i)_{i=1}^p$ tends to $\mathcal{N}(0, \frac{\delta}{\delta - 1})$ (in the sense of weak convergence, asymptotically almost surely).

How nontrivial is the claimed result about the convergence of the empirical distribution to $\mathcal{N}(0, \frac{\delta}{\delta-1})$? Actually, by rotation invariance it is easy to see that the following vectors are equal in distribution:

$$\hat{\theta} \overset{\text{distribution}}{=} \|\hat{\theta}\|_2 \mathbf{u} \tag{6.2}$$

where $\mathbf{u} \in \mathbb{R}^{\mathbf{p}}$ is a vector sampled from the unit sphere uniformly at random and independent of $\hat{\theta}$ (see Exercise 24). Thus, since $\|\hat{\theta}\|_2$ is concentrated around its mean and the empirical distribution of the coordinates of $\mathbf{u}$ is close to Gaussian (with high probability), it is not hard to see that the limit law must be Gaussian. On the other hand, it is nontrivial to show that the variance is $\frac{\delta}{\delta-1}$. In the next section, we will present a general result on the asymptotic empirical distribution via the *leave-one-out technique*, which, in particular, implies that the variance is $\frac{\delta}{\delta-1}$.

Before turning to the leave-one-out analysis, let us describe how to show that the variance is $\frac{\delta}{\delta-1}$ using the *Marchenko-Pastur Law* about the asymptotic empirical distribution of the singular values of random matrices, which is perhaps the most natural idea for this toy problem. Note that

$$\begin{aligned}
\frac{1}{p}\mathbb{E}[\hat{\theta}^\top \hat{\theta} | \mathbf{A}] &= \frac{1}{p} \operatorname{tr}\left((\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \operatorname{Cov}(\mathbf{w}) [(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top]^\top\right) && (6.3)\\
&= \frac{n}{p} \operatorname{tr}((\mathbf{A}^\top \mathbf{A})^{-1}) && (6.4)\\
&= \frac{1}{p} \operatorname{tr}((\frac{1}{n}\mathbf{A}^\top \mathbf{A})^{-1}) && (6.5)
\end{aligned}$$

$$= \int \frac{1}{x} \mu_p(\mathrm{d}x), \tag{6.6}$$

where $\mu_p$ denotes the empirical distribution of the eigenvalues of $\frac{1}{n}\mathbf{A}^\top\mathbf{A}$. By the Marchenko-Pastur law from the random matrix theory (e.g. [Tao12]), we have that $\mu_p$ convergences to the following limiting distribution (weakly and asymptotically almost surely):

$$\mu(\mathrm{d}x) = \frac{1}{2\pi} \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x} \mathrm{d}x \tag{6.7}$$

where $\lambda := \delta^{-1} < 1$ and $\lambda_\pm := (1 \pm \sqrt{\lambda})^2$. Using change of variables $(1-\lambda)^2/x \to t$ we get

$$\int \frac{1}{x} \mu(\mathrm{d}x) = \frac{\delta}{\delta - 1} \int \mu(\mathrm{d}t) = \frac{\delta}{\delta - 1} \tag{6.8}$$

which is the promised formula for the asymptotic variance. The above derivation is simple, but it employed the nontrivial Marchenko-Pastur law from random matrix theory. Many of the techniques we discuss in this and the next a few chapters have deep connections to random matrix theory. In fact, one derivation of the Marchenko-Pastur law based on "predecessor comparison" [Tao12, P172-177] can be viewed as a version of the leave-one-out analysis.

## 6.2 Optimization with Random Instances

Let us consider **A**, **w** as defined in (1)-(3) from the previous section. Define

$$\epsilon := \frac{1}{\sqrt{n}}\mathbf{w} \in \mathbb{R}^n \tag{6.9}$$

and

$$\hat{\beta} := \text{argmin}_\beta \sum_{i=1}^{n} \rho(\epsilon_i - A_i^\top \beta), \tag{6.10}$$

where $\rho\colon \mathbb{R} \to [0, \infty)$ is a given convex function, and $A_i$ denotes the $i$-th sample (the transpose of the $i$-th row of $\mathbf{A}$). With such rescaling, $\epsilon_i$, $\epsilon_i - A_i^\top \beta$ and $\|\beta\|_2$ all have the order of $\Theta(1)$ as $p \to \infty$. Clearly, the previous least squares example is the special case where $\rho(t) = t^2$ and $\hat{\theta} = \sqrt{n}\hat{\beta}$.

### 6.2.1 Asymptotic Distribution: A General Claim

In the remainder of the chapter, we discuss a result of El Karoui et al. [EKBB$^+$13], which characterized the limit of $\|\hat{\beta}\|_2$ as $p \to \infty$.

First, let us define the prox operator which is useful in convex optimizations:

$$\text{prox}_c(\rho)(x) := \text{argmin}_y\{\rho(y) + \frac{1}{2c}(x - y)^2\} \tag{6.11}$$

for any $c > 0$ and $x \in \mathbb{R}$. Intuitively, $\text{prox}_c(\rho)(x)$ is a point close to $x$, but also regularized by $\rho$. As examples, taking $c$ and $\rho$ appropriately, we can recover the soft and hard thresholding operators (Exercise!)

The leave-one-out analysis in [EKBB$^+$13] shows that the asymptotic limit of $\|\beta\|_2$ can be computed from a pair of fixed point equations.

**Theorem 36.** *Let the scalar $\epsilon \sim \mathcal{N}(0,1)$ and let $Z \sim \mathcal{N}(0,r^2)+\epsilon$, where $r := \lim_{p\to\infty} \|\hat{\beta}\|$ and $\hat{\beta}$ is as in* (6.10)*. Then, there exists some $c > 0$ such that*

$$\mathbb{E}[(\text{prox}_c(\rho))'(Z)] = 1 - \delta^{-1}; \tag{6.12}$$

$$\delta^{-1} r^2 = \mathbb{E}[(Z - \text{prox}_c(\rho)(Z))^2]. \tag{6.13}$$

For a general convex $\rho$, we can solve (6.12) and (6.13) to find $c$ and $r$. For the particular case of least squares where $\rho(x) = x^2$, we have

$$\text{prox}_c(\rho)(x) := \text{argmin}_y\{y^2 + \frac{1}{2c}(x-y)^2\} \tag{6.14}$$

$$= \frac{x}{2c+1}; \tag{6.15}$$

$$Z \sim \mathcal{N}(0, 1+r^2). \tag{6.16}$$

Therefore the fixed-point equations become

$$\begin{cases} \frac{1}{2c+1} = 1 - \frac{1}{\delta} \\ \delta^{-1} r^2 = \left(\frac{2c}{2c+1}\right)^2 \cdot (1+r^2) \end{cases} \tag{6.17}$$

from which we derive that

$$c = \frac{1}{2(\delta - 1)}; \tag{6.18}$$

$$r^2 = \frac{1}{\delta - 1}. \tag{6.19}$$

Now the empirical distribution of $\hat{\theta} = \sqrt{n}\hat{\beta}$ converges to

$$\mathcal{N}\left(0, \frac{n}{p} \cdot r^2\right) = \mathcal{N}\left(0, \frac{n}{p} \cdot \frac{1}{\delta - 1}\right) = \mathcal{N}\left(0, \frac{\delta}{\delta - 1}\right). \tag{6.20}$$

### 6.2.2 Normal Equation

In the remainder of the chapter, we provide the derivation of Theorem 36. Technical justifications of some steps are swept under the rug; we mainly focus on how to "see" the formula in Theorem 36 from the analysis.

The first step is to write the normal equation, that is, the gradient of the objective function equals zero at $\hat{\beta}$. Letting

$$\psi(x) := \rho'(x), \quad x \in \mathbb{R}, \tag{6.21}$$

we have

$$\sum A_i \psi(\epsilon_i - A_i^\top \hat{\beta}) = 0. \tag{6.22}$$

Also define the residues

$$R_i = \epsilon_i - A_i^\top \hat{\beta}. \tag{6.23}$$

### 6.2.3 Leaving out an Observation

Let $\hat{\beta}_{(i)}$ be the leave one out estimator where the $i$-th sample is not used in solving the regression. It satisfies the following normal equation:

$$\sum_{j \neq i} A_j \psi(\epsilon_j - A_j^\top \hat{\beta}_{(i)}) = 0. \tag{6.24}$$

Also define, for $1 \leq j \leq n$,

$$r_{j,(i)} := \epsilon_j - A_j^\top \hat{\beta}_{(i)}. \tag{6.25}$$

Remark that $r_{i,(i)}$ may be understood as an estimate of the prediction error, which is useful in cross-validation.

Taking the difference of (6.22) and (6.24), and Taylor expanding $\psi(\cdot)$, we obtain

$$A_i \psi(\epsilon_i - A_i^\top \hat{\beta}) + \sum_{j \neq i} \psi'(r_{j,(i)}) A_j A_j^\top (\hat{\beta}_{(i)} - \hat{\beta}) \approx 0. \tag{6.26}$$

Defining the matrix

$$S_i := \sum_{j \neq i} \psi'(r_{j,(i)}) A_j A_j^\top, \tag{6.27}$$

we can rewrite the above as

$$\hat{\beta} - \hat{\beta}_{(i)} \approx S_i^{-1} A_i \psi(\epsilon_i - A_i^\top \hat{\beta}). \tag{6.28}$$

Then

$$\begin{aligned} R_i - r_{i,(i)} &= -A_i^\top(\hat{\beta} - \hat{\beta}_{(i)}) && (6.29)\\ &\approx -A_i^\top S_i^{-1} A_i \psi(\epsilon_i - A_i^\top \hat{\beta}) && (6.30)\\ &\approx -\operatorname{tr}(S_i^{-1})\psi(R_i) && (6.31) \end{aligned}$$

with high probability. Here, Step (6.31) follows since $S_i$ is independent of $A_i$, implying $\mathbb{E}[A_i^\top S_i^{-1} A_i | S_i] = \operatorname{tr}(S_i^{-1})$. Concentration of measure implies that $A_i^\top S_i^{-1} A_i$ is close to its mean with high probability (see [EKBB$^+$13] for details), hence (6.31). Also, by symmetry, $\operatorname{tr}(S_i^{-1})$, $i = 1, \ldots, n$ should have approximately the same value. Therefore we have

$$c \approx \operatorname{tr}(S_i^{-1}) \qquad (6.32)$$

for some $c > 0$. Thus we have the following (approximate) equation linking the residue and the prediction risk:

$$R_i - r_{i,(i)} \approx -c\psi(R_i). \qquad (6.33)$$

### 6.2.4 Leaving out a Predictor

Let us also consider leaving out a predictor. By symmetry of the distribution of $A_i$, it is without loss of generality that the $p$-th column of $\mathbf{A}$ is what is left out.

For convenience, let us use the following notations:

$$A_i = \begin{bmatrix} V_i \\ A_i(p) \end{bmatrix}, \quad V_i \in \mathbb{R}^{p-1}; \tag{6.34}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_{\setminus p} \\ \hat{\beta}_p \end{bmatrix}, \quad \hat{\beta}_{\setminus p} \in \mathbb{R}^{p-1}. \tag{6.35}$$

Moreover, let

$$\hat{\gamma} := \operatorname{argmin}_{\gamma \in \mathbb{R}^{p-1}} \sum_i \rho(\epsilon_i - V_i^\top \gamma) \tag{6.36}$$

be the optimal regression vector by leaving out the $p$-th feature. Writing out the normal equation, we have

$$\sum_i V_i \psi(\epsilon_i - V_i^\top \hat{\gamma}) = 0. \tag{6.37}$$

Let us define the leave-out residue

$$r_{i,[p]} := \epsilon_i - V_i^\top \gamma. \tag{6.38}$$

Taking the difference of (6.37) and (6.22), we obtain (by looking at the first $(p-1)$ coordinates and the last coordinate respectively),

$$\sum_i V_i[\psi(R_i) - \psi(r_{i,[p]})] = 0_{p-1}; \tag{6.39}$$

$$\sum_i A_i(p)\psi(R_i) = 0. \tag{6.40}$$

Note that by definition, $R_i - r_{i,[p]} = V_i^\top(\hat{\gamma} - \hat{\beta}_{\setminus p}) - A_i(p)\hat{\beta}_p$. Taylor expanding $\psi(R_i)$ around $\psi(r_{i,[p]})$ in the above equations, we have

$$\left[\sum_i \psi'(r_{i,[p]}) V_i V_i^\top\right](\hat{\gamma} - \hat{\beta}_{\setminus p}) - \hat{\beta}_p \sum_i \psi'(r_{i,[p]}) V_i A_i(p) \approx 0_{p-1}; \tag{6.41}$$

$$\sum_i A_i(p)\left[\psi(r_{i,[p]}) + \psi'(r_{i,[p]})(V_i^\top(\hat{\gamma} - \hat{\beta}_{\setminus p}) - A_i(p)\hat{\beta}_p)\right] \approx 0 \tag{6.42}$$

where here and below, ≈ means the remainder term is higher-order smallness. Rearranging (6.42) yields

$$\hat{\beta}_p \approx \frac{\sum_i A_i(p)[\psi(r_{i,[p]}) + \psi'(r_{i,[p]})V^\top(\hat{\gamma} - \hat{\beta}_{\setminus p})]}{\sum_i A_i^2(p)\psi'(r_{i,[p]})}. \tag{6.43}$$

Moreover, defining

$$G_p := \sum_i \psi'(r_{i,[p]}) V_i V_i^\top, \tag{6.44}$$

$$u_p := \sum_i \psi'(r_{i,[p]}) V_i A_i(p), \tag{6.45}$$

we can rewrite (6.41) as

$$\hat{\gamma} - \hat{\beta}_{\setminus p} \approx \hat{\beta}_p G_p^{-1} u_p. \tag{6.46}$$

Using (6.43) and (6.46) to cancel $\hat{\gamma} - \hat{\beta}_{\backslash p}$, we obtain

$$\hat{\beta}_p \approx \frac{\sum_i A_i(p)\psi(r_{i,[p]})}{\sum_i A_i^2(p)\psi'(r_{i,[p]}) - u_p^\top G_p^{-1} u_p}. \tag{6.47}$$

As a side remark, $\hat{\beta}_p \approx \frac{\sum_i A_i(p)\psi(r_{i,[p]})}{\sum_i A_i^2(p)\psi'(r_{i,[p]})}$ is what we would obtain in the classical theory where $p/n \to 0$.

### 6.2.5 Simplifying $u_p^\top G_p^{-1} u_p$

Let us introduce the notation of the diagonal matrix

$$D := \operatorname{diag}([\psi'(r_{i,[p]})]_{i=1}^n), \tag{6.48}$$

and let

$$A(p) := \begin{bmatrix} A_1(p) \\ \vdots \\ A_n(p) \end{bmatrix}, \tag{6.49}$$

$$V := \begin{bmatrix} V_1^\top \\ \vdots \\ V_n^\top \end{bmatrix}. \tag{6.50}$$

With these notations,

$$\begin{aligned} u_p^\top G_p^{-1} u_p &= [A(p)^\top DV][V^\top DV]^{-1}[V^\top DA(p)] && (6.51) \\ &\approx \operatorname{tr}(DV[V^\top DV]^{-1}V^\top D) && (6.52) \\ &= \sum_i D_{ii} P_{ii} && (6.53) \end{aligned}$$

where the last step again used measure concentration and the fact that $A(p)$ is standard Gaussian independent of $D, V$. Moreover, we defined $P_{ii}$'s as the diagonal values of the projection matrix

$$P := D^{1/2}V[V^\top DV]^{-1}V^\top D^{1/2}. \tag{6.54}$$

As a result, we obtain the following approximation of (6.47):

$$\hat{\beta}_p \approx \frac{\sum_i A_i(p)\psi(r_{i,[p]})}{\sum_i A_i^2(p)D_{ii}(1-P_{ii})}. \tag{6.55}$$

### 6.2.6 Work on $\sum_i A_i^2(p)D_{ii}(1-P_{ii})$

Since $A_i(p)$ is independent of $r_{i,[p]}$ and $P_{ii}$, we can expect, from the law of large numbers,

$$\sum_i A_i^2(p)D_{ii}(1-P_{ii}) \approx \sum_i D_{ii}(1-P_{ii}) \tag{6.56}$$

where we replaced $A_i^2(p)$ by its expected value. To compute $P_{ii}$, the first try might be $P_{ii} = D_{ii}V_i^\top G_p^{-1}V_i \approx D_{ii}\,\mathrm{tr}(G_p^{-1})$, which is unfortunately incorrect since $V_i$ is not independent of $G_p$. Instead, we shall decompose $G_p$ into the sum of an independent component and a "coherent" rank-1 component. Define

$$G_p(i) := \sum_{j\neq i} D_{jj}V_jV_j^\top. \tag{6.57}$$

Using the Sherman-Morrison formula (Exercise 25) for the rank-1 update for matrix inversion, we have

$$\begin{aligned}[V^\top DV]^{-1} &= [G_p(i) + D_{ii}V_iV_i^\top]^{-1} && (6.58)\\ &= G_p(i)^{-1} - \frac{D_{ii}G_p(i)^{-1}V_iV_i^\top G_p(i)^{-1}}{1 + D_{ii}V_i^\top G_p(i)^{-1}V_i}. && (6.59)\end{aligned}$$

Note that the terms in (6.59) are matrices, therefore although the second term is higher oder of smallness in terms of the trace, it cannot be neglected at this point, as the two terms will be comparable after left and right multiplying $V_i^\top$ and $V_i$. Then using (6.59) we find

$$\begin{aligned}1 - P_{ii} &= 1 - D_{ii}V_i^\top[V^\top DV]^{-1}V_i && (6.60)\\ &= \frac{1}{1 + D_{ii}V_i^\top G_p(i)^{-1}V_i}. && (6.61)\end{aligned}$$

By the independence of $V_i$ and $G_p(i)$ and concentration of measure again, we have

$$\begin{aligned}V_i^\top G_p(i)^{-1}V_i &\approx \operatorname{tr}(G_p(i)^{-1}) && (6.62)\\ &\approx \operatorname{tr}(G_p^{-1}) && (6.63)\\ &\approx \operatorname{tr}(S_i^{-1}) && (6.64)\\ &\approx c && (6.65)\end{aligned}$$

where the approximation errors are up to a multiplicative factor of $1 + o(1)$. Thus, summing (6.61) over $i$,

$$\sum_i \frac{1}{1 + D_{ii}c} \approx \sum_i [1 - P_{ii}] \qquad (6.66)$$

$$= n - \mathrm{tr}(P) \tag{6.67}$$

$$= n - (p-1) \tag{6.68}$$

where the last step use the fact that $P$ is a projection matrix of rank $p-1$ (almost surely). Now collecting all the above in this section, we can simply the denominator in (6.55) as follows:

$$\begin{aligned}
\sum_i A_i^2(p) D_{ii}(1-P_{ii}) &\approx \sum_i D_{ii}(1-P_{ii}) && (6.69)\\
&\approx \sum_i \frac{D_{ii}}{1+cD_{ii}} && (6.70)\\
&= \frac{1}{c}\sum_i \left[1 - \frac{1}{1+cD_{ii}}\right] && (6.71)\\
&\approx \frac{1}{c}\sum_i P_{ii} && (6.72)\\
&\approx \frac{p}{c}. && (6.73)
\end{aligned}$$

Thus by (6.55),

$$\hat{\beta}_p \approx \frac{c}{p}\sum_i A_i(p)\psi(r_{i,[p]}). \tag{6.74}$$

Since $r_{i,[p]}$ is defined by $(V, \epsilon)$ which is independent of $\{A_i(p)\}_{i=1}^n$,

$$\mathbb{E}[\hat{\beta}_p^2 | V, \epsilon] = \frac{c^2}{p^2}\sum_i \psi^2(r_{r_{i,[p]}}) \tag{6.75}$$

$$\approx \frac{c^2}{p^2} \sum_i \psi^2(R_i) \tag{6.76}$$

where the last step used the fact that $r_{r_{i,[p]}} \approx R_i$ and the continuity of $\psi$. While our analysis focused on the $p$-th feature, the result extends to the others by symmetry. Thus

$$\mathbb{E}[\|\hat{\beta}\|^2] \approx \frac{c^2}{p} \sum_i \psi^2(R_i). \tag{6.77}$$

### 6.2.7 Collecting Terms

So far, we have derived three independent approximate equations:

$$\begin{cases} \hfill r_{i,(i)} & = R_i + c\psi(R_i), \\ \frac{1}{n}\sum_i \frac{1}{1+\psi'(R_i)} & \approx 1 - \frac{p}{n}, \\ \hfill \mathbb{E}[\|\hat{\beta}\|^2] & \approx \frac{n}{p}\left[\frac{1}{n}\sum_i c^2\psi^2(R_i)\right]. \end{cases} \tag{6.78}$$

Note that by symmetry, $(r_{i,(i)}, R_i)$ have the same distribution for different $i$. Moreover, for large $p$ we can assume that $\|\hat{\beta}_{(i)}\|$ is nearly deterministic by measure concentration. Since $r_{i,(i)} = \epsilon_i - A_i^\top \hat{\beta}_{(i)}$ and $\epsilon_i$ and $A_i$ are independent, we see that $r_{i,(i)}$ is the independent sum of the Gaussian $\epsilon_i \sim \mathcal{N}(0,1)$ and the Gaussian $\mathcal{N}(0, \|\hat{\beta}_{(i)}\|)$; in particular, its distribution is determined by $\|\hat{\beta}_{(i)}\| \approx \|\hat{\beta}\|$. Therefore there are three independent unknowns in (6.78): $\|\hat{\beta}\|$, The distribution of $R_i$, and $c$.

### 6.2.8 Rewriting the Fixed Point Equations

The fixed point equations (6.78) can be rewritten as two independent equations while suppressing the notation $R_i$ in the meantime. Define a function $g_c$ by

$$g_c(x) := x + c\psi(x). \tag{6.79}$$

Differentiating the objective in (6.14) and setting it zero, we easily see the functional identity:

$$g_c^{-1} = \mathrm{prox}_c(\rho). \tag{6.80}$$

Moreover, the first equation in (6.78) can be rewritten as

$$R_i = g_c^{-1}(r_{i,(i)}), \tag{6.81}$$

therefore,

$$\begin{aligned} \frac{1}{1+\psi'(R_i)c} &= \frac{1}{g_c'(R_i)} && (6.82) \\ &= \frac{1}{g_c'(g_c^{-1}(r_{i,(i)}))} && (6.83) \\ &= (g_c^{-1})'(r_{i,(i)}). && (6.84) \end{aligned}$$

Using (6.81) and (6.84) we can rewrite the last two equations in (6.78) as the two fixed point equations in Theorem 36.

# Chapter 7

# Replica Method

In this chapter we introduce the replica method, another useful technique for analyzing high dimensional optimization problems with random instances. The replica method often allows us to obtain solutions (relatively quickly); however, the method is non-rigorous. It boils down to the simple fact that

$$\lim_{k \downarrow 0} \frac{1}{k}(z^k - 1) = \ln z, \quad \forall z > 0. \tag{7.1}$$

Now if $Z > 0$ is a random variable, and its $k$-th moment $\mathbb{E}[Z^k]$ can be computed and equals $f(k)$ where $f$ is some analytic function, $k = 1, 2, 3 \ldots$, then we may expect to use (7.1) to compute $\mathbb{E}[\ln Z]$ as $\lim_{t \downarrow 0} \frac{1}{t}[f(t) - 1]$. This argument is only heuristic since the $k$-th moment is only computed for integer $k$. Nevertheless, this argument often gives the right answer (as can be verified later by other techniques). Once we understand the asymptotic behavior of

$\mathbb{E}[\ln Z]$ (called the free-energy, which is a function of the temperature), many useful information about the high dimensional random object can be deduced.

The replica method was originated from statistical physics since 1970's; later it was applied to linear inverse problems (see e.g. [Tan02] for multiuser systems, [GV05] for the minimum mean square estimator; [RFG12] for the maximum a posteriori probability estimator; [BM11a, JM14] for the Lasso). The calculations involved in those papers are often rather complicated. In this chapter, we show the basic idea of the replica method through the simple least squares problem in Section 6.1, including the intuitions for the Gaussian limit and the mysterious $\frac{\delta}{\delta-1}$ asymptotic empirical variance.

## 7.1 Proof of the Pedagogical Example

Recall the problem of finding the asymptotic limit of the empirical distribution of $(\hat{\theta}_i)_{i=1}^p$ in the least squares regression. In the previous chapter, we showed how this problem can be solved using two methods, either utilizing the Marchenko-Pastur law of random matrices as a blackbox result, or deriving fixed point equations by the leave-one-out technique (which actually characterizes the limiting distribution in more general convex optimization problems).

Here, we will use the replica method to compute the variance $\frac{\delta}{\delta-1}$, and we will also mention how asymptotic Gaussianity must hold in view of the *maximal entropy property.*

The analysis here is slightly different from [Tan02] and [GV05]; we shall touch on this later after the proof. We will also remark how the proof will change when one wants to extend this simple least square example to the $\ell_1$-regularized (Lasso) case, as was done in [BM11a, JM14].

### 7.1.1 The Free Energy

Given $\delta$, we can view the least squares

$$\hat{\theta} := \text{argmin}_{\theta \in \mathbb{R}^p} \|\mathbf{w} - \mathbf{A}\theta\|_2 \tag{7.2}$$

as an optimization problem indexed by $n$ and with random instances $\mathbf{w}$ and $\mathbf{A}$. Let us also introduce a parameter $\beta > 0$, called the *inverse temperature*, and define

$$Z(\beta, n) := \int e^{-\frac{\beta}{2n}\|\mathbf{w}-\mathbf{A}\theta\|^2} \mathrm{d}\theta, \tag{7.3}$$

called the *partition function.* The *energy* of the state $\theta$ is

$$E(\theta) = \frac{1}{2n}\|\mathbf{w} - \mathbf{A}\theta\|^2 \tag{7.4}$$

The partition function encodes much useful information about the system. For example, it is easy to see that

$$-\frac{\partial}{\partial\beta} \ln Z(\beta, n) = \frac{1}{Z(\beta, n)} \int E(\theta) e^{-\beta E(\theta)} \mathrm{d}\theta \tag{7.5}$$

is the expectation of the energy with respect to the Gibbs measure $\frac{1}{Z(\beta,n)}e^{-\beta E(\theta)}\mathrm{d}\theta$. As $\beta \to \infty$, the Gibbs measure will become concentrated around (7.2), the solution to the optimization problem.

Why not define the partition function as $Z(\beta, n) := \int e^{-\frac{\beta}{2}\|\mathbf{w}-\mathbf{A}\theta\|^2}\mathrm{d}\theta$, but with $\frac{1}{n}$ in the exponent in (9.13)? Note that we generally assume that $\beta > 0$ does not scale with $n$. Experience with large deviation theory tells us that we should normalized the exponent so that with high probability, the exponent $-\frac{\beta}{2n}\|\mathbf{w} - \mathbf{A}\theta\|^2$ is order $\Theta(n)$.

The free *energy* is defined as

$$F(\beta) := -\frac{1}{\beta}\log Z(\beta, n), \tag{7.6}$$

and as with the partition function, it also encodes useful information about the system. The key idea of the proof is thus to compute the expectation of the free enegry. While it is relatively easy to see that $F(\beta)$ is order $\Theta(n)$, the nontrivial part is to use the replica method to determine the prefactor. This is equivalent to finding the speed of exponential decay of $Z(\beta, n)$ for *typical* $(\mathbf{w}, \mathbf{A})$, which, by the large deviation theory, is strictly faster than the speed of exponential decay of $\mathbb{E}[Z(\beta, n)]$.

### 7.1.2 $\mathbb{E}_{\mathbf{w},\mathbf{A}}$

As mentioned around (7.1), our strategy is to compute the moment $\mathbb{E}[Z^k]$ and then pass it to the limit $k \downarrow 0$ to obtain $\mathbb{E}[\ln Z]$.

$$\begin{aligned}
\mathbb{E}[Z^k] &= \mathbb{E}_{\mathbf{w},\mathbf{A}} \int e^{-\frac{\beta}{2n}\sum_{a=1}^k \|\mathbf{w}-\mathbf{A}\theta^a\|^2} \mathrm{d}\theta^1 \ldots \mathrm{d}\theta^k && (7.7)\\
&= \int \left(\mathbb{E} e^{-\frac{\beta}{2n}\sum_{a=1}^k (w_1 - A_1^\top \theta^a)^2}\right)^n \mathrm{d}\theta^1 \ldots \mathrm{d}\theta^k && (7.8)\\
&= \int \det{}^{-n/2}(\mathbf{I} + \beta\mathbf{E} + \frac{\beta}{n}\sum_{j=1}^p \theta_j^{[k]}\theta_j^{[k]\top}) \mathrm{d}\theta^1 \ldots \mathrm{d}\theta^k && (7.9)
\end{aligned}$$

where we used the property of Gaussian distribution in (7.31), and for each $j = 1, \ldots, p$

$$\theta_j^{[k]} := \begin{bmatrix} \theta_j^1 \\ \theta_j^2 \\ \vdots \\ \theta_j^k \end{bmatrix} \in \mathbb{R}^k. \tag{7.10}$$

Note that $Z^k$ can be understood as the partition function of $k$-replicated systems, with energy $\frac{1}{2n}\sum_{a=1}^k \|\mathbf{w} - \mathbf{A}\theta^a\|^2$, hence the name replica method.

### 7.1.3 Substitute $\frac{1}{p}\sum_{j=1}^p \theta_j^{[k]}\theta_j^{[k]\top} \longrightarrow \mathbf{Q}$

Note that the integrand in (7.9) only depends on the empirical second moment $\mathbf{Q} := \frac{1}{p}\sum_{j=1}^p \theta_j^{[k]}\theta_j^{[k]\top}$, hence we can understand the asymptotic behavior of the integral using large deviations theory. If you are not familiar with large deviations, you may first take a look at the stand-alone Section 7.3 ahead.

Let us make the ansatz that we can compute the exponent of (7.9) by supremizing over the empirical distribution $\mu$ of $\theta_j^{[k]}$. This corresponds to the method-of-types in information theory (even though the domain is $\mathbb{R}$ instead of finite sets). Suppose the empirical distribution of $(\theta_j^1, \dots, \theta_j^k)_{j=1}^p$ is $\mu$, and let $\Theta \sim \mu$ be a random variable on $\mathbb{R}^k$. Note that the second moment $\mathbf{Q}$ imposes a constraint on $\mu$. From the large deviation theory (method of types) we have

$$\begin{aligned} \mathrm{d}\theta^1 \dots \mathrm{d}\theta^k &\doteq \exp\left\{ p \sup_{\mu:\, \mathbb{E}_\mu[\Theta\Theta^\top]=\mathbf{Q}} h(\mu) \right\} \mathrm{d}\mathbf{Q} && (7.11) \\ &= (2\pi e)^{pk/2} \exp\left\{ \frac{p}{2} \log\det(\mathbf{Q}) \right\} \mathrm{d}\mathbf{Q}, && (7.12) \end{aligned}$$

where $\doteq$ denotes equivalence up to a factor of $\exp(o(p))$, and $\mathrm{d}\mathbf{Q}$ denotes the Lebesgue measure on $\mathbb{R}^{k(k+1)/2}$. Moreover,

$$h(\mu) := \mathbb{E}\left[\log \frac{1}{p(X)}\right]. \tag{7.13}$$

denotes the differential entropy, where $p$ is the density of $\mu$ and $X \sim \mu$.

**Here we can see that the supremizing $\mu$ yields a Gaussian measure, since it maximizes the differential entropy under the second moment constraint.**

### 7.1.4 $p \to \infty$

Recall that we have shown that

$$\mathbb{E}[Z^k] = \int \det{}^{-n/2}(\mathbf{I} + \beta\mathbf{E} + \frac{\beta}{n}\sum_{j=1}^{p} \theta_j^{[k]}\theta_j^{[k]\top})\mathrm{d}\theta^1 \dots \theta^k \tag{7.14}$$

$$\doteq (2\pi e)^{pk/2} \int \det{}^{-n/2}(\mathbf{I} + \beta\mathbf{E} + \frac{\beta}{\delta}\mathbf{Q}) \cdot \det{}^{p/2}(\mathbf{Q})\mathrm{d}\mathbf{Q}, \tag{7.15}$$

therefore,

$$\lim_{p\to\infty} \frac{1}{p} \log \mathbb{E}[Z^k]$$
$$= \frac{k}{2}\log(2\pi e) + \sup_{\mathbf{Q} \text{ is psd}} \left\{ -\frac{\delta}{2}\log\det(\mathbf{I} + \beta\mathbf{E} + \frac{\beta}{\delta}\mathbf{Q}) + \frac{1}{2}\log\det(\mathbf{Q}) \right\}. \tag{7.16}$$

### 7.1.5 Replica Symmetry Ansatz

Optimizing $\mathbf{Q}$ in (7.16) amounts to optimizing $k(k+1)/2$ variables, which is not easily tractable analytically. The Replica Symmetry Ansatz referring to the a common assumption that the optimizer is symmetric, which, in this example, means that we assume that the optimizer is of the form

$$\mathbf{Q} = r\mathbf{I} + q\mathbf{E}. \tag{7.17}$$

In general, replica symmetry ansatz is correct in many basic examples, but can also fail in many other examples (which is called *replica symmetry breaking*).

Using the formula (7.30) for the determinant, we compute (7.16) as

$$\begin{aligned}
&\lim_{p\to\infty}\frac{1}{p}\ln\mathbb{E}[Z^k] \\
&=\sup_{r,q}\left\{-\frac{(k-1)\delta}{2}\ln(1+\beta r/\delta)-\frac{\delta}{2}\ln(1+\beta r/\delta+\beta qk/\delta+\beta k)\right. \\
&\quad\left.+\frac{k-1}{2}\ln\frac{r}{\delta}+\frac{1}{2}\ln\frac{r+qk}{\delta}\right\}+\frac{k}{2}\ln(2\pi e). \qquad (7.18)
\end{aligned}$$

### 7.1.6 Take $k \to 0$

$$\begin{aligned}
&\lim_{p\to\infty}\frac{1}{p}\mathbb{E}[\ln Z] \\
&=\lim_{p\to\infty}\lim_{k\downarrow 0}\frac{1}{p}\frac{1}{k}\ln\mathbb{E}[Z^k] \qquad (7.19)\\
&=\lim_{k\downarrow 0}\frac{1}{k}\lim_{p\to\infty}\frac{1}{p}\ln\mathbb{E}[Z^k] \qquad (7.20)\\
&=\sup_{q,r}\left\{-\frac{\delta}{2}\ln(1+\beta r/\delta)-\frac{\beta q+\beta\delta}{2(1+\beta r/\delta)}+\frac{1}{2}\ln\frac{r}{\delta}+\frac{q}{2r}\right\} \\
&\quad+\frac{1}{2}\ln(2\pi e). \qquad (7.21)
\end{aligned}$$

In (7.20), we assumed that the order of limits can be switched, which is not rigorous.

### 7.1.7 $\sup_{r,q}$

To find the supremizers in (7.21), we take the partial derivatives in $r$ and $q$ respectively, set them to equal 0, and get

$$r = \frac{\delta}{\beta(\delta-1)}; \tag{7.22}$$

$$q = \frac{\delta}{\delta-1}. \tag{7.23}$$

Substituting these, compute the supremum value, and take $\lim_{\beta\to\infty}\frac{1}{\beta}$, we get

$$\lim_{\beta\to\infty}\frac{1}{\beta}\lim_{p\to\infty}\frac{1}{p}\mathbb{E}[\ln Z] = -\frac{1}{2}(\delta-1). \tag{7.24}$$

Note, however, that (Exercise 28)

$$\lim_{\beta\to\infty}\frac{1}{\beta}\mathbb{E}[\ln Z] = -\frac{1}{2n}\min_{\theta}\|\mathbf{w}-\mathbf{A}\theta\|^2, \tag{7.25}$$

Therefore the result of (7.24) agrees with what we can compute directly:

$$\frac{1}{n^2}\mathbb{E}[\|\mathbf{w}-\mathbf{A}\theta\|^2] = n^{-1}\mathrm{Tr}[\mathbf{I}-\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top] \tag{7.26}$$

$$= \frac{\delta-1}{\delta}. \tag{7.27}$$

Moreover, the result of (7.23) suggests that the empirical variance of $\theta$ equals $\frac{\delta}{\delta-1}$. Which agrees with what we can compute from the M-P law from random matrix theory or the leave-one-out analysis in the previous Chapter.

### 7.1.8 Remarks

If the least squares problem is replaced by LASSO, then a debiased version of $\theta$ also tends to a Gaussian channel. The replica proof of this can be extracted from [Tan02][GV05] in the multiuser literature or [JM14] directly stated for LASSO. In that case, we need to replace the partition function by

$$Z(\beta, n) := \int e^{-\frac{\beta}{2n}\|\mathbf{w}-\mathbf{A}\theta\|^2-\beta\lambda\|\theta\|_1}\mathrm{d}\theta.$$

Subsequently, (7.12) will be replaced by

$$\mathrm{d}\theta^{[k]} \doteq \exp\left\{p\left(\sup_{\mu\colon\,\mathbb{E}_\mu[\Theta\Theta^\top]=\mathbf{Q}} h(\mu) - \beta\lambda\mathbb{E}[|\Theta|]\right)\right\}\mathrm{d}\mathbf{Q}. \tag{7.28}$$

From here, [Tan02] proceeds by expressing the exponent using the moment generating function of the prior distribution (in this case the Laplace distribution):

$$\tilde{Q} \mapsto \mathbb{E}[\exp(\mathrm{Tr}[\tilde{Q}\Theta\Theta^\top])]. \tag{7.29}$$

More precisely, by the Chernoff bound, the exponent equals the Legendre transform of the log moment generating function. Then an important insight: **assume that the max over $\tilde{Q}$ in the computation of the Legendre transform is achieved by a replica-symmetric $\tilde{Q}$.** Thus, the optimization previously in Section 7.1.7 is now replaced by optimization over the symmetric parameters in both $Q$ and $\tilde{Q}$. [JM14] used the Fourier transform instead of Legendre transform, and a similar replica symmetric assumption of the dual matrix $\tilde{Q}$ was made.

## 7.2 Appendix: Useful Facts

- Let $\mathbf{E}$ be the matrix whose entries are all 1;

  $$\det(\lambda\mathbf{I} + r\mathbf{E}) = \lambda^{k-1}(\lambda + rk), \tag{7.30}$$

  where $k$ is the dimension.

- If $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, then

  $$\mathbb{E}[e^{-\frac{1}{2}\sum Y_i^2}] = \det{}^{-1/2}(\mathbf{I} + \mathbf{\Sigma}). \tag{7.31}$$

- Inverse formula

  $$(\mathbf{I} + a\mathbf{E})^{-1} = \mathbf{I} - \frac{a}{1 + a\dim}\mathbf{E}. \tag{7.32}$$

## 7.3 Appendix: Large Deviations

In probability theory, the theory of large deviations concerns the asymptotic behavior of remote tails of sequences of probability distributions [dem10]. If you have no experience with large deviations, you may learn it through the following basic example, which has a similar flavor as the calculation in Section 7.1.3.

Let $X_1$,...,$X_n$ be a sequence of i.i.d. Bernoulli random variables with expectation 1/2. Let $\beta > 0$ be a constant independent of $n$. What is the limit

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}[e^{-\beta \sum_{i=1}^{n} X_i}]? \tag{7.33}$$

A large deviations problem like this can be solved by the *method of types*, which is a popular technique in information theory [CK11]. The *type* is synonymous with the empirical distribution of $\{X_i\}_{i=1}^{n}$. The set

$$\mathcal{P}_n := \{P \colon \exists\, x^n \in \{0,1\}^n, P = \widehat{P}_{x^n}\} \tag{7.34}$$

has cardinality $n + 1$, which, in particular, is polynomial in $n$. We have

$$\begin{aligned} \mathbb{E}[e^{-\beta \sum_{i=1}^{n} X_i}] &= \sum_{P \in \mathcal{P}_n} e^{-\beta n \mathbb{E}_P[X]} \mathbb{P}[\widehat{P}_{X^n} = P] \\ &\doteq \sum_{P \in \mathcal{P}_n} e^{-\beta n \mathbb{E}_P[X]} \exp(-nD(P\|Q)) \end{aligned} \tag{7.35}$$

where $Q := \mathrm{Ber}(1/2)$, and we recall that $\doteq$ means equivalence up to $\exp(o(n))$ factor. The KL divergence term in (7.35) is typical in large deviations theory (Exercise 29), which is similar to the differential entropy term in (7.11). Since there are only polynomially many summands in (7.35), we have

$$\lim_{n\to\infty}\frac{1}{n}\ln\mathbb{E}[e^{-\beta\sum_{i=1}^n X_i}] = \max_P\{-\beta\mathbb{E}_P[X] - D(P\|Q)\} \tag{7.36}$$

where the max is over distribution $P$ on $\{0,1\}$, which is parameterized by one number.

# Chapter 8

# Iterative Algorithms

The Lasso problem is a convex optimization, and can be solved using the glmnet package quickly when the dimension of the unknown signal is on the order of a few thousands. However, there are applications in which the dimension is on the order of millions, which motivates the search for faster algorithms [IR08, TG07, HGT06]. Iterative hard and soft thresholding algorithms are often good options in this situation: they are faster than solving $\ell_0$ and $\ell_1$ regularized minimizations, yet the accuracies are generally better than greedy algorithms such as orthogonal matching pursuits.

Remarkably, with an additional (and apparently mysterious) decorrelation term for the iterative thresholding algorithm, one obtains the approximate message passing algorithm (AMP). Asymptotically, AMP achieves identical accuracy as Lasso for select (e.g. i.i.d. Gaussian) matrix ensembles. Moreover, the asymptotic behaviors of AMP can be analyzed precisely. Thus this sometimes provides rigorous proofs to claims about large systems suggested by replica calculations.

Unless otherwise noted, in this chapter we consider the setting where we seek to reconstruct an unknown vector $x^\star \in \mathbb{R}^p$ from linear observations

$$y = Ax^\star + w \in \mathbb{R}^n. \tag{8.1}$$

# 8.1 Iterative thresholding

## 8.1.1 Derivation From Regularized Least Squares

Let $\rho\colon \mathbb{R} \to \mathbb{R}$ be a "nice" function, and consider the following general optimization problem

$$\hat{x} := \mathrm{argmin}_{x\in\mathbb{R}^p} \frac{1}{2}\|y - Ax\|_2^2 + \rho(x) \tag{8.2}$$

where, by an abuse of notation, $\rho(x) := \sum_{j=1}^p \rho(x_j)$. The normal equation reads as

$$-(y - Ax)^\top A + \rho'(x) = 0. \tag{8.3}$$

In the case of Lasso, we should choose $\rho(x) = |x|$ and hence $\rho'(x) = \mathrm{sign}(x)$. Unfortunately, (8.3) does not have an analytic solution in that case. Let us rewrite (8.3) using the prox operator notation in (6.80). That is, define functions $g, \eta$ by

$$g(x) := x + \rho'(x); \tag{8.4}$$

$$\eta(x) := g^{-1}(x). \tag{8.5}$$

(Here, we used $\eta$ to denote the $\text{prox}_1(\rho)$ in (6.80).) Now (8.3) is equivalent to

$$-x - (y - Ax)^\top A + g(x) = 0, \tag{8.6}$$

and, in turn,

$$x = \eta(x + (y - Ax)^\top A). \tag{8.7}$$

This leads to the following iterative algorithm: start with $x^0 = 0$ and proceed by

$$x^{t+1} = \eta(A^\top z^t + x^t), \tag{8.8}$$

$$z^t = y - Ax^t. \tag{8.9}$$

## Iterative Soft Thresholding

Taking $\rho(t) := \lambda|t|$, we obtain $\eta(t) = \text{sign}(t)\max\{|t| - \lambda, 0\}$. In this case, the iterations (8.8)-(8.9) are called Iterative Soft Thresholding.

**Iterative Hard Thresholding**

From (8.5) we see that $\eta(y) = \text{argmin}_t\{\frac{1}{2}|y-t|^2+\rho(t)\}$. In particular, taking $\rho(t) := \lambda 1\{t \neq 0\}$ we get $\eta(y) = y1_{|y|>\sqrt{2\lambda}}$. In this case, the iterations (8.8)-(8.9) are called Iterative Hard Thresholding.

### 8.1.2 The Issue of Convergence

In general, the iterative algorithm (8.8)-(8.9) is not guaranteed to convergence. To see this, consider the simplest example where $\rho(\cdot)$ is constant, in which case $\eta$ is the identity function. Then the iterative algorithm reads as

$$x^{t+1} = A^\top y + (I_p - A^\top A)x^t. \tag{8.10}$$

We see that $x^t$ converges if all eigenvalues of $I_p - A^\top A$ are in $(-1, 1)$ (equivalently, all singular values of $A$ are in $(0, \sqrt{2})$). But otherwise, $x^t$ may not converge for some initialization and $y$.

On the other hand, if $A$ is an orthogonal matrix and $\rho$ is a general convex function, we see that

$$\begin{aligned} x^{t+1} &= \eta(A^\top y + (I_p - A^\top A)x^t) && (8.11)\\ &= \eta(A^\top y) && (8.12) \end{aligned}$$

converges in just one iteration!

With these examples of extreme cases in mind, it may not come as a surprise that

**Theorem 37.** *Assume that $\rho(t) = |t|^q$, $q \geq 1$, and $A$ has operator norm strictly smaller than* 1 *and a trivial null space. Then $x^t$ converges to the solution of* (8.2).

Indeed, under the conditions of Theorem 37, the mapping $x \mapsto \eta(A^\top y + (I_p - A^\top A)x)$ is a (strict) contraction under the $\ell_2$ norm, therefore the convergence to the unique fixed point follows immediately from the Banach fixed point theorem. A reference for the proof of Theorem 37 (along with some generalizations to the Banach space settings) can be found in [DDDM04].

In practice, Theorem 37 suggests that one should perform a certain "batch normalization" for the the sample matrix $A$, before running the iterative thresholding algorithm.

While the cost function in Theorem 37 is general enough to cover the $\ell_1$ norm (take $q = 1$), unfortunately, this result is not applicable to the case of $p > n$ where $A$ has nontrivial null-space. Consequently, for the general sparse recovery problem with $p > n$ and $k < n$, the soft-thresholding algorithm does not produce the same solution as Lasso; indeed, this is noted in the numerical experiments [DMM09]. Nevertheless, iterative *hard* thresholding still seems to enjoy competitive statistical accuracy and computational efficiency, despite the lack of theoretical convergence guarantees.

## 8.2 Approximate Message Passing

### 8.2.1 The Algorithm

The Approximate Message Passing (AMP) algorithm starts with $x^0 = x^{-1} = 0$ ,$z^{-1} = 0$, and computes

$$z^t = y - Ax^t + \frac{1}{\delta} z^{t-1} \left\langle \eta'_{t-1}(A^\top z^{t-1} + x^{t-1}) \right\rangle , \tag{8.13}$$

$$x^{t+1} = \eta_t(A^\top z^t + x^t), \tag{8.14}$$

for $t = 0, 1, 2, \ldots$ Here, $\langle x \rangle := \frac{1}{p} \sum_{j=1}^{p} x_p$ denotes the average value of the coordinates of a vector. $\eta_t$ is a (possibly nonlinear) Lipschitz function depending on the iteration index $t$. Thus, AMP can be interpreted as an iterative thresholding algorithm with an additional correction term $\frac{1}{\delta} z^{t-1} \left\langle \eta'_{t-1}(A^\top z^{t-1} + x^{t-1}) \right\rangle$. The motivations for including this correction term are

- Note that the term $\langle \ldots \rangle$ in (8.13) is a scalar. Therefore, by the same argument as Section 8.1.1, it is clear that when $\eta_t$ is chosen as a soft-thresholding operator, the limit point of the iterations (assuming convergence) is the solution of Lasso. The correspondence between the parameter in AMP (i.e. threshold in $\eta_t$) and in Lasso is through the *calibration map* in [BM11b]. In contrast to the iterative thresholding algorithm, the mean square deviation between the AMP solution and the Lasso solution is asymptotically zero, regardless of $\delta$ and the noise variance. This was shown empirically in [DMM09] and theoretically in [BM11b, Theorem 1.8] for Gaussian matrices and assuming the weak convergence of the empirical distributions $(\widehat{P}_{x^\star})_{p=1}^\infty$.

- Asymptotically, the correction cancels some annoying correlations in the state evolution analysis, so that the asymptotic behavior of AMP can be precisely characterized. Then by the previous point, [BM11b] also obtained a rigorous proof of the weak limit of $(\widehat{P}_{\hat{x}^{Lasso}})_{p=1}^\infty$ (cf. Section 7.1.8).

### 8.2.2 Asymptotic Results

Consider the following setting:

- $\{A(p)\}_{p\geq 1}$ is a sequence of random matrices indexed by $p$. We may drop the argument $p$ later when no confusion. $A \in \mathbb{R}^{n\times p}$, with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, 1/n)$, and assume that $n/p \to \delta \in (0, \infty)$;

- $\{x^\star(p)\}_{p\geq 1}$, and the empirical distribution of its entries converge weakly to a probability measure $P_{X^\star}$ on $\mathbb{R}$ with bounded $(2k-2)$-th moment, and $\mathbb{E}_{\widehat{P}_{x^\star(p)}}[X^{2k-2}] \to \mathbb{E}_{P_{X^\star}}[X^{2k-2}]$, $p \to \infty$, for some $k \geq 2$;

- The noise $w$ has i.i.d. entries with distribution $P_W$ which has bounded $(2k-2)$-th moment, and variance $\sigma^2$;

- $\psi \colon \mathbb{R}^2 \to \mathbb{R}$ is a *pseudo-Lipschitz function of order $k$* (see [BM11a]);
- $\{\eta_t\}_{t \geq 0}$ is a sequence of scalar functions where each $\eta_t \colon \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous (hence almost everywhere differentiable).

**Theorem 38.** *Assume the setting in the itemized above. Let $y$ be generated from the observation model* (8.1)*, and let $x^t$ be the solution produced by the AMP iterations* (8.14)-(8.13)*. Then for any $t \geq 0$ fixed (independent of $p$),*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi(x_i^{t+1}, x_i^\star) = \mathbb{E}[\psi(\eta_t(X^\star + \tau_t Z), X^\star)] \tag{8.15}$$

*where $X^\star \sim P_{X^\star}$ and $Z \sim \mathcal{N}(0,1)$ are independent, and the scalars $\tau_t$ are defined as follows: $\hat{\tau}_0^2 = \mathbb{E}[(X^\star)^2]$, and*

$$\tau_t^2 = \sigma^2 + \delta^{-1}\hat{\tau}_t^2; \tag{8.16}$$

$$\hat{\tau}_{t+1}^2 = \mathbb{E}[|\eta_t(X^\star + \tau_t Z) - X^\star|^2] \tag{8.17}$$

*for $t = 0, 1, 2\ldots$*

*Remark* 12. It is expected that the result of Theorem 38 continues to hold when the measurement matrix $A$ has i.i.d. entries not necessarily Gaussian but only satisfying milder conditions (i.e. universality) [BM11a]. However, as with many similar problems in the random matrix theory, characterizing all distributions in the universality class is an outstanding mathematical problem.

*Remark* 13. By (8.15), the asymptotic MSE $\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\|x_i^{t+1}-x^\star\|_2^2$ is finite iff $\tau_t^2$ is finite. By letting $t\to\infty$, letting $P_{X^\star}$ be a distribution with an atom at zero, and in view of the asymptotic equivalence between Lasso and AMP solutions [BM11b], Theorem 38 yields a characterization for the *noise sensitivity phase transition*, that is, the iterations (8.16) and (8.17) must converge to some finite value. Interestingly, an alternative characterization of the noise sensitivity transition was found in [DT05, Don06] via combinatorial geometry, which is equivalent but has an apparently different form [DMM09].

### 8.2.3 The Least Squares Example

Take $\eta_t$ as the identity scalar function, and it is easy to see that the fixed point of the AMP iterations is the least squares solution. In (8.15), take $X^\star$ as constant 0, and we see

$$\lim_{p\to\infty}\frac{1}{p}\|x^\infty\|_2^2=\tau_\infty^2 \tag{8.18}$$

while computing the fixed point in (8.16)-(8.17) gives $\tau_\infty^2=\frac{\sigma^2\delta}{\delta-1}$, which is the same solution we previously computed using the leave-one-out method and the replica method.

### 8.2.4 A Proof via the Path Representation

In this section, we give a proof of Theorem 38. Some computations of bounding the approximation errors will not be shown, we mostly focus on showing how the solution emerges. The proof given here is by rewriting the key quantities using paths on a bipartite graph, which appears to be more transparent than the *conditioning technique* in the original paper [BM11a].

We will see by inductions that

$$\tau_t^2 \approx \frac{1}{n}\|z^t\|_2^2, \tag{8.19}$$

$$\hat{\tau}_t^2 \approx \frac{1}{p}\|x^t\|_2^2. \tag{8.20}$$

Here and below, $\approx$ means approximation up to a term of mean and variance $o(1)$. On the other hand, the first a few iterations in (8.16)-(8.17) are

$$\tau_0^2 = \sigma^2, \tag{8.21}$$

$$\hat{\tau}_1^2 = \mathbb{E}|\eta_0(\tau_0 Z)|^2 \tag{8.22}$$

$$\tau_1^2 = \sigma^2 + \delta^{-1}\hat{\tau}_1^2 \tag{8.23}$$

$$\cdots \tag{8.24}$$

In the AMP iterations (8.13)-(8.14), we start with $z^0 = w$. Therefore for any $a \in [n]$,

$$z_a^0 = w_a \tag{8.25}$$

which verifies that $z_a^0$ is approximately $\mathcal{N}(0, \tau_0^2)$ in distribution, and (8.19) holds for $t = 0$. Next, for any $i \in [p]$,

$$x_i^1 = \eta_0(\sum_{a \in [n]} A_{ai} w_a). \tag{8.26}$$

By the central limit theorem, $\sum_{a \in [n]} A_{ai} w_a$ is close to Gaussian with variance $n \cdot \frac{1}{n} \sigma^2$. This shows that $x_i^1$ is close to $\eta_0(\tau_0 Z)$ in distribution, and hence (8.20) is verified for $t = 1$. Next, for any $b \in [n]$,

$$z_b^1 = w_b - \sum_{i \in [p]} A_{bi} x_i^1 + \delta^{-1} z_b^0 \left\langle \eta_0'(x^1) \right\rangle. \tag{8.27}$$

If $x^1$, $w$ and $A$ were independent, we would have that $w_b - \sum_{i \in [p]} A_{bi} x_i^1$ approximates $\sigma Z' + \sqrt{\frac{p}{n} \cdot \frac{1}{p} \|x^1\|_2^2} Z \approx \sigma Z' + \sqrt{\delta^{-1} \hat{\tau}_1^2} Z$ in distribution, where $Z, Z'$ are i.i.d. $\mathcal{N}(0, 1)$, which verifies (8.19) for $t = 1$ since $\tau_1^2 := \sigma^2 + \delta^{-1} \hat{\tau}_1^2$. In reality, however, independence is not true, while the last correction term in (8.27) cancels the correlation so that the above intuition works through. To see this, consider

$$z_b^1 = w_b - \sum_{i \in [p]} A_{bi} \eta_0(\sum_{a \in [n]} A_{ai} w_a) + \delta^{-1} z_b^0 \left\langle \eta_0'(x^1) \right\rangle \tag{8.28}$$

$$\approx w_b - \sum_{i \in [p]} A_{bi} \left[ \eta_0(\sum_{a \neq b} A_{ai} w_a) + \eta_0'(\sum_{a \neq b} A_{ai} w_a)) A_{bi} w_b \right] + \delta^{-1} z_b^0 \left\langle \eta_0'(x^1) \right\rangle \tag{8.29}$$

$$
\begin{aligned}
&= w_b - \sum_{i\in[p]} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a) \\
&\quad - \sum_{i\in[p]} A_{bi}^2\eta_0'(\sum_{a\neq b} A_{ai}w_a))w_b + \delta^{-1}z_b^0 \left\langle \eta_0'(x^1) \right\rangle . \qquad (8.30)
\end{aligned}
$$

Remark that although $\eta_0(\sum_{a\neq b} A_{ai}w_a)$ dominates $\eta_0'(\sum_{a\neq b} A_{ai}w_a))A_{bi}w_b$, the $\sum_{i\in[p]} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a)$ is comparable to $\sum_{i\in[p]} A_{bi}^2\eta_0'(\sum_{a\neq b} A_{ai}w_a))w_b$, since the former is zero mean with variance of order $p/n$, and the latter has mean $p/n$. In other words, in the Taylor expansion step we decomposed $\eta_0(\sum_{a\in[n]} A_{ai}w_a)$ into a main term uncorrelated with $A_{bi}$ and another term smaller but correlated with $A_{bi}$.

Since $\eta_0(\sum_{a\neq b} A_{ai}w_a)$ is approximately $x_i^1$, we see that the two terms on the first line in (8.30) is approximately $\sigma Z' + \sqrt{\delta^{-1}\hat{\tau}_1^2}Z$ as in the idealized setting before, so it remains to show that the two terms in the last line in (8.30) cancel. To see the cancellation, note that by concentration of measure and independence, we have

$$
\sum_{i\in[p]} A_{bi}^2\eta_0'(\sum_{a\neq b} A_{ai}w_a))w_b \approx \mathbb{E}\left[\sum_{i\in[p]} A_{bi}^2\eta_0'(\sum_{a\neq b} A_{ai}w_a))\right] \qquad (8.31)
$$

$$
= \sum_{i\in[p]} \mathbb{E}\left[A_{bi}^2\right]\mathbb{E}\left[\eta_0'(\sum_{a\neq b} A_{ai}w_a))\right] \qquad (8.32)
$$

$$= \frac{1}{n}\sum_{i\in[p]} \mathbb{E}\left[\eta_0'(\sum_{a\neq b} A_{ai}w_a))\right] \tag{8.33}$$

$$\approx \delta^{-1}\left\langle \eta_0'(x^1)\right\rangle . \tag{8.34}$$

Thus indeed we have the following which justifies the cancellation:

$$z_b^1 \approx w_b - \sum_{i\in[p]} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a). \tag{8.35}$$

Consider one more iteration: for each $j \in [p]$,

$$x_j^2 = \eta_1(\sum_{b\in[n]} A_{bj}z_b^1 + x_j^1). \tag{8.36}$$

Inside the parenthesis,

$$\begin{aligned}
&\sum_{b\in[n]} A_{bj}z_b^1 + x_j^1 \\
&\approx \sum_{b\in[n]} A_{bj}\left(w_b - \sum_{i\in[p]} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a)\right) + x_j^1 && (8.37) \\
&= \sum_{b\in[n]} A_{bj}\left(w_b - \sum_{i\neq j} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a)\right) \\
&\quad - \sum_{b\in[n]} A_{bj}^2\eta_0(\sum_{a\neq b} A_{aj}w_a) + x_j^1 && (8.38)
\end{aligned}$$

$$\approx \sum_{b\in[n]} A_{bj} \left( w_b - \sum_{i\neq j} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a) \right). \tag{8.39}$$

The approximation in (8.38) follows since

$$\begin{aligned}\sum_{b\in[n]} A_{bj}^2\eta_0(\sum_{a\neq b} A_{aj}w_a) &\approx \sum_{b\in[n]} A_{bj}^2\eta_0(\sum_{a\in[n]} A_{aj}w_a) && (8.40)\\ &\approx \eta_0(\sum_{a\in[n]} A_{ai}w_a)\\ &= x_i^1. && (8.41)\end{aligned}$$

By the independence of $\{A_{bj}\}_{b\in[n]}$ and $\left\{w_b - \sum_{i\neq j} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a)\right\}_{b\in[n]}$, (8.39) is approximately Gaussian with variance

$$\begin{aligned}&\frac{1}{n}\left\|\left\{w_b - \sum_{i\neq j} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a)\right\}_{b\in[n]}\right\|_2^2 && (8.42)\\ &\approx \frac{1}{n}\left\|\left\{w_b - \sum_{i\in[p]} A_{bi}\eta_0(\sum_{a\neq b} A_{ai}w_a)\right\}_{b\in[n]}\right\|_2^2 && (8.43)\\ &\approx \frac{1}{n}\left\|z^1\right\|_2^2 && (8.44)\\ &\approx \tau_1^2. && (8.45)\end{aligned}$$

Thus, returning to (8.36), $x_j^2$ is approximately $\eta_t(\tau_1 Z)$ in distribution, which verifies (8.20) for $t = 2$.

The pattern of the induction is clear now. In the special case of $\eta_t$ being the identity function (i.e., the least squares example), we can approximate the iterations by the explicit formulae

$$x_i^t \approx \sum_{s=1}^{t} \frac{1-(-1)^s}{2} \sum_{\gamma \in \Gamma_i^s} L(\gamma); \tag{8.46}$$

$$z_a^t \approx \sum_{s=0}^{t} \frac{1+(-1)^s}{2} \sum_{\bar{\gamma} \in \bar{\Gamma}_a^s} L(\bar{\gamma}). \tag{8.47}$$

Here, $\Gamma_i^s$ is the set of all $s$-step *none-returning* paths on the compelte bipartite graph $K_{p,n}$ starting from the left vertex $i$, and $L(\gamma)$ is the cost of the walk defined by $A$ and $w$. Non-returning means $\gamma(s+1) \neq \gamma(s-1)$ for each time $s$. For example, $j \to b \to i \to a$ is a path in $\Gamma_j^3$ if $i \neq j$, $a \neq b$, and the associated cost is

$$L(\gamma) = A_{bj} A_{bi} A_{ai} w_a. \tag{8.48}$$

Similarly, $\bar{\Gamma}_a^s$ is the set of non-returning paths starting from the right vertex $a$ of length $s$. For example, $\bar{\gamma} \in \bar{\Gamma}_c^4$ if it is a path $c \to j \to b \to i \to a$ where $b \neq c$, $i \neq j$, $a \neq b$, and the associated cost is

$$L(\bar{\gamma}) = A_{cj} A_{bj} A_{bi} A_{ai} w_a. \tag{8.49}$$

We can verify that (8.46)-(8.47) satisfy the AMP iterations. Non-returning paths enter the picture because once the path returns, the corresponding terms cancel with the $x^t$ or the $\delta^{-1}z^{t-1}\left\langle\eta'_{t-1}(x^t)\right\rangle$ term.

*Remark* 14. The path representation is also useful in random matrix theory: it can be used to compute all the moments of the spectral measure (Exercise 32), and then characterize the limiting law of the spectrum of the random matrix.

# Chapter 9

# Problem Set

[updated weekly. A total of $\geq 25$ points of problems due before the midterm and another $\geq 25$ points of problems due before the final.]

SGD query lower bound, variable selection

**Exercise 1** (2 points)**.** Use the union bound to prove the estimate of the Gaussian max in (1.11). In fact, note that the bound holds even when $\xi_i$'s are correlated.

**Exercise 2** (2 points)**.** Show that the soft thresholding estimator (see (1.21)) achieves the oracle $\ell_2$ error up to a logarithmic factor.

**Exercise 3** (2 points)**.** In the section on the James-Stein estimator, suppose that the noise may be biased, i.e., $\xi_i \sim \mathcal{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$. Find an analogue of the James-Stein estimator in this setting, so that the naive estimator $\hat{\theta}_{\mathsf{naive}} = Y - \mu$ is shown to be inadmissible.

**Exercise 4** (4 points)**.** [Gaussian integration by parts] Suppose that $f$ is a smooth function with compact support on $\mathbb{R}^d$. Let $Y \sim \mathcal{N}(\theta, \sigma^2 I_{d\times d})$. Use integration by parts to show that

$$\mathbb{E}_\theta[(\theta_i - Y_i)f(Y)] = -\epsilon^2 \mathbb{E}_\theta\left[\frac{\partial f}{\partial y_i}(Y)\right] \tag{9.1}$$

for $i = 1, \ldots, d$.
*Note: once the formula is proven for nice (smooth, compactly supported function), it can be extended to more general class of functions by standard approximation arguments.*

**Exercise 5** (2 points)**.** Show that when $n = d$ and $\mathbb{X} = I_d$, BIC and Lasso are reduced to the hard-thresholding and the soft thresholding estimators for the Gaussian sequence model.

**Exercise 6** (4 points)**.** Suppose that the basis pursuit algorithm in (1.62) is replaced by the following[1]

$$\hat{\theta}^{BP} \in \text{argmin}_{\theta:\, Y=\mathbb{X}\theta} \|\theta\|_q. \tag{9.2}$$

where $q \in (0, 1]$ and $\|\theta\|_q := (\sum_{j=1}^d |\theta_j|^q)^{1/q}$. For such general $q$, write a counterpart of the null space condition (Definition 9), and prove a counterpart of the equivalence result Theorem 10.

[1] Note this is no longer a convex optimization when $q < 1$. In practice we can try to solve this optimization by gradient descent and the performance turns out to be quite good empirically, although there is no theoretical guarantees of finding the global minimum.

**Exercise 7** (2 points)**.** Use Chernoff's inequality to prove the following Chi-square tail bound: for $X^d \sim \mathcal{N}(0, I_d)$ and any $t > 0$,

$$\mathbb{P}[| \|X^d\|_2^2 - d| \geq 2\sqrt{dt} + 2t] \leq 2e^{-t}. \tag{9.3}$$

*[Hint: p. 43 [BLM13]]*

**Exercise 8** (4 points)**.** Observe that Gordon's theorem (Theorem 13) shows that whenever $w(\mathcal{K}) \leq 0.5\sqrt{n}$ and $n$ is large, than $\mathcal{K} \cap \nu = \emptyset$ with high probability. Find an example of $\mathcal{K}$ satisfying $w(\mathcal{K}) \leq \sqrt{n}$ yet $\mathcal{K} \cap \nu \neq \emptyset$ almost surely. (This shows the sharpness of Gordon's theorem.)

**Exercise 9** (3+3 points)**.** Let $\mathcal{K}$ be a convex cone in $\mathbb{R}^d$. Show the following relation between the statistical dimension and the Gaussian width:

$$w^2(\mathcal{K} \cap S^{d-1}) \leq \delta(\mathcal{K}) \leq w^2(\mathcal{K} \cap S^{d-1}) + 1 \tag{9.4}$$

where $S^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. Each side of the inequality is worth 3 points. [*Hint: the left inequality follows relatively easily from Jensen's inequality. To show the right inequality, we need* $\mathrm{Var}(\sup_{z \in \mathcal{K} \cap S^{d-1}} z^\top G) \leq 1$*, which can be shown using the fact that a 1-Lipschitz function of a Gaussian vector has variance* $\leq 1$ *(Gaussian poincare inequality).*]

**Exercise 10** (4 points)**.** Recall the descent cone $\mathcal{D}(\theta)$ defined in (1.95). Show that if $\theta \in \mathbb{R}^d$ is $k$-sparse ($d \geq 2k$) then the statistical dimension has the order $k \log \frac{d}{2k} \lesssim \delta(\mathcal{D}(\theta)) \lesssim k \log d$. [*Hint: similar to Theorem 14. The difference is that the set of interest here is not union over all k-sparse vectors.*]

**Exercise 11** (2 points)**.** Suppose that $X \in \mathbb{R}^p$ follows the distribution $\mathcal{N}(0, \Theta^{-1})$. Let $u, v \in \{1, \ldots, p\}$ and $u \neq v$. Show that the distribution of $(X_u, X_v)$ given $X_{\{1,\ldots,p\}\setminus\{u,v\}} = 0$ is $\mathcal{N}(0, \bar{\Theta}^{-1})$ where $\bar{\Theta} \in \mathbb{R}^{2\times 2}$ denotes the restriction of $\Theta$ on the $u$-th and $v$-th rows and columns. Then verify the claim in (2.3).

**Exercise 12** (2 points)**.** Let $S$ be a rank-deficient positive-semidefinite matrix. Show that the maximum likelihood estimate of the precision matrix in the Gaussian graphical model (see (2.6)) does not exist (i.e. there is no $\Theta$ achieving the maximum).

**Exercise 13** (2 points)**.** Verify that graphical lasso (2.7) is convex optimization.

**Exercise 14** (2+2 points)**.** Let $\mathbb{Y} \in \mathbb{R}^{n\times T}$ be a given matrix.

- Let $k \leq \min\{n, T\}$. Find an explicit formula for

  $$\operatorname{argmin}_{\Theta\in\mathbb{R}^{n\times T},\, \operatorname{rank}(\Theta)\leq k} \|\mathbb{Y} - \Theta\|_F^2.$$

  [*Hint: consider the singular value decomposition of* $\mathbb{Y}$.]

- Let $\mathbb{X} \in \mathbb{R}^{n\times d}$ be another given matrix. Let $k \leq \min\{d, T\}$. Find an explicit formula for

  $$\operatorname{argmin}_{\Theta\in\mathbb{R}^{d\times T},\, \operatorname{rank}(\Theta)\leq k} \|\mathbb{Y} - \mathbb{X}\Theta\|_F^2.$$

[*Hint: use the decomposition* $\|\mathbb{Y}-\mathbb{X}\Theta\|_F^2 = \|\mathbb{Y}-\bar{\mathbb{Y}}\|_F^2+\|\bar{\mathbb{Y}}-\mathbb{X}\Theta\|_F^2$ *where* $\bar{\mathbb{Y}} := \mathbb{X}(\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top\mathbb{Y}$ *denotes the orthogonal projection of* $\mathbb{Y}$ *onto the image space of* $\mathbb{X}$.]

**Exercise 15** (3 points)**.** In Theorem 18, suppose that $\Lambda_j$ and $\Lambda_j^*$ are all smooth convex functions. Show that $u_1, \ldots, u_k$ achieve the infimum if there exists $v$ such that any one of the following is true:

- $\Lambda_j^*(v) + \Lambda_j(u_j) = \langle v, u_j\rangle$ for each $j$.
- $\nabla\Lambda_j(u_j) = v$ for each $j$.
- $\nabla\Lambda_j^*(v) = u_j$ for each $j$.

**Exercise 16** (2+2 points)**.** Prove the relation of the coherence parameters in Definition 24-25

$$\mu_1 \leq \mu_2 \leq \mu_1^2 r. \tag{9.5}$$

[*Hint: each inequality follows by a one-line proof. You may also try to find it in [CLMW11, Che15]*]

**Exercise 17** (4 points)**.** Let

$$L_0 = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n^2} \tag{9.6}$$

where $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix with positive entries. Show that the subdifferential of the Schatten 1-norm at $L_0$ takes the form

$$\partial\|L_0\|_{\mathsf{s1}} = \left\{ \begin{pmatrix} I_r & 0 \\ 0 & W \end{pmatrix} \right\} \tag{9.7}$$

where $W \in \mathbb{R}^{(n-r)\times(n-r)}$ is any matrix with operator norm bounded by 1.

**Exercise 18** (2 points)**.** Show the equivalence of (5.9) and (5.10)

**Exercise 19** (2 points)**.** Show the data processing inequality for the KL divergence, that is, justify the step (5.15) in the proof of Fano's inequality.

**Exercise 20** (2 points)**.** Use the definition of the KL divergence to give a proof of the step (5.20).

**Exercise 21** (2 points)**.** Justify (5.23); more precisely, show that

$$|\mathcal{B}| = \Theta\left(\frac{1}{\sqrt{d}}\exp(dh(k/d))\right) \tag{9.8}$$

where $\mathcal{B}$ denotes the Hamming ball in $\{0,1\}^d$ of radius $k$. [*Hint: Let $X_1, \ldots, X_d$ be i.i.d.* $\mathrm{Ber}(k/d)$*. Show that $\mathbb{P}[\rho(X^d) = k] = \Theta(\frac{1}{\sqrt{d}})$; show that $-\log\mathbb{P}[X^d = x^d] = dh(k/d)$ for any $x^d \in \mathcal{B}$.*]

**Exercise 22** (4 points)**.** Let $P_0$ and $P_1$ be two probability measures. Show that for any test $\psi$,

$$\max_{j=0,1} P_j[\psi \neq j] \geq \frac{1}{4}\exp(-D(P_0 \| P_1)). \tag{9.9}$$

[*Hint: Consider the data processing inequality for the KL divergence.*]

**Exercise 23** (5 points)**.** In the Gaussian sequence model, suppose that the parameter space is $\mathcal{B}_1(R) = \{\theta \colon \|\theta\|_1 \leq R\}$, where $R$ is polynomial in $d$. Show that for any $\delta \in (0,1)$, there exists $c > 0$ such that

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathcal{B}_1(R)} \mathbb{P}[\|\hat{\theta} - \theta\|_2^2 \geq cR\sigma\sqrt{\log \frac{d\sigma}{R}}] \geq 1 - \delta. \tag{9.10}$$

[*Hint: Choose a packing in which vectors are $k$-sparse and each have 1-norm $R$. The squared risk is then $R^2/k$ whereas the mutual information is order $\frac{R^2}{k\sigma^2}$ which should equal $k \log \frac{d}{k}$ by the GV bound. The latter shows that we should pick $k \sim \frac{R}{\sigma}\sqrt{\frac{1}{\log \frac{d}{k}}}$.*]

**Exercise 24** (3 points)**.** Prove the equivalence in distribution in (6.2). [*Hint: First show the rotation invariance of $\hat{\theta}$: if $Q \in \mathbb{R}^{p \times p}$ is any (deterministic) orthogonal matrix, then $\hat{\theta}$ and $Q\hat{\theta}$ have the same distribution.*]

**Exercise 25** (2 points)**.** Given a proof of the Sherman-Morrison formula for the rank-1 update for matrix inversion:

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u} \tag{9.11}$$

where $A$ is an invertible matrix and $u, v$ are vectors.

**Exercise 26** (3 points)**.** In the example of Section 6.1, show that the test error converges to $\frac{\delta}{\delta-1}$. That is, when there is new data $w_{\text{new}}, A_{\text{new}}$, we have

$$\mathbb{E}[|A_{\text{new}}\hat{\theta} - w_{\text{new}}|^2] \to \frac{\delta}{\delta - 1} \tag{9.12}$$

as $p \to \infty$.

**Exercise 27** (4 points)**.** Suppose that, in lieu of (9.13), we define the partition function as

$$Z(\beta, \gamma, n) := \int e^{-\frac{\beta}{2n}\|\mathbf{w}-\mathbf{A}\theta\|^2 + \gamma\sum_{j=1}^{p}\phi(\theta_j)}\mathrm{d}\theta, \tag{9.13}$$

where both $\beta, \gamma > 0$ are constants independent of $n$, and $\phi(\cdot)$ is a given function on $\mathbb{R}$. Show that

$$\lim_{\beta\to\infty}\lim_{p\to\infty}\frac{1}{p}\frac{\partial}{\partial\gamma}\mathbb{E}[\ln Z(\beta, \gamma, n)]\bigg|_{\gamma=0} = \lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\phi(\hat{\theta}_j) \tag{9.14}$$

where $\hat{\theta}$ denotes the solution to the least squares problem (7.2).

**Exercise 28** (2 points)**.** In the setting of Section 7.1.7, show that

$$\lim_{\beta \to +\infty} \frac{1}{\beta} \mathbb{E}[\ln Z(\beta, n)] = -\frac{1}{2n} \min_{\theta} \|\mathbf{w} - \mathbf{A}\theta\|^2 \tag{9.15}$$

**Exercise 29** (4 points)**.** Justify the step (7.35), that is, show that

$$\mathbb{P}[\widehat{P}_{X^n} = P] \doteq \exp(-nD(P\|Q)). \tag{9.16}$$

[*Hint: first show that for any* $x^n$ *satisfying* $\widehat{P}_{x^n} = P$, *we have* $\mathbb{P}[X^n = x^n] = \exp(-n\mathbb{E}_P[\log \frac{1}{Q(X)}])$. *Then use the result of Exercise 21 to show the cardinality estimate* $|\{x^n \colon \widehat{P}_{x^n} = P\}| \doteq \exp(-n\mathbb{E}_P[\log P(X)])$.]

**Exercise 30** (3 points)**.** In deriving (7.36), why did we need the fact that there are only polynomially many summands in (7.35)?

**Exercise 31** (3 points)**.** Show that it is information-theoretically impossible to find a planted clique of size $2(1-\delta)\log_2 n$ with error probability $1-\delta$ in an Erdos-Renyi graph of size $n$ with edge connection probability 1/2. [*Hint: The number of* $k$*-cliques is* $\binom{n}{k}$. *Moreover,* $D(P\|Q) = \log 2^{\frac{k(k-1)}{2}}$, *where* $P$ *is the edge distribution in any planted* $k$*-clique graph, and* $Q$ *is the edge distribution in the Erdos-Renyi graph. Then use the Fano inequality.*]

**Exercise 32** (4 points)**.** Let $A^{n\times p}$ be a random matrix with i.i.d. entries, $A_{ij} \sim \mathcal{N}(0,1)$, and $n/p = \delta$ is fixed. Use either the

Marchenko-Pastur Law or the path method to compute the limit $\mathbb{E}[\mathrm{tr}(A^\top AA^\top A)]$ as $p \to \infty$. [*Hint: for the Marchenko-Pastur Law, try to apply change of variables and use the fact that* $\int \mu = 1$. *For the path method, it should be similar to the computation in Section 8.2.4, but simpler.*] [*Note: once all moments of the spectral measure are calculated, it is not far from deriving the Marchenko-Pastur Law; this is called the moment method [AGZ10].*]

**Exercise 33** (3 points)**.** Let $A = A^\top \in \mathbb{R}^{n\times n}$ be the Wigner random matrix where the $A_{i,j}$ are independent for $j \geq i$, $A_{i,j} \sim \mathcal{N}(0,1)$ for $i \neq j$, and $A_{i,i} \sim \mathcal{N}(0,2)$. Set $x^0$ be the all 1's vector, and consider the iterations

$$x^{t+1} = Ax^t - x^{t-1}. \tag{9.17}$$

Show that for each fixed $t$, the distribution of $x_1^t$ converges to $\mathcal{N}(0,1)$ as $n \to \infty$. [*Hint: similar to the computation in Section 8.2.4, but simpler. Work with an undirected graph instead of a bipartite graph. This exercise was also mentioned in* ***https://www.youtube.com/watch?v=ZlTNcXzcemA&t=1523s***]

# Bibliography

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Number 118. Cambridge university press `http://www.wisdom.weizmann.ac.il/~zeitouni/cupbook.pdf` see also the summary `https://www.math.harvard.edu/media/feier.pdf`, 2010.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BM11a] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[BM11b] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[BR$^+$13] Quentin Berthet, Philippe Rigollet, et al. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(4):1780–1815, 2013.

[CDD09] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best $k$-term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

[CHE98] SS CHEN. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[Che15] Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[CK11] Imre Csiszár and János Körner. *Information theory:*

*coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[CMW20] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

[CR09] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.

[CSPW11] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[DDDM04] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[dem10] *Large Deviations Techniques and Applications*. Springer, 2010.

[DH06] DL Donoho and X Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2006.

[DMM09] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[Don06] David L Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete & Computational Geometry*, 35(4):617–652, 2006.

[DT05] David L Donoho and Jared Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102(27):9452–9457, 2005.

[EB02] Michael Elad and Alfred M Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 48(9):2558–2567, 2002.

[EKBB$^+$13] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[FN03] Arie Feuer and Arkadi Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579–1581, 2003.

[Gor88] Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in ? n. In *Geometric aspects of functional analysis*, pages 84–106. Springer, 1988.

[GV05] D Guo and S Verdu. Randomly spread cdma: asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, 2005.

[HGT06] Kyle K Herrity, Anna C Gilbert, and Joel A Tropp. Sparse approximation via iterative thresholding. In *2006 IEEE International Conference on Acoustics*

*Speech and Signal Processing Proceedings*, volume 3, pages III–III. IEEE, 2006.

[IR08] Piotr Indyk and Milan Ruzic. Near-optimal sparse recovery in the l1 norm. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 199–207. IEEE, 2008.

[JM14] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[Kar77] Richard M Karp. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Mathematics of operations research*, 2(3):209–224, 1977.

[KLT$^+$11] Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[LCCV18] Jingbo Liu, Thomas A Courtade, Paul W Cuff, and Sergio Verdú. A forward-reverse brascamp-lieb inequality: Entropic duality and gaussian optimality. *Entropy*, 20(6):418, 2018.

[LJG15] Jingbo Liu, Jian Jin, and Yuantao Gu. Robustness of sparse recovery via $f$-minimization: A topological viewpoint. *IEEE Transactions on Information Theory*, 61(7):3996–4014, 2015.

[LMN$^+$20] Martin Lotz, Michael B McCoy, Ivan Nourdin, Giovanni Peccati, and Joel A Tropp. Concentration of the intrinsic volumes of a convex body. In *Geometric Aspects of Functional Analysis*, pages 139–167. Springer, 2020.

[LR19] Jingbo Liu and Philippe Rigollet. A perspective on false discovery rate control via knockoffs. In *Proc. Neural Information Processing Systems (NIPS)*, 2019.

[LW13] Po-Ling Loh and Martin J Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, pages 3022–3049, 2013.

[MH12] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

[MT14] Michael B McCoy and Joel A Tropp. From steiner formulas for cones to concentration of intrinsic volumes. *Discrete & Computational Geometry*, 51(4):926–963, 2014.

[RFG12] Sundeep Rangan, Alyson K Fletcher, and Vivek K Goyal. Asymptotic analysis of map estimation via the replica method and applications to compressed sensing. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 58(3), 2012.

[Rig15] Philippe Rigollet. High dimensional statistics lecture notes. ***https://ocw.mit.edu/courses/mathematics/18-s997-high-dimensional-statistics-spring-2015/lecture-notes/***, 2015.

[RV08] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045, 2008.

[RWY10] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[SVH19] Yair Shenfeld and Ramon Van Handel. Mixed volumes and the bochner method. *Proceedings of the American Mathematical Society*, 147(12):5385–5402, 2019.

[SW08] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Springer Science & Business Media, 2008.

[Tan02] Toshiyuki Tanaka. A statistical-mechanics approach to large-system analysis of cdma multiuser detectors. *IEEE Transactions on Information theory*, 48(11):2888–2910, 2002.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[TG07] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[TLR21] P Turner, J Liu, and P Rigollet. A statistical perspective on coreset density estimation. In *The 24th In-*

*ternational Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[Tsy08] Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Science & Business Media, 2008.

[vH14] Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

[WL$^+$08] Tong Tong Wu, Kenneth Lange, et al. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.

[YS18] Christina Lee Yu and Devavrat Shah. What do we know about matrix estimation? *ISIT 2018 tutorial*, 2018.

[Yu97] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.