A Tutorial of Optimization

最佳化基礎介紹

CHUN-JEN SHIH

revised by Jian-Jiun Ding

施淳仁 著　　丁建均老師 修訂

Graduate Institute of Communication Engineering

National Taiwan University

國立台灣大學電信工程學研究所

Contents

Abstract

In this tutorial, we give a fruitful introduction about concepts of optimization in multiple aspects with mathematical rigorous. Chapter one provides some fundamental knowledge of convex analysis, which is mostly devoted when making mathematical derivations in chapter three. Chapter two provides some fundamental knowledge of linear algebra, which is mostly devoted to chapter four and five. In chapter three, we introduce abundant algorithms for solving unstructured optimization problems. The paradigms of those algorithms can be seen in some important viewpoints. The EM algorithm explained in the last section is an important algorithm to deal with maximum likelihood estimation problems, which can also be seen as a kind of unstructured optimization problem. In chapter four, we focus on the ℓ_0 minimization problem with which the compressive sensing problems and the sparse representation tasks have close relationship. ℓ_1 minimization approaches, greedy algorithms and hard-thresholding-based algorithms are covered. In chapter five, we elaborate a few crucial techniques, e.g., the L-curve method, model order selection and dictionary screening, which prove to be helpful when tackling optimization problems. Issues about standard procedures to conduct experiments and evaluate performance for the compressive sensing problems are also briefly discussed.

Notation

1. $\mathbb{N}$: the set of natural numbers
2. $\mathbb{R}^n$: the set of n-dimensional real vectors
3. $Re\cdot$: the operation of taking the real part of a complex scalar
4. $\langle \cdot, \cdot \rangle$: the operation of taking the inner product of two complex vectors
5. $\inf \cdot$: the operation of taking the infimum of an objective function
6. $\sup \cdot$: the operation of taking the supremum of an objective function
7. A^* : the conjugate transpose of a matrix A
8. f^* : the convex conjugate of a function f
9. ∇f : the gradient vector of a differentiable function f
10. $\nabla^2 f$: the Hessian matrix of a twice differentiable function f
11. ∂f : the subdifferential of a function f
12. $[N]$: the set $\{i \in \mathbb{N} \,|\, i \leq N\}$
13. $a[i]$: If $a \in \mathbb{R}^m$ and i is an integer less than or equal to m, then $a[i]$ denotes the i-th entry of a
14. $x^{(k)}$: In an iterative algorithm, we use $x^{(k)}$ to denote the iterate x at the k-th iteration
15. x_s : x is an n-dimensional vector and s is a positive integer. In the context of compressive sensing or sparse representation, we use x_s to denote two possible vectors. One is an s-dimensional vector whose entries are the s largest components of x in modulus. The other one is an n-dimensional vector obtained by holding the s largest components of x in modulus

unchanged and setting the remaining components of x to be zero. It depends on the context to determine which vector we denote.

16. $x|_T$: x is an n-dimensional vector and T is a subset of $[n]$. In the context of compressive sensing or sparse representation, we use $x|_T$ to denote two possible vectors. One is obtained by retaining all $x[i]$, $i \in T$ and setting all the other components to be zero. That is, $x|_T = \begin{cases} x[i] & i \in T \\ 0 & otherwise \end{cases}$. The other one is obtained by retaining all $x[i]$, $i \in T$ and removing all the other components. Thus, the resulting vector $x|_T$ has dimension equal to the cardinality of T. It depends on the context to determine which vector we denote.

17. A_T : A is an $m \times n$-dimensional matrix and T is a subset of $[n]$. In the context of compressive sensing or sparse representation, we denote A_T as the column submatrix of A whose columns are listed in the set T

18. card(S) : the cardinality of a set S

19. $\equiv$: be equivalent to

20. $\triangleq$: be defined as

21. $:=$: be defined as

22. $\succeq$: vector inequality or component-wise inequality. If a vector $a \in \mathbb{R}^m \succeq$ another vector $b \in \mathbb{R}^m$, then $a[i] \geq b[i]$ $\forall i \in [m]$

23. $\Rightarrow$: implies

24. $\xrightarrow{D}$: convergence in distribution to

25. $\xrightarrow{p}$: convergence in probability to

26. $N(\mu, \Sigma)$: the Gaussian probability density function with mean μ and covariance Σ. We will also use the notation $N(y; \mu, \Sigma)$ if we want to point out the random variable y that follows the Gaussian probability density function $N(\mu, \Sigma)$.

27. $sgn(\cdot)$: the sign function. If z is a complex scalar,

$$sgn(z) := \begin{cases} \frac{z}{|z|} & , when\ z \neq 0 \\ 0 & , when\ z = 0 \end{cases}$$

If $z \in \mathbb{C}^n$, we denote $sgn(z)$ as the vector with components $sgn(z[j])$, $j \in [n]$.

Chapter 1

Fundamental Knowledge of Convex Analysis

1.1 Optimality Condition

Theorem 1.1.1 (optimality condition)**.** *Assume $f : \mathbb{R}^n \to (-\infty, \infty]$ is a convex and differentiable function. Let $\mathcal{X} \subseteq \mathbb{R}^D$ be a closed and convex set. Then $x^* \in \underset{x\in\mathcal{X}}{argin f}\, f(x)$ if and only if*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in \mathcal{X} \tag{1.1}$$

Proof. Assume 1.1 holds. Because f is a convex function,

$$f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle \geq f(x^*)$$

Hence, $x^* \in \underset{x\in\mathcal{X}}{argin f}\, f(x)$. Assume 1.1 does not hold for some $y \in \mathcal{X}$. Let $\phi(a) := f(x^* + a(y - x^*))$, $a \in [0, 1]$. $\phi(0) = f(x^*)$ and $\phi'(0) = \langle \nabla f(x^*), y - x^* \rangle < 0$. Hence, we have $f(x^* + a(y - x^*)) < f(x^*)$ for some small enough a. Therefore, we prove the "only if " part. □

An important application of the optimality condition is the projection theorem.

Theorem 1.1.2 (projection theorem)**.** *Let $\mathcal{X}$ be a closed and convex subset of $\mathbb{R}^n$ and let $z \in \mathbb{R}^n$. There exists a unique vector $x^* \in \mathcal{X}$ that minimizes $\|z-x\|_2$ over $x \in \mathcal{X}$, called the projection of z onto $\mathcal{X}$. Moreover, x^* is the projection of z onto $\mathcal{X}$ if and only if*

$$\langle z - x^*, x - x^* \rangle \leq 0, \quad \forall x \in \mathcal{X} \tag{1.2}$$

Proof. Minimizing $\|z-x\|_2$ is equivalent to minimizing $f \triangleq \frac{1}{2}\|z-x\|_2^2$. Since f is a convex and differentiable function, by the optimality con-

dition, the necessary and sufficient condition for x^* to be the projection of z onto $\mathcal{X}$ is clearly 1.2. Assume there are such vectors x_1^* and x_2^*. To prove the uniqueness, we need to verify that $x_1^* = x_2^* = x^*$. Indeed, since x_1^* is the projection of z onto $\mathcal{X}$, $\langle z - x_1^*, x_2^* - x_1^* \rangle \leq 0$ because of 1.2. Similarly, $\langle z - x_2^*, x_1^* - x_2^* \rangle \leq 0$. Adding these two inequalities, we get $\|x_2^* - x_1^*\|_2^2 \leq 0$, which implies $x_1^* = x_2^* = 0$. □

Because of the projection theorem, we can verify the non-expansive property of the projector.[1]Assume there are two vectors $z_1 \in \mathbb{R}^n$ and $z_2 \in \mathbb{R}^n$. We want to prove that

$$\|x_1^* - x_2^*\|_2 \leq \|z_1 - z_2\|_2 \tag{1.3}$$

where $x_1^* \in \mathcal{X}$ and $x_2^* \in \mathcal{X}$ are the projections of z_1 and z_2 onto $\mathcal{X}$. Because of 1.2, we have

$$\begin{aligned} \langle z_1 - x_1^*, x_2^* - x_1^* \rangle &\leq 0 \\ \langle z_2 - x_2^*, x_1^* - x_2^* \rangle &\leq 0 \end{aligned}$$

Adding these two inequalities, we get $\langle x_2^* - x_1^* + z_1 - z_2, x_2^* - x_1^* \rangle \leq 0$. As a result,

$$\begin{aligned} \|x_1^* - x_2^*\|_2^2 &\leq \langle z_1 - z_2, x_1^* - x_2^* \rangle \\ &\leq \|z_1 - z_2\|_2 \|x_1^* - x_2^*\|_2 \\ \Rightarrow \|x_1^* - x_2^*\|_2 &\leq \|z_1 - z_2\|_2 \end{aligned}$$

The second inequality is due to the Hölder's inequality 2.6.

1.2 Strong Convexity

Definition 1.2.1 (strong convexity)**.** A function $f : \mathbb{R}^n \to (-\infty, \infty]$ is m-strongly convex for some $m > 0$ if one of the following equivalent conditions holds

1. The function f is twice differentiable and satisfies

$$\nabla^2 f(x) \geq mI \quad \forall x \in \mathbb{R}^n \tag{1.4}$$

[1]We will introduce the concept of non-expansive mapping in more detail in section 1.9.

2. The function f is differentiable and satisfies

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq m\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n \tag{1.5}$$

3. The function f is differentiable and satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n \tag{1.6}$$

An important consequence of strong convexity is that a differentiable strongly convex function has a unique minimizer on a closed convex set. To prove this, we assume $f : \mathbb{R}^n \to (-\infty, \infty]$ is a m-strongly convex function and $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set. First, we need to prove the existence of a minimizer. Assume $y \in \mathcal{X}$

$$\begin{aligned} f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{m}{2}\|x - y\|_2^2 \quad \forall x \in \mathcal{X} \\ &\geq f(y) - \|\nabla f(y)\|_2\|x - y\|_2 + \frac{m}{2}\|x - y\|_2^2 \\ &\geq f(y) \quad if \ \|x - y\|_2 > \frac{2\|\nabla f(y)\|_2}{m} \end{aligned}$$

Hence, the set $\{x \in \mathcal{X} | f(x) \leq f(y)\}$ is bounded. According to the extreme value theorem, a continuous function attains its maximum and minimum on a closed and bounded set. Next, assume a minimizer of f is x^*.

$$\begin{aligned} f(x) &\geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{m}{2}\|x - x^*\|_2^2 \quad \forall x \in \mathcal{X} \\ &\geq f(x^*) + \frac{m}{2}\|x - x^*\|_2^2 \quad \because 1.1 \end{aligned}$$

Hence, $f(x) = f(x^*)$ if and only if $x = x^*$, which indicates the uniqueness of the minimizer.

1.3 Smoothness

Definition 1.3.1 (smoothness)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is M-smooth for some $M > 0$ if one of the following equivalent conditions holds

1. The function f is twice differentiable and satisfies

$$\nabla^2 f(x) \leq MI \quad \forall x \in \mathbb{R}^n \tag{1.7}$$

2. The function f is differentiable and satisfies

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq M\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n \tag{1.8}$$

3. The function f is differentiable and satisfies

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n \tag{1.9}$$

Furthermore, if the gradient of f is M-Lipschitz with respect to the ℓ_2 norm, i.e.,

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq M\|y - x\|_2 \quad \forall x, y \in \mathbb{R}^n \tag{1.10}$$

then

$$\begin{aligned} &\langle \nabla f(y) - \nabla f(x), y - x \rangle \\ \leq &\|\nabla f(y) - \nabla f(x)\|_2 \|y - x\|_2 \\ \leq &M\|y - x\|_2^2 \end{aligned}$$

Hence, f is M-smooth.

1.4 Relative Smoothness

Definition 1.4.1 (Bregman divergence)**.** Let $h : \mathbb{R}^n \to (-\infty, \infty]$ be a convex function. The Bregman divergence associated with h is defined as

$$D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle \tag{1.11}$$

Since h is convex, $D_h(y, x)$ is non-negative. If $h(x) = \frac{1}{2}\|x\|_2^2$, then

$$D_h(y, x) = \frac{1}{2}\|y - x\|_2^2 \tag{1.12}$$

If $h(x) = \sum_{i=1}^{n} x[i] \log x[i]$, which is called the negative entropy, then

$$D_h(y, x) = \sum_{i=1}^{n} y[i] \log \frac{y[i]}{x[i]} \tag{1.13}$$

which is the relative entropy.

Definition 1.4.2 (relative smoothness)**.** A function is called M-smooth relative to a convex function h for some $M > 0$ if and only if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + M D_h(y, x) \tag{1.14}$$

If $h(x) = \frac{1}{2}\|x\|_2^2$, then the relative smoothness coincides with smoothness introduced in section 1.3. Furthermore, it can be easily verified that a function f is M-smooth relative to a convex function h if and only if the function $Mh - f$ is convex.

1.5 Subdifferential

Definition 1.5.1 (subdifferential)**.** Assume $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. The subdifferential of f at $x \in \mathbb{R}^n$ is defined as

$$\partial f(x) := \{v \in \mathbb{R}^n : f(z) \geq f(x) + \langle v, z - x \rangle \ \forall z \in \mathbb{R}^n\} \tag{1.15}$$

The elements of $\partial f(x)$ are called the subgradients of f at x.

Theorem 1.5.1. *A vector x^* is a minimizer of a function $f : \mathbb{R}^n \to \mathbb{R}$ if and only if*

$$0 \in \partial f(x^*) \tag{1.16}$$

Proof. First, we prove the "if" part. Since $0 \in \partial f(x^*)$, $f(z) \geq f(x^*)$ $\forall z \in \mathbb{R}^n$. Hence x^* is a minimizer of f. Next, we prove the "only if" part. Since x^* is a minimizer of f, $f(z) \geq f(x^*)$ $\forall z \in \mathbb{R}^n$. 0 belongs to $\partial f(x^*)$ because $f(z) \geq f(x^*) + \langle 0, z - x^* \rangle$ $\forall z \in \mathbb{R}^n$ □

Theorem 1.5.2. *Assume $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$. A sufficient condition for x^* to be a minimizer of $f_1 + f_2$ is*

$$\exists v^* \in \partial f_1(x^*) \text{ and } -v^* \in \partial f_2(x^*) \tag{1.17}$$

Proof. If $v^* \in \partial f_1(x^*)$, $f_1(z) \geq f_1(x^*) + \langle v^*, z - x^* \rangle \ \forall z \in \mathbb{R}^n$
If $-v^* \in \partial f_2(x^*)$, $f_2(z) \geq f_2(x^*) + \langle -v^*, z - x^* \rangle \ \forall z \in \mathbb{R}^n$
Hence, $f_1(z) + f_2(z) \geq f_1(x^*) + f_2(x^*) \ \forall z \in \mathbb{R}^n$ □

Theorem 1.5.3. *Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and any $x, x', v, v' \in \mathbb{R}^n$ such that $v \in \partial f(x)$ and $v' \in \partial f(x')$, we have $\langle x - x', v - v' \rangle \geq 0$*

Proof. $\because v \in \partial f(x) \therefore f(x') \geq f(x) + \langle v, x' - x \rangle$
$\because v' \in \partial f(x') \therefore f(x) \geq f(x') + \langle v', x - x' \rangle$
$\Rightarrow f(x') + f(x) \geq f(x) + f(x') + \langle v - v', x' - x \rangle$
$\Rightarrow \langle x - x', v - v' \rangle \geq 0$ □

Theorem 1.5.4. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are convex functions and g is differentiable. Then $\forall x \in \mathbb{R}^n$*

$$\partial(f + g)(x) = \partial f(x) + \nabla g(x) = \{y + \nabla g(x) | y \in \partial f(x)\} \tag{1.18}$$

Proof. See theorem 23.8 and 25.1 of [34]. □

This theorem conforms to the intuition that adding a differentiable convex function g to the convex function f simply translates the set of subgradients at each point x by $\nabla g(x)$

Theorem 1.5.5. *Given any function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $z \in \mathbb{R}^n$, and a scalar $c > 0$, there is at most one way to write $z = x + cv$, where $v \in \partial f(x)$. Furthermore, if f is proper, closed and convex[2], then there is exactly one way to write $z = x + cv$, where $v \in \partial f(x)$*

Proof. See lemma 3 and proposition 6 of [10]. □

[2] A function is called proper if its domain is non-empty and its range does not include $-\infty$. A function is called closed if its epigraph is a closed set.

1.6 Convex Conjugate

Definition 1.6.1 (convex conjugate)**.** Given a function $F : \mathbb{R}^n \to (-\infty, \infty]$, the convex conjugate function of F is the function F^* defined as follows

$$F^*(y) := \sup_{x \in \mathbb{R}^n} \{\langle x, y\rangle - F(x)\} \tag{1.19}$$

Assume $\theta \in [0,1], y_1 \in \mathbb{R}^n$ and $y_2 \in \mathbb{R}^n$

$$\begin{aligned}
&F^*(\theta y_1 + (1-\theta)y_2) \\
=&\sup_{x \in \mathbb{R}^n} \{\langle x, \theta y_1 + (1-\theta)y_2\rangle - F(x)\} \\
=&\sup_{x \in \mathbb{R}^n} \{[\langle x, \theta y_1\rangle - \theta F(x)] + [\langle x, (1-\theta)y_2\rangle - (1-\theta)F(x)]\} \\
\leq&\sup_{x \in \mathbb{R}^n} \{\langle x, \theta y_1\rangle - \theta F(x)\} + \sup_{x \in \mathbb{R}^n} \{\langle x, (1-\theta)y_2\rangle - (1-\theta)F(x)\} \\
=&\theta F^*(y_1) + (1-\theta)F^*(y_2)
\end{aligned}$$

Hence, although F may not be convex, F^* is always a convex function.

1.7 Proximal Mapping

Definition 1.7.1 (proximal mapping)**.** The proximal mapping associated with a function $F : \mathbb{R}^n \to (-\infty, \infty]$ is defined as

$$P_F(z) := \underset{x \in \mathbb{R}^n}{argin}f\ F(x) + \frac{1}{2}\|x - z\|_2^2 \tag{1.20}$$

We will illustrate in the following some examples which will be quite helpful in the derivations of some optimization algorithms .

1. $F(x) = \frac{1}{2}\|x\|_2^2$, then $P_F(z) = \frac{1}{2}z$
2. $F(x)$ is the characteristic function of a closed convex set $\mathcal{X} \subseteq \mathbb{R}^D$, then
$$P_F(z) = proj_{\mathcal{X}}(z) := \underset{x \in \mathcal{X}}{argin}f\ \frac{1}{2}\|x - z\|_2^2 \tag{1.21}$$

3. $F(x) = \lambda|x|$, then

$$P_F(z) = S_\lambda(z) = \begin{cases} 0 & |z| < \lambda \\ z - sgn(z)\lambda & |z| \geq \lambda \end{cases} \tag{1.22}$$

where $S_\lambda(z)$ is called the soft thresholding operator

4. $F(x) = \lambda\|x\|_1$, then $P_F(z)$ is the entry-wise soft thresholding operator operating on each entry of z

Assume $x = P_F(z)$, then

$$\begin{aligned} 0 &\in \partial F(x) + x - z \\ \Rightarrow z &\in x + \partial F(x) \end{aligned} \tag{1.23}$$

Hence, $z \in (I_d + \partial F)(x)$, which means

$$P_F(\cdot) = (I_d + \partial F)^{-1}(\cdot) \tag{1.24}$$

Besides, in the following chapters, we will adopt the notation $P_F(\sigma, \cdot)$ to denote $P_{\sigma F}(\cdot)$, i.e.,

$$P_F(\sigma, \cdot) := P_{\sigma F}(\cdot) \tag{1.25}$$

1.8 Moreau Envelope

Definition 1.8.1 (Moreau envelope)**.** Let $g : \mathbb{R}^D \to (-\infty, \infty]$ be a proper closed convex function. The Moreau envelope of g is defined as

$$g_\eta(x) := \inf_{y \in \mathbb{R}^D} g(y) + \frac{1}{2\eta}\|y - x\|_2^2 \tag{1.26}$$

As a comparison, the proximal mapping associated with the function ηg is $P_{\eta g}(x) = \underset{y \in \mathbb{R}^D}{arginf}\, \eta g(y) + \frac{1}{2}\|y - x\|_2^2$. The Moreau envelope can be proved to be convex, differentiable and $1/\eta$-smooth. Furthermore,

$$\nabla g_\eta(x) = \frac{1}{\eta}(x - P_{\eta g}(x)) = \frac{1}{\eta}(I_d - (I_d + \eta\partial g)^{-1})(x) \tag{1.27}$$

Intuitively, we can approximate the operator $(I_d + \eta\partial g)^{-1}$ by $I_d - \eta\partial g$. Hence, $\frac{1}{\eta}(I_d - (I_d + \eta\partial g)^{-1})(x) \approx \frac{1}{\eta}(I_d - (I_d - \eta\partial g))(x) = \partial g(x)$. Due

to this intuition, the operator $\frac{1}{\eta}(I_d - (I_d + \eta\partial g)^{-1})$ is called the Yosida approximation of ∂g. The reader can refer to [25] and [42] for detailed and thorough introductions.

1.9 Non-expansive Mapping

Definition 1.9.1 (non-expansive mapping)**.** A non-expansive mapping $N : \mathbb{R}^n \to \mathbb{R}^n$ is a mapping that satisfies

$$\|N(x) - N(x')\| \leq \|x - x'\| \quad \forall x, x' \in \mathbb{R}^n \tag{1.28}$$

Theorem 1.9.1. *Given two closed, proper and convex functions $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$ and a scalar $c > 0$, the mapping $N_{cf_1} : \mathbb{R}^n \to \mathbb{R}^n$ is given by $N_{cf_1}(z) = x_1 - cv_1$ where $x_1, v_1 \in \mathbb{R}^n$ satisfy $v_1 \in \partial f_1(x_1)$ and $z = x_1 + cv_1$; the mapping $N_{cf_2} : \mathbb{R}^n \to \mathbb{R}^n$ is given by $N_{cf_2}(z) = x_2 - cv_2$ where $x_2, v_2 \in \mathbb{R}^n$ satisfy $v_2 \in \partial f_2(x_2)$ and $z = x_2 + cv_2$*

1. *N_{cf_1} and N_{cf_2} are both everywhere uniquely defined and non-expansive*
2. *The fixed points of N_{cf_1} are precisely the minimizers of f_1*
3. *The fixed points of N_{cf_2} are precisely the minimizers of f_2*
4. *The composite $N_{cf_1} \circ N_{cf_2}$ is also non-expansive*
5. *The fixed points of $N_{cf_1} \circ N_{cf_2}$ are the set of points $\{x + cv | v \in \partial f_1(x), -v \in \partial f_2(x)\}$*

Proof. For the first, second and third results, see proposition 7 of [10] for the proof. For the fourth result, it can be verified as follows

$$\begin{aligned}\|N_{cf_1}(N_{cf_2}(x)) - N_{cf_1}(N_{cf_2}(x'))\| &\leq \|N_{cf_2}(x) - N_{cf_2}(x')\| \\ &\leq \|x - x'\| \ \forall x, x' \in \mathbb{R}^n\end{aligned}$$

assuming the correctness of the first result. For the fifth result, see lemma 14 of [10] for the proof. □

Hence, according to the fifth result, finding a fixed point of $N_{cf_1} \circ N_{cf_2}$ is equivalent to finding $x, v \in \mathbb{R}^n$ satisfying $v \in \partial f_1(x)$ and $-v \in \partial f_2(x)$. Such x is a minimizer of $f_1 + f_2$ as have been proved in theorem 1.5.2

Theorem 1.9.2. *Let the sequence* $\{\rho_k\}$ *satisfy* $\inf_k\{\rho_k\} > 0$ *and* $\sup_k\{\rho_k\} < 2$. *Starting from some arbitrary* $x^{(0)} \in \mathbb{R}^n$, *if* N *has any fixed points and* $\{x^{(k)}\}$ *follows the iteration rule that*

$$x^{(k+1)} = \frac{\rho_k}{2} N(x^{(k)}) + (1 - \frac{\rho_k}{2}) x^{(k)} \tag{1.29}$$

, then $\{x^{(k)}\}$ *converges to a fixed point of* N

Proof. See theorem 10 of [10] for details □

Note that if the Lipschitz modulus of N is less than 1, then the iterate sequence $\{x^{(k)}\}$ can converge to a fixed point by simply iterating the non-expansive mapping recursively with the iteration rule $x^{(k+1)} = N(x^{(k)})$. However, if the Lipschitz modulus happens to be 1, then the iterates could simply orbit at fixed distance from each other without converging. Theorem 1.9.2 states how we can avoid such situation.

Chapter 2

Fundamental Knowledge of Linear Algebra

2.1 Vector and Matrix Norms

2.1.1 Metric, Norm and ℓ_p-Norm

Definition 2.1.1 (metric)**.** Let X be a set. A function $d : X \times X \to [0, \infty)$ is called a metric if

1. $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)\ \forall x, y \in X$
3. $d(x, z) \leq d(x, y) + d(y, z)\ \forall x, y, z \in X$

If only the second and the third conditions hold, then d is called a pseudometric. The set X endowed with a metric d is called a metric space.

Definition 2.1.2 (norm)**.** Let X be a set. A function $\|\cdot\| : X \to [0, \infty)$ is called a norm if

1. $\|x\| = 0$ if and only if $x = 0$ (definiteness)
2. $\|\lambda x\| = |\lambda|\|x\|$ for all scalars λ and all vectors $x \in X$ (homogeneity)
3. $\|x + y\| \leq \|x\| + \|y\|\ \forall x, y \in X$ (triangle inequality)

If the definiteness condition does not hold, then $\|\cdot\|$ is called a seminorm.
If the triangle inequality does not hold but is replaced by the weaker quasitriangle inequality

$$\|x + y\| \leq C(\|x\| + \|y\|) \tag{2.1}$$

for some constant $C \geq 1$, then $\|\cdot\|$ is called a quasinorm. The smallest constant C is called its quasinorm constant.
The set X endowed with a norm $\|\cdot\|$ is called a normed space.
A norm $\|\cdot\|$ on X induces a metric on X by $d(x, y) = \|x - y\|$ and a seminorm induces a pseudometric in the same way.

Definition 2.1.3 (ℓ_p-norm)**.** The ℓ_p norm (or p-norm) on $\mathbb{R}^n$ is defined for $1 \leq p < \infty$ as

$$\|x\|_p := \left(\sum_{j=1}^{n} |x[j]|^p \right)^{1/p} \tag{2.2}$$

and for $p = \infty$ as

$$\|x\|_\infty := \max_{j \in [n]} |x[j]| \tag{2.3}$$

For $0 < p < 1$, the expression 2.2 only defines a quasinorm with the quasinorm constant $C = 2^{1/p-1}$. This can be proved via the p-triangle inequality

$$\|x + y\|_p^p \leq \|x\|_p^p + \|y\|_p^p \tag{2.4}$$

Hence, the ℓ_p-quasinorm induces a metric via $d(x, y) = \|x - y\|_p^p$ for $0 < p < 1$.

We want to verify that the ℓ_p norm 2.2 for $p \geq 1$ is indeed a norm function. The definiteness condition and the homogeneity condition are trivial. Our main concern is whether the ℓ_p-norm satisfies the triangle inequality; that is, whether the following inequality is true or not.

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p \quad for\ p \geq 1 \tag{2.5}$$

Such inequality is called the Minkowski's inequality. To prove the Minkowski's inequality, we need to introduce another important inequality - Hölder's inequality in advance. The Hölder's inequality is expressed as follows.

$$|\langle x, y \rangle| \leq \|x\|_p \|y\|_q \quad \forall x, y \tag{2.6}$$

for $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$. The proof for the Hölder's inequality relies on a simple generalized form of the arithmetic-geometric mean inequality

$$A^{\theta} B^{1-\theta} \leq \theta A + (1-\theta) B \quad if\ A, B \geq 0, and\ 0 \leq \theta \leq 1 \tag{2.7}$$

Proof. Assume B $\neq 0$, and replace A by AB, we see that it suffices to prove that $A^{\theta} \leq \theta A + (1-\theta)$. Let $f(x) = x^{\theta} - \theta x - (1-\theta)$. It is simple to verify that $f(x)$ attains a maximum at $x = 1$, where $f(1) = 0$. Hence, $f(A) \leq 0$. □

With the generalized arithmetic-geometric mean inequality at hand, we can prove the Hölder's inequality as follows.

Proof. If either $\|x\|_p$ or $\|y\|_q$ is 0, then $|\langle x, y \rangle| = 0$, which obviously verifies the Hölder's inequality. Hence, we assume that neither of these norms vanish, and after replacing x by $x/\|x\|_p$ and y by $y/\|y\|_q$, we further assume that $\|x\|_p$=1 and $\|y\|_q$=1. Now it suffices to prove that $|\langle x, y \rangle| \leq 1$. We apply the generalized arithmetic-geometric mean inequality by setting A to be $(x[j])^p$, B to be $(y[j])^q$, and θ to be $1/p$ $\forall j \in [n]$. Therefore,

$$\begin{aligned}
x[j]y[j] &\leq \frac{1}{p}(x[j])^p + \frac{1}{q}(y[j])^q \\
\Rightarrow |x[j]y[j]| &\leq |\frac{1}{p}(x[j])^p + \frac{1}{q}(y[j])^q| \leq \frac{1}{p}|x[j]|^p + \frac{1}{q}|y[j]|^q \\
|\langle x, y \rangle| &= |\sum_{j=1}^{n} x[j]y[j]| \\
&\leq \sum_{j=1}^{n} |x[j]y[j]| \\
&\leq \sum_{j=1}^{n} \left[\frac{1}{p}|x[j]|^p + \frac{1}{q}|y[j]|^q \right] \\
&= \frac{1}{p}\|x\|_p^p + \frac{1}{q}\|y\|_q^q = \frac{1}{p} + \frac{1}{q} = 1
\end{aligned}$$

□

We make a remark that a useful inequality relating the ℓ_p-norm and ℓ_q-norm of an n-dimensional vector x :

$$\|x\|_p \leq n^{1/p-1/q}\|x\|_q \tag{2.8}$$

can be derived using the Hölder's inequality.

Proof. We replace the role of x, y, p, and q by $[|x[1]|^p, |x[2]|^p, \cdots, |x[n]|^p]^T$, $[1, 1, \cdots, 1]^T$, q/p and $q/(q-p)$ respectively. Then we can come up with the inequality : $\|x\|_p^p \leq n^{(q-p)/q}\|x\|_q^p$, which is equivalent to 2.8. □

Specifically, we can get $\|x\|_1 \leq \sqrt{n}\|x\|_2 \leq n\|x\|_\infty$ with 2.8. Let's get down to business to prove the Minkowski's inequality as follows.

Proof. For p=1,

$$\begin{aligned}
&|x[j] + y[j]| \leq |x[j]| + |y[j]| \quad \forall j \in [n] \\
\Rightarrow &\sum_{j=1}^{n} |x[j] + y[j]| \leq \sum_{j=1}^{n} |x[j]| + \sum_{j=1}^{n} |y[j]| \\
\Rightarrow &\|x + y\|_1 \leq \|x\|_1 + \|y\|_1
\end{aligned}$$

When p>1,

$$\begin{aligned}
|x[j] + y[j]|^p &= |x[j] + y[j]||x[j] + y[j]|^{p-1} \quad \forall j \in [n] \\
&\leq (|x[j]| + |y[j]|)|x[j] + y[j]|^{p-1} \\
&= |x[j]||x[j] + y[j]|^{p-1} + |y[j]||x[j] + y[j]|^{p-1} \\
\Rightarrow \sum_{j=1}^{n} |x[j] + y[j]|^p &\leq \sum_{j=1}^{n} |x[j]||x[j] + y[j]|^{p-1} + \sum_{j=1}^{n} |y[j]||x[j] + y[j]|^{p-1} \\
&\leq (\|x\|_p + \|y\|_p) \left(\sum_{j=1}^{n} |x[j] + y[j]|^{q(p-1)} \right)^{1/q} \\
&\quad (\textit{by the Hölder's inequality})
\end{aligned}$$

Hence, $\|x + y\|_p^p \leq (\|x\|_p + \|y\|_p)\|x + y\|_p^{p/q} \because \frac{1}{p} + \frac{1}{q} = 1$
$\Rightarrow \|x + y\|_p \leq \|x\|_p + \|y\|_p$ □

Now we turn our attention to the case when $0 < p < 1$. First, we give a proof for the p-triangle inequality as follows.

Proof.

$$\|x+y\|_p^p = \sum_{j=1}^{n} |x[j]+y[j]|^p, \|x\|_p^p = \sum_{j=1}^{n} |x[j]|^p, \|y\|_p^p = \sum_{j=1}^{n} |y[j]|^p$$

It suffices to prove that $|x[j]+y[j]|^p \leq |x[j]|^p + |y[j]|^p$ for $\forall j \in [n]$. We prove by contradiction. If $|x[j]+y[j]|^p > |x[j]|^p + |y[j]|^p$, then it implies that $|x[j]+y[j]|^{1-p} > |x[j]|^{1-p} + |y[j]|^{1-p}$ since $0 < 1-p < 1$. In this way,

$$\begin{aligned} |x[j]+y[j]| &> (|x[j]|^p + |y[j]|^p)(|x[j]|^{1-p} + |y[j]|^{1-p}) \\ &= |x[j]| + |y[j]| + |x[j]|^p |y[j]|^{1-p} + |y[j]|^p |x[j]|^{1-p} \\ &\geq |x[j]| + |y[j]| \end{aligned}$$

which contradicts with the triangle inequality for ℓ_1 norm. Hence, $|x[j]+y[j]|^p \leq |x[j]|^p + |y[j]|^p \ \forall j \in [n]$ for $0 < p < 1$. □

Then, we can use the p-triangle inequality to prove that the quasinorm constant for the ℓ_p-quasinorm is indeed $2^{1/p-1}$ as follows

Proof.

$$\begin{aligned} \|x+y\|_p &= \left(\sum_{j=1}^{n} |x[j]+y[j]|^p \right)^{1/p} \\ &\leq \left(\sum_{j=1}^{n} |x[j]|^p + \sum_{j=1}^{n} |y[j]|^p \right)^{1/p} \\ &= 2^{1/p} \left(\frac{1}{2} \sum_{j=1}^{n} |x[j]|^p + \frac{1}{2} \sum_{j=1}^{n} |y[j]|^p \right)^{1/p} \end{aligned}$$

$$\leq 2^{1/p}\left[\frac{1}{2}\left(\sum_{j=1}^{n}|x[j]|^p\right)^{1/p}+\frac{1}{2}\left(\sum_{j=1}^{n}|y[j]|^p\right)^{1/p}\right]$$
$$= 2^{1/p-1}\left(\|x\|_p+\|y\|_p\right)$$

□

The first inequality is due to the p-triangle inequality and the monotonicity of the $\ell_{1/p}$-norm. The second inequality is due to the convexity of the $\ell_{1/p}$-norm and the application of the Jenson's inequality.

2.1.2 Dual Norm

Definition 2.1.4 (dual norm)**.** Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Its dual norm $\|\cdot\|_*$ is defined by

$$\begin{aligned}\|x\|_* :&= \sup_{\|y\|\leq 1}|\langle y,x\rangle| \quad ,x\in\mathbb{R}^n\\ &= \sup_{y\neq 0}\frac{|\langle y,x\rangle|}{\|y\|}\end{aligned} \tag{2.9}$$

From the definition of the dual norm, we can easily derive the useful inequality

$$|\langle y,x\rangle|\leq\|y\|\|x\|_*, \forall x,y\in\mathbb{R}^n \tag{2.10}$$

Furthermore, we want to introduce two important properties of the dual norm. The first one is that the dual of the dual norm is the norm itself and the second one is that the dual of the ℓ_p-norm is the ℓ_q-norm with $\frac{1}{p}+\frac{1}{q}=1$. From the second one, we find that the ℓ_2-norm is self-dual and the dual norm of the ℓ_1-norm is the ℓ_∞-norm. In the following, we give proofs for the two properties respectively.

Proof for the first property. Consider the problem :

$$\min_y \|y\| \quad \textit{subject to } y=x$$

Its optimal value is obviously $\|x\|$. Its Lagrangian is

$$L(y, \nu) = \|y\| + \nu^T(x - y)$$

and the Lagrange dual function is

$$\begin{aligned} g(\nu) &= \min_y \|y\| + \nu^T x - \nu^T y \\ &= \nu^T x - \max_y \nu^T y - \|y\| \end{aligned}$$

Assume $z_* = \underset{\|z\| \leq 1}{argmax}\, \nu^T z$, then $\|\nu\|_* = \nu^T z_*$. If $\|\nu\|_* > 1$, we can select $y = tz_*$, $t > 0$ so that $\nu^T y - \|y\| = t(\|\nu\|_* - \|z_*\|)$ approaches ∞ as t approaches ∞. If $\|\nu\|_* \leq 1$, then $\nu^T y \leq \|y\|\ \forall y \in \mathbb{R}^n$; that is, $\nu^T y - \|y\| \leq 0\ \forall y \in \mathbb{R}^n$ and we can obtain the maximum 0 when $y = 0$. Therefore, the Lagrange dual function is

$$g(\nu) = \begin{cases} \nu^T x & , when\ \|\nu\|_* \leq 1 \\ -\infty & , when\ \|\nu\|_* > 1 \end{cases}$$

The dual problem will be

$$\max_\nu \nu^T x \quad subject\ to\ \|\nu\|_* \leq 1$$

Its optimal value is $\|x\|_{**}$. By the strong duality, we have $\|x\| = \|x\|_{**}$. □

Note that this proof involves the application of the duality theory. The readers can refer to section 3.1.1.

Proof for the second property. We want to show that $\|z\|_p = \sup_{\|x\|_q \leq 1} \langle x, z \rangle$. We may assume without generality that $z \neq 0$; otherwise, $\|z\|_p = \sup_{\|x\|_q \leq 1} \langle x, z \rangle$ is trivially true. Let $x \in \mathbb{R}^n$ satisfy $\|x\|_q \leq 1$.

$$\langle x, z \rangle \leq |\langle x, z \rangle| \leq \|x\|_q \|z\|_p \leq \|z\|_p$$

by the Hölder's inequality. Hence, $\sup_{\|x\|_q \leq 1} \langle x, z \rangle \leq \|z\|_p$. In order to show that $\sup_{\|x\|_q \leq 1} \langle x, z \rangle$ is exactly $\|z\|_p$, it suffices to find an $y \in \mathbb{R}^n$

with $\|y\|_q \leq 1$ such that $\langle y, z\rangle = \|z\|_p$. Let $x \in \mathbb{R}^n$ be a vector with each component $x[j] = sgn(z[j])|z[j]|^{p-1}$.

$$\langle x, z\rangle = \sum_{j=1}^{n} |z[j]|^p = \|z\|_p^p$$

and

$$\|x\|_q^q = \sum_{j=1}^{n} |x[j]|^q = \sum_{j=1}^{n} |z[j]|^{(p-1)q} = \sum_{j=1}^{n} |z[j]|^p = \|z\|_p^p$$

Now choose $y = \frac{x}{\|x\|_q}$ ($\because z \neq 0 \therefore \|x\|_q = \|z\|_p^{p/q} \neq 0$).

$$\langle y, z\rangle = \langle \frac{x}{\|x\|_q}, z\rangle = \frac{\|z\|_p^p}{\|x\|_q} = \frac{\|z\|_p}{\|z\|_p^{p/q}} = \|z\|_p^{p-p/q} = \|z\|_p$$

Hence, we successfully find an $y \in \mathbb{R}^n$ with $\|y\|_q \leq 1$ such that $\langle y, z\rangle = \|z\|_p$. □

2.1.3 Operator Norm

Definition 2.1.5 (operator norm)**.** Let $A : X \to Y$ be a linear mapping between two normed spaces $(X, \|\cdot\|)$ and $(Y, |||\cdot|||)$. The operator norm of A is defined as

$$\|A\| := \sup_{\|x\|\leq 1} |||Ax||| = \sup_{x\neq 0} \frac{|||Ax|||}{\|x\|} \tag{2.11}$$

In particular, let $A \in \mathbb{R}^{m\times n}$, X be the ℓ_p space and Y be the ℓ_q space, $1 < p, q \leq \infty$. We define the matrix norm (operator norm) between ℓ_p and ℓ_q as

$$\|A\|_{p\to q} := \sup_{\|x\|_p\leq 1} \|Ax\|_q = \sup_{x\neq 0} \frac{\|Ax\|_q}{\|x\|_p} \tag{2.12}$$

From 2.11, we can easily derive the inequality

$$|||Ax||| \leq \|A\|\|x\| \quad \forall x \in X \tag{2.13}$$

From 2.12, we easily derive the inequality

$$\begin{aligned} \|AB\|_{p\to r} &= \sup_{\|x\|_p\leq 1} \|ABx\|_r \\ &\leq \sup_{\|x\|_p\leq 1} \|A\|_{q\to r}\|Bx\|_q \\ &= \|A\|_{q\to r}\|B\|_{p\to q} \end{aligned} \tag{2.14}$$

where $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times k}$ and $1 \leq p, q, r \leq \infty$. Combining 2.12, the definition of dual norm and the duality between ℓ_p and $\ell_{p'}$ norms (ℓ_q and $\ell_{q'}$ norms) where $\frac{1}{p} + \frac{1}{p'} = \frac{1}{q} + \frac{1}{q'} = 1$, we can derive a useful inequality as follows

$$\|A\|_{p\to q} = \|A^*\|_{q'\to p'} \tag{2.15}$$

We verify it as follows

$$\begin{aligned} \|A\|_{p\to q} &= \sup_{\|x\|_p\leq 1} \|Ax\|_q \\ &= \sup_{\|x\|_p\leq 1} \sup_{\|y\|_{q'}\leq 1} \langle y, Ax\rangle \\ &= \sup_{\|x\|_p\leq 1} \sup_{\|y\|_{q'}\leq 1} \langle x, A^*y\rangle \\ &= \sup_{\|y\|_{q'}\leq 1} \|A^*y\|_{p'} \\ &= \|A^*\|_{q'\to p'} \end{aligned}$$

In the following, we excerpt some important results regarding the matrix norm from lemma A.5, lemma A.7, lemma A.9 and remark A.10 of [13].

Theorem 2.1.1. *Let $A \in \mathbb{R}^{m\times n}$.*

1.

$$\|A\|_{2\to 2} = \sqrt{\lambda_{max}(A^*A)} = \sigma_{max}(A) \tag{2.16}$$

where $\lambda_{max}(A^*A)$ denotes the largest eigenvalue of A^*A and $\sigma_{max}(A)$ the largest singular value of A. In particular, if $B \in$

$\mathbb{R}^{n\times n}$ *is self-adjoint, then*

$$\|B\|_{2\to 2} = \max_{j\in[n]} |\lambda_j(B)| \tag{2.17}$$

where $\lambda_j(B)$, $j \in [n]$ *denotes the eigenvalues of* B.

2. For $1 \le p \le \infty$,

$$\|A\|_{1\to p} = \max_{k\in[n]} \|a_k\|_p \tag{2.18}$$

where a_k *represents the k-th columns of* A. *In particular,*

$$\|A\|_{1\to 1} = \max_{k\in[n]} \sum_{j=1}^{m} |A_{j,k}| \tag{2.19}$$

$$\|A\|_{1\to 2} = \max_{k\in[n]} \|a_k\|_2 \tag{2.20}$$

3.

$$\|A\|_{\infty\to\infty} = \max_{j\in[m]} \sum_{k=1}^{n} |A_{j,k}| \tag{2.21}$$

Theorem 2.1.2. *Let* $A \in \mathbb{R}^{m\times n}$.

$$\|A\|_{2\to 2} = \sup_{\|y\|_2\le 1} \sup_{\|x\|_2\le 1} |\langle Ax, y\rangle| \tag{2.22}$$

If $B \in \mathbb{R}^{n\times n}$ *is self-adjoint, then*

$$\|B\|_{2\to 2} = \sup_{\|x\|_2\le 1} |\langle Bx, x\rangle| \tag{2.23}$$

Theorem 2.1.3. *The operator norm* $\|\cdot\|_{p\to q}$ $(1 \le p, q \le \infty)$ *of a submatrix is bounded by the whole matrix. More precisely, if* $A \in \mathbb{R}^{m\times n}$ *has the form*

$$A = \begin{bmatrix} A^{(1)} & A^{(2)} \\ A^{(3)} & A^{(4)} \end{bmatrix}$$

then $\|A^{(\ell)}\|_{p\to q} \le \|A\|_{p\to q}$ *for* $\ell = 1, 2, 3, 4$. *In particular, any entry of* A *satisfies* $|A_{j,k}| \le \|A\|_{p\to q}$.

Definition 2.1.6 (Frobenius norm)**.** The Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as

$$\|A\|_F := \sqrt{tr(AA^*)} = \sqrt{tr(A^*A)} = \left(\sum_{j\in[m],k\in[n]} |A_{j,k}|^2\right)^{1/2} \tag{2.24}$$

The Frobenius norm can give an upper bound on the operator norm $\|\cdot\|_{2\to 2}$; that is,

$$\|A\|_{2\to 2} \le \|A\|_F \tag{2.25}$$

Proof.

$$\begin{aligned}
\|Ax\|_2^2 &= \sum_{j=1}^{m}\left(\sum_{k=1}^{n} A_{j,k}x[k]\right)^2 \\
&\le \sum_{j=1}^{m}\left(\sum_{k=1}^{n}|x[k]|^2\right)\left(\sum_{\ell=1}^{n}|A_{j,\ell}|^2\right) \\
&\qquad (\textit{by the Hölder's inequality}) \\
&= \|A\|_F^2\|x\|_2^2
\end{aligned}$$

Therefore, $\forall x \in \mathbb{R}^n\backslash\{0\}$, $\|A\|_F \ge \frac{\|Ax\|_2}{\|x\|_2}$; that is, $\|A\|_F \ge \sup_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2} = \|A\|_{2\to 2}$ □

2.2 The Singular Values and Eigenvalues

2.2.1 The Singular Value Decomposition

Theorem 2.2.1 (the singular value decomposition)**.** *For $A \in \mathbb{R}^{m\times n}$, there exist unitary matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, and uniquely defined non-negative numbers $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{\min\{m,n\}} \ge 0$, called singular values of A, such that*

$$A = U\Sigma V^* \quad \textit{where } \Sigma = diag[\sigma_1, \cdots, \sigma_{\min\{m,n\}}] \in \mathbb{R}^{m\times n} \tag{2.26}$$

The column vectors of U are called the left singular vectors while those of V are called right singular vectors.

The readers can refer to proposition A.13 of [13] for the proof of this theorem. The equation 2.26 is famous for being the singular value decomposition of the matrix A. Assume A has r positive singular values. Sometimes, it is more convenient to work with the reduced singular value decomposition.

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^* = \sum_{j=1}^{r} \sigma_j u_j v_j^* \tag{2.27}$$

where $\tilde{\Sigma} = diag(\sigma_1, \cdots, \sigma_r) \in \mathbb{R}^{r\times r}$, $\tilde{U} = [u_1|\cdots|u_r] \in \mathbb{R}^{m\times r}$ and $\tilde{V} = [v_1|\cdots|v_r] \in \mathbb{R}^{n\times r}$ are submatrices of $U = [u_1|\cdots|u_m] \in \mathbb{R}^{m\times m}$ and $V = [v_1|\cdots|v_n] \in \mathbb{R}^{n\times n}$. We want to make an important connection between the singular value decomposition of a matrix A and its four fundamental subspaces, i.e., the row space, the null space, the column space and the left null space. The row space of A is the set of all $x \in \mathbb{R}^n$ such that Ax is nonzero. The null space of A is the set of all $x \in \mathbb{R}^n$ such that $Ax = 0$. The column space of A is the set of all $y \in \mathbb{R}^m$ such that A^*y is nonzero. The left null space of A is the set of all $y \in \mathbb{R}^m$ such that $A^*y = 0$. Hence, we see that the row space and null space are orthogonal to each other and together span the whole $\mathbb{R}^n$. The column space and left null space are also orthogonal to each other and together span the whole $\mathbb{R}^m$. From the relations 2.26 and 2.27, we know that

$$\begin{aligned}
Av_i &= \sigma_i u_i \quad , i = 1, 2, \cdots, r \\
Av_i &= 0 \quad , i = r+1, \cdots, n \\
A^*u_j &= \sigma_j v_j \quad , j = 1, 2, \cdots, r \\
A^*u_j &= 0 \quad , j = r+1, \cdots, m
\end{aligned}$$

Hence,

1. $[v_1|v_2|\cdots|v_r] \in \mathbb{R}^{n\times r}$ forms an orthonormal basis of the row space.
2. $[v_{r+1}|\cdots|v_n] \in \mathbb{R}^{n\times(n-r)}$ forms an orthonormal basis of the null

space.

3. $[u_1|u_2|\cdots|u_r] \in \mathbb{R}^{m\times r}$ forms an orthonormal basis of the column space.
4. $[u_{r+1}|\cdots|u_m] \in \mathbb{R}^{m\times(m-r)}$ forms an orthonormal basis of the left null space.

In the following, we list some important properties about the singular values.

1.
$$\sigma_{max}(A) = \|A\|_{2\to 2} = \max_{\|x\|_2=1} \|Ax\|_2 \tag{2.28}$$
$$\sigma_{min}(A) = \min_{\|x\|_2=1} \|Ax\|_2 \tag{2.29}$$

Proof. see proposition A.13 of [13]. □

2.
$$\sigma_j(A) = \sqrt{\lambda_j(A^*A)} = \sqrt{\lambda_j(AA^*)} \quad j \in [\min\{m,n\}] \tag{2.30}$$

Proof.
$$A^*A = V\Sigma^2V^*$$
$$AA^* = U\Sigma^2U^*$$

□

3. If A has rank r, then its r largest singular values $\sigma_1 \geq \cdots \geq \sigma_r$ are positive, while $\sigma_{r+1} = \sigma_{r+2} = \cdots = 0$.

Proof. First we prove that $\text{rank}(A) = \text{rank}(A^*A)$ as follows. If $x \in \mathbb{R}^n$ lies in the null space of A (i.e., $Ax = 0$), then it is also in the null space of A^*A since $A^*Ax = 0$. On the other hand, if $x \in \mathbb{R}^n$ lies in the null space of A^*A (i.e., $A^*Ax = 0$), then it is also in the null space of A since $x^*A^*Ax = 0$, which implies $\|Ax\|_2^2 = 0$ (hence, $Ax = 0$). Therefore, the nullity of A and A^*A are the same, which implies the rank of A and A^* are also the same. Now since $A^*A = V\Sigma^2V^*$, $\text{rank}(A^*A) = \text{rank}(V\Sigma^2V^*) =$

$\text{rank}(\Sigma^2)$. Since $\text{rank}(A) = r$, there are exactly r positive singular values $\sigma_1, \cdots, \sigma_r$. □

4. The largest and smallest singular values are 1-Lipschitz functions with respect to the operator norm and the Frobenius norm. That is,

$$|\sigma_{max}(A) - \sigma_{max}(B)| \leq \|A - B\|_{2\to 2} \leq \|A - B\|_F \tag{2.31}$$
$$|\sigma_{min}(A) - \sigma_{min}(B)| \leq \|A - B\|_{2\to 2} \leq \|A - B\|_F \tag{2.32}$$

for all matrices A and B of equal dimensions.

Proof. see proposition A.16 of [13]. □

5.

$$\|A^*A - Id\|_{2\to 2} \leq \delta \quad for\ some\ \delta \in [0, 1] \tag{2.33}$$

if and only if

$$\sigma_{max}(A) \leq \sqrt{1+\delta} \quad and \quad \sigma_{min} \geq \sqrt{1-\delta} \tag{2.34}$$

Proof.

$$\begin{aligned}
\|A^*A - Id\|_{2\to 2} &= \max_{j\in[n]} |\lambda_j(A^*A - Id)| \\
&= \max_{j\in[n]} |\lambda_j(A^*A) - 1| \\
&= \max_{j\in[n]} |\sigma_j^2(A) - 1| \\
&= \max\{\sigma_{max}^2(A) - 1, 1 - \sigma_{min}^2(A)\}
\end{aligned}$$

If $\|A^*A - Id\|_{2\to 2} \leq \delta$, then

$$\begin{aligned}
\sigma_{max}^2(A) - 1 &\leq \delta \\
1 - \sigma_{min}^2(A) &\leq \delta
\end{aligned}$$

which implies 2.34. If $\sigma_{max}(A) \leq \sqrt{1+\delta}$ and $\sigma_{min} \geq \sqrt{1-\delta}$, then

$$\begin{aligned}
\sigma_{max}^2(A) - 1 &\leq \delta \\
1 - \sigma_{min}^2(A) &\leq \delta
\end{aligned}$$

which implies 2.33. □

6. For two matrices $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times n}$, any $k \in [\ell]$, where $\ell := \min\{m, n\}$,

$$\sum_{j=1}^{k} |\sigma_j(X) - \sigma_j(Y)| \leq \sum_{j=1}^{k} \sigma_j(X - Y) \tag{2.35}$$

Proof. see lemma A.18 of [13]. □

2.2.2 The Schatten p-norm

Definition 2.2.1 (Schatten p-norm)**.** For a matrix $A \in \mathbb{R}^{m\times n}$ and all $1 \leq p \leq \infty$, the Schatten p-norm of A is defined as

$$\|A\|_{S_p} := \left[\sum_{j=1}^{min\{m,n\}} \sigma_j(A)^p\right]^{1/p} \tag{2.36}$$

The Schatten p-norm is indeed a norm function; that is, it satisfies the definiteness condition, the homogeneity condition and the triangle inequality. One can easily verify the definiteness condition and the homogeneity condition. As for the triangle inequality, although we do not give a proof here, we can use the inequality 2.35 to verify the Schatten 1-norm case as follows.

$$\begin{aligned}\sum_{j=1}^{\ell} \sigma_j(X) - \sum_{j=1}^{\ell} \sigma_j(Y) &= \sum_{j=1}^{\ell} |\sigma_j(X)| - |\sigma_j(Y)| \\ &\leq \sum_{j=1}^{\ell} |\sigma_j(X) - \sigma_j(Y)| \\ &\leq \sum_{j=1}^{\ell} \sigma_j(X - Y)\end{aligned}$$

Choose $X = A + B$ and $Y = B$, we can derive the triangle inequality for the Schatten 1-norm. Specifically, we will denote the Schatten 1-

norm as

$$\|A\|_* := \sum_{j=1}^{\min\{m,n\}} \sigma_j(A) \tag{2.37}$$

and call it the nuclear norm. Further note that the Schatten p-norm reduces to the Frobenius norm for $p = 2$ and to the operator norm $\|\cdot\|_{2\to 2}$ for $p = \infty$. The reasons are simple. For $p = 2$

$$\begin{aligned}
\|A\|_{S_2} &= \sqrt{\sum_{j=1}^{\min\{m,n\}} \sigma_j^2(A)} \\
&= \sqrt{\sum_{j=1}^{\min\{m,n\}} \lambda_j(A^*A)} \\
&= \sqrt{tr(A^*A)} \\
&= \|A\|_F
\end{aligned}$$

For $p = \infty$, $\|A\|_{S_\infty} = \max_{j\in[\ell]} \sigma_j(A) = \sigma_{max}(A) = \|A\|_{2\to 2}$.

2.2.3 Moore-Penrose Pseudo-Inverse

Definition 2.2.2 (Moore-Penrose pseudo-inverse)**.** Let $A \in \mathbb{R}^{m\times n}$ of rank r with reduced singular value decomposition

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^* = \sum_{j=1}^{r} \sigma_j(A) u_j v_j^*$$

then its Moore-Penrose pseudo-inverse $A^\dagger \in \mathbb{R}^{n\times m}$ is defined as

$$A^\dagger = \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^* = \sum_{j=1}^{r} \sigma_j^{-1}(A) v_j u_j^* \tag{2.38}$$

We list some important properties of the Moore-Penrose pseudo-inverse (or simply pseudo-inverse).

1. $A^\dagger$ has the same rank r as A

2.

$$\sigma_{max}(A^\dagger) = \|A^\dagger\|_{2\to 2} = \sigma_r^{-1}(A) \tag{2.39}$$

3. If A is an invertible square matrix, then $A^\dagger = A^{-1}$

4. If A^*A is invertible (implying $m \geq n$), then

$$A^\dagger = (A^*A)^{-1}A^* \tag{2.40}$$

If AA^* is invertible (implying $n \geq m$), then

$$A^\dagger = A^*(AA^*)^{-1} \tag{2.41}$$

The first and second property can be easily verified from the definition of the pseudo-inverse. For the third property, we can verify it by checking that $AA^\dagger = A^\dagger A = I_m$ (when $m = n$). As for the fourth property,

$$\begin{aligned}
(A^*A)^{-1}A^* &= (\tilde{V}\tilde{\Sigma}^2\tilde{V}^*)^{-1}\tilde{V}\tilde{\Sigma}\tilde{U}^* \\
&= \tilde{V}\tilde{\Sigma}^{-2}\tilde{V}^*\tilde{V}\tilde{\Sigma}\tilde{U}^* \\
&= \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^* \\
&= A^\dagger \\
A^*(AA^*)^{-1} &= \tilde{V}\tilde{\Sigma}\tilde{U}^*(\tilde{U}\tilde{\Sigma}^2\tilde{U}^*)^{-1} \\
&= \tilde{V}\tilde{\Sigma}\tilde{U}^*\tilde{U}\tilde{\Sigma}^{-2}\tilde{U}^* \\
&= \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^* \\
&= A^\dagger
\end{aligned}$$

2.2.4 Gershgorin's Disk Theorem

Theorem 2.2.2 (Gershgorin's disk theorem)**.** *Let λ be an eigenvalue of a square matrix $A \in \mathbb{R}^{m\times n}$. There exists an index $j \in [n]$ such that*

$$|\lambda - A_{j,j}| \leq \sum_{\ell\in[n]\setminus\{j\}} |A_{j,\ell}| \tag{2.42}$$

The readers can refer to theorem A.11 of [13] for the proof. The

Gershgorin's disk theorem states that for a given eigenvalue λ, there is at least one diagonal entry of A whose distance with λ is within the absolute sum of all the other off-diagonal entries of the same row.

2.3 Least-Squares Problems

In this section, we consider two kinds of least squares problems. The first one has the following objective

$$\underset{x}{minimize} \, \|Ax - y\|_2 \tag{2.43}$$

where $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$ $(m \geq n)$ has full rank n. We introduce two methods to solve it. One method involves projection of y onto the column space of A. Assume the orthogonal projection of y onto the column space of A be Ax for some $x \in \mathbb{R}^n$. The residual vector $y - Ax$ will be orthogonal to the column space of A (hence, to all the n column vectors of A). Therefore, we can write

$$\begin{aligned} A^*(y - Ax) &= 0 \\ \Rightarrow A^*Ax &= A^*y \end{aligned} \tag{2.44}$$

The equation 2.44 is called the normal equation and we can come up with the solution $x = (A^*A)^{-1}A^*y = A^\dagger y$ from it. The other method transforms the original problem to an equivalent quadratic problem. Precisely speaking,

$$\begin{aligned} &\underset{x}{arginf} \, \|Ax - y\|_2 \\ &= \underset{x}{arginf} \, \|Ax - y\|_2^2 \\ &= \underset{x}{arginf} \langle A^*Ax, x\rangle - 2\langle Ax, y\rangle \end{aligned}$$

Take the first derivative of $\langle A^*Ax, x\rangle - 2\langle Ax, y\rangle$ with respect to x and we can exactly get the normal equation 2.44. As a result, we derive that the orthogonal projection of y onto the column space of A is $AA^\dagger y$. We can define the orthogonal projection matrix onto the column space

of A as

$$P_A := AA^{\dagger} \tag{2.45}$$

If we compute the SVD of A as $U\Sigma V^*$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are two orthonormal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ has rank r. In section 2.2, we have verified that $C \triangleq [u_1|\cdots|u_r] \in \mathbb{R}^{m\times r}$ forms a basis for the column space of A, $L \triangleq [u_{r+1}|\cdots|u_m] \in \mathbb{R}^{m\times(m-r)}$ forms a basis for the left null space of A, $R \triangleq [v_1|\cdots|v_r] \in \mathbb{R}^{n\times r}$ forms a basis for the row space of A and $N \triangleq [v_{r+1}|\cdots|v_n]\,\mathbb{R}^{n\times(n-r)}$ forms a basis for the null space of A. If we plug the matrices C, L, R and N into 2.45, we can have the resulting important result : CC^*, LL^*, RR^* and NN^* are all orthogonal projection matrices onto the column space, left null space, row space and null space of A, respectively.

The second kind of least-squares problem has the following objective

$$\underset{x}{minimize}\,\|x\|_2 \quad subject\ to\ Ax = y \tag{2.46}$$

where $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$ $(n \geq m)$ has full rank m. We can solve it using the duality theory, which will be introduced in section 3.1.1. Note that we can transform the problem 2.46 to an equivalent quadratic problem

$$\underset{x}{minimize}\,\|x\|_2^2 \quad subject\ to\ Ax = y$$

We can construct the Lagrangian $L(x,\nu) = \|x\|_2^2 + \nu^T(Ax - y)$. The gradient of the Lagrangian at the primal-dual optimal pair (x^*, ν^*) should be zero. That is,

$$2x^* + (A^*\nu^*) = 0$$

Hence, $x^* = -\frac{1}{2}A^*\nu^*$. Plugging it into the equality $Ax^* = y$, we can get $-\frac{1}{2}AA^*\nu^* = y$. Therefore, $\nu^* = -2(AA^*)^{-1}y$ and $x = A^*(AA^*)^{-1}y = A^{\dagger}y$.

2.4 The Generalized Singular Value Decomposition (GSVD)

Assume $P \in \mathbb{R}^{\ell\times\ell}$ is an unitary matrix. We partition P as

$$P = \begin{matrix}\begin{bmatrix} P_1 & P_2 \end{bmatrix} \\ k \quad q\end{matrix} = \begin{matrix}\begin{bmatrix} P_3 \\ P_4 \end{bmatrix} \begin{matrix} m \\ p \end{matrix} \\ \end{matrix} = \begin{matrix}\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{matrix} m \\ p \end{matrix} \\ k \quad q\end{matrix} \quad , \ell = k + q = m + p$$

$P_1^*P_1 = P_{11}^*P_{11} + P_{21}^*P_{21} = I_k$, where $I_k \in \mathbb{R}^{k\times k}$ denotes the identity matrix. Assume $u_i \in \mathbb{R}^k$ is an eigenvalue of $P_{11}^*P_{11}$ corresponding to the eigenvalue λ_i. Therefore,

$$\begin{aligned} & P_{11}^*P_{11}u_i = \lambda_i u_i \\ \Rightarrow & (P_1^*P_1 - P_{21}^*P_{21})u_i = \lambda_i u_i \\ \Rightarrow & P_{21}^*P_{21}u_i = (1-\lambda_i)u_i \end{aligned}$$

Hence, P_{11} and P_{21} have the same right singular vectors. Assume P_{11} has singular values 1 of multiplicity r, α_i $(i = 1, 2, \cdots, s,\ 1 > \alpha_1 \geq \cdots \geq \alpha_s > 0)$. It follows that P_{21} has singular values 1 of multiplicity $k-r-s$, β_i $(i = 1, 2, \cdots, s, 0 < \beta_1 \geq \cdots \geq \beta_s < 1$ and $\alpha_i^2 + \beta_i^2 = 1)$. Assume SVD of P_{11} is $W\Sigma_{11}U^*$, where $W \in \mathbb{R}^{m\times m}$, $U \in \mathbb{R}^{k\times k}$ and

$$\Sigma_{11} = \begin{matrix} \\ m-r-s \end{matrix} \begin{matrix} r \\ s \\ \end{matrix} \begin{bmatrix} I & & \\ & C & \\ & & 0_c \end{bmatrix} \in \mathbb{R}^{m\times k},\ \text{C=diag}(\alpha_1, \cdots, \alpha_s),$$
$$\begin{matrix} \\ r \end{matrix} \begin{matrix} s \end{matrix} \begin{matrix} k-r-s \end{matrix}$$

I denotes the identity matrix and 0_c denotes the matrix with all entries 0. Assume SVD of P_{21} is $Z\Sigma_{21}U^*$, where $Z \in \mathbb{R}^{p\times p}$ and $\Sigma_{21} =$

$$\begin{matrix} p-k+r \\ s \\ k-r-s \end{matrix} \begin{bmatrix} 0_s & & \\ & S & \\ & & I \end{bmatrix} \in \mathbb{R}^{p\times k},\ \text{S=diag}(\beta_1, \cdots, \beta_s) \text{ and } 0_s \text{ also}$$
$$\begin{matrix} r \end{matrix} \begin{matrix} s \end{matrix} \begin{matrix} k-r-s \end{matrix}$$

denotes the matrix with all entries 0.

$P_3P_3^* = P_{11}P_{11}^* + P_{12}P_{12}^* = I_m$, where $I_m \in \mathbb{R}^{m\times m}$ denotes the identity matrix. Similarly, we can derive that P_{11} and P_{12} have the same left singular vectors and assume SVD of P_{12} is $W\Sigma_{12}V^*$, where

$$\Sigma_{12} = \begin{matrix} r \\ s \\ m-r-s \end{matrix}\begin{bmatrix} 0_s^* & & \\ & S & \\ & & I \end{bmatrix} \in \mathbb{R}^{m\times q} \text{ and } V \in \mathbb{R}^{q\times q}.$$
$$\begin{matrix} p-k+r & s & m-r-s \end{matrix}$$

Finally, we can derive that P_{22} and P_{21} have the same left singular vectors and P_{22} and P_{12} have the same right singular vectors. SVD of P_{22} can be expressed as $Z\Sigma_{22}V^*$, where

$$\Sigma_{22} = \begin{matrix} p-k+r \\ s \\ k-r-s \end{matrix}\begin{bmatrix} Q & & \\ & T & \\ & & 0_c^* \end{bmatrix} \in \mathbb{R}^{p\times q},\ Q \text{ may be } \pm I$$
$$\begin{matrix} p-k+r & s & m-r-s \end{matrix}$$

and T may be $\pm C$.

Hence, we can decompose P as $P = D\Sigma E^*$, where

$$D = \begin{matrix} m \\ p \end{matrix}\begin{bmatrix} W & 0 \\ 0 & Z \end{bmatrix} \tag{2.47}$$
$$\begin{matrix} m & p \end{matrix}$$

$$\Sigma = \begin{matrix} r \\ s \\ m-r-s \\ \\ p-k+r \\ s \\ k-r-s \end{matrix}\left[\begin{array}{ccccccc} I & & & \vdots & 0_s^* & & \\ & C & & \vdots & & S & \\ & & 0_c & \vdots & & & I \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0_s & & & \vdots & Q & & \\ & S & & \vdots & & T & \\ & & I & \vdots & & & 0_s^* \end{array}\right]$$
$$\begin{matrix} r & s & k-r-s & & p-k+r & s & m-r-s \end{matrix} \tag{2.48}$$

$$E^* = \begin{matrix} k \\ q \end{matrix}\begin{bmatrix} U^* & 0 \\ 0 & V^* \end{bmatrix} \tag{2.49}$$
$$\begin{matrix} k & q \end{matrix}$$

Since P, W, Z, U, V are all unitary, Σ is also unitary. Therefore, $CS + ST = 0$, which means T should be $-C$. Furthermore, since $P_{12} = W\Sigma_{12}V^*$, we know that $P_{12}v_i = 0$ for $i = 1, 2, \cdots, p-k+r$, where v_i denotes the i-th column of V. Since $P_{22} = Z\Sigma_{22}V^*$, $P_{22}v_i = -z_i$

if $Q = -I$, where z_i denotes the i-th column of Z. If this is the case, we can simply change v_i to $-v_i$ and then we can change Q to I. As a summary, we can decompose each block P_{11}, P_{12}, P_{21} and P_{22} of an unitary matrix P as

1.
$$P_{11} = W\Sigma_{11}U^* = W \begin{bmatrix} I & & \\ & C & \\ & & 0_c \end{bmatrix} U^* \tag{2.50}$$

2.
$$P_{12} = W\Sigma_{12}V^* = W \begin{bmatrix} 0_s^* & & \\ & S & \\ & & I \end{bmatrix} V^* \tag{2.51}$$

3.
$$P_{21} = Z\Sigma_{21}U^* = Z \begin{bmatrix} 0_s & & \\ & S & \\ & & I \end{bmatrix} U^* \tag{2.52}$$

4.
$$P_{22} = Z\Sigma_{22}V^* = Z \begin{bmatrix} I & & \\ & -C & \\ & & 0_c^* \end{bmatrix} V^* \tag{2.53}$$

Given two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$, in the following we will demonstrate how to derive a generalized singular value decomposition (GSVD) of A and B step by step.

1. Form $F = \begin{bmatrix} A \\ B \end{bmatrix} \in \mathbb{R}^{(m+p)\times n}$. Let k denote the rank of F.

2. Compute SVD of F. Assume it is $P \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix} Q^*$, where $P \in \mathbb{R}^{(m+p)\times(m+p)}$ and $Q \in \mathbb{R}^{n\times n}$ are two orthonormal matrices. $R \in \mathbb{R}^{k\times k}$ is non-singular with diagonal entries being the singular values of F.

3. Partition P as $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$, where $P_{11} \in \mathbb{R}^{m\times k}$ and $P_{21} \in \mathbb{R}^{p\times k}$

4. Compute SVD of P_{11}. Assume it is $W\Sigma_A U^*$, where $W \in \mathbb{R}^{m\times m}$ and $U \in \mathbb{R}^{k\times k}$ are two orthonormal matrices. $\Sigma_A \in \mathbb{R}^{m\times k}$ is $\begin{matrix} r \\ s \\ m-r-s \end{matrix}\begin{bmatrix} I_A & & \\ & S_A & \\ & & 0_A \end{bmatrix}$, where S_A is a diagonal matrix $\begin{matrix} r & s & k-r-s \end{matrix}$ with all entries being the non-one singular values of A, 0_A denotes the matrix with all entries 0 and I_A denotes an identity matrix.

5. Let $S_B \in \mathbb{R}^{s\times s}$ be $I_s - S_A^2$, where I_s denotes an identity matrix. Form $\Sigma_B \in \mathbb{R}^{p\times k}$ as $\begin{matrix} p-k+r \\ s \\ k-r-s \end{matrix}\begin{bmatrix} 0_B & & \\ & S_B & \\ & & I_B \end{bmatrix}$, where 0_B also $\begin{matrix} r & s & k-r-s \end{matrix}$ denotes the matrix with all entries 0 and I_B also denotes an identity matrix.

6. Compute $Z = P_{21}\Sigma_B^* U$. Then SVD of P_{21} can be expressed as $Z\Sigma_B U^*$

7. $F = \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix} Q^*$

   $\Rightarrow \begin{bmatrix} A \\ B \end{bmatrix} Q = \begin{bmatrix} P_{11}R & 0 \\ P_{21}R & 0 \end{bmatrix} = \begin{bmatrix} W\Sigma_A U^* R & 0 \\ Z\Sigma_B U^* R & 0 \end{bmatrix}$

   Therefore, we have the following two equations which denote a GSVD of A and B.

$$W^* AQ = \Sigma_A \begin{bmatrix} U^*R & 0 \end{bmatrix} \tag{2.54}$$

$$Z^* BQ = \Sigma_B \begin{bmatrix} U^*R & 0 \end{bmatrix} \tag{2.55}$$

If $p = n$ and B is non-singular, then $k = n$. Hence,

$$W^* AQ = \begin{bmatrix} S_A & \\ & 0_A \end{bmatrix} U^*R, \quad Z^* BQ = \begin{bmatrix} S_B & \\ & I_B \end{bmatrix} U^*R$$

We can derive that

$$\begin{aligned}\begin{bmatrix} S_B & \\ & I_B \end{bmatrix} W^*AQ &= \begin{bmatrix} S_B & \\ & I_B \end{bmatrix} \begin{bmatrix} S_A & \\ & 0_A \end{bmatrix} U^*R \\ &= \begin{bmatrix} S_B & \\ & I_B \end{bmatrix} \begin{bmatrix} S_A & \\ & 0_A \end{bmatrix} \begin{bmatrix} S_B^{-1} & \\ & I_B \end{bmatrix} Z^*BQ \\ &= \begin{bmatrix} S_A & \\ & 0_A \end{bmatrix} Z^*BQ\end{aligned}$$

Therefore, we have

$$\Sigma_B W^*A = \Sigma_A Z^*B \tag{2.56}$$

In this section, we make detailed explanations about the GSVD and decomposition of an unitary matrix, which are introduced in [30].

2.5 Tikhonov Regularized Least-Squares Problems

Given $L \in \mathbb{R}^{p\times n}$ and $A \in \mathbb{R}^{m\times n}$, $m \geq n \geq p$. Define $F \triangleq \begin{bmatrix} L \\ A \end{bmatrix}$. Assume rank $L = p$ and rank $F = n$. From the discussion of the last section, we can construct a generalized singular value decomposition of the matrices L and A as

$$W^*LQ = \Sigma_L U^*R$$
$$Z^*AQ = \Sigma_A U^*R$$

where $\Sigma_L \in \mathbb{R}^{p\times n}$ is $\begin{matrix} r \\ s \end{matrix}\begin{bmatrix} I_L & & 0 \\ & S_L & 0 \end{bmatrix} = \begin{bmatrix} M_p & 0 \end{bmatrix}$ $(M_p = diag(\mu_i))$

$$\begin{matrix} r & s & n-r-s \end{matrix}$$

and $\Sigma_A \in \mathbb{R}^{m\times n}$ is $\begin{matrix} m-n+r \\ s \\ n-p \end{matrix}\begin{bmatrix} 0_A & & \\ & S_A & \\ & & I_A \end{bmatrix} = \begin{matrix} m-n \end{matrix}\begin{bmatrix} 0 & 0 \\ \Sigma_p & 0 \\ 0 & I_{n-p} \end{bmatrix}$

$$\begin{matrix} r & s & n-p \end{matrix}$$

$(\Sigma_p = diag(\sigma_i))$. The $\mu_i's$ and $\sigma_i's$ satisfy $0 \leq \sigma_1 \leq \cdots \leq \sigma_p$, $1 \geq \mu_1 \geq$

$\cdots \geq \mu_p > 0$, $\sigma_i^2 + \mu_i^2 = 1$. [1] Hence, we can write $A = Y\Sigma X^{-1}$ and $L = W\Sigma_L X^{-1}$, where $Y \in \mathbb{R}^{m\times n}$ is $[z_{m-n+1}\, z_{m-n+2} \cdots z_m]$, $\Sigma \in \mathbb{R}^{n\times n}$ is $\begin{bmatrix} \Sigma_p & 0 \\ 0 & I_{n-p} \end{bmatrix}$ and $X \in \mathbb{R}^{n\times n}$ is $(U^*RQ^*)^{-1} = QR^{-1}U$.

Consider a least-squares problem with Tikhonov regularization as follows.

$$\min_x \{\|Ax - b\|^2 + \lambda^2 \|Lx\|^2\} \tag{2.57}$$

where $b \in \mathbb{R}^m$ and λ is the regularization parameter that controls the weight given to minimization of the seminorm $\|Lx\|$ relative to minimization of the residual norm $\|Ax - b\|$. Let x_λ denote the Tikhonov regularized solution of the least-squares problem. We can analytically solve the least-squares problem by assigning the gradient of the objective function to zero. That is,

$$2A^*Ax_\lambda - 2A^*b + 2\lambda^2 L^*Lx_\lambda = 0$$

As a result,

$$\begin{aligned} x_\lambda &= (A^*A + \lambda^2 L^*L)^{-1} A^*b \\ &= \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \lambda^2 \mu_i^2} (y_i^* b) x_i + \sum_{i=p+1}^{n} (y_i^* b) x_i \end{aligned}$$

Define $\beta_i \triangleq y_i^* b$ and the generalized singular values γ_i as $\gamma_i \triangleq \frac{\sigma_i}{\mu_i}$. We can derive that

$$x_\lambda = \sum_{i=1}^{p} \frac{\gamma_i^2}{\gamma_i^2 + \lambda^2} \frac{\beta_i}{\sigma_i} x_i + \sum_{i=p+1}^{n} \beta_i x_i \tag{2.58}$$

We call $\phi_i \triangleq \frac{\gamma_i^2}{\gamma_i^2 + \lambda^2}$ as the filter factors for Tikhonov regularization, which can dampen or filter out the contributions to x_λ corresponding to the generalized singular values γ_i smaller than λ. Note that if there are indeed perturbation errors added on b, $y_i^* b$ will deviate from its original correct value. The largest corresponding perturbation to the

[1] Note that L corresponds to the matrix A and A corresponds to the matrix B of the last section. Besides, we interchange the notation of p and m, that is, the p in this section corresponds to the m of the last section and the m in this section corresponds to the p of the last section.

ordinary least-squares solution is associated with the smallest σ_i (and thus associated with the smallest γ_i). The existence of the regularization parameter λ effectively alleviates such error.

Define $\delta_0 \triangleq \|(I_m - YY^*b)\|$. It can be derived that

$$\|Lx_\lambda\|^2 = \sum_{i=1}^{p} \left(\frac{\gamma_i^2}{\gamma_i^2 + \lambda^2} \frac{\beta_i}{\sigma_i} \right)^2 \tag{2.59}$$

$$\|Ax_\lambda - b\|^2 = \sum_{i=1}^{p} \left(\frac{\lambda^2}{\gamma_i^2 + \lambda^2} \beta_i \right)^2 + \delta_0^2 \tag{2.60}$$

The physical meaning of δ_0 is that it is the norm of the component of b which is outside the range of A. The system $Ax = b$ is consistent if $\delta_0 = 0$; hence, δ_0 is also viewed as the incompatibility measure. We further define x_0 as the solution of the unregularized least-squares problem; that is, $x_0 \equiv \lim_{\lambda \to 0} x_\lambda = X\Sigma^\dagger Y^* b$ and $\delta_\infty \equiv \delta_0 + \|Y_p Y_p^* b\|$ where $Y_p = [y_1, \cdots, y_p]$. We observe that when there is no regularization, i.e., $\lambda = 0$, the residual norm $\|Ax_0 - b\| = \delta_0$ and the seminorm is $\|Lx_0\|$. When there is infinite regularization, i.e., $\lambda \to \infty$, the residual norm is δ_∞ and the seminorm is 0. Moreover, For $0 < \lambda < \infty$, we can verify that $\|Ax_\lambda - b\|$ and $\|Lx_\lambda\|$ simultaneously the following two equations :

$$\|Ax_\lambda - b\| = \min_x \|Ax - b\| \quad \textit{subject to } \|Lx\| \leq \|Lx_\lambda\| \tag{2.61}$$

$$\|Lx_\lambda\| = \min_x \|Lx\| \quad \textit{subject to } \|Ax - b\| \leq \|Ax_\lambda - b\| \tag{2.62}$$

The reason is as follows. We know that $x_\lambda = \underset{x}{argmin}\{\|Ax - b\|^2 + \lambda^2 \|Lx\|^2\}$. Assume there exists an x_1 such that $\|Ax_1 - b\| < \|Ax_\lambda - b\|$ and $\|Lx_1\| \leq \|Lx_\lambda\|$. Then $\|Ax_1 - b\|^2 + \lambda^2\|Lx_1\|^2 < \|Ax_\lambda - b\|^2 + \lambda^2\|Lx_\lambda\|^2$. This contradicts with the fact that $x_\lambda = \underset{x}{argmin}\{\|Ax - b\|^2 + \lambda^2\|Lx\|^2\}$. Assume there exists an x_2 such that $\|Lx_2\| < \|Lx_\lambda\|$ and $\|Ax_2 - b\| \leq \|Ax_\lambda - b\|$. Then $\|Ax_2 - b\|^2 + \lambda^2\|Lx_2\|^2 < \|Ax_\lambda - b\|^2 + \lambda^2\|Lx_\lambda\|^2$. This again contradicts with the fact that $x_\lambda = \underset{x}{argmin}\{\|Ax - b\|^2 + \lambda^2\|Lx\|^2\}$.

In this section, we introduce the Tikhonov regularization least-squares problem and some properties regarding the seminorm $\|Lx_\lambda$ and the residual norm $\|Ax_\lambda - b\|$. Actually, it is useful to plot the seminorm $\|Lx_\lambda\|$ versus the residual norm $\|Ax_\lambda - b\|$ to visualize the compromise between the minimization of these two quantities. The corresponding graph is called the L-curve, which consists of all points $(\|Ax_\lambda - b\|, \|Lx_\lambda\|)$ for $\lambda \in [0, \infty)$. In section 5.1, we will delve deeply into the properties of the L-curve and present the L-curve method, as well as other methods, e.g., the discrepancy principle, for determining a good regularization parameter λ.

2.6 Sparsity and Compressiblity

The support of a vector $x \in \mathbb{R}^n$ is the index set of its nonzero entries, which is defined as

Definition 2.6.1 (support)**.**

$$supp(x) := \{j \in [N] : x[j] \neq 0\} \tag{2.63}$$

x is called s-sparse if at most s of its entries are nonzero. We quantify the concept of sparsity using the ℓ_0 norm defined as follows.

Definition 2.6.2 (sparsity)**.**

$$\|x\|_0 := card(supp(x)) \tag{2.64}$$

Hence, if x is s-sparse, $\|x\|_0 \leq s$. The notation $\|\cdot\|_0$ comes from the observation that the p-th power of the ℓ_p-quasinorm of x (i.e., $\|x\|_p^p = \sum_{j=1}^{n} |x[j]|^p$ approaches $card(supp(x))$ as p approaches zero. Therefore, the notation $\|\cdot\|_0^0$ is actually more appropriate but $\|\cdot\|_0$ is customary to use in the literature. Furthermore, the term ℓ_0 norm is actually a misuse since it is neither a norm nor a quasinorm but it is also customary to call it ℓ_0 "norm" in the literature. If x is not s-sparse,

we may want to "measure" how close it is to be s-sparse, which results in the concept of best s-term approximation.

Definition 2.6.3 (best s-term approximation)**.** For $p > 0$, the ℓ_p error of best s-term approximation to a vector $x \in \mathbb{R}^n$ is defined by

$$\sigma_s(x)_p := \inf\{\|x - z\|_p, z \in \mathbb{R}^n \ is \ s - sparse\} \tag{2.65}$$

Clearly, the infimum is achieved by x_s, which is an s-sparse vector whose nonzero entries equal the s largest (in modulus) components of x. We call x_s the best s-term approximation to x. Note that $\sigma_s(x)_p$ satisfies the following useful inequality.

$$\sigma_s(x)_p \leq \frac{1}{s^{1/q-1/p}}\|x\|_q \tag{2.66}$$

for any $p > q > 0$ and any $x \in \mathbb{R}^n$. This inequality is the same as that stated in the proposition 2.3 of [13]. Actually, we can easily verify this inequality by the inequality 2.8. We may also introduce the notations $H_s(x)$ for the best s-term approximation to x and $L_s(x)$ for the support of it. That is,

$$L_s(x) := index \ set \ of \ s \ largest \ absolute \ entries \ of \ x \tag{2.67}$$

$$H_s(x) := x|_{L_s(x)} \tag{2.68}$$

We call the nonlinear operator $H_s(x)$ the hard thresholding operator of order s.

The sparsity condition may somehow be too stringent for a signal to satisfy. We may rather consider the compressibility condition. First, we introduce the non-increasing rearrangement of a vector x.

Definition 2.6.4 (non-increasing rearrangement)**.** The non-increasing rearrangement of a vector $x \in \mathbb{R}^n$ is the vector $x^* \in \mathbb{R}^n$ for which

$$\begin{aligned} x^*[1] \geq x^*[2] \geq \cdots \geq x^*[n] \geq 0 \\ and \ x^*[j] = |x[\mathcal{I}(j)]| \ \forall j \in [n] \end{aligned} \tag{2.69}$$

where $\mathcal{I}$ indexes the sorted components of x.

Then we define the compressibility as follows.

Definition 2.6.5 (compressibility)**.** An n-dimensional vector x is said to be r-compressible with magnitude G if the sorted components of the vector obey the power law :

$$x^*[i] \leq Gi^{-1/r} \tag{2.70}$$

where $x^* \in \mathbb{R}^n$ denotes the non-increasing rearrangement of x and $r \in (0, 1]$.

We can approximate an r-compressible vector with a sparse signal and bound the approximation error as follows

$$\begin{aligned}
\sigma_s(x)_p &= \left(\sum_{i=s+1}^{n} |x[i]|^p \right)^{1/p} \\
&\leq \left(\sum_{i=s+1}^{n} (Gi^{-1/r})^p \right)^{1/p} \\
&\leq G \left(\int_s^n i^{-p/r} di \right)^{1/p} \\
&\leq G \left(\int_s^\infty i^{-p/r} di \right)^{1/p} \\
&= G \left(\frac{r}{p-r} \right)^{1/p} s^{\frac{r-p}{pr}} \\
&\leq G \left(\frac{p}{p-r} \right)^{1/p} s^{\frac{r-p}{pr}} \\
&= G(rk)^{-1/p} s^{-k}
\end{aligned}$$

where $k \triangleq \frac{1}{r} - \frac{1}{p}$ and $p \geq 1$.

2.7 Random Matrices

A random matrix has random variables as their entries. It plays an important role in the field of compressive sensing since compared with a deterministic matrix, a random matrix can satisfy the restricted isometry property with asymptotically fewer number of rows, which in turn leads to fewer number of measurements required for reconstruction of a signal. We will cover the restricted isometry property in the next section and also introduce how a random matrix behaves in the aspect of it. In this section, we will introduce two important categories of random matrices, which are subgaussian random matrices and structured random matrices respectively.

2.7.1 Subgaussian Random Matrices

To introduce subgaussian random matrices, we need to introduce what a subgaussian random variable is first.

Definition 2.7.1 (subgaussian random variable)**.** A random variable X is called subgaussian if $\exists$ constants $\beta, \kappa > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t^2} \ \forall t > 0 \tag{2.71}$$

From the definition, we know that a random variable X is subgaussian if its tail distribution is dominated by that of the standard Gaussian random variable. A standard Gaussian random variable is subgaussian with $\beta = 1$ and $\kappa = 1/2$. We can prove it as follows. Let

g be a standard Gaussian random variable.

$$\begin{aligned}
\mathbb{P}(|g| \geq t) &= \frac{2}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2} du \\
&= \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-(u+t)^2/2} du \\
&= \frac{2}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-tu} e^{-u^2/2} du \\
&\leq \frac{2}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-u^2/2} du \ \because e^{-tu} \leq 1 \, for \, t, u \geq 0 \\
&= e^{-t^2/2}
\end{aligned}$$

A zero-mean and bounded random variable is also subgaussian. Indeed, due to theorem 7.20 of [13], we know that it is subgaussian with $\beta = 2$ and $\kappa = \frac{1}{2B^2}$ if $|X| \leq B$. A Rademacher variable (sometimes also called symmetric Bernoulli variable) is an important example of a zero-mean and bounded random variable, which is defined as follows.

Definition 2.7.2 (Rademacher variable)**.** A Rademacher variable is a random variable ϵ that takes the values +1 and -1 with equal probability, i.e.

$$\mathbb{P}(\epsilon = +1) = \mathbb{P}(\epsilon = -1) = \frac{1}{2} \tag{2.72}$$

Hence, a Rademacher variable is a subgaussian random variable with $\beta = 2$ and $\kappa = 1/2$. Besides the definition 2.7.1, the proposition 7.24 of [13] presents another equivalent condition related to the moment generating function of a subgaussian random variable X.

Theorem 2.7.1. *Let X be a random variable.*

1. *If X is subgaussian with mean zero, then there exists a constant c (depending only on β and κ) such that*

$$\mathbb{E}[\exp(\theta X)] \leq \exp(c\theta^2) \quad \forall \theta \in \mathbb{R} \tag{2.73}$$

2. *Conversely, if 2.73 holds, then the mean of X is zero and X is subgaussian with parameters $\beta = 2$ and $\kappa = 1/(4c)$.*

Any valid constant c in 2.73 is called a subgaussian parameter of X. Of course, it is preferable to choose the minimal possible c. It can be derived that the moment generating function of a standard Gaussian random variable is

$$\mathbb{E}[\exp(\theta g)] = \exp(\theta^2/2) \tag{2.74}$$

where g is a standard Gaussian random variable and $\theta \in \mathbb{R}$. Therefore, $c = 1/2$ is a valid subgaussian parameter for a standard Gaussian random variable. As for a zero-mean and bounded random variable, its moment generating function has been derived in the theorem 7.20 of [13] to be

$$\mathbb{E}[\exp(\theta X)] \leq \exp(\theta^2 B^2/2) \tag{2.75}$$

where X is a zero-mean and bounded random variable with $|X| \leq B$. Therefore, $c = B^2/2$ is a valid subgaussian parameter for a zero-mean and bounded random variable, which implies that $c = 1/2$ is a valid subgaussian parameter for a Rademacher variable.

Let's pay our attention back to the subgaussian random matrices.

Definition 2.7.3 (subgaussian random matrices)**.** Let $A \in \mathbb{R}^{m\times n}$ be a random matrix. If the entries of A are independent zero-mean subgaussian random variables with variance 1 and same subgaussian parameters β, κ in 2.71, i.e.,

$$\mathbb{P}(|A_{j,k}| \geq t) \leq \beta e^{-\kappa t^2} \quad \forall t > 0, j \in [m], k \in [n] \tag{2.76}$$

then A is called a subgaussian random matrix. Specifically,

1. If the entries of A are independent Rademacher variables, then A is called a Bernoulli random matrix.
2. If the entries of A are independent standard Gaussian random variable, then A is called a Gaussian random matrix.

Note that the entries of a subguassian random matrix do not necessarily have to be identically distributed. They only need to satisfy the equation 2.76.

2.7.2 Structured Random matrices

Structured random matrix is a structured matrix generated by a random choice of parameters. Why do we need structured random matrices? In some applications, the measurement matrices have certain structures due to physical or other constraints and because of this fact, fast matrix-vector algorithms, e.g. fast Fourier transform (FFT), are allowed. Furthermore, since a structured matrix is generated by a number of parameters much fewer than the matrix entries, so that it consumes less space to store a structured random matrix than a unstructured one. In this section, we will introduce an important kind of structured random matrices - random sampling matrices, which are associated with bounded orthonormal systems (BOSs).

Definition 2.7.4 (bounded orthonormal system)**.** Let $\mathcal{D} \subset \mathbb{R}^d$ be endowed with a probability measure ν. $\Phi = \{\phi_1, \cdots, \phi_n\}$ is called a bounded orthonormal system (BOS) of complex-valued functions on $\mathcal{D}$ with constant K if

1. $\int_{\mathcal{D}} \phi_j(t)\overline{\phi_k(t)}d\nu(t) = \delta_{j,k}$ (orthonormal)
2. $\|\phi_j\|_\infty := \sup_{t\in\mathcal{D}} |\phi_j(t)| \leq K$, $\forall j \in [n]$ (bounded)
   where K should be independent of n and be no less than one ($K \geq 1$).

Definition 2.7.5 (random sampling matrix)**.** A random sampling matrix $A \in \mathbb{R}^{m\times n}$ associated with a BOS with constant K is constructed by

$$A_{\ell,k} = \phi_k(t_\ell) \quad \ell \in [m], k \in [n] \tag{2.77}$$

where $t_1, \cdots, t_m$ are sampling points selected independently at random according to the probability measure ν.

Hence, A has stochastically independent rows, but the entries within each row are not independent. Indeed, for fixed ℓ, the entries $A_{\ell,k}$, $k \in [n]$, all depend on the single random sampling point t_ℓ. In

the following, we will introduce some important BOSs and random sampling matrices associated with them.

Trigonometric Polynomials Systems

A trigonometric polynomials system is defined as

$$\{\phi_k(t) = e^{j\,2\pi kt}\,, k \in \mathbb{Z}\} \tag{2.78}$$

The domain $\mathcal{D}$ of each element is $[0, 1]$ and the probability measure is chosen as the uniform distribution. Clearly, the boundedness condition is satisfied with the constant $K = 1$. The orthonormal condition can also be easily verified since $\int_0^1 \phi_k(t)\overline{\phi_j(t)}dt = \delta_{j,k}$ for $j, k \in \mathbb{Z}$. Therefore, a trigonometric polynomials system is indeed a BOS. Random sampling matrices associated with it can be constructed by

$$A_{\ell,k} = e^{j\,2\pi kt_\ell} \quad \ell \in [m], k \in \Gamma \subset \mathbb{Z}\ of\ size\ n \tag{2.79}$$

The corresponding random sampling matrices are called non-equispaced Fourier matrices and a common choice of Γ is $\{-q, -q+1, \cdots, q-1, q\}$ $(n = 2q + 1)$.

Discrete Orthonormal Systems

A discrete orthonormal system is defined as

$$\{\sqrt{n}u_k\,, k \in [n]\} \tag{2.80}$$

where u_k are columns of a unitary matrix U. The domain $\mathcal{D}$ of each element is $[n]$ and the probability measure is chosen as the uniform distribution. The orthonormal condition is satisfied since

$$\frac{1}{n}\sum_{t=1}^{n} \sqrt{n}u_k(t)\overline{\sqrt{n}u_\ell(t)} = \langle u_k, u_\ell\rangle = \delta_{k,\ell}\ , k, \ell \in [n]$$

Therefore, a discrete orthonormal system is a BOS if the boundedness condition is satisfied with a constant K; that is,

$$\max_{k,t\in[n]} |\sqrt{n}u_k(t)| = \max_{k,t\in[n]} |\sqrt{n}U_{t,k}| \le K$$

Random sampling matrices associated with it can be constructed by selecting rows of $\sqrt{n}U$ independently and uniformly at random, i.e.,

$$a_k = \sqrt{n} R_T u_k \, , k \in [n] \tag{2.81}$$

where a_k denotes the k-th column of A, $T = \{t_1, \cdots, t_m\}$ and $R_T : \mathbb{R}^n \to \mathbb{R}^m$ denotes the random subsampling operator

$$(R_T u_k)[\ell] = (u_k)[t_\ell] \, , \ell \in [m] \tag{2.82}$$

Note that it may happen that a row of $\sqrt{n}U$ is selected more than once. Hence, A may have repeated rows.

If we choose the unitary matrix U to be the discrete Fourier matrix F with entries

$$F_{\ell,k} = \frac{1}{\sqrt{n}} e^{j\, 2\pi(\ell-1)(k-1)/n} \;\; \ell, k \in [n] \tag{2.83}$$

then a BOS with the constant 1 can be established. The corresponding random sampling matrix is famous for being the random partial Fourier matrix. It can be viewed as a special case of the non-equispaced Fourier matrix with the points t_ℓ chosen from the grid $\mathbb{Z}_n/n := \{0, 1/n, \cdots, (n-1)/n\}$ instead of the whole interval $[0, 1]$. Taking measurements Ax with a random partial Fourier matrix A amounts to observing m random frequencies of the signal x. A crucial advantage of the random partial Fourier matrix is that it possesses a fast matrix-vector multiplication, namely, the FFT.

Suppose we choose U to be

$$U = W^* V \tag{2.84}$$

where $W \in \mathbb{R}^n$ and $V \in \mathbb{R}^n$ are two unitary matrices. That is, their columns form two orthonormal bases of $\mathbb{R}^n$. Since

$$U^* U = U U^* = I_n$$

U is indeed an unitary matrix. A BOS thus can be established if the boundedness condition is satisfied with a constant K; that is,

$$\max_{\ell,k\in[n]} |\sqrt{n}\langle v_\ell, w_k\rangle| \leq K \tag{2.85}$$

where v_ℓ denotes the ℓ-th column of V and w_k denotes the k-th column of W. The bases $\{v_\ell, \ell \in [n]\}$ and $\{w_k, k \in [n]\}$ are called incoherent if K can be chosen small. Note that the random partial Fourier matrix falls into this setting by choosing one of the bases as the canonical basis, say $W = I_n$ and we can observe that the Fourier basis and the canonical basis are maximally incoherent since $K = 1$.

2.8 Restricted Isometry Property

The s-th restricted isometry constant of a sampling matrix $A \in \mathbb{R}^{m\times n}$ is the smallest number δ_s such that

$$(1-\delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_s)\|x\|_2^2 \tag{2.86}$$

for any s-sparse vector $x \in \mathbb{R}^n$. The intuition is that we want the geometry of sparse signals to be preserved under the action of the sampling matrix. We can also view this inequality relation in another aspect. Consider two vectors $a_1 \in \mathbb{R}^2$ and $a_2 \in \mathbb{R}^2$ with unit norm. Construct a vector $y = x_1a_1 + x_2a_2 = Ax$, where $A = [a_1, a_2]$, $x = [x_1, x_2]^T$ and $\sqrt{x_1^2 + x_2^2} = 1$. $\|Ax\|_2^2 = x_1^2 + x_2^2 + 2\langle x_1a_2, x_2a_2\rangle = 1 + 2x_1x_2\langle a_1, a_2\rangle$. Hence if the correlation of a_1 and a_2 are smaller, the restricted isometry constant will be smaller. Intuitively speaking, if vectors correlate with each other lesser, then they are "dissimilar" to each other more. In this way, given a linear combination $y = Ax$, it is more possible that x is the unique combination coefficient that can give rise to y (or at least, if there exists another vector z satisfying $y = Az$, then $\|z - x\|_2$ is small enough). Rigorously speaking, if there is a subset of 2s (we denote such index set as Ω_3) columns of A with nonzero nullity, then there exists a nonzero 2s-sparse vector x_3 satisfying $Ax_3 = 0$. Arbitrarily choose a subset of s entries of

Ω_3 and denote it as Ω_1. In this way, $x_3 = (x_3)_{\Omega_1} - (-x_3)_{\Omega_1^c}$. Let $(x_3)_{\Omega_1}$ be denoted as x_1 and $(-x_3)_{\Omega_1^c}$ be denoted as x_2. Hence, $Ax_3 = A(x_1 - x_2) = 0$, i.e. $Ax_1 = Ax_2$. If x_3 happens to be an 1-sparse vector, then we choose $x_1 = \frac{1}{2}x_3$ and $x_2 = -\frac{1}{2}x_3$. Therefore, to uniquely recover an s-sparse signal, it is necessary that every subset of 2s columns of A be independent (equivalently, $\delta_{2s} < 1$), which in turn requires that $m \geq 2s$.

In the following, we collect some important properties and results relating to the restricted isometry constant.

1. For any two integers $s \leq t$, $\delta_s \leq \delta_t$. That is, the restricted isometry constant is monotonically increasing with the sparsity level.

   Proof. This can be easily verified by the definition of the restricted isometry constant. □

2. Let c and s be two positive integers. Then $\delta_{cs} \leq c\delta_{2s}$.

   Proof. see corollary 3.4 of [26] □

   This property states that δ_{2s} can give an upper bound of higher-order restricted isometry constants.

3. Suppose A has s-th restricted isometry constant δ_s. Let T be a set of s indices or fewer. Then the singular values of A_T lie between $\sqrt{1-\delta_s}$ and $\sqrt{1+\delta_s}$

   - The singular values of A_T^* is the same as those of A_T. Hence,

   $$\sqrt{1-\delta_s}\|u\|_2 \leq \|A_T^* u\|_2 \leq \sqrt{1+\delta_s}\|u\|_2 \quad \forall u \in \mathbb{R}^m \quad (2.87)$$

   - The singular values of $A_T^\dagger$ is the reciprocal of those of A_T. Hence,

   $$\frac{1}{\sqrt{1+\delta_s}}\|u\|_2 \leq \|A_T^\dagger u\|_2 \leq \frac{1}{\sqrt{1-\delta_s}}\|u\|_2 \quad \forall u \in \mathbb{R}^m \quad (2.88)$$

   - The singular values of $A_T^* A_T$ is the square of those of A_T. Hence,

   $$(1-\delta_s)\|u\|_2 \leq \|A_T^* A_T u\|_2 \leq (1+\delta_s)\|u\|_2 \quad \forall u \in \mathbb{R}^{|T|} \quad (2.89)$$

- The singular values of $(A_T^* A_T)^{-1}$ is the reciprocal of those of $A_T^* A_T$. Hence,

$$\frac{1}{1+\delta_s}\|u\|_2 \leq \|(A_T^* A_T)^{-1} u\|_2 \leq \frac{1}{1-\delta_s}\|u\|_2 \quad \forall u \in \mathbb{R}^m \tag{2.90}$$

4. Suppose $\|Ax\|_2 \leq \sqrt{1+\delta_s}\|x\|_2$ for any s-sparse vectors $x \in \mathbb{R}^n$. Then for every signal $v \in \mathbb{R}^n$,

$$\|Av\|_2 \leq \sqrt{1+\delta_s}(\|v\|_2 + \frac{1}{\sqrt{s}}\|v\|_1) \tag{2.91}$$

Proof. see proposition 3.5 of [26]. □

5. Suppose A has s-th restricted isometry constant δ_s. Let S and T be disjoint sets of indices whose combined cardinality does not exceed s. Then

$$\|A_S^* A_T\|_{2\to 2} \leq \delta_s \tag{2.92}$$

Proof. see proposition 3.2 of [26]. □

Assume a vector x has support $supp(x)$ satisfying the cardinality of the union of T and $supp(x)$ does not exceed s. Then

$$\begin{aligned}\|A_T^* A x|_{T^c}\|_2 &= \|A_T^* A x|_S\|_2, \textit{ where } S \triangleq supp(x)\backslash T \\ &= \|A_T^* A_S x|_S\|_2 \leq \|A_T^* A_S\|_{2\to 2}\|x|_S\|_2 \\ &\leq \delta_s \|x|_{T^c}\|_2\end{aligned}$$

Let S and T be disjoint sets of indices whose combined cardinality does not exceed s and $a \in \mathbb{R}^{|S|}$ and $b \in \mathbb{R}^{|T|}$ be two vectors. It can be verified that

$$|\langle A_S a, A_T b\rangle| \leq \delta_s \|a\|_2 \|b\|_2 \tag{2.93}$$

6. Let $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, $I, J \subset [n]$ be two disjoint sets. Assume $y \in span(A_I)$ and let the projection of y onto $span(A_J)$ be $y_p =$

$A_J A_J^{\dagger} y$. Then

$$\|y_p\|_2 \leq \frac{\delta_{|I|+|J|}}{1-\delta_{max(|I|,|J|)}}\|y\|_2 \tag{2.94}$$

Let $y_r = y - y_p$. We can further derive that

$$(1 - \frac{\delta_{|I|+|J|}}{1-\delta_{max(|I|,|J|)}})\|y\|_2 \leq \|y_r\|_2 \leq \|y\|_2 \tag{2.95}$$

Proof. see lemma 2 of [6]. □

In the following, we will introduce the restricted isometry property of the subgaussian random matrices and the random sampling matrices. We excerpt some theorems in section 9.1, 9.3 and 12.5 of [13].

Theorem 2.8.1. *Let $A \in \mathbb{R}^{m\times n}$ be a subgaussian random matrix. Then there exists a constant $C > 0$ (depending only on the subgaussian parameters β, κ) such that the restricted isometry constant of $\frac{1}{\sqrt{m}}A$ satisfies $\delta_s \leq \delta$ with probability at least $1-\epsilon$ provided*

$$m \geq C\delta^{-2}\left(s\ln(eN/s) + \ln(2\epsilon^{-1})\right) \tag{2.96}$$

Setting $\epsilon = 2\exp(-\delta^2 m/(2C))$ yields the condition

$$m \geq 2C\delta^{-2}\ln(eN/s) \tag{2.97}$$

In this way, $\delta_s \leq \delta$ with probability at least $1 - 2\exp(-\delta^2 m/(2C))$.

Theorem 2.8.2. *Let $A \in \mathbb{R}^{m\times n}$ be a Gaussian random matrix. For $\eta, \epsilon \in (0,1)$, assume that*

$$m \geq 2\eta^{-2}\left(s\ln(eN/s) + \ln(2\epsilon^{-1})\right) \tag{2.98}$$

Then with probability at least $1-\epsilon$, the restricted isometry constant δ_s of $\frac{1}{\sqrt{m}}A$ satisfies

$$\delta_s \leq 2\left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)\eta + \left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)^2\eta^2 \leq C\eta \tag{2.99}$$

where $C = 2(1+\sqrt{1/2}) + (1+\sqrt{1/2})^2 \approx 6.3284$.

Therefore, if

$$\begin{aligned} m &\geq 2C^2\delta^{-2}\left(s\ln(eN/s) + \ln(2\epsilon^{-1})\right) \\ &\approx 80.098\delta^{-2}\left(s\ln(eN/s) + \ln(2\epsilon^{-1})\right) \end{aligned}$$

Then with probability at least $1-\epsilon$, the restricted isometry constant δ_s of $\frac{1}{\sqrt{m}}A$ satisfies

$$\delta_s \leq C(\frac{\delta}{C}) = \delta$$

Theorem 2.8.3. *Let* $A \in \mathbb{R}^{m\times n}$ *be a random sampling matrix associated to a BOS with constant* $K \geq 1$. *For* $\epsilon, \eta_1, \eta_2 \in (0,1)$, *if*

$$\frac{m}{\ln(9m)} \geq C_1\eta_1^{-2}K^2 s\ln^2(4s)\ln(8n) \tag{2.100}$$

$$m \geq C_2\eta_2^{-2}K^2 s\ln(\epsilon^{-1}) \tag{2.101}$$

then with probability at least $1-\epsilon$, *the restricted isometry constant* δ_s *of* $\frac{1}{\sqrt{m}}A$ *satisfies* $\delta_s \leq \eta_1 + \eta_1^2 + \eta_2$. *The constants may be chosen as* $C_1 \approx 5576$ *and* $C_2 = 32/3$.

Note that 2.100 and 2.101 can be implied by

$$m \geq C'K^2\delta^{-2}s\max\{ln^2(s)ln(K^2\delta^{-2}sln(n))ln(n), ln(\epsilon^{-1})\} \tag{2.102}$$

for some constant $C' > 0$. If ϵ is chosen to be $n^{-ln^3(n)}$, then $\delta_s \leq \delta$ with probability at least $1 - n^{-ln^3(n)}$ if

$$m \geq C''K^2\delta^{-2}sln^4(n) \tag{2.103}$$

for some constant $C'' > 0$.

Besides the common definition 2.86 of the restricted isometry property. Some authors adopt the alternative definition as follows, e.g., the restricted isometry condition introduced in [27] and [28]. The s-th restricted isometry constant is defined as the smallest number $\hat{\delta}_s$ (to differentiate this definition from 2.86, we adopt the notation $\hat{\delta}_s$ here)

such that

$$(1-\hat{\delta}_s)\|x\|_2 \leq \|Ax\|_2 \leq (1+\hat{\delta}_s)\|x\|_2 \tag{2.104}$$

for any s-sparse vector $x \in \mathbb{R}^n$. Assume a sampling matrix $A \in \mathbb{R}^{m\times n}$ satisfies $\hat{\delta}_{2s} = \epsilon$, then it has been proved in the proposition 3.2 of [27] that

1. For every s-sparse vector $x \in \mathbb{R}^n$ and every set $T \subset [n]$, $|T| \leq s$, the vector $u = A^*Ax$ satisfies

$$\|u|_T - x|_T\|_2 \leq 2.03\epsilon\|x\|_2 \tag{2.105}$$

2. For any vector $z \in \mathbb{R}^n$ and every set $T \subset [n]$, $|T| \leq 2s$, we have

$$\|(A^*z)|_T\|_2 \leq (1+\epsilon)\|z\|_2 \tag{2.106}$$

3. Consider two disjoint sets $I, J \subset [n]$, $|I \cup J| \leq 2s$. Let P_I and P_J denote the orthogonal projections in $\mathbb{R}^n$ onto range(A_I) and range(A_J), respectively. Then

$$\|P_IP_J\|_{2\to 2} \leq 2.2\epsilon \tag{2.107}$$

Because of the restricted isometry property, for any set $T \subset [n]$ with $|T| \leq s$, submatrices A_T are almost isometries. Therefore, $u = A^*Ax$ approximates x locally when restricted to a set of cardinality s. We thus call the first result the local approximation property. The third result states that range(A_I) and range(A_J) are almost orthogonal. Hence, range(A_I) is close to the orthogonal complement of range(A_J) in range$(A_{I\cup J})$.

Chapter 3

Unstructured Optimization

In this chapter, we focus on optimization problems that are unstructured, which means they do not take specific problem formulations. Instead, we only assume the objective functions and the constraint functions to possess particular mathematical properties. For instance, the differentiability, the convexity or the smoothness. Based on these mathematical assumptions, we manage to design effective algorithms in order to meet the purpose of optimization and also analyze them with mathematical rigorous accordingly. On the contrary, in chapter four, we study ℓ_0 minimization problems, which take specific problem formulations and thus do not fall in the category of unstructured optimization.

3.1 "Minorization-Maximization" Viewpoint

In this section, we will introduce the important duality theory based on the concept of "minorization-maximization". First, we find a lower bound of the objective function. Then we try to maximize the lower bound hoping that the maximizer can ideally be a solution of the original optimization problem. This is the reason why such design methodology is called "minorization-maximization". Proceeding with the duality theory, we introduce some algorithms, including the Chambolle and Pock's primal-dual algorithm, the augmented Lagrangian method and ADMM, that makes use of the theory of duality.

3.1.1 Duality Theory

We consider a general optimization problem as follows.

$$\begin{aligned} &\inf_x f_0(x) \\ &\textit{subject to} \quad f_i(x) \leq 0, i = 1, \cdots, m \\ &\qquad\qquad\quad h_i(x) = 0, i = 1, \cdots, p \end{aligned} \tag{3.1}$$

The domain is $\mathcal{D} = \left(\bigcap_{i=0}^{m} dom\, f_i\right) \cap \left(\bigcap_{i=1}^{p} dom\, h_i\right)$. Define

$$I_-(u) := \begin{cases} 0 & , u \leq 0 \\ \infty & , u > 0 \end{cases}$$

and

$$I_0(u) := \begin{cases} 0 & , u = 0 \\ \infty & , u \neq 0 \end{cases}$$

Then 3.1 is equivalent to

$$\inf_{x \in \mathcal{D}} f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) + \sum_{i=1}^{p} I_0(h_i(x)) \tag{3.2}$$

To lower-bound 3.2, we replace $I_-(u)$ with the linear functions $\lambda_i u$, $i = 1, \cdots, m$, $\lambda_i \geq 0$ and $I_0(u)$ with the linear functions $\nu_i u$, $i = 1, \cdots, p$ (since $I_-(u) \geq \lambda_i u$, $I_0(u) \geq \nu_i u \; \forall u$). Hence we define a function called Lagrangian as follows to be a lower bound of 3.2

Definition 3.1.1 (Lagrangian)**.**

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{3.3}$$

$\lambda_i, i = 1, \cdots, m$ and $\nu_i, i = 1, \cdots, p$ are called Lagrange multipliers and λ and ν are called dual variables or Lagrange multiplier vectors.

We further define the Lagrangian dual function as follows.

Definition 3.1.2 (Lagrangian dual function)**.**

$$g(\lambda, \nu) := \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \tag{3.4}$$

The domain of g is $dom\, g := \{(\lambda, \nu) | g(\lambda, \nu) > -\infty\}$.

Denoting the optimal value of 3.1 as p^*, we observe that for any $\lambda \succeq 0$ and any ν $g(\lambda, \nu)$ will always be smaller than or equal to p^*, i.e., the dual function yields a lower bound on the optimal value of 3.1. Hence, we are motivated to maximize the Lagrangian dual function over the variables λ and ν hoping to achieve p^*. This motivation gives rise to defining the Lagrange dual problem associated with 3.1 as follows

Definition 3.1.3 (Lagrange dual problem)**.**

$$\sup_{\lambda, \nu} g(\lambda, \nu) \quad subject\ to\ \lambda \succeq 0 \tag{3.5}$$

As a counterpart, 3.1 is called the primal problem.

Note that since the Lagrangian dual function will always be concave, the Lagrange dual problem (or simply dual problem) is always a convex optimization problem no matter whether the primal problem is a convex optimization problem or not.

Suppose $\hat{x}$ satisfies $f_i(\hat{x}) \leq 0,\ i = 1, \cdots, m$ and $h_i(\hat{x}),\ i = 1, \cdots, p$, then we call $\hat{x}$ primal feasible. As a counterpart, a pair $(\hat{\lambda}, \hat{\nu})$ is called dual feasible if $\hat{\lambda} \succeq 0$ and $(\hat{\lambda}, \hat{\nu}) \in dom\, g$. Suppose x^* is optimal for the primal problem (hence, certainly primal feasible), we call x^* primal optimal ($f_0(x^*) = p^*$). As a counterpart, if (λ^*, ν^*) is optimal for the dual problem (hence, certainly dual feasible), we call (λ^*, ν^*) dual optimal. Let d^* denote the optimal value of the dual problem. Then $g(\lambda^*, \nu^*) = d^*$. As a summary, we can have the following inequality equation.

$$g(\hat{\lambda}, \hat{\nu}) \leq g(\lambda^*, \nu^*) = d^* \leq p^* = f_0(x^*) \leq f_0(\hat{x}) \tag{3.6}$$

We call the relation

$$d^* \leq p^* \tag{3.7}$$

weak duality, which always holds. The corresponding difference

$$p^* - d^* \tag{3.8}$$

is called the optimal duality gap. Attractively, we say that strong duality holds if

$$p^* = d^* \tag{3.9}$$

In such situation, the optimal duality gap is zero. Besides, if $f_0(\hat{x}) - g(\hat{\lambda}, \hat{\nu}) = \epsilon$, then we say that $\hat{x}$ is ϵ-suboptimal for the primal problem since $f_0(\hat{x}) - p^* \leq f_0(\hat{x}) - g(\hat{\lambda}, \hat{\nu}) = \epsilon$ and $(\hat{\lambda}, \hat{\nu})$ is ϵ-suboptimal for the dual problem since $d^* - g(\hat{\lambda}, \hat{\nu}) \leq f_0(\hat{x}) - g(\hat{\lambda}, \hat{\nu}) = \epsilon$

Let x^* be primal optimal and (λ^*, ν^*) be dual optimal. Suppose the strong duality holds, then

$$\begin{aligned} f_0(x^*) = g(\lambda^*, \nu^*) &= \inf_x \{ f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i^*(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \} \\ &\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\ &\leq f_0(x^*) \end{aligned}$$

The first equality is established because of the strong duality. The second equality is established because of the definition of the Lagrange dual function. The first inequality is established clearly and holds with equality if x^* minimizes the Lagrangian $L(x, \lambda^*, \nu^*)$ over x. The second inequality is established because λ^* is feasible (hence, $\lambda_i^* \geq 0$) and x^* is feasible (hence, $f_i(x^*) \leq 0$ and $h_i(x^*) = 0$). The second inequality holds with equality if $\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$. Since $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$, $\lambda_i^* f_i(x^*) = 0 \ \forall i \in [m]$, which is called the complementary slackness condition. As a summary, we conclude that for any optimization problem with differentiable objective and constraint functions for which strong duality holds, any pair of primal and dual optimal points (x^*, λ^*, ν^*) should satisfy the following necessary conditions

$$f_i(x^*) \leq 0, i \in [m] \ and \ h_i(x^*) = 0, i \in [p] \tag{3.10}$$

$$\lambda_i^* \geq 0, i \in [m] \tag{3.11}$$

$$\lambda_i^* f_i(x^*) = 0, i \in [m] \tag{3.12}$$

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0 \tag{3.13}$$

Condition 3.10 is the primal feasible condition. Condition 3.11 is the dual feasible condition. Condition 3.12 is the complementary slackness condition. Condition 3.13 is made because $\nabla L(x^*, \lambda^*, \nu^*) = 0$. As a whole, conditions 3.10, 3.11, 3.12 and 3.13 are jointly called the Karush-Kuhn-Tucker (KKT) conditions.

Two questions arise naturally. One is that when KKT conditions are sufficient conditions for primal and dual optimal points. The other is that when the strong duality can hold. To answer these two questions, we consider the optimization problem to be convex in the subsequent discussion. That is, $f_i, i = 0, 1, \cdots, m$ are convex and $h_i, i \in [p]$ are affine. Assume a primal feasible point $\hat{x}$ and a pair of dual feasible points $(\hat{\lambda}, \hat{\nu})$ satisfy the KKT conditions. That is,

$$\begin{aligned} f_i(\hat{x}) &\leq 0, i \in [m] \\ h_i(\hat{x}) &= 0, i \in [p] \\ \hat{\lambda}_i &\geq 0, i \in [m] \\ \hat{\lambda}_i f_i(\hat{x}) &= 0, i \in [m] \\ \nabla f_0(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla f_i(\hat{x}) + \sum_{i=1}^{p} \hat{\nu}_i \nabla h_i(\hat{x}) &= 0 \end{aligned}$$

Since $\hat{\lambda}_i \geq 0$, $L(x, \hat{\lambda}, \hat{\nu})$ is convex in x. The last KKT condition states that its gradient with respect to x vanishes at $x = \hat{x}$, so $\hat{x}$ minimizes $L(x, \hat{x}, \hat{\nu})$ over x. Therefore

$$\begin{aligned} g(\hat{\lambda}, \hat{\nu}) &= L(\hat{x}, \hat{\lambda}, \hat{\nu}) \\ &= f_0(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i f_i(\hat{x}) + \sum_{i=1}^{p} \hat{\nu}_i h_i(\hat{x}) \\ &= f_0(\hat{x}) \end{aligned}$$

The last equality holds because $\hat{\lambda}_i f_i(\hat{x}) = 0$ and $h_i(\hat{x}) = 0$. Hence, $(\hat{x}, \hat{\lambda}, \hat{\nu})$ is primal and dual optimal with zero duality gap. As a result, for any convex optimization problem with differentiable objective and constraint functions, any points that satisfy the KKT conditions are primal and dual optimal, and have zero duality gap. KKT conditions provide necessary and sufficient conditions for optimality. As for the second question, we introduce the Slater's condition first. A convex optimization problem is said to satisfy the Slater's condition if there exists an x in the relative interior of $\mathcal{D}$ such that it is strictly feasible; that is,

$$\begin{aligned} f_i(x) < 0, i \in [m] \\ h_i(x) = 0, i \in [p] \end{aligned} \tag{3.14}$$

The Slater's theorem states that if the Slater's condition holds for a convex optimization problem, then the strong duality holds. Hence, the Slater's condition is a sufficient condition for the strong duality. Furthermore, if the first k constraint functions $f_1, f_2, \cdots, f_k$ are affine, we can refine the Slater's condition so that $f_i(x)$ can equal zero for these k constraint functions. As a special case, refined Slater's condition reduces to feasibility when all the constraints are affine functions.

Based on the concept of strong duality, we may consider solving the primal problem 3.1 by dealing with the following alternative optimization problem

$$\sup_{\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p} \inf_{x \in \mathcal{D}} f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{3.15}$$

The reader can refer to [5] for references of more details of the theory of duality.

3.1.2 Chambolle and Pock's Primal-Dual Algorithm

Assume $A \in \mathbb{R}^{m \times n}$ and $F : \mathbb{R}^m \to (-\infty, \infty]$ and $G : \mathbb{R}^n \to (-\infty, \infty]$ are two convex functions. Primal-dual algorithm deals with

the following objective

$$\inf_{x\in\mathbb{R}^n} F(Ax) + G(x) \tag{3.16}$$

$$\equiv \inf_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} F(z) + G(x) \quad subject\ to\ Ax = z \tag{3.17}$$

We call 3.17 the primal problem. The associated Lagrange function is defined as

$$L(x, z, \xi) := F(z) + G(x) + \langle \xi, Ax - z \rangle \tag{3.18}$$

The Lagrange dual function is defined as

$$H(\xi) := \inf_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} L(x, z, \xi) = -F^*(\xi) - G^*(-A^*\xi) \tag{3.19}$$

where F^* and G^* denote the convex conjugate of the function F and G respectively.
Then we can come up with the dual problem defined as

$$\sup_{\xi\in\mathbb{R}^m} H(\xi) \tag{3.20}$$

x and z are called the primal variables and ξ is called the dual variable. By the strong duality, the optimal objective function values of the primal and dual problem are the same, i.e., if x^* and z^* are a pair of minimizers of the primal problem and ξ^* is a minimizer of the dual problem, then $F(z^*) + G(x^*) = H(\xi^*)$. Furthermore, as stated in theorem B.30 of [13], (x^*, ξ^*) is a solution to the following saddle-point problem

$$\inf_{x\in\mathbb{R}^n} \sup_{\xi\in\mathbb{R}^m} \langle Ax, \xi \rangle + G(x) - F^*(\xi) \tag{3.21}$$

The objective function of the saddle-point problem 3.21 is the fundamental concern of designing the primal-dual algorithm.

In the following, we will first explicitly describe the primal-dual algorithm introduced in the chapter 15 of [13] and explain it using the fixed-point iteration rules.[1]

[1] An optimization algorithm can be abstractly interpreted as a mapping $M : \Omega \to \Omega$, where Ω is the input domain and the output range of the mapping. A point $\theta^* \in \Omega$ is called a fixed point of the algorithm if $\theta^* = M(\theta^*)$. Fixed point iteration is a method of designing an algorithm

Algorithm 1 Primal-Dual Algorithm

Input: $A \in \mathbb{R}^{m\times n}$; convex functions $F : \mathbb{R}^m \to (-\infty, \infty]$ and $G : \mathbb{R}^n \to (-\infty, \infty]$
Parameters : $\theta \in [0, 1]$, $\tau, \sigma > 0$ such that $\|A\|^2_{2\to 2} < 1$
Initialization : $x^{(0)} \in \mathbb{R}^n$, $\xi^{(0)} \in \mathbb{R}^m$, $\bar{x}^{(0)} = x^{(0)}$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$\xi^{(n+1)} := P_{F^*}(\sigma, \xi^{(n)} + \sigma A\bar{x}^{(n)}) \qquad (PD_1)$$
$$x^{(n+1)} := P_G(\tau, x^{(n)} - \tau A^*\xi^{(n+1)}) \qquad (PD_2)$$
$$\bar{x}^{(n+1)} := x^{(n+1)} + \theta(x^{(n+1)} - x^{(n)}) \qquad (PD_3)$$

Output: $\xi^{(\bar{k})}$, $x^{(\bar{k})}$

Assume (x^*, ξ^*) is a solution pair of the saddle-point problem. On one hand,

$$\begin{aligned} \xi^* &= \underset{\xi\in\mathbb{R}^m}{argsup}\langle Ax^*, \xi\rangle - F^*(\xi) \\ &= \underset{\xi\in\mathbb{R}^m}{arginf} -\langle Ax^*, \xi\rangle + F^*(\xi) \end{aligned} \tag{3.22}$$

On the other hand,

$$x^* = \underset{x\in\mathbb{R}^n}{arginf}\langle Ax, \xi^*\rangle + G(x) \tag{3.23}$$

Because of 3.22, we can make the following derivations

$$\begin{aligned} &0 \in -Ax^* + \partial F^*(\xi^*) \\ \Rightarrow &\sigma Ax^* \in \sigma\partial F^*(\xi^*) \\ \Rightarrow &\xi^* + \sigma Ax^* \in \xi^* + \sigma\partial F^*(\xi^*) \end{aligned} \tag{3.24}$$

where σ is some positive constant
Because of 3.23, we can make the following derivations

$$\begin{aligned} &0 \in A^*\xi^* + \partial G(x^*) \\ \Rightarrow &-\tau A^*\xi^* \in \tau\partial G(x^*) \\ \Rightarrow &x^* - \tau A^*\xi^* \in x^* + \tau\partial G(x^*) \end{aligned} \tag{3.25}$$

where τ is some positive constant
The results of 3.24 and 3.25 indicate two equations that a minimizer pair (x^*, ξ^*) should satisfy. Based on those two equations, we can derive the following fixed-point iteration rules

$$\xi^{(n)} + \sigma A\bar{x}^{(n)} \in \xi^{(n+1)} + \sigma\partial F^*(\xi^{(n+1)}) \tag{3.26}$$

with the aim that the output of the algorithm will converge to a fixed point in the long run.

$$x^{(n)} - \tau A^* \xi^{(n+1)} \in x^{(n+1)} + \tau \partial G(x^{(n+1)}) \tag{3.27}$$

Equation 3.26 can be expressed as PD_1 (P_F^* is a proximal mapping associated with the function F^*) and equation 3.27 can be expressed as PD_2 (P_G is a proximal mapping associated with the function G), whereas PD_3 is trivially a fixed-point iteration

Finally, a practical stopping criterion can be based on the primal-dual gap

$$E(x, \xi) := F(Ax) + G(x) + G^*(-A^*\xi) + F^*(\xi) \tag{3.28}$$

We can terminate the algorithm once $E(x^{(k)}, \xi^{(k)}) \leq \eta$ for some positive prescribed tolerance η with the iterates of the algorithm being $x^{(k)}$ and $\xi^{(k)}$

3.1.3 The Augmented Lagrangian Method

Assume $A \in \mathbb{R}^{m \times n}$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a closed, proper and convex function. The augmented Lagrangian method deals with the following convex optimization problem

$$\inf_{x \in \mathbb{R}^n} f(x) \quad \textit{subject to } Ax = b \tag{3.29}$$

We can construct the dual problem of the primal problem 3.29 as follows

$$\begin{aligned} &\sup_{\lambda \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^n} f(x) + \langle \lambda, Ax - b \rangle \\ =& - \inf_{\lambda \in \mathbb{R}^m} \sup_{x \in \mathbb{R}^n} -f(x) - \langle \lambda, Ax - b \rangle \\ =& - \inf_{\lambda \in \mathbb{R}^m} d(\lambda) \end{aligned} \tag{3.30}$$

where $d(\lambda) := \sup_{x \in \mathbb{R}^n} -f(x) - \langle \lambda, Ax - b \rangle$. According to lemma 8 of [10], $d(\lambda)$ is proved to be closed and convex. Assume the primal problem 3.29 is feasible; that is, there exists an $x \in \mathbb{R}^n$ satisfying $Ax = b$. Then by the strong duality, we can solve the primal problem 3.29 by solving its dual problem 3.30.

Given any $\mu \in \mathbb{R}^m$, suppose d is proper; that is, it is finite for

at least one choice of $\lambda \in \mathbb{R}^m$, we can compute $\bar{x} = \underset{x\in\mathbb{R}^n}{argin}f\{f(x) + \langle\mu, Ax-b\rangle+\frac{c}{2}\|Ax-b\|_2^2\}$. Let $\lambda = \mu+c(A\bar{x}-b)$ and $\nu = b-A\bar{x} \in \mathbb{R}^m$. Then $\mu = \lambda + c\nu$ and it is proved in the proposition 9 of [10] that $\nu \in \partial d(\lambda)$. We can define the mapping $N_{cd}(\mu) = \lambda - c\nu = \mu + 2c(A\bar{x} - b)$. Since $d(\lambda)$ is proper, closed and convex, by theorem 1.9.1, N_{cd} is non-expansive and any fixed point of N_{cd} is a minimizer of d, i.e., an optimal solution to the dual problem 3.30.

Suppose for some scalar sequence $\{\rho_k\}$ satisfying $\inf_k\{\rho_k\} > 0$ and $\sup_k\{\rho_k\} < 2$. The sequences $\{x^{(k)}\} \subset \mathbb{R}^n$ and $\{\lambda^{(k)}\} \subset \mathbb{R}^m$ follow the iteration rules

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^n}{argin}f\{f(x) + \langle\lambda^{(k)}, Ax - b\rangle + \frac{c}{2}\|Ax - b\|_2^2\} \\ \lambda^{(k+1)} &= \frac{\rho_k}{2}N_{cd}(\lambda^{(k)}) + (1 - \frac{\rho_k}{2})\lambda^{(k)} \\ &= \frac{\rho_k}{2}(\lambda^{(k)} + 2c(Ax^{(k+1)} - b)) + (1 - \frac{\rho_k}{2})\lambda^{(k)} \\ &= \lambda^{(k)} + \rho_k c(Ax^{(k+1)} - b) \end{aligned} \tag{3.31}$$

If the dual problem 3.30 possesses an optimal solution, then by theorem 1.9.2, we know that $\{\lambda^{(k)}\}$ converges to one of them. Furthermore, as have been proved in the proposition 11 of [10], all limit points of $\{x^{(k)}\}$ are optimal solutions to the primal problem 3.29. We can describe the augmented Lagrangian method as the following algorithm If we set ρ_k

Algorithm 2 The Augmented Lagrangian Method

Input: $b \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; $f : \mathbb{R}^n \to \mathbb{R}$: closed, proper and convex
Parameter : $c > 0$
Initialization : $\lambda^{(0)} \in \mathbb{R}^m$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^N}{argin}f\{f(x) + \langle\lambda^{(k)}, Ax - b\rangle + \frac{c}{2}\|Ax - b\|_2^2\} \\ \lambda^{(k+1)} &= \lambda^{(k)} + \rho_k c(Ax^{(k+1)} - b) \quad where\ 0 < \rho_k < 2 \end{aligned}$$

Output: $x^{(\bar{k})}$ and $\lambda^{(\bar{k})}$

to be 1 at each iteration, then the iteration rules 3.31 will become

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^n}{argin}f\{f(x) + \langle\lambda^{(k)}, Ax - b\rangle + \frac{c}{2}\|Ax - b\|_2^2\} \\ \lambda^{(k+1)} &= \lambda^{(k)} + c(Ax^{(k+1)} - b) \end{aligned} \tag{3.32}$$

In this case, the augmented Lagrangian method minimizes the augmented Lagrangian $L_c(x, \lambda) = f(x) + \langle\lambda, Ax - b\rangle + \frac{c}{2}\|Ax - b\|_2^2$ with respect to x followed by a maximization over the dual variable λ. Note that the iteration rule of λ is called the dual ascent step, which can be seen as a fixed point iteration.

3.1.4 Alternating Direction Method of Multipliers (ADMM)

Assume $A \in \mathbb{R}^{m\times n}$, $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ are two convex functions. ADMM deals with the following convex optimization problem

$$\begin{aligned} &\inf_{x\in\mathbb{R}^n} f(x) + g(Mx) \\ \equiv &\inf_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} f(x) + g(z) \quad subject\ to\ Mx = z \end{aligned} \tag{3.33}$$

We can construct the dual problem of the primal problem 3.33 as follows

$$\begin{aligned} &\sup_{\lambda\in\mathbb{R}^m}[\inf_{x\in\mathbb{R}^n}\{f(x) + \langle\lambda, Mx\rangle\} + \inf_{z\in\mathbb{R}^m}\{g(z) + \langle\lambda, -z\rangle\}] \\ =& -\inf_{\lambda\in\mathbb{R}^m}[\sup_{x\in\mathbb{R}^n}\{-f(x) - \langle\lambda, Mx\rangle\} + \sup_{z\in\mathbb{R}^m}\{-g(z) - \langle\lambda, -z\rangle\}] \\ =& -\inf_{\lambda\in\mathbb{R}^m} d_1(\lambda) + d_2(\lambda) \end{aligned} \tag{3.34}$$

where $d_1(\lambda) := \sup_{x\in\mathbb{R}^n}\{-f(x) - \langle\lambda, Mx\rangle\}$ and $d_2(\lambda) := \sup_{z\in\mathbb{R}^m}\{-g(z) - \langle\lambda, -z\rangle\}$. According to lemma 8 of [10], $d_1(\lambda)$ and $d_2(\lambda)$ are both closed and convex. By the strong duality, we can solve the primal problem 3.33 by solving its dual problem 3.34.

Given $y^{(0)} \in \mathbb{R}^m$, a scalar sequence $\{\rho_k\}$ satisfying $\inf_k\{\rho_k\} > 0$ and $\sup_k\{\rho_k\} < 2$, we want to produce the iteration rule : $y^{(k+1)} =$

$\frac{\rho_k}{2}N_{cd_1}(N_{cd_2}(y^{(k)}))+(1-\frac{\rho_k}{2})y^{(k)}$. In this way, by theorem 1.9.2, $\{y^{(k)}\}$ can converge to a fixed point $y^{(\infty)}$ of $N_{cd_1}\circ N_{cd_2}$. Then by theorem 1.9.1, we can find a minimizer $\lambda^{(\infty)}$ of d_1+d_2, where $y^{(\infty)}=\lambda^{(\infty)}+c\nu^{(\infty)}$ and $\nu^{(\infty)}\in\partial d_1(\lambda^{(\infty)})$, $-\nu^{(\infty)}\in\partial d_2(\lambda^{(\infty)})$. We can produce such iteration rule as the following four steps

1. express $y^{(k)}$ as $y^{(k)}=\lambda^{(k)}+c\nu^{(k)}$, where $\nu^{(k)}\in\partial d_2(\lambda^{(k)})$
2. apply the mapping N_{cd_2}
$$y^{(k+1)}=\frac{\rho_k}{2}N_{cd_1}(\lambda^{(k)}-c\nu^{(k)})+(1-\frac{\rho_k}{2})(\lambda^{(k)}+c\nu^{(k)})$$
3. express $\lambda^{(k)}-c\nu^{(k)}$ as $\lambda^{(k)}-c\nu^{(k)}=\mu^{(k)}+cw^{(k)}$, where $w^{(k)}\in\partial d_1(\mu^{(k)})$
4. apply the mapping N_{cd_1}
$$\begin{aligned}y^{(k+1)}&=\frac{\rho_k}{2}(\mu^{(k)}-cw^{(k)})+(1-\frac{\rho_k}{2})(\lambda^{(k)}+c\nu^{(k)})\\&=\rho_k\mu^{(k)}+(1-\rho_k)\lambda^{(k)}+c\nu^{(k)}\end{aligned}$$

We can further organize the four steps into the following two steps. Given $\lambda^{(0)},\nu^{(0)}\in\mathbb{R}^m$ such that $y^{(0)}=\lambda^{(0)}+c\nu^{(0)}$.

1. Find $\mu^{(k)},w^{(k)}\in\mathbb{R}^m$ such that $\lambda^{(k)}-c\nu^{(k)}=\mu^{(k)}+cw^{(k)}$ and $w^{(k)}\in\partial d_1(\mu^{(k)})$
2. Find $\lambda^{(k+1)},\nu^{(k+1)}\in\mathbb{R}^m$ such that $y^{(k+1)}=\lambda^{(k+1)}+c\nu^{(k+1)}=\rho_k\mu^{(k)}+(1-\rho_k)\lambda^{(k)}+c\nu^{(k)}$ and $\nu^{(k+1)}\in\partial d_2(\lambda^{(k+1)})$

By the proposition 9 of [10], we can implement step 1 as follows

$$\begin{aligned}x^{(k+1)}&=\underset{x\in\mathbb{R}^n}{argin}f\{f(x)+\langle\lambda^{(k)}-c\nu^{(k)},Mx\rangle+\frac{c}{2}\|Mx\|_2^2\}\\\mu^{(k)}&=\lambda^{(k)}-c\nu^{(k)}+cMx^{(k+1)}\\w^{(k)}&=-Mx^{(k+1)}\end{aligned}$$

and we can implement step 2 as follows

$$z^{(k+1)} = \underset{z\in\mathbb{R}^m}{argin}f\{g(z) + \langle \rho_k\mu^{(k)} + (1-\rho_k)\lambda^{(k)} + c\nu^{(k)}, -z\rangle + \frac{c}{2}\| - z\|_2^2\}$$
$$\lambda^{(k+1)} = \rho_k\mu^{(k)} + (1-\rho_k)\lambda^{(k)} + c\nu^{(k)} - cz^{(k+1)}$$
$$\nu^{(k+1)} = z^{(k+1)}$$

We can summarize them as follows

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^n}{argin}f\{f(x) + \langle \lambda^{(k)} - cz^{(k)}, Mx\rangle + \frac{c}{2}\|Mx\|_2^2\} \\ &= \underset{x\in\mathbb{R}^n}{argin}f\{f(x) + \langle \lambda^{(k)}, Mx\rangle + \frac{c}{2}\|Mx - z^{(k)}\|_2^2\} \end{aligned}$$
$$\mu^{(k)} = \lambda^{(k)} - cz^{(k)} + cMx^{(k+1)}$$
$$z^{(k+1)} = \underset{z\in\mathbb{R}^m}{argin}f\{g(z) + \langle \rho_k\mu^{(k)} + (1-\rho_k)\lambda^{(k)} + cz^{(k)}, -z\rangle + \frac{c}{2}\|z\|_2^2\}$$
$$\lambda^{(k+1)} = \rho_k\mu^{(k)} + (1-\rho_k)\lambda^{(k)} + cz^{(k)} - cz^{(k+1)}$$

Since $\mu^{(k)} = \lambda^{(k)} - cz^{(k)} + cMx^{(k+1)}$,

$$\begin{aligned} &\rho_k\mu^{(k)} + (1-\rho_k)\lambda^{(k)} + cz^{(k)} \\ =&\rho_k(\lambda^{(k)} - cz^{(k)} + cMx^{(k+1)}) + (1-\rho_k)\lambda^{(k)} + cz^{(k)} \\ =&\lambda^{(k)} + c(\rho_k Mx^{(k+1)} + (1-\rho_k)z^{(k)}) \end{aligned}$$

we can further summarize them as follows.

$$x^{(k+1)} = \underset{x\in\mathbb{R}^n}{argin}f\{f(x) + \langle \lambda^{(k)}, Mx\rangle + \frac{c}{2}\|Mx - z^{(k)}\|_2^2\} \tag{3.35}$$
$$z^{(k+1)} = \underset{z\in\mathbb{R}^m}{argin}f\{g(z) - \langle \lambda^{(k)}, z\rangle + \frac{c}{2}\|\rho_k Mx^{(k+1)} + (1-\rho_k)z^{(k)} - z\|_2^2\} \tag{3.36}$$
$$\lambda^{(k+1)} = \lambda^{(k)} + c(\rho_k Mx^{(k+1)} + (1-\rho_k)z^{(k)} - z^{(k+1)}) \tag{3.37}$$

We describe ADMM as the following algorithm. If we set ρ_k to be 1 at each iteration and add some redundant terms, 3.35, 3.36 and 3.37 will

Algorithm 3 ADMM

Input: $A : \mathbb{R}^{m\times n}$; $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ are two convex functions
Parameter : $c > 0$
Initialization : $\lambda^{(0)} \in \mathbb{R}^m$ and $z^{(0)} \in \mathbb{R}^m$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = \underset{x\in\mathbb{R}^n}{arginf}\{f(x) + \langle\lambda^{(k)}, Mx\rangle + \frac{c}{2}\|Mx - z^{(k)}\|_2^2\}$$
$$z^{(k+1)} = \underset{z\in\mathbb{R}^m}{arginf}\{g(z) - \langle\lambda^{(k)}, z\rangle + \frac{c}{2}\|\rho_k Mx^{(k+1)} + (1-\rho_k)z^{(k)} - z\|_2^2\} \quad where\ 0 < \rho_k < 2$$
$$\lambda^{(k+1)} = \lambda^{(k)} + c(\rho_k Mx^{(k+1)} + (1-\rho_k)z^{(k)} - z^{(k+1)})$$

Output: $x^{(\bar{k})}$, $z^{(\bar{k})}$ and $\lambda^{(\bar{k})}$

become

$$x^{(k+1)} = \underset{x\in\mathbb{R}^n}{arginf}\{f(x) + g(z^{(k)}) + \langle\lambda^{(k)}, Mx - z^{(k)}\rangle + \frac{c}{2}\|Mx - z^{(k)}\|_2^2\} \tag{3.38}$$

$$z^{(k+1)} = \underset{z\in\mathbb{R}^m}{arginf}\{f(x^{(k+1)}) + g(z) + \langle\lambda^{(k)}, Mx^{(k+1)} - z\rangle + \frac{c}{2}\|Mx^{(k+1)} - z\|_2^2\} \tag{3.39}$$

$$\lambda^{(k+1)} = \lambda^{(k)} + c(Mx^{(k+1)} - z^{(k+1)}) \tag{3.40}$$

In this case, ADMM minimizes the augmented Lagrangian $L_c(x, z, \lambda) = f(x) + g(z) + \langle\lambda, Mx - z\rangle + \frac{c}{2}\|Mx - z\|_2^2$ with respect to x and z sequentially followed by a maximization over the dual variable λ. Note that as we can see from algorithm 3, ADMM has a merit that it decouples the composite of the functions f and g, which enables us to exploit the individual structure of the two functions.

Up to now, we find that the derivation of the ADMM is mainly an application of the fixed-point algorithm to the non-expansive mapping $N_{cd_1} \circ N_{cd_2}$. Proposition 15 of [10] gives a theoretical guarantee as follows

Theorem 3.1.1. *Suppose there exists some optimal primal-dual solution pair $((x^*, z^*), \lambda^*)$ to the original problem such that*

1. *x^* minimizes $f(x) + \langle\lambda^*, Mx\rangle$ with respect to x*
2. *z^* minimizes $g(z) - \langle\lambda^*, z\rangle$ with respect to z*
3. *$Mx^* = z^*$*

Assume also that all subgradients of the function $d_1(\lambda) = \sup_{x \in \mathbb{R}^n} \{-f(x) - \langle \lambda, Mx \rangle\}$ *at each point* $\lambda \in \mathbb{R}^m$ *take the form* $-M\bar{x}$*, where* $\bar{x}$ *attains the stated supremum over* x*. Then if the sequences* $\{x^{(k)}\} \subset \mathbb{R}^n$*,* $\{z^{(k)}\} \subset \mathbb{R}^m$*,* $\{\lambda^{(k)}\} \subset \mathbb{R}^m$ *follow the iteration rules, where* $\inf_k\{\rho_k\} > 0$ *and* $\sup_k\{\rho_k\} < 2$*, then* $\lambda^{(k)}$ *converges to* $\lambda^{(\infty)}$*,* $z^{(k)}$ *converges to* $z^{(\infty)}$*, and* $Mx^{(k)}$ *converges to* $Mx^{(\infty)} = z^{(\infty)}$*, where* $((x^{(\infty)}, z^{(\infty)}), \lambda^{(\infty)})$ *is an optimal primal-dual solution pair to the original problem.*

3.2 "Majorization-Minimization" Viewpoint

In this section, we will introduce the Bregman proximal gradient method and some methods that are special cases of it. Assume $f : \mathbb{R}^n \to (-\infty, \infty]$ is a convex and differentiable function and $g : \mathbb{R}^n \to (-\infty, \infty]$ is a convex function. The Bregman proximal gradient method deals with the following objective

$$\inf_{x \in \mathbb{R}^n} F(x) = f(x) + g(x) \tag{3.41}$$

Algorithm 4 Bregman Proximal Gradient Method

Input: $f : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
$g : \mathbb{R}^n \to (-\infty, \infty]$:a convex function
$h : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
Parameters : $\eta > 0$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = \underset{x \in \mathbb{R}^n}{arginf}\, f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{\eta} D_h(x, x^{(k)}) + g(x) \tag{3.42}$$

Output: $x^{(\bar{k})}$

The Bregman proximal gradient method is described as algorithm 4 as above ($D_h(x, x^{(k)})$ is the Bregman divergence associated with h). Assume f is M-smooth relative to h, then

$$F(x) = f(x) + g(x) \leq f(x^{(k)}) + \langle \nabla f(x^{(k)}, x - x^{(k)} \rangle + M D_h(x, x^{(k)}) + g(x)$$

Hence, if we choose the parameter η to be $1/M$, then the iteration 3.42 is just to minimize the upper bound of the objective function $F(x)$. Such idea of designing an optimization algorithm is called " majorization-minimization". In the following, we will introduce six special cases of the Bregman proximal gradient method. Readers can refer to [2],[22] and [36].

3.2.1 Proximal Gradient Method

This is a special case of the Bregman proximal gradient method when $h(x) = \frac{1}{2}\|x\|_2^2$. Hence, 3.42 becomes

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^n}{argin f}\ f(x^{(k)}) + \nabla f(x^{(k)})^T(x - x^{(k)}) + \frac{1}{2\eta}\|x - x^{(k)}\|_2^2 + g(x) \\ &= (I + \eta\partial g)^{-1}(I - \eta\nabla f)(x^{(k)}) \\ &= P_g(\eta, x^{(k)} - \eta\nabla f(x^{(k)})) \end{aligned}$$

The proximal gradient method is described as the following algorithm.

Algorithm 5 Proximal Gradient Method

Input: $f : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
$g : \mathbb{R}^n \to (-\infty, \infty]$:a convex function
Parameters : $\eta > 0$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = P_g(\eta, x^{(k)} - \eta\nabla f(x^{(k)})) \tag{3.43}$$

Output: $x^{(\bar{k})}$

We make a remark that the proximal gradient method is the same as the forward-backward splitting method. We can split 3.43 into a forward step and a backward step. The forward step is $z^{(k)} = x^{(k)} - \eta\nabla f(x^{(k)})$. It is a gradient descent method which will be introduced later. We can think of it as forwarding to $z^{(k)}$ from the current value $x^{(k)}$. The backward step is $x^{(k+1)} = P_g(\eta, z^{(k)})$. It is a proximal point method which will also be introduced later. Since $x^{(k+1)} = (I + \eta\partial g)^{-1}(z^{(k)})$, $z^{(k)} = (I + \eta\partial g)(x^{(k+1)})$. We can think of it as backwarding to $z^{(k)}$ from the next value $x^{(k+1)}$.

3.2.2 Proximal Point Method

The proximal point method is a special case of the Bregman proximal gradient method when $h(x) = \frac{1}{2}\|x\|_2^2$ and $f(x) = 0$ (or simply, a special case of the proximal gradient method when $f(x) = 0$). Hence, 3.42 becomes

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^n}{argin}f\ g(x) + \frac{1}{2\eta}\|x - x^{(k)}\|_2^2 \\ &= (I + \eta\partial g)^{-1}(x^{(k)}) \\ &= P_g(\eta, x^{(k)}) \end{aligned}$$

The proximal point method is described as the following algorithm.

Algorithm 6 Proximal Point Method

Input: $g : \mathbb{R}^n \to (-\infty, \infty]$: a convex function
Parameters : $\eta > 0$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = P_g(\eta, x^{(k)}) \tag{3.44}$$

Output: $x^{(\bar{k})}$

3.2.3 Mirror Descent Method

Mirror descent method is a special case of the Bregman proximal gradient method when $g(x)$ is the characteristic function for a non-empty, closed and convex set $\mathcal{X} \subseteq \mathbb{R}^n$. Therefore, the algorithm can be described as follows

Algorithm 7 Mirror Descent Method

Input: $f : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
$h : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
$\mathcal{X} \subseteq \mathbb{R}^n$: a non-empty, closed and convex set
Parameters : $\eta > 0$
Initialization : $x^{(0)} \in \mathcal{X}$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = \underset{x\in\mathcal{X}}{argin}f\ f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)}\rangle + \frac{1}{\eta}D_h(x, x^{(k)}) \tag{3.45}$$

Output: $x^{(\bar{k})}$

3.2.4 Projected Gradient Descent Method

Projected gradient descent method is a special case of the Bregman proximal gradient method when $h(x) = \frac{1}{2}\|x\|_2^2$ and $g(x)$ is the

characteristic function for a non-empty, closed and convex set $\mathcal{X} \subseteq \mathbb{R}^n$ (or simply, a special case of the mirror descent method when $h(x) = \frac{1}{2}\|x\|_2^2$). Hence, 3.42 becomes

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathcal{X}}{argmin}\, f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\eta}\|x - x^{(k)}\|_2^2 \\ &= \underset{x\in\mathcal{X}}{argmin}\, \frac{1}{2}\|x - (x^{(k)} - \eta\nabla f(x^{(k)}))\|_2^2 \\ &= proj_{\mathcal{X}}(x^{(k)} - \eta\nabla f(x^{(k)})) \end{aligned}$$

where $proj_{\mathcal{X}}(\cdot)$ is the projection operator defined in 1.21. We can describe the projected gradient descent method as the following algorithm

Algorithm 8 Projected Gradient Descent Method

Input: $f : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
$\mathcal{X} \subseteq \mathbb{R}^n$: a non-empty, closed and convex set
Parameters : $\eta > 0$
Initialization : $x^{(0)} \in \mathcal{X}$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = proj_{\mathcal{X}}(x^{(k)} - \eta\nabla f(x^{(k)})) \tag{3.46}$$

Output: $x^{(\bar{k})}$

3.2.5 Entropy Mirror Descent Method

Entropy mirror descent method is a special case of the Bregman proximal gradient method when $h(x)$ is the negative entropy and $g(x)$ is the characteristic function for the probability simplex Δ (a vector belonging to Δ has the sum of its values of entries equal to one). Hence, 3.42 becomes

$$x^{(k+1)} = \underset{x\in\Delta}{argin f}\, f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{\eta}\sum_{i=1}^{n} x[i] \log \frac{x[i]}{(x^{(k)})[i]}$$

$$\Rightarrow (x^{(k+1)})[i] = \frac{(x^{(k)})[i]e^{-\eta(\nabla f(x^{(k)})[i]}}{\sum_{i=1}^{n}(x^{(k)})[i]e^{-\eta(\nabla f(x^{(k)})[i]}} \quad \forall i \in [n]$$

Therefore, we can describe the entropy mirror descent method as the following algorithm

Algorithm 9 Entropy Mirror Descent Method

Input: $f : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
Parameters : $\eta > 0$
Initialization : $x^{(0)} \in \Delta$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$(x^{(k+1)})[i] = \frac{(x^{(k)})[i]e^{-\eta(\nabla f(x^{(k)})[i]}}{\sum_{i=1}^{n}(x^{(k)})[i]e^{-\eta(\nabla f(x^{(k)})[i]}} \quad \forall i \in [n] \tag{3.47}$$

Output: $x^{(\bar{k})}$

3.2.6 Gradient Descent Method

Gradient descent method is a special case of the Bregman proximal gradient method when $h(x) = \frac{1}{2}\|x\|_2^2$ and $g(x) = 0$. Hence, 3.42 becomes

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^n}{argmin} f\ f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\eta}\|x - x^{(k)}\|_2^2 \\ &= x^{(k)} - \eta \nabla f(x^{(k)}) \end{aligned}$$

Hence, we can describe the gradient descent method as the following algorithm

Algorithm 10 Gradient Descent Method

Input: $f : \mathbb{R}^n \to (-\infty, \infty]$: a convex and differentiable function
Parameters : $\eta > 0$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = x^{(k)} - \eta \nabla f(x^{(k)}) \tag{3.48}$$

Output: $x^{(\bar{k})}$

We can make a connection between the gradient descent method and the proximal point method. Let g be a convex function. Since g is not necessarily differentiable, we cannot apply gradient descent method directly to g. However, the Moreau envelope g_η is convex, differentiable and ∇g_η approximates ∂g (the reader can refer to section 1.8). We may consider applying the gradient descent method to g_η instead. In this

way, 3.48 becomes

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \eta \nabla g_\eta(x^{(k)}) \\ &= x^{(k)} - \eta[\frac{1}{\eta}(I - (I + \eta \partial g)^{-1})(x^{(k)})] \\ &= (I + \eta \partial g)^{-1}(x^{(k)}) \end{aligned}$$

which coincides with the iteration rule 3.44 of the proximal point method.

3.3 "Minimization of First/Second Order Approximation" Viewpoint

In this section, our main concern is the steepest descent method, which is designed based on the methodology of minimizing the first-order approximation of the objective function. Gradient descent method and Newton's method can both be viewed as special cases of the steepest descent method. Furthermore, Newton's method can also be interpreted as a method of minimizing the second-order approximation of the objective function. Before introducing the steepest descent method, we need to go through the concept of a general descent algorithm and the idea of line search. Finally, we close this section by discussing how to solve an equality-constrained minimization problem, including an extension of the Newton's method to solve it. The readers can refer to [5] for references.

3.3.1 General Descent Algorithm

Suppose the objective function $f : \mathbb{R}^n \to \mathbb{R}$ of an unconstrained minimization problem is convex and twice differentiable. Denote x^* as an optimal point (i.e., minimizer) of the minimization problem and $p^* = f(x^*)$ as the optimal value. Let $x^{(0)}, x^{(1)}, \cdots \in dom\, f$ be a minimizing sequence; that is, $f(x^{(k+1)}) \leq f(x^{(k)})\ \forall k = 0, 1, \cdots$ and $f(x^{(k)}) \to p^*$ as $k \to \infty$. For descent methods, minimizing sequences

are chosen to be

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)} \tag{3.49}$$

$\Delta x^{(k)} \in \mathbb{R}^n$ is called a step or search direction and $t^{(k)} \geq 0$ is called a step size. From convexity of f, we know that

$$f(x^{(k+1)}) \geq f(x^{(k)}) + \nabla f(x^{(k)})^T(x^{(k+1)} - x^{(k)}) = f(x^{(k)}) + t^{(k)}\nabla f(x^{(k)})^T\Delta x^{(k)}$$

Since $f(x^{(k+1)}) \leq f(x^{(k)})$, $\nabla f(x^{(k)})^T\Delta x^{(k)}$ should be smaller than or equal to zero. Such search direction is called a descent direction of f at $x^{(k)}$. A general descent algorithm can be described as follows.

Algorithm 11 General Descent Algorithm

Input: f : $\mathbb{R}^n \to \mathbb{R}$: convex and twice differentiable
Initialization : $x^{(0)} \in dom\, f$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

determine a descent direction $\Delta x^{(k)}$
line search : choose a step size $t^{(k)} \geq 0$
update : $x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$

Output: $x^{(\bar{k})}$

Different ways of determining a descent direction correspond to different descent algorithms. We cover this issue in the next subsection. In the following, we introduce two line search methods. The first one is called exact line search.

$$t^{(k)} = \underset{s \geq 0}{arginf}\, f(x^{(k)} + s\Delta x^{(k)}) \tag{3.50}$$

The second one is called backtracking line search. It can be described as the following algorithm

Algorithm 12 Backtracking Line Search

Input: $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$
Initialization : $t^{(k)} = 1$
  while $f(x^{(k)}) + t^{(k)}\Delta x^{(k)}) > f(x^{(k)}) + \alpha t^{(k)}\nabla f(x^{(k)})^T\Delta x^{(k)}$ **do**
    $t^{(k)} := \beta t^{(k)}$
  end while
Output: : $t^{(k)}$

We can deduce from figure 3.1 that the output $t^{(k)}$ will be 1 or belong to $(\beta t_0, t_0)$. Also note that $f(x^{(k)} + t^{(k)}\Delta x^{(k)}) \approx f(x^{(k)}) + t^{(k)}\nabla f(x^{(k)})^T\Delta x^{(k)} < f(x^{(k)}) + \alpha\nabla f(x^{(k)})^T\Delta x^{(k)}$ for t small enough.

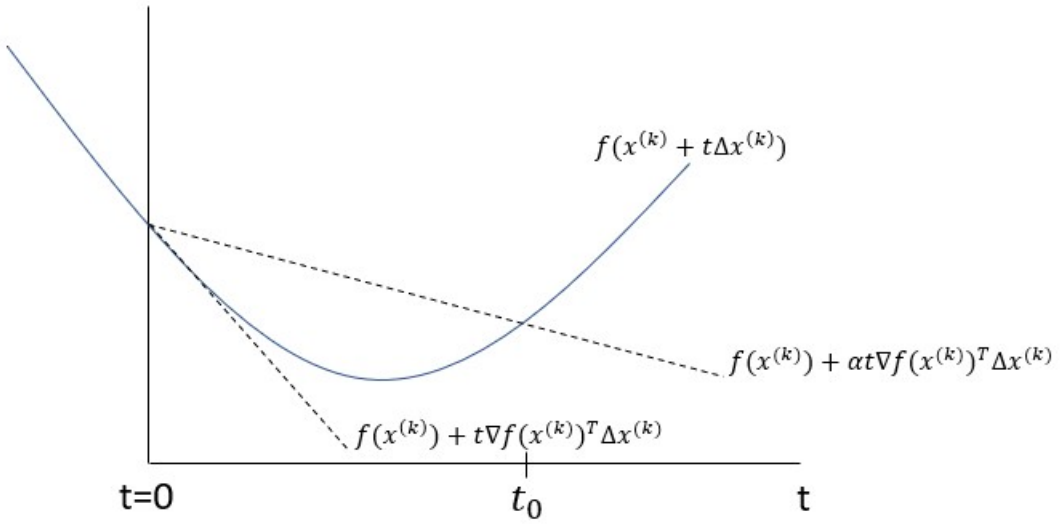

Figure 3.1: backtracking line search

Hence, the backtracking line search is guaranteed to terminate eventually.

3.3.2 Steepest Descent Method

First, we introduce the normalized steepest descent method. It is a descent algorithm whose descent direction is determined by

$$\begin{aligned} \Delta x_{nsd}^{(k)} &= \underset{v}{argin}f\{f(x^{(k)}) + \nabla f(x^{(k)})^T v \mid \|v\| = 1\} \\ &= \underset{v}{argin}f\{\nabla f(x^{(k)})^T v \mid \|v\| = 1\} \end{aligned} \tag{3.51}$$

where $\|\cdot\|$ denotes some norm of a vector. Note that because $\nabla f(x^{(k)})^T v$ is a linear function of v, we need to add some constraint on v (i.e., $\|v\| = 1$) or there will be no infimum. By Taylor series expansion, we can approximate $f(x)$ as $f(x^{(k)}) + \nabla f(x^{(k)})^T (x - x^{(k)})$ when x is near $x^{(k)}$. Hence, $\Delta x_{nsd}^{(k)}$ can be interpreted as a search direction that minimizes the first order approximation of $f(x)$ at $x^{(k)}$. Furthermore, it is indeed a descent direction because $\nabla f(x)^T \Delta x_{nsd} = -\|\nabla f(x)\|_* \leq 0$ ($\|\cdot\|_*$ denotes the dual norm of a vector with respect to norm $\|\cdot\|$). The descent direction of the steepest descent method is obtained by multiplying $\Delta x_{nsd}^{(k)}$ by $\|\nabla f(x^{(k)})\|_*$. That is,

$$\Delta x_{sd}^{(k)} = \|\nabla f(x^{(k)})\|_* \Delta x_{nsd}^{(k)} \tag{3.52}$$

Once again, we can verify that $\Delta x_{sd}^{(k)}$ is indeed a descent direction because $\nabla f(x^{(k)})^T \Delta x_{sd} = -\|\nabla f(x^{(k)})\|_*^2 \leq 0$

If we choose the norm to be Euclidean ℓ_2 norm, then the steepest descent coincides with the gradient descent method whose descent direction is

$$\Delta x_{gd}^{(k)} = -\nabla f(x^{(k)}) \tag{3.53}$$

Suppose f is m-strongly convex, then as have been introduced in section 1.2.

$$f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|_2^2 \leq f(y)$$
$$\Rightarrow f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \leq p^*$$

If $\|\nabla f(x)\|_2 \leq \sqrt{2m\epsilon}$ for some prescribed threshold $\epsilon > 0$, then $f(x) - p^* \leq \epsilon$. Hence, a natural stopping criterion can be set as $\|\nabla f(x)\|_2 \leq \eta$ for some prescribed threshold $\eta > 0$. Furthermore, if f is also M-smooth, it can be proved that (see [5])

$$f(x^{(k)}) - p^* \leq c^k (f(x^{(0)} - p^*) \tag{3.54}$$

For exact line search, $c = 1 - \frac{m}{M} < 1$ and for backtracking line search, $c = 1 - \min\{2m\alpha, 2\beta\alpha\frac{m}{M}\} < 1$. Hence, the convergence rate of the gradient descent method depends greatly on the condition number M/m of the Hessian matrix $\nabla^2 f(x)$.

If we choose the norm to be the quadratic P norm defined as

$$\|z\|_P = (z^T P z)^{1/2} = \|P^{1/2} z\|_2 \tag{3.55}$$

where $z \in \mathbb{R}^n$ is a vector and $P \in \mathbb{R}^{n \times n}$ is a positive definite matrix. The descent direction of the normalized steepest descent for the quadratic P norm can be derived as

$$\Delta x_{nsd}^{(k)} = -(\nabla f(x^{(k)})^T P^{-1} \nabla f(x^{(k)}))^{-1/2} P^{-1} \nabla f(x^{(k)}) \tag{3.56}$$

and that of the steepest descent for the quadratic P norm is

$$\Delta x_{sd}^{(k)} = -P^{-1} \nabla f(x^{(k)}) \tag{3.57}$$

The requirement of a search direction to be a descent direction rationalizes the positive definiteness of P ($\because \nabla f(x^{(k)})^T \Delta x_{sd}^{(k)}$ $= -\nabla f(x^{(k)})^T P^{-1} \nabla f(x^{(k)}) < 0$ if and only if P is positive definite). We want to make an important connection between the steepest descent for the quadratic P norm and the gradient descent in the following. Let $\bar{x} = P^{1/2}x$ and $\bar{f}(\bar{x}) = f(P^{-1/2}\bar{x}) = f(x)$. Suppose we apply the gradient descent method to $\bar{f}(\bar{x})$. The descent direction will be

$$\Delta \bar{x}_{gd}^{(k)} = -\nabla \bar{f}(\bar{x}^{(k)}) = -P^{-1/2} \nabla f(P^{-1/2}\bar{x}^{(k)}) = -P^{-1/2} \nabla f(x^{(k)})$$

and the update step will become

$$\begin{aligned} \bar{x}^{(k+1)} &= \bar{x}^{(k)} + t^{(k)} \Delta \bar{x}_{gd}^{(k)} \\ \Rightarrow x^{(k+1)} &= x^{(k)} + t^{(k)}(-P^{-1} \nabla f(x^{(k)})) \\ &= x^{(k)} + t^{(k)} \Delta x_{sd}^{(k)} \end{aligned}$$

Therefore, the steepest descent method in the quadratic P norm is equivalent to the gradient descent method applied to the problem after the change of coordinates $\bar{x} = P^{1/2}x$. The Hessian matrix after the associated change of coordinates is

$$\begin{aligned} \nabla^2 \bar{f}(\bar{x}^{(k)}) &= (P^{-1/2})^T \nabla^2 f(p^{-1/2}\bar{x}^{(k)}) P^{-1/2} \\ &= (P^{-1/2})^T \nabla^2 f(x^{(k)}) P^{-1/2} \end{aligned}$$

Since the convergence rate of the gradient method depends on the condition number of the Hessian matrix, the steepest descent method in the quadratic P norm converges very rapidly when the condition number of the Hessian matrix $(P^{-1/2})^T \nabla^2 f(x^{(k)}) P^{-1/2}$ is small.

The Newton's method is a special case of the steepest descent for the quadratic P norm; however, instead of using a fixed positive definite matrix P for every iterations, the Newton's method uses different P matrices depending on $x^{(k)}$ for every iterations. Concretely speaking, the Newton's method uses the matrix $\nabla^2 f(x^{(k)})$ for each iteration. Hence the descent direction of the Newton's method is defined as

$$\Delta x_{nt}^{(k)} = -(\nabla^2 f(x^{(k)}))^{-1} \nabla f(x^{(k)}) \tag{3.58}$$

$\Delta x_{nt}^{(k)}$ is also called the Newton step of f at $x^{(k)}$. Note that the Hessian matrix $\nabla^2 f(x)$ of f should be positive definite so that $\Delta x_{nt}^{(k)}$ is indeed a descent direction. Also note that since the Hessian matrix of each iteration, which is $(\nabla^2 f(x^{(k)})^{-1/2})\nabla^2 f(x^{(k)})\nabla^2 f(x^{(k)})^{-1/2} = I$, has the smallest condition number one, we can expect the Newton's method converges very rapidly. Aside from interpreting the Newton's method as a special case of the steepest descent, there are two other important interpretations of the Newton's method. According to the optimality condition introduced in section 1.1, we know that $\nabla f(x^*) = 0$ at the optimal point x^* (the equality holds because there are no constraints on x, i.e., $x \in \mathbb{R}^N$). We hope that $x^{(k)}+v$, where v is a small perturbation from $x^{(k)}$, can be an optimal point. Hence, we set the gradient vector of f at $x^{(k)}+v$ to be zero; that is, $\nabla f(x^{(k)}+v) = 0$. Since v is a small perturbation from $x^{(k)}$, we can linearize the gradient vector so that

$$\nabla f(x^{(k)} + v) \approx \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})v = 0$$

In this way, $v = -\nabla^2 f(x^{(k)})^{-1}\nabla f(x^{(k)})$, which is just the Newton step. The Newton step can also be verified as a vector that minimizes the second-order approximation of f at $x^{(k)}$, i.e.,

$$\Delta x_{nt}^{(k)} = \underset{v}{argin}f\ f(x^{(k)}) + \nabla f(x^{(k)})^T v + \frac{1}{2}v^T\nabla^2 f(x^{(k)})v \qquad (3.59)$$

Denote the second-order approximation as

$$\hat{f}_k(v) = f(x^{(k)}) + \nabla f(x^{(k)})^T v + \frac{1}{2}v^T\nabla^2 f(x^{(k)})v \qquad (3.60)$$

We find that an estimate of the distance from $f(x^{(k)})$ to the optimal

value p^* based on the second-order approximation is

$$\begin{aligned}
&f(x^{(k)}) - p^* \\
=& f(x^{(k)}) - \inf_x f(x) \\
\approx& f(x^{(k)}) - \inf_v f(x^{(k)}) + \nabla f(x^{(k)})^T v + \frac{1}{2} v^T \nabla^2 f(x^{(k)}) v \\
=& f(x^{(k)}) - \hat{f}_k(\Delta x_{nt}^{(k)}) \\
=& \frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)
\end{aligned}$$

We define the Newton decrement at $x^{(k)}$ as

$$\lambda(x^{(k)}) := (\nabla f(x^{(k)})^T \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}))^{1/2} \tag{3.61}$$

Hence, $f(x^{(k)}) - p^* = \frac{1}{2}\lambda(x^{(k)})$. A natural stopping criterion can be set as $\frac{1}{2}\lambda^2(x^{(k)}) \leq \epsilon$ for some prescribed threshold $\epsilon > 0$. Besides, the Newton decrement can also be verified as equivalent to

$$((\Delta x_{nt}^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)})^{1/2} = \|\Delta x_{nt}^{(k)}\|_{\nabla^2 f(x^{(k)})} \tag{3.62}$$

which is the quadratic $\nabla^2 f(x^{(k)})$ norm of $\Delta x_{nt}^{(k)}$. Finally, we present some convergence results of the Newton's method. Since the Hessian matrix of f is required to be positive definite, f is m-strongly convex for some $m > 0$. Suppose f is also M-smooth and the Hessian of f is L-Lipschitz, i.e., $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$. It can be proved that (see [5]) there exist two constants η and γ satisfying $0 < \eta \leq \frac{m^2}{L}$ and $\gamma > 0$ such that

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma \quad if \ \|\nabla f(x^{(k)})\|_2 \geq \eta \tag{3.63}$$

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq (\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2)^2 \quad if \ \|\nabla f(x^{(k)})\|_2 < \eta \tag{3.64}$$

If $\|\nabla f(x^{(k)})\|_2 \geq \eta$, we call $x^{(k)}$ is at the damped Newton phase; otherwise, we call $x^{(k)}$ is at the quadratically convergent phase. Although the Newton's method converges fast, the cost of forming and storing the Hessian matrix and that of computing the Newton step are high.

3.3.3 Equality-Constrained Minimization Problems

In this subsection, we consider solving the following equality constrained minimization problem

$$\inf_x f(x) \quad subject\ to\ Ax = b \tag{3.65}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and twice differentiable function and $A \in \mathbb{R}^{p\times n}$ with rank equal to $p < n$. In the following, we introduce two methods of solving 3.65. The first method solves 3.65 by converting it to an unconstrained minimization problem. Concretely speaking, we can express the constraint set $\{x|Ax = b\}$ as

$$\{\hat{x} + Fz|z \in \mathbb{R}^{n-p}\} \tag{3.66}$$

where $A\hat{x} = b$ and $F \in \mathbb{R}^{n\times(n-p)}$ is any matrix whose range is the nullspace of A. Therefore, we can solve the following unconstrained minimization problem instead

$$\inf_z \tilde{f}(z) = f(Fz + \hat{x}) \tag{3.67}$$

The second method uses an extension of the Newton's method to solve 3.65. Concretely speaking,

$$\begin{aligned}\Delta x_{nt}^{(k)} = \underset{v}{argmin} f(x^{(k)}) + \nabla f(x^{(k)})^T v + \frac{1}{2} v^T \nabla^2 f(x^{(k)}) v \\ subject\ to\ A(x^{(k)} + v) = b\end{aligned} \tag{3.68}$$

KKT conditions should be satisfied. They are listed as follows.

$$A(x^{(k)} + \Delta x_{nt}^{(k)}) = b \tag{3.69}$$

$$\nabla^2 f(x^{(k)})\Delta x_{nt}^{(k)} + \nabla f(x^{(k)}) + A^T w = 0 \tag{3.70}$$

3.69 corresponds to the primal feasible condition. Since $x^{(k)}$ is feasible (i.e., $Ax^{(k)} = b$), $A\Delta x_{nt}^{(k)} = 0$. Such search direction is called a feasible direction. Furthermore, it can be deduced that $\nabla f(x^{(k)})^T \Delta x_{nt}^{(k)} = -(\Delta x_{nt}^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)} \leq 0$. Hence, the Newton step $\Delta x_{nt}^{(k)}$ is a descent direction as desired. 3.70 corresponds to the condition that the gradient of the Lagrangian should be zero (w is the associated dual

variable for the primal problem 3.68). We can combine conditions 3.69 and 3.70 into a KKT linear system as follows.

$$\begin{bmatrix} \nabla^2 f(x^{(k)}) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt}^{(k)} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x^{(k)}) \\ 0 \end{bmatrix} \tag{3.71}$$

We call $\begin{bmatrix} \nabla^2 f(x^{(k)}) & A^T \\ A & 0 \end{bmatrix}$ the KKT matrix. In the following, we also define the second-order approximation $\hat{f}_k(v)$ as 3.60. We also derive an estimate of the distance from $f(x^{(k)})$ to the optimal value p^* based on the second-order approximation as follows

$$\begin{aligned}
& f(x^{(k)}) - p^* \\
=& f(x^{(k)}) - \inf_x f(x) \\
\approx& f(x^{(k)}) - \inf_v f(x^{(k)}) + \nabla f(x^{(k)})^T v + \frac{1}{2} v^T \nabla^2 f(x^{(k)}) v \\
=& f(x^{(k)}) - \hat{f}_k(\Delta x_{nt}^{(k)}) \\
=& -\nabla f(x^{(k)})^T \Delta x_{nt}^{(k)} - \frac{1}{2} (\Delta x_{nt}^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)} \\
& \because \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)} + A^T w = -\nabla f(x^{(k)}) \\
& \therefore (\Delta x_{nt}^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)} + (\Delta x_{nt}^{(k)})^T A^T w = -(\Delta x_{nt}^{(k)})^T \nabla f(x^{(k)}) \\
& \because A \Delta x_{nt}^{(k)} = 0 \Rightarrow (\Delta x_{nt}^{(k)})^T A^T = 0 \\
& \therefore (\Delta x_{nt}^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)} = -(\Delta x_{nt}^{(k)})^T \nabla f(x^{(k)}) = -\nabla f(x^{(k)})^T \Delta x_{nt}^{(k)} \\
& \therefore f(x^{(k)}) - \hat{f}_k(\Delta x_{nt}^{(k)}) = \frac{1}{2} (\Delta x_{nt}^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)}
\end{aligned}$$

We also define the Newton decrement of f at $x^{(k)}$ as

$$\lambda(x^{(k)}) := ((\Delta x_{nt}^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x_{nt}^{(k)})^{1/2} = \|\Delta x_{nt}^{(k)}\|_{\nabla^2 f(x^{(k)})} \tag{3.72}$$

A natural stopping criterion can also be set as $\frac{1}{2}\lambda^2(x^{(k)}) \leq \epsilon$ for some prescribed threshold $\epsilon > 0$. Last, we prove that the Newton step of f at $x^{(k)}$ derived by method 1 (denoted as $(\Delta x_{nt}^{(k)})_1$) is the same as that derived by method 2 (denoted as $(\Delta x_{nt}^{(k)})_2$). The Newton step of $\tilde{f}$ at

$z^{(k)}$, denoted as $(\Delta z_{nt}^{(k)})_1$, is calculated as

$$(\Delta z_{nt}^{(k)})_1 = -\nabla^2 \tilde{f}^{-1}(z^{(k)})\nabla \tilde{f}(z^{(k)}) = -(F^T \nabla f(x^{(k)})F)^{-1}F^T \nabla f(x^{(k)})$$
$$\because\ x^{(k)} = Fz^{(k)} + \hat{x}\ and\ x^{(k)} + (\Delta x_{nt}^{(k)})_1 = F(z^{(k)} + (\Delta z_{nt}^{(k)})_1) + \hat{x}$$
$$\therefore\ (\Delta x_{nt}^{(k)})_1 = F(\Delta z_{nt}^{(k)})_1 = -F(F^T \nabla f(x^{(k)})F)^{-1}F^T \nabla f(x^{(k)})$$

As for method 2, since $A(\Delta x_{nt}^{(k)})_2 = 0$, $(\Delta x_{nt}^{(k)})_2 = F(\Delta z_{nt}^{(k)})_2$ for some vector $(\Delta z_{nt}^{(k)})_2$. Hence,

$$\nabla^2 f(x^{(k)})F(\Delta z_{nt}^{(k)})_2 + A^T w = -\nabla f(x^{(k)})$$
$$\because\ AF = 0 \Rightarrow F^T A^T = 0$$
$$\therefore\ F^T \nabla^2 f(x^{(k)})F(\Delta z_{nt}^{(k)})_2 = -F^T \nabla f(x^{(k)})$$
$$\therefore\ (\Delta z_{nt}^{(k)})_2 = -(F^T \nabla^2 f(x^{(k)})F)^{-1}F^T \nabla f(x^{(k)})$$
$$\therefore\ (\Delta x_{nt}^{(k)})_2 = F(\Delta z_{nt}^{(k)})_2 = -F(F^T \nabla^2 f(x^{(k)})F)^{-1}F^T \nabla f(x^{(k)})$$

Therefore, we successfully verify that $(\Delta x_{nt}^{(k)})_1 = (\Delta x_{nt}^{(k)})_2$. Furthermore, we can verify as follows that the Newton decrement of $\tilde{f}$ at $z^{(k)}$ (denoted as $\tilde{\lambda}^2(z^{(k)})$) is the same as the Newton decrement of f at $x^{(k)}$ (denoted as $\lambda^2(x^{(k)})$).

$$\begin{aligned}
\tilde{\lambda}^2(z^{(k)}) &= (\Delta z_{nt}^{(k)})_1^T \nabla^2 \tilde{f}(z^{(k)})(\Delta z_{nt}^{(k)})_1 \\
&= (\Delta z_{nt}^{(k)})_2^T \nabla^2 \tilde{f}(z^{(k)})(\Delta z_{nt}^{(k)})_2 \\
&= (\Delta z_{nt}^{(k)})_2^T F^T \nabla^2 f(x^{(k)})F(\Delta z_{nt}^{(k)})_2 \\
&= (\Delta x_{nt}^{(k)})_2 \nabla^2 f(x^{(k)})(\Delta x_{nt}^{(k)})_2 \\
&= \lambda^2(x^{(k)})
\end{aligned}$$

For this subsection, the readers can refer to [5] for references.

3.4 Barrier Method

Barrier method (also called log barrier method or interior point method) focuses on the following constrained minimization problem

$$\begin{aligned} \inf_x f_0(x) \quad & subject\ to\ f_i(x) \leq 0, i = 1, 2, \cdots, m \\ & and\ Ax = b \end{aligned} \tag{3.73}$$

where $f_0, f_1, \cdots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex and twice differentiable functions and $A \in \mathbb{R}^{p \times n}$ has rank $p < n$. 3.73 is equivalent to

$$\inf_x f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) \quad subject\ to\ Ax = b \tag{3.74}$$

where $I_-(u) = \begin{cases} 0 & u \leq 0 \\ \infty & u > 0 \end{cases}$. Since $I_-(u)$ is not differentiable, we consider approximating it by a convex and differentiable function $\hat{I}_-(u) = -\frac{1}{t}\log(-u)$, where $t > 0$ is a parameter for the approximation accuracy. As can be seen in figure 3.2, $\hat{I}_-(u)$ can approximate $I_-(u)$ more and more accurately as t increases. Note that in section 3.1.1, we replace $I_-(u)$ with a lower bound function $\lambda_i(u)$, which results in a "minorization-maximization" method. However, $\hat{I}_-(u)$ introduced in this section is not a lower bound for $I_-(u)$ since $\hat{I}_-(u)$ is over $I_-(u)$ when u is in the interval of -1 and 0. Hence, we do not cover the barrier method in section 3.1.

As a result, the barrier method considers solving the following alternative minimization problem instead of 3.73

$$\begin{aligned} & \inf_x f_0(x) + \sum_{i=1}^{m} -\frac{1}{t}\log(-f_i(x)) \quad subject\ to\ Ax = b \\ \equiv & \inf_x t f_0(x) + \phi(x) \quad subject\ to\ Ax = b \end{aligned} \tag{3.75}$$

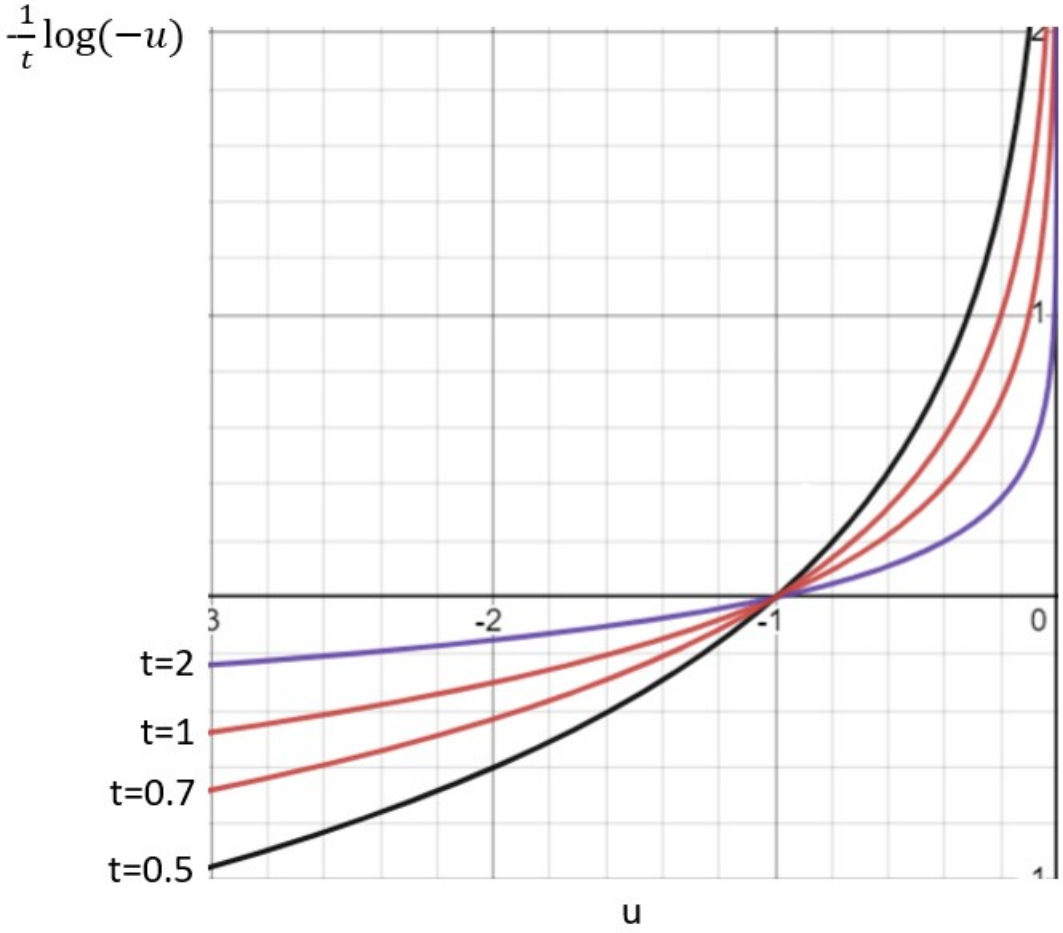

Figure 3.2: $-\frac{1}{t}\log(-u)$

where $\phi(x) = -\sum_{i=1}^{m}\log(-f_i(x))$ with the domain $dom\,\phi = \{x \in \mathbb{R}^n | f_i(x) < 0, i = 1, \cdots, m\}$[2]. $\phi(x)$ is called the logarithmic barrier (or log barrier) for the original problem 3.73 and the alternative problem 3.75 is called the centering problem. A minimizer $x^*(t)$ of the centering problem is called a central point and a central path associated with the original problem is defined as

$$C = \{x^*(t) \,|\, t > 0\} \tag{3.76}$$

Assume $x^*(t)$ is primal optimal and $\nu^*(t)$ is dual optimal for the centering problem. The following KKT conditions should be satisfied.

$$Ax^*(t) = b \tag{3.77}$$

[2]since $f_i(x)$ is constrained to be strictly smaller than zero, feasible solutions of 3.75 always lie in the interior of the constraint region of 3.73. This is why the barrier method is also called the interior point method.

$$f_i(x^*(t)) < 0 \tag{3.78}$$

$$t\nabla f_0(x^*(t)) + \nabla\phi(x^*(t)) + A^T\nu^*(t) = 0$$

$$\Rightarrow t\nabla f_0(x^*(t)) + \sum_{i=1}^{m} \frac{1}{-f_i(x^*(t))}\nabla f_i(x^*(t)) + A^T\nu^*(t) = 0 \tag{3.79}$$

Conditions 3.77 and 3.78 correspond to the primal feasible conditions and condition 3.79 corresponds to the requirement that the Lagrangian of the centering problem be zero. Condition 3.79 is also called the centrality condition.

Let $\tilde{\lambda}_i(t) = \frac{1}{-tf_i(x^*(t))}, i = 1, \cdots, m$ and $\tilde{\nu}(t) = \nu^*(t)/t$. Since $t > 0$ and $f_i(x^*(t)) < 0$, $\tilde{\lambda}_i(t) > 0$. Furthermore, from the centrality condition, $\nabla f_0(x^*(t)) + \sum_{i=1}^{m} \tilde{\lambda}_i(t)\nabla f_i(x^*(t)) + A^T\tilde{\nu}(t) = 0$. Hence $x^*(t)$ minimizes the Lagrangian associated with the primal problem 3.73 $L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T(Ax - b)$ for $\lambda = \tilde{\lambda}(t)$ and $\nu = \tilde{\nu}(t)$, which implies that $g(\tilde{\lambda}(t), \tilde{\nu}(t)) > -\infty$. Actually, $g(\tilde{\lambda}(t), \tilde{\nu}(t)) = f_0(x^*(t)) + \sum_{i=1}^{m} \tilde{\lambda}_i(t) f_i(x^*(t)) + \tilde{\nu}(t)^T(Ax^*(t) - b) = f_0(x^*(t)) - \frac{m}{t}$. Therefore, $(\tilde{\lambda}(t), \tilde{\nu}(t))$ is dual feasible for the original primal problem 3.73. Considering the KKT conditions for the original primal problem 3.73, we have verified that $x^*(t)$ is primal feasible, $(\tilde{\lambda}(t), \tilde{\nu}(t))$ is dual feasible and $\nabla L(x^*(t), \tilde{\lambda}(t), \tilde{\nu}(t)) = 0$. However, because $\tilde{\lambda}_i(t) f_i(x^*(t)) = -\frac{1}{t} \neq 0, i = 1, \cdots, m$, the complementary slackness condition is not satisfied. Hence, $x^*(t)$ is not primal optimal and $(\tilde{\lambda}(t), \tilde{\nu}(t))$ is not dual optimal for the original problem 3.73. We can conclude that

$$g(\tilde{\lambda}(t), \tilde{\nu}(t)) < d^* \leq p^* < f_0(x^*(t))$$

$$\Rightarrow f_0(x^*(t)) - p^* < f_0(x^*(t)) - g(\tilde{\lambda}(t), \tilde{\nu}(t)) = \frac{m}{t}$$

where p^* represents the optimal value for the original problem 3.73 and d^* represents the optimal value for the dual problem of 3.73. As a result, we find that $x^*(t)$ is m/t-suboptimal and converges to an optimal point as t increases.

Based on the above discussion, we can introduce the barrier method as the following algorithm now.

Algorithm 13 Barrier Method

Input: $f_i(x) : \mathbb{R}^n \to \mathbb{R}, i = 0, 1, \cdots, m$: convex and twice differentiable functions
$A \in \mathbb{R}^{p \times n}$: a matrix with rank $p < n$
Parameters : $\mu > 1$; tolerance $\epsilon > 0$
Initialization : $x^{(0)}$: strictly feasible (i.e., $f_i(x^{(0)}) < 0, i = 1, 2, \cdots, m$) ; $t^{(0)} > 0$
Iteration : repeat until $m/t^{(\bar{k})} < \epsilon$ at $k = \bar{k}$

$$\begin{aligned} &\textit{centering step} : \textit{compute } x^*(t^{(k)}) \textit{ by minimizing } t^{(k)} f_0(x) + \phi(x) \textit{ subject to } Ax = b \\ &\qquad\qquad \textit{starting at } x^{(k)} \textit{ using the Newton's method} \\ &\textit{update} \ : \ x^{(k+1)} = x^*(t^{(k)}) \\ &\textit{increase } t \ : \ t^{(k+1)} = \mu t^{(k)} \end{aligned}$$

Output: $x^{(\bar{k})}$

The barrier method solves a convex optimization problem with linear equality constraints and inequality constraints by reducing it to a sequence of linear equality constrained problems, with twice differentiable objective. The barrier method then applies the Newton's method to solve them. As what we have introduced in section 3.3, The Newton's method solves a linear equality constrained optimization problem with twice differentiable objective by reducing it to a sequence of linear equality constrained quadratic optimization problems. Each linear equality constrained quadratic optimization problem can be solved analytically by solving the KKT systems. As a whole, the barrier method solves a convex optimization with linear equality constraints and inequality constraints by tackling a hierarchy of optimization problems iteratively and progressively. Note that at each iteration within the Newton's method, we have a primal feasible point for the original problem 3.73. However, we do not have a dual feasible point until the end of the execution of the Newton's method (i.e., the end of the centering step). What's more, if we do not compute an exact minimizer of each centering problem (inexact centering), then we still cannot have a dual feasible point even at the end of each centering step.

Finally, we give a short remark on the choice of parameter μ. If we choose a small μ, then a large number of centering steps are required to meet the stopping criterion. But at each centering step, only a small

number of iterations are required for the Newton's method since we start at the optimal point of last centering step and since the objective does not change much due to small μ. Hence the iterates closely follow the central path, which results in a path-following method. In contrast, if we choose a large μ, a small number of centering steps are required while the number of iterations for each centering step are large. Hence, the iterates veer away off the central path. Empirically, a good choice of μ is within 10 and 100. The readers can refer to [5] for materials of the barrier method introduced in this section.

3.5 EM Algorithm

For signal processing, we often receive data from various sources. In a parametric setting (or model-based setting), we assume the data to be stochastic and generated from a probabilistic model. The probabilistic model is governed by a probability density function which depends on a set of unknown parameters. An important task is thus to estimate the values of the parameters from the data (or called statistics) so that we can have better understanding about the underlying generating mechanism. Let the observation data vector $y \in \mathbb{R}^{d_y}$ be a realization of the random vector $Y \in \mathbb{R}^{d_y}$. Assume Y depends on a set of unknown parameters $\theta_0 \in \Omega$, which is the parameter space. The density of y given a random parameter $\theta \in \Omega$[3] is denoted as

$$p(y|\theta) = p(Y = y|\theta) \tag{3.80}$$

which is also called the likelihood function of y. Clearly, our goal is to estimate θ_0 given y, which is known as an estimation problem. A famous approach called the maximum likelihood estimation (MLE) approach resorts to solving the following optimization problem in order

[3]Note that in this tutorial, we may interchangeably use the notation θ and θ_0 depending on the context of usage. If we want to express the deterministic nature of the true underlying parameter, then we use θ_0. On the other hand, if we are doing maximum likelihood estimation or some other mathematical derivations, then we use θ to denote the random nature of the unknown parameter.

to deal with the estimation problem.

$$\hat{\theta}_{MLE} = \underset{\theta \in \Omega}{argsup}\, L(\theta) := p(y|\theta) = \underset{\theta \in \Omega}{argsup}\, \ell(\theta) := \log(p(y|\theta)) \quad (3.81)$$

We call $\hat{\theta}_{MLE}$ the MLE estimator of the true parameter vector θ_0.

In some situations, y comes from the complete data $x \in \mathbb{R}^{d_x}$ which is a realization of the random vector $X \in \mathbb{R}^{d_x}$. The generation of X also depends on the true parameter vector θ_0. We cannot observe X directly; however, we known the density function of x given a random parameter $\theta \in \Omega$, which is

$$p(x|\theta) = p(X = x|\theta) \quad (3.82)$$

The support of X, denoted as $\mathcal{X}$, is thus the closure of the set of x where $p(x|\theta) > 0$. Since y comes from x, we also call y the incomplete data as a counterpart of the complete data x. A full understanding of x implies a full understanding of x. Hence, we can express the conditional density of x given y and θ as

$$p(x|y, \theta) = \frac{p(x|\theta)}{p(y|\theta)} \quad (3.83)$$

The support of X conditioned on y, denoted as $\mathcal{X}(y)$, is thus the closure of the set of x where $p(x|y, \theta) > 0$. We can incorporate the information of x (i.e., $p(x|\theta)$) to deal with the maximum likelihood estimation problem 3.81. We will introduce an example in the following section to illustrate such cases. In some situations, because it is difficult to directly solve the optimization problem 3.81 we are urged to incorporate $p(x|\theta)$ or even construct fictitious such x if actually y does not come from a complete data x. We will also introduce an example in the following section to illustrate such cases. In either cases, we can make use of the expectation-maximization (EM) algorithm to fulfill the estimation task.

In the following sections, we first introduce the EM algorithm in section 3.5.1. In section 3.5.2, we define some notations about the score statistics and information matrices and also point out some important

properties and notions regarding them. Those definitions and concepts are useful and required for the analysis of convergence. In section 3.5.3, we then present some convergence issues of the EM algorithm, including the convergence of the likelihood function, the rate of convergence, etc. In section 3.5.4, we introduce some useful variants of the EM algorithm. Lastly, in section 3.5.5, we give some examples regarding the use of the EM algorithm. The readers can refer to [8], [17], [20], [24] and [29] for references of the EM algorithm.

3.5.1 The EM Algorithm

The EM algorithm is described as the following algorithm. As can

Algorithm 14 EM Algorithm

Input: $p(y|\theta)$: the likelihood function of y given θ
$p(x|\theta)$: the likelihood function of x given θ
Parameters : $\epsilon > 0$: prescribed tolerance threshold
Initialization : $\theta^{(0)} \in \Omega$ as an initial estimate for θ_0
Iteration : repeat until $|\theta^{(\bar{k})} - \theta^{(\bar{k}-1)}| < \epsilon$ or $|\ell(\theta^{(\bar{k})}) - \ell(\theta^{(\bar{k}-1)})| < \epsilon$

1. E-step (Expectation)
   (a) formulate the conditional probability density function $p(x|y, \theta^{(k)})$ for the complete data x
   (b) form the conditional expected log-likelihood, which is called the Q-function

$$Q(\theta|\theta^{(k)}) = \int_{\mathcal{X}(y)} \log(p(x|\theta)) p(x|y, \theta^{(k)}) dx = \mathbb{E}_{X|y,\theta^{(k)}}[\log(p(X|\theta))] \tag{3.84}$$

2. M-step (Maximization)

$$\theta^{(k+1)} = \underset{\theta \in \Omega}{argsup}\, Q(\theta|\theta^{(k)}) \tag{3.85}$$

Output: $\theta^{(\bar{k})}$

be seen, the EM algorithm solves for θ that maximizes the expected log-likelihood of X. And once we have an estimate for θ_0, we can make a better guess about the complete data x.

3.5.2 Score Statistics and Information Matrices

First, we make a list of some definitions of score statistics and information matrices

1. The incomplete-data score statistic :

$$S(y; \theta) := \partial \log p(y|\theta) / \partial \theta \tag{3.86}$$

2. The complete-data score statistic :

$$S(x;\theta) := \partial \log p(x|\theta)/\partial\theta \tag{3.87}$$

3. The incomplete-data observed information matrix :

$$I(y;\theta) = -\partial^2 \log p(y|\theta)/\partial\theta\partial\theta^T \tag{3.88}$$

4. The incomplete-data expected information matrix :

$$\mathcal{I}_Y(\theta) = \mathbb{E}_{Y|\theta}[I(Y;\theta)] \tag{3.89}$$

5. The complete-data observed information matrix :

$$I(x;\theta) = -\partial^2 \log p(x|\theta)/\partial\theta\partial\theta^T \tag{3.90}$$

6. The complete-data expected information matrix :

$$\mathcal{I}_X(\theta) = \mathbb{E}_{X|\theta}[I(X;\theta)] \tag{3.91}$$

7. The complete-data conditional expected information matrix given y :

$$\mathcal{I}_X(\theta|y) = \mathbb{E}_{X|\theta,y}[I(X;\theta)] \tag{3.92}$$

8. The missing information matrix :

$$\mathcal{I}_m(\theta|y) = \mathbb{E}_{X|\theta,y}[-\partial^2 \log p(X|y,\theta)/\partial\theta\partial\theta^T] \tag{3.93}$$

We can make a connection between the incomplete-data score statistic and the complete-data score statistic in the following derivation.

$$\begin{aligned} S(y;\theta) &= \partial \log p(y|\theta)/\partial\theta \\ &= p'(y|\theta)/p(y|\theta) \\ &= (\int_{\mathcal{X}(y)} p'(x|\theta)dx)/p(y|\theta) \\ &= \int_{\mathcal{X}(y)} (\partial \log p(x|\theta)/\partial\theta)(p(x|\theta)/p(y|\theta))dx \\ &= \int_{\mathcal{X}(y)} S(x;\theta)p(x|y,\theta)dx \\ &= \mathbb{E}_{X|y,\theta}[S(X;\theta)] \end{aligned} \tag{3.94}$$

We can also make a connection among different information matrices in the following derivation.

$$\begin{aligned} &\log p(y|\theta) = \log p(x|\theta) - \log p(x|y,\theta) \\ &\Rightarrow \frac{\partial^2 \log p(y|\theta)}{\partial\theta\partial\theta^T} = \frac{\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 \log p(x|y,\theta)}{\partial\theta\partial\theta^T} \\ &\Rightarrow I(y;\theta) = I(x;\theta) + \frac{\partial^2 \log p(x|y,\theta)}{\partial\theta\partial\theta^T} \\ &\Rightarrow \mathbb{E}_{X|y,\theta}[I(Y;\theta)] = \mathbb{E}_{X|y,\theta}[I(X;\theta)] + \mathbb{E}_{X|y,\theta}[\frac{\partial^2 \log p(x|y,\theta)}{\partial\theta\partial\theta^T}] \\ &\Rightarrow I(y;\theta) = \mathcal{I}_X(\theta|y) - \mathcal{I}_m(\theta|y) \end{aligned} \tag{3.95}$$

The final result of 3.95 corresponds to the missing information principle, which states that the observed information equals the conditional expected complete-data information minus the missing information.

3.5.3 Convergence Analysis

The EM algorithm serves as an iterative approach to solve the maximum likelihood estimation problem 3.81 so it is important to verify

that the likelihood do increases iteratively. We prove it as follows

$$
\begin{aligned}
&\because p(x|\theta) = p(x|y,\theta)p(y|\theta) \\
&\therefore \log L(\theta) = \log p(x|\theta) - \log p(x|y,\theta) \\
\Rightarrow &\mathbb{E}_{X|y,\theta^{(k)}}[\log L(\theta)] = \mathbb{E}_{X|y,\theta^{(k)}}[\log p(x|\theta)] - \mathbb{E}_{X|y,\theta^{(k)}}[\log p(x|y,\theta)] \\
\Rightarrow &\log L(\theta) = Q(\theta;\theta^{(k)}) - H(\theta;\theta^{(k)}) \\
&where\, H(\theta;\theta^{(k)}) := \mathbb{E}_{X|y,\theta^{(k)}}[\log p(x|y,\theta)] \\
&\log L(\theta^{(k+1)}) - \log L(\theta^{(k)}) \\
= &(Q(\theta^{(k+1)};\theta^{(k)}) - Q(\theta^{(k)};\theta^{(k)})) - (H(\theta^{(k+1)};\theta^{(k)}) - H(\theta^{(k)};\theta^{(k)})) \\
\geq &H(\theta^{(k)};\theta^{(k)}) - H(\theta^{(k+1)};\theta^{(k)}) \\
= &\mathbb{E}_{X|y,\theta^{(k)}}[\log \frac{p(x|y,\theta^{(k)})}{p(x|y,\theta^{(k+1)})}] \\
\geq &0
\end{aligned}
$$

The first inequality is due to the operation of the M-step 3.85. The second inequality is due to the non-negativity of the KL divergence. Hence, the likelihood indeed monotonically increases. If the likelihood sequence $\{L(\theta^{(k)})\}$ is a bounded sequence, then $L(\theta^{(k)})$ converges monotonically to some L^*. If $L^* = L(\theta^*)$ for some point θ^* at which $\frac{\partial L(\theta)}{\partial \theta} = 0$, or equivalently, $\frac{\partial \log L(\theta)}{\partial \theta} = 0$, i.e., $S(y;\theta^*) = 0$, then L^* is called a fixed value. Furthermore, we can prove that $[\partial Q(\theta;\theta^*)/\partial\theta]|_{\theta=\theta^*} = S(y;\theta^*)$. Therefore $[\partial Q(\theta;\theta^*)/\partial\theta]|_{\theta=\theta^*} = 0$, which implies that θ^* is a fixed point of the EM algorithm. We can prove it in two ways as follows. The first way of derivation is

$$
\begin{aligned}
&\log L(\theta) = Q(\theta;\theta^*) - H(\theta;\theta^*) \\
\Rightarrow &S(y;\theta^*) = [\partial Q(\theta;\theta^*)/\partial\theta]|_{\theta=\theta^*} - [\partial H(\theta;\theta^*)/\partial\theta]|_{\theta=\theta^*} = 0 \\
\Rightarrow &S(y;\theta^*) = [\partial Q(\theta;\theta^*)/\partial\theta]|_{\theta=\theta^*} = 0
\end{aligned}
$$

The last equality is due to the fact that $H(\theta;\theta^*) \leq H(\theta^*;\theta^*)\ \forall\theta \in \Omega$ (a result of the non-negativity of the KL divergence). The second way

of derivation is

$$\begin{aligned} S(y;\theta) &= \mathbb{E}_{X|y,\theta}[S(X;\theta)] \\ &= \int_{\mathcal{X}(y)} (\partial \log p(x|\theta)/\partial\theta) p(x|y,\theta) dx \\ &= \partial[\int_{\mathcal{X}(y)} \log p(x|\theta) p(x|y,\theta) dx]/\partial\theta \\ &= [\partial Q(\Theta;\theta)/\partial\Theta]|_{\Theta=\theta} \\ \Rightarrow S(y;\theta^*) &= [\partial Q(\Theta;\theta^*)/\partial\Theta]|_{\Theta=\theta^*} = [\partial Q(\theta;\theta^*)/\partial\theta]|_{\theta=\theta^*} = 0 \end{aligned}$$

A natural question that when $L(\theta^{(k)})$ converges to a fixed value thus arises. To answer this question, we need to introduce the concept of EM mapping and the regularity conditions established in [20]. The EM algorithm implicitly defines a mapping $\theta \to M(\theta)$ from the parameter space Ω to itself such that $\theta^{(k+1)} = M(\theta^{(k)})$. The function M is called the EM mapping. As for the regularity conditions, they are the following three conditions

1. Ω is a subset in d-dimensional Euclidean space $\mathbb{R}^d$
2. $\Omega_{\hat{\theta}} := \{\theta \in \Omega | L(\theta) \geq L(\hat{\theta})\}$ is compact for any $L(\hat{\theta}) > -\infty$
3. $L(\theta)$ is continuous in Ω and differentiable in the interior of Ω

A consequence of the regularity conditions is that any sequence $\{L(\theta^{(k)})\}$ is bounded above for any $\theta^{(0)} \in \Omega$. Besides, If $\Omega_{\hat{\theta}}$ is in the interior of Ω for any $\hat{\theta} \in \Omega$, then in each M-step $\theta^{(k+1)}$ is a solution of the equation $\partial Q(\theta;\theta^{(k)})/\partial\theta = 0$. An important result, proved in [20], about the convergence of $L(\theta^{(k)})$ to a fixed value is described as the following theorem

Theorem 3.5.1. *If regularity conditions hold and $M(\theta^{(k)})$ is closed* [4] *over the complement of S, the set of fixed points in the interior of Ω, then all the limit points of $\{\theta^{(k)}\}$ are fixed points and $L(\theta^{(k)})$ converges monotonically to $L^* = L(\theta^*)$ for some fixed point $\theta^* \in S$*

[4] A mapping is said to be closed at $\theta = \theta_0$ if $\theta^{(k)} \to \theta_0, \theta \in \Omega$ and $\phi^{(k)} \to \phi_0, \phi^{(k)} \in M(\theta^{(k)})$ implies that $\phi_0 \in M(\theta_0)$

Note that a sufficient condition for the closedness of the EM mapping is that $Q(\Theta;\theta)$ is continuous in both Θ and θ, which is easier to verify. Hence, we can have the following corollary

Corollary 3.5.1.1. *If regularity conditions hold and $Q(\Theta;\theta)$ is continuous in both Θ and θ, then all the limit point of $\{\theta^{(k)}\}$ are fixed points and $L(\theta^{(k)})$ converges monotonically to $L^* = L(\theta^*)$ for some fixed point $\theta^* \in S$*

We need to make a remark that the convergence of $L(\theta^{(k)})$ to some value L^* does not automatically imply the convergence of the corresponding sequence of iterates $\{\theta^{(k)}\}$ to the point θ^*, where $L^* = L(\theta^*)$. Then when does $\theta^{(k)}$ also converge to a fixed point θ^*? First, we define $S(a) \triangleq \{\theta \in S : L(\theta) = a\}$ to be the subset of S of fixed points in the interior of Ω at which $L(\theta) = a$ and $\mathcal{L}(a) \triangleq \{\theta \in \Omega : L(\theta) = a\}$ to be the subset of Ω at which $L(\theta) = a$. [20] and [24] give some results regarding the convergence of $\theta^{(k)}$. We present them as follows.

Theorem 3.5.2. *If regularity conditions hold, $M(\theta^{(k)})$ is closed over the complement of S and $S(L^*)$ consists of the single point θ^* (that is, there cannot be two different fixed points with the same value L^*), where L^* is the limit of $L(\theta^{(k)})$, then $\theta^{(k)}$ converges to θ^**

Theorem 3.5.3. *If regularity conditions hold, $M(\theta^{(k)})$ is closed over the complement of S, $S(L^*)$ is discrete and $\|\theta^{(k+1)} - \theta^{(k)}\| \to 0$ as $k \to \infty$, then $\theta^{(k)}$ converges to some θ^* in $S(L^*)$*

Theorem 3.5.4. *If regularity conditions hold, $[\frac{\partial Q(\theta;\theta^{(k)})}{\partial \theta}]|_{\theta=\theta^{(k+1)}} = 0$ and $\partial Q(\Theta;\theta)/\partial\Theta$ is continuous in Θ and θ, then $\theta^{(k)}$ converges to a fixed point θ^* with $L(\theta^*) = L^*$, if either $\mathcal{L}(L^*) = \{\theta^*\}$ or $\|\theta^{(k+1)} - \theta^{(k)}\| \to 0$ as $k \to \infty$ and $\mathcal{L}(L^*)$ is discrete*

From theorem 3.5.4, we can come up with the following useful corollary

Corollary 3.5.4.1. *If regularity conditions hold,* $[\frac{\partial Q(\theta;\theta^{(k)})}{\partial\theta}]\big|_{\theta=\theta^{(k+1)}} = 0$, $\partial Q(\Theta;\theta)/\partial\Theta$ *is continuous in* Θ *and* θ, *and* $L(\theta)$ *is unimodal in* Ω *with* θ^* *being the only fixed point, then any EM sequence* $\{\theta^{(k)}\}$ *converges to the unique maximizer* θ^* *of* $L(\theta)$*;that is, it converges to the unique MLE of* θ

Finally, we make a derivation of the rate of convergence $r \triangleq \lim_{k\to\infty} \|\theta^{(k+1)} - \theta^*\|/\|\theta^{(k)} - \theta^*\|$

1. $S(y;\theta) \approx S(y;\theta^{(k)}) - I(y;\theta^{(k)})(\theta - \theta^{(k)})$ when θ is near $\theta^{(k)}$
   Assume $k \to \infty$ and $\theta^{(k)}$ converges to a fixed point θ^*
   then $S(y;\theta^*) = 0 \approx S(y;\theta^{(k)}) - I(y;\theta^{(k)})(\theta^* - \theta^{(k)})$
   $\Rightarrow \theta^* \approx \theta^{(k)} + I^{-1}(y;\theta^{(k)})S(y;\theta^{(k)})$
2. $\because \theta^{(k+1)}$ is near $\theta^{(k)}$ as $k \to \infty$
   $\therefore 0 = [\frac{\partial Q(\theta;\theta^{(k)})}{\partial\theta}]\big|_{\theta=\theta^{(k+1)}}$
   $\approx [\frac{\partial Q(\theta;\theta^{(k)})}{\partial\theta}]\big|_{\theta=\theta^{(k)}} + [\frac{\partial^2 Q(\theta;\theta^{(k)})}{\partial\theta\partial\theta^T}]\big|_{\theta=\theta^{(k)}}(\theta^{(k+1)} - \theta^{(k)})$
   $\Rightarrow 0 \approx S(y;\theta^{(k)}) - \mathcal{I}_X(\theta^{(k)}|y)(\theta^{(k+1)} - \theta^{(k)})$
   $\Rightarrow S(y;\theta^{(k)}) \approx \mathcal{I}_X(\theta^{(k)}|y)(\theta^{(k+1)} - \theta^{(k)})$

Combining 1. and 2., we can get

$$\begin{aligned}
&\theta^* - \theta^{(k)} \approx I^{-1}(y;\theta^{(k)})\mathcal{I}_X(\theta^{(k)}|y)(\theta^{(k+1)} - \theta^{(k)}) \\
&\Rightarrow \theta^{(k+1)} - \theta^* \approx [I_d - \mathcal{I}_X^{-1}(\theta^{(k)}|y)I(y;\theta^{(k)})](\theta^{(k)} - \theta^*) \\
&\qquad\qquad \approx [I_d - \mathcal{I}_X^{-1}(\theta^*|y)I(y;\theta^*)](\theta^{(k)} - \theta^*) \\
&\qquad\qquad = [\mathcal{I}_X^{-1}(\theta^*|y)\mathcal{I}_m(\theta^*|y)](\theta^{(k)} - \theta^*)
\end{aligned}$$

Hence, the rate of convergence is the largest eigenvalue of the information ratio matrix $\mathcal{I}_X^{-1}(\theta^*|y)\mathcal{I}_m(\theta^*|y)$. A larger value of r implies slower convergence, which may be somehow counter-intuitive. We can alternatively define the speed of convergence $s \triangleq 1 - r$. In this way, a larger value of s corresponds to faster convergence.

3.5.4 Variants of The EM Algorithm

In this section, we will introduce three useful and important variants of the EM algorithm.

MAP EM

MAP EM takes into account or imposes some prior information on θ. The E-step remains the same and the M-step is modified to maximize the posterior rather than the likelihood. Concretely speaking, the M-step becomes

$$\theta^{(m+1)} = \underset{\theta \in \Omega}{argmax}\, Q(\theta|\theta^{(m)}) + \log(p(\theta)) \tag{3.96}$$

where $p(\theta)$ is the prior probability density function of θ

Monte Carlo EM (MCEM)

In E-step, we need to calculate $Q(\theta; \theta^{(k)}) = \mathbb{E}_{X|y,\theta^{(k)}}[\log p(X|\theta)] = \int_{\mathcal{X}(y)} \log p(x|\theta) p(x|y, \theta^{(k)}) dx$. If we cannot obtain a closed-form solution to the Q function, then we might consider using the technique of Markov Chain Monte Carlo (MCMC). First we make m independent draws $x^{1_k}, x^{2_k}, \cdots, x^{m_k}$ from $p(x|y, \theta^{(k)})$ using MCMC. Then we approximate $Q(\theta; \theta^{(k)})$ by $Q_m(\theta; \theta^{(k)}) = \frac{1}{m} \sum_{j=1}^{m} \log p(x^{j_k}|\theta)$. Therefore, in M-step, we take maximization on the function $Q_m(\theta; \theta^{(k)})$ rather than on $Q(\theta; \theta^{(k)})$. The readers can refer to [14] for materials of the MCMC.

Expectation-Conditional Maximization (ECM) algorithm

One of major reasons for the popularity of the EM algorithm is that the M-step involves only complete-data ML estimation, which is often computationally simple. But if the complete-data ML estimation is rather complicated, then the EM algorithm is less attractive. If this is the case, then we may consider taking the maximization process of the M-step conditionally on parts of the parameters under estimation, which is often relatively simple. The ECM algorithm takes advantage of the simplicity of complete-data conditional maximization by replac-

ing a complicated M-step of the EM algorithm with several computationally simpler conditional maximization (CM) steps. Concretely speaking, we partition θ into S parts, namely, $\theta = (\theta_1, \theta_2, \cdots, \theta_S)$. After the E-step, we perform the following S CM steps rather than the original one M-step.

$$CM\ step\, 1 : \theta_1^{(k+1)} = \underset{\theta_1}{argmax}\, Q(\theta_1, \theta_2^{(k)}, \cdots, \theta_S^{(k)}; \theta^{(k)})$$

$$CM\ step\, 2 : \theta_2^{(k+1)} = \underset{\theta_2}{argmax}\, Q(\theta_1^{(k+1)}, \theta_2, \theta_3^{(k)}, \cdots, \theta_S^{(k)}; \theta^{(k)})$$

$$\vdots$$

$$CM\ step\, s : \theta_s^{(k+1)} = \underset{\theta_s}{argmax}\, Q(\theta_1^{(k+1)}, \cdots, \theta_{s-1}^{(k+1)}, \theta_s, \theta_{s+1}^{(k)} \cdots, \theta_S^{(k)}; \theta^{(k)})$$

$$\vdots$$

$$CM\ step\, S : \theta_S^{(k+1)} = \underset{\theta_S}{argmax}\, Q(\theta_1^{(k+1)}, \cdots, \theta_{S-1}^{(k+1)}, \theta_S; \theta^{(k)})$$

Hence, in the ECM algorithm, the s-th CM step requires the maximization of Q function with respect to θ_s with the other $S-1$ subvectors held fixed at their current values. If we update the Q function after each CM step; that is, perform one E-step before each CM step, then the corresponding algorithm is called the multicycle ECM. We define a cycle as one E-step followed by one CM step. Concretely speaking,

we perform the multicycle ECM as follows

$$
\begin{aligned}
&E\,step : Q(\theta;\theta^{(k)})\\
&CM\,step : \theta_1^{(k+1)} = \underset{\theta_1}{argmax}\, Q(\theta_1, \theta_2^{(k)}, \cdots, \theta_S^{(k)};\theta^{(k)})\\
&E\,step : Q(\theta;\theta_1^{(k+1)}, \theta_2^{(k)}, \cdots, \theta_S^{(k)})\\
&CM\,step : \theta_2^{(k+1)} = \underset{\theta_2}{argmax}\, Q(\theta_1^{(k+1)}, \theta_2, \theta_3^{(k)}, \cdots, \theta_S^{(k)};\theta_1^{(k+1)}, \theta_2^{(k)}, \cdots, \theta_S^{(k)})\\
&E\,step : Q(\theta;\theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_3^{(k)}, \cdots, \theta_S^{(k)})\\
&CM\,step : \theta_3^{(k+1)} = \underset{\theta_3}{argmax}\, Q(\theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_3, \theta_4^{(k)}, \cdots, \theta_S^{(k)}\\
&\qquad\qquad ;\theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_3^{(k)}, \cdots, \theta_S^{(k)})\\
&\vdots
\end{aligned}
$$

An obvious disadvantage of using a multicycle ECM algorithm is the extra computation at each iteration. However, as a tradeoff, one might expect it to result in larger increases in the log likelihood function per iteration since the Q-function is updated more often.

3.5.5 Examples

A Toy Example

This example was originally introduced in [8] and later elaborated on with more details in [17]. Assume there are n kids asked to choose a toy out of five choices. Let $X = [X_1, X_2, X_3, X_4, X_5]^T$ denote the histogram of their n choices, i.e., X_i denotes the number of the kids that chose toy i, $i = 1, 2, 3, 4, 5$. However, we cannot observe X directly; instead, we observe $Y = [Y_1, Y_2, Y_3, Y_4]^T$, where $Y_1 = X_1 + X_2$ denote the sum of the number of kids that chose toy 1 or 2, $Y_2 = X_3$, $Y_3 = X_4$ and $Y_4 = X_5$. If the probability by which a kid chose toys is modeled as $p = [p_1, p_2, p_3, p_4, p_5]^T$, then we can model the distribution of X

and Y as multinomial distributions; namely,

$$p(x|p) = \frac{n!}{\prod_{i=1}^{5} x_i!} \prod_{i=1}^{5} p_i^{x_i} \; ; \; p(y|p) = \frac{n!}{\prod_{i=1}^{4} y_i!} (p_1 + p_2)^{y_1} p_3^{y_2} p_4^{y_3} p_5^{y_4}$$

If p is parameterized as $[\frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta]^T, \theta \in (0,1), \theta \in (0,1)$, then

$$p(x|\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} (\frac{1}{2})^{x_1} (\frac{\theta}{4})^{x_2+x_5} (\frac{1-\theta}{4})^{x_3+x_4}$$

$$p(y|\theta) = \frac{n!}{\prod_{i=1}^{4} y_i!} (\frac{1}{2} + \frac{\theta}{4})^{y_1} (\frac{1-\theta}{4})^{y_2+y_3} (\frac{\theta}{4})^{y_4}$$

To apply the EM algorithm, we need to calculate the Q function and maximize it; that is,

$$\begin{aligned}
\theta^{(m+1)} &= \underset{\theta \in (0,1)}{argmax}\, Q(\theta|\theta^{(m)}) \\
&= \underset{\theta \in (0,1)}{argmax}\, \mathbb{E}_{X|y,\theta^{(m)}}[\log(p(X|\theta))] \\
&= \underset{\theta \in (0,1)}{argmax}\, \mathbb{E}_{X|y,\theta^{(m)}}[(X_2 + X_5)\log\theta + (X_3 + X_4)\log(1-\theta)]
\end{aligned}$$

We need to calculate the conditional probability $p(x|y, \theta^{(m)})$, which is

$$\begin{aligned}
& p(x|y, \theta^{(m)}) \\
&= \frac{p(x|\theta^{(m)})}{p(y|\theta^{(m)})} \\
&= \frac{y_1!}{x_1 x_2!} \left(\frac{\frac{1}{2}}{\frac{1}{2} + \frac{\theta^{(m)}}{4}} \right)^{x_1} \left(\frac{\frac{\theta^{(m)}}{4}}{\frac{1}{2} + \frac{\theta^{(m)}}{4}} \right)^{x_2} \mathbf{1}\{x_1 + x_2 = y_1\} \prod_{i=3}^{5} \mathbf{1}\{x_i = y_{i-1}\} \\
&= \frac{y_1!}{x_1 x_2!} \left(\frac{2}{2 + \theta^{(m)}} \right)^{x_1} \left(\frac{\theta^{(m)}}{2 + \theta^{(m)}} \right)^{x_2} \mathbf{1}\{x_1 + x_2 = y_1\} \prod_{i=3}^{5} \mathbf{1}\{x_i = y_{i-1}\}
\end{aligned}$$

Hence, $\mathbb{E}_{X|y,\theta^{(m)}}[X] = [\frac{2}{2+\theta^{(m)}}y_1, \frac{\theta^{(m)}}{2+\theta^{(m)}}y_1, y_2, y_3, y_4]^T$. We can get

$$\begin{aligned}\theta^{(m+1)} &= \underset{\theta\in(0,1)}{argmax}\left[\left(\frac{\theta^{(m)}}{2+\theta^{(m)}}y_1 + y_4\right)\log\theta + (y_2+y_3)\log(1-\theta)\right] \\ &= \frac{\frac{\theta^{(m)}}{2+\theta^{(m)}}y_1 + y_4}{\frac{\theta^{(m)}}{2+\theta^{(m)}}y_1 + y_2 + y_3 + y_4}\end{aligned}$$

In this simple illustrative example, we can simply perform maximum likelihood estimation on $p(y|\theta)$. However, in the following two examples, directly applying maximum likelihood estimation on the likelihood functions becomes intractable, which in turn forces us to resort to the EM algorithm instead.

Gaussian Mixtures

Assume there are two groups, one with p-dimensional Gaussian distribution density $N(y;\mu_1,\Sigma)$ and the other with p-dimensional Gaussian distribution density $N(y;\mu_2,\Sigma)$, where $\mu_1,\mu_2 \in \mathbb{R}^p$ is the mean vectors and $\Sigma \in \mathbb{R}^{p\times p}$ is the common covariance matrix. We randomly sample data $y \in \mathbb{R}^p$ from one of the two groups. The probability that a sample is from the former group is $1-\pi$ and from the latter group is π. Hence the probability density function of y follows the Gaussian mixture density $(1-\pi)N(y;\mu_1,\Sigma) + \pi N(y;\mu_2,\Sigma)$. Given n i.i.d. observations $y_1, y_2, \cdots, y_n$ from the Gaussian mixture density, we want to estimate the unknown true parameters $\theta_0 = (\pi, \mu_1, \mu_2, \Sigma)$, which is a two-group discrimination problem or a statistical pattern recognition problem. The likelihood function is

$$L(\theta|y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n}[(1-\pi)N(y_i;\mu_1,\Sigma) + \pi N(y_i;\mu_2,\Sigma)]$$

We find it difficult to analytically compute either $\underset{\theta}{argmax}\, L(\theta|y_1, y_2, \cdots, y_n)$ or $\underset{\theta}{argmax}\log L(\theta|y_1, y_2, \cdots, y_n)$ because of the bundle of two Gaussian densities. If we can know which group each y_i comes from, then the

conditional probability density of each y_i given this information will be simply a single Gaussian density, whose related maximum likelihood estimation is tractable. With this observation, we think of applying the EM algorithm. We can fictitiously construct the missing data z_j, which is an indicator variable identifying the j-th observation y_j as coming from the first ($z_j = 0$) or the second ($z_j = 1$) group. Then we can obtain the complete data $X = (Z, Y)$, where $Z = (z_1, z_2, \cdots, z_n)$ and $Y = (y_1, y_2, \cdots, y_n)$. To calculate the Q function, we need to derive the probability density function $p(x_j|\theta)$ and $p(x_j|y_j, \theta)$.

1. $p(X_j = (0, y_j)|\theta) = p(X_j = (0, y_j)|Z_j = 0, \theta)p(Z_j = 0|\theta)$
$= N(y_j; \mu_1, \Sigma)(1 - \pi)$

2. $p(X_j = (1, y_j)|\theta) = p(X_j = (1, y_j)|Z_j = 1, \theta)p(Z_j = 1|\theta)$
$= N(y_j; \mu_2, \Sigma)\pi$

3. $p(X_j = (0, y_j)|Y_j = y_j, \theta)$
$= p(X_j = (0, y_j)|\theta)/p(Y_j = y_j|\theta)$
$= (1 - \pi)N(y_j; \mu_1, \Sigma)/[(1 - \pi)N(y_j; \mu_1, \Sigma) + \pi N(y_j; \mu_2, \Sigma)]$

4. $p(X_j = (1, y_j)|Y_j = y_j, \theta)$
$= p(X_j = (1, y_j)|\theta)/p(Y_j = y_j|\theta)$
$= \pi N(y_j; \mu_2, \Sigma)/[(1 - \pi)N(y_j; \mu_1, \Sigma) + \pi N(y_j; \mu_2, \Sigma)]$

Hence, in E-step, we can calculate the Q function as follows

$$
\begin{aligned}
&Q(\theta; \theta^{(k)}) \\
&= \mathbb{E}_{X|y,\theta^{(k)}}[\log p(X|\theta)] \\
&= \sum_{j=1}^{n} \left[\log[N(y_j; \mu_1, \Sigma)(1 - \pi)](1 - \tau_j^{(k)}) + \log[N(y_j; \mu_2, \Sigma)\pi]\tau_j^{(k)}\right]
\end{aligned}
$$

$$
where\, \tau_j^{(k)} := \frac{\pi^{(k)} N(y_j; \mu_2^{(k)}, \Sigma^{(k)})}{(1 - \pi^{(k)})N(y_j; \mu_1^{(k)}, \Sigma^{(k)}) + \pi^{(k)} N(y_j; \mu_2^{(k)}, \Sigma^{(k)})}
$$

In M-step, we apply four CM steps instead of one M-step. Each CM step maximizes the Q function over one of π, μ_1, μ_2 and Σ.

1. $\pi^{(k+1)} = \underset{\pi}{argmax}\, Q(\theta; \theta^{(k)}) = \sum_{j=1}^{n} \tau_j^{(k)}/n$

2. $\mu_1^{(k+1)} = \underset{\mu_1}{argmax}\, Q(\theta; \theta^{(k)}) = \sum_{j=1}^{n}(1-\tau_j^{(k)})y_j/(n - \sum_{j=1}^{n}\tau_j^{(k)})$

3. $\mu_2^{(k+1)} = \underset{\mu_2}{argmax}\, Q(\theta; \theta^{(k)}) = \sum_{j=1}^{n}\tau_j^{(k)}y_j/\sum_{j=1}^{n}\tau_j^{(k)}$

4. $$\begin{aligned}\Sigma^{(k+1)} &= \underset{\Sigma}{argmax}\, Q(\theta; \theta^{(k)}) \\ &= \sum_{j=1}^{n}[(1-\tau_j^{(k)})(y_j - \mu_1^{(k+1)})(y_j - \mu_1^{(k+1)})^T \\ &\quad +\tau_j^{(k)}(y_j - \mu_2^{(k+1)})(y_j - \mu_2^{(k+1)})^T]/n\end{aligned}$$

Hence, we can view this example as an application of the ECM algorithm.

Baum-Welch Algorithm

Baum-Welch algorithm is an application of EM algorithm to learn model parameters of a hidden Markov model (HMM). Let $X = \{1, 2, \cdots, N\}$ denote the space of observation. Let $X^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_T^{(i)})$ denote an observation sequence with length T. Suppose there are total D observation sequences $\mathcal{X} = (X^{(1)}, X^{(2)}, \cdots, X^{(D)})$ and each observation is drawn independently and identically distributed (i.i.d.). Let $Z = \{1, 2, \cdots, M\}$ denote the space of hidden states. Let $Z^{(i)} = (z_1^{(i)}, z_2^{(i)}, \cdots, z_T^{(i)})$ denote a hidden state sequence with length T. Suppose there are total D hidden state sequences $\mathcal{Z} = (Z^{(1)}, Z^{(2)}, \cdots, Z^{(D)})$. Assume an HMM is parameterized by $\theta = (\pi, A, B)$, which represents the initial state vector, the state transition matrix and the emission matrix respectively. Concretely speaking,

1. $\pi_i = P(z_1 = i)$, $i \in [M]$

2. $A_{ij} = P(z_{t+1} = j | z_t = i)$, $i, j \in [M]$

3. $B_i(j) = P(x_t = j | z_t = i)$, $i \in [M]$, $j \in [N]$

The joint probability of $\mathcal{X}$ and $\mathcal{Z}$ conditioning on θ is

$$P(\mathcal{X}, \mathcal{Z}|\theta) = \prod_{d=1}^{D} \left(\pi_{z_1^{(d)}} B_{z_1^{(d)}}(x_1^{(d)}) \prod_{t=2}^{T} A_{z_{t-1}^{(d)} z_t^{(d)}} B_{z_t^{(d)}}(x_t^{(d)}) \right)$$

The learning problem is nontrivial because the hidden state sequences $\mathcal{Z}$ is not available; otherwise, we can directly compute the MLE $\theta^* = \underset{\theta}{argsup}\, P(\mathcal{X}, \mathcal{Z}|\theta)$. Without $\mathcal{Z}$, if we still want to compute the MLE, then we need to deal with $\theta^* = \underset{\theta}{argsup}\, P(\mathcal{X}|\theta) = \underset{\theta}{argsup} \sum_{z \in \mathcal{Z}} P(\mathcal{X}, z|\theta) = \underset{\theta}{argsup} \sum_{z \in \mathcal{Z}} \prod_{d=1}^{D} \left(\pi_{z_1^{(d)}} B_{z_1^{(d)}}(x_1^{(d)}) \prod_{t=2}^{T} A_{z_{t-1}^{(d)} z_t^{(d)}} B_{z_t^{(d)}}(x_t^{(d)}) \right)$, which is intractable since there are DT^M different values of z to try and different parameters are bundled together by the operation of multiplications. However, if we view $(\mathcal{X}, \mathcal{Z})$ as the complete data, $\mathcal{X}$ the incomplete data and $\mathcal{Z}$ the missing data, then we can apply the EM algorithm to estimate θ^*. The resulting iterative procedure is called the Baum-Welch algorithm. It is described as follows

1. E-step : compute $Q(\theta, \theta^{(k)}) = \sum_{z \in \mathcal{Z}} [\log p(\mathcal{X}, \mathcal{Z}|\theta)] p(\mathcal{Z}|\mathcal{X}, \theta^{(k)})$
2. M-step : $\theta^{(k+1)} = \underset{\theta}{argsup}\, Q(\theta, \theta^{(k)})$

As we can see, the EM algorithm is powerful for the usage of the log function can effectively disentangle the bundle of different parameters. After some derivations (see [38] for details), we can come up with the following iteration rules.

1. $\pi_i^{(k+1)} = \frac{1}{D} \sum_{d=1}^{D} p(z_1^{(d)} = i|X^{(d)}, \theta^{(k)})$
2. $A_{ij}^{(k+1)} = \frac{\sum_{d=1}^{D} \sum_{t=2}^{T} p(z_{t-1}^{(d)}=i, z_t^{(d)}=j|X^{(d)}, \theta^{(k)})}{\sum_{d=1}^{D} \sum_{t=2}^{T} p(z_{t-1}^{(d)}=i|X^{(d)}, \theta^{(k)})}$
3. $B_i^{(k+1)}(j) = \frac{\sum_{d=1}^{D} \sum_{t=1}^{T} p(z_t^{(d)}=i|X^{(d)}, \theta^{(k)}) I(x_t^{(d)}=j)}{\sum_{d=1}^{D} \sum_{t=1}^{T} p(z_t^{(d)}=i|X^{(d)}, \theta^{(k)})}$

Note that $p(z_t|X, \theta)$ and $p(z_{t-1}, z_t|X, \theta)$ are both quantities which can be computed efficiently by the forward-backward algorithm.

Chapter 4

ℓ_0 Minimization Problem

In this chapter, we will discuss a specific problem formulation related to the ℓ_0 norm. Precisely, we consider the following problem

$$\inf_{x \in \mathbb{R}^n} \|x\|_0 \quad subject\ to\ Ax = y \tag{4.1}$$

where $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$ $(m < n)$. In the context of sparse representation, we call y the signal, A the dictionary and x the coefficient vector. The dictionary has n m-dimensional column vectors (called atoms, denoted as $a_j, j \in [n]$), which leads to an underdetermined linear system $Ax = y$ since $m < n$ (more often $m \ll n$). Hence, there may be infinite possible coefficient vectors x that are feasible for the linear system. Among them, our interest is the sparsest ones (i.e., ones with smallest ℓ_0 norm). That is, our aim is to use as few dictionary atoms as possible to represent the signal y. Besides 4.1, we also consider the following problem

$$\inf_{x \in \mathbb{R}^n} \|x\|_0 \quad subject\ to\ \|Ax - y\|_2 \leq \epsilon \tag{4.2}$$

where $\epsilon > 0$. In this problem, we introduce a tolerance scalar ϵ to relax the strict constraint $Ax = y$. We can tolerate some extent of mismatch between Ax and y to gain some possible improvements of sparsity.

In the context of compressive sensing (compressive sampling), we assume there is an s-sparse n-dimensional signal vector x. Since it is sparse, we hope that we can compressively sample it with as few samples (or called measurements) as possible using the sampling matrix (or called the measurement matrix) $A \in \mathbb{R}^{m \times n}$ $(m < n)$. After the sampling operation, we will get $y = Ax$. We call y the sample vector (or called the measurement vector). We want y to contain sufficient information of x so that we can recover x with y and A. Note that it has been proved in [13] that the following three statements are equivalent.

1. The vector x can be reconstructed as the unique solution of the problem $\inf_{z\in\mathbb{R}^n} \|z\|_0$ subject to $Az = Ax = y$
2. The vector x is the unique s-sparse solution of $Az = y$ with $y = Ax$; that is, $\{z \in \mathbb{R}^n | Az = Ax, \|z\|_0 \leq s\} = \{x\}$
3. Every set of 2s columns of A is linearly independent.

Hence, it is reasonable to consider the ℓ_0 minimization problem 4.1[1] in order to achieve reconstruction. Furthermore, because every set of 2s columns of A should be linearly independent, the minimal number of measurements to recover an s-sparse vector is 2s. To better understand the properties of A in the context of compressive sensing, we will adopt the notion of restricted isometry property. In a more general setting, we will get $y = Ax + e$; that is, the sample vector y is contaminated with a noise vector $e \in \mathbb{R}^m$. Hence it is also reasonable to consider the problem 4.2[2]. Actually, the major algorithmic challenge in compressive sampling is to recover a signal given a vector of noisy samples.

However, it has been proved that solving both 4.1 and 4.2 are NP-hard (the readers can refer to section 2.3 of [13] for the NP-hardness of ℓ_0 minimization problem). Therefore, many algorithms are designed to pursuit the objective of 4.1 and 4.2 as close as possible while can be implemented efficiently. Also note that in the context of sparse representation, our major concern would be the difference between the true signal y and the linear combination $Ax^{(\bar{k})}$, where $x^{(\bar{k})}$ denotes the output vector of some algorithm, while in the context of compressive sensing, our major concern would be the difference between the true sparse signal x and the reconstruction vector $x^{(\bar{k})}$. However, we will call the difference the residual vector in either case. As for the notation, we use the capital $R^{(\bar{k})}$ to denote $\|y - Ax^{(\bar{k})}\|_2$ and the lower case $r^{(\bar{k})}$ to denote $\|x - x^{(\bar{k})}\|_2$. In the following sections, we will introduce some

[1]We may change the notation of 4.1 in order to make clear of the concept. Precisely, we consider the same problem of different notation : $\inf_{z\in\mathbb{R}^n} \|z\|_0$ *subject to* $Az = y = Ax$

[2]We also change the notation of the problem 4.2 in order to make clear of the concept : $\inf_{z\in\mathbb{R}^n} \|z\|_0$ *subject to* $\|Az - y\|_2 = \|Az - Ax - e\|_2 \leq \epsilon$

algorithms designed based on different methodologies respectively.

4.1 Minimization of Alternative Diversity Measures

We can view the ℓ_0 norm as a sparsity measure. The smaller the ℓ_0 norm of a vector, the greater the sparsity of the vector. Opposed to sparsity, we can also view the ℓ_0 norm as a diversity measure. The larger the ℓ_0 norm of a vector, the greater the diversity of the vector. However, the minimization of ℓ_0 norm is an NP-hard problem. Hence, a type of algorithms resort to alternative diversity measures. Minimization of those diversity measures is computationally tractable while it can be verified that the resulting coefficient vectors are sparse enough.

4.1.1 Iteratively Reweighted Least Squares (IRLS)-Type Algorithms

Let $D(x)$, $x \in \mathbb{R}^n$ be some convex diversity measure other than the ℓ_0 norm. The minimization problem we consider now becomes the following convex optimization problem.

$$\inf_{x \in \mathbb{R}^n} D(x) \quad subject\ to\ Ax = y \tag{4.3}$$

Assume x^* is a minimizer of the problem 4.3 and λ^* is a dual optimal point of the dual problem of 4.3. The following KKT conditions are necessary to satisfy

$$\nabla_x D(x^*) + A^T \lambda^* = 0$$
$$Ax^* = y$$

Assume the gradient of the diversity measure has a factored representation, which is

$$\nabla_x D(x) = \alpha(x) \Pi(x) x \tag{4.4}$$

where $\alpha(x)$ is a scalar and $\Pi(x)$ is a diagonal matrix. In this way, the KKT conditions become

$$\alpha(x^*)\Pi(x^*)x^* + A^T\lambda^* = 0$$
$$Ax^* = y$$

We can in turn come up with the following two fixed-point equations

$$x^* = \Pi^{-1}(x^*)A^T(A\Pi^{-1}(x^*)A^T)^{-1}y$$
$$\lambda^* = -\alpha(x^*)(A\Pi^{-1}(x^*)A^T)^{-1}y$$

From the first equation, we can design a algorithm based on the following fixed-point iteration rule

$$x^{(k+1)} = \Pi^{-1}(x^{(k)})A^T(A\Pi^{-1}(x^{(k)})A^T)^{-1}y \tag{4.5}$$

With such iteration rule, the KKT conditions of the minimization problem of each iteration become

$$\alpha(x^{(k)})\Pi(x^{(k)})x^{(k+1)} + A^T\lambda^{(k+1)} = 0$$
$$Ax^{(k+1)} = y$$

Hence, the algorithm equivalently solves the following minimization problem at each iteration

$$\begin{aligned} x^{(k+1)} &= \underset{x\in\mathbb{R}^n}{argin}f\ \frac{\alpha(x^{(k)})}{2}x^T\Pi(x^{(k)})x \quad subject\ to\ Ax = y \\ &= \underset{x\in\mathbb{R}^n}{argin}f\ x^T\Pi(x^{(k)})x \quad subject\ to\ Ax = y \\ &= \underset{x\in\mathbb{R}^n}{argin}f\ \|W_k^{-1}x\|_2^2 \quad subject\ to\ Ax = y \\ &\quad where\ W_k := \Pi(x^{(k)})^{-1/2} \end{aligned} \tag{4.6}$$

We call such algorithm a reweighted minimum norm algorithm or a iteratively reweighted least squares (IRLS) algorithm. The term " reweighted" refers to the fact that x is reweighted by W_k^{-1}. The term "minimum norm" and "least squares" refer to the minimization of ℓ_2 norm. Let $q := W_k^{-1}x$. The minimization problem 4.6 is equivalent to

$$q^{(k+1)} = \underset{q\in\mathbb{R}^n}{argin}f\ \|q\|_2^2 \quad subject\ to\ AW_kq = y \tag{4.7}$$

Since q is an affine scaling transformation (AST) of x using the matrix W_k, we also call such algorithm an AST-based algorithm. Note that we can analytically solve the problem 4.7; namely, $q^{(k+1)} = (AW_k)^{\dagger}y$, where the symbol † denotes the Moore-Penrose pseudoinverse of a vector. In the following, we will introduce three kinds of diversity measures and the corresponding IRLS algorithms based on minimizing them. The three diversity measures are the p-norm like diversity measure $E^p(x)$, the Gaussian entropy diversity measure $H_G(x)$ and the Shannon entropy diversity measure $H_S(x)$. The readers can refer to [32] for details.

The p-norm Like Diversity Measure $E^p(x)$

The p-norm like diversity measure $E^p(x)$ is defined as

Definition 4.1.1 (p-norm like diversity measure)**.**

$$\begin{aligned} E^p(x) &= sgn(p)\sum_{i=1}^{n}|x[i]|^p, p \leq 1 \\ &= \begin{cases} \sum_{i=1}^{n}|x[i]|^p & 0 \leq p \leq 1 \\ -\sum_{i=1,x[i]\neq 0}^{n}|x[i]|^p & p < 0 \end{cases} \end{aligned} \tag{4.8}$$

The gradient vector of $E^p(x)$ can be expressed as a factored representation; namely, $\nabla_x E^p(x) = \alpha(x)\Pi(x)x$, where $\alpha(x) = |p|$ and $\Pi(x) = diag(|x[i]|^{p-2})$. Hence, based on minimizing the p-norm like diversity measure, we can design an algorithm as follows

Algorithm 15 IRLS Based on $E^p(x)$

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$

1. Let $W_k = diag(|x^{(k)}[i]|^{1-p/2})$
2. $q^{(k+1)} = (AW_k)^{\dagger}y$
3. $x^{(k+1)} = W_k q^{(k+1)}$

Output: $x^{(\bar{k})}$

The Gaussian Entropy Diversity Measure $H_G(x)$

The Gaussian entropy diversity measure $H_G(x)$ is defined as

Definition 4.1.2 (Gaussian entropy diversity measure)**.**

$$H_G(x) = \sum_{i=1}^{n} ln|x[i]|^2 = 2\sum_{i=1}^{n} ln|x[i]| \tag{4.9}$$

The gradient vector of $H_G(x)$ can be expressed as a factored representation; namely, $\nabla_x H_G(x) = \alpha(x)\Pi(x)x$, where $\alpha(x) = 2$ and $\Pi(x) = diag(\frac{1}{x[i]^2})$. Hence, based on minimizing the Gaussian entropy diversity measure, we design an algorithm as follows

Algorithm 16 IRLS Based on $H_G(x)$

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$

1. Let $W_k = diag(|x^{(k)}[i]|)$
2. $q^{(k+1)} = (AW_k)^{\dagger}y$
3. $x^{(k+1)} = W_k q^{(k+1)}$

Output: $x^{(\bar{k})}$

Note that it is equivalent to the p-norm-like case when p is set to be zero (although the gradient of $E^p(x)$ is not defined for $p = 0$). Hence, there must be some close relationships between $H_G(x)$ and $E^p(x)$ when p approaches zero. Indeed, we can present two relationships as follows. The first one is based on the arithmetic-geometric inequality.

$$\left(\prod_{i=1}^{n}|x[i]|^p\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n}|x[i]|^p$$

$$\Rightarrow \left(\frac{1}{n}E^{p_-}(x)\right)^{1/p_-} \leq (e^{H_G(x)})^{1/2n} \leq (\frac{1}{n}E^{p_+}(x))^{1/p_+}$$

$$where\ p_- \leq 0\ and\ p_+ \geq 0$$

We have equality as p approaches zero. Hence $e^{\frac{1}{2n}H_G(x)} = \lim_{p\to 0}(\frac{1}{n}E^p(x))^{1/p}$. The second one is based on the first-order Taylor approximation of

$E^p(x)$ when p approaches zero.

$$\begin{aligned} E^p(x) &\approx E^0(x) + \frac{dE^p(x)}{dp}p \\ &= E^0(x) + (\sum_{i=1}^{n} |x[i]|^p ln|x[i]|)p \\ &= E^0(x) + \frac{1}{2}H_G(x) \quad \because p \to 0 \end{aligned}$$

Hence, as p gets smaller, $E^p(x)$ begins to behave like $H_G(x)$ except at sparsity points where the diversity measure $E^0(x)$ jumps discontinuously. Finally, we make a remark that the IRLS algorithm based on the Gaussian entropy diversity measure is equivalent to the basic FOCUSS algorithm introduced in [16]. We will delve deeply into the FOCUSS algorithm in the subsequent section.

The Shannon Entropy Diversity Measure $H_S(x)$

The Shannon entropy diversity measure $H_S(x)$ is defined as

Definition 4.1.3 (Shannon entropy diversity measure)**.**

$$H_S(x) = -\sum_{i=1}^{n} \tilde{x}[i] ln(\tilde{x}[i]) \ , where \ \tilde{x}[i] = \frac{x[i]^2}{\|x\|_2^2} \tag{4.10}$$

The gradient vector of $H_S(x)$ can be expressed as a factored representation; namely, $\nabla_x H_S(x) = \alpha(x)\Pi(x)x$, where $\alpha(x) = \frac{2}{\|x\|_2^2}$ and $\Pi(x) = -diag(H_S(x) + ln(\tilde{x}[i]))$. Note that since $\Pi(x)$ is indefinite, simply mimicking the way we construct algorithms 15 and 16 cannot ensure us a provably convergent algorithm. If we still mimic the way we construct algorithms 15 and 16, we will get the following iteration rules

1. Let $W_k = -diag((H_S(x^{(k)}) + ln(\tilde{x}^{(k)}[i]))^{-1/2})$
2. $q^{(k+1)} = (AW_k)^\dagger y$
3. $x^{(k+1)} = W_k q^{(k+1)}$

In order to ensure $H_S(x^{(k+1)}) \leq H_S(x^{(k)})$, we change the notation of the output of step 3 to $x_r^{(k+1)} = W_k q^{(k+1)}$ and let $x^{(k+1)} = x^{(k)} + \mu_k(x^{(k)} - x_r^{(k+1)})$. It has been proved in [32] that it is sufficient to choose μ_k such that

$$(x^{(k+1)})^T \Pi(x^{(k)}) x^{(k+1)} = -\mu_k(\mu_k + 2)(x_r^{(k+1)})^T \Pi(x^{(k)})(x_r^{(k+1)}) \leq 0$$

If $(x_r^{(k+1)})^T \Pi(x^{(k)})(x_r^{(k+1)}) \leq 0$, then we need to choose μ_k such that $\mu_k(\mu_k + 2) \leq 0$. Since $\mu_k(\mu_k + 2)$ attains its minimum at $\mu_k = -1$, we can construct $x^{(k+1)} = x^{(k)} - (x^{(k)} - x_r^{(k+1)}) = x_r^{(k+1)}$ accordingly. If $(x_r^{(k+1)})^T \Pi(x^{(k)})(x_r^{(k+1)}) > 0$, then any positive value of μ_k is acceptable. If we choose $\mu_k = 1$, then we can construct $x^{(k+1)} = x^{(k)} + (x^{(k)} - x_r^{(k+1)}) = 2x^{(k)} - x_r^{(k+1)}$ accordingly. As a summary, we establish an extra step 4 as follows

$$x^{(k+1)} = \begin{cases} x_r^{(k+1)} & (x_r^{(k+1)})^T \Pi(x^{(k)})(x_r^{(k+1)}) \leq 0 \\ 2x^{(k)} - x_r^{(k+1)} & (x_r^{(k+1)})^T \Pi(x^{(k)})(x_r^{(k+1)}) > 0 \end{cases}$$

We can describe the IRLS algorithm based on $H_S(x)$ as follows

Algorithm 17 IRLS Based on $H_S(x)$

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m \times n}$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$

1. Let $W_k = -diag((H_S(x^{(k)}) + ln(\tilde{x}^{(k)}[i]))^{-1/2})$, where $\tilde{x}^{(k)}[i] = \frac{x^{(k)}[i]^2}{\|x^{(k)}\|_2^2}$
2. $q^{(k+1)} = (AW_k)^\dagger y$
3. $x_r^{(k+1)} = W_k q^{(k+1)}$
4. $x^{(k+1)} = \begin{cases} x_r^{(k+1)} & (x_r^{(k+1)})^T \Pi(x^{(k)})(x_r^{(k+1)}) \leq 0 \\ 2x^{(k)} - x_r^{(k+1)} & (x_r^{(k+1)})^T \Pi(x^{(k)})(x_r^{(k+1)}) > 0 \end{cases}$, where $\Pi(x^{(k)}) = W_k^{-2}$

Output: $x^{(\bar{k})}$

Note that the fixed points of algorithm 17 cannot generally be completely sparse; however, they do tend to have a large number of entries with very small (albeit nonzero) amplitudes.

Lastly, we relax the strict constraint $Ax = y$ so that we can tolerate some differences between Ax and y just as problem 4.2 does. However, here we consider adding a Tikhonov regularization term in-

stead of imposing an inequality constraint and we limit our discussion to the p-norm like diversity measure. Concretely speaking, we consider the following optimization problem

$$\inf_x J(x) = \gamma E^p(x) + \|Ax - b\|_2^2 \tag{4.11}$$

where γ is a positive regularization parameter and $\|Ax - b\|_2^2$ is the Tikhonov regularization term. γ controls the tradeoff between quality of fit $\|Ax - b\|$ and the degree of sparsity. Larger values of γ lead to sparser solutions while smaller values of γ lead to better fit, i.e., smaller error $\|Ax-b\|$. Besides the regularization viewpoint, we can also adopt a maximum a posteriori (MAP) interpretation of 4.11. Assume the signal model to be $y = Ax + e$, where $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ and $e \in \mathbb{R}^m$ are all considered to be random. We model the noise vector v as a Gaussian random vector with i.i.d. elements, i.e.,

$$P(e[i] = u) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{u^2}{2\sigma^2}}$$

where σ^2 is the noise variance. We assume the coefficient vector x to be independent of e and to be sparse. Hence, probability density functions that are concentrated near zero but also have heavy tails are appropriate to model x. Specifically, x is modeled as a random vector with i.i.d. elements that have a generalized Gaussian distribution, which is defined as

Definition 4.1.4 (generalized Gaussian distribution)**.**

$$P(x[i] = u) = \frac{p}{2\sqrt[p]{2}\beta\Gamma(\frac{1}{p})} \exp(-sgn(p)\frac{|u|^p}{2\beta^p}) \tag{4.12}$$

where $p \in \mathbb{R}$, $\beta > 0$, $\Gamma(\cdot)$ is the gamma function and $sgn(\cdot)$ is the sign function.

β is the generalized variance and p controls the shape of the generalized Gaussian distribution. When $p = 1$, the generalized Gaussian distribution reduces to the Laplacian distribution while when $p = 2, \beta = 1$, it reduces to the standard normal distribution. As we can see

from figure 4.1, the generalized Gaussian distribution moves towards a uniform distribution as $p \to \infty$ and towards a very peaky distribution as $p \to 0$. Given the vector $y \in \mathbb{R}^m$, we want to perform MAP

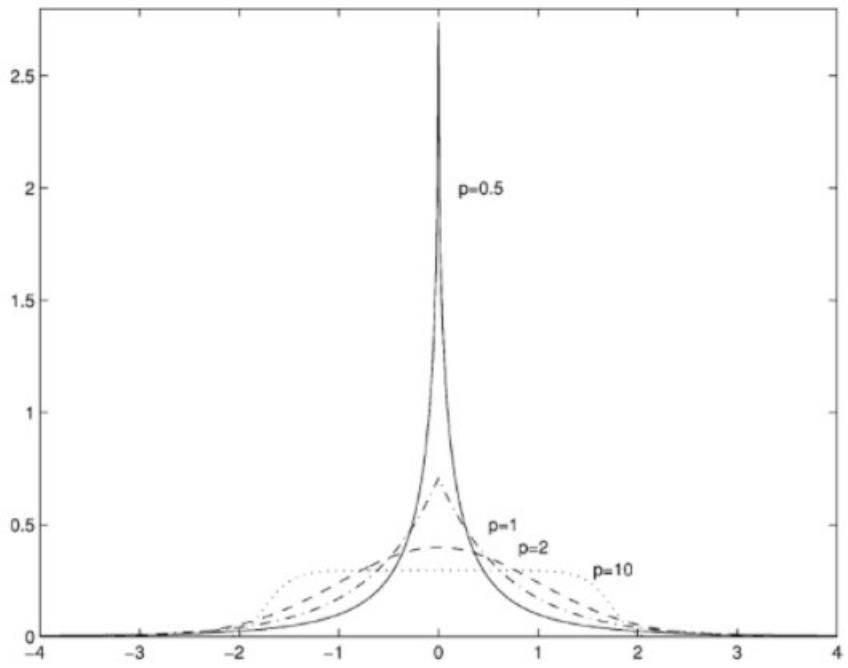

Figure 4.1: generalized Gaussian distribution

estimation of the vector $x \in \mathbb{R}^n$ as follows

$$\begin{aligned} x_{MAP} &= \underset{x\in\mathbb{R}^n}{argsup} \ln(p(x|y)) \\ &= \underset{x\in\mathbb{R}^n}{argsup}\{\ln(p(y|x)) + \ln(p(x))\} \\ &= \underset{x\in\mathbb{R}^n}{argsup}\{\ln(p_v(y - Ax)) + \ln(p(x))\} \\ &= \underset{x\in\mathbb{R}^n}{arginf}\, J(x) = \gamma E^p(x) + \|Ax - b\|_2^2 \quad with\ \gamma = \frac{\sigma^2}{\beta^p} \end{aligned}$$

which is equivalent to our objective 4.11. $p = 2$, which corresponds to the Gaussian prior for x, gives rise to a regularized least squares problem. With $p \leq 1$, it can be shown that the local minima of $J(x)$ is sparse.

Then how can we solve the problem 4.11? The gradient vector of $J(x)$ evaluated at a fixed point x^* can be computed as follows

$$\nabla_x J(x^*) = 2A^T Ax^* - 2A^T y + 2\lambda\Pi(x^*)x^* = 0$$

where $\lambda = \frac{|p|}{2}\gamma$ and $\Pi(x) = diag(|x[i]|^{p-2})$. From such equation for a fixed point, we can come up with the following fixed-point iteration rule.

$$2A^TAx^{(k+1)} - 2A^Ty + 2\lambda\Pi(x^{(k)})x^{(k+1)} = 0 \tag{4.13}$$

Hence, an algorithm designed based on 4.13 is equivalent as solving the following optimization problem for each iteration

$$x^{(k+1)} = \underset{x\in\mathbb{R}^n}{argin}f\, Q^{(k+1)}(x) = \lambda\|W_k^{-1}x\|_2^2 + \|Ax - y\|_2^2 \tag{4.14}$$

where $W(x) = \Pi(x)^{-1/2} = diag(|x[i]|^{1-p/2})$ and $W_k = W(x^{(k)})$. Note that 4.14 is just the same as replacing the strict equality constraint of 4.6 with the Tikhonov regularization term $\|Ax-y\|_2^2$. For this relation, we call the resulting algorithm a regularized IRLS algorithm. Indeed, as λ approaches zero, the algorithm reduces to the IRLS algorithm. From equation 4.13, we can analytically compute 4.14 as

$$x^{(k+1)} = W_k(A_k^TA_k + \lambda I)^{-1}A_k^Ty \tag{4.15}$$

where $A_k = AW_k$. Since $A_k^T(A_kA_k^T + \lambda I)^{-1} = (A_k^TA_k + \lambda I)^{-1}A_k^T$,

$$x^{(k+1)} = W_kA_k^T(A_kA_k^T + \lambda I)^{-1}y \tag{4.16}$$

We can describe the regularized IRLS algorithm based on the p-norm like diversity measure as follows. The readers can refer to [33] for details.

Algorithm 18 Regularized IRLS Based on $E^p(x)$

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$

1. Let $W_k = diag(|x^{(k)}[i]|^{1-p/2})$ and $A_k = AW_k$
2. $x^{(k+1)} = W_k(A_k^TA_k + \lambda I)^{-1}A_k^Ty = W_kA_k^T(A_kA_k^T + \lambda I)^{-1}y$

Output: $x^{(\bar{k})}$

4.1.2 FOCal Underdetermined System Solver (FOCUSS) Algorithm

It is easier to deal with minimization of ℓ_2 norm compared with minimization of ℓ_1 and ℓ_0 norm; however, simply minimizing $\|x\|_2$ subject to $Ax = y$ cannot produce sparse solutions. To deal with problem 4.1, the FOCUSS algorithm considers the following objective function at each iteration instead.

$$x^{(k+1)} = \underset{x \in \mathbb{R}^n}{argmin} \sum_{i=1, x^{(k)}[i] \neq 0}^{n} \left(\frac{x[i]}{x^{(k)}[i]}\right)^2 \quad subject\ to\ Ax = y \tag{4.17}$$

The intuition is that the relatively large entries in $x^{(k)}$ reduce the contribution of the corresponding elements of $x^{(k+1)}$ to the objective function. In this way, larger entries in $x^{(k)}$ result in larger corresponding entries in $x^{(k+1)}$ if the respective columns in A are significant in fitting b as compared to the rest of the columns of A. Hence, minimizing such objective function at each iteration gradually reinforces some of the already prominent entries in x while suppressing the rest. As we have mentioned in the last section, FOCUSS algorithm is just the same as the IRLS algorithm based on the Gaussian entropy diversity measure. We can verify this fact by observing equation 4.6 and algorithm 16. In [16], it generalizes the basic FOCUSS algorithm to a generalized FOCUSS algorithm. The generalized FOCUSS algorithm is described as follows

Algorithm 19 Generalized FOCUSS Algorithm

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m \times n}$; a positive integer ℓ ; $W_{ak} \in \mathbb{R}^{n \times n}$: the scaling matrix
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$

1. Let $W_{pk} = diag((x^{(k)})^\ell)$
2. Let $W_k = W_{ak} W_{pk}$
3. $q^{(k+1)} = (AW_k)^\dagger y$
4. $x^{(k+1)} = W_k q^{(k+1)}$

Output: $x^{(\bar{k})}$

We introduce an additional parameter ℓ so that the original affine scaling transformation matrix can be more flexible (which becomes the W_{pk}

matrix). We also introduce an additional scaling matrix W_{ak}, which is independent of W_{pk} and can carry a priori information. We can make the algorithm be more general by letting W_{pk} be $diag(\prod_{i=1}^{k}(x^{(i)})^{\ell_i})$, where ℓ_i, $i = 1, \cdots, k$ are all positive integers. In the following, we will discuss some convergence issues regarding the FOCUSS algorithm.

Convergence Issues

We need to introduce some terms in advance.

1. Fixed points : We have introduce the concept of fixed points in footnote 1. We can further classify fixed points into three categories, which are stable fixed points, saddle fixed points and unstable fixed points respectively.
   (a) Stable fixed points : fixed points to which the algorithm converges from anywhere within some closed neighborhood around such points
   (b) Saddle fixed points : fixed points to which the algorithm converges only along some special trajectories
   (c) Unstable fixed points : fixed points from which an algorithm moves away given any perturbation
2. Basin of attraction : the largest neighborhood of points from which an algorithm converges to a given stable fixed point
3. Phase space : a collection of trajectories that trace the temporal evolution of an algorithm from different initial points
4. Fixed point convergence / Absolute convergence / Absolute stability : these terms mean that an algorithm converges to a fixed point from any starting point

For a fixed-point convergent algorithm, its entire phase space is divided up by the basins of attraction containing stable fixed points. The borders separating individual basins do not belong to any of the basins. These borders can consist of trajectories leading to saddle fixed

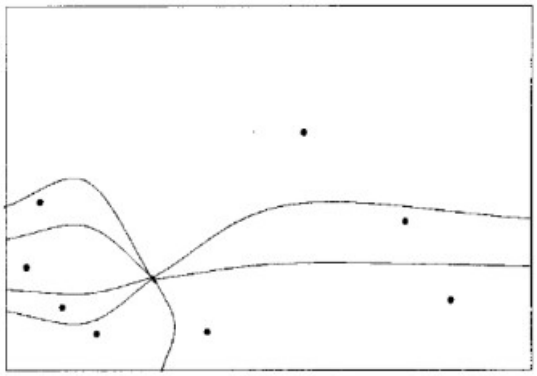

Figure 4.2: Visualization of a phase space of FOCUSS

points or to infinity, or they can be a dense set of unstable fixed points, or they can be a combination of those. As a result, an absolutely convergent algorithm is not guaranteed to converge to a stable fixed point from any initialization point; that is, it may converge to a saddle fixed point or get stuck at an unstable fixed point. However, since the saddle fixed points are reached only along special trajectories whose total number has measure zero, an algorithm converges to these solutions with probability 0 . The unstable fixed points also have measure zero; therefore, an algorithm returns these points with probability 0. Hence, an absolutely convergent algorithm converges to stable fixed points with probability 1. We can refer to a figure in [16] (i.e., figure 4.2) for a visualization of a phase space of FOCUSS. The phase space is divided up by basins of attraction. The dots represent the stable fixed points of the algorithm and the lines indicate the boundaries of the basins. The FOCUSS algorithm is proved to be absolutely convergent. Furthermore, sparse solutions in $\mathbb{R}^r$, $r \leq m$, are proved to be the stable fixed points of the FOCUSS algorithm. Non sparse solutions in $\mathbb{R}^r$, $m < r < n$, are saddle fixed points of the FOCUSS algorithm. Non sparse solutions in $\mathbb{R}^n$ are unstable fixed points of the FOCUSS algorithm. Therefore, the FOCUSS algorithm converges to a sparse solution with probability 1. Besides, since the entries of $x^{(k)}$ converges to zero or nonzero values and $x^{(k+1)} = W_k q^{(k+1)}$, where $W_k = diag(x^{(k)})$ (assuming the basic FOCUSS algorithm), the entries of $q^{(k)}$ converge to zeros or ones. Hence $q^{(k)}$ can serve as an indicator when k approaches the infinity, which indicates the support of a sparse solution. As we

expect, the basins of attraction play an important role in the convergence issues. In the following, we discuss some factors that affect the basins of attraction.

Factors That Affect The Basins of Attraction

First, we discuss the effect of the dictionary A on the basins. Let $A = A_n N$, where A_n is normalized version of A (i.e., ℓ_2 norm of each column of A_n is normalized to 1) and N is a diagonal matrix whose diagonal entries are the corresponding normalizing factors. Hence, A_n affects the solution only through the degree of correlation of individual columns with the vector y. Note that

$$\begin{aligned} &\inf_x \|W_{pk}^{-1} x\|_2^2 \quad \textit{subject to } Ax = y \\ \equiv &\inf_q \|q\|_2^2 \quad \textit{subject to } A_n N W_{pk} q = b \end{aligned} \tag{4.18}$$

where W_{pk} is the same as the matrix defined in algorithm 19. As a result, we can view 4.18 as the generalized FOCUSS algorithm applied to the system $A_n x = y$ with W_{ak} being N. Originally, the size of each basin is equal due to the characteristic of A_n. However, with such W_{ak}, stable fixed points whose support containing locations of large entries of N are favored, which in turn makes the corresponding basins larger. Therefore, to alleviate such intrinsic bias of the dictionary A, we had better normalize the ℓ_2 norm of each column of A in advance before applying the FOCUSS algorithm.

Second, we discuss the effect of the number of sparse solutions on the basins. If the number of sparse solutions to a given problem is larger, the greater the fragmentation of the phase space of the FOCUSS algorithm is greater, which leads to smaller basins. Therefore, in the context of compressive sensing, as the sizes of individual basins diminish, the algorithm must start closer to the real sparse signal in order to converge to it.

Lastly, we discuss the effect of the dimensions of stable fixed points on the basins. If all stable fixed points are m-dimensional, each m-dimensional subspace has an equal probability of being the solution

(if the intrinsic bias has been removed); that is, all basins are equal. However, if a stable fixed point x of r-dimension ($r < m$) exists, those solutions of m-dimension, whose support contains the support of x, will no longer be present and an initialization that would have led to one of those solutions now leads to x. Hence, the basin of x is larger than any basins of m-dimensional sparse solution. We can conclude that the smaller the dimension of a stable fixed point, the larger the size of the corresponding basin. Therefore, the maximally sparse solution has the largest basin, which implies that the FOCUSS algorithm favors the maximally sparse solution and the smaller the dimension of the maximally sparse solution, the greater the likelihood of convergence to it. On the other hand, if the dimension of the sparse solutions are larger, the sizes of the corresponding basins are smaller. As a result, as the dimension of the sparse solution increases, the FOCUSS algorithm gradually starts to favor the solution near its initialization and it must start closer to the real sparse signal in order to lie in the right basin.

Finally, we list three useful tips when performing the FOCUSS algorithm. The first tip is that we can initialize $x^{(0)}$ as the solution of the minimum ℓ_2 norm problem (i.e., $\inf_{x \in \mathbb{R}^n} \|x\|_2^2$ subject to $Ax = y$). In this way, $x^{(0)} = A^\dagger y$. The second tip is that we can eliminate diminishing entries of $x^{(k)}$ that are indicated by $q^{(k)}$ at each iteration. The third tip is to implement a hard thresholding operation to obtain the final result once the convergence pattern becomes clear. The readers can refer to [16] for details.

4.1.3 ℓ_1 Convex Relaxation

In section 4.1.1, we have introduced IRLS-type algorithms with $D(x)$ being the p-norm like diversity measure $E^p(x)$. In particular, if we choose $p = 1$, we will obtain $E^1(x) = \|x\|_1$ and problem 4.3 becomes

$$\inf_{x \in \mathbb{R}^n} \|x\|_1 \quad \textit{subject to } Ax = y \tag{4.19}$$

In the field of sparse representation and compressive sensing, this problem, known as the basis pursuit, gains much attention for it can be an adequate surrogate of the ℓ_0-minimization problem. Indeed, if the basis pursuit has an unique solution $x^\sharp$, then theorem 3.1 of [13] proves that the system $\{a_j, j \in supp(x^\sharp)\}$ is linearly independent, which implies $x^\sharp$ must be m-sparse. Another reason to consider 4.19 is due to the convexity of the ℓ_1 norm. It can be verified that ℓ_p quasinorm for $0 < p < 1$ is non-convex and ℓ_p norm for $p \geq 1$ is convex. Solving a non-convex ℓ_p-minimization problem for $0 < p < 1$ is NP-hard in general. Hence, it is natural to consider the ℓ_1-minimization problem 4.19 for it is the "closest" convex problem to the non-convex ℓ_0-minimization problem. For this reason, we call the ℓ_1-minimization problem the convex relaxation of the ℓ_0-minimization problem.

Parallel to the extension of the problem 4.1 to the problem 4.2, it is also natural to extend the problem 4.19 to the following problem 4.20.

$$\inf_{x \in \mathbb{R}^n} \|x\|_1 \quad subject\ to\ \|Ax - y\|_2 \leq \epsilon \tag{4.20}$$

We call it the quadratically-constrained basis pursuit. Note that the quadratically-constrained basis pursuit has strong connections with another two famous problems, which are the basis pursuit denoising 4.21 and the least absolute shrinkage and selection operator (LASSO) 4.22, respectively.

$$\inf_{x \in \mathbb{R}^n} \lambda\|x\|_1 + \frac{1}{2}\|Ax - y\|_2^2\ , \lambda \geq 0 \tag{4.21}$$

$$\inf_{x \in \mathbb{R}^n} \|Ax - y\|_2 \quad subject\ to\ \|x\|_1 \leq \tau\ , \tau \geq 0 \tag{4.22}$$

Indeed, proposition 3.2 of [13] manifests some links among the three problems.

Theorem 3.1 of [13] gives an intuition of why we want to consider the basis pursuit. In the following, we will continue to delve into more involved mathematical discussion about the behavior of the basis pursuit and the quadractically-constrained basis pursuit. First, we define the ℓ_q-robust null space property of a matrix A as follows.

Definition 4.1.5 (ℓ_q-robust null space property)**.** Given $q \geq 1$, the matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy the ℓ_q-robust null space property of order s with respect to the norm $\|\cdot\|$ with constants $0 < \rho < 1$ and $\tau > 0$ if, for any set $S \subset [n]$ with $card(S) \leq s$,

$$\|v|_S\|_q \leq \frac{\rho}{s^{1-1/q}}\|v|_{\overline{S}}\|_1 + \tau\|Av\| \quad \forall v \in \mathbb{R}^n \tag{4.23}$$

In particular, if $q = 1$, then the ℓ_q-robust null space property reduces to the robust null space property.

Definition 4.1.6 (robust null space property)**.** The matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy the robust null space property of order s with respect to the norm $\|\cdot\|$ with constants $0 < \rho < 1$ and $\tau > 0$ if, for any set $S \subset [n]$ with $card(S) \leq s$,

$$\|v|_S\|_1 \leq \rho\|v|_{\overline{S}}\|_1 + \tau\|Av\| \quad \forall v \in \mathbb{R}^n \tag{4.24}$$

If 4.24 holds only for all v that belongs to the null space of A ($ker\, A$), then the robust null space property reduces to the stable null space property.

Definition 4.1.7 (stable null space property)**.** The matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy the stable null space property of order s with constant $0 < \rho < 1$ if, for any set $S \subset [n]$ with $card(S) \leq s$,

$$\|v|_S\|_1 \leq \rho\|v|_{\overline{S}}\|_1 \quad \forall v \in ker\, A \tag{4.25}$$

If we further choose ρ to be infinitely close to 1, then the stable null space property reduces to the null space property.

Definition 4.1.8 (null space property)**.** The matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy the null space property of order s if, for any set $S \subset [n]$ with $card(S) \leq s$,

$$\|v|_S\|_1 < \|v|_{\overline{S}}\|_1 \quad \forall v \in ker\, A\backslash\{0\} \tag{4.26}$$

If we add $\|v|_S\|_1$ on the both sides of 4.26, we will get an equivalent

condition

$$2\|v|_S\|_1 < \|v\|_1 \quad \forall v \in ker\, A \backslash \{0\} \tag{4.27}$$

If we add $\|v|_{\overline{S}}\|_1$ on the both sides of 4.26 and choose S to be $L_s(v)$ as defined in 2.66, we will get another equivalent condition

$$\|v\|_1 < 2\sigma_s(v)_1 \quad \forall v \in ker\, A \backslash \{0\} \tag{4.28}$$

Those properties have a lot to do with the effectiveness of the basis pursuit and quadratically-constrained basis pursuit as being the surrogate problems of 4.1 and 4.2. We excerpt some important results introduced in chapter 4 of [13] in the following. Once again, as have been remarked in footnotes 1 and 2, to make clear of the concept and to avoid conflicts of use of notations, we modify the notations in 4.19 and 4.20 to

$$\inf_{z \in \mathbb{R}^n} \|z\|_1 \quad subject\ to\ Az = y$$
$$\inf_{z \in \mathbb{R}^n} \|z\|_2 \quad subject\ to\ \|Az - y\|_2 = \|A(z - x) - e\|_2 \leq \epsilon$$

First, we excerpt the theorem 4.25 of [13] as follows.

Theorem 4.1.1. *Given $1 \leq p \leq q$, suppose that the matrix $A \in \mathbb{R}^{m \times n}$ satisfies the ℓ_q-robust null space property of order s with respect to the norm $\|\cdot\|$ with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x, z \in \mathbb{R}^n$,*

$$\|z - x\|_p \leq \frac{C}{s^{1-1/p}} \left(\|z\|_1 - \|x\|_1 + 2\sigma_s(x)_1\right) + Ds^{1/p-1/q}\|A(z - x)\| \tag{4.29}$$

where $C := (1 + \rho)^2/(1 - \rho)$ and $D := (3 + \rho)\tau/(1 - \rho)$

If we let q to be 2 and z to be a solution $x^\sharp$ of the quadractically-constrained basis pursuit problem, we can have the following corollary.

Corollary 4.1.1.1. *Suppose that the matrix $A \in \mathbb{R}^{m \times n}$ satisfies the ℓ_2-robust null space property of order s with respect to the ℓ_2 norm with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x \in \mathbb{R}^n$, a solution $x^\sharp$ of the quadractically-constrained basis pursuit, $y =$*

$Ax + e$, and $\|e\|_2 \leq \epsilon$ approximates the vector x with ℓ_p-error

$$\|x - x^\sharp\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\epsilon \quad , 1 \leq p \leq 2 \tag{4.30}$$

where $C := 2(1+\rho)^2/(1-\rho)$ and $D := 2\tau(3+\rho)/(1-\rho)$

We excerpt another theorem related to the robust null space property as follows.

Theorem 4.1.2. *The matrix $A \in \mathbb{R}^{m\times n}$ satisfies the robust null space property of order s with respect to the norm $\|\cdot\|$ with constants $0 < \rho < 1$ and $\tau > 0$ if and only if, for any set $S \subset [n]$ with $card(S) \leq s$,*

$$\|z-x\|_1 \leq \frac{1+\rho}{1-\rho}\left(\|z\|_1 - \|x\|_1 + 2\|x|_{\overline{S}}\|_2\right) + \frac{2\tau}{1-\rho}\|A(z-x)\| \tag{4.31}$$

for all vectors $x, z \in \mathbb{R}^n$.

If we choose S to be $L_s(x)$ and z to be a solution $x^\sharp$ of the quadractically-constrained basis pursuit problem, we can have the following corollary.

Corollary 4.1.2.1. *Suppose that a matrix $A \in \mathbb{R}^{m\times n}$ satisfies the robust null space property of order s with respect to the ℓ_2 norm with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x \in \mathbb{R}^n$, a solution $x^\sharp$ of the quadractically-constrained basis pursuit, $y = Ax + e$, and $\|e\|_2 \leq \epsilon$ approximates the vector x with ℓ_1-error*

$$\|x - x^\sharp\|_1 \leq \frac{2(1+\rho)}{(1-\rho)}\sigma_s(x)_1 + \frac{4\tau}{1-\rho}\epsilon \tag{4.32}$$

If we restrict the x and z in theorem 4.1.2 to satisfy $Az = Ax$, then we can come up with the following corollary.

Corollary 4.1.2.2. *The matrix $A \in \mathbb{R}^{m\times n}$ satisfies the stable null space property of order s with constant $0 < \rho < 1$ if and only if,*

for any set $S \subset [n]$ *with* $card(S) \leq s$,

$$\|z - x\|_1 \leq \frac{1+\rho}{1-\rho} \left(\|z\|_1 - \|x\|_1 + 2\|x|_{\overline{S}}\|_1 \right) \tag{4.33}$$

for all vectors $x, z \in \mathbb{R}^n$ *with* $Az = Ax$.

If we choose S to be $L_s(x)$ and z to be a solution $x^\sharp$ of the basis pursuit problem, we can have the following corollary.

Corollary 4.1.2.3. *Suppose that a matrix* $A \in \mathbb{R}^{m \times n}$ *satisfies the stable null space property of order* s *with constant* $0 < \rho < 1$. *Then, for any* $x \in \mathbb{R}^n$, *a solution* $x^\sharp$ *of the basis pursuit with* $y = Ax$ *approximates the vector* x *with* ℓ_1*-error*

$$\|x - x^\sharp\|_1 \leq \frac{2(1+\rho)}{(2-\rho)} \sigma_s(x)_1 \tag{4.34}$$

Finally we excerpt theorem 4.5 of [13] as follows.

Theorem 4.1.3. *Given a matrix* $A \in \mathbb{R}^{m \times n}$, *every* s*-sparse vector* $x \in \mathbb{R}^n$ *is the unique solution of the basis pursuit with* $y = Ax$, *i.e.,* $x^\sharp = x$, *if and only if* A *satisfies the null space property of order* s.

This theorem shows that exact sparse recovery can be achieved if the sampling matrix A satisfies the null space property.

Section 6.2 of [13] presents some theorems that use constraints on the restricted isometry constant as sufficient conditions for satisfaction of those null space properties so that some error bounds are guaranteed. Precisely, theorem 6.9 of [13] states that if $\delta_{2s} < 1/3$, then the null space property of order s is satisfied. Hence, combining the theorem 4.1.3, we can have the following result.

Theorem 4.1.4. *Suppose that the 2s-th restricted isometry constant of the matrix* $A \in \mathbb{R}^{m \times n}$ *satisfies*

$$\delta_{2s} < \frac{1}{3} \tag{4.35}$$

Then every s-sparse vector $x \in \mathbb{R}^n$ is the unique solution of the basis pursuit.

Theorem 6.13 of [13] states that if $\delta_{2s} < \frac{4}{\sqrt{41}} \approx 0.6246$, then the ℓ_2-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$ depending only on δ_{2s} is satisfied. Combining with the corollary 4.1.1.1, we can have the following result.

Theorem 4.1.5. *Suppose that the 2s-th restricted isometry constant of the matrix $A \in \mathbb{R}^{m\times n}$ satisfies*

$$\delta_{2s} < \frac{4}{\sqrt{41}} \approx 0.6246 \tag{4.36}$$

Then, for any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ with $\|Ax - y\|_2 \leq \epsilon$, a solution $x^\sharp$ of the quadractically-constrained basis pursuit approximates the vector x with ℓ_p-error

$$\|x - x^\sharp\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\epsilon \quad , 1 \leq p \leq 2$$

where the constants $C, D > 0$ depend only on δ_{2s}.

Up to now, we have provided solid mathematical foundation on ℓ_1-minimization problems. We can solve the basis pursuit, the quadractically-constrained basis pursuit, the basis pursuit denoising and the LASSO by resorting to various algorithms we have introduced, e.g. the Chambolle and Pock's primal-dual algorithm, the ADMM and the proximal gradient method, in chapter 3 or the IRLS-type algorithms and FOCUSS algorithm in chapter 4. In particular, we want to introduce the use of the proximal gradient method to solve the basis pursuit denoising problem because it results in a famous algorithm called the iterative shrinkage-thresholding algorithm (ISTA) or the iterative soft-thresholding.

We apply algorithm 5 to the basis pursuit denoising problem 4.21. Let $f(x) = \frac{1}{2}\|Ax - y\|_2^2$ and $g(x) = \lambda\|x\|_1$. The gradient of $f(x)$ is $\nabla f(x) = A^*(Ax - y)$. The proximal mapping for $g(x)$ is the soft

thresholding operator S_λ introduced in 1.22 and is applied entry-wise. Therefore, the iteration rule 3.43 becomes

$$x^{(k+1)} = S_{\eta\lambda}(x^{(k)} + \eta A^*(y - Ax^{(k)})) \tag{4.37}$$

The resulting algorithm is called the ISTA or the iterative soft-thresholding.

4.2 Greedy Algorithms

For a sparse representation or compressive sensing problem, an important and challenging task is to identify the support of the sparse representation coefficient vector (or the sparse signal). Greedy Algorithms attempt to identify indices of some nonzero components at each iteration based on some greedy rules (typically involving the calculation of the correlations of the residual vector with columns of the dictionary (or sampling matrices) or equivalently involving the formation of the correlation vector) . During the progress of iterations, our aim is to obtain residual vectors with gradually decreasing norms. Many effective greedy algorithms have been proved to achieve such goal. In this section, we will introduce various such effective greedy algorithms.

4.2.1 Matching Pursuit (MP)

Let $y \in \mathbb{R}^m$ be the signal vector and $A \in \mathbb{R}^{m\times n}$ $(m < n)$ be the dictionary. Let $a_j \in \mathbb{R}^m$ represent the j-th column vector of A and $\|a_j\|_2^2 = \langle a_j, a_j\rangle = 1\ \forall j \in [n]$. As have been mentioned in the introduction of this chapter, our objective is to achieve a sparse representation of y so that we can use as few atoms as possible to approximate it. That is, we want to decompose the signal y into a linear expansion of atoms that are selected from the redundant dictionary A and these atoms are chosen in order to best match the signal structure.

In the following, we will first introduce the matching pursuit algorithm and then we will provide a theoretical analysis of it.

Algorithm 20 Matching Pursuit

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; a prescribed threshold parameter $0 < \epsilon < 1$
Initialization :

1. Compute $\langle a_j, a_k\rangle$ for $j, k \in [n]$
2. Compute $\langle y, a_j\rangle$ for $j \in [n]$
3. $R^{(0)} = y$

Iterations : from $k = 0$ **until** $\sum_{k=0}^{\bar{k}} |\langle R^{(k)}, a_{j^{k+1}}\rangle|^2 \geq (1-\epsilon^2)\|y\|_2^2$ **at** $k = \bar{k}$

1. $j^{k+1} = \underset{j\in[n]}{argmax}\, |\langle R^{(k)}, a_j\rangle|$
2. $R^{(k)} = \langle R^{(k)}, a_{j^{k+1}}\rangle a_{j^{k+1}} + R^{(k+1)} \Rightarrow R^{(k+1)} = R^{(k)} - \langle R^{(k)}, a_{j^{k+1}}\rangle a_{j^{k+1}}$
3. Compute $\langle R^{(k+1)}, a_j\rangle = \langle R^{(k)}, a_j\rangle - \langle R^{(k)}, a_{j^{k+1}}\rangle\langle a_{j^{k+1}}, a_j\rangle$ for $j \in [n]$

$$y = \sum_{k=0}^{\bar{k}} (R^{(k)} - R^{(k+1)}) + R^{(\bar{k}+1)} = \sum_{k=0}^{\bar{k}} \langle R^{(k)}, a_{j^{k+1}}\rangle a_{j^{k+1}} + R^{(\bar{k}+1)}$$

Perform back-projection

1. Let $X = [x^{(k)}], k = 1, \cdots, \bar{k}+1$; $Y = [\langle R^{(\bar{k}+1)}, a_{j^{k+1}}\rangle], k = 0, \cdots, \bar{k}$; $G \in \mathbb{R}^{(\bar{k}+1)\times(\bar{k}+1)}$, where $G_{r\ell} = \langle a_{j^r}, a_{j^\ell}\rangle$, $r, \ell = 1, \cdots, \bar{k}+1$
2. Compute the linear system $Y = GX$ to get X

Output: $(j^{k+1}, \langle R^{(k)}, a_{j^{k+1}}\rangle + x^{(k+1)})$ for $k = 0, 1, \cdots, \bar{k}$

Hence, the matching pursuit chooses at each iteration an atom that best matches the residue $R^{(k)}$ and subdecomposes $R^{(k)}$ by projecting it on such best atom. We analyze the matching pursuit in details as follows. First, from the algorithm 20, we know that

$$R^{(k)} = \langle R^{(k)}, a_{j^{k+1}}\rangle a_{j^{k+1}} + R^{(k+1)} \tag{4.38}$$

$$y = \sum_{k=0}^{\bar{k}} \langle R^{(k)}, a_{j^{k+1}}\rangle a_{j^{k+1}} + R^{(\bar{k}+1)} \tag{4.39}$$

We want to derive similar relations but in the form of energy. We make some derivations as follows.

$$\begin{aligned} R^{(k)} &= \langle R^{(k)}, a_{j^{k+1}}\rangle a_{j^{k+1}} + R^{(k+1)} \\ \Rightarrow \langle R^{(k)}, a_{j^{k+1}}\rangle &= \langle R^{(k)}, a_{j^{k+1}}\rangle\langle a_{j^{k+1}}, a_{j^{k+1}}\rangle + \langle R^{(k+1)}, a_{j^{k+1}}\rangle \\ &= \langle R^{(k)}, a_{j^{k+1}}\rangle + \langle R^{(k+1)}, a_{j^{k+1}}\rangle \end{aligned}$$

Hence, $\langle R^{(k+1)}, a_{j^{k+1}}\rangle = 0$, which means $R^{(k+1)}$ is orthogonal to $a_{j^{k+1}}$. Therefore, we can come up with the following equation.

$$\|R^{(k)}\|_2^2 = |\langle R^{(k)}, a_{j^{k+1}}\rangle|^2 + \|R^{(k+1)}\|_2^2 \tag{4.40}$$

Furthermore, since we can express $\|y\|_2^2$ as $\|y\|_2^2 = \sum_{k=0}^{\bar{k}}(\|R^{(k)}\|_2^2 - \|R^{(k+1)}\|_2^2) + \|R^{(\bar{k}+1)}\|_2^2$, we can derive another relation below from 4.40.

$$\|y\|_2^2 = \sum_{k=0}^{\bar{k}} |\langle R^{(k)}, a_{j^{k+1}}\rangle|^2 + \|R^{(\bar{k}+1)}\|_2^2 \tag{4.41}$$

If we define $\lambda(y) = \sup_{j\in[n]} \frac{|\langle y,a_j\rangle|}{\|y\|_2}$ as the correlation ratio of the signal y with respect to the dictionary. Since $|\langle R^{(k)}, a_{j^{k+1}}\rangle| = \sup_{j\in[n]} |\langle R^{(k)}, a_j\rangle|$ and 4.40, we can derive that

$$\|R^{(k+1)}\|_2^2 = (1 - \lambda(R^{(k)})^2)\|R^{(k)}\|_2^2 \tag{4.42}$$

We can further derive that

$$\|R^{(\bar{k}+1)}\|_2 = \|y\|_2 \prod_{k=0}^{\bar{k}} (1 - \lambda(R^{(k)})^2)^{1/2} \tag{4.43}$$

Therefore, the norm of $R^{(\bar{k}+1)}$ depends on the correlation between the residues and the dictionary atoms. If the signal y is the sum of several high energy components that belong to the dictionary, the correlation ratios of y and its residues is large so that the norm of residues decay quickly. These components can be viewed as "coherent structures" with respect to the dictionary. On the other hand, if the residues of y have low correlation ratios, then their norms decay slowly. In such situation, y must be expanded over many atoms in order to approximate y well, which means the information of y is spread over the dictionary.

Let $V_{\bar{k}}$ denote the space spanned by the vectors $a_{j^{k+1}}, k = 0, 1, \cdots, \bar{k}$ and $P_{V_{\bar{k}}}$ denote the orthogonal projector on $V_{\bar{k}}$. Let $W_{\bar{k}}$ denote the orthogonal complement of $V_{\bar{k}}$ and $P_{W_{\bar{k}}}$ denote the orthogonal projector on $W_{\bar{k}}$. After we choose $\bar{k}+1$ vectors $a_{j^{k+1}}$, the closest vector to y is $P_{V_{\bar{k}}}y$. Hence, we perform an additional back projection step. Because

of 4.39, we can derive

$$P_{V_{\bar{k}}} y = \sum_{k=0}^{\bar{k}} \langle R^{(k)}, a_{j^{k+1}} \rangle a_{j^{k+1}} + P_{V_{\bar{k}}} R^{(\bar{k}+1)} \tag{4.44}$$

Let $P_{V_{\bar{k}}} R^{(\bar{k}+1)} = \sum_{k=0}^{\bar{k}} x^{(k+1)} a_{j^{k+1}}$.

$$\begin{aligned} \sum_{k=0}^{\bar{k}} x^{(k+1)} \langle a_{j^{k+1}}, a_{j^{\ell}} \rangle &= \langle P_{V_{\bar{k}}} R^{(\bar{k}+1)}, a_{j^{\ell}} \rangle \\ &= \langle P_{V_{\bar{k}}} R^{(\bar{k}+1)}, P_{V_{\bar{k}}} a_{j^{\ell}} \rangle \\ &= \langle R^{(\bar{k}+1)}, a_{j^{\ell}} \rangle \ , \ell = 1, \cdots, \bar{k}+1 \end{aligned}$$

Hence, we can derive the back projection step as what is described in the algorithm 20. The resulting approximation error is $y - P_{V_{\bar{k}}} y = P_{W_{\bar{k}}} y = P_{W_{\bar{k}}} R^{(\bar{k}+1)}$ and its energy is $\|P_{W_{\bar{k}}} R^{(\bar{k}+1)}\|_2^2 = \|R^{(\bar{k}+1)}\|_2^2 - \|P_{V_{\bar{k}}} R^{(\bar{k}+1)}\|_2^2$. Note that before performing the back projection, the approximation error is $\|R^{(\bar{k}+1)}\|_2^2$. Therefore, the reduction of the approximation error depends on $\|P_{V_{\bar{k}}} R^{(\bar{k}+1)}\|_2^2$.

Finally, we make a remark on the stopping criterion. The number of times we subdecompose the residues, i.e., $\bar{k}+1$ depends on a prescribed precision threshold ϵ. We require that

$$\|R^{(\bar{k}+1)}\|_2 = \|y - \sum_{k=0}^{\bar{k}} \langle R^{(k)}, a_{j^{k+1}} \rangle a_{j^{k+1}} \|_2 \leq \epsilon \|y\|_2 \tag{4.45}$$

, equivalently, we require that $\|R^{(\bar{k}+1)}\|_2^2 = \|y\|_2^2 - \sum_{k=0}^{\bar{k}} |\langle R^{(k)}, a_{j^{k+1}} \rangle|^2 \leq \epsilon^2 \|y\|_2^2$. Hence, the stopping criterion would be

$$\sum_{k=0}^{\bar{k}} |\langle R^{(k)}, a_{j^{k+1}} \rangle|^2 \geq (1 - \epsilon^2) \|y\|_2^2 \tag{4.46}$$

at $k = \bar{k}$.

The readers can refer to [23] for materials introduced in this sec-

tion.

4.2.2 Orthogonal Matching Pursuit (OMP)

In the context of sparse representation, let $y \in \mathbb{R}^m$ be the signal vector and $A \in \mathbb{R}^{m \times n} (m < n)$ be the dictionary. Let $a_j \in \mathbb{R}^m$ represent the j-th column vector of A and $\|a_j\|_2^2 = \langle a_j, a_j \rangle = 1 \; \forall j \in [n]$. The same as we have been mentioned in the last section, we hope to achieve compact signal coding so that we can use as few dictionary atoms as possible to approximate y. These atoms are chosen in order to best match the signal structure. Assume at the k-th iteration, we have chosen k atoms $\{a_{j^1}, a_{j^2}, \cdots, a_{j^k}\}$. Let V_k denote the space spanned by the set of vectors $\{a_{j^1}, a_{j^2}, \cdots, a_{j^k}\}$ and P_{V_k} denote the orthogonal projector on V_k. Let W_k denote the orthogonal complement of V_k and P_{W_k} denote the orthogonal projector on W_k. The best approximation of y using the set of vectors is $P_{V_k} y$. That is,

$$y = P_{V_k} y + P_{W_k} y = \sum_{n=1}^{k} x_n^k a_{j^n} + R^{(k)} \tag{4.47}$$

with $\langle R^{(k)}, a_{j^n} \rangle = 0$ for $n = 1, \cdots, k$. x_n^k denotes the coefficient for a_{j^n} at the k-th iteration and $R^{(k)}$ denotes the residue at the k-th iteration. Assume at the k+1-th iteration, we choose the atom $a_{j^{k+1}}$. Then the best approximation of y using $\{a_{j^1}, a_{j^2}, \cdots, a_{j^{k+1}}\}$ is $P_{V_{k+1}} y$. Similarly, we can get

$$y = P_{V_{k+1}} y + P_{W_{k+1}} y = \sum_{n=1}^{k+1} x_n^{k+1} a_{j^n} + R^{(k+1)} \tag{4.48}$$

with $\langle R^{(k+1)}, a_{j^n} \rangle = 0$ for $n = 1, \cdots, k+1$.

How can we update the coefficient for a_{j^n} from x_n^k to x_n^{k+1} and also obtain the coefficient for $a_{j^{k+1}}$, i.e., x_{k+1}^{k+1} ? We can expand $a_{j^{k+1}}$ as

$$a_{j^{k+1}} = P_{V_k} a_{j^{k+1}} + P_{W_k} a_{j^{k+1}} = \sum_{n=1}^{k} b_n^k a_{j^n} + \gamma^k \tag{4.49}$$

with $\langle \gamma^k, a_{j^n} \rangle = 0$ for $n = 1, \cdots, k$. In this way,

$$\sum_{n=1}^{k} x_n^k a_{j^n} + R^{(k)} - \sum_{n=1}^{k+1} x_n^{k+1} a_{j^n} - R^{(k+1)} = 0$$

$$\Rightarrow \sum_{n=1}^{k} (x_n^k - x_n^{k+1}) a_{j^n} - x_{k+1}^{k+1} a_{j^{k+1}} + R^{(k)} - R^{(k+1)} = 0$$

$$\Rightarrow \sum_{n=1}^{k} (x_n^k - x_n^{k+1} - b_n^k x_{k+1}^{k+1}) a_{j^n} - \gamma^k x_{k+1}^{k+1} + R^{(k)} - R^{(k+1)} = 0$$

$$\Rightarrow \sum_{n=1}^{k} (x_n^k - x_n^{k+1} - b_n^k x_{k+1}^{k+1}) \langle a_{j^n}, a_{j^{k+1}} \rangle - x_{k+1}^{k+1} \langle \gamma^k, a_{j^{k+1}} \rangle + \langle R^{(k)}, a_{j^{k+1}} \rangle = 0$$

$$\Rightarrow \begin{cases} x_n^k - x_n^{k+1} - b_n^k x_{k+1}^{k+1} = 0 \\ -x_{k+1}^{k+1} \langle \gamma^k, a_{j^{k+1}} \rangle + \langle R^{(k)}, a_{j^{k+1}} \rangle = 0 \end{cases}$$

Hence, $x_{k+1}^{k+1} = \frac{\langle R^{(k)}, a_{j^{k+1}} \rangle}{\langle \gamma^k, a_{j^{k+1}} \rangle}$ and $x_n^{k+1} = x_n^k - b_n^k x_{k+1}^{k+1}$ for $n = 1, \cdots, k$. How can we compute the coefficients b_n^k for $n = 1, \cdots, k$? From 4.49, we can get

$$\langle a_{j^{k+1}}, a_{j^i} \rangle = \sum_{n=1}^{k} b_n^k \langle a_{j^n}, a_{j^i} \rangle + \langle \gamma^k, a_{j^i} \rangle = \sum_{n=1}^{k} b_n^k \langle a_{j^n}, a_{j^i} \rangle \quad , i = 1, \cdots, k$$

Let $B_k = [b_n^k]$, $n = 1, \cdots, k$; $Y_k = [\langle a_{j^{k+1}}, a_{j^i} \rangle]$, $i = 1, \cdots, k$; $G_k \in \mathbb{R}^{k \times k}$, $(G_k)_{r\ell} = \langle a_{j^r}, a_{j^\ell} \rangle$, $r, \ell = 1, \cdots, k$. Then we form a linear system $Y_k = G_k B_k$. Since the set $\{a_{j^1}, a_{j^2}, \cdots, a_{j^k}\}$ forms a basis for V_k and the orthogonal projection of $a_{j^{k+1}}$ onto V_k is unique, there is exactly one solution for B_k, i.e., $B_k = G_k^{-1} Y_k$. Note that $G_k = \begin{bmatrix} G_{k-1} & Y_{k-1} \\ Y_{k-1}^* & 1 \end{bmatrix}$. Using the block matrix inversion formula, we can get $G_k^{-1} = \begin{bmatrix} G_{k-1}^{-1} + \beta_{k-1} B_{k-1} B_{k-1}^* & -\beta_{k-1} B_{k-1} \\ -\beta_{k-1} B_{k-1}^* & \beta_{k-1} \end{bmatrix}$, where $\beta_{k-1} = \frac{1}{1 - Y_{k-1}^* B_{k-1}}$. Hence, G_k^{-1} can be obtained using G_{k-1}^{-1} 、B_{k-1} and Y_{k-1}. After getting B_k, we can determine γ^k via 4.49.

In the following, we summarize our discussion and present the resulting OMP algorithm.

Algorithm 21 Orthogonal Matching Pursuit

Input: $y \in \mathbb{R}^m$;$A \in \mathbb{R}^{m\times n}$;a prescribed threshold parameter $0 < \delta < 1$

Initialization :

1. compute $\langle a_j, a_\ell\rangle$ for $j, \ell \in [n]$
2. $R^{(0)} = y$
3. compute $\langle R^{(0)}, a_j\rangle$ for $j \in [n]$
4. compute $j^1 = \underset{j\in[n]}{argsup}\langle R^{(0)}, a_j\rangle$
5. $x_1^1 = \langle y, a_{j^1}\rangle$; $y^1 = x_1^1 a_{j^1}$; $R^{(1)} = y - y^1$; $D^1 = \{j^1\}$
6. compute $\langle R^{(1)}, a_j\rangle$ for $j = [n]\backslash D^1 \Rightarrow \langle R^{(1)}, a_j\rangle = \langle y - y^1, a_j\rangle = \langle y, a_j\rangle - x_1^1\langle a_{j^1}, a_j\rangle$
7. compute $j^2 = \underset{j\in[n]\backslash D^1}{argsup}\langle R^{(1)}, a_j\rangle$

If $\langle R^{(1)}, a_{j^2}\rangle < \delta$, then stops; otherwise

8. $Y_1 = [\langle a_{j^2}, a_{j^1}\rangle]$; $G_1 = [\langle a_{j^1}, a_{j^1}\rangle] = [1]$; $B_1 = [b_1^1] = [\langle a_{j^2}, a_{j^1}\rangle]$
9. $a_{j^2} = b_1^1 a_{j^1} + \gamma^1 \Rightarrow \gamma^1 = a_{j^2} - b_1^1 a_{j^1}$
10. $x_2^2 = \frac{\langle R^{(1)}, a_{j^2}\rangle}{\langle \gamma^1, a_{j^2}\rangle}$; $x_1^2 = x_1^1 - b_1^1 x_2^2$; $y^2 = x_1^2 a_{j^1} + x_2^2 a_{j^2}$; $R^{(2)} = y - y^2$; $D^2 = \{j^1, j^2\}$
11. compute $\langle R^{(2)}, a_j\rangle$ for $j = [n]\backslash D^2 \Rightarrow \langle R^{(2)}, a_j\rangle = \langle y - y^2, a_j\rangle = \langle y, a_j\rangle - \sum_{n=1}^{2} x_n^2\langle a_{j^n}, a_j\rangle$

Iteration : from $k = 2$ **until** $\langle R^{(\bar{k})}, a_{j^{\bar{k}+1}}\rangle < \delta$ **at** $k = \bar{k}$

1. compute $j^{k+1} = \underset{j\in[n]\backslash D^k}{argsup}\langle R^{(k)}, a_j\rangle$
2. $G_k^{-1} = \begin{bmatrix} G_{k-1}^{-1} + \beta_{k-1}B_{k-1}B_{k-1}^* & -\beta_{k-1}B_{k-1} \\ -\beta_{k-1}B_{k-1}^* & \beta_{k-1}\end{bmatrix}$, where $\beta_{k-1} = \frac{1}{1 - Y_{k-1}^* B_{k-1}}$
3. $Y_k = [\langle a_{j^{k+1}}, a_{j^i}\rangle]$, $i = 1, \cdots, k$
4. $B_k = G_k^{-1} Y_k$
5. $a_{j^{k+1}} = \sum_{n=1}^{k} b_n^k a_{j^n} + \gamma^k = B_k[a_{j^1}, a_{j^2}, \cdots, a_{j^k}]^T + \gamma^k \Rightarrow \gamma^k = a_{j^{k+1}} - B_k[a_{j^1}, a_{j^2}, \cdots, a_{j^k}]^T$
6. $x_{k+1}^{k+1} = \frac{\langle R^{(k)}, a_{j^{k+1}}\rangle}{\langle \gamma^k, a_{j^{k+1}}\rangle}$; $x_n^{k+1} = x_n^k - b_n^k x_{k+1}^{k+1}$ for $n = 1, \cdots, k$; $y^{k+1} = \sum_{n=1}^{k+1} x_n^{k+1} a_{j^n}$; $R^{(k+1)} = y - y^{k+1}$; $D^{k+1} = \{j^1, \cdots, j^{k+1}\}$
7. compute $\langle R^{(k+1)}, a_j\rangle$ for $j = [n]\backslash D^{k+1} \Rightarrow \langle R^{(k+1)}, a_j\rangle = \langle y - y^{k+1}, a_j\rangle = \langle y, a_j\rangle - \sum_{n=1}^{k+1} x_n^{k+1}\langle a_{j^n}, a_j\rangle$

Output: $(j^k, x_k^{\bar{k}+1})$ for $k = 1, \cdots, \bar{k} + 1$

At the k-th iteration, we have the best approximation we can get using the k vectors we have selected from the dictionary. Therefore if the dictionary has finite number of atoms (e.g. M atoms), OMP converges in no more than M iterations to the projection of y onto the span of the dictionary atoms.

Note that before performing the back-projection step, at each iteration, MP possesses the following energy conservation formula for

the residues $\|R^{(k)}\|_2^2 = |\langle R^{(k)}, a_{j^{k+1}}\rangle|^2 + \|R^{(k+1)}\|_2^2$. We also want to derive similar formula for the OMP. We make a derivation as follows.

$$\sum_{n=1}^{k} x_n^k a_{j^n} + R^{(k)} - \sum_{n=1}^{k+1} x_n^{k+1} a_{j^n} - R^{(k+1)} = 0$$

$$\begin{aligned}
\Rightarrow R^{(k)} - R^{(k+1)} &= \sum_{n=1}^{k}(x_n^{k+1} - x_n^k)a_{j^n} + x_{k+1}^{k+1} a_{j^{k+1}} \\
&= \sum_{n=1}^{k}(x_n^{k+1} - x_n^k)a_{j^n} + x_{k+1}^{k+1}(\sum_{n=1}^{k} b_n^k a_{j^n} + \gamma^k) \\
&= \sum_{n=1}^{k}(x_n^{k+1} - x_n^k + x_{k+1}^{k+1} b_n^k)a_{j^n} + x_{k+1}^{k+1}\gamma^k = x_{k+1}^{k+1}\gamma^k \\
\therefore R^{(k)} &= R^{(k+1)} + x_{k+1}^{k+1}\gamma^k \\
\Rightarrow \|R^{(k)}\|_2^2 &= \langle R^{(k+1)} + x_{k+1}^{k+1}\gamma^k, R^{(k+1)} + x_{k+1}^{k+1}\gamma^k\rangle \\
&= \|R^{(k+1)}\|_2^2 + \|x_{k+1}^{k+1}\gamma^k\|_2^2 + 2Re(\langle R^{(k+1)}, x_{k+1}^{k+1}\gamma^k\rangle)
\end{aligned}$$

For the $\|x_{k+1}^{k+1}\gamma^k\|_2^2$ term, we can get

$$x_{k+1}^{k+1} = \frac{\langle R^{(k)}, a_{j^{k+1}}\rangle}{\langle \gamma^k, a_{j^{k+1}}\rangle} = \frac{\langle R^{(k)}, a_{j^{k+1}}\rangle}{\langle \gamma^k, a_{j^{k+1}} - \sum_{n=1}^{k} b_n^k a_{j^n}\rangle} = \frac{\langle R^{(k)}, a_{j^{k+1}}\rangle}{\|\gamma^k\|_2^2}$$

$$\therefore \|x_{k+1}^{k+1}\gamma^k\|_2^2 = \frac{|\langle R^{(k)}, a_{j^{k+1}}\rangle|^2}{\|\gamma^k\|_2^2}$$

As for the $\langle R^{(k+1)}, x_{k+1}^{k+1}\gamma^k\rangle$ term, we can get

$$\langle R^{(k+1)}, x_{k+1}^{k+1}\gamma^k\rangle = \langle R^{(k+1)}, x_{k+1}^{k+1}(a_{j^{k+1}} - \sum_{n=1}^{k} b_n^k a_{j^n})\rangle = 0$$

Hence, we derive that

$$\|R^{(k)}\|_2^2 = \|R^{(k+1)}\|_2^2 + \frac{|\langle R^{(k)}, a_{j^{k+1}}\rangle|^2}{\|\gamma^k\|_2^2} \tag{4.50}$$

The reduce of energy of the residues for MP is $\|R^{(k)}\|_2^2 - \|R^{(k+1)}\|_2^2 =$

$|\langle R^{(k)}, a_{j^{k+1}}\rangle|^2$ while the reduce of energy of the residues for OMP is $\|R^{(k)}\|_2^2 - \|R^{(k+1)}\|_2^2 = \frac{|\langle R^{(k)}, a_{j^{k+1}}\rangle|^2}{\|\gamma^k\|_2^2}$. Note that $\|a_{j^{k+1}}\|_2^2 = \|\gamma^k + \sum_{n=1}^{k} b_n^k a_{j^n}\|_2^2 = \|\gamma^k\|_2^2 + \|\sum_{n=1}^{k} b_n^k a_{j^n}\|_2^2$. As a result, $\|\gamma^k\|_2^2 = 1 - \|\sum_{n=1}^{k} b_n^k a_{j^n}\|_2^2$, which is not larger than 1. Hence, OMP reduces more energy of the residues, which in turn leads to convergence in fewer iterations. However, the computational effort required may not be smaller either. It should depend on both the signals and the dictionary. The readers can see [31] for references of materials we introduce so far.

In [7], the author introduces another way to compute OMP based on the Gram-Schmidt process. We call it "orthogonal matching pursuit version 2" in the following algorithm.

Algorithm 22 Orthogonal Matching Pursuit Version 2

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$

Initialization :

1. $R^{(0)} = y$
2. compute $\langle R^{(0)}, a_j\rangle$ for $j \in [n]$
3. $j^1 = \underset{j\in[n]}{argsup}\langle R^{(0)}, a_j\rangle$
4. $u_1 = a_{j^1}$
5. $D^1 = \{j^1\}$
6. $y^1 = \frac{\langle y,u_1\rangle}{\|u_1\|_2^2}u_1 = \langle y, u_1\rangle u_1$
7. $R^{(1)} = y - y^1$

Iteration : from $k = 1$ until a stopping criterion is met at $k = \bar{k}$

1. compute $\langle R^{(k)}, a_j\rangle$ for $j \in [n]\backslash D^k$
2. $j^{k+1} = \underset{j\in[n]\backslash D^k}{argsup}\langle R^{(k)}, a_j\rangle$
3. Gram-Schmidt Process : $u_{k+1} = a_{j^{k+1}} - \sum_{n=1}^{k}\frac{\langle a_{j^{k+1}},u_n\rangle}{\|u_n\|_2^2}u_n$
4. $D^{k+1} = \{j^1, \cdots, j^{k+1}\}$
5. $y^{k+1} = \sum_{n=1}^{k+1}\frac{\langle y,u_n\rangle}{\|u_n\|_2^2}u_n \triangleq \sum_{n=1}^{k+1} c_n u_n$
6. $R^{(k+1)} = y - y^{k+1}$

Output: (c_k, u_k) for $k = 1, \cdots, \bar{k}+1$

In the following, we want to derive the energy conservation formula

similar to 4.50.

$$R^{(k)} = y - y^k = y - \sum_{n=1}^{k} \frac{\langle y, u_n \rangle}{\|u_n\|_2^2} u_n$$

$$R^{(k+1)} = y - y^{k+1} = y - \sum_{n=1}^{k+1} \frac{\langle y, u_n \rangle}{\|u_n\|_2^2} u_n$$

$$\Rightarrow R^{(k)} - R^{(k+1)} = \frac{\langle y, u_{k+1} \rangle}{\|u_{k+1}\|_2^2} u_{k+1}$$

$$\because \langle y^k, u_{k+1} \rangle = 0$$

$$\therefore R^{(k)} - R^{(k+1)} = \frac{\langle y - y^k, u_{k+1} \rangle}{\|u_{k+1}\|_2^2} u_{k+1} = \frac{\langle R^{(k)}, u_{k+1} \rangle}{\|u_{k+1}\|_2^2} u_{k+1}$$

$$\because \langle R^{(k)}, u_n \rangle = 0 \ for \ n = 1, \cdots, k$$

$$\therefore R^{(k)} - R^{(k+1)} = \frac{\langle R^{(k)}, a_{j^{k+1}} \rangle}{\|u_{k+1}\|_2^2} u_{k+1}$$

$$\Rightarrow \|R^{(k)}\|_2^2 = \|R^{(k+1)} + \frac{\langle R^{(k)}, a_{j^{k+1}} \rangle}{\|u_{k+1}\|_2^2} u_{k+1}\|_2^2$$

$$\because \langle R^{(k+1)}, u_{k+1} \rangle = 0$$

$$\therefore \|R^{(k)}\|_2^2 = \|R^{(k+1)}\|_2^2 + \|\frac{\langle R^{(k)}, a_{j^{k+1}} \rangle}{\|u_{k+1}\|_2^2} u_{k+1}\|_2^2$$

$$= \|R^{(k+1)}\|_2^2 + \frac{|\langle R^{(k)}, a_{j^{k+1}} \rangle|^2}{\|u_{k+1}\|_2^2}$$

Note that the span of $\{u_1, u_2, \cdots, u_k\}$ in [7] is the same as the span of $\{a_{j^1}, a_{j^2}, \cdots, a_{j^k}\}$ in [31]. Hence, the orthogonal projection onto $\{u_1, u_2, \cdots, u_k\}$ is the same as that onto $\{a_{j^1}, a_{j^2}, \cdots, a_{j^k}\}$, which implies $u_{k+1} = \gamma^k$. Therefore, the energy conservation formulas of the two different ways are identical as expected. We make a comparison with the matching pursuit (MP) algorithm in the last section. In MP,

$$y = \sum_{n=0}^{k} \langle R^{(n)}, a_{j^{n+1}} \rangle a_{j^{n+1}} + R^{(k+1)}$$

while in OMP,

$$\begin{aligned} y &= \sum_{n=1}^{k+1} \frac{\langle y, u_n \rangle}{\|u_n\|_2^2} u_n + R^{(k+1)} \\ &= \sum_{n=0}^{k} \frac{\langle y, u_{n+1} \rangle}{\|u_{n+1}\|_2^2} u_{n+1} + R^{(k+1)} \\ &= \sum_{n=0}^{k} \frac{\langle R^{(n)}, a_{j^{n+1}} \rangle}{\|u_{n+1}\|_2^2} u_{n+1} + R^{(k+1)} \end{aligned}$$

In MP,

$$\|y\|_2^2 = \sum_{n=0}^{k} |\langle R^{(n)}, a_{j^{n+1}} \rangle|^2 + \|R^{(k+1)}\|_2^2$$

while in OMP,

$$\|y\|_2^2 = \sum_{n=0}^{k} \frac{|\langle R^{(n)}, a_{j^{n+1}} \rangle|^2}{\|u_{n+1}\|_2^2} + \|R^{(k+1)}\|_2^2$$

Finally, we introduce some important sparse recovery results of OMP applied in the context of compressive sensing. The proposition 3.5 of [13] presents a result regarding the condition of exact sparse recovery. We excerpt it as follows.

Theorem 4.2.1. *Given a matrix $A \in \mathbb{R}^{m \times n}$, every nonzero vector $x \in \mathbb{R}^n$ supported on a set S of size s is recovered from $y = Ax$ after at most s iterations of OMP if and only if the matrix A_S is injective and*

$$\max_{j \in S} |(A^* r)_j| > \max_{\ell \in \bar{S}} |(A^* r)_\ell| \tag{4.51}$$

for all nonzero $r \in \{Az, supp(z) \subset S\}$

It is proved that the condition 4.51 can be equivalently expressed as the following more concise condition.

$$\|A_S^\dagger A_{\bar{S}}\|_{1 \to 1} < 1 \tag{4.52}$$

Theorem 6.25 of [13] presents another result regarding the restricted

isometry condition of robust sparse recovery (that is, with a sparsity defect and measurement noise). We also excerpt it as follows.

Theorem 4.2.2. *Suppose that $A \in \mathbb{R}^{m\times n}$ has restricted isometry constant*

$$\delta_{26s} < \frac{1}{6} \tag{4.53}$$

then there are constants $C, D > 0$ depending only on δ_{26s} such that, for all $x \in \mathbb{R}^n$ and $e \in \mathbb{R}^m$, the iterates $\{x^{(k)}, k = 1, 2, \cdots\}$ which can be derived from the output of the OMP algorithm 21 *with $y = Ax + e$ satisfies, for any $1 \leq p \leq 2$,*

$$\|x - x^{(24s)}\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\|e\|_2 \tag{4.54}$$

4.2.3 Regularized Orthogonal Matching Pursuit (ROMP)

In [27] and [28], an algorithm called the regularized orthogonal matching pursuit (ROMP) is presented and analyzed. In the following, we will first describe the algorithm and then make a thorough mathematical analysis of it. The ROMP algorithm is described as follows.

Algorithm 23 ROMP

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m \times n}$; sparsity level s
Initialization : Let the index set $I^{(0)} = \emptyset$ and the residual vector $R^{(0)} = y$
Repeat the following iteration steps until one of three possible stopping criteria is triggered :

1. $R^{(\bar{k})} = 0$ at some iteration $\bar{k}$
2. the algorithm has run for $\bar{k} = s$ iterations
3. the cardinality of the index set $|I^{(\bar{k})}| \geq 2s$ at some iteration $\bar{k}$

Iterations :

1. Identification : choose a set $J^{(k)}$ of the s biggest coordinates in magnitude of the correlation vector

$$u^{(k)} = A^* R^{(k)} \tag{4.55}$$

or all of its nonzero coordinates, whichever set is smaller.

2. Regularization : Among all subsets $J' \subset J^{(k)}$ with comparable coordinates :

$$|u^{(k)}[i]| \leq 2|u^{(k)}[j]| \quad \forall i, j \in J' \tag{4.56}$$

choose $J_0^{(k)}$ with the maximal energy; that is,

$$J_0^{(k)} = \underset{J' \subset J^{(k)}}{argmax} \|u|_{J'}\|_2 \tag{4.57}$$

3. Update : Add the set $J_0^{(k)}$ to the index set

$$I^{(k+1)} = I^{(k)} \cup J_0^{(k)} \tag{4.58}$$

and update the residual :

$$x^{(k)} = \underset{z \in \mathbb{R}^{I^{(k+1)}}}{argmin} \|y - Az\|_2 \tag{4.59}$$

$$R^{(k+1)} = y - Ax^{(k)} \tag{4.60}$$

Output: index set $I^{(\bar{k})} \subset [n]$ and reconstructed signal $x^{(\bar{k})}$

Assume the restricted isometry constant δ_{4s} of the sampling matrix A satisfies $\delta_{4s} = \epsilon$ and the measurement vector $y = Ax + e$, where x is an s-sparse signal vector and e is a noise vector. At the start of

each iteration, we define

$$H^{(k)} := range(A_{supp(x)\cup I^{(k)}}) \tag{4.61}$$

$$F^{(k)} := range(A_{I^{(k)}}) \tag{4.62}$$

$$E^{(k)} := (F^{(k)})^{\perp} \cap H^{(k)} \tag{4.63}$$

$$E_0^{(k)} := range(A_{supp(x)\setminus I^{(k)}}) \tag{4.64}$$

$$x_0^{(k)} := x|_{supp(x)\setminus I^{(k)}} \tag{4.65}$$

$$y_0^{(k)} := Ax_0^{(k)} \in E_0^{(k)} \tag{4.66}$$

$$u_0^{(k)} := A^* y_0^{(k)} \tag{4.67}$$

Note that $R^{(k)} = P_{(F^{(k)})^{\perp}} y$, where $P_{(F^{(k)})^{\perp}}$ denotes the orthogonal projector on $(F^{(k)})^{\perp}$. Since $Ax \in H^{(k)}$, $R^{(k)} = P_{E^{(k)}}(Ax) + P_{(F^{(k)})^{\perp}} e$. However, what we actually want is a vector in the range of $A_{supp(x)\setminus I^{(k)}}$ (i.e., $E_0^{(k)}$) so that we may successfully identify indices belonging to $supp(x)\setminus I^{(k)}$ from that vector. Luckily, due to the almost orthogonality property 2.107, the subspaces $F^{(k)}$ and $E_0^{(k)}$ are almost orthogonal, which implies the subspaces $E^{(k)}$ and $E_0^{(k)}$ may be close to each other. Indeed, we can derive that $R^{(k)}$ and $y_0^{(k)}$ are close to each other with the inequality

$$\|R^{(k)} - y_0^{(k)}\|_2 \leq 2.2\epsilon \|y_0^{(k)}\|_2 + \|e\|_2 \tag{4.68}$$

What's more, we can derive that the correlation vector $u^{(k)}$ is close to $u_0^{(k)}$ with the inequality

$$\|(u^{(k)} - u_0^{(k)})|_T\|_2 \leq 2.4\epsilon \|x_0^{(k)}\|_2 + (1+\epsilon)\|e\|_2 \tag{4.69}$$

where $T \in [n]$ is any set with $|T| \leq 3s$.

At the identification step, we can show that the energy of $u^{(k)}$ when restricted to $J^{(k)}$ is not too small. Indeed,

$$\|u^{(k)}|_{J^{(k)}}\|_2 \geq (1 - 4.43\epsilon)\|x_0^{(k)}\|_2 - (1+\epsilon)\|e\|_2 \tag{4.70}$$

At the regularization step, we can derive that

$$\|u^{(k)}|_{J_0^{(k)}}\|_2 \geq \frac{2}{5\sqrt{\log s}}((1-4.43\epsilon)\|x_0^{(k)}\|_2 - (1+\epsilon)\|e\|_2) \qquad (4.71)$$

with the help of the lemma 3.7 of [27] that guarantees us to obtain a subset with sufficient energy after the regularization operation. Precisely, the lemma is described as follows.

Lemma 4.2.3. *Let y be any vector in $\mathbb{R}^m$, $m > 1$. Then there exists a subset $A \in [m]$ with comparable coordinates :*

$$|y[i]| \leq 2|y[j]| \quad \forall i, j \in A$$

and with sufficient energy :

$$\|y|_A\|_2 \geq \frac{2}{5\sqrt{\log m}}\|y\|_2$$

Besides, we give an intuition about why it is good to perform regularization. Because of the comparability property and the maximal energy property of $J_0^{(k)}$, the minimum components of $J_0^{(k)}$ will not be too small (at least only be half smaller than the maximum components of $J_0^{(k)}$), which in turn may filters out some indices of $J^{(k)}$ that do not belong to the support of x. Therefore, after the identification step, $J^{(k)}$ collects some indices that are possible to lie in the support of x while after the regularization step, some possibly outliers (indices that do not belong to the support of x) may be filtered out and $J_0^{(k)}$ is obtained. We make a remark on the implementation issue of the regularization step. Note that although it seems like a combinatorial problem to get $J_0^{(k)} \subset J^{(k)}$, it can actually be done fast if one observes that $J_0^{(k)}$ is an interval in the non-increasing rearrangement of $J^{(k)}$. Indeed, it can be done in $\mathcal{O}(s)$ time.

First, we assume $e \neq 0$. Since $R^{(k)} \in (F^{(k)})^{\perp}$, the correlation vector $u^{(k)} = A^* R^{(k)}$ restricted to $I^{(k)}$ will be zero (i.e., $u^{(k)}|_{I^{(k)}} = 0$). Because the set $J^{(k)}$ contains only nonzero coordinates of $u^{(k)}$, we have $J^{(k)} \cap I^{(k)} = \emptyset$. Since $J_0^{(k)} \subset J^{(k)}$, $J_0^{(k)} \cap I^{(k)} = \emptyset$. Furthermore, utilizing the comparability property of the coordinates in $J_0^{(k)}$, assuming

$\epsilon \leq 0.01/\sqrt{\log s}$ and incorporating 4.71, we can prove that if $\|x_0^{(k)}\|_2 \geq 100\|e\|_2\sqrt{\log s}$, then $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ must be true; on the other hand, if $\|x_0^{(k)}\|_2 < 100\|e\|_2\sqrt{\log s}$, $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ may be true or not. Hence, at every iteration of ROMP, at least one of the two conditions $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ and $\|x_0^{(k)}\|_2 < 100\|e\|_2\sqrt{\log s}$ will hold true. The condition $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ means that at least half of the newly selected coordinates are from the support of the signal x.

Next, we consider the case when $e = 0$. Using the same argument of the case when $e \neq 0$, we can also have $J_0^{(k)} \cap I^{(k)} = \emptyset$. Furthermore, since $e = 0$, $\|x_0^{(k)}\|_2 \geq 100\|e\|_2\sqrt{\log s} = 0$ always holds, $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ must be true at every iteration. What's more, the condition that $\delta_{4s} = \epsilon \leq 0.01/\sqrt{\log s}$ can actually be relaxed to the condition $\delta_{2s} = \epsilon \leq 0.03/\sqrt{\log s}$ in the case when $e = 0$. Finally, at the regularization step, we have $\|u^{(k)}|_{J_0^{(k)}}\|_2 \geq \frac{2}{5\sqrt{\log s}}((1 - 4.43\epsilon)\|x_0^{(k)}\|_2 - (1 + \epsilon)\|e\|_2)$. When $e = 0$, then we have $\|u^{(k)}|_{J_0^{(k)}}\|_2 \geq \frac{2}{5\sqrt{\log s}}(1 - 4.43\epsilon)\|x_0^{(k)}\|_2$. If $R^{(k)} \neq 0$ at the start of each iteration, then $supp(x) \backslash I^{(k)} \neq \emptyset$. Hence, $x_0^{(k)} \neq 0$, which implies that $\|u^{(k)}|_{J_0^{(k)}}\|_2 \neq 0$. Then, $J_0^{(k)}$ won't be an empty set at the regularization step. In contrast, if $e \neq 0$, then $(1 - 4.43\epsilon)\|x_0^{(k)}\|_2 - (1 + \epsilon)\|e\|_2$ may be smaller than zero. Hence, we cannot deduce that $\|u^{(k)}|_{J_0^{(k)}}\|_2 \neq 0$. We summarize our discussion in the following theorems.

Theorem 4.2.4. *Assume a sampling matrix A satisfies $\delta_{2s} = 0.03/\sqrt{\log s}$. Let x be an s-sparse vector in $\mathbb{R}^n$ and $y \in \mathbb{R}^m$ a sample vector satisfying $y = Ax$. Then at any iteration of ROMP, after the regularization step, we have*

$$J_0^{(k)} \neq \emptyset \tag{4.72}$$

$$J_0^{(k)} \cap I^{(k)} = \emptyset \tag{4.73}$$

$$|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}| \tag{4.74}$$

Theorem 4.2.5. *Assume a sampling matrix A satisfies $\delta_{4s} = 0.01/\sqrt{\log s}$. Let x be an s-sparse vector in $\mathbb{R}^n$ and $y \in \mathbb{R}^m$ a sample vector satisfying $y = Ax + e$, where $e \in \mathbb{R}^m$ is a noise vector. Then at any iteration of ROMP, after the regularization step, we have $J_0^{(k)} \cap I^{(k)} = \emptyset$ and at least one of the following conditions hold :*

1. $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$
2. $\|x_0^{(k)}\|_2 < 100\|e\|_2\sqrt{\log s}$

Based on theorem 4.2.4, we know that at every iteration, ROMP finds at least one coordinate in the support of the signal x since $J_0^{(k)} \neq \emptyset$ and $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$. Furthermore, since $J_0^{(k)} \cap I^{(k)} = \emptyset$, ROMP can outputs a set $I^{(\bar{k})}$ such that $supp(x) \subset I^{(\bar{k})}$ in at most s iterations when the stopping criterion that $R^{(\bar{k})} = 0$ is triggered and due to $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ again, $|I^{(\bar{k})}| \leq 2s$. Hence, as a consequence of theorem 4.2.4, we can have the following theorem.

Theorem 4.2.6. *Assume a sampling matrix A satisfies $\delta_{2s} = 0.03/\sqrt{\log s}$. Let x be an s-sparse vector in $\mathbb{R}^n$ and $y \in \mathbb{R}^m$ a sample vector satisfying $y = Ax$. Then ROMP outputs a set I such that $supp(x) \subset I$ and $|I| \leq 2s$ in at most s iterations.*

such result guarantees exact sparse recovery of a signal. Indeed, since A_I has full rank due to the restricted isometry property of A, we can compute the signal x from its measurement y and the set I by $x = (A_I)^\dagger y$ and actually the output $x^{(\bar{k})}$ is exactly $(A_I)^\dagger y$. Moreover, such recovery is uniform because it holds for any s-sparse vector. In contrast, uniform recovery has been shown to be impossible for OMP. Also note that it has been unknown whether OMP gives sparse recovery for partial Fourier measurements (even with nonuniform guarantees). However, since partial Fourier matrices satisfy restricted isometry property, ROMP gives sparse recovery for these measurements, and even with uniform guarantees.

For the case when $e \neq 0$, we can derive that

$$\|x - x^{(k)}\|_2 \leq \frac{1+\epsilon}{1-\epsilon}\|x_0^{(k)}\|_2 + \frac{2}{1-\epsilon}\|e\|_2 \tag{4.75}$$

Due to theorem 4.2.5, there are three possible cases. The first case is that $\|x_0^{(k)}\|_2 < 100\|e\|_2\sqrt{\log s}$ occurs at some iteration. Since $\|x - x^{(k)}\|_2 \leq \frac{1+\epsilon}{1-\epsilon}\|x_0^{(k)}\|_2 + \frac{2}{1-\epsilon}\|e\|_2$, we can derive that $\|x - x^{(k)}\|_2 \leq 104\sqrt{\log s}\|e\|_2$. Also note that since $|I^{(k)}|$ is non-decreasing, if $\|x_0^{(k)}\|_2 < 100\|e\|_2\sqrt{\log s}$ occurs at some iteration, then it will hold for all subsequent iterations. Hence, when one of the three stopping criteria is triggered, $\|x_0^{(k)}\|_2 < 100\|e\|_2\sqrt{\log s}$ still holds and thus $\|x - x^{(k)}\|_2 \leq 104\sqrt{\log s}\|e\|_2$ also holds. The second case is that $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ occurs at every iteration and $J_0^{(k)} = \emptyset$ for some iteration. Since $J_0^{(k)} = \emptyset$, $u^{(k)} = A^*R^{(k)} = 0$. It can be derived that the inequality $\|x_0^{(k)}\|_2 < 100\|e\|_2\sqrt{\log s}$ also holds. Hence, the second case will reduce to the first case. The third case is that $|J_0^{(k)} \cap supp(x)| \geq \frac{1}{2}|J_0^{(k)}|$ occurs at every iteration and $J_0^{(k)}$ is always nonempty. In this way, ROMP identifies at least one coordinate in the support of the signal x at each iteration. Thus, if ROMP runs s iterations or until $|I^{(\bar{k})}| \geq 2s$, it must be that $supp(x) \subset I^{(\bar{k})}$, which means $x_0^{(\bar{k})} = x|_{supp(x)\setminus I^{(\bar{k})}} = 0$. Hence, the inequality $\|x_0^{(\bar{k})}\|_2 < 100\|e\|_2\sqrt{\log s}$ holds, which reduces the third case to the first case. Nonetheless, if the stopping criterion that $R^{(\bar{k})} = 0$ is triggered, then we have $u^{(\bar{k})} = A^*R^{(\bar{k})} = 0$. Hence, the third case will reduce to the second case. As a result, both the second and the third cases will reduce to the first case and the first case results in the inequality $\|x - x^{(\bar{k})}\|_2 \leq 104\sqrt{\log s}\|e\|_2$. Summarizing this discussion, we can come up with the following theorem.

Theorem 4.2.7. *Assume a sampling matrix A satisfies $\delta_{4s} = 0.01/\sqrt{\log s}$. Let x be an s-sparse vector in $\mathbb{R}^n$ and $y \in \mathbb{R}^m$ a sample vector satisfying $y = Ax + e$, where $e \in \mathbb{R}^m$ is a noise vector. Then when one of the stopping criteria is triggered, ROMP*

outputs $x^{(\bar{k})}$ *that satisfies*

$$\|x - x^{(\bar{k})}\|_2 \leq 104\sqrt{\log s}\|e\|_2 \tag{4.76}$$

Note that if $e = 0$, then $x^{(\bar{k})} = x$, which implies exact recovery. This result is consistent with that of the theorem 4.2.6.

Finally, we can extend theorem 4.2.7 to the most general case - the signal x is not exactly s-sparse but a general vector in $\mathbb{R}^n$. We can partition $y = Ax + e$ to $y = Ax_{2s} + (A(x - x_{2s}) + e)$. Suppose we tighten the restricted isometry property of A to $\delta_{8s} = 0.01/\sqrt{\log s}$. Since x_{2s} is 2s-sparse and $\delta_{8s} = 0.01/\sqrt{\log s}$, we can apply theorem 4.2.7, if we input the sparsity level as $2s$ to the ROMP algorithm, and obtain

$$\begin{aligned}\|x^{(\bar{k})} - x_{2s}\|_2 &\leq 104\sqrt{\log 2s}\|A(x - x_{2s}) + e\|_2 \\ &\leq 104\sqrt{\log 2s}(\|A(x - x_{2s})\|_2 + \|e\|_2)\end{aligned}$$

By property 4 of the restricted isometry constant introduced in section 2.8, we can derive that

$$\begin{aligned}\|A(x - x_{2s})\|_2 &\leq (1 + \epsilon)\left(\|x - x_{2s}\|_2 + \frac{\|x - x_{2s}\|_1}{\sqrt{8s}}\right) \\ &\leq (1 + \epsilon)\left(\|x - x_{2s}\|_2 + \frac{\|x - x_s\|_1}{\sqrt{s}}\right)\end{aligned}$$

By the following lemma :

Lemma 4.2.8. *Let* $w \in \mathbb{R}^n$. *Then*

$$\|w - w_s\|_2 \leq \frac{\|w\|_1}{2\sqrt{s}} \tag{4.77}$$

We can derive that $\|x - x_{2s}\|_2 \leq \frac{\|x - x_s\|_1}{2\sqrt{s}}$ by letting w to be $x - x_s$.

Hence, $\|A(x - x_{2s})\|_2 \leq 1.5(1+\epsilon)\frac{\|x-x_s\|_1}{\sqrt{s}}$. Therefore,

$$\begin{aligned}\|x^{(\bar{k})} - x_{2s}\|_2 &\leq 104\sqrt{\log 2s}\left(1.5(1+\epsilon)\frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2\right)\\ &\leq 159\sqrt{\log 2s}\left(\frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2\right)\end{aligned}$$

Moreover, since

$$\begin{aligned}\|x^{(\bar{k})} - x_{2s}\|_2 &= \|(x^{(\bar{k})} - x) + (x - x_{2s})\|_2\\ &\geq \|x^{(\bar{k})} - x\|_2 - \|x - x_{2s}\|_2\\ &\geq \|x^{(\bar{k})} - x\|_2 - \frac{\|x-x_s\|_1}{2\sqrt{s}}\end{aligned}$$

we can derive that

$$\|x^{(\bar{k})} - x\|_2 \leq 160\sqrt{\log 2s}\left(\frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2\right)$$

Corollary 3.2 of [28] presents the third error bound

$$\|(x^{(\bar{k})})_{2s} - x_{2s}\|_2 \leq 477\sqrt{\log 2s}\left(\frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2\right)$$

Due to this inequality, we can further derive that

$$\begin{aligned}\|(x^{(\bar{k})})_{2s} - x\|_2 &\leq \|(x^{(\bar{k})})_{2s} - x_{2s}\|_2 + \|x - x_{2s}\|_2\\ &\leq 477\sqrt{\log 2s}\left(\frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2\right) + \frac{\|x-x_s\|_1}{2\sqrt{s}}\\ &\leq 478\sqrt{\log 2s}\left(\frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2\right)\end{aligned}$$

Ultimately, the most general theorem can be presented as follows.

Theorem 4.2.9. *Assume a sampling matrix A satisfies $\delta_{8s} = 0.01/\sqrt{\log s}$. Let $x \in \mathbb{R}^n$ be a signal vector and $y \in \mathbb{R}^m$ a sampling vector satisfying $y = Ax + e$, where $e \in \mathbb{R}^m$ is a noise vector. Then*

ROMP outputs $x^{(\bar{k})}$ that satisfies

$$\|x^{(\bar{k})} - x_{2s}\|_2 \leq 159\sqrt{\log 2s}\left(\frac{\|x - x_s\|_1}{\sqrt{s}} + \|e\|_2\right) \tag{4.78}$$

$$\|x^{(\bar{k})} - x\|_2 \leq 160\sqrt{\log 2s}\left(\frac{\|x - x_s\|_1}{\sqrt{s}} + \|e\|_2\right) \tag{4.79}$$

$$\|(x^{(\bar{k})})_{2s} - x_{2s}\|_2 \leq 477\sqrt{\log 2s}\left(\frac{\|x - x_s\|_1}{\sqrt{s}} + \|e\|_2\right) \tag{4.80}$$

$$\|(x^{(\bar{k})})_{2s} - x\|_2 \leq 478\sqrt{\log 2s}\left(\frac{\|x - x_s\|_1}{\sqrt{s}} + \|e\|_2\right) \tag{4.81}$$

Hence, both $x^{(\bar{k})}$ and its best 2s-term approximation $(x^{(\bar{k})})_{2s}$ are close to the true signal x or its best 2s-term approximation x_{2s} within an distance proportional to $\sqrt{\log 2s}\left(\frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2\right)$.

The ℓ_1-minimization method has strong uniform guarantees of sparse recovery. However, the running time is not only polynomial in n but also in certain condition numbers of the objective, which may consume much time. On the other hand, OMP runs fast both theoretically and empirically, and is easier to implement than ℓ_1-minimization. However, OMP has weaker guarantees of sparse recovery. ROMP is good for possessing the advantages of both methods, i.e., strong uniform guarantees of sparse recovery, fast speed and transparency of the methodology. Furthermore, in the aspect of implementation, although ROMP requires the estimation about the sparsity level s, it does not require the knowledge of the error e as the ℓ_1-minimization methods requires. In some applications, it seems more natural to impose a sparsity requirement than an error constraint.

4.2.4 Compressive Sampling Matching Pursuit (CoSaMP)

In [26], it introduces a novel greedy algorithm called the compressive sampling matching pursuit (CoSaMP). In this section, we will first describe the algorithm, then analyze it theoretically, present some important results, and finally discuss some practical issues. The algorithm

is described as the following algorithm 24.

Algorithm 24 CoSaMP

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m \times n}$; sparsity level s

Initialization :

1. $x^{(0)} = 0$
2. $R^{(0)} = y$

Iteration : repeat until a stopping criterion is met at $k = \bar{k}$

1. compute $\langle R^{(k)}, a_j \rangle$ for $j \in [n]$ (equivalently, compute $c = A^*(R^{(k)})$)
2. Identification : $\Omega = supp(c_{2s})$
3. Merge support : $T = \Omega \cup supp(x^{(k)})$
4. Estimation : $b|_T = A_T^{\dagger} y$, $b|_{T^c} = 0$
5. Pruning : $x^{(k+1)} = b_s$
6. $R^{(k+1)} = y - Ax^{(k+1)}$

Output: $x^{(\bar{k}+1)}$

First, we consider the case when the signal $x \in \mathbb{R}^n$ is actually s-sparse. Assume the sampling matrix $A \in \mathbb{R}^{m \times n}$ satisfies the restricted isometry property $\delta_{4s} \leq 0.1$ and let the sample vector $y \in \mathbb{R}^m$ equals $Ax + e$, where $e \in \mathbb{R}^m$ is a noise vector. For the identification step, let $r^{(k)} = x - x^{(k)}$. We can have the inequality

$$\|r^{(k)}|_{\Omega^c}\|_2 \leq 0.2223\|r^{(k)}\|_2 + 2.34\|e\|_2 \tag{4.82}$$

The physical meaning of this inequality is that the energy of the difference between the true signal x and the current estimate $x^{(k)}$ on the set Ω^c is small compared with the total energy of such difference. For the support-merging step, we can have the inequality

$$\|x|_{T^c}\|_2 \leq \|r^{(k)}|_{\Omega^c}\|_2 \tag{4.83}$$

Hence, by the identification and the support-merging, we can come up with an index set outside which the components of the signal x have little energy. The estimation step solves a least-squares problem to obtain values for the coefficients in the set T. We can have a bound on the error of this approximation by the inequality

$$\|x - b\|_2 \leq 1.112\|x|_{T^c}\|_2 + 1.06\|e\|_2 \tag{4.84}$$

Finally, for the pruning step, we can bound the distance between x and

b_s by the inequality

$$\|x - b_s\|_2 \leq 2\|x - b\|_2 \tag{4.85}$$

Therefore, combining those inequalities, we can make the following derivation.

$$\begin{aligned}
\|x - x^{(k+1)}\|_2 &= \|x - b_s\|_2 \\
&\leq 2\|x - b\|_2 \\
&\leq 2(1.112\|x|_{T^c}\|_2 + 1.06\|e\|_2) \\
&\leq 2.224\|r^{(k)}|_{\Omega^c}\|_2 + 2.12\|e\|_2 \\
&\leq 2.224(0.2223\|r^{(k)}\|_2 + 2.34\|e\|_2) + 2.12\|e\|_2 \\
&< 0.5\|r^{(k)}\|_2 + 7.5\|e\|_2 \\
&= 0.5\|x - x^{(k)}\|_2 + 7.5\|e\|_2
\end{aligned}$$

Note that in the estimation step, we ideally assume we can precisely compute the least-squares approximation. However, in practice, we will apply some iterative least-squares solver to solve the least-squares problem. In this way, there must be some difference between the precise least-squares approximation and the computed solution. Hence, the inequality we have presented needs to be modified. We consider two iterative least-squares solvers, which are the Richardson's iteration method and the conjugate gradient method respectively. Richardson's iteration produces a sequence $\{z^{(\ell)}\}$ of iterates that satisfy

$$\|z^{(\ell)} - A_T^\dagger y\|_2 \leq (\delta_{3s})^\ell \|z^{(0)} - A_T^\dagger y\|_2 \leq 0.1^\ell \|z^{(0)} - A_T^\dagger y\|_2 \quad for\, \ell = 0, 1, 2, \cdots \tag{4.86}$$

Conjugate gradient produces a sequence $\{z^{(\ell)}\}$ of iterates that satisfy

$$\|z^{(\ell)} - A_T^\dagger y\|_2 \leq 2\rho^\ell \|z^{(0)} - A_T^\dagger y\|_2 \quad for\, \ell = 0, 1, 2, \cdots \tag{4.87}$$

where $\rho \leq 1 - \frac{2}{\sqrt{\frac{1+\delta_{3s}}{1-\delta_{3s}}}+1} \leq 1 - \frac{2}{\sqrt{\frac{1+0.1}{1-0.1}}+1} \approx 0.05$. It is natural to assign the current signal approximation $x^{(k)}$ as the initial iterate for the least-squares solver; that is, $z^{(0)} = x^{(k)}$. We can bound the ℓ_2

norm of $\|z^{(0)} - A_T^\dagger y\|_2 = \|x^{(k)} - A_T^\dagger y\|_2$ by the inequality

$$\|x^{(k)} - A_T^\dagger y\|_2 \leq 2.112\|x - x^{(k)}\|_2 + 1.06\|e\|_2 \tag{4.88}$$

Hence, $\|z^{(0)} - A_T^\dagger y\|_2$ is controlled by the current signal approximation error. Assume we run the iterative least-squares solver (Richardson's iteration or conjugate gradient) for three iterations, then

$$\|z^{(3)} - A_T^\dagger y\|_2 \leq 0.002112\|x - x^{(k)}\|_2 + 0.00106\|e\|_2 \tag{4.89}$$

Let $b|_T = z^{(3)}$. We can derive that

$$\begin{aligned}
\|x - b\|_2 &= \|(x - A_T^\dagger y) + (A_T^\dagger y - b)\|_2 \\
&\leq \|x - A_T^\dagger y\|_2 + \|b|_T - A_T^\dagger y\|_2 \\
&\leq (1.112\|x|_{T^c}\|_2 + 1.06\|e\|_2) + (0.00212\|x - x^{(k)}\|_2 + 0.00106\|e\|_2) \\
&\leq 1.112\|x|_{T^c}\|_2 + 0.0022\|x - x^{(k)}\|_2 + 1.062\|e\|_2
\end{aligned} \tag{4.90}$$

After taking into account the practical iterative least-squares issue, we can bound the signal approximation error $\|x - x^{(k+1)}\|_2$ as follows.

$$\begin{aligned}
\|x - x^{(k+1)}\|_2 &= \|x - b_s\|_2 \\
&\leq 2\|x - b\|_2 \\
&\leq 2(1.112\|x|_{T^c}\|_2 + 0.0022\|r^{(k)}\|_2 + 1.062\|e\|_2) \\
&\leq 2.224\|r^{(k)}|_{\Omega^c}\|_2 + 0.0044\|r^{(k)}\|_2 + 2.124\|e\|_2 \\
&\leq 2.224(0.2223\|r^{(k)}\|_2 + 2.34\|e\|_2) + 0.0044\|r^{(k)}\|_2 + 2.124\|e\|_2 \\
&< 0.5\|r^{(k)}\|_2 + 7.5\|e\|_2 \\
&= 0.5\|x - x^{(k)}\|_2 + 7.5\|e\|_2
\end{aligned}$$

Therefore, either in the ideal case or the practical case, the inequality that

$$\|x - x^{(k+1)}\|_2 \leq 0.5\|x - x^{(k)}\|_2 + 7.5\|e\|_2 \tag{4.91}$$

holds true.

Finally, we consider the case when the signal x is general. We also assume the restricted isometry constant of the sampling matrix A satisfies $\delta_{4s} \leq 0.1$. The key point is that we can express the noisy

sample vector y of a general signal x as the sample vector of a sparse signal x_s contaminated with a different noise vector. That is, $y = Ax + e = Ax_s + \tilde{e}$, where $\tilde{e} = A(x - x_s) + e$. Applying the triangle inequality and 2.91, we can derive

$$\|\tilde{e}\|_2 \leq 1.05 \left[\|x - x_s\|_2 + \frac{1}{\sqrt{s}} \|x - x_s\|_1 \right] + \|e\|_2 \tag{4.92}$$

Hence, using the result 4.91 we have proved for the sparse case, we have $\|x_s - x^{(k+1)}\|_2 \leq 0.5\|x_s - x^{(k)}\|_2 + 7.5\|\tilde{e}\|_2$. We can further make the following derivation

$$\begin{aligned}
\|x - x^{(k+1)}\|_2 &= \|(x - x_s) + (x_s - x^{(k+1)})\|_2 \\
&\leq \|x - x_s\|_2 + \|x_s - x^{(k+1)}\|_2 \\
&\leq \|x - x_s\|_2 + 0.5\|x_s - x^{(k)}\|_2 + 7.5\|\tilde{e}\|_2 \\
&\leq \|x - x_s\|_2 + 0.5(\|x_s - x\|_2 + \|x - x^{(k)}\|_2) + 7.5\|\tilde{e}\|_2 \\
&= 0.5\|x - x^{(k)}\|_2 + 7.5\|\tilde{e}\|_2 + 1.5\|x - x_s\|_2 \\
&< 0.5\|x - x^{(k)}\|_2 + 10\nu
\end{aligned} \tag{4.93}$$

where we define the unrecoverable energy in the signal as

$$\nu \triangleq \|x - x_s\|_2 + \frac{1}{\sqrt{s}} \|x - x_s\|_1 + \|e\|_2 \tag{4.94}$$

In the following, we will present some results based on the main iteration invariant inequality 4.93.

1.

$$\|x - x^{(k)}]\|_2 \leq 2^{-k}\|x\|_2 + 20\nu \tag{4.95}$$

Proof.

$$\begin{aligned}
\|x - x^{(k)}\|_2 &\leq 2^{-1}\|x - x^{(k-1)}\|_2 + 10\nu \\
&\leq 2^{-1}(2^{-1}\|x - x^{(k-2)}\|_2 + 10\nu) + 10\nu \\
&= 2^{-2}\|x - x^{(k-2)}\|_2 + (1 + 2^{-1})10\nu \\
&\leq 2^{-k}\|x - x^{(0)}\|_2 + (1 + 2^{-1} + \cdots + 2^{-(k-1)})10\nu \\
&\leq 2^{-k}\|x\|_2 + 20\nu
\end{aligned}$$

□

2. Define the signal-to-noise ratio (SNR) as SNR= $10\log_{10}(\frac{\|x\|_2}{\nu})$ and the reconstruction SNR after k iterations as R-SNR(k)= $10\log_{10}(\frac{\|x\|_2}{\|x-x^{(k)}\|_2})$. We can prove that R-SNR(k)$\geq \min\{3k, \text{SNR}-13\} - 3$

Proof. R-SNR(k)= $10\log_{10}(\frac{\|x\|_2}{\|x-x^{(k)}\|_2})$
$\geq 10\log_{10}\|x\|_2 - 10\log_{10}(2^{-k}\|x\|_2 + 20\nu)$
If $2^{-k}\|x\|_2 > 20\nu$, then R-SNR(k)$\geq 10\log_{10}\|x\|_2 - 10\log_{10}(2 * 2^{-k}\|x\|_2) = 3k - 3$.
If $2^{-k}\|x\|_2 \leq 20\nu$, then R-SNR(k)$\geq 10\log_{10}\|x\|_2 - 10\log_{10}(2 * 20\nu)$ =(SNR−13)−3.
Hence, R-SNR(k)$\geq \min\{3k, \text{SNR}-13\} - 3$ □

From this inequality, we know that CoSaMP increases the reconstruction SNR by about 3 decibels at each iteration before k exceeds the ceiling of (SNR−13)/3, which means to reduce the error to its minimal value, the number of iterations required is proportional to the SNR.

3. Suppose that A is an $m \times n$ sampling matrix with restricted isometry constant $\delta_{2s} \leq c$. Let $y = Ax + e$ be a sample vector of an arbitrary signal, contaminated with an arbitrary noise vector. For a given precision parameter η, CoSaMP produces an s-sparse approximation $x^{(\bar{k})}$ that satisfies

$$\begin{aligned}\|x - x^{(\bar{k})}\|_2 &\leq C(\eta + \nu) \\ &= C(\eta + \|x - x_s\|_2 + \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \|e\|_2)\end{aligned}$$

where $\bar{k} = \mathcal{O}(\log(\|x\|_2/\eta))$. Hence, in the absence of noise, CoSaMP can recover an s-sparse signal to arbitrarily high precision. However, if the signal is not s-sparse but only compressible or if there is noise, the performance of it will degrade. Furthermore, let $\mathcal{L}$ denote the cost of a multiplication with the sampling matrix A or A^*. It has been calculated in [26] that each iteration of CoSaMP

requires $\mathcal{O}(\mathcal{L})$ time and $\mathcal{O}(N)$ storage. Hence, totally it requires $\mathcal{O}(\mathcal{L} * \log(\|x\|_2/\eta))$ time and $\mathcal{O}(N)$ storage to run the algorithm. As a result, the algorithm can run substantially fast if there is fast multiplication of the sampling matrix.

Finally, we discuss some practical issues. The first one is about how to determine the input sparsity level. It is suggested that one can choose s to be $m/(2\log n)$. One can also run CoSaMP for several sparsity levels and select the one with smallest approximation error $\|y - Ax^{(\bar{k})}\|_2$. Next, we talk about other stopping criteria. We have proved that if we stop the algorithm after a fixed number of iterations (i.e., $\bar{k} = \mathcal{O}(\log(\|x\|_2/\eta))$), then $\|x - x^{(\bar{k})}\|_2 \leq C(\eta + \nu)$. In [26], two possible alternative stopping criteria are suggested (The signal x is assumed to be s-sparse. If it is not s-sparse, we can simply use the same technique we have applied to analyze the general case.)

1. If we stop the algorithm when

$$\|R^{(\bar{k})}\|_2 \leq \epsilon \tag{4.96}$$

then

$$\|x - x^{(\bar{k})}\|_2 \leq 1.06(\epsilon + \|e\|_2) \tag{4.97}$$

2. If we stop the algorithm when

$$\|A^*(R^{(\bar{k})})\|_\infty \leq \eta/\sqrt{2s} \tag{4.98}$$

then

$$\|x - x^{(\bar{k})}\|_\infty \leq 1.12\eta + 1.17\|e\|_2 \tag{4.99}$$

However, can the stopping criteria indeed be triggered ? It is also proved in [26] that

1. If $\|x - x^{(\bar{k})}\|_2 \leq 0.95(\epsilon - \|e\|_2)$, then $\|R^{(\bar{k})}\|_2 \leq \epsilon$.
2. If $\|x - x^{(\bar{k})}\|_\infty \leq \frac{0.45\eta}{s} - \frac{0.68\|e\|_2}{\sqrt{s}}$, then $\|A^*(R^{(\bar{k})})\|_\infty \leq \eta/\sqrt{2s}$

Since $\|x - x^{(k)}\|_2 \leq 2^{-k}\|x\|_2 + 20\nu$, the inequality $\|x - x^{(\bar{k})}\|_2 \leq 0.95(\epsilon - \|e\|_2)$ is desired to be satisfied after some $\bar{k}$ iterations. Hence, $\|R^{(\bar{k})}\|_2 \leq \epsilon$ can indeed be triggered at least when $\|x - x^{(\bar{k})}\|_2 \leq$

$0.95(\epsilon - \|e\|_2)$, which means it is an effective stopping criterion. Similarly, also because $\|x - x^{(\bar{k})}\|_2 \leq 2^{-k}\|x\|_2 + 20\nu$ (hence, $\|x - x^{(\bar{k})}\|_\infty \leq \|x - x^{(\bar{k})}\|_2 \leq 2^{-k}\|x\|_2 + 20\nu$), the inequality $\|x - x^{(\bar{k})}\|_\infty \leq \frac{0.45\eta}{s} - \frac{0.68\|e\|_2}{\sqrt{s}}$ is desired to be satisfied after some $\bar{k}$ iterations. Hence, $\|A^*(R^{(\bar{k})})\|_\infty \leq \eta/\sqrt{2s}$ can also be triggered at least when $\|x - x^{(\bar{k})}\|_\infty \leq \frac{0.45\eta}{s} - \frac{0.68\|e\|_2}{\sqrt{s}}$, which means it is an effective stopping criterion.

Lastly, we describe a simple effective variant of CoSaMP introduced in [26].

Algorithm 25 A Variant of CoSaMP

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m \times n}$; the sparsity level s

Initialization :

1. $x^{(0)} = 0$
2. $R^{(0)} = y$

Iteration : repeat until a stopping criterion is met at $k = \bar{k}$

1. compute $\langle R^{(k)}, a_j \rangle$ for $j \in [n]$ (equivalently, compute $c = A^*(R^{(k)})$)
2. Identification : $\Omega = supp(c_{2s})$
3. Estimation : $b|_\Omega = A_\Omega^\dagger(R^{(k)})$, $b|_{\Omega^c} = 0$
4. Approximation Merging : $c = x^{(k)} + b$
5. Pruning : $x^{(k+1)} = c_s$
6. $R^{(k+1)} = y - Ax^{(k+1)}$

Output: $x^{(\bar{k}+1)}$

When performing the estimation step, one can initialize the iterative least-squares solver with the zero vector to take advantage of the fact that $R^{(k)}$ is becoming smaller and smaller.

4.2.5 Subspace Pursuit (SP)

In this section, we will introduce an algorithm called the subspace pursuit. The readers can refer to [6]. First, we will present this algorithm, and then make a thorough mathematical analysis of it. The subspace pursuit algorithm is described as follows.

Algorithm 26 Subspace Pursuit

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m \times n}$; the sparsity level s
Initialization :

1. $T^{(0)} = L_s(A^* y)$
2. $R^{(0)} = y - A_{T^{(0)}} A^{\dagger}_{T^{(0)}} y$

Iteration : repeat until $\|R^{(\bar{k}+1)}\|_2 \geq \|R^{(\bar{k})}\|_2$ **at** $k = \bar{k} + 1$

1. compute $\langle R^{(k)}, a_j \rangle$ for $j \in [n]$ (equivalently, compute $c = A^*(R^{(k)})$)
2. Identification 1 : $\Omega = L_s(c)$
3. Merge Support : $\tilde{T}^{(k+1)} = T^{(k)} \cup \Omega$
4. Projection 1 : $(x_p)|_{\tilde{T}^{(k+1)}} = A^{\dagger}_{\tilde{T}^{(k+1)}} y$, $(x_p)|_{(\tilde{T}^{(k+1)})^c} = 0$
5. Identification 2 : $T^{(k+1)} = L_s(x_p)$
6. Projection 2 : $x^{(k+1)}|_{T^{(k+1)}} = A^{\dagger}_{T^{(k+1)}} y$, $x^{(k+1)}|_{(T^{(k+1)})^c} = 0$
7. $R^{(k+1)} = y - A x^{(k+1)}$

Output: $x^{\bar{k}}$

As we can see, the subspace pursuit algorithm iteratively refines the candidate support set of s indices in order to pursue the correct subspace y lies in. First, given $T^{(k)}$, s additional candidate indices are selected to form $\tilde{T}^{(k+1)}$. Then given $\tilde{T}^{(k+1)}$, s most promising indices are chosen out from it so that we will get $T^{(k+1)}$. We make a comparison between the OMP, the stagewise OMP and the ROMP with the subspace pursuit. An important difference between them lies in the way they construct the candidate support set. At each iteration, the former ones (OMP, stagewise OMP and ROMP) select one or several indices that are deemed to be possible to lie in the support set. Once an index is selected, it remains in the candidate set throughout the remainder of the reconstruction process. However, on the side of the subspace pursuit, an index is considered good at some iteration but deemed bad at a later iteration, or vise versa, can be removed from or added to the candidate support set. Such flexibility of the subspace pursuit may provide it with better performance. We also make a comparison between the subspace pursuit and the compressive sampling matching pursuit (CoSaMP). One of the differences is that at the first identification step, the subspace pursuit chooses s indices with the largest correlation magnitudes while the CoSaMP chooses $2s$ ones. The other difference is that after the estimation step of CoSaMP (which corresponds to what we call the first projection step of the sub-

space pursuit), the CoSaMP performs a pruning step which amounts to an identification step followed by a hard thresholding step while the subspace pursuit performs an identification step followed by another projection step.

In the following, we analyze the subspace pursuit algorithm mathematically. Let $x \in \mathbb{R}^n$ be an s-sparse signal and the measurement vector $y \in \mathbb{R}^m$ be $y = Ax + e$, where $e \in \mathbb{R}^m$ is the noise vector. It can be verified that we can express $R^{(k)}$ as

$$R^{(k)} = A_{T\backslash T^{(k)}} x|_{T\backslash T^{(k)}} + A_{T^{(k)}} v + e_r$$

, where $v \triangleq -A^{\dagger}_{T^{(k)}}(A_{T\backslash T^{(k)}} x|_{T\backslash T^{(k)}})$ and $e_r \triangleq e - A_{T^{(k)}} A^{\dagger}_{T^{(k)}} e$. Define $v' \triangleq v - A^{\dagger}_{T^{(k)}} e$. We can prove that

$$\|v'\|_2 - \frac{1}{\sqrt{1-\delta_s}}\|e\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}}\|x|_{T\backslash T^{(k)}}\|_2 \tag{4.100}$$

Hence, among all columns that jointly compose $R^{(k)}$, those indexed by $T\backslash T^{(k)}$ have energies that are not too small. As a consequence, after step 1 and 2 (which, combined together, are called a correlation maximization step), it is confident that Ω contains sufficient number of indices that belong to $T\backslash T^{(k)}$. Therefore, if we merge the two sets $T^{(k)}$ and Ω, it is hopeful that $\tilde{T}^{(k+1)}$ contains sufficient number of indices that belong to T. Indeed, it is proved that

$$\|x|_{T\backslash \tilde{T}^{(k+1)}}\|_2 \leq \frac{2\delta_{3s}}{(1-\delta_{3s})^2}\|x|_{T\backslash T^{(k)}}\|_2 + \frac{2\sqrt{1+\delta_s}}{1-\delta_s}\|e\|_2 \tag{4.101}$$

After the first projection step, we get x_p. Note that $(x_p)|_{\tilde{T}^{(k+1)}} = A^{\dagger}_{\tilde{T}^{(k+1)}} y$, we can further express it as

$$(x_p)|_{\tilde{T}^{(k+1)}} = x|_{\tilde{T}^{(k+1)}} + A^{\dagger}_{\tilde{T}^{(k+1)}} A_{T\backslash \tilde{T}^{(k+1)}} x|_{T\backslash \tilde{T}^{(k+1)}} + A^{\dagger}_{\tilde{T}^{(k+1)}} e$$

Let $\epsilon \triangleq (x_p)|_{\tilde{T}^{(k+1)}} - x|_{\tilde{T}^{(k+1)}}$. It is proved that

$$\|\epsilon\|_2 \leq \frac{\delta_{3s}}{1-\delta_{3s}}\|x|_{T\backslash \tilde{T}^{(k+1)}}\|_2 + \frac{1}{\sqrt{1-\delta_{2s}}}\|e\|_2 \tag{4.102}$$

Hence, we see that $(x_p)|_{\tilde{T}^{(k+1)}}$ is close to $x|_{\tilde{T}^{(k+1)}}$. Thus it is hopeful that $T^{(k+1)} = L_s(x_p)$ also contains sufficient number of indices that belong to T. Indeed, it is proved that

$$\|x|_{T\setminus T^{(k+1)}}\|_2 \leq \frac{1+\delta_{3s}}{1-\delta_{3s}}\|x|_{T\setminus \tilde{T}^{(k+1)}}\|_2 + \frac{2}{\sqrt{1-\delta_{2s}}}\|e\|_2 \tag{4.103}$$

Combining the two inequalities 4.101 and 4.103, we can come up with the inequality

$$\|x|_{T\setminus T^{(k+1)}}\|_2 \leq \frac{2\delta_{3s}(1+\delta_{3s})}{(1-\delta_{3s})^3}\|x|_{T\setminus T^{(k)}}\|_2 + \frac{4(1+\delta_{3s})}{(1-\delta_{3s})^2}\|e\|_2 \tag{4.104}$$

Suppose that $\delta_{3s} < 0.083$. If $\|e\|_2 \leq \delta_{3s}\|x|_{T\setminus T^{(k)}}\|_2$, one can prove that $\|R^{(k+1)}\|_2 \leq \|R^{(k)}\|_2$ using inequality 4.104. Lastly, lemma 3 of [6] guarantees that the distance between the true signal x and the result after the second projection step, i.e. $x^{(k+1)}$, is not too large. Precisely,

$$\|x - x^{(k+1)}\|_2 \leq \frac{1}{1-\delta_{3s}}\|x|_{T\setminus T^{(k+1)}}\|_2 + \frac{1+\delta_{3s}}{1-\delta_{3s}}\|e\|_2 \tag{4.105}$$

When the stopping criterion is triggered (i.e., $\|R^{(\bar{k}+1)}\|_2 \geq \|R^{(\bar{k})}\|_2$), $\|e\|_2 \geq \delta_{3s}\|x|_{T\setminus T^{(\bar{k})}}\|_2$. Therefore,

$$\|x - x^{(\bar{k})}\|_2 \leq \frac{1}{1-\delta_{3s}}\frac{1}{\delta_{3s}}\|e\|_2 + \frac{1+\delta_{3s}}{1-\delta_{3s}}\|e\|_2 = \frac{1+\delta_{3s}+\delta_{3s}^2}{\delta_{3s}(1-\delta_{3s})}\|e\|_2 \tag{4.106}$$

We can summarize our discussion as the following theorem.

Theorem 4.2.10. *Let $x \in \mathbb{R}^n$ be an s-sparse signal and the measurement vector $y \in \mathbb{R}^m$ be $y = Ax + e$, where $e \in \mathbb{R}^m$ is the noise vector. Suppose that the sampling matrix A satisfies $\delta_{3s} < 0.083$, then $\|x - x^{(\bar{k})}\|_2 \leq \frac{1+\delta_{3s}+\delta_{3s}^2}{\delta_{3s}(1-\delta_{3s})}\|e\|_2$.*

When $e = 0$, the theorem implies that $x^{(\bar{k})} = x$, which means we can achieve exact reconstruction when the algorithm terminates. Furthermore, theorem 6 of [6] states that the number of iterations need for the

algorithm (denoted as n_{it}) satisfies

$$n_{it} \leq \min\{1 + \frac{\log \rho_{min}}{\log c_s}, -\frac{1.5s}{\log c_s}\} \tag{4.107}$$

where $\rho_{min} \triangleq \frac{\min_{i\in[n]} |x[i]|}{\|x\|_2}$ and $c_s \triangleq \frac{2\delta_{3s}(1+\delta_{3s})}{(1-\delta_{3s})^3}$. Specifically, for zero-one sparse signals, $n_{it} \leq \frac{\log s}{2\log(1/c_s)}$. For compressible sparse signals, $n_{it} \leq \frac{\log s}{r\log(1/c_s)}(1 + o(1))$, where $o(1) \to 0$ when $s \to \infty$. For exponentially decaying signals, $n_{it} \leq \begin{cases} \frac{ps}{\log(1/c_s)}(1 + o(1)) & if\ 0 < p \leq 1.5 \\ \frac{1.5s}{\log(1/c_s)} & if\ p > 1.5 \end{cases}$. Hence, $n_{it} = \mathcal{O}(\log s)$ for zero-one sparse signals and compressible sparse signals, and $n_{it} = \mathcal{O}(s)$ for exponentially decaying signals. If x is not exactly s-sparse and $e \neq 0$, we can express y as $y = Ax_{2s} + A(x - x_{2s}) + e$. Assume $\delta_{6s} < 0.083$. By the theorem 4.2.10, we can get

$$\|x_{2s} - x^{(\bar{k})}\|_2 \leq \frac{1 + \delta_{6s} + \delta_{6s}^2}{\delta_{6s}(1 - \delta_{6s})}(\|A(x - x_{2s})\|_2 + \|e\|_2)$$

if we set the input sparsity level as $2s$. By 2.91, we can derive that

$$\|A(x - x_{2s})\|_2 \leq \sqrt{1 + \delta_{6s}}(\|x - x_{2s}\|_2 + \frac{\|x - x_{2s}\|_1}{\sqrt{6s}})$$

By lemma 4.2.8, we can get $\|x - x_{2s}\|_2 \leq \frac{\|x - x_s\|_1}{2\sqrt{s}}$. Therefore, we can derive that

$$\|x_{2s} - x^{(\bar{k})}\|_2 \leq \frac{1 + \delta_{6s} + \delta_{6s}^2}{\delta_{6s}(1 - \delta_{6s})}(\|e\|_2 + \sqrt{1 + \delta_{6s}}\frac{\|x - x_s\|_1}{\sqrt{s}}) \tag{4.108}$$

Furthermore,

$$\begin{aligned} \|x - x^{(\bar{k})}\|_2 &\leq \|x_{2s} - x^{(\bar{k})}\|_2 + \|x - x_{2s}\|_2 \\ &\leq (1 + \frac{1 + \delta_{6s} + \delta_{6s}^2}{\delta_{6s}(1 - \delta_{6s})})(\|e\|_2 + \sqrt{1 + \delta_{6s}}\frac{\|x - x_s\|_1}{\sqrt{s}}) \\ &= \frac{1 + 2\delta_{6s}}{\delta_{6s}(1 - \delta_{6s})}(\|e\|_2 + \sqrt{1 + \delta_{6s}}\frac{\|x - x_s\|_1}{\sqrt{s}}) \end{aligned} \tag{4.109}$$

We can summarize our discussion in the following corollary.

Corollary 4.2.10.1. *Let $x \in \mathbb{R}^n$ be the approximately s-sparse signal and let the measurement vector $y = Ax + e$, where $e \in \mathbb{R}^m$ is the noise vector. Suppose that the sampling matrix A satisfies $\delta_{6s} < 0.083$. Then*

1. $\|x_{2s} - x^{(\bar{k})}\|_2 \leq \frac{1+\delta_{6s}+\delta_{6s}^2}{\delta_{6s}(1-\delta_{6s})}(\|e\|_2 + \sqrt{1+\delta_{6s}}\frac{\|x-x_s\|_1}{\sqrt{s}})$
2. $\|x - x^{(\bar{k})}\|_2 \leq \frac{1+2\delta_{6s}}{\delta_{6s}(1-\delta_{6s})}(\|e\|_2 + \sqrt{1+\delta_{6s}}\frac{\|x-x_s\|_1}{\sqrt{s}})$

4.3 Hard-Thresholding-Based Algorithms

Throughout this whole chapter, our main focus is the rectangular system $y = Ax$. Solving $y = Ax$ is equivalent to solving $A^*Ax = A^*y$. With such normal equation, we can come up with the fixed-point equation $x_{fixed} = x_{fixed} + A^*(y - Ax_{fixed})$ and derive the fixed-point iteration

$$x^{(k+1)} = x^{(k)} + A^*(y - Ax^{(k)})$$

Similar to greedy algorithms, it also involves the computation of correlation vector $A^*(y - Ax^{(k)})$. However, the underlying mechanisms are different. For the greedy algorithms, they compute the correlation vector in order to capture additional atoms that would be highly possible to compose the signal vector y. On the other hand, thresholding-based algorithms add the correlation vector and the current estimate coefficient vector $x^{(k)}$ together to form a surrogate coefficient vector for x. In section 4.1.3, we have introduced the ISTA algorithm 4.37. It applies the soft thresholding operator on the surrogate coefficient vector to obtain the new estimate coefficient vector $x^{(k+1)}$. As for this section, hard-thresholding-based algorithms apply the hard thresholding operator on the surrogate coefficient vector. The difference of adding ℓ_1 norm or ℓ_0 norm of x in the objective functions lead to the difference of thresholding operators to be soft or hard.

4.3.1 Iterative Hard Thresholding (IHT)

In [3] and [4], two main problems are discussed.

$$\min\{\|Ax - y\|_2^2 + \lambda\|x\|_0\} \tag{4.110}$$
$$\min\{\|Ax - y\|_2^2\} \; subject\ to\ \|x\|_0 \leq s \tag{4.111}$$

Clearly, these two problems are quite related to problem 4.2. The first problem is called the ℓ_0 regularized minimization problem, which can be viewed as the ℓ_0-norm counterpart of the basis pursuit denoising problem 4.21. The second problem is called the s-sparse minimization problem, which can be viewed as the ℓ_0-norm counterpart of the LASSO problem 4.22. The authors proposed two algorithms, which we call the IHT_λ algorithm and the IHT_s algorithm respectively, to deal with these two problems. In the following, we introduce these two algorithms thoroughly.

For the first problem, if we directly deal with the objective function

$$C_r(x) \triangleq \|Ax - y\|_2^2 + \lambda\|x\|_0 \tag{4.112}$$

we need to calculate the derivative of $C_r(x)$ with respect to x. It is easily derived that

$$\frac{\partial C_r(x)}{\partial x[i]} = \begin{cases} 0 & if\ x[i] = 0 \\ -2\langle a_i, y\rangle + \lambda + 2\langle a_i, Ax\rangle & if\ x[i] \neq 0 \end{cases}$$

For $x[i] \neq 0$, we need to let $-2\langle a_i, y\rangle + \lambda + 2\langle a_i, Ax\rangle$ be 0. However, all entries of x are coupled, which hinders us from deriving the value of $x[i]$. It is the $\|Ax\|_2^2$ term that causes the problem. Thus, it motivates us to consider a surrogate objective function

$$C_r^S(x, z) \triangleq \|Ax - y\|_2^2 + \lambda\|x\|_0 - \|Ax - Az\|_2^2 + \|x - z\|_2^2 \tag{4.113}$$

The reason why we add an additional argument z is that we can assign x as the next iterate and z as the current iterate in an iterative algorithm. Such trick will be made clear soon. If $\|A\|_{2\to 2} < 1$, then this surrogate objective function is a majorization of the objective function and minimization of the surrogate function thus leads to a majorization

minimization (MM) algorithm. Similarly, we also calculate the partial derivative of the surrogate objective function with respect to x_i. In this case,

$$\frac{\partial C_r^S(x)}{\partial x[i]} = \begin{cases} 0 & if\ x[i] = 0 \\ 2x[i] - 2z[i] - \langle a_i, y\rangle - \langle a_i, Az\rangle & if\ x[i] \neq 0 \end{cases}$$

Now all entries of x are decoupled and we can derive that $x[i] = z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle$ when $x[i] \neq 0$. The corresponding cost is 0 if $x[i] = 0$ and $-(z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle)^2 + \lambda$ if $x[i] = z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle$. As a result, if $z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle < \lambda^{0.5}$, then we set $x[i] = 0$, the resulting cost being 0. If $z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle > \lambda^{0.5}$, then we set $x[i] = z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle$, the resulting cost being $-(z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle)^2 + \lambda$. If $z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle = \lambda^{0.5}$, $x[i]$ can be set as either 0 or $z[i] + \langle a_i, y\rangle - \langle a_i, Az\rangle$, the resulting cost being 0. To produce unique update, we let $x = H_{\lambda^{0.5}}(z + A^*(y - Az))$, where $H_{\lambda^{0.5}}$ is the element wise hard thresholding operator defined as

$$H_{\lambda^{0.5}}(x[i]) = \begin{cases} 0 & if\ |x[i]| \leq \lambda^{0.5} \\ x[i] & if\ |x[i]| > \lambda^{0.5} \end{cases} \tag{4.114}$$

Hence, we can derive the IHT_λ algorithm as follows.

Algorithm 27 IHT_λ

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; the regularization parameter λ
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = H_{\lambda^{0.5}}(x^{(k)} + A^*(y - Ax^{(k)})) \tag{4.115}$$

Output: $x^{(\bar{k})}$

As for the second problem, similar to the first problem, instead of directly dealing with the objective function

$$C_s(x) \triangleq \|Ax - y\|_2^2 \tag{4.116}$$

we consider the surrogate objective function

$$C_s^S(x, z) \triangleq \|Ax - y\|_2^2 - \|Ax - Az\|_2^2 + \|x - z\|_2^2 \tag{4.117}$$

Likewise, $\|A\|_{2\to 2}$ should be smaller than 1 so that we can derive an MM algorithm. After some calculations, we can derive the IHT_s algorithm as follows.

Algorithm 28 IHT_s

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; the sparsity level s
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$x^{(k+1)} = H_s(x^{(k)} + A^*(y - Ax^{(k)})) \tag{4.118}$$

Output: $x^{(\bar{k})}$

Note that H_s, which has been defined in 2.68, is the hard thresholding operator which only retains the s entries with the largest magnitude.

In the following, we list important results that have been proved in [3] about the two algorithms.

1. Assume $\|A\|_{2\to 2} < 1$ and let $x^{(k+1)} = H_{\lambda^{0.5}}(x^{(k)} + A^*(y - Ax^{(k)}))$, then the sequences $\{C_r(x^{(k)})\}$ and $\{C_r^S(x^{(k+1)}, x^{(k)})\}$ are non-increasing. Assume $\|A\|_{2\to 2} < 1$ and let $x^{(k+1)} = H_s(x^{(k)} + A^*(y - Ax^{(k)}))$, then the sequences $\{C_s(x^{(k)})\}$ and $\{C_s^S(x^{(k+1)}, x^{(k)})\}$ are non-increasing.

   - This bullet states that the cost does not increase from iteration to iteration, which implies that using the algorithms will only produce better results than not applying them.

2. Define two sets as follows

$$\Gamma_0 = \{i : x_{fixed}[i] = 0\} \tag{4.119}$$
$$\Gamma_1 = \{i : x_{fixed}[i] \neq 0\} \tag{4.120}$$

   A necessary and sufficient condition for a point x_{fixed} to be a fixed point of the IHT_λ algorithm is that for each column a_i of A,

$$|\langle a_i, y - Ax_{fixed}\rangle| \begin{cases} = 0 & if\ i \in \Gamma_1 \\ \leq \lambda^{0.5} & if\ i \in \Gamma_0 \end{cases} \tag{4.121}$$

   A necessary and sufficient condition for a point x_{fixed} to be a fixed

point of the IHT_s algorithm is that for each column a_i of A,

$$|\langle a_i, y - Ax_{fixed}\rangle| \begin{cases} = 0 & if\ i \in \Gamma_1 \\ \leq x^*_{fixed}[s] & if\ i \in \Gamma_0 \end{cases} \tag{4.122}$$

where x^*_{fixed} denotes the non-increasing rearrangement of x_{fixed}. Note that a fixed point of the IHT_λ satisfies $|x_{fixed}[i]| > \lambda^{0.5}$ for $i \in \Gamma_1$ and that of the IHT_s algorithm satisfies $|x_{fixed}[i]| \geq x^*_{fixed}[s]$ for $i \in \Gamma_1$.

- This bullet states the necessary and sufficient condition of a fixed point, which facilitates the following analysis about the local minima, the optimal solution and the convergence issues about the two problems and algorithms.

3. Assume $\|A\|_{2\to 2} < 1$, then a fixed point x_{fixed} of the IHT_λ / IHT_s algorithm is a local minimum of the first / second problem.

   - This bullet makes a connection between a fixed point of the algorithms and a local minimum of the problems. Note that if a point is referred to as a local minimum, then perturbing it by an infinitesimal amount (in any direction) will not decrease the cost function. Mathematically, we want to show that

   $$C_r(x_{fixed} + \partial h) > C_r(x_{fixed}) \tag{4.123}$$
   $$C_s(x_{fixed} + \partial h) > C_s(x_{fixed}) \tag{4.124}$$

   for any small perturbation $|\partial h[i]| < \epsilon$, where ϵ is some positive constant.

4. Let the optimal solution to the first problem (i.e., the global minimum) be x_{opt}.

   (a) $\forall i \in \Gamma_1$, $|x_{opt}[i]| \geq \lambda^{0.5}$

   (b) $\forall i \in \Gamma_0$, $\langle a_i, y - Ax_{opt}\rangle \leq \lambda^{0.5}$

   (c) $\forall i \in \Gamma_1$, $\langle a_i, y - Ax_{opt}\rangle = 0$

   - The third condition points out that the residual $R_{opt} \triangleq y - Ax_{opt}$ is orthogonal to all used atoms and forms small angles with all

unused atoms. Thus, the optimal solution corresponds to the situation when the signal is projected as orthogonally as possible onto the space spanned by the atoms of A while restricting the number of used atoms.

5. Assume $\|A\|_{2\to 2} < 1$, then the optimal solution to the first problem belongs to the fixed points of the IHT_λ algorithm.

   - Combining the second and fourth bullets, we can come up with this result.

6. $\forall \epsilon > 0$, $\exists K$ such that $\forall k > K$, $\|x^{(k+1)} - x^{(k)}\|_2^2 \leq \epsilon$; that is,

$$\lim_{k\to\infty} \|x^{(k+1)} - x^{(k)}\|_2^2 = 0 \tag{4.125}$$

   where $\{x^{(k)}\}$ is the iterate sequence of either the IHT_λ or IHT_s algorithm.

   - This bullet is an important result for the proof of convergence.

7. If $C_r(x^{(0)}) < \infty$, and if $\|A\|_{2\to 2} < 1$, then the sequence $\{x^{(k)}\}$ produced by the IHT_λ algorithm converges to a fixed point of it (thus a local minimum of the first problem).
   If $C_s(x^{(0)}) < \infty$, $\{a_i\}$ contains a basis for the signal space, $\|a_i\|_2 > 0$ and $\|A\|_{2\to 2} < 1$, then the sequence $\{x^{(k)}\}$ produced by the IHT_s algorithm converges to a fixed point of it (thus a local minimum of the second problem).

   - With the help of the sixth bullet, we can prove the convergence of both algorithms and thus lead to this result.

8. Assume that $\{a_i\}$ contains a basis for the signal space and that $\|a_i\|_2 > c > 0$, then there exists a constant $\beta(A) > 0$ such that $\sup_i |\langle a_i, y\rangle| \geq \beta(A)\|y\|_2$ holds for all y. If $\|A\|_{2\to 2} \leq 1$, then

$$\|y - Ax_{fixed}\|_2 \leq \frac{\lambda^{0.5}}{\beta(A)} \tag{4.126}$$

when x_{fixed} is a fixed point of the IHT_λ algorithm and

$$\|y - Ax_{fixed}\|_2 \leq \frac{x^*_{fixed}[s]}{\beta(A)} \tag{4.127}$$

when x_{fixed} is a fixed point of the IHT_s algorithm.

- This bullet gives an upper bound on the approximation error, i.e., the norm of the residual $R_{fixed} \triangleq y - Ax_{fixed}$.

9.

$$\|x^{(n)} - x_{fixed}\|_2 \leq \|I - A^*_{\Gamma_1} A_{\Gamma_1}\|^{(n-m)}_{2\to 2} \|x^{(m)} - x_{fixed}\|_2 \tag{4.128}$$

where x_{fixed} is the fixed point of either the IHT_λ or IHT_s algorithm.

- This bullet shows that the asymptotic convergence speed of both algorithms is linear.

There are two common usages of both the IHT_λ and IHT_s algorithms. One is to randomly initialize $x^{(0)}$ and start the algorithms directly. The other one is to use them to refine the solutions found with other methods. Concretely speaking, we first use methods like the MP to find a solution, then initialize $x^{(0)}$ with this solution and execute the algorithms. Since the cost is guaranteed not to increase from iteration to iteration, we are confident to get a better solution. Moreover, for IHT_λ, we can adaptively change the threshold depending on the current residual norm. As for IHT_s, we can adaptively change the parameter s from iteration to iteration. A suggested way is that we start from $s = 1$ and increase s by 1 every ℓ iterations. If $\ell = 1$, the corresponding algorithm functions like the MP. As ℓ gets larger, it becomes gradually more and more like the OMP. Note that an advantage of the IHT_s algorithm over the OMP is that the set of selected atoms is allowed to change from iteration to iteration. Another advantage is that its computational cost is so low that it is comparable to the MP.

In [4], the authors proposed to apply the IHT_s algorithm in the field of compressive sensing and analyzed its performance mathemati-

cally. We list the main result as follows.

Theorem 4.3.1. *Given a noisy observation $y = Ax+e$. Assume A has the restricted isometry property with $\delta_{3s} < 1/\sqrt{32}$. Initializing $x^{(0)}$ as the zero vector, then at iteration k, the iterate $x^{(k)}$ satisfies*

$$\|x^{(k)} - x\|_2 \leq 2^{-k}\|x\|_2 + c\nu \tag{4.129}$$

where $\nu \triangleq \|x-x_s\|_2 + \frac{\|x-x_s\|_1}{\sqrt{s}} + \|e\|_2$ is the unrecoverable energy the same as what has already been defined in [26]. c is equal to 6 for any arbitrary signal and can be improved to be 5 if x is s-sparse. Also note that ν reduces to $\|e\|_2$ if x is s-sparse.

From this result, we can derive that after at most $\bar{k} = \lceil \log_2(\frac{\|x\|_2}{\nu}) \rceil$ iterations, IHT_s can produce the output $x^{(\bar{k})}$ satisfying $\|x^{(\bar{k})} - x\|_2 \leq (c+1)\nu$. Thus, the overall number of iterations required to achieve a desirable accuracy depends on the logarithm of the signal-to-noise ratio (note that the "noise" here not only accounts for the measurement error but also the sparsity defect). From this result, it can also be proved that the output $x^{(\bar{k})}$ satisfies $\|x^{(\bar{k})} - x\|_2 \leq 1.11\epsilon + 2.41\nu$ if the stopping criterion is set as $\|y - Ax^{(\bar{k})}\|_2 \leq \epsilon$ for some $\epsilon > 0$.

Finally, the authors pointed out that such uniform performance guarantees regarding the restricted isometry property are suitable for analyzing the worst-case scenario but not for average performance. Indeed, numerical experiments which can only analyze the average behavior demonstrates that algorithms with similar uniform guarantees have discernible performance (i.e., some perform better while others perform worse). Furthermore, in the regime when the restricted isometry constant is too large, some algorithms still perform well on average, while others do not.

4.3.2 Hard Thresholding Pursuit (HTP)

Inspired by the IHT algorithm, in [12], the author proposed the hard thresholding pursuit (HTP) algorithm, described as follows.

Algorithm 29 HTP

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; the sparsity level s
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$(HTP_1) \quad S^{(k+1)} = L_s(x^{(k)} + A^*(y - Ax^{(k)})) \tag{4.130}$$

$$(HTP_2) \quad x^{(k+1)} = argmin\{\|y - Az\|_2, supp(z) \subseteq S^{(k+1)}\} \tag{4.131}$$

Output: $x^{(\bar{k})}$

Note that L_s is the operator defined in 2.67. The author also generalized the HTP algorithm and proposed some variants of it; namely, the hard thresholding pursuit μ (HTP^μ) algorithm, the normalized hard thresholding pursuit (NHTP) algorithm, and the fast hard thresholding pursuit μ ($FHTP^\mu$) algorithm. We describe them respectively as follows.

Algorithm 30 HTP^μ

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; the sparsity level s
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$(HTP_1^\mu) \quad S^{(k+1)} = L_s(x^{(k)} + \mu A^*(y - Ax^{(k)})) \tag{4.132}$$

$$(HTP_2^\mu) \quad x^{(k+1)} = argmin\{\|y - Az\|_2, supp(z) \subseteq S^{(k+1)}\} \tag{4.133}$$

Output: $x^{(\bar{k})}$

Algorithm 31 NHTP

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; the sparsity level s
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$(NHTP_1) \quad S^{(k+1)} = L_s(x^{(k)} + \mu_k A^*(y - Ax^{(k)})),\ where\ \mu_k = \frac{\|(A^*(y - Ax^{(k)}))_{S^{(k)}}\|_2^2}{\|A_{S^{(k)}}((A^*(y - Ax^{(k)}))_{S^{(k)}})\|_2^2} \tag{4.134}$$

$$(NHTP_2) \quad x^{(k+1)} = argmin\{\|y - Az\|_2, supp(z) \subseteq S^{(k+1)}\} \tag{4.135}$$

Output: $x^{(\bar{k})}$

Algorithm 32 $FHTP^\mu$

Input: $y \in \mathbb{R}^m$; $A \in \mathbb{R}^{m\times n}$; the sparsity level s
Initialization : $x^{(0)} \in \mathbb{R}^n$
Iteration : repeat until a stopping criterion is met at $k = \bar{k}$:

$$(FHTP_1^\mu) \quad S^{(k+1)} = supp(u^{(k+1,1)}),\ u^{(k+1,1)} := H_s(x^{(k)} + \mu A^*(y - Ax^{(k)})) \tag{4.136}$$

$$\begin{aligned}(FHTP_2^\mu) \quad & x^{(k+1)} = u^{(k+1,\ell+1)},\ u^{(k+1,n+1)} := (u^{(k+1,n)} + t_{k+1,n}A^*(y - Au^{(k+1,n)}))|_{S^{(k+1)}} \\ & where\ t_{k+1,n} := \frac{\|(A^*(y - Au^{(k+1,n)}))_{S^{(k+1)}}\|_2^2}{\|A_{S^{(k+1)}}((A^*(y - Au^{(k+1,n)}))_{S^{(k+1)}})\|_2^2}\ and\ n = 1, 2, \cdots, \ell + 1.\end{aligned} \tag{4.137}$$

Output: $x^{(\bar{k})}$

For the NHTP algorithm, we can prove that μ_k is chosen such that $\mu_k = \underset{\mu}{argmin} \|y - A_{S^{(k)}}[(x^{(k)} + \mu A^*(y - Ax^{(k)}))|_{S^{(k)}}]\|_2^2$.

$$
\begin{aligned}
&Let\ R^{(k)} \triangleq y - Ax^{(k)} = y - A_{S^{(k)}}x^{(k)}|_{S^{(k)}} \\
&\mu_k = \underset{\mu}{argmin} \|R^{(k)} - \mu A_{S^{(k)}}(A^*R^{(k)})|_{S^{(k)}}\|_2^2 \\
&\quad = \underset{\mu}{argmin}[-2\mu\langle R^{(k)}, A_{S^{(k)}}(A^*R^{(k)})|_{S^{(k)}}\rangle + \mu^2\|A_{S^{(k)}}(A^*R^{(k)})|_{S^{(k)}}\|_2^2] \\
&\quad = \underset{\mu}{argmin}[-2\mu\|(A^*R^{(k)})|_{S^{(k)}}\|_2^2 + \mu^2\|A_{S^{(k)}}(A^*R^{(k)})|_{S^{(k)}}\|_2^2] \\
&\quad = \frac{\|(A^*(y - Ax^{(k)}))_{S^{(k)}}\|_2^2}{\|A_{S^{(k)}}((A^*(y - Ax^{(k)}))_{S^{(k)}})\|_2^2}
\end{aligned}
$$

For the $FHTP^{\mu}$ algorithm, we can prove that $t_{k+1,n}$ is chosen such that $t_{k+1,n} = \underset{t}{argmin} \|y - A_{S^{(k+1)}}[(u^{(k+1,n)} + tA^*(y - Au^{(k+1,n)}))|_{S^{(k+1)}}]\|_2^2$. The derivation is similar as above. Hence, as we can see, what we do in $FHTP_2^{\mu}$ is actually applying the gradient descent with exact line search to solve the orthogonal projection minimization problem. Note that if $\mu = 1$, the corresponding algorithm is called the FHTP algorithm. Furthermore, if $k = 0$, it reduces to the IHT algorithm and if $k = \infty$, it approaches the HTP algorithm.

In the following, we present some results about these algorithms.

1. The iterates $\{x^{(k)}\}$ produced by HTP, HTP^{μ} or NHTP are eventually periodic since there is only a finite number of subsets of $[n]$ with size s and the result of orthogonal projection is fixed given the same support set.
   - Because of this fact, once we can prove the convergence of any of these algorithms, we can certify that the limit is exactly achieved within a finite number of iterations.
2. Applying the HTP^{μ} algorithm, we can deduce that
$$\|y - Ax^{(k+1)}\|_2^2 - \|y - Ax^{(k)}\|_2^2 \leq (1 + \delta_{2s} - 1/\mu)\|x^{(k)} - u^{(k+1)}\|_2^2 \tag{4.138}$$
3. If $\mu(1 + \delta_{2s}) < 1$, then $\{x^{(k)}\}$ produced by the HTP^{μ} algorithm

converges in a finite number of iterations.

Proof. Since $\mu(1+\delta_{2s}) < 1$, $\|y - Ax^{(k+1)}\|_2^2 - \|y - Ax^{(k)}\|_2^2 \leq (1 + \delta_{2s} - 1/\mu)\|x^{(k)} - u^{(k+1)}\|_2^2 \leq 0$. Hence, $\{\|y - Ax^{(k)}\|_2^2\}$ is a non-increasing sequence with lower bound 0, which implies it is convergent. Because $\{x^{(k)}\}$ is eventually periodic, $\|y - Ax^{(k)}\|_2^2$ must be a constant eventually. For the inequality of the second bullet to hold, $\|x^{(k)} - u^{(k+1)}\|_2^2$ must be zero eventually. When this happens, $u^{(k+1)} = x^{(k)}$, which implies $x^{(k+1)} = x^{(k)}$. □

4. For any $k \geq 0$, if $\mu(1+\delta_{2s}) < 1$ and $\mu \geq 1/2$, then $\{x^{(k)}\}$ produced by the $FHTP^{\mu}$ converges.

   - Note that convergence "in a finite number of iterations" is not guaranteed because $FHTP^{\mu}$ does not satisfy the first bullet.

5. Suppose the $3s$-th restricted isometry constant of A satisfies $\delta_{3s} < \frac{1}{\sqrt{3}} \approx 0.57735$. Then for any s-sparse $x \in \mathbb{R}^n$, the iterates $\{x^{(k)}\}$ produced by HTP with $y = Ax$ converges toward x at a linear rate given by

$$\|x^{(k)} - x\|_2 \leq \rho^k \|x^{(0)} - x\|_2, \textit{ where } \rho := \sqrt{\frac{2\delta_{3s}^2}{1 - \delta_{2s}^2}} < 1 \quad (4.139)$$

   Furthermore, since HTP satisfies the first bullet, $\{x^{(k)}\}$ is guaranteed to converge in a finite number of iterations. Indeed, it can converge in at most $\lceil \frac{\ln(\sqrt{2/3}\|x^{(0)} - x\|_2/\xi)}{\ln(1/\rho)} \rceil$ iterations, where ξ is the smallest nonzero entry of x in modulus.

6. Suppose the $3s$-th restricted isometry constant of A satisfies $\delta_{3s} < \frac{1}{\sqrt{3}} \approx 0.57735$. Then for any s-sparse $x \in \mathbb{R}^n$, the iterates $\{x^{(k)}\}$ produced by FHTP ($t_{k+1,n}$ is set as 1 instead of the optimal one in this analysis) with $y = Ax$ converges toward x at a linear rate

given by

$$\|x^{(k)} - x\|_2 \leq \rho^k \|x^{(0)} - x\|_2$$
$$where\ \rho := \sqrt{\frac{\delta_{3s}^{2k+2}(1 - 3\delta_{3s}^2) + 2\delta_{3s}^2}{1 - \delta_{3s}^2}} < 1 \tag{4.140}$$

- Note that the restricted isometry condition $\delta_{3s} < \frac{1}{\sqrt{3}}$ is independent of the number ℓ of descent iterations used in $FHTP_2$.
- When $k = 0$, the FHTP algorithm reduces to the IHT algorithm. Thus, this bullet also implies the IHT algorithm allows s-sparse recovery once $\delta_{3s} < \frac{1}{\sqrt{3}}$.
- Although $\{x^{(k)}\}$ produced by the FHTP algorithm is not guaranteed to converge in a finite number of iterations, it can be verified that the norm of the residual $r^{(\bar{k})} \triangleq x^{(\bar{k})} - x$ does not exceed ϵ once $\bar{k} \geq \lceil \frac{\ln(\|x^{(0)} - x\|_2/\epsilon)}{\ln(1/\rho)} \rceil$.

7. Suppose the $3s$-th restricted isometry constant of A satisfies $\delta_{3s} < \frac{1}{\sqrt{3}} \approx 0.57735$. Then for any $x \in \mathbb{R}^n$ and any $e \in \mathbb{R}^m$, if S denotes an index set of s largest (in modulus) entries of x, the iterates $\{x^{(k)}\}$ produced by HTP with $y = Ax + e$ satisfies

$$\|x^{(k)} - x_s\|_2 \leq \rho^k \|x^{(0)} - x_s\|_2 + \tau \frac{1 - \rho^k}{1 - \rho} \|Ax|_{\overline{S}} + e\|_2$$
$$where\ \rho := \sqrt{\frac{2\delta_{3s}^2}{1 - \delta_{2s}^2}} < 1 and$$
$$\tau := \frac{\sqrt{2(1 - \delta_{2s})} + \sqrt{1 + \delta_s}}{1 - \delta_{2s}} \leq 5.15 \tag{4.141}$$

8. Suppose the $6s$-th restricted isometry constant of A satisfies $\delta_{6s} < \frac{1}{\sqrt{3}}$. Then for any $x \in \mathbb{R}^n$ and any $e \in \mathbb{R}^m$, every cluster point $x^\star$ of the iterates $\{x^{(k)}\}$ produced by HTP with s replaced with $2s$

and with $y = Ax + e$ satisfies

$$\|x - x^\star\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\|e\|_2,\ 1 \leq p \leq 2$$

where C and D depend only on δ_{6s}.

(4.142)

- This bullet is derived based on the results of the seventh bullet, the inequalities 2.8 and 2.66 and the following lemma excerpted from the lemma 6.10 of [13].

  Lemma 4.3.2. *Given $q > p > 0$, if $u \in \mathbb{R}^s$ and $v \in \mathbb{R}^t$ satisfy $\max_{i\in[s]} |u_i| \leq \min_{j\in[t]} |v_j|$, then*

$$\|u\|_q \leq \frac{s^{1/q}}{t^{1/p}}\|v\|_p \tag{4.143}$$

9. Suppose the $3s$-th restricted isometry constant of A and μ satisfy $\frac{1-1/\sqrt{3}}{1-\delta_{3s}} < \mu < \frac{1}{1+\delta_{3s}}$. Then the iterates $\{x^{(k)}\}$ produced by HTP^μ with $y = Ax + e$ converges eventually and satisfies the same inequality as 4.141 with δ_s, δ_{2s} and δ_{3s} in the expressions of ρ and τ be replaced with $\delta_s(\sqrt{\mu}A), \delta_{2s}(\sqrt{\mu}A)$ and $\delta_{3s}(\sqrt{\mu}A)$.

   Proof. Due to the third bullet, we know that a sufficient condition for $\{x^{(k)}\}$ to converge is $\mu(1 + \delta_{2s}) < 1$. As for the sparse recovery result with sparsity defect and measurement error, we observe that applying HTP^μ with input $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$ is the same as applying HTP with input $y' = \sqrt{\mu}y \in \mathbb{R}^m$ and $A' = \sqrt{\mu}A \in \mathbb{R}^{m\times n}$. Hence, a sufficient condition would be $\delta_{3s}(\sqrt{\mu}A) < 1/\sqrt{3}$.

   Note that $\mu(1 + \delta_{3s}) > \mu(1 + \delta_{2s})$. Thus the condition that $\mu < \frac{1}{1+\delta_{3s}}$ would imply $\mu(1 + \delta_{2s}) < 1$. On the other hand, according to the definition of δ_{3s}, we know that $(1 - \delta_{3s})\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_{3s})\|x\|_2^2$ for any s-sparse $x \in \mathbb{R}^n$. As a result, $\mu(1 - \delta_{3s})\|x\|_2^2 \leq \|\sqrt{\mu}Ax\|_2^2 \leq \mu(1 + \delta_{3s})\|x\|_2^2$. Let $\delta_{3s}(\sqrt{\mu}A)$ denote the $3s$-th restricted isometry constant of $\sqrt{\mu}A$. We know that $(1-\delta_{3s}(\sqrt{\mu}A))\|x\|_2^2 \leq \|\sqrt{\mu}Ax\|_2^2 \leq (1+\delta_{3s}(\sqrt{\mu}A))\|x\|_2^2$. Let

$1-c_1 = \mu(1-\delta_{3s})$, i.e., $c_1 = 1-\mu(1-\delta_{3s})$ and $1+c_2 = \mu(1+\delta_{3s})$, i.e., $c_2 = \mu(1+\delta_{3s})-1$. According to the definition of $\delta_{3s}(\sqrt{\mu}A)$, $\delta_{3s}(\sqrt{\mu}A)$ would be smaller than $\max\{c_1, c_2\}$. Thus if we enforce the condition that $1/\sqrt{3} > \max\{c_1, c_2\}$, i.e., $\frac{1-1/\sqrt{3}}{1-\delta_{3s}} < \mu < \frac{1+1/\sqrt{3}}{1+\delta_{3s}}$, then we can ensure that $\delta_{3s}(\sqrt{\mu}A) < 1/\sqrt{3}$. Hence, combining the two conditions, we can get a sufficient condition that $\frac{1-1/\sqrt{3}}{1-\delta_{3s}} < \mu < \frac{1}{1+\delta_{3s}}$. □

In summary, for the HTP algorithm, the fifth bullet states a result about the exact sparse recovery (and thus convergence in a finite number of iterations) while the seventh and eighth bullet state results about the sparse recovery with sparsity defect and measurement error. For the HTP^{μ} algorithm, the third bullet states the condition for convergence in a finite number of iterations while the ninth bullet states the condition for both convergence in a finite number of iterations and sparse recovery with sparsity defect and measurement error. For the FHTP algorithm, the sixth bullet states a result about the exact sparse recovery (and thus convergence). For the $FHTP^{\mu}$ algorithm, the fourth bullet states the condition for convergence.

Chapter 5

Some Important Issues and Techniques

5.1 The L-Curve Method

In section 2.5, we introduce the Tikhonov regularization least-squares problem and also mention the L-curve. In this section, we will delve deeply into the properties and applications of the L-curve. As far as we have discussed in section 2.5, we can have the following theorem (which is the theorem 1 of [18]).

Theorem 5.1.1. *Let x_λ denote the regularized solution. $\|Lx_\lambda\|$ is a monotonically decreasing function of $\|Ax_\lambda - b\|$, and any point (δ, η) on the L-curve $(\|Ax_\lambda - b\|, \|Lx_\lambda\|)$ is a solution to the following two inequality-constrained least-squares problems :*

$$\delta = \min_x \|Ax - b\| \text{ subject to } \|Lx\| \leq \eta, 0 \leq \eta \leq \|Lx_0\| \tag{5.1}$$

$$\eta = \min_x \|Lx\| \text{ subject to } \|Ax - b\| \leq \delta, \delta_0 \leq \delta \leq \delta_\infty \tag{5.2}$$

According to this theorem, we know that the L-curve divides the first quadrant into two regions, one above the curve and the other below the curve. It is impossible to construct any other regularized solution x_{reg} (obtained by other regularization method, e.g., truncated GSVD method) that lies below the L-curve. Hence, the Tikhonov regularized solution x_λ is optimal in the sense that given δ (or η), there does not exist a solution with a smaller residual norm (or seminorm) than $\|Ax_\lambda - b\|$ (or $\|Lx_\lambda\|$). Therefore, it is reasonable to measure the distance between x_{reg} and x_λ. The smaller the distance (or the smaller the difference of the residual norm and seminorm), the better the x_{reg} in the "Tikhonov" sense. Theorem 3 of [18] states that given any two vectors x_1 and x_2 satisfying $\|Lx_i\| \leq \eta$ and $\|Ax_i - b\| \leq \delta$, $i = 1, 2$,

the following three inequalities are satisfied

$$\|x_1 - x_2\| \leq 2\|X\|_{2\to 2}\sqrt{\delta^2 + \eta^2} \tag{5.3}$$

$$\|L(x_1 - x_2)\| \leq 2\min\{\frac{\delta}{\gamma_1}, \eta\} \tag{5.4}$$

$$\|A(x_1 - x_2)\| \leq 2\min\{\delta, \eta\gamma_p\} \tag{5.5}$$

[1] If we set x_1 and x_2 as x_λ and x_{reg}, this theorem gives an upper bound of $\|x_\lambda - x_{reg}\|$, $\|L(x_\lambda - x_{reg})\|$ and $\|A(x_\lambda - x_{reg})\|$ suppose the seminorm and residual norm of x_λ and x_{reg} are bounded. Besides theorem 1, [18] further presents characterization 2 to delve more deeply into the characteristics of a L-curve. We excerpt it as follows.

Theorem 5.1.2. *Assume b can be written as $b = \bar{b} + e$, where e is the perturbation error. Suppose*

1. *$|y_i^*\bar{b}|$ on average decay to zero faster than γ_i*
2. *e has zero mean and covariance matrix $\sigma_0^2 I_m$*
3. *$\|e\| < \|\bar{b}\|$*

Then the L-curve $(\|Ax_\lambda - b\|, \|Lx_\lambda\|)$ exhibits a "corner" behavior as a function of λ, and the corner appears approximately at $(\sqrt{\sigma_0^2(m-n+p) + \delta_0^2}, \|L\bar{x}_0\|)$. Here $\bar{x}_0$ denotes the unregularized solution to the unperturbed problem (i.e., $e = 0$), and δ_0 is the incompatibility measure. The faster $|y_i^* b|$ decay to zero, the sharper the L-shaped corner. For small λ the behavior of the L-curve is entirely dominated by contributions from e, while for large λ it is completely dominated by those from $\bar{b}$. In between, there is a small region where both $\bar{b}$ and e contribute, and this region defines the L-shaped "corner" of the L-curve.

The first assumption is called the discrete Picard condition, which is necessary to ensure that a reasonable regularized solution exists. The second assumption is to ensure that x_λ has a reasonable covariance matrix. The third assumption is to ensure a reasonable signal-to-noise

[1] Note that $A = Y\Sigma X^{-1}$ and $L = W\Sigma_L X^{-1}$ are described in section 2.5

ratio. It is sensible to choose λ that corresponds to a regularized solution near the corner because it can achieve small residual norm while keeping the seminorm reasonably small. That is, it strikes a good balance between the residual norm and the seminorm.

Finally, we introduce various methods for determining a good regularization parameter so that we can attain a good regularized solution which corresponds to a point near the corner of the L-curve.

1. The discrepancy principle

   Suppose the system is consistent, i.e., $\delta_0 = 0$, we select the regularization parameter λ so that the residual norm is equal to an a priori upper bound δ_e. That is, $\|Ax_\lambda - b\| = \delta_e$, where $\|e\| \leq \delta_e$. If e satisfies the second assumption, then the expected value of $\|e\|$ is $\sqrt{m}\sigma_0$. Since the corner of the L-curve is approximately at $(\sqrt{\sigma_0^2(m-n+p)}, \|L\bar{x}_0\|)$, the chosen regularization parameter corresponds to a point a little to the right of the corner.

   Now we consider the most general case. We will take into account the error E in the matrix A and the incompatibility measure δ_0. Assume $\|e\| \leq \delta_e$ and $\|E\|_{2\to 2} \leq \delta_E$. We select the regularization parameter λ so that

   $$\|Ax_\lambda - b\| = \delta_0 + \delta_e + \Delta_{E,L}\|Lx_\lambda\| \tag{5.6}$$

   where $\Delta_{E,L} \triangleq \max_{Lx\neq 0}\{\|Ex\|/\|Lx\|\}$, the largest generalized singular value of the matrix pair (E, L). Particularly, if $L = I_n$, then $\Delta_{E,L} = \delta_E$. Hence, the chosen parameter corresponds to the point which is the intersection of the L-curve and the line $\|Ax_\lambda - b\| = \delta_0 + \delta_e + \Delta_{E,L}\|Lx_\lambda\|$.

   Overall, the discrepancy principle tends to select the regularized solution appearing to the right of the corner, which means slightly larger λ than the "best" one is chosen. Hence, the discrepancy principle over-regularizes the solution.

2. The quasi-optimality criterion

   We select the regularization parameter λ so that it is a mini-

mizer of the function

$$Q(\lambda) \triangleq \|\lambda^2 \frac{dx_\lambda}{d(\lambda^2)}\| = \frac{1}{2}\|\lambda \frac{dx_\lambda}{d\lambda}\| \tag{5.7}$$

We can derive that

$$\begin{aligned} Q(\lambda)^2 &= \sum_{i=1}^{p} \phi^2 (1-\phi_i)^2 \left(\frac{\beta_i}{\sigma_i}\right)^2 \\ &\approx \sum_{\gamma_i \geq \lambda} (1-\phi_i)^2 \left(\frac{\beta_i}{\sigma_i}\right)^2 + \sum_{\gamma_i < \lambda} \phi_i^2 \left(\frac{\beta_i}{\sigma_i}\right)^2 \\ &\approx \|L(\bar{x}_0 - \bar{x}_\lambda)\|^2 + \|L(\bar{x}_\lambda - x_\lambda)\|^2 \end{aligned}$$

Hence, the quasi-optimality criterion attempts to strike a good balance between the minimization of the regularization error $\bar{x}_0 - \bar{x}_\lambda$ and the perturbation error $\bar{x}_\lambda - x_\lambda$.

3. The generalized cross-validation (GCV) method

We select the regularization parameter λ so that it is a minimizer of the function

$$\mathcal{G}(\lambda) \triangleq \frac{\|Ax_\lambda - b\|^2}{(\mathcal{T}(\lambda))^2} \tag{5.8}$$

where $\mathcal{T}(\lambda) \triangleq trace(I_m - A(A^*A + \lambda^2 L^*L)^{-1}A^*) = m - n + \sum_{i=1}^{p} \frac{\lambda^2}{\gamma_i^2 + \lambda^2}$. It has been verified that $\mathcal{G}(\lambda)$ has a minimum corresponding to a point near the corner. Hence, the GCV method is effective to find a good regularization parameter.

4. The L-curve method

As stated in [19], the L-curve consists of a near vertical part and an adjacent part with smaller slope. The intersection of these two parts is the corner. The more horizontal part corresponds to solutions where the regularization parameter is larger while the more vertical part corresponds to solutions where the regularization parameter is smaller. Based on these characteristics, we can select the regularization parameter using two possible ways.

(a) seek the point on the curve closest to the origin. That is, we can compute $\|Ax_\lambda - b\|_2^2 + \lambda^2\|Lx_\lambda\|_2^2$ and choose λ with the smallest such value.

(b) choose the point on the curve where the curvature is maximum. Let $\rho(\lambda) \triangleq \|Ax_\lambda - b\|_2$ and $\eta(\lambda) \triangleq \|Lx_\lambda\|_2$. We can compute the curvature $\kappa(\lambda)$ by the formula :

$$\kappa(\lambda) = \frac{\rho(\lambda)'\eta(\lambda)'' - \rho(\lambda)''\eta(\lambda)'}{((\rho(\lambda)')^2 + (\eta(\lambda)')^2)^{3/2}} \tag{5.9}$$

where $'$ denotes differentiation with respect to the regularization parameter λ. Then we choose λ with the largest curvature.

Note that we can actually generalize the choices of $\rho(\lambda)$ and $\eta(\lambda)$. $\rho(\lambda)$ represents the main function wished to be minimized, e.g., the residual. $\eta(\lambda)$ represents the norm or function associated with the regularized solution vector x_λ, e.g. the seminorm. Different choices of $\rho(\lambda)$ and $\eta(\lambda)$ correspond to different regularization methods. The L-curve is just a parametric plot of $(\rho(\lambda), \eta(\lambda))$.

If we encounter the functions $\rho(\lambda)$ or $\eta(\lambda)$ that are not differentiable, we cannot directly use the curvature formula. If this is the case, we can only use discrete points corresponding to different values of the regularization parameter at which we have evaluated ρ and η. In [19], the authors proposed to fit a cubic spline curve to these discrete points of the L-curve. The resulting approximating curve is good for being twice differentiable and can be differentiated in a numerically stable way. An algorithm called "FITCURVE" is proposed. It consists of two steps.

(a) Perform a local smoothing of the L-curve points, in which each point is replaced by a new point obtained by fitting a low-degree polynomial to a few neighboring points. This step controls the extent of fine-grained details to be retained.

(b) Use the new smoothed points as control points for a cubic spline curve with knots 1,..., N + 4, where N is the number of L-curve points.

Empirically, fitting a straight line in the least squares sense to five points centered at the point to be smoothed is good. With this " FITCURVE" algorithm, an algorithm called "FINDCORNER" was proposed to compute a sequence of new regularized solutions whose associated points on the L-curve hopefully approaches the corner. The algorithm is described as follows.

(a) Start with a few points (ρ_i, η_i) on each side of the corner.

(b) Use the "FITCURVE" algorithm to find an approximating three-dimensional cubic spline curve S for the points $(\rho_i, \eta_i, \lambda_i)$, where λ_i is the regularization parameter that corresponds to (ρ_i, η_i).

(c) Let S_2 denote the first two coordinates of S, such that S_2 approximates the L-curve.

(d) Compute the point on S_2 with maximum curvature, and find the corresponding λ_0 from the third coordinate of S.

(e) Solve the regularization problem for λ_0 and add the new point $(\rho(\lambda_0), \eta(\lambda_0))$ to the L-curve.

(f) Repeat from step 2 until convergence.

Note that it is suggested that in step 1, the initial points can be generated by choosing very large and very small regularization parameters , for instance, λ equal to σ_1, $\frac{1}{10}\sigma_1$, $10\sigma_p$ and σ_p. In this way, we can have points that lie on each side of the corner. In step 2, it is necessary to introduce λ_i as a third coordinate of S so that we can associate a regularization parameter with computed point on S_2 in step 4.

Although we have these numerically stable methods, it is good to plot the L-curve so that we can have a basic understanding of the behavior of the problem and where the corner may locate. It is suggested that plotting the L-curve in a log-log scale is advantageous because the log-log scale can emphasize the corner. Furthermore, the L-curves for pure signal and pure noise are both steep in lin-lin scale as $\lambda \to 0$ while only the noise curve is steep in log-log scale. The following curve is an example of L-curve for Tikhonov regularization in [19]. The numbers

are the regularization parameters that correspond to the points on the L-curve.

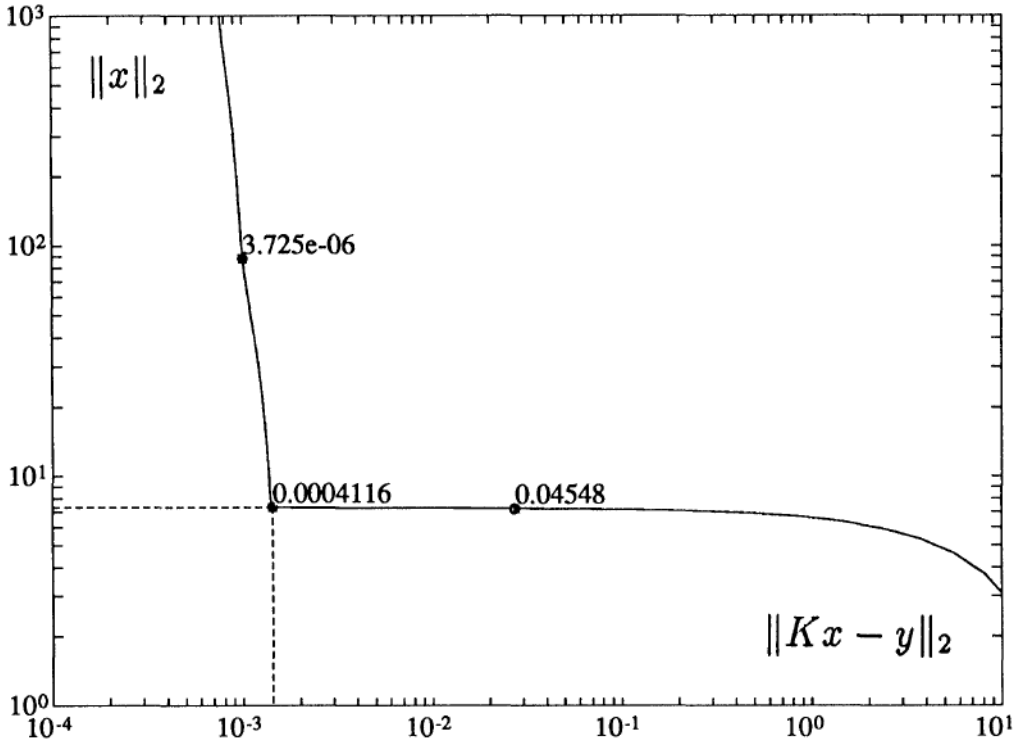

Figure 5.1: An example of L-curve

5.2 Model Order Selection

In a parametric (or model-based) approach, we assume the vector of available data $Y = (y_1, y_2, \cdots, y_N) \in \mathbb{R}^N$ has entries drawn identically and independently distributed (i.i.d.) from a model. The probability density function (pdf) of the model depends on some parameter vector. Assume there are $\bar{n}$ possible models with pdf $p_n(y|\theta_0^n)$, $n = 1, 2, \cdots, \bar{n}$. $\theta_0^n \in \mathbb{R}^n$ is the model parameter which is unknown and has size n. n is called the model order. An order selection rule is a method that determines which model most fits the data vector Y(hence, also determines the model order n). This is as important as the task of estimating the parameter vector θ_0^n from the data vector Y, which is often related to the maximum likelihood estimation (MLE) or the EM algorithm introduced in section 3.5. We concentrate on order selection rules that are associated with the maximum likelihood

method of parameter estimation.

Let the maximum likelihood estimator of θ_0^n be $\hat{\theta}^n \triangleq \underset{\theta^n}{argsup} \ln p_n(Y|\theta^n)$, where $p_n(Y|\theta^n) := \prod_{i=1}^{N} p_n(y_i|\theta^n)$. As $N \to \infty$, the pdf of $\hat{\theta}^n$ converges to the Gaussian pdf with mean θ_0^n and covariance equal to the reciprocal of the total expected information matrix (or called the Fisher information matrix or the Cramér-Rao bound matrix) defined as follows

Definition 5.2.1 (unit observed information matrix)**.**

$$\hat{i}(\theta^n) := -\frac{\partial^2 \ln p_n(y|\theta^n)}{\partial \theta^n \partial (\theta^n)^T} \tag{5.10}$$

Definition 5.2.2 (unit expected information matrix)**.**

$$\begin{aligned} i(\theta^n) &:= \int \hat{i}(\theta^n) p_n(y|\theta_0^n) dy \\ &:= \mathbb{E}_{y|\theta_0^n} \left[-\frac{\partial^2 \ln p_n(y|\theta^n)}{\partial \theta^n \partial (\theta^n)^T} \right] \end{aligned} \tag{5.11}$$

Definition 5.2.3 (total observed information matrix)**.**

$$\hat{J}^n(\theta^n) := N\hat{i}(\theta^n) \tag{5.12}$$

Definition 5.2.4 (total expected information matrix)**.**

$$J^n(\theta^n) := Ni(\theta^n) \tag{5.13}$$

We call such convergence phenomenon as the asymptotic normality of MLE and we denote it as $\hat{\theta}^n \xrightarrow{D} N(\theta_0^n, (J^n(\theta_0^n))^{-1})$. By the weak law of

large number, we can derive that

$$
\begin{aligned}
-\frac{1}{N}\frac{\partial^2 \ln p_n(Y|\theta^n)}{\partial\theta^n\partial(\theta^n)^T}|_{\theta^n=\theta_0^n} &= -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \ln p_n(y_i|\theta^n)}{\partial\theta^n\partial(\theta^n)^T}|_{\theta^n=\theta_0^n} \\
&\xrightarrow{p} \mathbb{E}_{y|\theta_0^n}[-\frac{\partial^2 \ln p_n(y|\theta^n)}{\partial\theta^n\partial(\theta^n)^T}|_{\theta^n=\theta_0^n}] \\
&= \frac{1}{N}J^n(\theta_0^n)
\end{aligned}
\tag{5.14}
$$

Furthermore, since MLE is consistent (i.e., $\hat{\theta}^n \xrightarrow{p} \theta_0^n$),

$$
\begin{aligned}
\frac{1}{N}\hat{J}^n(\hat{\theta}^n) &= -\frac{1}{N}\frac{\partial^2 \ln p_n(Y|\theta^n)}{\partial\theta^n\partial(\theta^n)^T}|_{\theta^n=\hat{\theta}^n} \\
&\xrightarrow{p} -\frac{1}{N}\frac{\partial^2 \ln p_n(Y|\theta^n)}{\partial\theta^n\partial(\theta^n)^T}|_{\theta^n=\theta_0^n} \\
&\to \frac{1}{N}J^n(\theta_0^n)
\end{aligned}
\tag{5.15}
$$

In the following discussion, we assume that $\frac{1}{N}J^n(\theta_0^n) = \mathcal{O}(1)$, where $\mathcal{O}$ denotes the big-O notation.

Let $p_0(Y)$ denote the true pdf of Y. We want to measure the discrepancy between $p_0(Y)$ and the pdf of each model $p_n(Y|\theta_0^n)$ using the KL divergence $D(p_0, p_n) := \int p_0(Y)\ln[\frac{p_0(Y)}{p_n(Y|\theta_0^n)}]dY := \mathbb{E}_Y[\ln[\frac{p_0(Y)}{p_n(Y|\theta_0^n)}]] = \mathbb{E}_Y[\ln p_0(Y)] - \mathbb{E}_Y[\ln p_n(Y|\theta_0^n)]$. The KL divergence can be viewed as showing the loss of information induced by the use of the model pdf $p_n(Y|\theta_0^n)$ in lieu of the true pdf $p_0(Y)$. Hence, the KL divergence is sometimes called the information function, and the order selection rules derived from it are called information criteria. Our aim is to minimize $D(p_0, p_n)$ with respect to the model order n; that is, to maximize the function $I(p_0, p_n) := \mathbb{E}_Y[\ln p_n(Y|\theta_0^n)]$, which is sometimes called the relative KL information. However, because the model parameter θ_0^n and the true pdf p_0 are unknown, we cannot directly calculate the expectation. In the following, we will introduce four approaches to deal with these two obstacles, which are the naive approach, the no-name rule, the Akaike information criterion (AIC) and the general informa-

tion criterion (GIC). The readers can refer to [35] for materials we are going to introduce below.

5.2.1 The Naive Approach

The first obstacle is that we do not know the model parameter θ_0^n. A naive approach is to replace $\ln p_n(Y|\theta_0^n)$ with $\ln p_n(Y|\hat{\theta}^n)$; hence, we replace $I(p_0, p_n)$ with $I(p_0, p_n(Y|\hat{\theta}^n))$. The second obstacle is that we do not know the true pdf p_0. A naive approach is thus to replace $I(p_0, p_n(Y|\hat{\theta}^n))$ with an unbiased estimate $\ln p_n(Y|\hat{\theta}^n)$ of it. Therefore, we come up with a totally naive approach, which is just to maximize $\ln p_n(Y|\hat{\theta}^n)$. However, this approach is not feasible. Since a model with more parameters reasonably has stronger power to fit the data, the likelihood of the data will monotonically increase with increasing model order. Hence, the order selection rule always chooses the largest model order $\bar{n}$.

5.2.2 The No-Name Rule

For the first obstacle, we approximate $\ln p_n(Y|\theta_0^n)$ by the Taylor series expansion as follows.

$$\begin{aligned}
&\ln p_n(Y|\theta_0^n) \\
\approx &\ln p_n(Y|\hat{\theta}^n) + (\theta_0^n - \hat{\theta}^n)^T [\frac{\partial \ln p_n(Y|\theta^n)}{\partial \theta^n}|_{\theta^n=\hat{\theta}^n}] \\
&+ \frac{1}{2}(\theta_0^n - \hat{\theta}^n)^T [\frac{\partial^2 \ln p_n(Y|\theta^n)}{\partial \theta^n \partial (\theta^n)^T}|_{\theta^n=\hat{\theta}^n}](\theta_0^n - \hat{\theta}^n) \\
= &\ln p_n(Y|\hat{\theta}^n) + \frac{1}{2}(\theta_0^n - \hat{\theta}^n)^T [\frac{\partial^2 \ln p_n(Y|\theta^n)}{\partial \theta^n \partial (\theta^n)^T}|_{\theta^n=\hat{\theta}^n}](\theta_0^n - \hat{\theta}^n) \\
\xrightarrow{p} &\ln p_n(Y|\hat{\theta}^n) - \frac{1}{2}(\theta_0^n - \hat{\theta}^n)^T J^n(\theta_0^n)(\theta_0^n - \hat{\theta}^n)
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}_Y[\ln p_n(Y|\theta_0^n)] &\approx \mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n)] - \frac{1}{2}\mathbb{E}_Y[(\theta_0^n - \hat{\theta^n})^T J^n(\theta_0^n)(\theta_0^n - \hat{\theta}^n)] \\
&= \mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n)] - \frac{1}{2}tr(\mathbb{E}_Y[(\theta_0^n - \hat{\theta^n})^T J^n(\theta_0^n)(\theta_0^n - \hat{\theta}^n)]) \\
&= \mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n)] - \frac{1}{2}\mathbb{E}_Y[tr((\theta_0^n - \hat{\theta^n})^T J^n(\theta_0^n)(\theta_0^n - \hat{\theta}^n))] \\
&= \mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n)] - \frac{1}{2}\mathbb{E}_Y[tr(J^n(\theta_0^n)(\theta_0^n - \hat{\theta}^n)(\theta_0^n - \hat{\theta^n})^T)] \\
&= \mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n)] - \frac{1}{2}tr(J^n(\theta_0^n)\mathbb{E}_Y[(\theta_0^n - \hat{\theta}^n)(\theta_0^n - \hat{\theta^n})^T])
\end{aligned}$$

If p_0 is a special case of p_n or $p_0 = p_n$, then $\mathbb{E}_Y[(\theta_0^n - \hat{\theta}^n)(\theta_0^n - \hat{\theta}^n)^T] = \mathbb{E}_{y|\theta_0^n}[(\theta_0^n - \hat{\theta}^n)(\theta_0^n - \hat{\theta}^n)^T] = (J^n(\theta_0^n))^{-1}$. In this way,

$$\begin{aligned}
\mathbb{E}_Y[\ln p_n(Y|\theta_0^n)] &= \mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n)] - \frac{1}{2}tr(I_n) \\
&= \mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n)] - \frac{1}{2}n
\end{aligned}$$

Even if p_n is just a good approximating model of p_0, it is widely accepted that the best estimator of $tr(J^n(\theta_0^n)\mathbb{E}_Y[(\theta_0^n - \hat{\theta}^n)(\theta_0^n - \hat{\theta^n})^T])$ is n. If p_n is really a poor approximating model of p_0, then $\ln p_n(Y|\hat{\theta}^n)$ will also be poor. Hence, we also do not choose this model. For the second obstacle, we will use an unbiased estimator of $\mathbb{E}_Y[\ln p_n(Y|\hat{\theta}^n) - \frac{1}{2}(\theta_0^n - \hat{\theta}^n)^T J^n(\theta_0^n)(\theta_0^n - \hat{\theta}^n)]$, which is $\ln p_n(Y|\hat{\theta}^n) - \frac{1}{2}n$. As a result, we come up with an order selection rule that chooses the model with the minimum value of

$$NN(n) = -2\ln p_n(Y|\hat{\theta}^n) + n \tag{5.16}$$

over $\bar{n}$ possible models. Although, compared to the naive approach, the no name rule has an additional penalty term n, it turns out that it does not penalize enough so that it tends to overfit (that is, selects a larger model than the true generating model).

5.2.3 The Akaike Information Criterion (AIC)

Let X denote a fictitious data vector with the same size N and the same pdf as Y but which is independent of Y. Let $\hat{\theta}_x^n = \underset{\theta^n}{argsup}\, p_n(X|\theta^n)$ and $\hat{\theta}_y^n = \underset{\theta^n}{argsup}\, p_n(Y|\theta^n)$. For the first obstacle, we replace the model pdf $\ln p_n(Y|\theta_0^n)$ with $\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_x^n)]$. That is to say, we want to maximize $\mathbb{E}_Y[\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_x^n)]]$. It has an interesting cross-validation interpretation : we use the samples X for estimation and the independent samples Y for validation of the resulting model's pdf. We approximate $\ln p_n(Y|\hat{\theta}_x^n)$ using the Taylor series expansion as follows.

$$\begin{aligned}
&\ln p_n(Y|\hat{\theta}_x^n) \\
\approx\ &\ln p_n(Y|\hat{\theta}_y^n) + (\hat{\theta}_x^n - \hat{\theta}_y^n)^T[\frac{\partial \ln p_n(Y|\theta^n)}{\partial \theta^n}|_{\theta^n=\hat{\theta}_y^n}] \\
&+ \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T[\frac{\partial^2 \ln p_n(Y|\theta^n)}{\partial \theta^n (\partial \theta^n)^T}|_{\theta^n=\hat{\theta}_y^n}](\hat{\theta}_x^n - \hat{\theta}_y^n) \\
\xrightarrow{p}\ &\ln p_n(Y|\hat{\theta}_y^n) - \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}_Y[\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_x^n)]] \\
\approx &\mathbb{E}_Y[\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_y^n) - \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)]]
\end{aligned}$$

We also assume p_0 is a special case of p_n, $p_0 = p_n$ or p_n is at least a good approximating model of p_0. In this way,

$$\begin{aligned}
&\mathbb{E}_Y[\mathbb{E}_X[\frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)]] \\
=&\frac{1}{2}\mathbb{E}_Y[\mathbb{E}_X[tr(J^n(\theta_0^n)[(\hat{\theta}_x^n - \theta_0^n) - (\hat{\theta}_y^n - \theta_0^n)][(\hat{\theta}_x^n - \theta_0^n) - (\hat{\theta}_y^n - \theta_0^n)]^T)]] \\
=&\frac{1}{2}tr(J^n(\theta_0^n)((J^n(\theta_0^n))^{-1} + (J^n(\theta_0^n))^{-1})) \\
=&\frac{1}{2}tr(2I_n) \\
=&n
\end{aligned}$$

For the second obstacle, we will use an unbiased estimator of $\mathbb{E}_Y[\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_y^n) - \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)]]$, which is $\ln p_n(Y|\hat{\theta}^n) - n$. As a result, we come up with an order selection rule that chooses the model with the minimum value of

$$AIC(n) = -2\ln p_n(Y|\hat{\theta}^n) + 2n \tag{5.17}$$

over $\bar{n}$ possible models. We make a remark that in [21] a corrected AIC rule, AIC_c is suggested

$$AIC_c = -2\ln p_n(Y|\hat{\theta}^n) + \frac{2N}{N-n-1}n \tag{5.18}$$

As N approaches ∞, AIC_c will approach AIC ; for finite values of N, the penalty term of AIC_c is larger than that of AIC so that AIC_c performs better than AIC with smaller risk of overfitting.

Besides the AIC order selection rule, we also want to introduce a useful quantity called the AIC difference which is defined below.

Definition 5.2.5 (AIC difference)**.**

$$\Delta_i = AIC_i - AIC_{min} \tag{5.19}$$

where AIC_i is the AIC of the ith model, and AIC_{min} is the lowest AIC value one obtains among all the candidate models.

From the definition, we know that the best candidate model has the AIC difference $\Delta_{min} = 0$. Note that it is not the absolute value of the AIC value that matters but the relative value, that is, the AIC difference that matters. A useful rule of thumb is listed as follows

- $0 \leq \Delta_i < 2$: there is substantial support for the ith model
- $2 \leq \Delta_i < 4$: there is strong support for the ith model
- $4 \leq \Delta_i < 7$: there is considerably less support for the ith model
- $\Delta_i > 10$: there is essentially no support for the ith model

It is reasonable to think of $\exp(-\frac{1}{2}\Delta_i)$ as the relative likelihood of the ith model. We can normalize it and come up with the Akaike weight of the ith model defined as follows

Definition 5.2.6 (Akaike weight)**.**

$$w_i \triangleq \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{\bar{n}} \exp(-\frac{1}{2}\Delta_r)} \tag{5.20}$$

It is considered as the weight of evidence in favor of model i being the best model or simply can be considered as the model probability. We can also compute the evidence ratio between the ith model and the jth model defined as w_i/w_j, which is the ratio of the Akaike weight of the ith model and the jth model. In particular, we are interested in w_1/w_j, where we label the best candidate model as the first model. Since $\Delta_1 = \Delta_{min} = 0$, $w_1/w_j = \exp(\frac{1}{2}\Delta_j)$. We list some values of δ_j and the corresponding evidence ratio w_1/w_j below.

- $\Delta_j = 2$: the evidence ratio is 2.7
- $\Delta_j = 4$: the evidence ratio is 7.4
- $\Delta_j = 8$: the evidence ratio is 54.6
- $\Delta_j = 10$: the evidence ratio is 148.4

Thus, the concept of evidence ratio somehow justifies the rule of thumb. The readers can refer to [1] for materials regarding the AIC difference, the Akaike weight and the evidence ratio.

5.2.4 The General Information Criterion (GIC)

GIC is a generalization of AIC. Both order selection rules adopt the notion of cross-validation. However, GIC uses a validation data vector Y longer than an estimation data vector X since the risk of overfitting will decrease in this way. We assume Y is ρ times the length of X, where $\rho \geq 1$. We also approximate $\ln p_n(Y|\hat{\theta}_x^n)$ using the Taylor series

expansion as follows

$$\begin{aligned}
&\ln p_n(Y|\hat{\theta}_x^n) \\
\approx\ &\ln p_n(Y|\hat{\theta}_y^n) + (\hat{\theta}_x^n - \hat{\theta}_y^n)^T [\frac{\partial \ln p_n(Y|\theta^n)}{\partial \theta^n}|_{\theta^n=\hat{\theta}_y^n}] \\
&+ \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T [\frac{\partial^2 \ln p_n(Y|\theta^n)}{\partial \theta^n (\partial \theta^n)^T}|_{\theta^n=\hat{\theta}_y^n}](\hat{\theta}_x^n - \hat{\theta}_y^n) \\
\xrightarrow{p}\ &\ln p_n(Y|\hat{\theta}_y^n) - \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J_y^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)
\end{aligned}$$

We add a subscript y to the total expected information matrix to highlight that we take expectation over the random vector Y. Hence,

$$\begin{aligned}
&\mathbb{E}_Y[\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_x^n)]] \\
\approx &\mathbb{E}_Y[\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_y^n) - \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J_y^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)]]
\end{aligned}$$

We also assume p_0 is a special case of p_n, $p_0 = p_n$ or p_n is at least a good approximating model of p_0. In this way,

$$\begin{aligned}
&\mathbb{E}_Y[\mathbb{E}_X[\frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J_y^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)]] \\
=&\frac{1}{2}\mathbb{E}_Y[\mathbb{E}_X[tr(J_y^n(\theta_0^n)[(\hat{\theta}_x^n - \theta_0^n) - (\hat{\theta}_y^n - \theta_0^n)][(\hat{\theta}_x^n - \theta_0^n) - (\hat{\theta}_y^n - \theta_0^n)]^T)]] \\
=&\frac{1}{2}tr(J_y^n(\theta_0^n)((J_x^n(\theta_0^n))^{-1} + (J_y^n(\theta_0^n))^{-1}))
\end{aligned}$$

From the definition of the total expected information matrix, we know that $J_y^n(\theta_0^n) = \rho J_x^n(\theta_0^n)$. Therefore,

$$\begin{aligned}
&\mathbb{E}_Y[\mathbb{E}_X[\frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J_y^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)]] \\
=&\frac{1}{2}tr(J_y^n(\theta_0^n)(\rho(J_y^n(\theta_0^n))^{-1} + (J_y^n(\theta_0^n))^{-1})) \\
=&\frac{1+\rho}{2}n
\end{aligned}$$

For the second obstacle, we will use an unbiased estimator of $\mathbb{E}_Y[\mathbb{E}_X[\ln p_n(Y|\hat{\theta}_y^n) - \frac{1}{2}(\hat{\theta}_x^n - \hat{\theta}_y^n)^T J_y^n(\theta_0^n)(\hat{\theta}_x^n - \hat{\theta}_y^n)]]$, which is $\ln p_n(Y|\hat{\theta}^n) - \frac{1+\rho}{2}n$. As a result, we come up with an order selection rule that

chooses the model with the minimum value of

$$GIC(n) = -2\ln p_n(Y|\hat{\theta}^n) + (1+\rho)n \tag{5.21}$$

over $\bar{n}$ possible models. GIC has smaller risk of overfitting than that of AIC. Values of ρ in the interval $[1, 5]$ is suggested to perform well.

So far, the four approaches introduced are derived based on minimizing the KL divergence. In the following we will introduce an approach that is based on the Bayesian setting. The resulting model order selection rule is called the Bayesian information criterion (BIC) rule.

5.2.5 The Bayesian Information Criterion (BIC)

In the Bayesian setting, the model parameter is considered to be random but depends on different models. Let the prior probability of the model parameter be $p_n(\theta_0^n|M_n)$, where M_n represents the model. We make three assumptions of $p_n(\theta_0^n|M_n)$ as follows

1. $p_n(\theta_0^n|M_n)$ is approximately constant around $\hat{\theta}^n$, where $\hat{\theta}^n = \underset{\theta^n}{argsup}\, p_n(Y|\theta^n, M_n)$
2. $p_n(\theta_0^n|M_n)$ is independent of N
3. $p_n(\hat{\theta}^n|M_n) \gg p_n(\theta_0^n|M_n)$ for θ_0^n outside the neighborhood of $\hat{\theta}^n$

Our aim is to maximize the model evidence, which is $p_n(Y|M_n) = \int p_n(Y|\theta_0^n, M_n)p_n(\theta_0^n|M_n)d\theta_0^n$. We approximate $\ln p_n(Y|\theta_0^n, M_n)$ using the Taylor series expansion as follows and get an approximation of $p_n(Y|\theta_0^n, M_n)$.

$$\ln p_n(Y|\theta_0^n, M_n) \approx \ln p_n(Y|\hat{\theta}^n, M_n) - \frac{1}{2}(\theta_0^n - \hat{\theta}^n)^T \hat{J}^n(\hat{\theta}^n)(\theta_0^n - \hat{\theta}^n)$$

$$\Rightarrow p_n(Y|\theta_0^n, M_n) \approx p_n(Y|\hat{\theta}^n, M_n)e^{-\frac{1}{2}(\theta_0^n - \hat{\theta}^n)^T \hat{J}^n(\hat{\theta}^n)(\theta_0^n - \hat{\theta}^n)}$$

Hence,

$$
\begin{aligned}
& p_n(Y|M_n) \\
& \approx p_n(Y|\hat{\theta}^n, M_n) \int e^{-\frac{1}{2}(\theta_0^n-\hat{\theta}^n)^T \hat{J}^n(\hat{\theta}^n)(\theta_0^n-\hat{\theta}^n)} p_n(\theta_0^n|M_n) d\theta_0^n \\
& \approx p_n(Y|\hat{\theta}^n, M_n) p_n(\hat{\theta}^n|M_n) \int_{\theta_0^n \, near \, \hat{\theta}^n} e^{-\frac{1}{2}(\theta_0^n-\hat{\theta}^n)^T \hat{J}^n(\hat{\theta}^n)(\theta_0^n-\hat{\theta}^n)} d\theta_0^n \\
& \approx \frac{p_n(Y|\hat{\theta}^n, M_n) p_n(\hat{\theta}^n|M_n)(2\pi)^{n/2}}{|\hat{J}^n(\hat{\theta}^n)|^{1/2}} \int_{\theta_0^n \, near \, \hat{\theta}^n} \frac{e^{-\frac{1}{2}(\theta_0^n-\hat{\theta}^n)^T \hat{J}^n(\hat{\theta}^n)(\theta_0^n-\hat{\theta}^n)}}{(2\pi)^{n/2}|\hat{J}^n(\hat{\theta}^n)|^{-1/2}} d\theta_0^n \\
& \approx \frac{p_n(Y|\hat{\theta}^n, M_n) p_n(\hat{\theta}^n|M_n)(2\pi)^{n/2}}{|\hat{J}^n(\hat{\theta}^n)|^{1/2}}
\end{aligned}
$$

The second approximation holds because of the second and the third assumptions. The last approximation holds since the exponential decays very fast away from $\hat{\theta}^n$ (the integral is roughly equal to 1). Maximizing the model evidence over $\bar{n}$ possible models is equivalent to maximize $\ln p_n(Y|\hat{\theta}^n, M_n) + \ln p_n(\hat{\theta}^n|M_n) + \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\hat{J}^n(\hat{\theta}^n)|$. For the last term,

$$
\begin{aligned}
\ln|\hat{J}^n(\hat{\theta}^n)| &= \ln|N \cdot \frac{1}{N}\hat{J}^n(\hat{\theta}^n)| \\
&= \ln N^n|\frac{1}{N}\hat{J}^n(\hat{\theta}^n)| \\
&= n\ln N + \ln|\frac{1}{N}\hat{J}^n(\hat{\theta}^n)| \\
&= n\ln N + \mathcal{O}(1)
\end{aligned}
$$

Since $p_n(\theta_0^n|M_n)$ is independent of N, $\ln p_n(Y|\hat{\theta}^n, M_n) + \ln p_n(\hat{\theta}^n|M_n) + \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\hat{J}^n(\hat{\theta}^n)| \approx \ln p_n(Y|\hat{\theta}^n, M_n) - \frac{n}{2}\ln N$. As a result, we come up with an order selection rule that chooses the model with the minimum value of

$$
BIC(n) = -2\ln p_n(Y|\hat{\theta}^n) + n\ln N \tag{5.22}
$$

over $\bar{n}$ possible models. Note that if $p_n(\theta_0^n|M_n)$ is dependent of N, then we cannot eliminate the $\ln p_n(\hat{\theta}^n|M_n)$ term when performing the

maximization, which results in a prior-dependent rule. Also note that the assumption that $\frac{1}{N}\hat{J}^n(\hat{\theta}^n) = \mathcal{O}(1)$ does not always hold true. For some models, a different normalization of $\hat{J}(\theta^n)$ is required to get a constant matrix as N approaches the infinity.

As a summary, all the model order selection rules have a common form, which is $-2\ln p_n(Y|\hat{\theta}^n) + \eta(n, N)n$.

1. The naive approach : $\eta(n, N) = 0$
2. The no-name rule : $\eta(n, N) = 1$
3. $AIC : \eta(n, N) = 2$
4. $AIC_c : \eta(n, N) = 2\frac{N}{N-n-1}$
5. $GIC : \eta(n, N) = \rho + 1,\ \rho \geq 1$
6. $BIC : \eta(n, N) = \ln N$

Order selection rules with smaller penalty term tend to choose larger model, which may results in overfitting. Hence, putting aside the first two poor order selection rules (i.e., the naive approach and the no-name rule), a crude ranking of the other four model selection rules may be : $BIC > GIC > AIC_c > AIC$ (note that AIC_c performs better than AIC in small samples whereas in medium or large samples the two selection rules perform almost the same). However, if the true model is more complex than all the candidate models, then AIC performs better instead for its tendency to select relatively larger models.

5.3 Experiments and Performance Evaluations for The Compressive Sensing Problem

In chapter four, we introduce the compressive sensing problem and various reconstruction algorithms to solve it. In this section, we will introduce standard experiment procedures, established in [6], to test the performance of a specific reconstruction algorithm. We described as follows.

1. Choose the dimension m and n of the sampling matrix A and a signal sparsity level s such that $m \geq 2s$
2. Randomly generate an $m \times n$ random matrix A (It may be a Gaussian random matrix, a Bernoulli random matrix or a random sampling matrix).
3. Select a support set T of size $|T| = s$ uniformly at random, and generate the sparse signal vector x by one of the following methods
   (a) A Gaussian signal : draw the elements of x restricted to T from the standard Gaussian distribution
   (b) A zero-one signal : set all entries of x supported on T to ones[2]
   (c) A sparse signal with power-law decaying entries (also known as compressible sparse signal) : the non-increasing rearrangement x^* of x satisfies $x^*[i] \leq Gi^{-1/r}$ for some $G > 0$, $r \in [0, 1]$ and $i \in [s]$
   (d) A sparse signal with exponentially decaying entries : the non-increasing rearrangement x^* of x satisfies $x^*[i] \leq Ge^{-pi}$ for some $G > 0$, $p > 0$ and $i \in [s]$
4. For the case when the signal is sparse and there are no measurement errors : compute $y = Ax$
   For the case when the signal is approximately sparse or there are measurement errors :
   (a) An approximately sparse signal : the signal entries in T remain unchanged but the signal entries outside of T are perturbed by i.i.d. Gaussian $\mathcal{N}(0, \sigma_s^2)$ samples. compute $y = Ax$
   (b) A measurement vector corrupted with errors : the error vector e is generated using a Gaussian distribution $\mathcal{N}(0, \sigma_e^2 I_m)$. compute $y = Ax + e$
5. Apply a reconstruction algorithm to obtain $x^{(\bar{k})}$, the output of the algorithm, and compute $x^{(\bar{k})}$ to x

[2]It is claimed that zero-one sparse signals are a particularly challenging case for OMP-type algorithms.

(a) For the case when the signal is sparse and there is no measurement errors, the algorithms are expected to provide exact reconstruction. Hence, we calculate the empirical frequency of exact reconstruction, i.e., the fraction of exactly recovered test signals. Note that the sparsity level at which the recovery rate drops below 100% is of particular interest. Such sparsity level is called the critical sparsity. If the sparsity level of the signal exceeds the critical sparsity, the reconstruction algorithm cannot exactly recover all sparse test signals.

(b) For the case when the signal is approximately sparse or there are measurement errors, we compute the reconstruction distortion $\|x - x^{(\bar{k})}\|_2$.

6. Repeat the process 500 times for each s , and then simulate the same algorithm for different values of m, n, σ_s and σ_e. Note that the performance of the algorithm is better when the empirical frequency of exact reconstruction is larger, the critical sparsity is larger and the reconstruction distortion is smaller.

In [9], a special performance evaluation method called the phase transition analysis is described. We consider the phase diagram, which is a two-dimensional graph with y-coordinates being the relative sparsity of x (number of non-zeros in x/number of rows in A, i.e., $\rho \triangleq s/m$) and x-coordinates being the indeterminacy of the system $y = Ax + e$ (number of rows in A/number of columns in A, i.e., $\delta \triangleq m/n$). ρ and δ are within $[0, 1]$. Hence, the phase diagram occupies the unit square. The value of each point at the phase diagram can be various performance evaluation metrics, e.g., the fraction of exact reconstruction, the averaged relative reconstruction errors or the fraction of reconstruction with errors within 10^{-4}. There are two phases in the phase diagram, the success phase and the failure phase. For small ρ, it is more possible to have high-accuracy reconstruction while for large ρ, reconstruction may fail more easily. The phase transition from success to failure occurs at different ρ for different values of δ. Asymptotically for large m, the transition gets perfectly sharp since for each δ, if ρ is smaller

than the threshold $\rho_{tr}(\delta)$, the probability of perfect reconstruction approaches 1 while it approaches 0 if $\rho > \rho_{tr}(\delta)$. The following figure is a phase diagram depicted in [9]. The variable k corresponds to our s, n

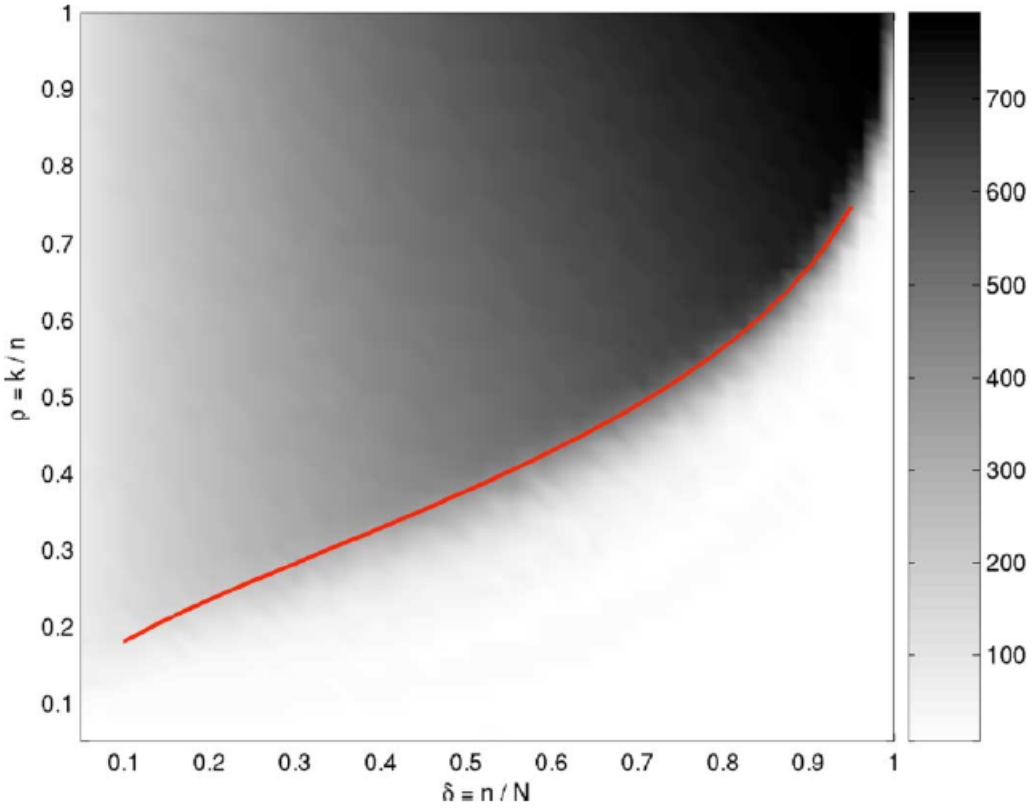

Figure 5.2: phase diagram

corresponds to our m and N corresponds to our n. The value of each point is the number of coordinates of reconstruction which differ from optimally sparse solution by more than 10^{-4}. The red curve is the theoretical phase transition boundary curve (i.e., the curve of threshold $\rho_{tr}(\delta)$). The phase diagram displays a sharp transition from perfect recovery (the region below the red curve) to perfect failure (the region above the red curve).

5.4 Dictionary Screening

In [41], two specific problem formulations are considered. They are the basis pursuit denoising and the non-negative basis pursuit denoising. The former has been introduced as 4.21[3] and the latter further

[3]In our tutorial, it is named as the basis pursuit denoising problem while in [41] it is called the LASSO problem.

constrains the variable $x \in \mathbb{R}^n$ to have non-negative entries; that is,

$$\inf_{x \in \mathbb{R}^n} \lambda \|x\|_1 + \frac{1}{2}\|y - Ax\|_2^2 \quad \textit{subject to } x \succeq 0 \tag{5.23}$$

To derive various screening tests, it is useful to consider the Lagrangian dual problems of these two primal problems. We make some derivations as follows. Let $z = y - Ax$. The basis pursuit denoising problem becomes

$$\min_{z \in \mathbb{R}^m, x \in \mathbb{R}^n} \frac{1}{2}\|z\|_2^2 + \lambda \|x\|_1 \quad \textit{subject to } z = y - Ax$$

The Lagrangian function would be

$$\mathcal{L}(z, x, \mu) \triangleq \frac{1}{2}\|z\|_2^2 + \lambda \|x\|_1 + \mu^T (y - Ax)$$

Setting the partial derivative of $\mathcal{L}$ with respect to z to be zero, we can get

$$\hat{z} = \mu$$

where $\hat{z}$ is a primal optimal point of z. As for a primal optimal point of x, denoted as $\hat{x}$, we make a discussion about three different cases. If $\hat{x}[i] > 0$, $i \in [n]$, we would get $\mu^T a_i = \lambda$ by setting the partial derivative of $\mathcal{L}$ with respect to $\hat{x}[i]$ to be zero. Similarly if $\hat{x}[i] < 0$, we would get $\mu^T a_i = -\lambda$ and if $\hat{x}[i] = 0$ we would get $|\mu^T a_i| \leq \lambda$. Thus, combining these three cases, we know that

$$|\mu^T a_i| \leq \lambda, i \in [n]$$

With these results, we can get the Lagrange dual function

$$g(\mu) \triangleq \mathcal{L}(\hat{z}, \hat{x}, \mu) = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|\mu - y\|_2^2$$

Hence, finally, we can come up with the Lagrange dual problem of the basis pursuit denoising problem as

$$\begin{aligned}&\max_{\mu\in\mathbb{R}^m} \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|\mu - y\|_2^2 \quad subject\ to\ |\mu^T a_i| \leq \lambda, i \in [n] \\ =&\max_{\theta\in\mathbb{R}^m} \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}\|\theta - y/\lambda\|_2^2 \quad subject\ to\ |\theta^T a_i| \leq 1, i \in [n]\end{aligned} \tag{5.24}$$

Furthermore, a primal optimal solution $\hat{x} \in \mathbb{R}^n$ and a dual optimal solution $\hat{\theta} \in \mathbb{R}^m$ satisfy

$$y = A\hat{x} + \lambda\hat{\theta} \tag{5.25}$$

where

$$\hat{\theta}^T a_i = \begin{cases} sgn(\hat{x}[i]) & if\ \hat{x}[i] \neq 0 \\ \gamma \in [-1, 1] & if\ \hat{x}[i] = 0 \end{cases} \tag{5.26}$$

Similar to the derivation of the dual problem of the basis pursuit denoising problem, we derive the dual problem of the non-negative basis pursuit denoising problem in the following. Let $z = y - Ax$. The non-negative basis pursuit denoising problem becomes

$$\min_{z\in\mathbb{R}^m, x\in\mathbb{R}^n} \frac{1}{2}\|z\|_2^2 + \lambda\|x\|_1 \quad subject\ to\ z = y - Ax\ and\ x \succeq 0$$

The Lagrangian function would be

$$\mathcal{L}(z, x, \mu, \phi) \triangleq \frac{1}{2}\|z\|_2^2 + \lambda\|x\|_1 + \mu^T(y - Ax - z) - \phi^T x$$

Setting the partial derivative of $\mathcal{L}$ with respect to z to be zero, we can get

$$\hat{z} = \mu$$

where $\hat{z}$ is a primal optimal point of z. As for a primal optimal point of x, denoted as $\hat{x}$, we make a discussion about three different cases. If $\hat{x}[i] > 0$, $i \in [n]$, we would get $\mu^T a_i = \lambda - \phi[i]$ by setting the partial derivative of $\mathcal{L}$ with respect to $\hat{x}[i]$ to be zero. Similarly if $\hat{x}[i] < 0$, we would get $\mu^T a_i = -\lambda - \phi[i]$ and if $\hat{x}[i] = 0$ we would get

$|\mu^T a_i + \phi[i]| \leq \lambda$. Thus, combining these three cases, we know that

$$|\mu^T a_i + \phi[i]| \leq \lambda, i \in [n]$$

With these results, we can get the Lagrange dual function

$$g(\mu) \triangleq \mathcal{L}(\hat{z}, \hat{x}, \mu) = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|\mu - y\|_2^2$$

Hence, finally, we can come up with the Lagrange dual problem of the non-negative basis pursuit denoising problem as

$$\begin{aligned} &\max_{\mu \in \mathbb{R}^m} \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|\mu - y\|_2^2 \\ &\quad \textit{subject to } |\mu^T a_i + \phi[i]| \leq \lambda \textit{ and } \phi[i] \geq 0, i \in [n] \\ &= \max_{\theta \in \mathbb{R}^m} \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|\mu - y\|_2^2 \quad \textit{subject to } \mu^T a_i \leq \lambda, i \in [n] \\ &= \max_{\theta \in \mathbb{R}^m} \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}\|\theta - y/\lambda\|_2^2 \quad \textit{subject to } \theta^T a_i \leq 1, i \in [n] \end{aligned} \tag{5.27}$$

Furthermore, a primal optimal solution $\hat{x} \in \mathbb{R}^n$ and a dual optimal solution $\hat{\theta} \in \mathbb{R}^m$ satisfy

$$y = A\hat{x} + \lambda\hat{\theta} \tag{5.28}$$

where

$$\hat{\theta}^T a_i = \begin{cases} 1 & if\ \hat{x}[i] > 0 \\ \gamma \in (-\infty, 1] & if\ \hat{x}[i] = 0 \end{cases} \tag{5.29}$$

Define the set $\mathcal{A}$, called the atom pool [4], as $\{\pm a_i, i \in [n]\}$ for the basis pursuit denoising problem and $\{a_i, i \in [n]\}$ for the non-negative basis pursuit denoising problem. In this way, the constraints of the two dual problems can be neatly expressed as

$$\theta^T a \leq 1\ \forall a \in \mathcal{A} \tag{5.30}$$

We further define $H(a) \triangleq \{\theta \mid \theta^T a \leq 1\}$. Hence, the set of feasible points $\mathcal{F}$ of the dual problems is the non-empty, closed and convex set formed by the intersection of the closed half spaces $H(a)\ \forall a \in \mathcal{A}$. De-

[4] In the original paper, the notation is $\mathcal{B}$ and is called the feature pool. We adapt the same concept to our used notation and name.

fine the set $\mathbb{A}(\hat{\theta})$, called the active constraints set at $\hat{\theta}$, as $\{i \,|\, |\hat{\theta}^T a_i| = 1\}$ for the basis pursuit denoising problem and as $\{i \,|\, \hat{\theta}^T a_i = 1\}$ for the non-negative basis pursuit denoising problem. If we know the dual optimal solution $\hat{\theta}$, then any point $\hat{x}$ satisfying the following equations is a primal optimal solution.

1. $A_{\mathbb{A}(\hat{\theta})}\hat{x}|_{\mathbb{A}(\hat{\theta})} = y - \lambda\hat{\theta}$
2. $\hat{x}[i](\hat{\theta}^T a_i) \geq 0,\ i \in \mathbb{A}(\hat{\theta})$
3. $\hat{x}[i] = 0,\ i \notin \mathbb{A}(\hat{\theta})$

To maximize the objective of the two dual problems, we need to project y/λ onto $\mathcal{F}$ to get the dual optimal solution $\hat{\theta}$. We call the set of points $\{\hat{\theta}(\lambda), \lambda > 0\}$ the dual regularization path. We can soon verify that the critical value of λ would be

$$\lambda_{max} \triangleq \max_{a \in \mathcal{A}} y^T a \tag{5.31}$$

We also denote a_{max} as a point belonging to $\underset{a \in \mathcal{A}}{argmax}\ y^T a$. That is,

$$a_{max} \in \underset{a \in \mathcal{A}}{argmax}\, y^T a \tag{5.32}$$

$\forall a \in \mathcal{A}$, $(y/\lambda_{max})^T a \leq y^T a_{max}/\lambda_{max} = 1$. Thus when $\lambda = \lambda_{max}$, y/λ_{max} is itself in $\mathcal{F}$, which implies $\hat{\theta} = y/\lambda_{max}$. If $\lambda \geq \lambda_{max}$, then $\forall a \in \mathcal{A}$, $(y/\lambda)^T a \leq (y/\lambda_{max}) \leq y^T a_{max}/\lambda_{max} = 1$. Again, y/λ is itself in $\mathcal{F}$, which implies $\hat{\theta} = y/\lambda$. If $\lambda < \lambda_{max}$, then $(y/\lambda)^T a_{max}$ would be larger than $(y/\lambda_{max})^T a_{max} = 1$. Hence, y/λ is not in $\mathcal{F}$. As a result, traversing along the dual regularization path from large $\lambda \geq \lambda_{max}$ to small $\lambda < \lambda_{max}$, the dual optimal solution $\hat{\theta}$ would first be equal to y/λ, moving in a straight line ($\hat{\theta} = yc$, where $c = 1/\lambda$) within $\mathcal{F}$ until $\lambda = \lambda_{max}$, where $\hat{\theta} = y/\lambda_{max}$ first lies on the boundary of $\mathcal{F}$. Then, as λ decreases below λ_{max}, y/λ moves away from $\mathcal{F}$ and $\hat{\theta}$ would be the unique projection of y/λ onto the boundary of $\mathcal{F}$. Furthermore, we can make a connection with the primal optimal solution $\hat{x}$. For $\lambda/\lambda_{max} > 1$, $\hat{\theta} = y/\lambda$. Since $-1 < (y/\lambda)^T a < 1\ \forall a \in \mathcal{A}$, $\hat{x} = 0$. For $0 < \lambda/\lambda_{max} < 1$, $\hat{\theta} \neq y/\lambda$. Since $y = A\hat{x} + \lambda\hat{\theta}$, $A\hat{x} = y - \lambda\hat{\theta} \neq 0$,

which implies $\hat{x} \neq 0$. For $\lambda = \lambda_{max}$, $\hat{\theta} = y/\lambda_{max}$ and $\hat{x}$ may be zero or non-zero. Conversely, if $\hat{x} = 0$, then $\hat{\theta} = (y - A\hat{x})/\lambda = y/\lambda$. However, if $\hat{x} \neq 0$, we cannot deduce that $\hat{\theta} \neq y/\lambda$ since $\hat{\theta} = y/\lambda_{max}$ when $\lambda = \lambda_{max}$ and $\hat{x}$ may jointly satisfy $y = A\hat{x} + \lambda\hat{\theta}$ for some $\hat{x} \neq 0$.

With these discussions, we have already had some geometric insights about the primal problems and dual problems, and the primal optimal solutions and the dual optimal solutions. Now we proceed to introduce the concepts of screening and screening tests. Screening refers to seeking an effective partition of the dictionary A to A_S and $A_{\overline{S}}$, where A_S is the sub-dictionary to be retained and $A_{\overline{S}}$ is the sub-dictionary to be rejected. That is, for all indices i belonging to S, the columns a_i will be retained while for all indices j not belonging to S (i.e., belonging to $\overline{S}$), the columns a_j will be rejected (removed). Then, what is an effective partition? Assume we solve the primal problems with the dictionary A and get a primal optimal solution $\hat{x}$. Let $\mathcal{N}$ denote the set $\{i \mid \hat{x}[i] \neq 0\}$. Clearly, if $\mathcal{N} \subseteq S$, then we can still recover the same primal optimal solution $\hat{x}$ by solving the primal problems with the sub-dictionary A_S. Thus, such S is an effective partition. If we know a dual optimal solution $\hat{\theta}$, then clearly $\mathcal{N} \subseteq \mathbb{A}(\hat{\theta})$ from the previous discussions. Hence, if we can ensure that $\mathbb{A}(\hat{\theta}) \subseteq S$, or equivalently $\overline{S} \subseteq \overline{\mathbb{A}}(\hat{\theta})$, then $\mathcal{N} \subseteq S$ as we desire.

To ensure that $\mathbb{A}(\hat{\theta}) \subseteq S$, we need to ensure that for all indices in $\mathbb{A}(\hat{\theta})$, they also belong to S. Note that $i \in \mathbb{A}(\hat{\theta})$ if and only if $|\hat{\theta}^T a_i| = 1$ for the basis pursuit denoising problem (or $\hat{\theta}^T a_i = 1$ for the non-negative basis pursuit denoising problem). Thus, if we assign to S all indices $i \in [n]$ that satisfy $|\hat{\theta}^T a_i| = 1$ (or $\hat{\theta}^T a_i = 1$), $\mathbb{A}(\hat{\theta}) \subseteq S$ is guaranteed. However, if we can know a dual optimal solution $\hat{\theta}$, there is no need to screen the dictionary. In practice, we can only try to bound $\hat{\theta}$ within a compact region $\mathcal{R}$. If $\hat{\theta}$ indeed belongs to $\mathcal{R}$, $\mathbb{A}(\hat{\theta})$ is clearly a subset of the set $\{i \mid \max_{\theta \in \mathcal{R}} |\theta^T a_i| \geq 1\}$ (or $\{i \mid \max_{\theta \in \mathcal{R}} \theta^T a_i \geq 1\}$). Hence, if we assign to S all indices $i \in [n]$ belonging to the set $\{i \mid \max_{\theta \in \mathcal{R}} |\theta^T a_i| \geq 1\}$ (or $\{i \mid \max_{\theta \in \mathcal{R}} \theta^T a_i \geq 1\}$), $\mathbb{A}(\hat{\theta}) \subseteq S$

is guaranteed. As a result, we can construct an effective partition

$$S = \{i \mid \max_{\theta \in \mathcal{R}} |\theta^T a_i| \geq 1\} \ (or \ \{i \mid \max_{\theta \in \mathcal{R}} \theta^T a_i \geq 1\}) \tag{5.33}$$

Moreover, we can equivalently construct a rejection test

$$T_{\mathcal{R}}(a_i) = \begin{cases} 1 & if \ i \in \overline{S} \\ 0 & otherwise \end{cases} \tag{5.34}$$

That is, a_i is rejected if $T_{\mathcal{R}}(a_i) = 1$ and is retained if $T_{\mathcal{R}}(a_i) = 0$. If both $\mathcal{R}_1$ and $\mathcal{R}_2$ contain $\hat{\theta}$ and $\mathcal{R}_1 \subseteq \mathcal{R}_2$, it can be easily verified that the rejection test $T_{\mathcal{R}_1}$ can potentially reject more atoms than the test $T_{\mathcal{R}_2}$. We say that $T_{\mathcal{R}_2}$ is weaker than $T_{\mathcal{R}_1}$ and denote such relation as $T_{\mathcal{R}_2} \preceq T_{\mathcal{R}_1}$. As a special case, $T_{\mathcal{R}_2} \preceq T_{\mathcal{R}_1} \preceq T_{\{\hat{\theta}\}}$.

Therefore, naturally arise two questions that how to find a bounding region $\mathcal{R}$ so that $\hat{\theta}$ can indeed lie in it and how to make $\mathcal{R}$ even tighter so as to boost the screening power of $T_{\mathcal{R}}$. The answers to these two questions lead to constructions of various screening tests. We will demonstrate in the following that $\mathcal{R}$ of our considered screening tests all have a sphere-hyperplane structure, which means it is the intersection of a spherical bound with a finite number of half spaces. Mathematically,

$$\mathcal{R} = \{\theta \mid \|\theta - q\|_2 \leq r\} \cap \cap_{i=1}^{m} \{\theta \mid n_i^T \theta \leq c_i\} \tag{5.35}$$

where q and r are the center and radius of a closed ball

$$S(q, r) = \{z \mid \|z - q\|_2 \leq r\} \tag{5.36}$$

and there are m half spaces

$$n_i^T \theta \leq c_i \ i \in [m] \tag{5.37}$$

With this region, we can construct the corresponding rejection test. Define $\mu_{\mathcal{R}}(a)$ as

$$\mu_{\mathcal{R}}(a) \triangleq \max_{\theta \in \mathcal{R}} \theta^T a \tag{5.38}$$

That is, $\mu_{\mathcal{R}}(a) = -\min_{\theta}(-\theta^T a)$ subject to $\|\theta - q\|_2^2 \leq r^2$ and $n_i^T - c_i \leq 0$, $i \in [m]$. Thus,

$$T_{\mathcal{R}}(a_i) = \begin{cases} 1 & if \ \max\{\mu_{\mathcal{R}}(a_i), \mu_{\mathcal{R}}(-a_i)\} < 1 \\ 0 & otherwise \end{cases} \tag{5.39}$$

for the basis pursuit denoising problem and

$$T_{\mathcal{R}}(a_i) = \begin{cases} 1 & if \ \mu_{\mathcal{R}}(a_i) < 1 \\ 0 & otherwise \end{cases} \tag{5.40}$$

for the non-negative basis pursuit denoising problem. Increasing m, i.e., using more half spaces, will make $\mathcal{R}$ tighter and thus increase the screening power. However, the cost is more computation time. In this paper, $m = 0$ (sphere tests), $m = 1$ (dome tests) and $m = 2$ (two hyperplanes tests) are introduced.

If a dual feasible point $\theta_f \in \mathcal{F}$ is known, we can use it to construct a spherical bound. Indeed, because of the optimality characteristic of $\hat{\theta}$, we know that

$$\|\hat{\theta} - y/\lambda\|_2 \leq \|\theta_f - y/\lambda\|_2 \tag{5.41}$$

As a result, this gives rise to the sphere

$$S(y/\lambda, \|\theta_f - y/\lambda\|_2) \tag{5.42}$$

In particular, we know that y/λ_{max} is dual feasible and thus

$$\|\hat{\theta} - y/\lambda\|_2 \leq \|y/\lambda_{max} - y/\lambda\|_2 = |1/\lambda - 1/\lambda_{max}|\|y\|_2 \tag{5.43}$$

We call such bound the default spherical bound. If a dual optimal solution $\hat{\theta}_0$ is known for a primal problem with $\lambda = \lambda_0$, then by the non-expansive property of the projection onto the convex set,

$$\|\hat{\theta} - \hat{\theta}_0\|_2 \leq \|y/\lambda - y/\lambda_0\|_2 = |1/\lambda - 1/\lambda_0|\|y\|_2 \tag{5.44}$$

In this case, the sphere is

$$S(\hat{\theta}_0, \|y/\lambda - y/\lambda_0\|_2) \tag{5.45}$$

In general, for $S(q,r) = \{\theta \mid \|\theta - q\|_2 \leq r\}$ and $a \in \mathbb{R}^m$, $\mu_{S(q,r)}(a)$ can be calculated as

$$\mu_{S(q,r)}(a) = q^T a + r\|a\|_2 \tag{5.46}$$

Hence, the following theorem states the important result of a sphere test.

Theorem 5.4.1. *The sphere test $ST(q,r)$ for the sphere $S(q,r)$ is $T_{S(q,r)}(a) = \begin{cases} 1 & if\ V_\ell(\|a\|_2) < q^T a < V_u(\|a\|_2) \\ 0 & otherwise \end{cases}$, where $V_u(t) = 1 - rt$ and for the basis pursuit denoising $V_\ell(t) = -V_u(t)$ and for the non-negative basis pursuit denoising $V_\ell(t) = -\infty$.*

In the literature, the Strong Rule introduced in [37] discards atom a_i if

$$|y^T a_i| < 2\lambda - \lambda_{max} \tag{5.47}$$

Indeed, it is a sphere test with center $q = y/\lambda$ and radius $r_{sr} = \lambda_{max}/\lambda - 1$. As we have pointed out, $\hat{\theta}$ is bounded within the default sphere whose center $q = y/\lambda$ and radius $r = (1/\lambda - 1/\lambda_{max})\|y\|_2$. $r_{sr} = r\lambda_{max}/\|y\|_2 \leq r\|a_{max}\|_2 \leq r$ if we normalize the norm of atoms to 1. Although the smaller sphere would lead to greater rejection power, it is not guaranteed to contain $\hat{\theta}$, which may result in false rejections. The Strong Sequential Rule introduced in [37] assumes a primal optimal solution $\hat{x}_0$ of the basis pursuit denoising problem with parameter λ_0 is available, where $\lambda_0 > \lambda$. It then forms the residual $R_0 = y - A\hat{x}_0$ and discards atom a_i if

$$|R_0^T a_i| < 2\lambda - \lambda_0 \tag{5.48}$$

Using the relation that $y = A\hat{x}_0 + \lambda_0\hat{\theta}_0$, screening test is equivalent as $|\hat{\theta}_0^T a_i| < 2\lambda/\lambda_0 - 1$. Indeed, it is a sphere test with center $q = \hat{\theta}_0$ and radius $r_{ssr} = 2 - 2\lambda/\lambda_0 = 2\lambda(1/\lambda - 1/\lambda_0)$. As we have pointed out, $\hat{\theta}$ is bounded with the sphere $S(q,r)$, where $q = \hat{\theta}_0$ and $r = (1/\lambda - 1/\lambda_0)\|y\|_2$. Thus, $r_{ssr} = \frac{2\lambda}{\|y\|_2} r \leq r$ if $\lambda \leq \frac{\|y\|_2}{2}$, which again inflicts risk of false rejections on the screening test. The SIS test introduced

in [11] can be interpreted as a default sphere test for a basis pursuit denoising problem with a specific parameter λ. Assume the atoms are normalized to have norm 1. We compute the correlation vector $\rho = A^T y$. Given $0 < \gamma < 1$, the SIS rejection test discard atom a_i if $|\rho[i]| < \rho^*[(\gamma m)]$, where ρ^* denotes the non-increasing rearrangement of ρ and (γm) denotes the integer part of γm. The test can be expressed as

$$|y^T a_i| < t_\gamma \triangleq \rho^*[(\gamma m)] \tag{5.49}$$

equivalently, $|(y/\lambda)^T a_i| < t_\gamma/\lambda$. Equating t_γ/λ with $1 - (1/\lambda - 1/\lambda_{max})\|y\|_2$, then we can verify that the SIS test is indeed a default sphere test for a basis pursuit denoising problem with $\lambda = \frac{\lambda_{max}(t_\gamma + \|y\|_2)}{\lambda_{max} + \|y\|_2} \leq \lambda_{max}$.

We have introduced the $m = 0$ case and we can even further confine the bounding region $\mathcal{R}$ by incorporating additional hyperplanes. Indeed, each constraint $\theta^T a_i \leq 1$ bounds $\hat{\theta}$. We can pick one of them and express it as $\{\theta \,|\, n^T\theta \leq c\}$, where n has norm 1. Combining it with some spherical region $S(q, r)$, we can come up with a dome bounding region $D(q, r; n, c)$. We can visualize $D(q, r; n, c)$ in the following graph. We call q_d the dome center and r_d the dome radius.

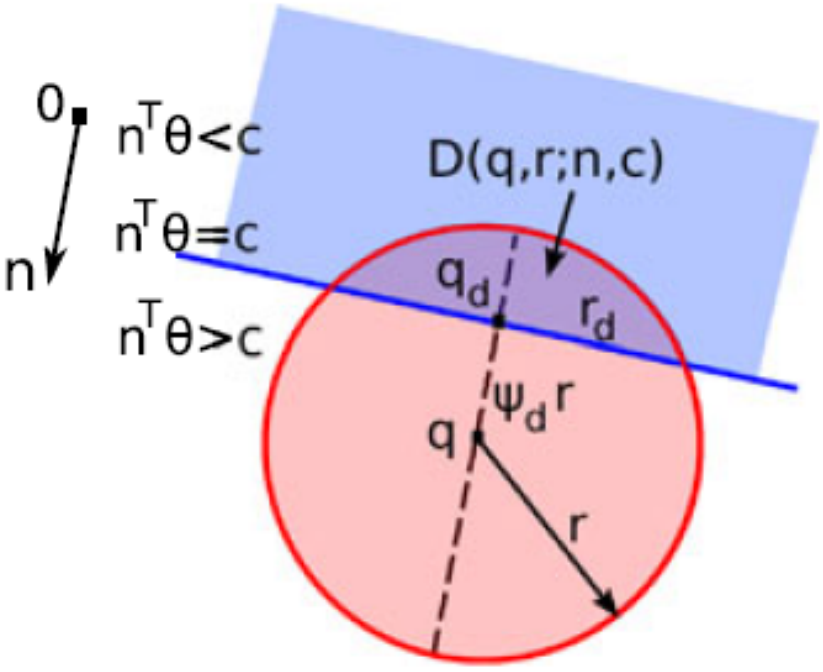

Figure 5.3: A general dome region $D(q, r; n, c)$ with $0 < \psi_d < 1$

The signed distance from q to q_d in the direction $-n$ is a fraction of the radius r of the sphere. We use ψ_d to denote such fraction. $n^T(q_d - q) = \|n\|_2\|q_d - q\|_2 \cos(0°) = \|q_d - q\|_2 = -\psi_d r$. Hence,

$$\psi_d = (n^T q - n^T q_d)/r = (n^T q - c)/r \tag{5.50}$$

$q_d - q = \psi_d r(-n)$. Hence,

$$q_d = q - \psi_d r n \tag{5.51}$$

Finally,

$$r_d = r\sqrt{1 - \psi_d^2} \tag{5.52}$$

To effectively confine the spherical region $S(q, r)$ to a smaller dome region $D(q, r; n, c)$, we require $-1 \leq \psi_d \leq 1$. Actually, we desire ψ_d to be as close to 1 as possible. Thus, when choosing the incorporated constraint, it is natural to select what can maximize ψ_d, say $\theta^T a_g \leq 1$. The corresponding hyperplane would be $n^T\theta \leq c$, where $n = a_g/\|a_g\|_2$ and $c = 1/\|a_g\|_2$. Hence,

$$a_g = \underset{a\in\mathcal{A}}{argmax}\, \frac{a^T q - 1}{\|a\|_2} \tag{5.53}$$

The corresponding dome will be

$$D(q, r; a_g/\|a_g\|_2, 1/\|a_g\|_2) \tag{5.54}$$

If we use the default sphere, i.e., $q = y/\lambda$ and $r = |1/\lambda - 1/\lambda_{max}|\|y\|_2$, then $a_g = a_{max}$. The corresponding dome is called the default dome. If we are given a dual optimal solution $\hat{\theta}_0$ for a primal problem with $\lambda = \lambda_0$, then we have another way to construct an effective hyperplane. By the projection theorem, for any $\theta \in \mathcal{F}$,

$$(y/\lambda_0 - \hat{\theta}_0)^T(\theta - \hat{\theta}_0) \leq 0 \tag{5.55}$$

That is, $(y/\lambda_0 - \hat{\theta}_0)^T\theta \leq (y/\lambda_0 - \hat{\theta}_0)^T\hat{\theta}_0$. Since $0 \in \mathcal{F}$, the right hand side is non-negative. Thus, this inequality defines an effective hyperplane $n_0^T\theta \leq c_0$, where $n_0 = \frac{y/\lambda_0 - \hat{\theta}_0}{\|y/\lambda_0 - \hat{\theta}_0\|_2}$ and $c_0 = n_0^T\hat{\theta}_0$. In this way, $\psi_d = \frac{n_0^T q - n_0^T\hat{\theta}_0}{r} = \frac{\|n_0\|_2\|q - \hat{\theta}_0\|_2\cos(\beta)}{r} = \frac{\|q - \hat{\theta}_0\|_2\cos(\beta)}{r}$, where β is the

angle between n_0 and $q - \hat{\theta}_0$. Since $\hat{\theta}_0 \in \mathcal{F}$, we can select the sphere to be $S(q, r)$, where $q = y/\lambda$ and $r = \|\hat{\theta}_0 - y/\lambda\|_2$. Then we can verify that ψ_d is indeed within -1 and 1. The corresponding dome can be expressed as

$$D(y/\lambda, \|\hat{\theta}_0 - y/\lambda\|_2; n_0, c_0) \tag{5.56}$$

In general, for $D(q, r; n, c) = \{\theta \mid \|\theta - q\|_2 \leq r, n^T\theta \leq c\}$, and $a \in \mathbb{R}^m$, $\mu_{D(q,r;n,c)}(a)$ can be calculated as

$$\mu_{D(q,r;n,c)}(a) = q^T a + M_1(n^T a, \|a\|_2) \tag{5.57}$$

where

$$M_1(t_1, t_2) = \begin{cases} rt_2 & if\ t_1 < -\psi_d t_2 \\ -\psi_d r t_1 + r\sqrt{t_2^2 - t_1^2}\sqrt{1 - \psi_d^2} & if\ t_1 \geq -\psi_d t_2 \end{cases} \tag{5.58}$$

Hence, the following theorem states an important result of a dome test.

Theorem 5.4.2. *The dome test $DT(q, r; n, c)$ for the dome $D(q, r; n, c)$ is $T_{D(q,r;n,c)}(a_i) = \begin{cases} 1 & if\ V_\ell(n^T a_i, \|a_i\|_2) < q^T a_i < V_u(n^T a_i, \|a_i\|_2) \\ 0 & otherwise \end{cases}$, where $V_u(t_1, t_2) = 1 - M_1(t_1, t_2)$ and for the basis pursuit denoising $V_\ell(t_1, t_2) = -V_u(-t_1, t_2)$ and for the non-negative basis pursuit denoising $V_\ell(t_1, t_2) = -\infty$.*

In the literature, the SAFE-LASSO test introduced in [15] is a dome test $D(y/\lambda, \|\hat{\theta}_0 - y/\lambda\|_2; n_0, c_0)$, where $\hat{\theta}_0$ is a dual optimal solution for a primal problem with $\lambda = \lambda_0$, $n_0 = \frac{y/\lambda_0 - \hat{\theta}_0}{\|y/\lambda_0 - \hat{\theta}_0\|_2}$, $c_0 = n_0^T\hat{\theta}_0$ and $\lambda < \lambda_0 \leq \lambda_{max}$. Due to the optimality characteristic of $\hat{\theta}$, for any $\theta_f \in \mathcal{F}$, $\|\hat{\theta} - y/\lambda\|_2 \leq \|\theta_f - y/\lambda\|_2$. Specifically, since $\hat{\theta}_0 \in \mathcal{F}$, $s\hat{\theta}_0$ also belongs to $\mathcal{F}$ for any $-1 \leq s \leq 1$. As a result, $\|\hat{\theta} - y/\lambda\|_2 \leq \hat{r} \triangleq \min_{-1 \leq s \leq 1} \|s\hat{\theta}_0 - y/\lambda\|_2$. The author proved that for $\lambda < \lambda_0$, $\hat{s}(\lambda) \triangleq \underset{-1 \leq s \leq 1}{argmin} \|s\hat{\theta}_0 - y/\lambda\|_2 = 1$. Indeed, by simple calculus, we know that $\hat{s}(\lambda) = \max\{-1, \min\{1, (y^T\hat{\theta}_0)/(\lambda\|\hat{\theta}_0\|_2^2)\}\}$. By the optimality characteristic of $\hat{\theta}_0$, $\hat{s}(\lambda_0) = 1$. Moreover, since

$\|\hat{\theta}_0 - y/\lambda\|_2^2 - \| - \hat{\theta}_0 - y/\lambda\|_2^2 = -4y^T\hat{\theta}_0/\lambda$, $y^T\hat{\theta}_0 \geq 0$ again by the optimality characteristic of $\hat{\theta}_0$. Hence, $(y^T\hat{\theta}_0/(\lambda\|\hat{\theta}_0\|_2^2)) \geq 0$, which means $\hat{s}(\lambda) = \min\{1, (y^T\hat{\theta}_0)/(\lambda\|\hat{\theta}_0\|_2^2)\}$. Finally, since $\hat{s}(\lambda_0) = 1$ and $\lambda < \lambda_0$, $\hat{s}(\lambda) = 1$.

Up to now, we have discussed the $m = 0$ and $m = 1$ cases. In the following, we will discuss the $m = 2$ case; that is, to incorporate an additional hyperplane compared with the dome tests. For the first hyperplane, we can do the same thing as we do to construct dome tests. We can select

$$a^{(1)} = \underset{a \in \mathcal{A}}{argmax} \frac{a^T q - 1}{\|a\|_2} \tag{5.59}$$

and form a hyperplane $n_1^T\theta \leq c_1$, where

$$n_1 = a^{(1)}/\|a^{(1)}\|_2 \tag{5.60}$$

and

$$c_1 = 1/\|a^{(1)}\|_2 \tag{5.61}$$

If we are given a dual optimal solution $\hat{\theta}_0$ for a primal problem with $\lambda = \lambda_0$, then as we have pointed out, we can form a hyperplane $n_1^T\theta \leq c_1$, where

$$n_1 = \frac{y/\lambda_0 - \hat{\theta}_0}{\|y/\lambda_0 - \hat{\theta}_0\|_2} \tag{5.62}$$

and

$$c_1 = n_1^T\hat{\theta}_0 \tag{5.63}$$

With a sphere $S(q, r)$ and the first hyperplane, we can construct a dome region $D(q, r; n_1, c_1)$. Ideally, we assume $\psi_d \geq 0$. In this way, it is clear that the sphere $S(q_d, r_d)$ is the circumsphere of the dome region. To be mathematically rigorous, we need to show that for every point p on the boundary of $D(q, r; n_1, c_1)$ is contained within $S(q_d, r_d)$. p can be expressed as $q_d + \alpha v + \beta n$, where v is a unit norm vector orthogonal to n and α, β are scalars with $\beta \leq 0$. Hence, we have to show that $\|p - q_d\|_2 = \alpha^2 + \beta^2 \leq r_d^2$. Note that $r^2 = \|p - q\|_2^2 = \|q_d - q + \alpha v + \beta n\|_2^2 = \|(-\psi_d r + \beta)n + \alpha v\|_2^2 = \psi_d^2 r^2 - 2\psi_d r\beta + \alpha^2 + \beta^2$. Since $\beta \leq 0$

and $\psi_d \geq 0$, $\alpha^2 + \beta^2 = r^2(1-\psi_d^2) + 2\psi_d r\beta \leq r^2(1-\psi_d^2) = r_d^2$. Hence, $S(q_d, r_d)$ is a tighter bounding region than $S(q, r)$. We can form the second hyperplane $n_2^T\theta \leq c_2$ based on $S(q_d, r_d)$ by selecting

$$a^{(2)} = \underset{a \in \mathcal{A} \setminus a^{(1)}}{argmax} \frac{a^T q_d - 1}{\|a\|_2} \tag{5.64}$$

thus

$$n_2 = a^{(2)}/\|a^{(2)}\|_2 \tag{5.65}$$

and

$$c_2 = 1/\|a^{(2)}\|_2 \tag{5.66}$$

Combining the spherical region and the two hyperplanes, we can come up with a bounding region $R(q, r; n_1, c_1; n_2, c_2)$. Note that if the first hyperplane is formed by maximizing $\frac{a^T q - 1}{\|a\|_2}$, the corresponding two hyperplanes test (THT) is called dictionary-based THT (D-THT). Also note that generally, to ensure the two half spaces intersect within the sphere so that the bounding region is effectively confined,

$$-1 \leq \psi_i = (n_i^T q - c_i)/r \leq 1, i = 1, 2 \tag{5.67}$$

and

$$\cos^{-1}(\psi_1) + \cos^{-1}(\psi_2) \geq \cos^{-1}(n_1^T n_2) \tag{5.68}$$

In general, for $R(q, r; n_1, c_1; n_2, c_2) = \{\theta \,|\, \|\theta - q\|_2 \leq r, n_1^T\theta \leq c_1, n_2^T\theta \leq c_2\}$ and $a \in \mathbb{R}^m$, $\mu_{R(q,r;n_1,c_1;n_2,c_2)}(a)$ can be calculated as

$$\mu_{R(q,r;n_1,c_1;n_2,c_2)}(a) = q^T a + M_2(n_1^T a, n_2^T a, \|a\|_2) \tag{5.69}$$

where

$$M_2(t_1, t_2, t_3) = \begin{cases} rt_3 & if\ (a) \\ -rt_2\psi_2 + r\sqrt{t_3^2 - t_2^2}\sqrt{1-\psi_2^2} & if\ (b) \\ -rt_1\psi_1 + r\sqrt{t_3^2 - t_1^2}\sqrt{1-\psi_1^2} & if\ (c) \\ -\frac{r}{1-\tau^2}[(\psi_1 - \tau\psi_2)t_1 + (\psi_2 - \tau\psi_1)t_2] + & \\ \frac{r}{1-\tau^2}h(\psi_1, \psi_2, 1)h(t_1, t_2, t_3) & otherwise \end{cases} \tag{5.70}$$

$\tau = n_1^T n_2$, $h(x, y, z) = \sqrt{(1-\tau^2)z^2 + 2\tau xy - x^2 - y^2}$ and conditions (a), (b) and (c) are given by

(a) $t_1 < -\psi_1 t_3 \,\&\, t_2 < -\psi_2 t_3$

(b) $t_2 \geq -\psi_2 t_3 \,\&\, (t_1 - \tau t_2)/\sqrt{t_3^2 - t_2^2} < (-\psi_1 + \tau\psi_2)/\sqrt{1-\psi_2^2}$

(c) $t_1 \geq -\psi_1 t_3 \,\&\, (t_2 - \tau t_1)/\sqrt{t_3^2 - t_1^2} < (-\psi_2 + \tau\psi_1)/\sqrt{1-\psi_1^2}$

Hence, the following theorem states the important result of a two hyperplanes test.

Theorem 5.4.3. *The two hyperplanes test $THT(q, r; n_1, c_1; n_2, c_2)$ for the region $R(q, r; n_1, c_1; n_2, c_2)$ (abbreviated as R) is*

$$T_R(a_i) = \begin{cases} 1 & if\ V_\ell(n_1^T a_i, n_2^T a_i, \|a_i\|_2) < q^T a_i < V_u(n_1^T a_i, n_2^T a_i, \|a_i\|_2) \\ 0 & otherwise \end{cases}$$

where $V_u(t_1, t_2, t_3) = 1 - M_2(t_1, t_2, t_3)$ and for the basis pursuit denoising $V_\ell(t_1, t_2, t_3) = -V_u(-t_1, -t_2, t_3)$ and for the non-negative basis pursuit denoising $V_\ell(t_1, t_2, t_3) = -\infty$.

When deriving the two hyperplanes tests, we introduced the concept of finding the circumsphere of a dome region. We can expand this idea more generally. Assume at step k, we have a bounding sphere $S_k = S(q_k, r_k)$. Then we want to find an effective hyperplane $n_k^T \theta \leq c_k$ so that we can confine the region to the dome $D_k = D(q_k, r_k; n_k, c_k)$. Lastly, if ideally $\psi_k \geq 0$, we can get the circumsphere $S_{k+1} = S(q_{k+1}, r_{k+1})$ of D_k. The way we find an effective hyperplane is by maximizing ψ_k; that is,

$$a^{(k)} = \underset{a \in \mathcal{A} \setminus \{a^{(1)}, \cdots, a^{(k-1)}\}}{argmax} \frac{a^T q_k - 1}{\|a\|_2} \tag{5.71}$$

$$n_k = a^{(k)} / \|a^{(k)}\|_2 \tag{5.72}$$

$$c_k = 1 / \|a^{(k)}\|_2 \tag{5.73}$$

The new center

$$q_{k+1} = q_k - \psi_k r_k n_k \tag{5.74}$$

and the new radius

$$r_{k+1} = r_k \sqrt{1 - \psi_k^2} \tag{5.75}$$

During the process of successive construction, we can get a sequence of spheres and domes : $S_1 \supset D_1 \subset S_2 \supset \cdots \supset S_{k-1} \supset D_{k-1} \subset S_k$. Since dome D_j is contained in S_j and S_{j+1}, each dome test is stronger than the sphere tests for the spheres that precede and succeed it. However, since S_{j+1} is not contained in S_j, D_{j+1} is not contained in D_j. Thus, we cannot deduce that the last dome test is the strongest. A test based on the region $\cap_{j=1}^{k-1} D_j$ is certainly the strongest. Nonetheless, such test would be too complex to compute. Alternatively, we can form a composite test

$$T_c \triangleq T_{D_1} \vee T_{D_2} \vee \cdots \vee T_{D_{k-1}} \tag{5.76}$$

Equivalently, an atom a_i is rejected by any of the tests $\{T_{D_j} \,|\, j \in [k-1]\}$. Such composite test is called the iteratively refined dome test (IRDT). IRDT is practically easy to implement. First, we apply T_{D_1} on all atoms. Next, we apply T_{D_2} on the remaining atoms. Then we apply T_{D_3}, T_{D_4}, $\cdots$. That is to say, we can sequentially apply the dome tests on gradually shrinking atom pools. However, we have to point out that T_c is weaker than $T_{\cap_{j=1}^{k-1} D_j}$. Indeed, if an atom a_i is rejected by T_c (assume it is rejected by T_{D_s} for some s), then $\max\limits_{\theta \in D_s} |\theta^T a_i| < 1$ for the basis pursuit denoising problem or $\max\limits_{\theta \in D_s} \theta^T a_i < 1$ for the non-negative basis pursuit denoising problem. Since $\cap_{j=1}^{k-1} \subseteq D_s$, $\max\limits_{\theta \in \cap_{j=1}^{k-1}} |\theta^T a_i| \leq \max\limits_{\theta \in D_s} |\theta^T a_i| < 1$ (or $\max\limits_{\theta \in \cap_{j=1}^{k-1}} \theta^T a_i \leq \max\limits_{\theta \in D_s} \theta^T a_i < 1$). Hence, a_i is also rejected by $T_{\cap_{j=1}^{k-1} D_j}$. In particular, the IRDT test using only two domes D_1 and D_2 is weaker than the D-THT test. In the literature, the sphere test ST3 introduced in [40] utilizes the concept of finding the circumsphere of a dome region and is based on a refined spherical bound,

which is $S(q, r) = S_2$.

So far, our introduced tests all fall in the category of "one-shot" screening tests, which means we screen the dictionary, then solve the reduced primal problem with parameter λ_t once and finish. Such paradigm is claimed to perform well for moderate to large values of λ/λ_{max} but often fail to support smaller values of λ/λ_{max} (empirically, for $\lambda/\lambda_{max} < 0.2$). A concept called sequential screening is proposed to deal with such issue. Choose a parameter $\lambda_1 < \lambda_{max}$, say $0.95\lambda_{max}$, and then select N points along the dual regularization path from λ_1 to $\lambda_N = \lambda_t$. To solve the primal problem with parameter λ_t, we first screen and solve the primal problem with parameter λ_1 and get a primal-dual optimal solution pair $(\hat{x}_1, \hat{\theta}_1)$. Then sequentially for $k = 2, \cdots, N$, we screen the dictionary with the help of previously obtained solution $(\lambda_{k-1}, \hat{x}_{k-1}, \hat{\theta}_{k-1})$ and solve the primal problem with parameter λ_k. Finally, when $k = N$, we can get the desired primal-dual optimal solution pair $(\hat{x}_N, \hat{\theta}_N)$. Sometimes, all solutions along the dual regularization path are of interest; for instance, for parameter selection. Note that at each step, we can use all the tests we have introduced, e.g. the sphere tests, the dome tests, the IRDT tests and the D-THT test, to screen the dictionary and use any possible convex optimization solvers to solve the primal problem. As for N and $\{\lambda_k \,|\, k = 2, \cdots, N-1\}$, we introduce two ways to determine them. One is to directly assign a value for N and select λ_k via geometric spacing :

$$\lambda_k = \alpha\lambda_{k-1} \; where \; \alpha = (\lambda_t/\lambda_1)^{1/(N-1)} \tag{5.77}$$

The other one determine them in an adaptive way with the following proposition from [39].

Proposition 5.4.4. *Let $D_k(q_k, r_k; n_k, c_k)$ be a dome bounded by the sphere $S(q_k, r_k)$ with $q_k = y/\lambda_k$ and $r_k = \|\hat{\theta}_{k-1} - y/\lambda_k\|_2$ and the hyperplane $n_k^T\theta \leq c_k$ with $n_k = \frac{y/\lambda_{k-1} - \hat{\theta}_{k-1}}{\|y/\lambda_{k-1} - \hat{\theta}_{k-1}\|_2}$ and $c_k = n_k^T\hat{\theta}_{k-1}$.*

Also let δ_k denote the diameter of D_k. Then

$$\delta_k = 2(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k-1}})\sqrt{y^T(I - n_k n_k^T)y} \tag{5.78}$$

Hence,

$$\frac{1}{\lambda_k} = \frac{1}{\lambda_{k-1}} + \frac{1/2\delta_k}{\sqrt{y^T(I - n_k n_k^T)y}} \tag{5.79}$$

Let $\delta_k = R > 0$ be a selectable parameter. In this way, we can adaptively determine λ_k and also directly control how tightly the dome bound D_k bounds $\hat{\theta}_k$ based on the choice of R. We continue this process until at step N, $\lambda_N \leq \lambda_t$ (if $\lambda_N < \lambda_t$, let $\lambda_N = \lambda_t$). Such screening method is called the data-adaptive sequential screening (DASS).

To evaluate the performance of each test, we adopt two metrics, which are the rejection fraction

$$\frac{|\overline{S}|}{n} \tag{5.80}$$

and the speedup factor

$$\frac{t_{solve}}{t_{screen} + t^r_{solve}} \tag{5.81}$$

respectively. Here t_{solve} is the time to solve the primal problem with non-screened dictionary, t_{screen} is the time to screen the dictionary, and t^r_{solve} is the time to solve the primal problem with screened dictionary. For one-shot screening tests, the THT tests perform significantly well beyond simpler tests in both rejection fraction and speedup factor, which means it is indeed worthwhile to seek more complex region tests. However, as we have pointed out, one-shot screening falls short of desirable performance when λ/λ_{max} is small. Sequential screening can effectively remedy such problem and operates successfully in a wide range of λ/λ_{max}. Specifically, DASS boasts adaptive and automatic selection of both the N and $\{\lambda_k \,|\, k = 1, 2, \cdots, N\}$. In brief, dictionary screening can effectively identify a subset of dictionary atoms which will not appear in a solution of the primal problem and thus can be

removed before we start to solve the primal problem. In this way, not only can the size of dictionary be reduced in order to save storage space but also can the primal problem be solved faster.

References

[1] D. Anderson and K. Burnham, *Model selection and multi-model inference*. Springer New York, NY, no. 2.

[2] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 330–348, 2017.

[3] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier analysis and Applications*, vol. 14, pp. 629–654, 2008.

[4] ——, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[5] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[6] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[7] G. M. Davis, S. G. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," *Optical engineering*, vol. 33, no. 7, pp. 2183–2191, 1994.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[9] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise

orthogonal matching pursuit," *IEEE transactions on Information Theory*, vol. 58, no. 2, pp. 1094–1121, 2012.

[10] J. Eckstein and W. Yao, "Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," *RUTCOR Research Reports*, vol. 32, no. 3, p. 44, 2012.

[11] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 70, no. 5, pp. 849–911, 2008.

[12] S. Foucart, "Hard thresholding pursuit: an algorithm for compressive sensing," *SIAM Journal on numerical analysis*, vol. 49, no. 6, pp. 2543–2563, 2011.

[13] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser New York, NY, 2013.

[14] C. J. Geyer, "Introduction to markov chain monte carlo," *Handbook of markov chain monte carlo*, vol. 20116022, p. 45, 2011.

[15] L. E. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination for the lasso and sparse supervised learning problems," *arXiv preprint arXiv:1009.4219*, 2010.

[16] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Transactions on signal processing*, vol. 45, no. 3, pp. 600–616, 1997.

[17] M. R. Gupta and Y. Chen, "Theory and use of the em algorithm," *Foundations and Trends® in Signal Processing*, vol. 4, no. 3, pp. 223–296, 2011. [Online]. Available: http://dx.doi.org/10.1561/2000000034

[18] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the l-curve," *SIAM review*, vol. 34, no. 4, pp. 561–580, 1992.

[19] P. C. Hansen and D. P. O' Leary, "The use of the l-curve in the regularization of discrete ill-posed problems," *SIAM journal on scientific computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[20] A. Hero, "On the convergence of the em algorithm," in *Proceedings. IEEE International Symposium on Information Theory*, 1993, pp. 187–187.

[21] C. M. Hurvich and C.-L. Tsai, "A corrected akaike information criterion for vector autoregressive model selection," *Journal of time series analysis*, vol. 14, no. 3, pp. 271–279, 1993.

[22] H. Lu, R. M. Freund, and Y. Nesterov, "Relatively smooth convex optimization by first-order methods, and applications," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 333–354, 2018.

[23] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[24] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, ser. Wiley Series in Probability and Statistics. Wiley, 2007. [Online]. Available: https://books.google.com.tw/books?id=NBawzaWoWa8C

[25] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.

[26] D. Needell and J. A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and computational harmonic analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[27] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of computational mathematics*, vol. 9, pp. 317–334, 2009.

[28] ——, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *IEEE Journal of selected topics in signal processing*, vol. 4, no. 2, pp. 310–316, 2010.

[29] S. Ng, T. Krishnan, and G. Mclachlan, "The em algorithm," *Handbook of Computational Statistics: Concepts and Methods*, 01 2004.

[30] C. C. Paige and M. A. Saunders, "Towards a generalized singular value decomposition," *SIAM Journal on Numerical Analysis*, vol. 18, no. 3, pp. 398–405, 1981.

[31] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40–44 vol.1.

[32] B. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 187–200, 1999.

[33] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE transactions on Signal processing*, vol. 51, no. 3, pp. 760–770, 2003.

[34] R. T. Rockafellar, *Convex analysis*. Princeton university press, 1997, vol. 11.

[35] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[36] M. Teboulle, "A simplified view of first order methods for optimization," *Mathematical Programming*, vol. 170, no. 1, pp. 67–96, 2018.

[37] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 74, no. 2, pp. 245–266, 2012.

[38] S. Tu, "Derivation of baum-welch algorithm for hidden markov models," *URL: https://people. eecs. berkeley. edu/∼ stephentu/ writeups/hmm-baum-welch-derivation. pdf*, 2015.

[39] Y. Wang, X. Chen, and P. J. Ramadge, "Feedback-controlled sequential lasso screening," *arXiv preprint arXiv:1608.06010*, 2016.

[40] Z. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," *Advances in neural information processing systems*, vol. 24, 2011.

[41] Z. J. Xiang, Y. Wang, and P. J. Ramadge, "Screening tests for lasso problems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 1008–1027, 2016.

[42] K. Yosida, *Functional analysis*. Springer Science & Business Media, 2012.