Dictionary Learning and Anti-Concentration

Broadening the Reach of Efficient, Gradient-Descent Algorithms for Learning Sparsely-Used, Incoherent Dictionaries

Max Simchowitz
Advised by Professor Sanjeev Arora
Department of Mathematics
Princeton University
United States
May 4, 2015

This thesis represents my own work in accordance with university regulations
- Max Simchowitz

Abstract

As central as *concentration of measure* is to statistics and machine learning, this thesis aims to motivate *anti-concentration* as a promising and under-utilized toolkit for the design and analysis of statistical learning algorithms. This thesis focuses on learning incoherent dictionaries A^* from observations $y = A^*x$, where x is a sparse coefficient vector drawn from a generative model. We impose an exceedingly simple anti-concentration property on the entries of x, which we call (C, ρ)-smoothness. Leveraging this assumption, we present the first computationally efficient, provably correct algorithms to approximately recover A^* even in the setting where *neither* the non-zero coordinates of x are guaranteed to be $\Omega(1)$ in magnitude, nor are the supports x chosen in a uniform fashion. As an application of our analytical framework, we present an algorithm which learns a class of randomly generated non-negative matrix factorization instances with run-time and sample complexity polynomial in the dimension and logarithmic in the desired precision.

Acknowledgements

First and foremost, I would like to express my deep gratitude to Professor Sanjeev Arora for his advice, support, and guidance over the course of writing this thesis. I would also like to thank Tengyu Ma and Rong Ge for their exceptionally helpful correspondences regarding the subtler points of their paper "Simple, Efficient, and Neural Algorithms for Sparse Coding" Arora et al. (2015), co-authored by Professor Arora and Professor Ankur Moitra. The algorithms in this thesis and the analyses thereof are heavily indebted to their joint work.

I am very grateful to Professor David Blei for first sparking my interest in the field of machine learning, and to Professor Philippe Rigollet for introducing me to the worlds of learning theory and high dimensional statistics. I also wish to thank Professor Ramon van Handel for always making time to point me to references in the concentration of measure literature, and to Professors Elad Hazan and Ankur Moitra, both of whom have deepened my appreciation for learning theory at large. Finally, I am immensely grateful to Professor Amit Singer, who advised my first research project at Princeton University, and who has graciously agreed to serve as the second reader of this thesis.

Contents

Chapter 1

Introduction

1.1 Anti-Concentration in Machine Learning

Both machine learning and theoretical computer science are in great debt to the *concentration of measure* phenomenom, which roughly states:

> *"If $X_1, \ldots, X_n$ are independent (or weakly dependent) random variables, then the random variable $f(X_1, \ldots, X_n)$ is 'close' to its mean $\mathbb{E}[f(X_1, ;X_n)]$, provided that $f(x_1, \ldots, x_n)$ is not too 'sensitive' to any of its coordinates"* - Chapter 1, van Handel (2014)

In statistics, concentration guarantees that the average $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$ of light-tailed, i.i.d. random vectors $x^{(i)}$ does not deviate too far from its mean. As an example from the analysis of algorithms, consider a procedure $\mathcal{P}$ which succeeds with probability p: e.g. $\mathcal{P}$ rounds a solution to a linear program, or balances tasks across many servers. If we repeat $\mathcal{P}$ for n trials, Chernoff concentration ensures that, despite the uncertainty of the procedure, $\mathcal{P}$ will succeed $\Omega(pn)$ times with high probabilty (Proposition C.1.6).

When we appeal to concentration of measure, we view randomness as a deviation; a source of error to be controlled. However, there are settings in which randomness can be regarded as more benign than adversarial. The quintessential example of a more favorable attitude towards randomness is the smoothed analysis of algorithms. In smoothed analysis, an algorithm $\mathcal{A}$ is fed a random perturbation $\tilde{\mathcal{I}}$ of a deterministic, and possibly adversarially chosen input $\mathcal{I}$. Even if $\mathcal{A}$ runs in worst case exponential time, many algorithms run polynomially on suitably perturbed instances, with high probability. Spielman and Teng (2001) establishes that the popular Simplex Algorithm for linear programming runs in smoothed polynomial time, despite its worst case exponential complexity. More recently, Bhaskara et al. (2014) shows that a class of worst-case intractable tensor decompositions can be recovered in polynomial time in the smoothed analysis framework.

Whereas smoothed analysis introduces extrinsic noise to "smooth out" particularly difficult instances, recently, Mendelson (2014) leverages *intrinsic* randomness in statistical learning problems to vastly improve known bounds on empirical risk minimization (ERM). In Mendelson's setup, we have a distribution $\mathcal{D}$ over data-label pairs $(x, y) \in X \times Y$, and aim approximate the function $f^* : X \to Y$ which minimizes the expected l_2 loss, or risk, $R(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$, over all f in some hypothesis class $\mathcal{F}$.

Mendelson's strategy is to lower bound the "small-ball probability"

$$Q(u) := \Pr(|f(x)| \geq u\sqrt{\mathbb{E}[f(x)^2]}) \tag{1.1}$$

uniformly over all $f \in \mathcal{F}$. Hence, if a particular f has large risk in expectation, then the small-ball probability estimates ensure that f has a large empirical risk $ER(f) := \frac{1}{n}\sum_{i=1}^{n}(f(x^{(i)}) - y^{(i)})^2$ over the samples $(x^{(i)}, y^{(i)})$ as well. Consequently, an estimation procedure can determine that f lies far from the true risk minimizer, f^*.

Fundamentally, small-ball probability measures the tendency of random quantities to disperse across their range, a phenenomon known more generally as *anti-concentration* Vershynin and Rudelson (2007). For an even simpler example, consider a standard normal random variable $X \sim \mathcal{N}(0, 1)$. Then $\sqrt{\mathbb{E}[X^2]} = 1$, and X's density function (with respect to the Lebesgue measure) is bounded above $\frac{1}{\sqrt{2\pi}}$. Integrating, we can conclude

$$\Pr(|X| \geq u\sqrt{\mathbb{E}[X^2]}) \;\geq\; 1 - \sqrt{\frac{2}{\pi}}u \tag{1.2}$$

In fact, we can say something much more generally about any continuously distributed real-valued random variable X with density $p(u)$ that is bounded uniformly above by a constant $\rho > 0$:

$$\text{If} \quad \sup_{u\in\mathbb{R}} p(u) \leq \rho \quad \text{then} \quad \Pr(X \in S) \leq \rho \,\text{vol}(S) \tag{1.3}$$

where $\text{vol}(S)$ denotes the volume, or Lebesgue measure, of the set S (assuming S is Lebesgue measurable). We call the property defined by Equation 1.3 *ρ-smoothness*, and it will play a central role in the remainder of this report. Like small-ball probability, ρ-smoothness aims to capture the tendency of a random variable to *anti-concentrate* on sets of small volume.

Further References for Anti-Concentration

ρ-smoothness is the only small-ball style assumption to which we will appeal in this thesis. It is at best an extreme simplification, and possibly even a trivialization of the rich theory of small-ball estimates to which we have been unable to do justice in this very short introduction. Both small-ball probability and anti-concentration are deeply connected to the fields of geometric functional analysis and convex geometry, and we point the curious reader to Vershynin and Rudelson (2007) and Nguyen and Vu (2013) for a more thorough treatment. As another example of the power of anti-concentration in learning theory, we direct the reader to Bresler (2014) which uses Erdos' anti-concentration bounds for the Littlewood-Offord problem to efficiently learn Ising models on graphs.

1.1.1 The Dictionary Learning Problem

In the present work, we leverage anti-concentration to improve the analyses of algorithms for the *sparse coding*, or *dictionary learning* problem. In dictionary learning, the goal is to find a collection of often overcomplete basis vectors, collectively referred to as the *dictionary*, for which any input data vector can be represented, or *encoded*, as a linear combination of only a few vectors from that basis. We call the coefficients of the input vector's encoding its *sparse representation*.

Sparse coding has been broadly applied in statistical signal processing Bruckstein et al. (2009), modeling brain states Olshausen and Field (1997), and in extracting meaningful features for a wide range of image processing applications Elad (2010). Beyond its domain-specific merits, the sparsity assumption can help to extract low dimensional, hidden structure from otherwise very high dimensional data. Furthermore, sparsity restrictions can help offset overfitting and ensure a degree of robustness to noise.

Sparse coding was first formalized by Olshausen and Field in Olshausen and Field (1997), and we adopt their notational conventions in the present work. Given p data vectors $y^{(1)}, \dots, y^{(p)} \in \mathbb{R}^n$, the task is to extract a collection of m dictionary vectors $A_1, \dots, A_m \in \mathbb{R}^n$ and sparse coefficient vectors $x^{(1)}, \dots, x^{(p)}$ such that

$$y^{(j)} \approx Ax^{(j)} \text{ for all } j \in \{1, \dots, p\} \tag{1.4}$$

where $A \in \mathbb{R}^{n \times m}$ is the dictionary matrix whose i-th column is A_i. Olshausen and Field propose an alternating gradient descent heuristic to learn the non-convex objective

$$F(A, X) := \sum_{j=1}^{p} \|y^{(j)} - Ax^{(j)}\|^2 + \sum_{j=1}^{p} \lambda(x^{(j)}) \tag{1.5}$$

where the regularization function λ encourages the $x^{(j)}$ to be sparse, and X is the matrix whose j-th column is $x^{(j)}$. Despite the problem's non-convexity, alternating minimizing procedures like the algorithm proposed by Olshausen and Field (1997), and similar heuristics procedures in Aharon et al. (2006) and Kreutz-Delgado et al. (2003), have been found to work remarkably well in practice.

Towards Provable Algorithms for Dictionary Learning

Recent research in sparse coding has sought to design *provably correct* algorithms for recovering a hidden dictionary A^* by specifying a generative model for the sparse representations $x^{(j)}$. In the under-complete ($n \leq m$) setting, Spielman et al. (2013) recovers A^* as long as the sparsity of x is no more than approximately $\sqrt{n}$. The independent works Arora et al. (2014) and Agarwal et al. (2013a) provide polynomial time algorithms to recover $\mu/\sqrt{n}$-incoherent dictionaries (defined in Section 1.2) based on overlapping community detection, but shed little light on the startling efficacy of gradient descent procedures. Barak et al. (2014) proposes a sum-of-squares approach which succeeds for a sparsity up to $n^{1-\gamma}$ for some fixed γ, but their algorithm runs exponentially in the desired precision. Furthermore, Luh and Vu (2015) introduce an efficient convex problem to recover square-dictionaries $A^* \in \mathbb{R}^{n \times n}$ using an almost information-theoretically optimal number of samples. However, their algorithm succeeds only in the very limited setting where the supports of the sparse representation x are i.d.d. Bernoulli, in the sense that $\Pr(x_i \neq 0) \sim \text{Bernoulli}(p)$.

Recently, Arora et al. (2015) introduce a simple gradient descent procedure which learns the true dictionary A^* up a columwise error of $\sqrt{k/n}$ using $\tilde{O}(mk)$ samples per round, once initialized with an estimate A for which $\|A - A^*\| \leq 2$, and each column of A has distance less than $1/C \log n$ from a corresponding column of A^*, where C is a suitably large constant. The algorithm assumes that A^* is $\mu/\sqrt{n}$, and succeeds as long as the sparsity is no more than roughly $\mu/\sqrt{n}$. Furthermore, Arora et al. (2015) provide a descent algorithm that

learn A^* up to an arbitrary inverse polynomial columnwise error, using only polynomially many samples. Both algorithms converge at geometric rates. The descent algorithms are complemented by a provably correct initialization procedure requiring $\tilde{O}(m^2/k^2)$ samples.

1.1.2 Dictionary Learning and ρ-Smoothness

Both the community detection algorithms in Agarwal et al. (2013b) and Arora et al. (2014), and the coordinate descent scheme in Agarwal et al. (2013a) and Arora et al. (2015) require that the nonnegative entries of the coefficient vectors x satisfy what we will refer to as the *Lower Boundedness Asumption*[1]:

Definition 1.1.1. We say that a real valued random variable Z is C-lower bounded if there exists some constant C for $|Z| \geq C$ almost surely whenever $Z \neq 0$. We say that a random vector x is C-lower bounded if $|x_i| > C$ for all $i \in \text{supp}(x)$, almost surely.

We can think of imposing the lower boundedness assumption on the sparse latent samples x in $y = A^*x$ as postulating some minimal signal strength necessary to "activate" a given column of A^*. This assumption holds some weight in, say, modeling a biological neural network, where neurons only respond to electrical impulses above a certain threshold voltage. However, in the many other applications to which Dictionary Learning has been applied, it is unclear that we can demand that our data satisfy such a restrictive condition. Indeed, the Lower Boundedness Assumption would rule out sparse vectors for which nonzero entries are independent, identically distributed, and Gaussian.

In this thesis, we present a framework for gradient-like algorithms which can make progress even when the lower boundedness assumption fails to hold. Rather than assuming that the entries of x are bounded below in magnitude, we will instead assume that there is some constant C below which the x_i have a smooth distribution, in the following sense:

Definition 1.1.2 ((C, ρ)-smoothly distributed)**.** Given $C > 0$ and $\rho > 0$, we say that a real valued random variable Z is (C, ρ) smoothly-distributed if there exists a constant C such that

$$\Pr(Z \in S - \{0\}) \leq \rho \cdot \text{vol}(S) \quad \forall S \in \mathcal{F}([-C, C]) \tag{1.6}$$

where $\mathcal{F}([-C, C])$ denotes the set of all Borel sets supported on the interval $[-C, C]$. We say that Z is ρ-smoothly distributed if it is (C, ρ)-smoothly distributed for any C. We say that a vector x is (C, ρ)-(resp. ρ-)smoothly distributed if its entries are (C, ρ)-(rep. ρ-)smoothly distributed conditioned on their support.

It is easy to see that if the entries of x satisfy the Lower Boundedness assumption with constant C, then then are $(C, 0)$-smoothly distributed. On the other hand, if the entries of x have a continuous density $p(x)$ conditioned on their support, with say $\sup_{x \in \mathbb{R}} p(x) = \rho$, then the entries of x are ρ-smoothly distributed. Hence, the class of (C, ρ) smoothly distributed random variables strictly and substantially generalizes the class of those which satisfy the lower boundedness assumption.

[1] While Luh and Vu (2015) and Spielman et al. (2013) do not impose such a restriction, their require that the supports of x be uniform or Bernoulli. In the present work, we can handle a more general distribution on the support of x, specified in Assumption 2

1.2 Notation and Assumptions

1.2.1 General Notation

Throughout this report, we will adopt the following notational conventions.

- For $m \in \mathbb{N}$, let $[m]$ denote the set $\{1, \dots, m\}$. We let $\binom{[m]}{i}$ be the set of all subsets consisting of i distinct elements of $[m]$, and $2^{[m]}$ be the set of all subsets of $[m]$.
- For a vector $v \in \mathbb{R}^d$, we use $\|v\|$ and $\|v\|_2$ to refer to the standard l_2 norm, $\|v\|_p$ to denote l_p norm. Given two vectors $v, w \in \mathbb{R}^d$, we will denote their inner product by either $\langle v, w \rangle$ or $v^T w$.
- For a matrix $M \in \mathbb{R}^{n \times m}$, we denote its spectral norm by $\|M\|$, its Frobenius norm $\|M\|_F$, and let $\|M\|_{l_p}$ denote its l_p norm viewed as a vector in $\mathbb{R}^{nm}$. We denote M_i to be the i-th column of M, and for a set $S \subset [m]$, we let M_S be the submatrix M formed by the columns of M which are indexed by the elements of S. Note that $\|M_S\| \le \|M\|$. Where ambiguous, we will interpret M_S^T as $(M_S)^T$.
- We denote the Moore-Penrose pseudoinverse of M by $M^\dagger$ (cite), and the orthogonal projection onto the column space of M by Proj_M.
- Given a set of elements $\{x_j\}_{j \in [n]}$, we set $\mathrm{vec}(x_j)$ to be the vector in $\mathbb{R}^n$ whose j-th element is x_j, and $\mathrm{diag}(x_j)$ to be the diagonal matrix in $\mathbb{R}^{n \times n}$ whose (j, j)-th element is x_j.
- Given a random event A, $\Pr(A)$ denotes the probability of A, and $\Pr(\cdot \big| A)$ and $\mathbb{E}[\cdot \big| A)$ denote the probability and expectation operators, conditional on A. For two events A and B, we let $A \wedge B$ be their disjunction, and $A \vee B$ denote their union.
- For a convex set $\mathcal{K} \subset \mathbb{R}^n$, we denote $\mathrm{Proj}_{\mathcal{K}}(z) = \min_{w \in \mathcal{K}} \|z - w\|$

There are also define the following functions which we will use throughout the present work:

- $\mathbb{1}(\cdot)$ denotes the indicator function
- $\mathrm{sign}_\tau(u) := \mathrm{sign}(u)\mathbb{1}(|u| > \tau)$, where τ is nonnegative and $u \in \mathbb{R}$
- For $v \in \mathbb{R}^n$, $\mathrm{Normalize}(v) = \frac{v}{\|v\|}$. For $M \in \mathbb{R}^{n \times m}$, $\mathrm{Normalize}(M) = M\mathrm{diag}(\frac{1}{\|M_i\|})$.
- $\mathrm{thres}_\tau(u) := \mathbb{1}(|u| > \tau)$, where τ is nonnegative and $u \in \mathbb{R}$
- $\mathrm{Thres}_\tau(u) := u\mathbb{1}(|u| > \tau)$, where τ is nonnegative and $u \in \mathbb{R}$

Asymptotic Notation

We adopt the standard conventions in computer science for asymptotic notation:

- We say that $f(n) = \mathrm{O}(g(n))$ if $f(n) \le Cg(n)$ for some constant $C > 0$, that $f(n) = \Omega(g(n))$ if $f(n) \ge Cg(n)$ for some constant $C > 0$, $f(n) = \Theta(g(n))$ if there are $C_1 g(n) \le f(n) \le C_2 g(n)$ for constant $C_1, C_2 \ge 0$.

- We write $f(n) = \mathrm{o}\left(g(n)\right)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$, and $f(n) = \omega\left(g(n)\right)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$
- We also use soft asymptotic notation: $f(n) = \tilde{\mathrm{O}}\left(g(n)\right)$ if $f(n) \le C\log^c(n)g(n)$ for some constants $C > 0$ and $c \in \mathbb{R}$, and $f(n) = \tilde{\Omega}\left(g(n)\right)$ if $f(n) \ge C\log^c(n)g(n)$ for some constants $C > 0$ and $c \in \mathbb{R}$.
- Following Arora et al. (2015), we let $f(n) = O^*(n)$ if $f(n) \le Cg(n)$ for a sufficiently small constant $C > 0$, and $f(n) = \Omega^*(n)$ if $f(n) \ge Cg(n)$ for a sufficiently large constant $C > 0$. Morover, we see that $f(n) = \Theta^*(g(n))$ if, for two specific constants C_1 and C_2, it holds that $c_1 g(n) \le f(n) \le c_2 g(n)$

Occasionally, the conventional CS asymptotic notation will prove burdensome and visually clumsy. Consequently, we introduce the following notation for approximate inequalities:

- We say that $f(n) \lesssim g(n)$ if $f(n) = O(g(n) + n^{-\omega(1)})$.
- We say that $f(n) \asymp g(n)$ if $f(n) = \Theta(g(n) + n^{-\omega(1)})$.

1.2.2 The Dictionary Learning Setup

Unless otherwise specified, we assume that we have access to random variables $y = A^*x$, where $x \in \mathbb{R}^m$ is a sparse coefficient vector, $A^* \in \mathbb{R}^{m\times n}$ is the true dictionary, and $y \in \mathbb{R}^n$ are our observations. We will frequently denote $S = \mathrm{supp}(x)$. We now place the following assumptions on x:

Assumption 1 (Properties of Sparse Coefficients)**.** *We assume that* $p_i := \mathbb{E}[|x_i|\big|i \in S] = \Theta(1)$ *for all* $i \in [m]$ *and that* $S = \mathrm{supp}(x)$ *satifies* $|S| \le k$ *almost surely. We also require that, for any* $S \in \binom{[m]}{k}$ *containing* i, $x_i\big|\mathrm{supp}(x) = S$ *is* $O(1)$*-subgaussian.*

Remark. The assumption $|S| \le k$ can easily be relaxed to the assumption $|S| \le k$ with probability at least $1 - n^{-\omega(1)}$. For example, if $k = \Omega\left(\log^2 n\right)$, entries of S are chosen independently with probability $p \le \frac{k}{m(1+c)}$ for any constant $c > 0$, then a multiplicative Chernoff Bound ensures that $|S| \le k$ with $1 - n^{-\omega(1)}$ probability.

Unlike Luh and Vu (2015), Agarwal et al. (2013a) and Spielman et al. (2013), we shall not require that the supports of x be chosen exactly uniformly. Instead, we require only that the supports on x are sufficiently even, in the following sense:

Assumption 2. *We assume* $q_i := \Pr\left(i \in S\right) = \Theta\left(k/m\right)$*. Furthermore, we require that threewise correlations are bounded, in the the sence that for all* $\{i,j,r\} \in \binom{[m]}{3}$*, that* $q_{i,j,r} := \Pr\left(\{i,j,r\} \subset S\right) = \mathrm{O}\left(k^3/m^3\right)$*. By marginalizing, it follows that* $q_{i,j} := \Pr\left(\{i,j\} \subset S\right) = \mathrm{O}\left(k^2/m^2\right)$*.*

For the sake of brevity, we all results in this paper will be stated *in the absence of noise.* Where possible, we will remark how to extend some of the results to isotropic, subgaussian noise vectors which are independent of the coefficient vectors x. We will also impose another relatively strong assumption, and again, where possible, we will remark on when this assumption can be removed or adjusted:

Assumption 3. *We assume that the entries of x are symmetrically distributed, and independent conditioned on $S = \mathrm{supp}(x)$. Moreover, the distribution of $x_i \big| i \in S$ is independent for any S containing i.*

We will also assume that A^* is an incoherent dictionary, which we define below.

Definition 1.2.1. We say that a dictionary A^* is μ-incoherent if $|\cos(A_i^*, A_j^*)| \leq \mu/\sqrt{n}$ for all $i \neq j \in [m]$. Note that if the columns of A^* have unit norm, then A^* is μ-incoherent precisely when $|\langle A_i^*, A_j^* \rangle| \leq \mu/\sqrt{n}$.

Incoherence is quite reasonable to impose, since it is satisfied by many common signal processing filters (see Elad (2010)), and by random dictionaries with high probability Candès and Wakin (2008). From the standpoint of algorithm analysis, incoherence works in tandem with sparsity to show that our model is identifiable. Indeed, if $\frac{\mu}{\sqrt{n}} \leq \frac{1}{2k}$, then an application of the Gershgorin Circle Theorem esures that any k-columns submatrix of A^* are well conditioned, and that any subset of $2k$ column-submatrix of A^* are linearly independent (see Lemma D.1.1). Consequently, if $y = A^*x$ for a $k \leq \frac{2\sqrt{n}}{\mu}$-sparse x, one can show that x is the unique k-sparse for which $y = A^*x$. This is observation serves as the cornerstone of Compressed Sensing and Sparse Recovery, and we direct the curious reader to Candes (2008) and Candes and Tao (2006) for further reading.

In the present work, we will assume that the columns of A^* are normalized to unit norm (this can be imposed without loss of generality by rescaling the distributions of the entryies of the coefficient vectors x). With the exception of Chapter 4, we will establish most of our results in the *overcomplete case*, where $n \leq m \leq n^2$. We summarize the key properties of A^* below:

Assumption 4 (Properties of an Overcomplete A^*)**.** *The columns of A^* have unit norm and A^* is μ-incoherent. Furthermore, $\|A^*\| = O\left(\sqrt{m/n}\right)$, where $n \leq m \leq n^2$. We also assume that $k = \omega(1)$, $k = \mathrm{o}(\sqrt{n})$, and that $k = O^*(\mu/\sqrt{n})$*

Remark. We remark that the assumptions in the present work are only slightly stronger than those in Arora et al. (2015), in that Assumption 2 requires controls on the 3-wise correlations in the distribution of the support, whereas Arora et al. (2015) only imposes assumptions on first- and second- order correlations. We also note that, while Arora et al. (2015) only states the assumption that the entries of x are *pairwise independent* conditioned on their support, the analysis appeals to concentration arguments which necessitate that the entries of x are jointly independent as well (again, after conditioning on the support).

A Measure of Closeness

Finally, we will measure the distance between two dictionaries by the distance between their columns, up to suitable permutations. We recall Definition 8 in Arora et al. (2015):

Definition 1.2.2. $[(\delta, \kappa)-$near$]$ We say that A is δ-close to A^* if there is a permutation $\pi : [m] \to [m]$ and assignment of signs $\sigma : [m] \to \{-1, 1\}$ for which $\|\sigma(i)A_{\pi(i)} - A_i^*\| \leq \delta$ for all $i \in [m]$. We say that A is (δ, κ)-near to A^* if, in addition, $\|A - A^*\| \leq \kappa$.

Throughout the paper, we will assume that $\delta = O^*(1/\log n)$. Hence, the incoherence of A^* and the fact that its columns have unit norm will ensure that there is a unique permutation π and sign-assigment σ for which $\|\sigma(i)A_{\pi(i)} - A_i^*\| \leq \delta$ for all $i \in [m]$. Hence, we can unambiguously reindex the columns of A, and flip the signs of A^* if necessary to ensure that $\|A_i^* - A_i\| \leq \delta$ for all $i \in [m]$.

1.2.3 γ-Notation and High Probability Events

Assumption 3 requires that the entries x_i of the coefficient vectors to be $O(1)$ sub-gaussian. By Lemma C.1.1, this implies that the tails of x_i decay super-polynomially, in the sense that there exists some $\sigma^2 = O(1)$ for which

$$\Pr\left(|x_i| \geq \sigma\sqrt{\log(1/\delta)}\right) 1 - \delta \text{ for all } \delta > 0 \tag{2.7}$$

As a consequence, it is straightforward to demonstrate that vector $y = Ax$ will also satisfy some sort of super-polynomial decay: that is, there will exist an $R = O(\text{poly}(\|A\|, m, n, k))$ which is a low-degree polynomial in n, and a constant $c > 0$ for which

$$\Pr(\|y\| \geq R(\log(1/\delta))^c) \geq 1 - \delta \text{ for all } \delta > 0 \tag{2.8}$$

In fact, it is routine to verify that essentially every random quanity Z encountered in this report will satisfy a tail bound similar to Equation 2.8 for some constant $c > 0$ and $R = O(\text{poly}(n))$. Hence, we will liberally use the following property of random variables with super-polynomial decay, proved formally in Proposition C.1.4:

Claim 1.2.1. *Let $C, C', c > 0$ be constants. If Z is a random variable for which*

$$\Pr(\|Z\| \geq R(\log(1/\delta))^c) \geq 1 - \delta \text{ for all } \delta > 0 \tag{2.9}$$

for $R \leq n^C$, then for any any variables X_1 and X_2 bounded above by $n^{(C')}$ almost surely such that $X_1 = X_2$ with probability $n^{-\omega(1)}$,

$$\|\mathbb{E}[Z \cdot X_1] - \mathbb{E}[Z \cdot X_2]\| \leq O(n^{-\omega(1)}) \tag{2.10}$$

In light of the above claim, we will follow Arora et al. (2015), in using the letter γ to denote quantities of norm no more than $n^{-\omega(1)}$.Hence, if Z, X_1, X_2 satisfy the conditions of Claim 1.2.1, and $X_1 = X_2$ with probability $n^{-\omega(1)}$, then we will write

$$\mathbb{E}[Z \cdot X_1] = \mathbb{E}[Z \cdot X_2] + \gamma \tag{2.11}$$

without comment. We will also use the phrase *with high probability* to refer to any event which occurs with probability at least $1 - n^{-\omega(1)}$.

1.3 Contributions

To our knowledge, this paper presents the first analysis of provably correct and computationally efficient algorithms for learning incoherent dictionaries under generative processes

where *neither* the non-zero coordinates of the sparse coefficients are $\Omega(1)$ in magnitude, *nor* are the supports chosen in a uniform fashion.

After reviewing the approximate coordinate descent framework presented in Arora et al. (2015), Chapter 2 establishes a Dictionary Learning Meta-Algorithm which encapsulates a broad class of coordinate descent rules. Chapter 3 then presents two coordinate descent algorithms for the learning dictionaries with (C, ρ)-smooth coefficient vectors, and we summarize their properties here:

Theorem 1.3.1. *If the coefficient vectors are (C, ρ)-smooth for $C = \Omega(1)$, then there is an algorithm which, when initialized with an estimate A^0 that is $(\delta, 2)$-near to A^* for $\delta = \mathrm{O}^*(1/\log n)$, and given $\tilde{\mathrm{O}}(mt)$-samples per-step, converges at a geometric rate to A^* until the column-wise error is $\mathrm{O}\left(1/\sqrt{t} + \rho k/n\right)$, as long as $t = \Omega(k)$. The run-time is $\mathrm{O}(mnp)$. Moreover, the algorithm can be parallelized in such a way that each server only needs to store one column of A at any given time.*

We call the algorithm in the above theorem the "Toy Rule", and the algorithm in the subsequent theorem the "Neural Rule":

Theorem 1.3.2. *If the coefficient vectors are (C, ρ)-smooth for $C = \Omega(1)$, then there is an algorithm which, when initialized with an estimate A^0 that is $(\delta, 2)$-near to A^* for $\delta = \mathrm{O}^*\left((\rho \log nk^{1/4})^{-1}\right)$, converges at a geometric rate to A^* until the column-wise error is $\mathrm{O}(\rho k/n + \mu/\sqrt{n})$. The run-time is $\mathrm{O}(mnp)$, where the algorithm uses $p = \tilde{\mathrm{O}}\left(mk^2\right)$ samples per step.*

The second algorithm we present is a slight modification of the Neural Update rule in Arora et al. (2015), which we believe also corrects a flaw in their analysis. We remark that the sample complexity suffers by a factor of k when transitioning from the C-lower bounded to the C-ρ-smooth setting. In the case where $k \ll \mu/\sqrt{n}$, we can use the first algorithm to initialize the second. This establishes that:

Theorem 1.3.3. *If the coefficient vectors are (C, ρ)-smooth for $C = \Omega(1)$, then there is a two-stage coordinate descent algorithm which, when initialized with an estimate A^0 that is $(\delta, 2)$-near to A^* for $\delta = \mathrm{O}^*(1/\log n)$, returns an estimate of A^* with column-wise error $\mathrm{O}(\rho k/n + \mu/\sqrt{n})$. Each step converges geometrically, has runtime $\mathrm{O}(mnp)$, and uses fewer than $p = \tilde{\mathrm{O}}\left(mk^2\right)$ samples.*

It also turns out that the initialization algorithm in Arora et al. (2015) successfully returns estimates within the radius of convergence of the Toy Algorithm described in Theorem 1.3.1 in the (C, ρ)-smooth context as well. We will not adress this point in further detail in the present work.

Following the analytic framework in Chapter 3, Chapter 4 introduces a projection based update rule which has neglible bias. The projection rule was a radius of convergence $\mathrm{O}^*\left(1/\sqrt{k}\right)$, and hence can be initialized using either the Neural Rule or the Toy Rule:

Theorem 1.3.4. *If the coefficient vectors are C-Lower Bounded for $C = \Omega(1)$, then there is an algorithm, which when initialized with an estimate A^0 that is δ-close to A^* for $\delta = \mathrm{O}^*\left(1/\sqrt{k}\right)$, and given $\tilde{\mathrm{O}}(m)$-samples per-step, converges at a geometric rate to A^* until the column-wise error is $\mathrm{O}\left(n^{-\omega(1)}\right)$.*

While reminiscent of the "unbiased-update rule" inArora et al. (2015), our Projection Rule has far superior sample complexity, requiring only $\tilde{\mathrm{O}}(m)$-samples per step in the noiseless setting to learn A^* up to arbitrary inverse polynomial accuracy (in the presence of noise, the sample complexity will grow to reflect the noise level).[2]

As an application of the anti-concentration framework introduced in this thesis, Chapter 5 analyzes algorithms for learning a class of random non-negative matrix factorization (NMF) instances that mimic the random NMF instances commonly used to benchmark gradient descent heuristics. While the exact NMF problem is NP-Hard general, alternating descent algorithms like the Multiplicative Updates Rule in Lee and Seung (2001) and Hierarchical Alternating Least Squares in Cichocki et al. (2007) seem to perform surprisingly well on certain randomly generated NMF instances Vavasis (2009).

Motivated by these experiments, we consider learning a Nonnegative Offset Incoherent Dictionary Learning or "NOID"[3] - an undercomplete matrix $B^* \in \mathbb{R}^{n\times m}$ which decomposes as

$$B^* = A^* + vc^T \quad \text{where } A^* \text{ is incoherent and } c \text{ is nonnegative} \tag{3.12}$$

from samples $y = B^*x$, where the coefficient vectors x are sparse, entrywise nonnegative, ρ-smooth, and satisfy the regularity assumptions laid out in Definition 5.1.5. Appealing to standard results about the incoherence of random matrices with mean zero, i.i.d. entries (see Proposition C.1.10, or Candès and Wakin (2008)), it is straightforward to verify that random matrices with non-negative i.i.d. entries are NOIDs with high probability.

Theorem 5.1.3 demonstrates a reduction from NOID Learning to Semi-Nonnegative Dictionary Learning (S-NDL), in which one learns an incoherent, undercomplete dictionary from nonnegative sparse coefficient vectors. Leveraging the ρ-smoothness property, Theorem 5.2.1 proves that Algorithm 8 learns sufficiently incoherent dictionaries in the S-NDL setting up to arbitrary inverse polynomial error, once suitably initialized. We complement this result with an initialization scheme borrowed from Arora et al. (2014), and hence Theorem 5.2.3 establishes that there is a polynomial time algorithm to recover dictionaries in the S-NDL setting under suitable sparsity restrictions and distributional assumptions, starting with only random samples. Finally, by feeding our S-NDL algorithm into the reduction from NOID Learning, we conclude with the following theorem:

Theorem 1.3.5 (Tractability of NOID Learning, Stated Roughly)**.** *Under reasonable sparsity restrictions and distributional assumptions, there is an algorithm which can learning a NOID B^* up to a Frobenius norm error of $\delta\|B^*\|_F$ with run-time and sample complexity on the order of poly$(n, \log(1/\delta))$.*

Theorem 5.1.1 states the above result in more precise language.

[2] It was the author's intention to also include guarantees for projection-based algorithms in the (C,ρ)-smooth setting. However, the analysis turned out to be quite involved, difficult to follow, and only reduces the systematic error in the Neural Rule by a factor of roughly $\sqrt{k/n}$. If the reader of this present work is curious about this sort of analysis, he or she may contact the author for a roughly edited sketch of that result.

[3] See Definition 5.1.3

Chapter 2

Approximate Gradient Descent for Dictionary Learning

2.1 Approximate Gradient Descent

In this section, we describe a generic framework for learning the optimal solutions to possibly non-convex problems, introduced independently in both Arora et al. (2015) and Candes et al. (2014). This section is mostly expository in flavor, and can be skimmed once the reader is comfortable with the main results and definitions.

As a motivating example, suppose that we wish to minimize a known convex function $f(\cdot)$ over a convex set $\mathcal{K}$, such that $\arg\min_{z\in\mathcal{K}} f(z) = z^*$. If we can evaluate the gradients $g(\cdot) := \nabla f(\cdot)$, one of the most popular, and perhaps the simplest minimization strategies is just *projected gradient descent:* start with an initial guess z^0, and update each successive iterate $z^{s+1} \leftarrow \text{Proj}_{\mathcal{K}}(z^s - \eta^s g(z^s))$ for appropriate step sizes η^s.

2.1.1 Review From Convex Analysis

Let's now assume further that f is β-smooth and α-strongly convex: that is, $f(x) - \frac{\alpha}{2}x^2$ is convex, f is continuously differentiable, and, given two points $x, y \in \mathbb{R}^n$, $\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|$. Then, a standard lemma in convex optimization shows that, for any $z \in \mathbb{R}$ the gradient of $f(z)$ points strongly in the direction of the optimal z^*

Lemma 2.1.1. *Bubeck (2014) Let f be a β-smooth, α-strongly convex function. Then*

$$\langle \nabla f(z), z - z^* \rangle \geq \frac{\alpha}{2}\|\nabla f(z)\|^2 + \frac{1}{2\beta}\|\nabla f(z)\|^2 \tag{1.1}$$

Remark. The archetypical 1-strongly convex, 1-smooth function is just $f(z) = \frac{1}{2}\|z - z^*\|^2$. We see that $\nabla f(z) = 2(z - z^*)$, so the gradient points exactly in the direction of the optimal solution. In this case, it is trivial to verify that Lemma 2.1.1 holds with $\alpha = \beta = 1$.

More generally, if f is twice continuously differentiable, then f is α-strongly convex and β-smooth as long as $\alpha I \preceq \nabla^2 f \preceq \beta I$. By taking a Taylor Expansion, we can think of f as locally resembling the sum of a linear function and well-conditioned quadratic form. Thus, up to a linear term, we can think of the strong convexity and smoothness properties of f as

measuring how closely f resembles a (possibly scaled) Euclidean distance. Indeed, if $\alpha = \beta$, then integrating twice shows that $f(z) = \frac{\alpha}{2}\|z\|^2$.

From Lemma 2.1.1, it is easy to prove the following guarantee for gradient descent:

Proposition 2.1.2 (Theorem 3.6 in Bubeck (2014)). *Let $\eta = \frac{1}{\beta}$. Then projected gradient descent algorithm $z^{s+1} \leftarrow \mathrm{Proj}_{\mathcal{K}}(z^s - \eta g(z^s))$ satisfies*

$$\|z^s - z^*\|^2 \quad \leq \quad (1 - \frac{\alpha}{\beta})^s \|z^0 - z^*\|^2 \tag{1.2}$$

Proof. Using Lemma 2.1.1 and the facts that Euclidean distances are non-increasing under projections onto convex sets, we have

$$\begin{aligned}
\|z^s - z^*\|^2 &= \|\mathrm{Proj}_{\mathcal{K}}(z^{s-1} - \frac{1}{\beta}\nabla(z^{s-1})) - z^*\|^2 \\
&\leq \|z^{s-1} - \frac{1}{\beta}\nabla(z^{s-1}) - z^*\|^2 \\
&\leq \|z^{s-1} - z^*\|^2 - \frac{2}{\beta}\nabla f(z^{s-1})^T(z^{s-1} - z^*) + \frac{1}{\beta^2}\|\nabla f(z^{s-1})\| \qquad (1.3) \\
&\leq (1 - \frac{\alpha}{\beta})\|z^{s-1} - z^*\|^2 \\
&\leq (1 - \frac{\alpha}{\beta})^s \|z^0 - z^*\|^2
\end{aligned}$$

□

2.1.2 Generalizing Gradient Descent

Upon examining the proof of Proposition 2.1.2, we notice that we never appealed to the convexity properties of f directly. Instead, we simply require that the gradient of f at each iterate z^s pointed roughly in the same direction as $z^s - z^*$. Thus, if our actual goal is to compute an accurate estimate of some vector z^* over a convex set $\mathcal{K}$, we can think of running gradient descent a smooth and strongly convex function f as a convenient proxy for minimizing the Euclidean Distance towards z^* (see Remark 2.1.1 for further discussion).

However, the proof of Proposition 2.1.2 suggests that all we really need is to run gradient descent with gradient vectors which satisfy a similar relation as the one in Lemma 2.1.1. We generalize this relation here

Definition 2.1.1 (Definition 5 in Arora et al. (2015)). Given a gradient descent iterate z^s, we say that vector g^s is $(\alpha, \beta, \epsilon)$-correlated with a desired solution z^* if

$$\langle g^s, z^s - z^* \rangle \geq \alpha\|z_s - z^*\|^2 + \beta\|g^s\|^2 - \epsilon_s \tag{1.4}$$

We say that a random vector g^s is $(\alpha, \beta, \epsilon)$-correlated-whp with a desired solution z^* if Equation 1.4 holds with probability $1 - n^{-\omega(1)}$,

Remark. Note that if g^s is $(\alpha, \beta, \epsilon)$-correlated with a desired solution z^*, then it is also $(\alpha', \beta', \epsilon'_s)$-well correlated for all $\alpha' \leq \alpha$, $\beta' \leq \beta$, and $\epsilon'_s \geq \epsilon_s$

It is easy to see from Lemma 2.1.1 that the gradients of $\nabla f(z^s)$ are $(\alpha/2, \frac{1}{2\beta}, 0)$-correlated with $z^* = \arg\min_{z\in\mathcal{K}} f(z)$ if f is β-smooth, α-strongly convex.

But the power of this definition lies in versatility. Indeed, the vectors g^s need not be the gradients of any convex function, and the ϵ paramater allows us to analyze gradient descent schemes in a framework that both tolerates, and makes precise the dependence of the algorithm on a systemic error ϵ. Specifically, we have the following theorem:

Theorem 2.1.3 (Adaptation of Theorems 6 and 40 in Arora et al. (2015))**.** *Let $\mathcal{B}$ be any convex set containing z^*. Suppose that for steps $s = 1, \dots, T$, the vector g^s is $(\alpha, \beta, \epsilon_s)$-correlated with a desired solution z^*. Then given an initial vector z^0, $\eta \in [0, 2\beta]$, and update rule $z^{s+1} = \mathrm{Proj}_{\mathcal{B}}(z^s - \eta g^s)$, it holds that*

$$\|z^{s+1} - z^*\|^2 \le (1 - 2\alpha\eta)\|z^s - z^*\|^2 + 2\eta\epsilon_s \tag{1.5}$$

for all $s \in [T]$. Consequently, for all $s \in [T]$

$$\|z^s - z^*\|^2 \le (1 - 2\alpha\eta)^s\|z^0 - z^*\|^2 + \frac{2}{\alpha}\max_{s\in[T]}\epsilon_s \tag{1.6}$$

In particular, if $\epsilon_s \le \alpha \cdot \mathrm{O}^*\left((1 - 2\alpha\eta)^s\|z^0 - z^*\|^2\right) + \epsilon$, then the updates converge geometrically to z^* with systemic error ϵ/α in the sense that

$$\|z^s - z^*\|^2 \le (1 - \alpha\eta)^s\|z^0 - z^*\|^2 + \frac{\epsilon}{\alpha} \tag{1.7}$$

If the vector g^s is $(\alpha, \beta, \epsilon_s)$-correlated-whp with a desired solution z^*, then as long $T \le$ poly(n), Equation 1.5 and Equation 1.6 hold simultaneously for all $s \in [T]$ with very high probability. Furthermore, if $\epsilon_s \le \alpha \cdot \mathrm{O}^*\left((1 - \alpha\eta)^s\|z^0 - z^*\|^2\right) + \epsilon$ with high probability, then Equation 1.7 holds with high probability as well.

A More Intuitive Characterization of Well-Correlatedness [1]

Establishing that that an arbitrary vector g^s is well-correlated with a solution z^* might seem rather opaque and unintuitive. Forturnately, the following lemma provides a very easy-to-grasp sufficient condition for well-correlatedness:

Lemma 2.1.4 (Lemma 15 in Arora et al. (2015))**.** *Suppose that $g^s = \alpha(z^s - z^*) + v^s$, where $\|v^s\| \le \frac{\alpha}{4}\|z^s - z^*\| + \zeta$, then g^s_i is $(\alpha/4, 1/25\alpha, 4\zeta^2/\alpha)$-correlated with z^*.*

Hence, if g^s satisfies the decomposition specified in Lemma 2.1.4 and the step size η is chosen appropriately, then Theorem 2.1.3 ensures that

$$\|z^s - z^*\|^2 \le (1 - \Omega(1))^s\|z^0 - z^*\|^2 + \mathrm{O}\left(\zeta^2/\alpha^2\right) \tag{1.8}$$

In other words, if g^s/α points *mostly in the direction of* $z^s - z^*$ - up to a systematic error of $\epsilon := \zeta/\alpha$ - then the iterates z^s will converge geometrically to z^* up to a systematic error of ϵ.

[1] This section was recently added to the present work after the oral defense, in order to make the approximate coordinate descent exposition more clear.

2.2 A Meta-Algorithm for Dictionary Learning

2.2.1 Review of Approach in Arora et al. (2015)

The high-level exposition of the sparse coding algorithms in Arora et al. (2015) draw analogies between approximate gradient descent in the dictionary learning problem, stochastic gradient descent and alternating minimization. Briefly, they imagine that that given N samples $y^{(1)} = A^* x^{(1)}, \dots, y^{(N)} = A^* x^{(N)}$, and attempt to compute the dictionary A which minimizes Equation 1.5.

If X^* is the matrix whose columns are the true sparse signals $x^{(1)}, \dots, x^{(N)}$, then computing A^* amounts to optimizing $F(A, X^*)$. Drawing analogies to stochastic gradient descent, casts the dictionary learning problem as the optimization of an *convex function* $F(A, X^*)$, as an *unkown convex function* of A. Their strategy is to alternate between moving along gradients of $F(A, \hat{X})$ with respect to A, and then refining estimates of the latent sparse signals by updating $\hat{X}$. The hope is that, as $\hat{X}$ gets closer to X^*, the gradients $\nabla_A F(A, \hat{X})$ will begin to resemble $\nabla_A F(A, X^*)$ more closely. This strategy motivates their Decode-and-Update alternating minimization approach for dictionary learning:

Algorithm 1: Alternating Minimization Framework in Arora et al. (2015)

Initialize A^0;
for $s = 1, 2, \dots, T$ **do**
  Decode: **for** $i = 1, 2, \dots, p$, **do**
    Find an approximate sparse solution to $A^s \hat{x}^{(i)} = y^{(i)}$;
    Set $\hat{X}^s$ to be the matrix whose i-th columns is A^s ;
  Update: $A^{s+1} = A^s - \eta g^s$, where $g^s = \nabla_A F(A^s, \hat{X}^s)$

2.2.2 A Meta-Algorithm Without Decoding

In this work, we still follow Arora et al. (2015) by learning A^* by simultaneously learning m vectors $A_1^*, \dots, A_m^*$ in $\mathbb{R}^n$ by approximate gradient descent. However, we bypass both the putative objective function and the analogy to alternating minimization entirely, and simply focus on designing sample dependent functions $f(y; A^s)$ such

$$g_i^s \approx \mathbb{E}[f(y; A^s)] \approx \alpha(A_i^* - A_i) \tag{2.9}$$

for some $\alpha > 0$. In light of Lemma 2.1.4, the *negative* of such gradients g_i^s will be correlated with A_i^*, and so the updates $A_i^{s+1} = A_i^s + \eta g_i^s$ will converge geometrically to A_i^*, up to some systematic error. We remark here that it is crucial to *add* g_i^s is A_i^s; Arora et al. (2015) in fact commits a sign error by subtracting g_i^s from A_i^s.

Though this exposition foregoes making connections to the convex optimization, encoding-decoding, alternating minimization, we believe it has three main advantages:

1. It does not require that we explicitly decode the coefficient vectors. This enables us to consider wider variety of gradient estimation procedures which do not have a clear interspects as the gradient of the objective function in Equation 1.5

2. It renders the analytical techniques in this paper more transparent.

3. It reinforces the generality of the approximate gradient descent framework for non-convex optimization and parameter estimation.

To compute such gradients, we introduce the *sign-thresholding* operation and the *projector matrix*. We can think of sign-thresholding as a sort of discretized decoding algorithm: Given a sample $y = A^*x$, threshold τ, and estimate A of the true dictionary A^*, the sign-thresholding operation computes estimates of the sign of the sparse representation x^* by regarding all coordinates x_i for which $|A_i^T y| \leq \tau$ as zero. Sign-thresholding is defined formally in Algorithm 2.

Algorithm 2: SignThreshold (A, y, τ)

Input: Dictionary A, Thresholds τ, Sample Sizes y

Initialize $\hat{\text{sign}}(x) \leftarrow 0 \in \mathbb{R}^m$, $\hat{S} \leftarrow \emptyset$ **for** $i = 1, 2, \ldots, m$ **do**

  $\hat{\text{sign}}(x)_i \leftarrow \text{sign}_\tau(\langle A_i, y\rangle)$

  $\hat{S} \leftarrow \hat{S} \cup \mathbb{1}(|\langle A_i, y\rangle| > \tau)$

Return $(\hat{S}, \hat{\text{sign}}(x))$

We define a *projector matrix* M as any collection of m matrix valued functions $M_i : \mathbb{R}^{n\times m} \times 2^{[m]} \to \mathbb{R}^{n\times m}$. Given an estimate A of the dictionary A^*, and a sample y with estimated support $\hat{S}$, we want to choose M such that $M_i(\hat{S}, A)$ to return a matrix for which

$$M_i(\hat{S}, A) \approx A_i^* - A_i \tag{2.10}$$

Indeed, if $\|A_i^*\| = \|A_i\| = 1$, then the projection onto the orthogonal complement of A_i is precisely $I - A_i A_i^T$, and satisfies

$$\left(I - A_i A_i^T\right) A_i^* = A_i^* - \left(A_i^T A_i^*\right) A_i = A_i^* - A_i + O\left(\|A_i - A_i^*\|^2\right)$$

In general, we cannot guarantee that $\|A_i\| = 1$ and so we will not pick M_i to be an exact orthogonal projection; this is rememdied in the Toy Rule and Neural Update Rules described in the following section. However, we still want to preserve the intutition that M should roughly resemble a projection; hence the title "projector matrix".

With these definitions taken care of, we now formally define a Meta-Algorithm for Dictionary Learning that generalizes both the Neural Update Rule and the Unbiased Update Rule presented in Arora et al. (2015):

Algorithm 3: Meta-Algorithm (M)

Input: Projector Matrix M, Initial estimate A_0, , step size η, Number of Iterations T, Thresholds τ^s, Sample Sizes p

for $s = 1, 2, \ldots, T$ **do**

  Estimate Supports: $(\hat{S}^{(j)}, \hat{\sigma}^{(j)}) \leftarrow \text{SignTreshold}(y^{(j)}, A^s, \tau^s)$ for $j = 1, 2, \ldots, p$

  Update: $A_i^{s+1} = A_i^s + \eta \hat{g}_i^s$ where $\hat{g}_i^s = \frac{1}{p}\sum_{j=1}^p M_i(\hat{S}^{(j)}, A) y^{(j)} \cdot \hat{\sigma}_i^{(j)}$

In section A.1.1 in the appendix, we establish that our Meta-Algorithm strictly generalizes the neural update rule in Arora et al. (2015). Without diving too far into the details

of any particular instantiation just quite yet, we provide some basic intuition for why the Meta-Algorithm should work, at least when we can ensure that $\text{sign}_\tau((A_i^T y) = \text{sign}(x_i)$ with high probability. To this end, let's compute the expectation of g_i^s. To lighten the notation a bit, we temporarily drop the dependence on the superscripts s:

Lemma 2.2.1. *Let $y = A^* x$, and suppose that the sign thresholding returns a support estimate $\hat{S} = S$ and sign estimate $\text{sign}_\tau((A_i^T y) = \text{sign}(x_i)$ with probability $1 - n^{-\omega(1)}$. Then, for projector matrix $M_i = M_i(\hat{S})$ for which $\|M\| = O(m)$ almost surely, then*

$$\mathbb{E}[M_i(\hat{S}, A) y \text{sign}_\tau(A_i^T y)] = p_i q_i \mathbb{E}[M_i(S) \big| i \in S] A^*{}_i + \gamma \tag{2.11}$$

Proof. Given a $y = A^* x$. By assumption the random variables $M_i(S)\text{sign}(x_i)$ and $M_i(\hat{S})\text{sign}_\tau(A_i^T y)]$ differ with probability $1 - n^{-\omega(1)}$. With our γ-notation, we can write

$$\begin{aligned} \mathbb{E}[M_i(\hat{S}) y \text{sign}_\tau(A_i^T y)] &= \mathbb{E}[M_i(S) y \text{sign}(x_i)] \pm \gamma \\ &= q_i \cdot \mathbb{E}[M_i(S) y \text{sign}(x_i) \big| i \in S] \pm \gamma \end{aligned}$$

Next,

$$\begin{aligned} \mathbb{E}[M_i(S) y \text{sign}(x_i) \big| i \in S] &= \mathbb{E}[M_i(S) A_i^* x_i \text{sign}(x_i) \big| i \in S] \\ &+ \mathbb{E}[M_i(S) \sum_{j \neq i \in S} A_i^* x_j \text{sign}(x_i) \big| i \in S] \end{aligned}$$

where the second term has mean zero since $\mathbb{E}[x_j \text{sign}(x_i)] = 0$. Consequently,

$$\begin{aligned} \mathbb{E}[M_i(\hat{S}) y \text{sign}_\tau(A_i^T y)] &= \mathbb{E}[M_i(S) A^*{}_i x_i \text{sign}(x_i)] \pm \gamma \\ &= p_i \cdot \mathbb{E}[M_i(S) |x_i| \big| i \in S] A_i^* \pm \gamma \\ &= \Pr(i \in S) \cdot \mathbb{E}[|x_i| \big| i \in S] \cdot \mathbb{E}[M_i(S) \big| i \in S] A_i^* \pm \gamma \\ &= p_i q_i \mathbb{E}[M_i(S) \big| i \in S] A_i^* \pm \gamma \end{aligned}$$

where the last step makes use of Assumption 3 that $\mathbb{E}[|x_i| \big| i \in S] = p_i$ for any set S. □

In other words, if we can estimate $\hat{S}$ and $\text{sign}(x_i)$ accurately with very high probability, the gradients g_i in the Meta Algorithm are roughly proportional $\mathbb{E}[M_i(S)] A_i^*$. Hence, if we can ensure that $M_i(S) A_i^* \approx A_i^* - A_i$, then we should expect a gradient based algorithm with a suitable step size to converge reasonably well to A_i^*.

Chapter 3

Update Rules for (C, ρ)-Smooth Samples

3.1 Introduction

In this section, we present and analyze two approximate gradient descent algorithms for sparse coding in the (C, ρ)-smooth setting. After explaining precisely the difficulties that the Meta-Algorithm from Section 3 encounters when the C-lower boundedness conditioned is removed, we will devote the remainder of the section to demonstrating how to leverage (C, ρ)-smoothness assumption in its place. To facilitate clarity, we proceed in order of increasing complexity of analysis.

In Section 3.1.3,we outlike an analytical framework which will essentially automate the convergence analyses of the update rules in the remainded of the chapter. Subsequently, section 3.1.4, we establish a few high probability properties of the thresholding operation that remain true *for general* (C, ρ)*-smooth distributions.*

Following these more general remarks, we analyze a variant of the Meta-Algorithm we call the "Toy Rule", where the projector matrix is choosen to be $M_i = I - \frac{1}{\|A_i\|} A_i A_i^T$ deterministically. The precise statement of its convergence is given in Theorem 3.2.1.

Just as imperfect sign thresholding can introduce correlations into our update rule, imperfect estimation of $\hat{S}$ can cause the projector matrix $M(\hat{S})$ to correlate with the entries of $x\big|\text{supp}(x)$. For an arbitrary matrix valued function $M(\cdot) : 2^{[m]} \to \mathbb{R}^n$, there is very little we can say about how these correlations will affect our update rule. Hence, analyzing an update with a deterministic matrix will prove much more straightforward, and serve to crystallize the key innovations in the present works analysis. Finally, we will analyze the expectated gradients under more sophisticated algorithms that afford stronger sample complexity guarantees when k is very small. In particular, we describe rule similar to the Neural Rule in Arora et al. (2015) whose convergence is described by Theorem 3.2.2

Throughout the chapter, we will take Assumptions 1, 2, 3, and 4 for granted. We also impose the following additional assumptions:

Assumption 5. *There is a* $C = \Omega(1)$ *for which* x *are* (C, ρ)*-smoothly distributed and* $\mathbb{E}[\Pr(|x_i| > C)\big|i \in S] \geq 1/2$*. We assume that* $\rho = \Omega(1)$*. Furthermore, we require that* $\rho \leq \sqrt{k}$ *and* $\rho k/\sqrt{n} = o(1)$*.*

Before continuiting, we also establish some further notation:

3.1.1 Further Notation

Conditional Expectation

First of all, we will often need to take expectations and probabilities conditioned on the support S of x. Hence, we will denote

$$\mathbb{E}_i[\cdot] = \mathbb{E}_i[\cdot \big| i \in \mathrm{supp}(x)] \tag{1.1}$$

and similarly $\mathbb{E}_{i,j}[\cdot] = \mathbb{E}_i[\cdot|i,j \in \mathrm{supp}(x)]$ and $\mathbb{E}_{i,j,r}[\cdot] = \mathbb{E}_i[\cdot|i,j,r \in \mathrm{supp}(x)]$. We will define the conditional probability operators $\Pr_i, \Pr_{i,j}, \Pr_{i,j,r}$ similarly. We will also define

$$\mathbb{E}_S = \mathbb{E}[\cdot \big| \mathrm{supp}(x) = S] \tag{1.2}$$

To avoid confusion, the notation $\mathbb{E}_S$ will only refer to the expecation conditioned on the support. If we want to take the expectation of all supports of x, we will write

$$\mathbb{E}_{S:S\subset[m]}[\cdot] = \sum_{S\subset[m]} \Pr(\mathrm{supp}(x) = S)\mathbb{E}[\cdot \big| \mathrm{supp}(x) = S] = \sum_{S\subset[m]} \Pr(\mathrm{supp}(x) = S)\mathbb{E}_S[\cdot] \tag{1.3}$$

More generally, when we want to take the expecation of all sets with a property $\mathcal{P}$, we write

$$\begin{aligned} \mathbb{E}_{S:\mathcal{P}(S)} &= \sum_{S:\mathcal{P}(S)} \Pr(\mathrm{supp}(x) = S|\mathcal{P}(\mathrm{supp}(x)))\mathbb{E}[\cdot \big| \mathrm{supp}(x) = S] &(1.4)\\ &= \sum_{S:\mathcal{P}(S)} \Pr(\mathrm{supp}(x) = S|\mathcal{P}(\mathrm{supp}(x)))\mathbb{E}_S[\cdot] &(1.5)\end{aligned}$$

For example, we have

$$\mathbb{E}_{S:i\in S}[\cdot] = \sum_{S:i\in S} \Pr_i(S)\mathbb{E}_S[\cdot] = \mathbb{E}_i[\cdot] \tag{1.6}$$

Gradient Notation

To lighten notation, we will drop the dependence on the iterate set A^s throughout the section. In keeping with this notation, we fix a step threshold $\tau > 0$ and define exact gradients

$$g_i := \mathbb{E}[M_i y \ \mathrm{sign}_\tau(A_i^T y)] = \mathbb{E}[\hat{g}_i] \tag{1.7}$$

where $\hat{g}_i$ is as given in Algorithm 3. We assume that A is $(\delta, 2)$-near to A^*, where $\delta = O^*(1/\log n)$, but may be smaller when specified. Throughout the section, we will also let $n_i = \|A_i\|^{-1}$ (think n for "normalize"). Because $\|A_i - A_i^*\| \le \delta = o(1)$, we have that $n_i = 1 \pm 2\delta$. Finally, we will define the matrix X_i by

$$X_i := 1 - n_i A_i A_i^T \tag{1.8}$$

The key property of X_i is summarized in the following claim:

Claim 3.1.1.

$$X_i A_i^* = A_i^* - A_i + O(\|A_i^* - A_i\|^2) A_i \tag{1.9}$$

Proof. We have

$$X_i A_i^* \;=\; (I - \frac{1}{\|A_i\|} A_i A_i^T) A_i^* \;=\; A_i^* - \langle \frac{A_i}{\|A_i\|}, A_i^* \rangle A_i$$

Now, $u := \frac{A_i}{\|A_i\|}$ and $v := A_i^*$ are unit vectors, so $\langle u, v \rangle = 1 + \|u - v\|^2$. Moreover, $\|u - v\| = \|A_i^* - \frac{A_i}{\|A_i\|}\| = O(\|A_i^* - A_i\|)$. Hence, $X_i A_i^* = A_i^* - A_i + O(\|A_i - A_i^*\|^2)$ □

Indexing Notation

For a sample $y = A^* x$, we will let $y_{-i} := y - A^*{}_i x_i$. Note that y_{-i} *does not* refer to the vector consisting of all entries of y except the i-th, which was the notation we established in the Section 1.2.1. To avoid ambiguity, the notation y_{-i} will always denote $y_{-i} = y - A^*{}_i x_i$ for the vectors y, and the $-i$ subscript will be interpreted as in the introduction for all other vector and matrix valued objects in the work (e.g, x, A^*, A, etc..).

3.1.2 Pitfalls for the Meta-Algorithm under Imperfect Sign Thresholding

The key hurdle to overcome for (C, ρ)-smoothly distributed vectors is that there is no clear way to guarantee correct sign thresholding with very high probability. Following the Meta-Algorithm, suppose that, given an estimate A of A^* and sample $y = Ax$, we estimate $\hat{S} := \{i : |A_i^T y| \geq \tau\}$ and $\text{sign}(x_i) = \text{sign}_\tau(A_i^T y)$. Then

$$\begin{aligned} A_i^T y &= A_i^T A_i^* x_i + \sum_{j \in S - \{i\}} A_i^T A_j^* \\ &= A_i^T A_i^* x_i + \sum_{j \in S - \{i\}} (A_i - A_i^*)^T A_j^* x_j + \sum_{j \in S - \{i\}} A_i^{*T} A^*{}_j x_j \\ &= A_i^T A^*{}_i x_i + (A_i - A_i^*)^T \sum_{j \in S - \{i\}} A_j^* x_j + \sum_{j \in S - \{i\}} A_i^{*T} A_j^* x_j \end{aligned} \tag{1.10}$$

The first term in in the above sum is $\Theta(x_i)$ (recall that $A_i^T A_i^* = 1 \pm o(1)$), but the examining the proof of Lemma 3.1.4 shows that the second term can contribute as much as $\tilde{\Omega}\left(\delta + \sqrt{k}\mu/\sqrt{n}\right)$, with non-trivial probability. Thus, if x is not C-lower bounded, then whenever we happen to draw a coefficient vector x for which $x_i = o\left(\delta + \sqrt{k}\mu/\sqrt{n}\right)$, the sign of $A_i^T y$ will be often be closer to the sign of $A_i^T(y - A^*{}_i x_i)$ than to x_i.

The challenge here isn't so much that support estimation and sign thresholding are noisy. Indeed, suppose for the sake of argument that $\hat{S}$ could be estimated perfectly, but that $\hat{\text{sign}}(x_i) = \text{sign}(x_i) + \xi$, where ξ is some noise depending on x_i but no other entries of x. Then, since $\mathbb{E}[\xi x_j] = 0$, it would be straightforward to adapt the proof of Lemma 2.2.1 to show that, still,

$$\mathbb{E}[M_i(\hat{S}) \hat{\text{sign}}(x_i) y] \propto \mathbb{E}[M_i(S)] A^*{}_i + \gamma \tag{1.11}$$

Rather, the obstacle is that innacurate sign thresholding causes the entries of x (given $S = \mathrm{supp}(x)$) to correlate. In particular, for a set threshold τ, suppose that we happen to draw a sample $y = Ax$ for which $x_i = \tau \pm \tilde{\Omega}\left(\delta + \sqrt{k}\mu/\sqrt{n}\right)$. Then, whether $|A_i^T y|$ makes it over the threshold τ or not depends very heaviliy on $A_i^T(y - A^*{}_i x_i)$.

3.1.3 Automating the Analysis of the Meta-Algorithm

Because we are analyzing many simultaneous update rules, we find expedient to automate much of our analysis.

Definition 3.1.1 ((α, δ, ζ)-true and α-nearness-well-conditioned)**.** For a set of gradient vectors g_i, let g be the matrix whose columns are g_i. We say that that the gradient g_i is $(\alpha_i, \delta, \zeta)$-true if there is a positive $\alpha_i = \Theta(1)$ for which

$$g_i \;/\; q_i = \alpha_i(A^*{}_i - A) + v \tag{1.12}$$

where $\|v\| \le \zeta + \mathrm{O}^*\left(\alpha_i \|A^*{}_i - A\|\right) + \mathrm{o}\left(\delta\right)$. We say g is α-nearness well-conditioned if we can write

$$g\ \mathrm{diag}(q_i)^{-1} = (A^* - A)\mathrm{diag}(\alpha_i) + A\ \mathrm{diag}(\beta_i) + \tilde{G} \tag{1.13}$$

where $\alpha_i = \Omega\left(1\right)$ are are positive, $\beta_i = \mathrm{o}\left(1\right)$, and $\|\tilde{G}\| = \mathrm{o}\left(\|A^*\|\right)$.

We first state two trivial consequences of the above definition:

Lemma 3.1.2. *If g is nearness well conditioned, then $g \pm o(\|A^*\|)$ is also nearness well conditioned. Furthermore, if g_i is $(\alpha_i, \delta, \zeta)$-true, then cg_i is $(c\alpha_i, \delta, \zeta)$-true for $c = \Theta\left(1\right)$.*

We can now state the theorem which automates the convergence analysis of approximate gradient descent for true and nearness-well-conditioned gradients:

Theorem 3.1.3 (Automated Analysis for Gradient Descent)**.** *Let A^0 be $(\delta_0, 2)$ near to A^*, and let (A^s, g^s) be a sequence of iterates and gradient vectors for $s \in \{1, \dots, T\}$ such that*

1. *$A^{s+1} = A^s + g^s \cdot \mathrm{diag}(\eta_i^s)$*
2. *Whenever A^* is $(\delta^s, 2)$-near to A^*, g_i^s is $(\alpha_i^s, \delta^s, \zeta)$-true and g^s is α^s-nearness-well-conditioned.*
3. *$\alpha_{\min} = \min_{i,s} \alpha_i^s$ and $\alpha_{\max} = \max_{i,s} \alpha_i^s$ are both $\Theta(1)$*

Then, as long as the step sizes satisfy

$$\eta_0 \;\le\; \frac{\eta_i^s}{q_i} \;\le\; \frac{2}{25\alpha_{\max}} \quad \text{for all } i \in [m],\ s \in \{1, \dots, T\} \tag{1.14}$$

It holds that for each $s \in \{1, \dots, T\}$ that A^s is $(\delta_0, 2)$-near to A^* and

$$\|A_i^s - A^*\|^2 \le (1 - \alpha_{\min}\eta_0/4)^s \| + 64\zeta^2/\alpha_{\min}^2 \tag{1.15}$$

In fact, the theorem holds in the more general setting when we only assume that the g^s are $(\alpha_i^s, \delta^s + e^s, \epsilon)$-true, whenever

$$e_s^2 \leq (1 - \alpha_{\min}\eta_0/4)^s \delta_0^2 + 64\zeta^2/\alpha_{\min}^2 \tag{1.16}$$

If the g_i^s and g^s are true and near with high probability, and if the and η_i^s satisfy Equation 1.14 with high probability, and $T = poly(n)$, then the conclusions of the theorem remain true with high probability. If g_i^s is $(\alpha_i^s, \delta^s, \epsilon)$-true whenever A^s is simply δ_s-close to A^* (possibly only with high probability), then Equation 1.15 holds as well (possibly again with high probability).

Remark. We made no attempt to be tight with constants in the above proof of the above theorem. Constants can be improved substantially if we require more strongly that the vectors g_i^s be $(\alpha_i, 0, \zeta)$-true. Note also that the previous Theorem gives guarantees for choosing appropriate gradient descent step sizes.

3.1.4 Sign and Support Recovery

While section 3.1.2 highlighted some of the difficulties of thresholding sparse representations which are not C lower bounded, we will still be able to establish a few very useful properties of the sign-thresholding operation for (C, ρ)-smooth distributions.

Definition 3.1.2. Suppose that A is δ-close to A^*. We say the threshold τ is δ-*suitable* if $(\delta + \sqrt{k}\mu/\sqrt{n}) \log n \leq \tau$, and that τ is (δ, C)-suitable if in addition $\tau \geq C/2$, where C is given in Assumption 5. We also define the *estimated support*

$$\hat{S} := \{i : |\text{sign}_\tau(A_i^T y)| \geq \tau\} \tag{1.17}$$

The following Lemma shows that $\hat{S} \subset S$ with high probability, and that $\text{sign}_\tau(A_i^T y)$accurately indentifies the signs of x_i for all $i \in \hat{S}$:

Lemma 3.1.4. *If τ is (δ, C)-suitable, then with probability $1 - n^{-\omega(1)}$,*

1. *$\hat{S} \subset S$*
2. *For all $i \in \hat{S}$ we have $\text{sign}_\tau(A_i^T x_i) = \text{sign}(x_i)$*
3. *If $|x_i| \geq C$, then $\text{sign}_\tau(A_i^T y) = \text{sign}(x_i)$.*

In other words, the only errors that the sign thresholding makes (whp) are in possibly mistaking the case $|x_i| \leq C/2$ for the case $i \notin S$.

As a consequence, we can compute g_i^s by conditioning on the event $i \in S$. To this end, we define the *conditional gradient*

$$G_i := \mathbb{E}_i[M_i y \text{sign}_\tau(A_i^T y)] \tag{1.18}$$

and state the following corrolary of Lemma 3.1.4.

Corollary (Reduction to Conditional Gradients)**.** *Suppose that τ is (δ, C)-suitable. Then g_i is $(\alpha_i, \delta, \zeta + n^{-\omega(1)})$-true if and only if G_i satisfies the decomposition given in Equation 1.12. Similarly, g is nearness well conditioned if and only if G satisfies the right hand side of Equation 1.13.*

Proof.

$$p^{-1}g_i = \frac{1}{p_i}\mathbb{E}[M_i y \text{sign}_\tau(A_i^T y)\mathbb{1}(i \in S)] + \frac{1}{p_i}\mathbb{E}[M_i y \text{sign}_\tau(A_i^T y)\mathbb{1}(i \notin S)] \quad (1.19)$$

The first term on the right is precisely G_i, and the second is $O(\gamma)$ since $\text{sign}_\tau(A_i^T y)\mathbb{1}(i \notin S) = 0$ whp by Lemma 3.1.4 □

3.2 Update Rules for (C, ρ)-Smooth Sparse Coding

In this section, we establish finite sample convergence guarantees for both the Toy Rule and the Neural Rule. For the Toy Rule, we have

Theorem 3.2.1 (Convergence of the Toy Rule)**.** *Suppose that A^* is $(\delta, 2)$-near to A^*, that $\eta = \Theta^*(m/k)$, and that Assumptions 1-5 hold. Then if the update step in the Meta-Algorithm with $M_i = X_i$ uses $p = \tilde{\Omega}(mt)$ fresh samples at each iteration, we have*

$$\|A_i^s - A_i^*\|^2 \le (1-\lambda)^s \|A_i^0 - A^*\|_i^2 + \mathrm{O}\left(1/t + \rho^2 k^2/n^2\right) \quad (2.20)$$

for some $\lambda \in (0, 1/2)$, and $t = \Omega k$, and for all $s = 1, 2, \ldots, T$. In particular, A^* converges geometrically until the column-wise error is $O(1/\sqrt{t} + \rho k/n)$.

and, for the Neural Rule,

Theorem 3.2.2 (Convergence of the Neural Rule)**.** *Suppose that A^* is $(\delta, 2)$-near to A^* for $\delta = \mathrm{O}^*\left((\rho \log nk^{1/4})^{-1}\right)$, the threshold $\tau = \Theta\left((\rho k^{1/4})^{-1}\right)$, the step size $\eta = \Theta^*(m/k)$, and Assumptions 1-5 hold. Then if the update step in the Algorithm with $M_i(A, \hat{S}) = I - \sum_{j \in \hat{S}} n_j A_j A_j^T$ uses $p = \tilde{\Omega}(mk^2)$ fresh samples at each iteration, we have*

$$\|A_i^s - A_i^*\|^2 \le (1-\lambda)^s \|A_i^0 - A^*\|_i^2 + \mathrm{O}\left(\rho^2 k^2/n^2 + \frac{\mu^2}{n}\right) \quad (2.21)$$

for some $\lambda \in (0, 1/2)$, and for any $s = 1, 2, \ldots, T$. In particular, A^* converges geometrically until the column-wise error is $\mathrm{O}\left(\rho k/n + \mu/\sqrt{n}\right)$.

In this section, we prove that the expected gradients $g_i^s = \mathbb{E}[\hat{g}_i^s]$ are $(\Theta(1), 0, k/n)$-true, and that the gradient matrices are $\Theta(1)$-well conditioned. While Theorem 3.1.3 renders these statements are sufficient for convergence in the infinite-sample regime, the complete proofs fo Theorems 3.2.1 and 3.2.2 will have to wait till a thorough sample complexity analysis in the Appendix, section A.2. We remark that the sample complexity of the Neural Rule suffers by a factor of k when we remove the C-lower boundedness assumption (performance of the Toy Rule, however, does not increase under C-lower bounded conditions).

3.2.1 Analysis of the Toy Rule

We will devote this section to the proof of the following theorem:

Theorem 3.2.3. *Suppose that A is $(2, \delta)$-near to A^* and τ is (δ, C)-suitable. Then, given then the toy update rule*

$$g_i := \mathbb{E}[X_i y \text{sign}_\tau(A_i^T y)] \tag{2.22}$$

g_i are $(a_i, 0, \mathrm{O}(\rho k/n))$-true and g is a-nearness well conditioned, where

$$a_i := \mathbb{E}[x_i \text{sign}_\tau(A_i^T y) \big| i \in S] = \Theta(1) \tag{2.23}$$

Recalling Theorem 3.1.3, we have the immediate corrolary

Theorem 3.2.4. *Suppose that A^* is $(\delta, 2)$-near to A^*, that the threshold $\tau = \Theta^*(C)$, the step size $\eta = \Theta^*(m/k)$, and Assumptions 1-5 hold. Then, if update step in the Meta Algorithm uses the projector matrix $M_i = X_i$ and an infinite number of samples per iteration, we have*

$$\|A_i^s - A_i^*\|^2 \le (1 - \lambda)^s \|A_i^0 - A^*\|_i^2 + \mathrm{O}\left(\rho^2 k^2 / n^2\right) \tag{2.24}$$

for some $\lambda \in (0, 1/2)$, and for any $s = 1, 2, \ldots, T$. In particular, A^* converges geometrically until the column-wise error is $\mathrm{O}(\rho k/n)$.

Because of its relative simplicity, the analysis of the toy algorithm will help elucidate the key techniques necessary for adapting gradient descent algorithms to the (C, ρ)-smooth case. The first step is a straightforward computation of G_i:

Claim 3.2.5.

$$G_i := \mathbb{E}_i[X_i y \text{sign}_\tau(A_i^T Y)] = \alpha_i (A^*{}_i - A_i + O(\|A^*{}_i - A_i\|^2) + E_i \tag{2.25}$$

where $\alpha_i := \mathbb{E}[x_i \text{sign}_\tau(A_i^T y) \big| i \in S]$ and

$$E_i := M_i A^*_{-i} \text{vec}(\Pr(j \in S | i \in S) \cdot \mathbb{E}[x_j \text{sign}_\tau(A_i^T y) \big| i, j \in S]) \tag{2.26}$$

Proof. Since $M_i = X_i$ deterministically, we may write

$$\begin{aligned} \mathbb{E}[X_i y \cdot \text{sign}_\tau(A_i^T y) \big| i \in S] = X_i \mathbb{E}[\sum_{j \in S} A^*{}_j x_j \cdot \text{sign}_\tau(A_i^T y) \big| i \in S] \\ + X_i A^*{}_i \mathbb{E}[x_i \text{sign}_\tau(A_i^T y) \cdot \big| i \in S] \end{aligned} \tag{2.27}$$

The first term is precisely E_i. Moreover, by the definition of α_i and Claim 3.1.1,

$$X_i A^*{}_i \mathbb{E}[x_i \text{sign}_\tau(A_i^T y) \cdot \big| i \in S] = a_i X_i A^*{}_i = a_i (A^*{}_i - A_i) + O(\|A^*{}_i - A_i\|^2) A_i \tag{2.28}$$

□

It is rather straightforward to show $\alpha_i = \Omega(1)$, so we will defer that argument to the end of the section. The more central innovation how we control E_i. From Claim 3.2.5, we see that bounding E_i amounts to controlling the terms

$$\xi_{i,j} := \mathbb{E}[x_j \text{sign}_\tau(A_i^T y) \big| i, j \in S]) \tag{2.29}$$

The $\xi_{i,j}$ are not a coincidental artificact of our algorithm or method of analysis. Instead, the $\xi_{i,j}$ capture how much influence x_j exerts on the sign on the estimated sign of x_i. In the C-lower bounded case, we could ensure that $\xi_{i,j} = n^{-\omega(1)}$ by thresholding at $\tau = C/2$. With (C, ρ) boundedness, the best we have is the following:

Proposition 3.2.6. *Let* $\xi_{i,j} := \mathbb{E}[x_j \text{sign}_\tau(A_i^T y) \big| i, j \in S])$. *Then,*

$$\xi_{i,j} \lesssim \rho |A_i^T A_j^*| + \gamma \tag{2.30}$$

Equivalently, there are real numbers $s_{i,j} = O(1)$ *for which*

$$\xi_{i,j} = \rho s_{i,j} A_j^{*T} A_i + \gamma \tag{2.31}$$

We refer to the property in Proposition 3.2.6 as a form of *non-correlation*, which follows from the *anti-concentration* assumption encoded in (C,ρ)-smoothness. The moral intuition is roughly:

If $X_1, \dots, X_n$ *are* ρ*-smooth independent (or weakly dependent) random variables, then the random variable* $f(X_1, \dots, X_n)$ *is roughly independent of* X_i *provided that* $f(x_1, \dots, x_n)$ *is not too sensitive to* x_i*, but is sufficiently sensitive to at least one of its coordinates*

Proof of Proposition 3.2.6. Let's first show that if the proposition is true for ρ smooth distributions, then it holds for (C,ρ)-smooth distributions.

Reduction to ρ-Smooth-Distributions

We have that

$$\begin{aligned}
\xi_{i,j} &= \mathbb{E}_{i,j}[x_j \text{sign}_\tau(A_i^T y)\mathbb{1}(x_i \in [-C,C])] + \mathbb{E}_{i,j}[x_j \text{sign}_\tau(A_i^T y)\mathbb{1}(x_i \in [-C,C])\mathbb{1}(|x_i| > C)] \\
&= \Pr(x_i \in [-C,C])\mathbb{E}_{i,j}[x_j \text{sign}_\tau(A_i^T y)\big| x_i \in [-C,C]] \\
&+ \Pr(|x_i| > C)\mathbb{E}_{i,j}[x_j \text{sign}_\tau(A_i^T y)\big| |x_i| > C]
\end{aligned}$$

Note that whenever $|x_i| > C$, then $\text{sign}_\tau(A_i^T y) = \text{sign}(x_i)$ whp by Lemma 3.1.4. Hence, we have

$$\begin{aligned}
\Pr(|x_i| > C)\mathbb{E}_{i,j}[x_j \text{sign}_\tau(A_i^T y)\big| |x_i| > C)] &= \gamma + \Pr(|x_i| > C)\mathbb{E}_{i,j}[x_j \text{sign}(x_i)\big| |x_i| > C)] \\
&= \gamma + \mathbb{E}_{i,j}[x_j \text{sign}(x_i)\mathbb{1}(|x_i| > C)] \\
&= \gamma
\end{aligned}$$

where $\mathbb{E}[\text{sign}(x_i)\mathbb{1}(|x_i| > C)] = 0$ follows from the fact that x_i is symmetrically distributed, and $x_i \perp x_j$. Letting $\tilde{x}_i = x_i \big| x_i \in [-C,C]$, and $\tilde{y} = y \big| x_i \in [-C,C]$, we have

$$\xi_{i,j} = \gamma + \Pr(x_i \in [-C,C])\mathbb{E}_{i,j}[x_j \text{sign}_\tau(A_i^T \tilde{y})] \tag{2.32}$$

Now, note that $\tilde{x}_i$ is $\frac{\rho}{\Pr(x_i \in [-C,C])}$ smooth, since for $S \subset [-C,C]$

$$\Pr(\tilde{x}_i \in S) = \frac{\Pr(x_i \in S)}{\Pr(x_i \in [-C,C])} \leq \frac{\rho \text{ vol}(S)}{\Pr(x_i \in [-C,C])} \tag{2.33}$$

Moreover $\tilde{x}_i$ satisfies the same key distributional assumptions as x_i (independence from the other x_j's, symmetrically distributed, etc.), so if the current proposition holds for the ρ smooth case, then

$$\begin{aligned}
\xi_{i,j} &= \gamma + \Pr(x_i \in [-C,C])\mathbb{E}_{i,j}[x_j \text{sign}_\tau(A_i^T \tilde{y})] \\
&= \gamma + O(\Pr(x_i \in [-C,C]) \frac{\rho}{\Pr(x_i \in [-C,C])} |A_i^T A^*{}_j|) \\
&= \gamma + O(\rho |A_i^T A^*{}_j|)
\end{aligned}$$

Proof for ρ-Smooth Distributions

If x_j is ρ-smooth, then we have

$$\begin{aligned}\mathbb{E}[x_j \text{sign}_\tau(A_i^T y)\big| i, j \in S]) &= \mathbb{E}[x_j \text{sign}_\tau(A_i^T \sum_{r \neq j} A_r^* x_r)\big| i, j \in S]) \\ &+ \mathbb{E}[x_j \left(\text{sign}_\tau(A_i^T \sum_{r \in S} A_r^* x_r) - \text{sign}_\tau(A_i^T \sum_{r \neq j} A_r^* x_r) \right) \big| i, j \in S])\end{aligned}$$

By the independence of x_j and $\{x_r\}_{r \neq j}$, we have $\mathbb{E}[x_j \text{sign}_\tau(A_i^T \sum_{r \neq j} A_r^* x_r)\big| i, j \in S]) = 0$. It therefore suffices to control the term on the second line, for which we note that:

$$\begin{aligned}|\mathbb{E}[x_j \text{sign}_\tau(A_i^T y)\big| i, j \in S])| &= |\mathbb{E}[x_j (\text{sign}_\tau(A_i^T \sum_{r \neq j} A_r^* x_r) - \text{sign}_\tau(A_i^T \sum_{r \in S} A_r^* x_r))\big| i, j \in S])| \\ &\leq \mathbb{E}[|x_j| \cdot |\text{sign}_\tau(A_i^T \sum_{r \in S} A_r^* x_r) - \text{sign}_\tau(A_i^T \sum_{r \neq j} A_r^* x_r)| S])| \\ &\leq \sup_{S : S \supset \{i,j\}} \mathbb{E}[|x_j| \cdot |\text{sign}_\tau(A_i^T \sum_{r \in S} A_r^* x_r) - \text{sign}_\tau(A_i^T \sum_{r \in S} A_r^* x_r)| S])| \\ &\leq \mathbb{E}[|x_j| |\text{sign}_\tau(A_i^T \sum_{r \neq j} A_r^* x_r) - \text{sign}_\tau(A_i^T \sum_{r \in S^*} A_r^* x_r)| S^*])|\end{aligned}$$

where S^* is the set that attains the supremum in the above display. Without loss of generality, we may relabel the indices so that S^* is supported on the first k indices of m, and that $i = 1$ and $j = 2$, and the constants $a_r := A_1^T A_r^*$ and the random variables $Z_r = x_r \big| S^*$. Because we are shifting indices around, let's remember to keep in mind that $a_2 = A_i^T A_j^*$ in our original indexing shcheme. Switching notation, we write

$$\begin{aligned}|\mathbb{E}[x_j \text{sign}_\tau(A_i^T y)\big| i, j \in S])| &\leq \mathbb{E}[x_j |\text{sign}_\tau(\sum_{r \neq 2} a_r x_r) - \text{sign}_\tau(\sum_r a_r x_r)| S^*])| \\ &= \mathbb{E}[|Y_j| |\text{sign}_\tau(\sum_{r \neq 2} a_r Z_r) - \text{sign}_\tau(\sum_r a_r Z_r)|\end{aligned}$$

Concluding The Proof

The basic intuition want to use is that removing a variable of magnitude roughly a_2 from the sum of random variables with magnitude at least roughly a_1 only introduces a correlation on the order a_2/a_1. The following lemma, proved in the appendix, makes this "non-correlation" precise:

Lemma 3.2.7. *Let $Z_1, \ldots, Z_k$ be real random variables such that $Z_1 \perp (Z_2, \ldots, Z_k)$, and Z_1 is ρ-smoothly distributed. Then, for any measurable function $f(\cdot)$ and any $\tau \in \mathbb{R}$, and any vector $a \in \mathbb{R}^k$, it holds that*

$$\mathbb{E}\left[|f(Z_2)| \left| \text{sign}_\tau(\sum_{i \neq 2}^n a_r Z_r) - \text{sign}_\tau(\sum_{i=1}^k a_i Z_i) \right| \right] \lesssim \left| \rho \frac{a_2}{a_1} \right| \mathbb{E}\left[|Z_2 f(Z_2)| \right] \tag{2.34}$$

With this lemma at our disposal, we can easily conclude:

$$|\mathbb{E}[x_j \text{sign}_\tau(A_i^T y)|i, j \in S])| \lesssim \rho \frac{|a_2|}{|a_1|}\mathbb{E}[Y_2^2] \asymp \rho \frac{|a_2|}{|a_1|} \asymp \rho |A_i^T A^*{}_j|$$

where we used the fact that $a_2 = A_i^T A^*{}_j$, and $a_1 = A_i^T A^*{}_i = 1 \pm o(1)$. □

Plugging in the just-proved lemma into the definition of E_i yields the following corollary:

Corollary.

$$E_i = \rho M_i A^*_{-i} \text{diag}(\Pr(j \in S | i \in S) \cdot s_{i,j})(A^*_{-i})^T A_i \tag{2.35}$$

where the numbers $s_{i,j}$ which are $O(1)$ in magnitude.

We are can finally control E_i with the following claim:

Claim 3.2.8.

$$\|E_i\| = O\left(\frac{\rho k}{m}\|A^*\|^2\right) = O\left(\frac{\rho k}{n}\right) \tag{2.36}$$

Consequently,

$$\|E\| = O\left(\frac{\rho k}{\sqrt{n}}\|A^*\|\right) = o\left(\|A^*\|\right) \tag{2.37}$$

Proof. For the first claim, since M_i is $O(1)$, it follows that E_i is on the same order as the following display:

$$\begin{aligned} \|\rho A^*_{-i} \text{diag}(\Pr(j \in S | i \in S) \cdot s_j)(A^*_{-i})^T A_i\| &\leq \rho \|A^*_{-i} \text{diag}(\Pr(j \in S | i \in S) \cdot s_j)(A^*_{-i})^T\| \\ &\leq \rho \|A^*_{-i}\|^2 \|\text{diag}(\Pr(j \in S | i \in S) \cdot s_j)\| \\ &= \rho \|A^*_{-i}\|^2 \max_{j \neq i} s_j \Pr(j \in S | i \in S) \\ &\lesssim \frac{\rho k \|A^*_{-i}\|^2}{m} \end{aligned}$$

Since $\|A^*_{-i}\|^2 = O(\sqrt{m/n})$, it follows that $E_i = O(\rho k/n)$ For the second claim,

$$\|E\|_2 \leq \|E\|_F = \sqrt{\sum_{i \in [m]} \|E_i\|^2} \leq \sqrt{m} \max_{i \in [m]} \|E_i\| \lesssim \frac{\rho k \sqrt{m}}{m}\|A^*\|^2 \tag{2.38}$$

Using the fact that $\|A^*\| = O(\frac{\sqrt{m}}{\sqrt{n}})$ and $\frac{\rho k}{\sqrt{n}} = o(1)$, it follows that $\|E\| = o(\|A^*\|)$. □

Finally, we establish that a_i is $\Theta(1)$:

Claim 3.2.9. $a_i \geq \mathbb{E}_i[|x_i|]/2 - \gamma = \Theta(1)$

Proof. Since $\text{sign}_\tau(x_i) \neq -\text{sign}(x_i)$ whp, $\text{sign}_\tau(x_i) = \text{sign}(x_i)$ whp whenever $|x_i| > C$, and $\Pr(|x_i| > C) \geq \frac{1}{2}$, it holds that

$$\begin{aligned}
\mathbb{E}[x_i \text{sign}_\tau(A_i^T y)] &= \mathbb{E}[|x_i| \mathbb{1}(\text{sign}_\tau(A_i^T y) = \text{sign}(x))] + \mathbb{E}[-|x_i| \mathbb{1}(\text{sign}_\tau(A_i^T y) = -\text{sign}(x))] \\
&= \mathbb{E}[|x_i| \mathbb{1}(\text{sign}_\tau(A_i^T y) = \text{sign}(x))] - \gamma \\
&\geq \mathbb{E}[|x_i| \mathbb{1}(|x_i| > C) \mathbb{1}(\text{sign}_\tau(A_i^T y) = \text{sign}(x))] - \gamma \\
&\geq \mathbb{E}[|x_i| \mathbb{1}(|x_i| > C)] - \gamma \\
&\geq \mathbb{E}[|x_i|]/2 - \gamma
\end{aligned}$$

□

Proof of Theorem 3.2.3. The Theorem now follows readily by combining Claims 3.2.5, Claim 3.2.9, 3.2.8, and Corrolary 3.1.4. □

Computational Efficiency

Since the toy rule only requires an estimate A_i^s to compute updates for A_i^{s+1}, iterations of the toy rule can be parallelized columnwise, so that a server need only maintain A_i^s at each step. Moreover, these serves need not even communicate after each iteration. This is useful when the number of columns m far exceeds memory limitations. However, we will see in the next section that simplicity and efficiency of the toy rule come at the cost of sample complexity.

3.2.2 Analysis of Neural Update Rules

While simple to analyze and efficient to implement, the toy rule has rather poor sample complexity in the $k \ll n^{1/2}$ regime. We can think of this issue as a matter of signal-to-noise ration. In this report, we will refer informally to the *signal* as the term $X_i A^*{}_i \approx A^*{}_i - A_i = O(\delta)$. In the toy rule, X_i roughly projects onto the orthogonal complement of A_i, so if $\|A_i - A^*{}_i\| = \delta$, then $A_i^T A_j^* = O(\delta + \mu/\sqrt{n}) = o(1)$. Hence, $X_i A_j^* = \Omega(1)$ for all $j \neq i$. Thus, for a sample $y = A^* x$ with $i \in \text{supp}(x)$, we will have that the *noise* term $X_i(y - A^*{}_i x_i)$ will be considerably large

$$\|X_i(y - A^*{}_i x_i)\| = \|X_i \sum_{j \neq i} A^*{}_j x_j\| \approx \|y\| \approx \sqrt{k} \gg \delta \tag{2.39}$$

In the finite sample setting, this means that the signal $X_i A_i^*$ gets drowned out by contributions from the entries $j \neq i$.

To build algorithms with better sample complexity, we transition to a more general class of algorithms which projector M_i for which $M_i A_j^*$ is small even for $j \neq i$. Motivated by the neural algorithm, we will consider update rules with projector matrices that depend on the estimate support $\hat{S}$ linearly:

Definition 3.2.1. We say that the projector M_i is an *additively decomposable projector* if there exists a matrix $B^{(i)} \in \mathbb{R}^{m \times n}$ for which

$$M_i = X_i - \sum_{r \in \hat{S} - \{i\}} B_r^{(i)} (B_r^{(i)})^T = X_i - \sum_{r \in [m] - \{i\}} B_r^{(i)} (B_r^{(i)})^T \mathbb{1}(r \in \hat{S}) \tag{2.40}$$

and $B^{(i)} = O(\|A^*\|)$. We define the *neural rule* as the update rule with the additively decomposable projector $M_i = I - \sum_{r \in \hat{S}} n_r A_r A_r^T$

Remark. We remark that the neural rule is very similar to the neural rule in Arora et al. (2015), the only difference being that the latter uses $M_i = \sum_{r \in \hat{s}} A_r A_r^T$. We believe, however, that the slight rescaling by the n_r necessarily corrects a flaw in their analysis of convergence, but we omit any further discussion here.

The following proposition gives a general analysis of additively decomposable update rules:

Proposition 3.2.10. *Suppose that A is $(\delta, 2)$-near to A^* for an appropriately small δ, that τ is (δ, C)-suitable, and that $\tau^2 \rho^2 k^{1/2} = O(1)$. Then if M_i is an additively decomposable projector, the the expected gradients*

$$g_i := \mathbb{E}[M_i y \text{sign}_\tau(A_i^T y)] \tag{2.41}$$

are $(a_i, 0, \mathrm{O}(\rho k/n))$-true and g is a-nearness well conditioned, where again.

$$a_i := \mathbb{E}[x_i \text{sign}_\tau(A_i^T y) \big| i \in S] = \Theta(1) \tag{2.42}$$

In particular, g_i are $(a_i, 0, \mathrm{O}(\rho k/n))$-true and g is a-nearness-well-conditioned when the projector matrix is chosen according to the neural update rule.

Again, applying Theorem 3.1.3 gives an immediate corollary:

Theorem 3.2.11. *Suppose that A^* is $(\delta, 2)$-near to A^* for $\delta = \mathrm{O}^*\left((\rho \log nk^{1/4})^{-1}\right)$, the threshold $\tau = \Omega\left((\rho k^{1/4})^{-1}\right)$, the step size $\eta = \Theta^*(m/k)$, and Assumptions 1-5 hold. Then, if update step in the Meta Algorithm uses the projector matrix $M_i = I - \sum_{j \in \hat{S}} n_j A_j A_j^T$ and an infinite number of samples per iteration, we have*

$$\|A_i^s - A_i^*\|^2 \le (1 - \lambda)^s \|A_i^0 - A^*\|_i^2 + \mathrm{O}\left(\rho^2 k^2 / n^2\right) \tag{2.43}$$

for some $\lambda \in (0, 1/2)$, and for any $s = 1, 2, \ldots, T$. In particular, A^* converges geometrically until the column-wise error is $O(\rho k/n)$. In fact, the theorem holds whenever M_i is an additively decomposable projector.

Proof. We verify that if $\delta = \mathrm{O}^*\left((\rho \log nk^{1/4})^{-1}\right) = o(1)$, then the threshold $\tau = \Omega\left((\rho k^{1/4})^{-1}\right)$ is (δ, C)-suitable. Moreover, under the stated scaling of τ, we have that $\tau^2 \rho^2 k^{1/2} = O(1)$. The conditions of Proposition 3.2.10 are met, and the result follows immediately from Theorem 3.1.3. □

Proof of Proposition 3.2.10. Let $Y_{i,j} = x_j \text{sign}_\tau(A_i^T y)$, and $Y_i = x_i \text{sign}_\tau(A_i^T y)$. As in the proof of the toy rule, its suffices to control G_i

$$G_i = \mathbb{E}_i[M_i A^*{}_i Y_i] + \mathbb{E}_i[M_i \sum_{j \neq i} Y_{i,j} A^*{}_j] \tag{2.44}$$

Lets control each term separately.

Controlling the "Signal" Term

First off, since $\|A_i^*\| = 1$, we have

$$\begin{aligned}\mathbb{E}_i[M_i A^*{}_i Y_i] &= \mathbb{E}_i[X_i A^*{}_i x_i \mathrm{sign}_\tau(A_i^T y)] + \mathbb{E}_i[A^*{}_i \sum_{j\in\hat{S}} A^*{}_i x_i \mathrm{sign}_\tau(A_i^T y)] \\ &= \alpha_i X_i A^*{}_i + O(\|\mathbb{E}_i[\sum_{j\in\hat{S}} A_j A_j^T x_i \mathrm{sign}_\tau(A_i^T y)]\|)\end{aligned}$$

Next, note that whenever $\hat{S} \subset S$, $\pm\left(\sum_{j\in\hat{S}} A_j A_j^T x_i \mathrm{sign}_\tau(A_i^T y)\right) \preceq \sum_{j\in S} A_j A_j^T |x_i|$. As $\hat{S} \subset S$ whp, Lemma D.2.2 yields that

$$\left\|\mathbb{E}_i[\sum_{j\in\hat{S}} A_j A_j^T x_i \mathrm{sign}_\tau(A_i^T y)]\right\| \le \left\|\mathbb{E}_i[\sum_{j\in S} A_j A_j^T |x_i|]\right\| + \gamma \lesssim \frac{k\|A\|^2}{m}\gamma \lesssim \frac{k}{n} + \gamma \tag{2.45}$$

Controlling the "Noise" Terms E_1, E_2, E_3

Next, up we have

$$\mathbb{E}_i[M_i \sum_{j\neq i} Y_{i,j} A^*{}_j] = E_1(i) + E_2(i) + E_3(i) \tag{2.46}$$

Where we have

$$\begin{aligned}E_1(i) &= -n_i A_i A_i^T \mathbb{E}_i[\sum_{j\neq i} \mathbb{1}(j \notin \hat{S}) A^*{}_j Y_{i,j}] Y_{i,j} \\ E_2(i) &= \mathbb{E}_i[\sum_{j\neq i} (I - B_j^{(i)}(B_j^{(i)})^T)\mathbb{1}(j\in\hat{S}) A^*{}_j Y_{i,j}] \\ E_3(i) &= -\mathbb{E}[\sum_{r\neq i} B_r^{(i)}(B_r^{(i)})^T \sum_{j\neq i,r} A^*{}_j Y_{i,j}\mathbb{1}(r\in\hat{S})]\end{aligned} \tag{2.47}$$

and let E_1, E_2, E_3 be the matrices whose i-th columns are are $E_1(i)$, $E_2(i)$, $E_3(i)$. Following the proof of Theorem 3.2.3, it is straightforward to show that

$$\|E_1(i)\| = O\left(\frac{\|A^*\|\rho k}{\sqrt{nm}}\right) \tag{2.48}$$

For E_3, we have the following claim

Claim 3.2.12.

$$\|E_3(i)\| \lesssim \frac{k^2\|A^*\|}{n\sqrt{m}} \tag{2.49}$$

Proof. By Lemma A.1.4 (a more sophisticated version of Claim 3.1.1), we can write

$$\mathbb{E}_i[Y_{i,j}\mathbb{1}(r\in\hat{S})] \lesssim \frac{\rho k^2}{m^2}(|A_i^T A^*{}_j| + |A_r^T A^*{}_j|) \tag{2.50}$$

That is, there are real numbers $s^1_{i,j,r}$ and $s^2_{i,j,r}$ which are $O(1)$in magnitude for which

$$\begin{aligned}
-E_3(i) &= \mathbb{E}[\sum_{r\neq i} B_r^{(i)}(B_r^{(i)})^T \sum_{j\neq i,r} A^*{}_j Y_{i,j}\mathbb{1}(r\in\hat{S})] \\
&= \sum_{r\neq i} B_r^{(i)}(B_r^{(i)})^T \sum_{j\neq i,r} A^*{}_j \mathbb{E}[Y_{i,j}\mathbb{1}(r\in\hat{S})] \\
&= \frac{\rho k^2}{m^2}\sum_{r\neq i} B_r^{(i)}(B_r^{(i)})^T \sum_{j\neq i,r} A^*_j s^1_{i,j,r} A^{*T}_j A_i + A^*_j s^2_{i,j,r} A^{*T}_j A^*{}_r \\
&= \frac{\rho k^2}{m^2}\sum_{r\neq i} B_r^{(i)}(B_r^{(i)})^T \left(A^*_{-i,r}\mathrm{diag}(s^1_{i,j,r})(A^*_{-i,r})^T A_i + A^*_{-i,r}\mathrm{diag}(s^1_{i,j,r})(A^*_{-i,r})^T A^*{}_j\right)
\end{aligned}$$

Now, define $v(i)$ to be the vector

$$v(i)_r := (B_r^{(i)})^T\left)A^*_{-i,r}\mathrm{diag}(s^1_{i,j,r})(A^*_{-i,r})^T A_i + A^*_{-i,r}\mathrm{diag}(s^1_{i,j,r})(A^*_{-i,r})^T A^*{}_j\right) \tag{2.51}$$

so that $E_i(i) = \frac{\rho k^2}{m^2}\sum_{r\neq i} B_r^{(i)} v_r(i)$. Since $s^1_{i,j,r}$, $s^2_{i,j,r}$, $\|A_i\|$, $\|B_i\|$ and $\|A^*{}_j\|$ are $O(1)$ in magntitude, using the multiplicativity of the spectral norm gives:

$$|v(i)_r| \lesssim \|A^*\|^2 \quad \text{and hence} \quad \|v(i)_r\|_2 \lesssim \sqrt{m}\|A^*\|^2 \tag{2.52}$$

Using the fact that $\|B^{(i)}\| = O(\|A^*\|)$, we have

$$\|E_3(i)\| \leq \frac{\rho k^2}{m^2}\|B^{(i)}\|\|v(i)_r\| \lesssim \frac{\rho k^2}{m^{3/2}}\|A^*\|^3 \lesssim \frac{\rho k^2\|A^*\|}{n\sqrt{m}} \tag{2.53}$$

□

Claim 3.2.13.

$$\|E_2(i)\| \lesssim \frac{\rho\|A^*_j\|^2 k}{m} + \frac{k^{3/2}\rho^2\tau^2\|A^*\|}{\sqrt{mn}} \tag{2.54}$$

Proof.

$$E_2(i) = \mathbb{E}_i[\textstyle\sum_{j\neq i}(I - B_j^{(i)}(B_j^{(i)})^T)\mathbb{1}(j\in\hat{S})A^*{}_j Y_{i,j}] \tag{2.55}$$

$$\tag{2.56}$$

We write $E_2(i) = T_1(i) + T_2(i)$, where

$$T_1(i) = \sum_{j\neq i} A^*{}_j \mathbb{E}_i[Y_{i,j}\mathbb{1}(j\in\hat{S})Y_{i,j}] \tag{2.57}$$

$$T_2(i) = \sum_j B_j^{(i)}(B_j^{(i)})^T A^*_j \mathbb{E}_i[\mathbb{1}(j\in\hat{S})Y_{i,j}] \tag{2.58}$$

We will bound T_1; T_2 can be conrolled using the exact same arguments after noting that $(B_j^{(i)})^T A^*{}_j) = O(1)$ and that $\|B^{(i)}\| \leq O(\|A^*\|)$. By Lemma A.1.5, we have

$$|\mathbb{E}_i[\mathbb{1}(j\in\hat{S})Y_{i,j}]| \lesssim \frac{k}{m}\left(\rho|A_i^T A^*{}_j| + \tau\rho|A_j^T A^*{}_i| + \rho^2\tau^2\sqrt{\frac{k}{n}}\right) \tag{2.59}$$

Hence, there are $O(1)$ real numbers $s^1_{i,j}, s^2_{i,j}, s^3_{i,j}$ for which

$$\begin{aligned} T_1(i) &= \frac{k}{m}\sum_{j\neq i} A^*_j \mathbb{E}_i[Y_{i,j}\mathbb{1}(j\in\hat{S})Y_{i,j}] \\ &= \rho\frac{k}{m}\sum_{j\neq i} A^*_j s^1_{i,j} A^{*T}_j A_i + \tau\rho\frac{k}{m}\sum_{j\neq i} A^*{}_j s^2_{i,j} A^T_j A^*{}_i + \frac{k}{m}\sum_{j\neq i} A^*_j s^3_{i,j}\rho^2\tau^2\sqrt{\frac{k}{n}} \end{aligned}$$

We bound the three sums on the last line. For the first term, we have

$$\begin{aligned} \|\rho\frac{k}{m}\sum_{j\neq i} A^*_j s^1_{i,j} A^{*T}_j A_i\| &\leq \frac{\rho k}{m}\|A^*_j\|^2 \max_j s^1_{i,j}\|A_i\| &(2.60)\\ &\lesssim \frac{\rho\|A^*_j\|^2 k}{m} &(2.61)\end{aligned}$$

Similarly, the second sum on the second-to-last display is controlled by $\frac{\rho\|A^*_j\|^2 k\tau}{m}$. Finally,

$$\begin{aligned} \|\sum_{j\neq i} A^*_j s^3_{i,j}\rho^2\tau^2\sqrt{\frac{k}{n}}\| &= \frac{k^{3/2}}{m\sqrt{n}}\|A^*\|\|\mathrm{vec}(s^3_{i,j}\rho^2\tau^2)\| &(2.62)\\ &= \sqrt{m}\|A^*\|\max_j(\mathrm{vec}(s^3_{i,j}\rho^2\tau^2)) &(2.63)\\ &\lesssim \frac{k^{3/2}\rho^2\tau^2\|A^*\|}{\sqrt{mn}} &(2.64)\end{aligned}$$

□

Putting together the bounds on $E_1(i), E_2(i), E_3(i)$, we see that as long as $\tau^2\rho^2 k^{1/2} = O(1)$, we have that

$$E_1(i) + E_2(i) + E_3(i) = O\left(\frac{k}{n} + \frac{k^{3/2}\rho^2\tau^2 k}{n}\right) = O(k/n) \quad (2.65)$$

and that

$$\begin{aligned} \|E_1 + E_2 + E_3\| &\leq \sqrt{m}\max_i \|E_1(i) + E_2(i) + E_3(i)\| \\ &\leq \|A^*\|\left(\frac{k}{\sqrt{n}} + \frac{k^{3/2}\rho^2\tau^2}{n}\right) \\ &= o(\|A^*\|)\end{aligned}$$

□

Chapter 4

A Projection-Based Algorithm for Sparse Coding

4.1 The Projection Rule

4.1.1 Motivation for Projection Rule

In the previous section, we introduced a Meta-Algorithm for dictionary learning and analyzed a two instantiations thereof. Recall that the Meta-Algorithm ran approximate gradient descent on the columns of dictionary iterates $A^{(s)}$, with expected gradients

$$g_i^s = \mathbb{E}[M_i(A^{(s)}, \hat{S})y\text{sign}(\langle A_i^{(s)}, y\rangle)] \tag{1.1}$$

where M_i was the "projector matrix". To facilitate a simple analysis under imperfect sign-thresholding, we analyzed a "toy update" rule where $M_i^{(s)}$ was chosen deterministically as $M_i^{(s)} = X_i : I - \frac{1}{\|A_i^{(s)}\|}A_i^{(s)}(A_i^{(s)})^T$. Despite its simplicity, or in fact because of it, the toy rule suffered from poor sample complexity guarantees. Consequently, we considered a neural update where the marix $M_i^{(s)}(y)$ dependended linearly on the estimated support $\hat{S}$ of y. The idea was that, for all $j \in \hat{S}$, $M_i^{(s)}(A^*{}_j)$ was small, so the variance of the gradient was reduced.

In this section, we take this same intuition one step further: We design a projector matrix M_i^{prj} which depends on $\hat{S}$ in a highly nonlinear way, but ensures that, for all $j \in \hat{S}$,

$$M_i^{\text{prj}}A_j = M_i^{\text{prj}}(\hat{S})A_j = O(\|A_j^* - A_j^{(s)}\|) \tag{1.2}$$

even when $\|A_j^* - A_j^{(s)}\|$ is made arbitrarily small. With M_i^{prj} defined in this manner, we establish the following theorem for C-Lower Bounded distributions:

Theorem 4.1.1. *Suppose that A^* is δ-close to A^* for $\delta = \text{O}^*\left(1/\sqrt{k}\right)$, that the sparse coeffcient vectors are C-lower bounded, the threshold $\tau = C/2$, the step size $\eta = \Theta^*(m/k)$, and Assumptions 1-4 hold. Then, if update step in the Meta Algorithm uses the projector matrix M_i^{prj} and and $p = \tilde{\Omega}(m)$ samples, we have*

$$\|A_i^s - A_i^*\|^2 \leq (1-\lambda)^s\|A_i^0 - A^*\|_i^2 + n^{-\omega(1)} \tag{1.3}$$

for some $\lambda \in (0, 1/2)$, and for any $s = 1, 2, \dots, T$. In particular, A^* converges geometrically to an arbitrary inverse polynomial error.

Following the analysis the toy and neural rules, this section proves of convergence of the projection rule in the infinite sample setting. The proof of the finite sample result is deferred to Section A.2 We also remark that the projection rule has substantialy superior sample complexity to the "unbiased algorithm" in Arora et al. (2015), where the sample complexity grows quadratically in the desired precision.

4.1.2 Definition of M_i^{prj}

The strategy is to make M_i^{prj} "almost" an orthogonal projection onto the complement of $A_{\hat{S}}^{(s)}$. Before defining M_i^{prj}, we need to establish some further notation. As in the previous chapter, drop the superscripts s to avoid notational clutter, and restrict our attention to one iteration of the update rule. Next, for $S \subset [m]$, let

$$Q_S = \mathrm{Proj}_{A_S} \quad \text{and} \quad P_S = \mathrm{Proj}_{A_S^\perp} \tag{1.4}$$

That is, Q_S is the projection onto the span of the columns of S, and P_S is the projection onto their complement. We will also let Q_i denote the projection onto the span of A_i, and P_S denote the projection onto the span onto $A_i^\perp$. As in the previous chapter, set $n_i = \frac{1}{\|A_i\|}$, $X_i = I - n_i A_i A_i^T$. We now define the "Projection Update Rule", which uses projector matrix

$$M_i^{\mathrm{prj}}(\hat{S}) = P_{\hat{S}} + (n_i^2 - n_i) A_i A_i^T \tag{1.5}$$

As noted above, the first term is just the orthogonal projection onto $A_{\hat{S}}$, and the second term is a subtle correction for when $\|A_i\|$ is not exactly a unit vector. Before continuing, lets establish some simple facts about $M_i(\hat{S})$:

Proposition 4.1.2 (Properties of M_i^{prj})**.** *Whenever $i \in \hat{S}$,*

$$M_i^{\mathrm{prj}}(\hat{S}) = X_i + (Q_{\hat{S}} - Q_i) = X_i + P_i(Q_{\hat{S}} - Q_i)P_i \tag{1.6}$$

Furthermore, suppose that the columns of A are δ-close to the columns of A^*. Then, for any $j \in \hat{S}$.

$$\|M_i^{\mathrm{prj}}(\hat{S}) A_j^*\| \le 2\delta \tag{1.7}$$

Proof. For the first point, write $P_{\hat{S}} = I - Q_{\hat{S}} = I - Q_i - (Q_{\hat{S}} - Q_i)$. Now, $I - Q_i = I - \frac{1}{\|A_i\|^2} A_i A_i^T = I - n_i^2 A_i A_i^T$. Hence,

$$\begin{aligned} P_{\hat{S}} &= I - n_i^2 A_i A_i^T - (Q_{\hat{S}} - Q_i) + n_i^2 A_i A_i^T - n_i A_i A_i^T \\ &= I - n_i A_i A_i^T - (Q_{\hat{S}} - Q_i) \;=\; X_i - (Q_{\hat{S}} - Q_i) \end{aligned}$$

By Lemma D.1.2 Part 4, $P_i(Q_{\hat{S}} - Q_i)P_i = Q_{\hat{S}} - Q_i$ (the basic idea is that the kernel of P_i, which is exactly the span of A_i, is contained in the kernel of $Q_{\hat{S}} - Q_i$). For the second point, we have that

$$\|M_i^{\mathrm{prj}}(\hat{S}) A_j^*\| \;\le\; \|P_{\hat{S}} A_j^*\| + |n_i^2 - n_i| \|A_i A_i^T A_j^*\|$$

Since $A_j \in \ker P_{\hat{S}}$ for $j \in \hat{S}$, we have $\|P_{\hat{S}} A_j^*\| = \|P_{\hat{S}}(A_j^* - A_j)\| \le \|P_{\hat{S}}\|\|A_j^* - A_j\| \le \delta$, since $\|A_j^* - A_j\| \le \delta$ and $\|P_{\hat{S}}\| = 1$. Secondly,

$$\begin{aligned} |n_i^2 - n_i|\|A_i A_i^T A_j^*\| &\le |n_i^2 - n_i|\|A_i\|^2 \quad \text{since } \|A_j^*\| = 1 \\ &= \left| \frac{1}{\|A_i\|^2} - \frac{1}{\|A_i\|} \right| \|A_i\|^2 \\ &= |1 - \|A_i\||| \\ &\le \delta \end{aligned}$$

where the last step follows since $\|A_i - A_i^*\| \le \delta$, and $\|A^*{}_i\| = 1$. □

4.1.3 Analyzing the Projection Update Rule

We now present an analysis of the Projection Rule under quite general conditions. Keeping the same notational conventions of Section 3.1, we begin by examining the expected conditional gradients

$$G_i := \mathbb{E}_i[M_i^{\text{prj}} y \ \text{sign}_\tau(A_i^T y)] \tag{1.8}$$

Again, we break up $y = A_i^* x_i + y_{-i}$, where $y_{-i} = \sum_{j \in S - \{i\}} A_j^* x_j$, and regard $G_{1,i} := \mathbb{E}[M_i^{\text{prj}} A_i^* x_i \text{sign}_\tau(A_i^T y)]$ as the desired "signal", whilst viewing $G_{2,i} := \mathbb{E}[M_i^{\text{prj}} y_{-i} \text{sign}_\tau(A_i^T y)]$ as systemic noise. Our first proposition shows that the signal $G_{1,i}$ is well correlated with $A^*{}_i$ as long as the columns of A are sufficiently near to those of A^*:

Proposition 4.1.3. *Suppose that Assumptions 1, 2, and 4 hold (but not necessarily Assumption 3), and that A is δ-close to A^*, where $\delta = \mathrm{O}^*\left(1/\sqrt{k}\right)$. Then as long as $\hat{S} \subset S$ with probability $1 - n^{-\omega(1)}$, we have*

$$\mathbb{E}_i[M_i^{\text{prj}} A_i^* x_i \text{sign}_\tau(A_i^T y)] = a_i(A_i^* - A_i) + \mathrm{O}^*\left(\|A_i^* - A_i\|\right) + \gamma \tag{1.9}$$

where again $a_i := \mathbb{E}_i[x_i \text{sign}_\tau(A_i^T y)]$. The $\mathrm{O}^*\left(\|A_i^* - A_i\|\right)$ may be replaced by $\mathrm{o}\left(\|A_i^* - A_i\|\right)$ when $\delta = \mathrm{o}\left(1/\sqrt{k}\right)$.

Proof. Define $v_i = P_i A_i^*$. Using Proposition 4.1.2 and Claim 3.1.1 together with the definitions of v_i, we have

$$\begin{aligned} \mathbb{E}_i[M_i^{\text{prj}} A_i^* x_i \text{sign}_\tau(A_i^T y)] &= \mathbb{E}_i[X_i A_i^* x_i \text{sign}_\tau(A_i^T y)] + \mathbb{E}_i[P_i(Q_i) P_i A_i^* x_i \text{sign}_\tau(A_i^T)] \\ &= a_i(A_i^* - A_i) + \mathrm{o}\left(a_i \|A_i^* - A_i\|\right) + E \end{aligned}$$

where $E \le \|P_i \mathbb{E}_i[(Q_{\hat{S}} - Q_i) x_i \text{sign}_\tau(A_i^T y)] v_i\|$. Since $A_i \in \ker(P_i)$, we have

$$\|v_i\| = \|P_i A_i^*\| = \|P_i(A_i^* - A_i)\| \le \|P_i\|\|A_i^* - A_i\| \le \|A_i^* - A_i\| \tag{1.10}$$

Thus, it simply suffices to show that

$$\|\mathbb{E}_i[(Q_{\hat{S}} - Q_i) x_i \text{sign}_\tau(A_i^T y)]\| = \mathrm{O}^*(1) \tag{1.11}$$

Computing a closed form expression for the above display is rather challenging, since $Q_{\hat{S}}$ depends on the estimated support $\hat{S}$ in a highly non-linear way. However, if we can find a PSD matrix Y for which $Y \succeq (Q_{\hat{S}} - Q_i$ with probability $1 - n^{-\omega(1)}$, then Lemma D.2.2 let's us bound left hand side of Equation 1.11 above by $\|\mathbb{E}_i[Y|x_i \mathrm{sign}_\tau(A_i^T y)|]\| \leq \|\mathbb{E}_i[Y|x_i]\|$. Hence, once we find an an appropriate Y, it will suffice to show that $\|\mathbb{E}[Y]\| = \mathrm{O}^*(1)$. We proceed as follows:

Removing Dependence on $\hat{S}$

First, we replace $Q_{\hat{S}} - Q_i$ by a matrix which depends on the true support $S = \mathrm{supp}(x)$ rather than the estimated support $\hat{S}$. Whenever $\hat{S} \subset S$, then the span of the columns of $A_{\hat{S}}$ lie in the span of A_S, and so $Q_S \succeq Q_{\hat{S}}$. Assuming that $\hat{S} \subset S$ with very high probability, we have established the following claim:

Claim 4.1.4. *$Q_{\hat{S}} - Q_i \preceq Q_S - Q_i$ with probability $1 - n^{-\omega(1)}$.*

Controlling $Q_S - Q_i$ by a Sum of Rank One Terms

Since A is δ-close to A^*, it follows that $\sigma_{\min}(A) \leq \sigma_{\min}(A^*) - \delta\sqrt{k} \geq 1/2 - \mathrm{O}^*(1)) \geq 1/4$, where we have that $\sigma_{\min}(A^*) \geq 1/2$ by the Gershgorin Circle Theorem (see Lemma D.1.1). We now appeal to a general theorem which both makes precise and generalizes the intutition that $A_S A_S^T \approx Q_S$ for incoherent matrices A_S:

Theorem 4.1.5. *If $\sigma_{\min}(A_S) \geq c$,*

$$Q_S - Q_V \preceq c^{-2} P_V (A_{S-V} A_{S-V}^T) P_V \tag{1.12}$$

Plugging in $c = 1/4$, we have

$$Q_S - Q_V \preceq 16 P_i (A_{S-i} A_{S-i}^T) P_i \preceq \mathrm{O}\left(P_i (A_{S-i} A_{S-i}^T) P_i\right) \tag{1.13}$$

with high probability.

Completing the Proof

To wrap up, we compute

$$\begin{aligned}
\|\mathbb{E}_i[(Q_{\hat{S}} - Q_i) x_i \mathrm{sign}_\tau(A_i^T y)]\| &\lesssim \|\mathbb{E}_i[P_i A_{S-i} x_i \mathrm{sign}_\tau(A_i^T y) A_{S-i}^T P_i]\| \\
&\leq \|P_i\|^2 \|\mathbb{E}_i[A_{S-i} A_{S-i}^T x_i \mathrm{sign}_\tau(A_i^T y)]\| \\
&= \|\mathbb{E}_i[A_{S-i} A_{S-i}^T x_i \mathrm{sign}_\tau(A_i^T y)]\| \\
&= \|A \mathrm{diag}(\mathbb{E}_i[\mathbb{1}(j \in S) x_i \mathrm{sign}_\tau(A_i^T y)]) A^T\| \\
&\lesssim \frac{k}{m} \|A\|^2
\end{aligned}$$

Recall that we *do not* assume here that $\|A\| = O(\|A^*\|)$. Instead, we have that

$$\begin{aligned}\|A\| &\leq \|A - A^*\| + \|A^*\| \\ &\leq \|A - A^*\|_F + \|A^*\| \\ &\leq \sqrt{m}\max_i \|A_i - {A^*}_i\|^2 + \|A^*\| \\ &= \sqrt{m}\delta + \|A^*\| \\ &= \mathrm{O}^*\left(\sqrt{m/(k+n)}\right) = \mathrm{O}^*\left(\sqrt{m/k}\right)\end{aligned}$$

Hence, $\frac{k}{m}\|A\|^2 = \mathrm{O}^*\left(\frac{k}{m}\cdot\frac{m}{k}\right) = \mathrm{O}^*(1)$, as needed. We note that if $\delta = \mathrm{o}\left(1/\sqrt{k}\right)$, then we would have $\frac{k}{m}\|A\|^2 = \mathrm{o}(1)$.

□

We emphasize that Proposition 4.1.3 *does not require that* $\|A^* - A\| \lesssim \|A^*\|$. Consequently, in many settings, the convergence proofs for the Projection Update Rule will *not require* us to prove $(\delta, 2)$ nearness at each step. Unfortunately, the generality of Proposition 4.1.3 comes at the cost of a small radius of convergence, as is adressed in the following remark:

Remark. The assumption that $\delta = o(1/\sqrt{k})$ is tight, as the following example shows. Suppose that $A^* \in \mathbb{R}^{n\times n}$ is the identity matrix, $S = [k]$, and that $A_S = A^*_S - \frac{\delta}{\sqrt{k}}1_k 1_k^T$, where 1_k and 1_n denote the ones vectors in $\mathbb{R}^k$ and $\mathbb{R}^n$ respectively. Then A is δ-close to A^*. However,

$$\|A_S(1_k/\sqrt{k})\| = (1 - \sqrt{k}\delta)(1_k/\sqrt{k}) \tag{1.14}$$

which can be made, arbitrarily close to 0 (but nonzero) by choosing $\delta = (1+\epsilon)/\sqrt{k}$ for some arbitrary $\epsilon > 0$. Hence, $A_S A_S^T$ and Q_S have the same span, but $\sigma_{\min}(A_S A_S^T) \leq \epsilon^2$. Thus, $Q_S \preceq A_S A_S^T$ holds only when C is as large as ϵ^{-2}.

We now state a corollary for C-lower bounded distributions:

Corollary. *If Assumptions 1- 4 and are satisfied, the x is C-lower bounded, and $\tau = C/2$, then*

$$G_i := \mathbb{E}_i[M_i^{\mathrm{prj}} y \mathrm{sign}_\tau(A_i^T y)] = \mathbb{E}_i[|x_i|](A^*_i - A_i) + \mathrm{O}^*\left(\|A^*_i - A_i\|\right) + \gamma \tag{1.15}$$

and hence g_i is $(\mathbb{E}_i[|x_i|], 0, n^{-\omega(1)})$-true. The result holds more generally as long as τ is (δ, C)-suitable.

Proof. When τ is (δ, C)-suitable the Lemma 3.1.4 ensures that $\mathrm{sign}_\tau(A_i^T y) = \mathrm{sign}_{C/2}(A_i^T y) = \mathrm{sign}(x_i)$ with probability $1-\gamma$ and $\hat{S} = S$ with probability $1-\gamma$ (even when conditioned on $i \in S$, since $i \in S$ occurs).

Now, write $G_i = G_{i,1} + G_{i,2}$. From Proposition 4.1.3, we have

$$G_{i,1} = \alpha_i(A^*_i - A_i) + o(\|A^*_i - A_i\|) + \gamma \tag{1.16}$$

where $\alpha_i = \mathbb{E}_i[x_i \text{sign}_{C/2}(A_i^T y)] = \mathbb{E}_i[x_i \text{sign}(x_i)] + \mathbb{E}_i[x_i(\text{sign}(x_i) - \text{sign}_{C/2}(A_i^T y)|] = \mathbb{E}_i[|x_i|] + \gamma$. By the same token,

$$\begin{aligned} G_{2,i} &= \mathbb{E}[M_i^{\text{prj}}(\hat{S}) y_{-i} \text{sign}_\tau(A_i^T y)] \\ &= \mathbb{E}[M_i^{\text{prj}}(S) y_{-i} \text{sign}(x_i)] + \mathbb{E}[\Big(M_i^{\text{prj}}(\hat{S}) \text{sign}_\tau(A_i^T y) - M_i^{\text{prj}}(S) \text{sign}(x_i)\Big) y_{-i}] \\ &= \mathbb{E}[\Big(M_i^{\text{prj}}(\hat{S}) \text{sign}_\tau(A_i^T y) - M_i^{\text{prj}}(S) \text{sign}(x_i)\Big) y_{-i}] \\ &= \gamma \end{aligned}$$

where the second to last line follows since $y_{-i} \perp x_i$ conditioned on $\text{supp}(x)$, the last from the fact that y_i is $O(k)$-subgaussian, while $\|M_i^{\text{prj}}(\hat{S}) \text{sign}_\tau(A_i^T y) - M_i^{\text{prj}}(S) \text{sign}(x_i)\|$ is $O(1)$ almost surely, and is 0 with probability $1 - n^{-\omega(1)}$. □

Applying Theorem 3.1.3, we immediately establish infinite sample convergence of the projection rule:

Theorem 4.1.6. *Suppose that A^* is δ-close to A^* for $\delta = \text{O}^*\left(1/\sqrt{k}\right)$, that the sparse coefficients x are C lower bounded, that the threshold $\tau = C/2$, the step size $\eta = \Theta^*(m/k)$, and Assumptions 1-4 hold. Then, if update step in the Meta Algorithm uses the projector matrix $M_i = X_i$ and an infinite number of samples per iteration, we have*

$$\|A_i^s - A_i^*\|^2 \leq (1 - \lambda)^s \|A_i^0 - A^*\|_i^2 + n^{-\omega(1)} \tag{1.17}$$

for some $\lambda \in (0, 1/2)$, and for any $s = 1, 2, \ldots, T$. In particular, A^* converges geometrically to an arbitrary inverse polynomial error.

Chapter 5

Learning Random NMF Instances with Dictionary Learning

5.1 NMF and NOID Learning

5.1.1 Motivation

In this section, we leverage the analytical and algorithmic machinery presented thus far to explain an experimentally well-documented, though hitherto theoretically unaccounted for phenomenom: that coordinate descent algorithms perform surprisingly well on certain randomly generated instances of *Nonnegative Matrix Factorization*, or *NMF*.

The rank-m NMF problem consists of expressing an entrywise nonnegative matrix $Y \in \mathbb{R}_{\geq 0}^{n \times p}$ as the product of two low rank, entrywise nonegative matrices $B \in \mathbb{R}_{\geq 0}^{m \times n}$ and $X \in \mathbb{R}_{\geq 0}^{m \times p}$. In contrast to dictionary learning, NMF is an *undercomplete factorization* in which $m \ll p, n$. NMF has been applied in numerous disciplines, ranging from image segmentation (Lee and Seung (2001)), to neuroscientific research (Cichocki et al. (2009)), to the famous "Netflix Problem" in recommendation systems (Koren et al. (2009)).

In its most general setting, Vavasis (2009) establishes that computing an exact Non-Negative Matrix Factorization is NP-Hard in general. Briefly, the proof consists is showing that NMF is equivalent to the "Intermediate Simplex" in polynomial combinatorics for which there is a polynomial time reduction from MAX 3-SAT.

One common complaint with Intermediate Simplex reduction is that it is exceedingly brittle. In practice, we do not expect NMF instances to be drawn adversarially. Indeed, there are many settings in which an NMF instance $Y = BX$ which is drawn from certain generative process, or that satisfies other favorable, structural assumptions, can be learned in polynomial time.

For example, Anandkumar et al. (2014) present a tensor spectral algorithm learning community structure in the Mixed Membership Stochastic Block Model can be intepreted as learning a symmetric factorization $X^T X$ from noisy observations $Y_{i,j} \sim \text{Bernoulli}(X_i^T X_j)$, where the columns of X are drawn from a Dirichlet distribution. Unfortunately, the spectral algorithm relies heavily on properties of Dirichlet moments, and is less robust when applied to real-world data.

If B satisfies the so-called *separability* condition from Donoho and Stodden (2003), Arora

et al. (2012b) and Recht et al. (2012) provide efficient and provably correct algorithms for recovering non-negative factorizations exactly in the noiseless setting, an approximately in the presence of noise. Arora et al. (2012a) and Bansal et al. (2014) extend the separable NMF literature to the "topic modeling" setting, where X is drawn from a suitable generative process, and Y_i is drawn a sparse discrete distribution whose expectation is BX_i.

Despite the recent advances in provably correct factorization algorithms, many of today's practioners still learn NMF instance with alternating descent algorithms like the Multiplicative Update Rule (MU) (Lee and Seung (2001)) and Hierachical Alternating Least Squares (HALS) (Cichocki et al. (2007)). Though these methods lack theoretical guarantees, numerical experiments find that coordinate descent algorithms converge to rapidly to local optima (Lin (2007)), and in one particular case, tend to learn exact factorizations when the components B and X are randomly generated (Vandaele et al. (2014)).

5.1.2 Our Contribution

While our ultimate goal would be to present rigorous guarantees for both the MU and HALS on a large class of NMF instances (note that demonstrating convergence for *all* instances is unlikely, as it would entail that $\mathbf{P} = \mathbf{NP}$), this chapter contents itself with a baby step in that general direction. We demonstrate conditions under which randomly generated NMF instances can be learned by the sorts of sparse coding gradient descent algorithms studied in this report. The hope is to establish that certain randomly generated NMF instances are "easy to factor", in the sense that the true factorization can be recovered approximately in polynomial time, with high probability. Furthermore, the approximate gradient approach in this chapter bears a much stronger resemblence to the MU and HALS procedures than the combinatorially and geometrically flavored algorithms introduced for learning separable factorizations.

The setting for our analysis is motivated by the numerical experiments in Vandaele et al. (2014), which generate both factor matrices B and X with sparse, random entries. Hence, we will treat X as a matrix whose columns correspond to samples from some favorable sparse distribution:

Definition 5.1.1. We say that the samples x are drawn from a (C, k)-*favorable distribution* if

1. x is entrywise nonnegative and C-lower bounded
2. x is k-sparse with probability $1 - n^{-\omega(1)}$
3. The entries of x are independent conditioned on their support, and $x_i \big| \text{supp}(x) = S$ has the same distribution for any $S \subset [m]$.
4. The support of S satisfies Assumption 2

We say that the samples are drawn from a (C, k, ρ)-*favorable distribution* if, in addition, the samples are ρ-smoothly distributed.

The full strength of our results will only hold for a more restricted class of distributions:

Definition 5.1.2 (Uniformly Favorable Distributions)**.** [1] We say that the samples x are drawn from a *uniformly* (C,k)*-favorable* (uniformly (C,k,ρ)-favorable)distribution if x is drawn from a (C,k) (resp. (C,k,ρ)) favorable distribution, $x_i\big|S$ has the same distribution for any $S \subset [m]$ containing i, if $q_{i,j} - q_i q_j \leq O(k^2/m)$, if $q_{i,j,r} - q_{ir}q_{ij} \leq \mathrm{O}\left(q_i k^2/m\right)$.

We say that the samples x are drawn from a *uniformly* (C,k)*-favorable (resp* (C,k,ρ)*-favorable), well-conditioned distribution* if, in addition, the covariance matrix $\Sigma := \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T]$ is well conditioned, in the sense that $\sigma_{\min}(\Sigma)$ and $\sigma_{\max}(\Sigma)$ are both $\Theta(k/m)$.

Example 5.1.1 (Uniform Supports yield Uniformly Favorable Distributions)**.** [2] Suppose that $\mathrm{supp}(x)$ is chosen uniformly. Then it is straightforward to verify that x are (C,k)-uniformly favorable. Moreover, if all $x_i\big|i \in S \sim \mathrm{Unif}([1,2])$, then the diagonals of Σ are

$$\Sigma_{ii} = \mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2 = (1 - \mathrm{o}\,(1))\mathbb{E}[x_i]^2 \tag{1.1}$$

where

$$\mathbb{E}[x_i^2] = \Pr(i \in S) \cdot \mathbb{E}_{Z\sim\mathrm{Unif}([1,2])}[Z^2] = \frac{7k}{3m} \tag{1.2}$$

The off diagonals Σ_{ij} are given by

$$\mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j] = \{\Pr(i \in S) - \Pr(i,j \in S)\}\mathbb{E}_{Z\sim\mathrm{Unif}([1,2])}[Z]^2 = \frac{9k}{4m}(\frac{k}{m} - \frac{k-1}{m-1}) \tag{1.3}$$

which have absolute value no more than $\frac{9k}{4m^2}$. Hence, $\sum_{ij}|\Sigma_{ij}| \leq (1 - \Omega(1))\Sigma_{ii}$, so by the Gershgorin Circle Theorem, we can conclude that Σ is well conditioned. Hence, x come from a uniformly (C,k)-favorable, well-conditioned distribution.

It is straightforward to show that the result holds more generally when the support of x is uniformly, and all of $x_i\big|i \in S$ satisfy $\mathbb{E}[x_i^2\big|i \in S] \geq (1 + \Omega(1))\mathbb{E}[x_i x_j\big|i,j \in S]$, e.g. example $\mathbb{E}[x_i\big|i \in S]$ is identically distributed for $i \in [m]$, and is not a point mass.

Remark. We remark that we can prove many modified versions of Theorems 5.1.1 and 5.1.3 under under various modified assumptions. While we will not discuss such other settings at length,

We will now deviate from Vandaele et al. (2014) by considering the case where B is non-sparse and non-negative, but nevertheless satisfies a sort of "hidden" incoherence property, called Nonnegative Offset Incoherence, defined in the following section. The reason is two fold: first, leveraging a modified version of incoherence makes it more straightforward to directly map ideas from sparse coding onto the NMF setting. Second, many alternating descent algorithms report rapid convergence to good local optima on *dense* factor matrices, and we hope to provide some insight into why this is possible. Hence, our setting can be viewed as sort of hybrid between Vandaele et al. (2014) and Lin (2007), where one factor matrix B is dense, and the other factor X is sparse. While we can establish a range of partial results for NMF instance with (C,k)-favorable distributions, we were only able to verify a complete polynomial time algorithm for learning generative NMF instances in the setting where the latent samples are uniformly (C,k,ρ)-favorable:

[1] This definition was recently modified since the May 4th draft to adress some errors in the earlier version of that work.

[2] This example was added after the May 4th draft was submitted

Theorem 5.1.1 (Provable Algorithm for Random NMF Instances)**.** *Given samples $y = B^*x$, where x^* has uniformly (C, k, ρ)-favorable, well-conditioned distribution, $B^* \in \mathbb{R}^{n\times m}$ is a μ-incoherent NOID (see section 5.1.3) for which $\frac{\mu}{\sqrt{n}} = \mathrm{O}^*\left(k^2 + k\log n\right)$, and $k = \mathrm{O}^*\left(\min\{m^{2/5}, m/\log n\}\right)$, then there is an algorithm with runtime and sample complexity $O(poly(m, n, k, \log(1/\epsilon))$ which learns a non-negative matrix B such that $\|B^* - B\|_F \leq \epsilon\|B^*\|_F$ for any inverse polynomial error ϵ.*

The intutition here is that, even though the problem superficially resembles an NMF, it is actually a disguised dictionary learning problem, which can be extracted by Algorithm 4, and then learned by the algorithm in Theorem 5.2.3.

Proof of Theorem 5.1.1. By Theorem 5.1.3, we may reduce to S-NDL where the matrix A^* is $\sqrt{n}/m + \mu$-incoherent and our desired tolerance is $\delta = \epsilon/\sqrt{m(k + \log n)}$. It is straightforward to check that, under the assumptions of the present theorem, A^* is sufficiently incoherent, and x^* are sufficiently sparse to satisfy the assumptions of Theorem 5.2.3. The latter theorem yields a polynomial time algorithm to learn A^* up to arbitrary inverse polynomial precision, and the geometric convergence ensured thereby guarantees that the run time and sample complexity is logarithmic in the desired precision. □

Remark. An analogue of the above theorem can be demonstrated in the case where the samples are not uniformly (C, k, ρ), but are far more sparse. In fact, the scalings in the above theorem depend on the expected norms of various covariance matrices, and it would be possible (in, say a later work) to provide a statement of Theorem 5.1.1 which interpolates between the norm of these covariance matrices and the required sparsity k/m. Moreover, we could also relax some assumptions on the covariance matrix Σ (for example, our reduction can achieve polynomial complexity and run-time if $\sigma_{\min}(\Sigma) = \Omega(n^{-C})$ for some fixed constant $C > 0$).

For example, if one modifies Algorithm 7 to not subtract off sample means, and sacrifices a factor of k in the sample complexity and precision, one can prove recovery guarantees in the setting where the *second moment matrix* rather than the covaraince matrix satisfies

$$\Omega\left(k/m\right) \cdot I \preceq \mathbb{E}[xx^T] \preceq \mathrm{O}\left(k^2/m\right) I \tag{1.4}$$

Such bounds are easy to establish when the x's have a (C, k)-favorable distribution, and the off-diagonal co-ocurence matrix Q with $Q_{ij} = q_{ij}$ if $i \neq j$ and $Q_{ii} = 0$ otherwise is rank one. In particular, this holds when the x's have a (C, k)-favorable distribution, $k^2 = \mathrm{O}^*\left(m\right)$, and the supports are chosen uniformly at random (or just that q_{ij} is the same for all $i \neq j$).

We also mention that the author of this report was able to establish some results for the case where both factors are sparse, but the tools are slightly different and not included in this report in the interest of concision. The setting where B and X are dense seems far less promising to analyze, since the absence of sparsity and non-convexity of the NMF objective lead to issues of identifiability in both theory (see Donoho and Stodden (2003)) and in practice (Lin (2007)).

Differences Between Dictionary Learning and Conventional NMF Algorithms

Although MU, HALS, and the approximate gradient descent algorithms for sparse coding all can be viewed as forms of gradient descent, we emphasize MU and HALS updates differ

substantially from the algorithms in this report. For one, MU and HALS iterate over the rows of B, and while sparse coding moves column-wise. But there is a more general difference. Because general matrix factorizations of the form $Y = BX$ are non-unique, the common motivation behind both sparse coding and and NMF is to impose meaningful structural contraints which can ensure uniqueness. In sparse coding, the update rules apply thresholds $\text{sign}_\tau(u) = \text{sign}(u)\mathbb{1}(|u| > \tau)$ to leverage the knowledge that the latent samples are sparse; in MU and HALS, the update rules instead leverages the structural knowledge that the entries of B and X are nonzero.

5.1.3 Offset Incoherent Dictionaries

In general, we should not expect non-negative matrices to have incoherent columns. Indeed, consider a matrix $B \in \mathbb{R}^{n\times 2}$, where $B_{pq} \sim \text{Bernoulli}(1/2)$. Then $\mathbb{E}[\langle B_i, B_j\rangle] = \frac{n}{4}$, while $\mathbb{E}[\|A_i\|^2] = \mathbb{E}[\|A_j\|^2] = n/2$. With a couple Chernoff bounds, we can conclude that $\cos(B_i, B_j) = \Theta(1)$ with $e^{-\Theta(n)}$ probability.

On the other hand, it is well know that dictionaries with light-tailed, independent *mean - zero* entries are $\tilde{O}(1)$ incoherent with high probability Candès and Wakin (2008). Consequently, if B has indepedent subgaussian entries, $A := B - \mathbb{E}[B]$ should be sufficiently incoherent with high probability. If the entries of B are iid, or more generally if the expectations of the columns $\mathbb{E}[B_i]$ are scalar multiples of one another, then $\mathbb{E}[B]$ is rank one. With these sorts of generative processes in mind, we introduce consider a generalization of dictionary learning where we learn dictionaries which are rank-one perturbation away from being incoherent:

Definition 5.1.3 (Offset Incoherent Dictionaries)**.** We say that a matrix $B \in \mathbb{R}^{n\times m}$ is an Offset Incoherent Dictionary (OID) with parameter μ if B can be written in the form $A+vc^T$ for any columns A_i, A_j of A, $|\cos(A_i, A_j)| \le \mu/\sqrt{n}$ and $|\cos(A_i, v)| \le \mu/\sqrt{n}$. We say that B is an Nonnegative-Offset Incoherent Dictionary (NOID) if c is entrywise nonnegative.

Because we want to reconstruct non-negative factors, we shall require a sign-sensitive notion of closeness: We shall also need the notion of signed-closeness:

Definition 5.1.4. [δ-signed-close] We say that A is δ-signed-close to A^* if there is a permutation $\pi : [m] \to [m]$ for which $\|A_{\pi(i)} - A_i^*\| \le \delta$ for all $i \in [m]$. We note that we do now allow ourselves to flip the signs of A_i.

With these two definition in place, we set up the Offset-Incoherent Nonnegative Dictionary Learning problem accordingly:

Problem 5.1.1. **δ-Nonnegative-Offset-Incoherent-Dictionary Learning** (δ-NOIDL): Learn an estimate B which is δ-signed-close to a NOID $B^* = A^* + cv^T \in \mathbb{R}^{n\times m}$ with incoherence parameter μ, given samples $y = B^*x$, where x are drawn from a (C, k)-favorable nonnegative distribution.

As we shall soon see, the decomposition $B = A + cv^T$ for offset incoherent matrices is not unique in general, and different decomposition parameters yield different incoherence paramters. Nevertheless, if we compute a naive average $\hat{v}$ of samples $y = B^*x$, then with high probability

$$\tilde{A}^* := \text{Normalize}(\text{Proj}_{\hat{v}^\perp} B^*) \text{ is } O\left(\frac{\sqrt{n}}{m} + \mu\right) \text{ incoherent} \tag{1.5}$$

Hence, learning $\tilde{A}^*$ from samples $\tilde{y} := \text{Proj}_{\hat{v}^\perp} y$ will amount to an instance of what we call Semi-Nonnegative Dictionary Learning (δ-S-NDL):

Problem 5.1.2. Semi-Nonnegative Dictionary Learning (δ-S-NDL) Learn an estimate A which is δ-signed-close to a μ' incoherent dictionary A^* for which $\|A^*\| = O(1)$ given samples $y = Ax$, where x are drawn from a (C, k)-favorable nonnegative distribution.

Once we have an accurate columwise estimate of A^*, then we can decode random samples $y = B^*x$ by feeding in $\text{Proj}_{\hat{v}^\perp} y$ Algorithm 6, and effectively "invert" those decodings to learn the original dictionary using Algorithm 7. The precise reduction is described in Algorithm 4. In sum, we will be able to demonstrate a reduction from OI-NDL reduces to S-NDL, which we state informally as follows, which we formalize in section 5.1.4

Non-Uniqueness of the NOID Decomposition

Proposition 5.1.2. *Let $1_d \in \mathbb{R}^d$ denote the vector of all ones, draw a matrix $B^* \in \mathbb{R}^{n\times m}$ be $B^* \overset{iid}{\sim} \mathcal{N}(1, 1/n)$, set $A^* = B^* - 1_n 1_m^T$. Then with high probability, A^* is $10\log(n)$ incoherent and B^* is entrywise nonnegative. Moreover, with probability at least $3/16$, there exists another vector v such that*

$$\|v - 1_n\|^2 \geq \frac{1}{2048m} \text{ and } B^* - v1_m^T \text{ is } \ 10\log(n) \ \text{ incoherent} \tag{1.6}$$

Proof. This proposition is really just a statement about estimating the means of m Gaussian vectors in $\mathbb{R}^n$. To make this more precise, fix an $\alpha > 0$ to be chosen later, and let $\Theta_0 = \Theta_0(\alpha) = \{w \in \mathbb{R}^n : 0 \leq w_i \leq \frac{\alpha}{\sqrt{4n}} \forall 1 \leq i \leq n\}$, and let $\Theta = 1_n + \Theta_0 = \{1_n + w : w \in \Theta_0\}$. Let $A \in \mathbb{R}^{n\times m}$ be a random matrix with entries $A_{i,j} \sim \mathcal{N}(0, 1/n)$, and for all $\theta \in \Theta$ and let $B_\theta := A + \theta 1_m^T$. Note that the matrix B^* correspond simply to B_{1_n}.

Now let $\mathcal{P}$ be the statistical procedure which exhaustively searches over all $v \in \mathbb{R}^n$, and returns any v for which the following decomposition is satisfied:

$$\begin{aligned} B_\theta &= \tilde{A} + v1_m^T \\ &s.t.\ \tilde{A} \text{ is } 10\log n - \text{incoherent} \end{aligned} \tag{1.7}$$

Furthermore, let $p_\theta(C_1)$ denote the probability that Equation 1.7 is satified for θ, and is only for vectors $v : \|v - \theta\|^2 \leq \frac{1}{C_1 m}$. Since B_θ depends linearly in θ, we can see that $p_\theta(C_1)$ is independent of θ, so we will write $p(C_1)$. Then if $p(C_1) \geq p^*$, $\mathcal{P}$ successfully returns a vector $\hat{\theta} : \|\hat{\theta} - \theta\|^2 \leq \frac{1}{C_1 m}$.

In other words, $\mathcal{P}$ estimates the mean of m Gaussian random vectors in $\mathbb{R}^n$ with variance $\sigma^2 = 1/n$. However, it is shown in Rigollet (2014) that for for any $\alpha > 0$ and any estimator $\hat{\theta}$ of θ.

$$\sup_{\theta \in \Theta_0(\alpha)} \Pr_\theta \left(\|\theta - \hat{\theta}\| \geq \frac{\alpha}{256} \cdot \frac{\sigma^2 n}{m} \right) \geq \frac{1}{2} - 2\alpha \tag{1.8}$$

Translating Θ_0 by Θ and plugging in $\alpha = 1/8$ and $\sigma^2 = 1/n$ gives that

$$\sup_{\theta \in \Theta_0(1/8)} \Pr_\theta \left(\|\theta - \hat{\theta}\| \geq \frac{1}{2048m} \right) \geq \frac{1}{4} \tag{1.9}$$

Hence, it follows that with probability at least $\frac{1}{4}$, then $\mathcal{P}$ fails. Thus, either θ does not satisfy the decomposition in Equation 1.7, or there is another vector v satisfying the decomposition in Equation 1.7 such that $\|v-\theta\| \geq \frac{1}{2048m}$. However, A is $10\log n$ incoherent with very high probability by Proposition C.1.10, so θ satisfies the decomposition given by Equation 1.7 with probability at least, say $1/16$. Hence, with probability at least $3/16$, the decomposition in Equation 1.7 holds for any v of distance at least $\sqrt{\frac{1}{2048m}}$ away from θ. To conclude the proof, note that the entries of A are bounded by $n^{-1/4}$ with very high probability, and hence B_θ is nonnegative for all $\theta \in \Theta$ with very high probability. □

5.1.4 Formalizing The Reduction

To formalize the reduction, we need to introduce a bit more notation:

Definition 5.1.5 ((C,k)-Favorable Distribution)**.** We say that $\mathcal{O}$ is (C,k)-favorable (resp (C,k,ρ)-favorable) nonnegative sample oracle for a matrix A if $\mathcal{O}$ can be queried for samples $y = Ax$, where x is are drawn independently from a (C,k) favorable distribution $\mathcal{D}$. We define uniformly (C,k)-favorable and uniformly (C,k,ρ)-favorable oracles similarly. We write $y \sim \mathcal{O}$ denote that y is drawn from the oracle $\mathcal{O}$. We will use the notation $\tilde{\mathcal{O}} = f(\mathcal{O})$ to denote that the $\tilde{\mathcal{O}}$ is the oracle which returns samples $f(y)$.

Moreover, we give the definition of an *approximate S-NDL* algorithm:

Definition 5.1.6 (Uniformly (C,k)-Favorable Distribution)**.** We say that an algorithm $\mathcal{A}$ is an (μ,δ_0)-approximate S-NDL algorithm if, given a (C,ρ,k)-favorable nonnegative sample oracle $\mathcal{O}$ for a μ incoherent matrix A^* and an $\delta \leq \delta_0$, then there is a constant C for which $\mathcal{A}(\mathcal{O},\delta)$ returns a dictionary $\tilde{A}$ whose columns are δ-signed-close to those of A^* with high probability, with run time and sample complexity on the order of $O\left(n^C \mathrm{polylog}(\delta)\right)$.

We now present the reduction which formalizes the steps in section 5.1.3 It precise guarantees are given by the following theorem:

Algorithm 4: Reduction from NOIDL to S-NDL

Data: (C,k)-favorable oracle $\mathcal{O}$ for μ-NOID Dictionary $B^* = A^* + vc^T$; Tolerance δ; $(\delta_0, O^*(\mu+\sqrt{m}/n)$-Approximate S-NDL Algorithm $\mathcal{A}$
Result: Nonnegative matrix B which is $O(\delta)$-close to B^* up to a $\Theta(1)$ re-scaling of columns
Get the S-NDL oracle $(\tilde{\mathcal{O}}, \hat{v}) \leftarrow$ **AverageInit**$(\mathcal{O}, p_1)$
for dictionary $\tilde{A}^*$: Normalize$(\mathrm{Proj}_{\hat{v}^\perp}(B^*))$, with $p_1 = \tilde{\Omega}(m/k)$
Learn approximate dictionary $A \leftarrow \mathcal{A}(\tilde{\mathcal{O}}, \delta)$
$(Y, \hat{X}) \leftarrow$ **Decode**$(A, \mathcal{O}, .8C, p_2)$ for $p_2 = \tilde{\Omega}(m)$
return $B \leftarrow$ **InvertAndThreshold**$(Y, \hat{X})$

Theorem 5.1.3 (Analysis of Reduction from NOIDL to S-NDL)**.** *Given an $(\mu/\sqrt{n}, \delta_0)$-approximate S-NDL algorithm, then for all $\delta \leq \delta_0$ for which $\delta = \mathrm{o}\left(1/k + 1/\sqrt{k}\log n\right)$, then, given a sample $y = B^*x \sim \tilde{O}$, the decoding step in Algorithm 3 can recover x up to an error of $\delta(k+\sqrt{k}\log n)$ with high probability.*

Furthermore, if the spectrum of the empirical covariance matrix $\Sigma^* := \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T$ *is* $\Theta(k/m)$*, and* $\delta = \mathrm{o}\left(1/m + \sqrt{k}/m \log n\right)$*, then Algorithm 4 returns an estimate* B *of* B^* *which is non-negative, and satisfies*

$$\|B - B^* D\|_F \lesssim \delta \sqrt{m(k + \log n)} \|B^*\|_F \tag{1.10}$$

where D *is a diagonal matrix with* $\Theta(1)$ *entries along its diagonal.*

The proof follows straightforwardly from combining the guarantees for the three subroutines AverageInit, Decode, and InvertSample provided in Theorems 5.1.4, 5.1.7, 5.1.8 respectively. We remark that learning B^* up to a $\Theta(1)$ rescaling of its columns is an artificact of our proof, but makes little difference due to the non-identifiability of the model up to scalings of the samples entries/dictionary columns by constants.

5.1.5 Analysis of the Averaging Step

This section presents an analysis of the Averaging Algorithm: The exact guarantees are

Algorithm 5: AverageInit$(\mathcal{O}, p)$

Data: (C, k)-favorable oracle $\mathcal{O}$ for $\mu/\sqrt{n}$-NOID B^*; Sample Number p
Average: Estimate rank one component by $\hat{v} = \frac{1}{p}\sum_{j=1}^{p} y^{(j)}$, where $y^{(j)} \sim \tilde{O}$
Project: Return an S-NDL sample oracle $\tilde{\mathcal{O}} := \mathrm{Proj}_{\hat{v}^\perp}(\mathcal{O})$
return $(\tilde{\mathcal{O}}, \hat{v})$

as follows:

Theorem 5.1.4. *Suppose that* B^* *is a* $\mu/\sqrt{n}$ *is a NOID with decomposition* $B^* = A^* + vc^T$*, and that* $\mathcal{O}$ *is a* (C, k)*-favorable nonnegative oracle for* B^**, where all* $c_i = \Theta(1)$*. Let* $\hat{v} = \frac{1}{p}\sum_{j=1}^{p} y^{(j)}$*. Then then*

$$\tilde{A}^* := \mathrm{Normalize}(\mathrm{Proj}_{\hat{v}^\perp} B^*) \text{ is an } O\left(\frac{\sqrt{n}}{m} + \mu\right) \text{ incoherent dictionary} \tag{1.11}$$

and the oracle $\tilde{\mathcal{O}} := \mathrm{Proj}_{\hat{v}^\perp}\mathcal{O}$ *is a* $(.9C, \rho, k)$*-favorable nonnegative oracle for* $\tilde{A}^*$*. Furthermore,* $\|\tilde{A}^*\| \le 1.2\|A^*\|$*.*

The proof of Theorem 5.1.4 hinges on a more fundamental, though later elementary proposition about the average of columns of a NOID. We state a simplified version of the proposition here; the proposition is expressed in its more precise form in Proposition B.1.1.

Proposition 5.1.5. *Suppose that* B *is a nonnegative offset incoherent dictionary with decomposition* $B = A + vc^T$ *and incoherence parameter* μ *for which* $\|A\|_i = \Theta(1)$ *adn* $c_i = \Theta(1)$*. for all* $i \in [m]$*, and* $\|v\| = \Theta(1)$*. Define* $\hat{v} = \sum_i B_i$ *and* $P := Proj_{\hat{v}^\perp}$*. Then if* $\frac{1}{m} + \frac{\mu}{\sqrt{n}} = \mathrm{O}^*(1)$*, then*

$$\cos(PB_i, PB_j) \le \mathrm{O}\left(\frac{\mu}{\sqrt{n}} + \frac{1}{m}\right) \tag{1.12}$$

and $\|PB_i\|^2 \ge (9/10)\|A_i\|^2$*.*

The next lemma leverages Proposition 5.1.5 to prove Theorem 5.1.4 for deterministic, weighted averages of the columns of B^*:

Lemma 5.1.6. *Suppose that B^* is a $\mu-$ is a NOID with decomposition $B^* = A^* + vc^T$, and that $\mathcal{O}$ is a (C,k)-favorable nonnegative oracle for B^*. Then if $\hat{v} = \frac{1}{m}\sum_{i=1}^m w_i B_i^* + \epsilon$. Then if $\|\epsilon\| = o(1/m)$, $c_i = \Theta(1)$, $w_i = \Theta(1)$, and $\max(\frac{1}{m}, \mu/\sqrt{n})$ is bounded above by some universal constant, then*

$$\tilde{A}^* := \text{Normalize}(\text{Proj}_{\hat{v}^\perp} B^*) \text{ is an } O\left(\mu + \sqrt{n}/m\right) \text{ incoherent dictionary} \tag{1.13}$$

and the oracle $\tilde{O} := \text{Proj}_{\hat{v}^\perp}\mathcal{O}$ is a $(.9C,k)$-favorable nonnegative oracle for $\tilde{A}^*$. Furthermore, $\|\tilde{A}^*\| \le 1.2\|A^*\|$

Proof. We prove the result for when $\epsilon = 0$; elementary computation extend to the case with error. We write

$$B' = B^*\text{diag}(w_i) = A^*\text{diag}(w_i) + (c \circ w)v_i^T \tag{1.14}$$

and note that $\sum_i B_i' = \hat{v}$. Then, since $\|A^*\| = 1$, and both $c_i = \Theta(1)$ and $w_i = \Theta(1)$, then, B' is $(\Theta(1), \Theta(1))-$well conditioned. Consequently, if $\max(\frac{1}{m}, \mu/\sqrt{n})$ is sufficently small, it follows from Proposition B.1.1 that the columns of $A' := \text{Proj}_{\hat{v}^\perp} C$ are at least $\sqrt{9/10}w_i$ in magnitude, and that $\cos(A_i', A_j') = O(\frac{1}{m} + \frac{\mu}{\sqrt{n}})$. Hence, $\tilde{A} = \text{Normalize}(A_i)$ is a $O(\frac{1}{m} + \frac{\mu}{\sqrt{n}})$ dictionary. Furthermore, the samples $\tilde{y} := \text{Proj}_{\hat{v}^\perp}\tilde{O}$ can be expressed as

$$\begin{aligned}
\tilde{y} &= \text{Proj}_{\hat{v}^\perp} B^* x \\
&= \text{Proj}_{\hat{v}^\perp} B'\text{diag}(w_i)x \\
&= A'\text{diag}(w_i)x \\
&= \tilde{A}\text{diag}(\|A'\|_i^{-1} w_i)x \\
&= \tilde{A}\tilde{x}
\end{aligned}$$

where $\tilde{x} = \text{diag}(\|A'\|_i^{-1}w_i)x$. Since $\|A_i'\| \in [\sqrt{9/10}w_i, w_i]$, $\|A'\|_i^{-1}w_i \in [\sqrt{9/10}, 1] \subset [.9, 1]$, so $\tilde{x}$ has a $(.9C,k)$-favorable nonnegative distribution. □

We are now ready to complete the proof of Theorem 5.1.4

Proof of Theorem 5.1.4. Projecting onto $\hat{v}$ is equivalent to projecting onto a rescaling $\frac{1}{k}\hat{v} = \frac{1}{kp}\sum_{j=1}^p y^{(j)} = \frac{1}{m}\sum_{i=1}^m B_i^*(\frac{m}{kp}\sum_{j=1}^p x_i^{(j)})$. In light of Lemma 5.1.6, it suffices to show that, with high probability, $\frac{m}{kp}\sum_{j=1}^p x_i^{(j)} = \Theta(1)$ for all $i \in [m]$. To this end, fix an i in m, and let $W = \{j : x_i^{(j)} \neq 0\}$.

By a Chernoff bound, $pq_i/2 \le |W| \le 2pq_i$ with high probability given $p = \tilde{\Omega}(k/m)$ samples, where we recall that $q_i = \Pr(i \in \text{supp}(x))$. Moreover, if $|W| = \tilde{\Omega}(1)$, then Subgaussian concentration yields that $\sum_{i\in W} x_i^{(j)} \in [|W|\mathbb{E}_i[x_i]/2, 2|W|\mathbb{E}_i[x_i]]$ with high probability. Consequently,

$$\frac{m}{kp}\sum_{j=1}^p x_i^{(j)} = \frac{m}{kp}\sum_{j=1}^W x_i^{(j)} \in \left(\frac{m}{kp}pq_i\mathbb{E}_i[X_i]\right)\cdot[1/4, 4] = \left(\frac{mq_i}{k}\mathbb{E}_i[X_i]\right)\cdot[1/4,4] \tag{1.15}$$

The result now follows since $\frac{mq_i}{k}\mathbb{E}_i[X_i] = \Theta(1)$. We also remark that, by increasing the number of log factors in the sample size if necessary, we can even ensure that

$$\frac{m}{kp}\sum_{j=1}^{p} x_i^{(j)} \in \left(\frac{mq_i}{k}\mathbb{E}_i[X_i]\right)\cdot[1-o(1),1+o(1)] \tag{1.16}$$

with high probability. □

5.1.6 Analysis of Decoding Step

This section uses standard results from perturbation analysis to establish the efficacy of the decoding step:

Algorithm 6: Decode($\mathcal{O}, A, C, p$)

Data: (C,k)-favorable oracle $\mathcal{O}$ for Incoherent Dictionary A^*; Estimate A which is δ-close to A^*, Sample Number p

Initialize $Y \in \mathbb{R}^{n\times p}, X \in \mathbb{R}^{m\times p}$;

for $j = 1, 2, \dots, p$ **do**

Sample $y^{(j)} = A^*x^{(j)} \sim \mathcal{O}$, and set $Y_j = y^{(j)}$

Estimate $\mathrm{supp}(x^{(j)})$ by $\hat{S} = \{i : |A_i^T y^{(j)}| > C/2\}$

Estimate $x^{(j)}$ by first letting $\hat{x} = ((A_{\hat{S}})^\dagger y^{(j)})$, then setting all the entries of $\hat{x}$ which not in $\hat{S}$ to zero, and letting $\hat{X}_j = \hat{x}$

return $(\hat{X}, Y)$

Theorem 5.1.7 (Decoding)**.** *Let $\mathcal{O}$ be (C,k)-favorable oracle for $\tilde{A}^*$, and suppose that A is δ-(signed)-close to A^* for $\delta = \mathrm{o}\left(1/\sqrt{k}\right)$. If $x \sim \mathcal{O}$ and $\hat{x}$ as computed in Algorithm 6, then*

$$\|\hat{x} - x\| = \mathrm{O}\left(\sqrt{k}\|x_i\|\delta\right) = \mathrm{O}\left((k+\sqrt{k}\log n)\delta\right) \tag{1.17}$$

and $\mathrm{supp}(\hat{x}) \subset \mathrm{supp}(x)$ with high probability. Hence, if Algorithm 6 returns the decoding pair $(\hat{X}, Y)$, where $Y = A^*X$, then $\hat{X}$ is $(k+\sqrt{k}\log n)\delta$-(signed)-close to $\hat{X}$. We remark here that the scalings of the entries of x and $\hat{x}$ are as to ensure that the columns of $\tilde{A}^*$ have unit norm.

Proof. First we show that $\|A_S^\dagger - (\tilde{A}^*_S)^\dagger\| \le \sqrt{k}\delta$. Using a standard matrix perturbation analysis (see Lemma 2.7.1 in Golub and Van Loan (2012)) , and the fact that $\kappa(\tilde{A}^*_S) = \sigma_{\max}(A^*_S)/\sigma_{\min}(A^*_S) = \Theta(1)$ by Lemma D.1.1, it holds that

$$\|\|A_S^\dagger - (A^*_S)^\dagger\| \le \mathrm{O}\left(\|A_S - A^*_S\|\right) \tag{1.18}$$

whenever $\|A_S - \tilde{A}^*_S\| = \mathrm{o}(1)$. Since A is δ-close to $\tilde{A}^*$, then $\|A_S - \tilde{A}^*_S\| \le \sqrt{k}\delta = o(1)$ for $\delta = o(1/\sqrt{k})$. To conclude, note that with high probability $S = \hat{S}$. Furthermore, let set $\mathcal{S}$ of all vectors in $\mathbb{R}^m$ whose support is S is convex, and so

$$\|\mathrm{Proj}_{\hat{S}}\hat{x}\| \overset{whp}{=} \|\mathrm{Proj}_S\hat{x} - x\| \le \|\hat{x} - x\| \tag{1.19}$$

But $\|\mathrm{Proj}_{\hat{S}}\hat{x}\|$ is precisely the operation which sets to zero all entries of $\hat{x}$ which are not in $\hat{S}$. The bound on $\|x_S\|$ follows from B.3.1. □

Remark. [3] Since we are working in the under complete setting, we could also naively decode samples Y without sign-thresholding by taking $\hat{X} = A^\dagger Y$. Using similar arugments, we can show that if $\delta = \mathrm{o}(1/m)$, then $\|\hat{X} - X\| \le m\delta\|X\|$, where X is the matrix whose j-th column is the j-th sparse coefficient $x^{(j)}$. It turns out that this method of decoding can be used to prove slightly sharper guarantees than those in Theorem 5.1.3. Nevertheless, the fact that the naive decoding (wihout sign-thresholding) foregoes sign thresholding may risk greater sensitivity to noise, and so we stick with the more robust decoding scheme given by Algorithm 6.

5.1.7 Analysis of the Inversion Algorithm

Given decodings of label samples, we use the following algorithm to recover the non-negative dictionary B^*:

Algorithm 7: InvertSample($(Y, \hat{X})$)

Data: Approximately Decoded Pair $(Y, \hat{X}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^{m\times p}$
Average: Set $\bar{y} \leftarrow \frac{1}{p}\sum_{i=1}^p Y_i$ and $\bar{x} \leftarrow \frac{1}{p}\hat{X}_i$.
Subtract Means: Set $U = Y - \bar{y}1^T$ and $\hat{W} = \hat{X} - \bar{x}1^T$.
Inverset: $B \to W\hat{U}^\dagger$
Threshold: For $i \in [m]$, $B_i \to \mathrm{Proj}_{\mathbb{R}^n_{\ge 0}}(B_i)$

The analysis of this subroutine is as follows:

Theorem 5.1.8. *Suppose that we are given N samples $x^{(1)}, \dots, x^{(N)}$ with approximate decoding $\hat{x}^{(1)}, \dots, \hat{x}^{(N)}$ for which $\max_j \|x^{(j)} - \hat{x}^{(j)}\| \le \eta$. Let $\hat{W}$ is the matrix whose j-th column is $\hat{x}^{(j)} - \frac{1}{N}\sum_{r=1}^N \hat{x}^{(r)}$, and U be the matrix whose j-th column is $y^{(j)} - \frac{1}{N}\sum_{r=1}^N y^{(r)}$).*

Then, if $\eta = \mathrm{o}(k/m)$, $N = \tilde{\Omega}(m)$, and all the eigenvalues of the population covariance matrix $\Sigma^* := \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T$ are $\Theta(k/m)$, then with high probability

$$\|\tilde{B^*} - U\hat{W}^\dagger\|_F \lesssim \eta\sqrt{m/k}\|B^*\|_F \tag{1.20}$$

where $\tilde{B^*}$ is obtained by rescaling the columns of B^* by the diagonal matrix $D := \mathrm{diag}(\|\mathrm{Proj}_{\hat{v}}B^*\|)$, where $\hat{v}$ is as defined as in Algorithm 5. Recall from Theorem 5.1.4 that the spectrum of D is $\Theta(1)$ with high probability

Proof. First, define the sample average $\mu = \frac{1}{N}\sum_j x^{(j)}$ is the sample average of the hidden samples, the empirical covariance matrix $\tilde{\Sigma} = \frac{1}{N}\sum_j (x^{(j)} - \mu)(x^{(j)} - \mu)^T$. The following lemma establishes that $\tilde{\Sigma}$ is well-conditioned

Lemma 5.1.9. *Suppose that Σ^* is well conditioned, in the sense that there are constants $0 < c_1 \le c_2$*

$$\frac{c_1 k}{m} I \preceq \Sigma^* \preceq \frac{c_2 k}{m} I \tag{1.21}$$

[3] This remark was added after the May 4th draft

Then, if $N = \tilde{\Omega}(m)$, $\tilde{\Sigma}$ *is also well conditioned, in the sense that*

$$\frac{c_1' k}{m} I \preceq \tilde{\Sigma} \preceq \frac{c_2' k}{m} I \tag{1.22}$$

for possibly different constants $0 < c_1' \leq c_2'$

The proof relies on a matrix concentration bound from Vershynin (2010), and we defer the proof to the appendix. Now let W be the matrix whose j-th column is $(x^{(j)} - \mu)/\sqrt{N}$. It is easy to verify that if $\|x^{(j)} - \hat{x}^{(j)}\| \leq \eta$, then $\|W_j - \hat{W}_j\| \leq 2\eta/\sqrt{N}$, so that $\|W - \hat{W}\| \leq 2\eta$.

Next, note that WW^T is precisely Σ. Hence, given $N = \tilde{\Omega}$ samples, it follows that $\sigma_{\min}(W)$ and $\sigma_{\max}(W)$ are both $\Theta(k/m)$. Hence, as long as $\eta = \mathrm{o}(\sigma_{\min}(\|W\|)) = \mathrm{o}(k/m)$, we have by the perturbation analysis from Lemma 2.7.1 in Golub and Van Loan (2012) yields

$$\begin{aligned}
\|\hat{B} - \tilde{B}^*\|_F &= \|U(\hat{W}^\dagger - \hat{W}^\dagger)\|_F \\
&= \|\hat{W}^\dagger - (\hat{W})^\dagger\|_2 \|U\|_F \\
&\lesssim \|W\|^{-2} \|W - \hat{W}\|_2 \|\tilde{B}^* W\|_F \\
&\leq \|W\|^{-1} \|W - \hat{W}\|_2 \|\tilde{B}^*\|_F \\
&\leq \|W\|^{-1} \eta \|\tilde{B}^*\|_F \\
&\lesssim \eta \sqrt{m/k} \|B^*\|_F
\end{aligned}$$

To conclude, we note that $\tilde{B}^*$ lies in the convex set of nonnegative real matrices, and hence $\|\mathrm{Proj}_{\{\mathbb{R}^{n\times m}_{\geq 0}\}} \hat{B} - \tilde{B}^*\|_F \leq \mathrm{Proj}_{\hat{B}} \|\hat{B} - \tilde{B}^*\|_F$ since the Frobenius norm is a Euclidean distance. □

5.2 Semi-Nonnegative Dictionary Learning

In this section, we present a projection based algorithm for learning δ-S-NDL, and a complementary intialization algorithm. In what follows, we assume that our samples will be drawn from a (C, k, ρ)-favorable oracle $\mathcal{O}$.

Before continuing, we note that the S-NDL learning problem is sign-sensitive, and so an $S-NDL$ algorithm will need to learn the signs of A_i^* correctly. Our approach will be first to learn an estimate of A^* which is δ-close to A_i^*, and then to use the algorithms in section 5.2.4 to get an estimate which is A^*-signed-δ-close. To simplify the exposition, the projection algorithms in section 5.2.3 and initialization scheme in section 5.2.5 will establish δ-closeness results (rather than signed closeness), and with therefore assume that the columns of A_i^* are permuted and sign-flipped in such away that $\|A_i - A_i^*\| \leq \delta$. Section 5.2.2 and 5.2.4 will be more precise in stating results in terms of the signs of $A_i^T A^*{}_i$.

We are nwo ready to state the main convergence result for algorithm 8

Theorem 5.2.1 (Convergence for Algorithm 8)**.** *Suppose that* A *is* δ*-near to* A^**, where* $\delta = \mathrm{O}^*(1/\rho\lambda_i)$*, and that* A^* *is* $\mathrm{o}\left(\rho^{-1}\lambda_i^{-2}\sqrt{n}\right)$*-incoherent, where* λ_i *is the parameter defined in Equation 2.30. Then, Algorithm 8 with sample parameter* $p = \tilde{\Omega}\left(k^2\right)$ *and step-size* $\eta_i = \Theta^*(1)$ *satisfies*

$$\|A_i^s - A_i^*\|^2 \leq (1-\nu)^s \|A_i^0 - A^*\|_i^2 + n^{-\omega(1)} \tag{2.23}$$

for some $\nu \in (0, 1/2)$*, and for any* $s = 1, 2, \ldots, T$*, using* $\tilde{\Omega}(mk)$ *total samples at each step. In particular,* A^* *converges geometrically to an arbitrary inverse polynomial error. Furthermore, after the sign correction step,* $A^{(T)}$ *is* $\delta^{(T)}$*-signed-close to* A^**, where* $\delta^{(T)} \le (1-\nu)^{T/2}\|A_i^0 - A^*\|_i$*. For arbitrary* (C, k, ρ)*-favorable distributions, it suffices that* $\delta = \mathrm{O}^*(1/\rho k)$ *and* A^* *is* $\mathrm{o}\left(\rho^{-1}k^{-2}\sqrt{n}\right)$*-incoherent. For* uniformly (C, k, ρ)*-favorable distributions, it suffices that* $\delta = \mathrm{O}^*\left(1/\rho\sqrt{k}\right)$ *and* A^* *is* $\mathrm{o}\left(\rho^{-1}k^{-1}n\right)$*.*

We defer of the proof of the the theorem to section B.4. The specification in the second paragraph of the theorem comes from the following lemma, which follows from straightforward computations in the same vein as the example given in section 5.2.1:

Lemma 5.2.2 (Bounds on λ)**.** *For any* (C, k)*-favorable distribution, then* $\lambda_i = O(k)$*. If the distribution is* (C, k)*-uniformly favorable, then* $\lambda_i = O(\sqrt{k})$*.*

We remark that, for uniformly (C, k)-favorable distributions, the radius of convergence and incoherence requirements for A^* are essentially the same as those in Theorem 4.1.1, up to an extra factor of ρ^{-1}.

The main technical obstacle is that the sparse coefficients x are nonnegative, and $\mathbb{E}[x_i x_j] > 0$. As explained in Section 5.2.1, even taking differences between samples will yield a non-trivial covariance structure. As a compromise, use samples consists of the *differences* between two samples which share a common entry in their support. This observation lies at the heart of the S-NDL Projection Descent Algorithm

Algorithm 8: SNDL Projection Descent Algorithm (M)

Input: Initial estimate A_0, step size η, Number of Iterations T, Lower Bound C, Sample Sizes p, (C, k, ρ)-favorable oracle $\mathcal{O}$

for $s = 1, 2, \ldots, T$ **do**

Get Samples:

Initialize array of samples $\mathcal{Y}_i$, support estimates $\mathcal{S}_i$, and sample counts $p_j = 1$, for $i \in [m]$.

while $p_j \le 2p$ *for some* $i \in [m]$ **do**

Sample $y \sim \mathcal{O}$ and set $\hat{S} \leftarrow \{i \in [m] : |A_i^T y^{(1)}| > \tau\}$ **for** $i \in \hat{S}$ *for which* $p_j < 2p$ **do**

$\mathcal{Y}_i[p_j] \leftarrow y$ and $\mathcal{S}_i[p_j] \leftarrow \hat{S}$, and $p_j \leftarrow p_j + 1$

Update: $A_i^{s+1} = A_i^s - \eta \hat{g}_i^s$ where

$$\hat{g}_i^s = \frac{1}{p}\sum_{j=1}^{p} M_i^{\mathrm{prj}}(S^{(j,i)}, A) z^{(j,i)} \operatorname{sign}((A_i^s)^T z^{(j,i)}) \tag{2.24}$$

and $z^{(j,i)} = \mathcal{Y}_i[2j-1] - \mathcal{Y}_i[2j]$, $S^{(j,i)} = \mathcal{S}_i[2j-1] - \mathcal{S}_i[2j]$

return CorrectSigns$(A_i^{(T)}, \mathcal{O}, C)$ or CorrectSigns2$(A_i^{(T)}, \mathcal{O}, C)$

We also complement Algorithm 8 with a sketch of an initialization algorithm, described in section 5.2.5 and whose analysis is given by Theorem 5.2.13. The algorithm is a sim-

ple modification of the ideas in Arora et al. (2014). Combining this initialization with Theorem 5.2.1 gives the following theorem:

Theorem 5.2.3 (Polynomial Time Algorithm for S-NDL under Suitable Conditions)**.** *Given samples $y = A^*x$, where x come from a (C, k, ρ)-favorable for which $\lambda_i' = \mathrm{O}(k/m)$ distribution, $k = \mathrm{O}^*\left(m^{2/5}\right)$, and $A^* \in R^{n\times n}$ is $\kappa\sqrt{n}$-incoherent, where*

$$\kappa = \mathrm{o}\left(\rho^{-1}\lambda_i^{-2}\right) \quad \text{and} \quad \kappa = \mathrm{O}^*\left(k^2 + k\log^2 n\right) \tag{2.25}$$

Then, there is a two-stage polynomial time algorithm, using only $\tilde{\Omega}\left(m^2/k^2\right)$ samples up-front, and then $\tilde{\Omega}(mk)$ samples at each successive iterations, which converges at a geometric rate to A^* up to arbitrarily small inverse polynomial error. In particular, such an algorithm exists in the setting where the samples x are uniformly-(C, k, ρ)-favorable, *$k = \mathrm{O}^*\left(m^{2/5}\right)$, and $A^* \in R^{m/n}$ is $\kappa\sqrt{n}$-incoherent, where κ-satisfies Equation 2.25.*

Proof. It is easy to verify that, under the conditions of this theorem, Assumption 6 and $\delta = \mathrm{O}^*(1/\lambda_i)$ since $\lambda_i = O(k)$. Hence, the result follows from combining Theorem 5.2.1 and Theorem 5.2.13. When □

5.2.1 Challenges for Non-Negative Data

The algorithms presented in Chapters 1-4 required that the entries of the latent samples x are symmetrically distributed about zero, k-sparse and independent conditioned on their support. On the other hand, S-NDL presents us with samples $y = A^*x$ which are entrywise nonnegative. If we use samples samples $z := y^{(1)} - y^{(2)}$, then we can write $z = A^*w$, where $w := x^{(1)} - x^{(2)}$ is no more $2k$ sparse and symmetrically distributed. However, the columns of w are correlated. Indeed, suppose that the latent vectors x_i are binary, and lets compute $\mathbb{E}[w_iw_j$ for $i \neq j \in [m]$. We have

$$\begin{aligned}
\mathbb{E}[w_iw_j] &= \mathbb{E}[x_i^{(1)}x_j^{(1)} + x_i^{(2)}x_j^{(2)} - (x_i^{(1)}x_j^{(2)} + x_i^{(2)}x_j^{(1)})] \\
&= \mathbb{E}[x_i^{(1)}x_j^{(1)}] - \mathbb{E}[x_i^{(1)}x_j^{(1)}|\mathrm{supp}(w) = S] \\
&= \Pr(i, j \in \mathrm{supp}(x^{(1)})) - \Pr(i \in \mathrm{supp}(x^{(1)}), j \in \mathrm{supp}(x^{(2)})) = q_{i,j} - q_iq_j
\end{aligned}$$

where $q_{i,j}$ and q_i and q_j are as defined in the statement of Assumption 2. In general, we do not assume that $q_{i,j} - q_iq_j$, and even if the supports are drawn uniformly over supports of size k, we have $|q_{i,j} - q_iq_j| = |\frac{k(k-1)}{m(m-1)} - \frac{k^2}{m^2}| \leq \frac{-k}{m^2} \neq 0$. Through similar computations, one can also verify that w_i and w_j are still correlated, even after conditioning on the support of w!

One way to avoid correlations would be to only use samples $z = y^{(1)} - y^{(2)}$ for which the coefficient vectors $x^{(1)}$ and $x^{(2)}$ share the exact same support. Unfortunately, a sample with a given support S may occur exponentially infrequently: in the uniform case, such samples occur with probability $\binom{m}{k} \approx m^k$. As a compromise, we consider looking at samples $z[i] = y^{(1)} - y^{(2)}$ for which $x^{(1)}$ and $x^{(2)}$ both share the entry i.

To recap, our notation is as follows: $z := y^{(1)} - y^{(2)} = A^*w$, where $w := x^{(1)} - x^{(2)}$, $\mathrm{supp}(w) = \mathrm{supp}(x_1) \cup \mathrm{supp}(x_2)$, and $z[i]$ and $w[i]$ have the distribution of z and w conditioned on the event that $i \in \mathrm{supp}(x_1) \cap \mathrm{supp}(x_2)$. If Assumption 3 holds, then we immediately verify that:

Claim 5.2.4. *$w[i]_i$ is independent of $w[i]_j$ for all $j \neq i$, even after conditioning on* $\mathrm{supp}(w[i])$.

5.2.2 Sign Thresholding

Before analyzing the projection algorithm, we establish that sign thresholding is effective with high probability on the nonnegative samples. The key technical result is that

Lemma 5.2.5. *Suppose that A^* is $\kappa\sqrt{n}$ incoherent, then with high probability it holds that*

$$\begin{aligned} A^{*T}_i y &= \mathrm{sign}(A_i^T A_i^*)\mathbb{1}(i \in \mathrm{supp}(x) \pm O(k\kappa) \\ &= \mathrm{sign}(A_i^T A_i^*)C\mathbb{1}(I \in \mathrm{supp}(x)) \pm O((k + \log n)\kappa) \end{aligned} \tag{2.26}$$

Moreover, for any $v \in \mathbb{R}^n$, it holds with high probability that

$$v^T y \leq O(\|v\|(\sqrt{k} + \log n)) \tag{2.27}$$

which we prove in Section B.5 in the appendix. As a consequence, we have

Lemma 5.2.6. *If A^* is $O^*(\sqrt{n}/(k + \sqrt{k}\log n))$-incoherent and A is $O^*(1/\sqrt{k} + \log n/k)$-close to A, then with high probability, $|A_i^T y| \geq C/2$ precisely when $i \in S$. Furthermore, when we distinguish between the signs of A_i and $A^*{}_i$, then $\mathrm{sign}_{C/2}(A_i^T y) = \mathrm{sign}(A_i^T A_i^*)$.*

Proof of Lemma 5.2.5. Write $A_i^T y = A^{*T}_i y - (A^*{}_i - A_i)^T y$, and apply the following Claim noting that $\|A^*{}_i - A_i\| = O^*(1/\sqrt{k} + \log n/k)$ □

As a corollary, we have

Corollary. *Given samples $z = y^{(1)} - y^{(2)}$, we can estimate the support $\hat{S}$ of $w = x^{(1)} - x^{(2)}$ with high probability by setting*

$$\hat{S} = \{i \in [m] : |A_i^T y^{(1)}| > C/2 \text{ or } |A_i^T y^{(1)}| > C/2\} \tag{2.28}$$

5.2.3 A Projection Algorithm for S-NDL

To ease notation, we drop the dependence on the iteration. Following section 5.2.1, let $z = y^{(1)} - y^{(2)}$ and $y^{(r)} = A^* x^{(r)}$ for $r = 1, 2$, and let $z[i] \sim z \big| i \in S := \mathrm{supp}(x^{(1)} \cap \mathrm{supp}(x^{(2)})$. The main theorem of this subsection is that the expectation the gradients under correctly support recovery given by

$$G_i := \mathbb{E}_i[M_i^{\mathrm{prj}}(S) z[i] \mathrm{sign}(A_i^T z_i)] \tag{2.29}$$

are are $(\Theta(1), 0, \delta)$-true after rescaling, as are the gradients $\hat{g}_s^i$ from Algorithm 8:

Proposition 5.2.7. *Suppose that A is δ-near to A^*, where $\delta = \mathrm{O}^*(1/\rho\lambda_i)$, and that A^* is $\kappa\sqrt{n}$-incoherent, where $\kappa\rho\lambda_i^2 = o(1)$. If G_i is defined as in Equation 2.29, then $q_i G_i$ is $(\Theta(1), \delta, n^{-\omega(1)})$-true. Furthermore, if Algorithm 8 uses $p = n^{O(1)}$, then $q_i g_i := q_i \mathbb{E}[\hat{g}_i^s]$ is $(\Theta(1), \delta, n^{-\omega(1)})$-true as well.*

In the interest of brevity, we do not state an infinite sample version of theorem 5.2.1, and defer the proof of the finite sample result from the beginning of Section 5.2 to the appendix. Following the proof strategy for the projection rule for mean-zero distributions, we differentiate between a signal component $G_{1,i} := \mathbb{E}[M_i^{\text{prj}} A_i^* x_i \text{sign}_\tau(A_i^T y)]$ and the noise $G_{2,i} := \mathbb{E}[M_i^{\text{prj}} y_{-i} \text{sign}_\tau(A_i^T y)]$. First, we need to quantify the degree of support independence by the following parameter:

$$\lambda_i := \sqrt{\mathbb{E}_{S:i\in S} \|\mathbb{E}_{U(i)}[w_U w_U^T\|_{l_1}} \tag{2.30}$$

where, for a set $S \subset [m]$, let $U(i)$ denote the set $U := S - \{i\}$, and let $\mathbb{E}_{U(i)}[\cdot] := \mathbb{E}[\cdot \big| S = U \cup \{i\}]$. We remark that $\lambda_i = \Omega(\sqrt{k})$. We also use the notation $z_{-i} = A^*{}_U w_U$, similar to the convention for y_{-i} in the previous chapters, and $\mathbb{E}_i[\cdot\big|]$ to denote that $i \in \text{supp}(x_1) \cap \text{supp}(x_2)$. The following controls $\mathbb{E}[|A_i^T z_{-i}|]$ in terms of λ

Lemma 5.2.8. *If A^* is $\kappa\sqrt{n}$incoherent and A is δ-close to A^*, then*

$$\mathbb{E}_i[|A_i^T z_{-i}|] \le \sqrt{\mathbb{E}[|A_i^T A^*{}_U w_U|^2]} \lesssim \|A_i - A^*\| + \kappa\lambda \tag{2.31}$$

Proof. By Jensens inequality and the fact that $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\begin{aligned} \mathbb{E}_i[|A_i^T A^*{}_U w_U|] &\le \sqrt{\mathbb{E}[|A_i^T A^*{}_U w_U|^2]} \\ &\le 2\mathbb{E}_i[|(A_i - A^*)^T A^*{}_U w_U|^2] + 2\mathbb{E}_i[|A^{*T} A^*{}_U w_U|^2] \end{aligned}$$

Next,

$$\begin{aligned} \mathbb{E}_i[|(A_i - A^*)^T A^*{}_U w_U|^2] &= \mathbb{E}_i[(A_i - A^*)^T A^*{}_U w_U w_U^T A^*{}_U (A_i - A^*)] \\ &\le \|A_i - A^*{}_i\|^2 \|A^*{}_i\|^2 \|\mathbb{E}_i[w_U w_U^T]\| \\ &\lesssim \|A_i - A^*{}_i\|^2 O \end{aligned}$$

by noting that $\|A^*\| = O(1)$ in the under complete setting, and that the the diagonals of $\|w_U w_U^T\|$ are no more than $O(k/m)$, and the off diagonals no more than $O(k^2/m^2)$, so that $\|\mathbb{E}_i[w_U w_U^T]\| = O(k^2/m) = O(1)$. To control $\mathbb{E}_i \|M_i^{\text{prj}}(U) A^*{}_U w_U\|^2$, we can write

$$\begin{aligned} \mathbb{E}_i \|A_i^T A^*{}_U w_U\|^2 &= \mathbb{E}_{S:i\in S} \mathbb{E} \|A_i^T A^*{}_U w_U \big| U = S - \{i\}\|^2 \\ &= \mathbb{E}_{S:i\in S} \mathbb{E}[\|A_i^T A^*{}_U w_U \big| U = S - \{i\}\|^2] \\ &= \mathbb{E}_{S:i\in S} \mathbb{E}[\|\sum_i A_i^T A^*{}_r w_U \big| U = S - \{i\}\|^2] \\ &\le \tau^2 \mathbb{E}_{S:i\in S} \|\mathbb{E}[w_U w_U^T \big| U = S - \{i\}]\|_{l_1} \\ &= \tau^2 \lambda_i^2 \end{aligned}$$

□

The first major consequence of the above lemma is that

Lemma 5.2.9. *If A^* is $\kappa\sqrt{n}$incoherent and A is δ-close to A^*, and $\rho(\delta + \kappa\lambda) = \mathrm{O}^*(1)$. Then $a_i := \mathbb{E}[w_i \mathrm{sign}(A_i^T y)]$ is $\Theta(1)$.*

Proof. One the one hand, we have $\mathbb{E}[w_i \mathrm{sign}_\tau(A_i^T y)] \le \mathbb{E}[|w_i|]$. For the other direction, note that $w_i = x_i^{(1)} - x_i^{(2)}$ is ρ-smooth, and sine ${A^*}_i^T A_i = 1 - O(1)$, $A_i^T {A^*}_i w_i = (1 + o(1))\rho$ smooth

$$\begin{aligned}
\mathbb{E}[w_i \mathrm{sign}_\tau(A_i^T y)] &= \mathbb{E}[|w_i|] - \mathbb{E}[(\mathrm{sign}(A_i^T {A^*}_i w_i) - \mathrm{sign}(A_i^z y))w_i] \\
&\ge \mathbb{E}[|w_i|] - \mathbb{E}[|\mathrm{sign}(A_i^T {A^*}_i w_i) - \mathrm{sign}(A_i^T z)| \cdot |w_i|] \\
&\ge \mathbb{E}[|w_i|] - \rho(1 + o(1))\mathbb{E}[|A_i^T z_{-i}| \cdot |w_i|] \\
&\ge \mathbb{E}[|w_i|] - \rho(1 + o(1))\mathbb{E}[w_i^2] \cdot \mathbb{E}[|A_i^T z_{-i}|^2] \\
&\ge \mathbb{E}[|w_i|] - \mathrm{O}(\rho\delta + \rho\kappa\lambda)
\end{aligned}$$

The lemma now follows from the stated assumptions. □

Hence, we have the following analogue of Proposition 4.1.3

Corollary. *If $\delta = \mathrm{O}^*\left(1/\sqrt{k}\right)$, then*

$$\mathbb{E}_i[M_i^{\mathrm{prj}} {A^*}_i w_i \mathrm{sign}(A_i^T z)] = a_i(A^*{}_i - A_i) + o(\|A^*{}_i - A_i\|) + \gamma \tag{2.32}$$

where again $a_i := \Theta(1)$.

Note that we do not threshold in the sign-estimate step, since thresholding was already taken care of in the sample collection step. We can also use Lemma 5.2.8 to control the "noise term":

Proposition 5.2.10.

$$\|\mathbb{E}_i[M_i^{\mathrm{prj}}(S) {A^*}_U w_U]\| \lesssim \|A_i - A^*\|\rho\delta\lambda + \rho\delta\kappa\lambda^2 \tag{2.33}$$

Proof. By Cauchy Schwartz, we have

$$\begin{aligned}
\|\mathbb{E}_i[M_i^{\mathrm{prj}}(S) {A^*}_U w_U]\| &\le \mathbb{E}_i[\|M_i^{\mathrm{prj}}(U) {A^*}_U w_U\| \left|\mathbb{1}(\mathrm{sign}(A_i^T {A^*}_i w_i) - \mathrm{sign}(A_i^T y)\right|] \\
&\lesssim \rho\mathbb{E}_i[\|M_i^{\mathrm{prj}}(U) {A^*}_U w_U\| |A_i^T {A^*}_U w_U|] \\
&\le \rho\sqrt{\mathbb{E}_i[\|M_i^{\mathrm{prj}}(U) {A^*}_U w_U\|^2]\mathbb{E}[|A_i^T {A^*}_U w_U|^2]}
\end{aligned}$$

To control $\mathbb{E}_i\|M_i^{\mathrm{prj}}(U) {A^*}_U w_U\|^2$, we can write

$$\mathbb{E}_i\|M_i^{\mathrm{prj}}(U) {A^*}_U w_U\|^2 = \mathbb{E}_{S: i \in S}\mathbb{E}\|M_i^{\mathrm{prj}}(U) {A^*}_U w_U \big| U = S - \{i\}\|^2$$

Conditioned on U, the vectors $M_i^{\mathrm{prj}}(U) {A^*}_j$ for $j \in U$ are deterministic and bounded in norm by δ. Hence, Claim D.2.1 yields:

$$\mathbb{E}_i[\|M_i^{\mathrm{prj}}(U) {A^*}_U w_U \big| U\|^2] \le \delta\|w_U w_U^T \mathbb{E}[\big| U]\|_{l_1} = \delta\lambda \tag{2.34}$$

Consequently,

$$\|\mathbb{E}_i[M_i^{\mathrm{prj}}(S) {A^*}_U w_U]\| \le \|A_i - A^*\|\rho\delta\lambda + \rho\kappa\lambda^2 \tag{2.35}$$

□

Putting together Proposition 5.2.3 and Proposition 5.2.10 helps us demonstrate that $\mathbb{E}[\hat{g}_s^i]$ is $(\Theta(1), \delta, n^{-\omega(1)}$-true after a rescaling:

Proof of Proposition 5.2.7. Let $\tilde{z}^{(j)}$ be the distribution of the $j-$th sample selected for the update rule $\hat{g}_s^i$ in Algorithm 2.2.1, and let $E^{(j)}$ be the event that the the samples $z^{(j)}$ were just taken from the first $2j$ samples $y = A^*x$ for which $i \in \mathrm{supp}(x)$. Let $z[i] \sim z\big|i \in \mathrm{supp}(x^{(1)}) \cap \mathrm{supp}(x^{(2)})$. Since $E^{(j)}$ occurs with high probability, and $\tilde{z}^{(j)}\mathbb{1}(E^{(j)})$ and $z[i]\mathbb{1}(E^{(j)})$ have the same distribution. Hence,

$$\begin{aligned}
\mathbb{E}[\hat{g}_i] &= \mathbb{E}[M_i^{\mathrm{prj}}\tilde{z}\mathrm{sign}(A_i^T\tilde{z})] \\
&= \mathbb{E}[M_i^{\mathrm{prj}}\tilde{z}\mathrm{sign}(A_i^T\tilde{z})\mathbb{1}(E^{(j)}] + \gamma \\
&= \mathbb{E}[M_i^{\mathrm{prj}}z[i]\mathrm{sign}(A_i^Tz[i])\mathbb{1}(E^{(j)}] + \gamma \\
&= \mathbb{E}[M_i^{\mathrm{prj}}z[i]\mathrm{sign}(A_i^Tz[i])] + \gamma \\
&= \mathbb{E}_i[M_i^{\mathrm{prj}}z\mathrm{sign}(A_i^Tz)] + \gamma \\
&= G_i + \gamma
\end{aligned} \tag{2.36}$$

Noting that $\lambda = \Omega(\sqrt{k})$, and that $\rho = \Omega(1)$, the bound on G_i follows from Proposition 5.2.3 and Proposition 5.2.10 □

5.2.4 Correcting the Signs From the Projection Algorithm

Note that coordinate descent only learns A which whose columns are within δ of the columns of A^*, up to permutations *and sign flips*. However, the reduction in Algorithm 4 requires that A be δ-signed-close to A. However, this is not hard to do, given an estimate of A with is $\delta = \mathrm{o}\left(1/k + 1/\sqrt{k}\log n\right)$ close to A^*:

Proposition 5.2.11. *Suppose that A^* is $O^*(\sqrt{n}/(k+\sqrt{k}\log n))$-incoherent. Then, given an estimate an estimate A^* of A with is δ-close to A^* for $\delta = \mathrm{o}\left(1/k + 1/\sqrt{k}\log n\right)$, Algorithm 9 returns a dictionary which is δ-signed-close to A^* using at most $\tilde{\Omega}(m/k)$ samples.*

Proof. Let Σ be the diagonal matrix of signs such that $A\Sigma$ is δ-signed-close to A^*. By the sign thresholding analysis in Lemma 5.2.6 it holds with $\mathrm{sign}_\tau(A_i^Ty) = \mathrm{sign}(A_i^TA^*) = \Sigma_{ii}$ with high probability. Hence, each sign flip corrects the sign of the the i-th columns of A. To view a sample from every column, we need only $\tilde{\Omega}(m/k)$ samples. □

In the case where A^* is not as incoherent, we can use a slightly more complictated algorithm, which calls Algorithm 6 as a subroutine: Algorithm 10 matches the performance to Algorithm 9, and we prove the following proposition by adapting many of the same ideas:

Proposition 5.2.12. *Given an estimate an estimate A^* of A shi h is δ-close to A^* for $\delta = \mathrm{o}\left(1/k + 1/\sqrt{k}\log n\right)$, Algorithm 10 returns a dictionary which is δ-signed-close to A^* using at most $\tilde{\Omega}(m/k)$ samples.*

Algorithm 9: CorrectSigns$(A, \mathcal{O}, C)$

Data: Dictionary A which is δ-close to A^*
Data: Dictionary A which is δ-signed-close to
Initialize $\mathcal{S} = [m]$ **while** $\mathcal{S}$ *is not empty* **do**
  Query sample y
  for $i = 1, \dots, m$ **do**
    If $\text{sign}_i := \text{sign}_{C/2}(A_i^T y) > 0$, flip $A_i \leftarrow \text{sign}_i A_i$ and remove i from $\mathcal{S}$.
return A

Algorithm 10: CorrectSigns2$(A, \mathcal{O}, C)$

Data: Dictionary A which is δ-close to A^*, (C, k)-favorable sample oracle $\mathcal{O}$ for A^*
Data: Dictionary A which is δ-signed-close to
Initialize $\mathcal{S} = [m]$ **while** $\mathcal{S}$ *is not empty* **do**
  $(y_S, \hat{x}_S) = \text{Decode}(A, \mathcal{O}, C, 1)$ **for** $i = 1, \dots, m$ **do**
    If $\hat{x}_i \neq 0$, flip $A_i \leftarrow \text{sign}(x_i) A_i$ and remove i from $\mathcal{S}$.
return A

5.2.5 Sketch of an Initialization Algorithm

In the interest of brevity, we only sketch the initialization algorithm; the ideas are not new, and can be adapted mostly from the OverlappingCluster and OverlappingAverage algorithms in Arora et al. (2014) under the following assumptions:

Assumption 6. *Suppose that $k \leq \mathrm{O}^*\left(m^{2/5}\right)$, that is A^* is μ-incoherent, with $\mu/\sqrt{n} \leq O^*(1/(k^2 + k\log^2 n)$, and the sparse coefficient vectors are (C, k)-favorable, and that $\lambda_i' := \sigma_{\max}(\mathbb{E}[w_{-i} w_{-i}^T | i \in S]) = O(k/m)$. We remark that $\lambda_i' = O(k/m)$ when x come from a uniformly-(C, k, ρ)-favorable distribution.*

Under these conditions, we have

Theorem 5.2.13. *Under Asssumption 6, there is a polynomial is a time algorithm, which, when given $p = \tilde{\Omega}\left(m^2 \log^2 m/k^2\right)$ samples, returns an estimate of A which is $O(k/m)$ close to A^*.*

We devote the rest of the section to a sketch of the above theorem's proof. The only modification to the techniques in Arora et al. (2014) are that the sparse coefficients are nonnegative, and we need to learn A_i^* with the correct sign. First, we need to show that we can detect overlapping entries with high probability

Lemma 5.2.14 (Detecting Common Entries)**.** *If A^* is μ-incoherent, with $\mu/\sqrt{n} \leq O^*(1/(k^2 + k\log^2 n)$, then given two samples $y^{(1)} = A^* x^{(1)}, y^{(2)} = A^* x^{(2)}$, it holds with high probability that $\langle y^{(1)}, y^{(2)} \rangle \geq C^2/2$ precisely when $\text{supp}(x^{(1)}) \cap \text{supp}(x^{(2)}) \neq \emptyset$*

Proof. The proof is immediate from summing up the first display in Lemma 5.2.5. □

With these thresholding bounds, we apply OverlappingCluster algorithm in Arora et al. (2014), which has the following guarantees

Lemma 5.2.15. *If $k = \mathrm{O}^*\left(n^{2/5}\right)$, then given $p = \tilde{\Omega}\left(m^2 \log^2 m/k^2\right)$ samples $y^{(j)} = A^* x^{(j)}$, there is a polynomial time algorithm which can recover the sets $S_1, \ldots, S_m$, where $S_i = \{j : i \in \mathrm{supp}(x^{(j)})\}$ with high probability*

Finally, we have an SVD initialization based on the the "Overlapping Average" algorithm in Arora et al. (2014):

Lemma 5.2.16 (SVD Initialization)**.** *Let $\Lambda(i)$ be the matrix $\mathbb{E}[w_{-i}w_{-i}^T \big| i \in S]$. Suppose that $\|Q(i)\| = O(k/m)$, and for all $\{i,j,r\} \in \binom{[m]}{3}$, $\mathbb{E}_{i,r}[x_r]\mathbb{E}_{i,j}[x_j] = \mathbb{E}_{i,j,r}[x_j x_r]$. Let $z^{(j)}[i]$ have the distribution of $z^{(j)} \big| i \in S$. Then the top singular vector of*

$$M_i := \frac{1}{p}\sum_{j}^{n} (z[i]^{(j)})(z[i]^{(j)})^T \tag{2.37}$$

satisfies $\|v - \pm A^*{}_i\| = O(k/m)$ as long as $p = \tilde{\Omega}\left(m/k\right)$.

Sketch. For ease of notation, let $z^{(j)}$ have the distribution of $z^{(j)}[i]$ and $w^{(j)}$ the distribution of $w^{(j)} \big| i \in S$. Then

$$\begin{aligned} M_i := \frac{1}{p}\sum_{i=1}^{p} z^{(j)})((j))^T &= \frac{1}{p}\sum_{i=1}^{p} A^* w w^T A^{*T} \\ &= A_i A_i^T \frac{1}{p}\sum_{j} w_i^2 + A^*_{-i}\mathbb{E}[w_{-i}w_{-i}^T]A^*_{-i} \\ &+ A^*_{-i}(\frac{1}{n}\sum_{i=1}^{p} w_{-i}w_{-i}^T - \mathbb{E}[w_{-i}w_{-i}^T])(A^*_{-i})^T \end{aligned}$$

By standard subgaussian arguments, we see that $A_i A_i^T \frac{1}{p}\sum_j w_i^2 = \Omega(A_i A_i^T)$. Furthermore, by assumption, $A^*\mathbb{E}[w_{-i}w_{-i}^T]A^* = O(\|A^*\|^2 k/m) = \mathrm{O}\left(k/m\right)$. Using a bound like the one in C.1.1, with $p = \tilde{\Omega}(m^2/k^2)$, the random error is controlled by $\mathrm{o}\left(k/m\right)$. Finally, by Wedins Theorem (see Stewart (1998)), it holds that the top singular vector of M_i is distance $O(k/m)$ from A_i^*. The proof for v_i is similar, using Bernstein's inequality. □

Proof of Theorem 5.2.13. To prove the theorem, take $p = \tilde{\Omega}\left(m^2/k^2\right)$ samples, detecting pairwise support overlaps using the strategy in Lemma 5.2.14, and feed them to OverlappingCluster based algorithm. Next, we run the SVD initialization to get estimates of A_i of $A^*{}_i$ which are within $O(k/m)$, up to a choice of signs. Standard high probability arguments show that if $z[i]$ satisfy the concentration in 5.2.16, then so will the clustered samples. □

Conclusion

In the present work, we applied a very simple anti-concentration assumption to substantially generalize the scope of models in which efficient gradient descent algorithms can be shown to learn incoherent dictionaries. By utilizing more sophisticated forms of anti-concentration, we wonder if we can answer even more difficult questions that arise in the context of sparse coding. For example, we would like to investigate if anti-concentration might be the correct lens through which to help explain the mystery that sparse coding algorithms seem to converge when initializated with random samples.

More generally, recent work Hazan et al. (2015) has shown that certain non-convex functions can be optimized by effectively "smoothing out" the objective using techniques that resemble zeroth-order convex optimization. Though optimisitic, we hope that the tools from anti-concentration might one day demonstrate similar "smoothing-out" phenomena when learning broader classes of non-convex objectives under suitably anti-concentrated generative models.

Bibliography

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013a.

Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. A clustering approach to learn sparsely-used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2013b.

Michal Aharon, Michael Elad, and Alfred Bruckstein. ¡ img src=”/images/tex/484. gif” alt=”\ rm k”¿-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. *arXiv preprint arXiv:1212.4777*, 2012a.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. pages 145–162, 2012b.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 779–806, 2014. URL http://jmlr.org/proceedings/papers/v35/arora14.html.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *CoRR*, abs/1503.00778, 2015. URL http://arxiv.org/abs/1503.00778.

Trapit Bansal, Chiranjib Bhattacharyya, and Ravindran Kannan. A provable svd-based algorithm for learning topics in dominant admixture corpus. pages 1997–2005, 2014.

Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. *arXiv preprint arXiv:1407.1543*, 2014.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. pages 594–603, 2014.

Guy Bresler. Efficiently learning ising models on high degree graphs. *arXiv preprint arXiv:1411.6156*, 2014.

Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

Emmanuel Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52 (12):5406–5425, 2006.

Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. In *Independent Component Analysis and Signal Separation*, pages 169–176. Springer, 2007.

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, page None, 2003.

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Elad Hazan, Kfir Y Levy, and Shai Shalev-Swartz. On graduated optimization for stochastic non-convex problems. *arXiv preprint arXiv:1503.03712*, 2015.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

Kyle Luh and Van Vu. Dictionary learning with few samples and matrix concentration. *arXiv preprint arXiv:1503.08854*, 2015.

Shahar Mendelson. Learning without concentration. *arXiv preprint arXiv:1401.0304*, 2014.

Hoi H Nguyen and Van H Vu. Small ball probability, inverse theorems, and applications. pages 409–463, 2013.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. pages 1214–1222, 2012.

Philippe Rigollet. Lecture notes high dimensional statistics. 2014.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration.

Daniel Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. pages 296–305, 2001.

Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. pages 3087–3090, 2013.

Gilbert W Stewart. Perturbation theory for the singular value decomposition. 1998.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Ramon van Handel. Probability in high dimension. *Lecture Notes: ORF 570, Princeton University*, 2014.

Arnaud Vandaele, Nicolas Gillis, François Glineur, and Daniel Tuyttens. Heuristics for exact nonnegative matrix factorization. *arXiv preprint arXiv:1411.7245*, 2014.

Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin and Mark Rudelson. Anti-concentration inequalities. *Presentation at Phenomena in High Dimension, 3rd Annual Conference, Samos, Greece.*, 2007.

Appendix A

Support Results for Gradient Descent

A.1 Update Rule Computations

A.1.1 Meta-Algorithm Generalizes Neural Rule

Given a sample $y = A^*x$ where x is C-lower bounded, the neural update rule estimates the support $\hat{S} = \{i \in [m] : |(A^s)_i^T y|\} \geq C/2$. The algorithm then computes an empirical average in order to approximate

$$g_i^s = \mathbb{E}[(y - A^s\hat{x})\text{sign}(\hat{x})] = \mathbb{E}[(y - A^s\hat{x})\text{sign}(\hat{x})\big|i \in S]\Pr(i \in S) \tag{1.1}$$

where $\hat{x} = \text{Thres}_{C/2}((A^s)^T y)$. Hence $\text{sign}(\hat{x}) = \text{sign}_{C/2}((A_i^s)^T y)$ and

$$A^s\hat{x} = \sum_{r\in\hat{S}} A_r^s(A_r^s)^T y = A_{\hat{S}}^s(A_{\hat{S}}^s)^T \tag{1.2}$$

Thus, we have that

$$\frac{1}{\Pr(i \in S)} g_i^s = \mathbb{E}[(I - A_{\hat{S}}^s(A_{\hat{S}}^s)^T)\text{sign}_{C/2}((A_i^s)^T y)] \tag{1.3}$$

which shows that, up to a rescaling by $\Pr(i \in S)$ (which is handled by the choice of the step size η), the Neural Update Rule is an instantiation of our Meta-Algorithm.

A.1.2 Proofs for Automated Analysis and Sign-Thresholding

Proof of Theorem 3.1.3. We prove the first part of the proposition in the more general case where g^s are $(\alpha_i^s, \delta^s + e^s, \epsilon)$-true, where again

$$e_s^2 \leq (1 - \alpha_{\min}\eta_0/4)^s\delta_0^2 + 64\zeta^2/\alpha_{\min}^2 \tag{1.4}$$

The high probability analogue and the statement with δ^s closeness instead of $(\delta^s, 2)$-nearness will follow from similar arguments.

First, we need to show that the invariant $\|A^s - A^*\| \leq 2$ is preserved at each iteration. This is immediate from an induction on the following claim, whose proof we to the end of the section

Claim A.1.1. *Suppose that g is α-nearness-well-conditioned and A is $(\delta, 2)$ to A^*. Then, if the step sizes $\eta_i > 0$ are chosen so that $0 < \max_i 1 - \eta_i \Pr(i \in S) \leq 1 - \Omega(1)$. Then if $A' = A - g\mathrm{diag}(\eta_i)$, it holds that*

$$\|A^* - A'\| \leq 2\|A^*\| \tag{1.5}$$

Now, lets consider the transformed step sizes $\tilde{\eta}_i^s = q_i \eta_i^s$ and transformed gradient $\tilde{g}_i = q_i^{-1} g_i$, we see that the update rule can be expressed as $A^{s+1} = A^s - \tilde{g}^s \mathrm{diag}(\tilde{\eta}_i^s)$. Next, if A^s is $(\delta^2, 2)$ near to A^*, then by assumption g_s^i is $(\alpha_i, \delta^s + e^s, \zeta)$-true, then we can write

$$\tilde{g}_i^s = \alpha_i(A_i^* - A_i) + v \quad \text{where} \quad \|v\| \leq \mathrm{O}^*\left(\alpha_i \|A^*{}_i - A_i\|\right) + e^s + \delta^s \tag{1.6}$$

Hence, we can establish that $\tilde{g}_i^s$ is

$$\left(\alpha_i^s/4, 1/25\alpha_i^s, 8\left(\zeta^2 + o((\delta^s + e^s)^2)\right)/\alpha_i\right) - \text{correlated with } A_i^*$$

by Lemma 2.1.4 together with the inequality $(a+b)^2 \leq 2a^2 + 2b^2$. Following the remark after Definition 2.1.2, we can substitute in the lower bounds $\alpha_{\min}/4 \leq \alpha_i^s/4$ and $1/25\alpha_{\max} \leq /25\alpha_i^s$ to establish that $\tilde{g}_i^s$ is

$$\left(\alpha_{\min}/4, 1/25\alpha_{\max}, 8\left(\zeta^2 + o((\delta^s + e^s)^2)\right)/\alpha_{\min}\right) - \text{correlated with } A_i^*$$

Define $\alpha = \alpha_{\min}/4$, $\eta_0 = \min_{i,s}$ and let

$$\epsilon_s := 8\zeta^2/\alpha_{\min} + o((\delta^s + e^s)^2/\alpha_{\min})) \text{ and } \epsilon_0 = 16\zeta^2/\alpha_{\min}$$

We now proceed by induction, where the invariant we wish to preserve is that, for all columns i,

$$\|A_i^s - A^*\|^2 \leq (\delta^s)^2 \leq (1 - \alpha\eta_0)^s \|z_0 - z_s\|^2 + \epsilon_0/\alpha = e_s^2 \tag{1.7}$$

Indeed, if this holds then since $\alpha = \alpha_{\min}/4 = \Theta(1)$, it follows that

$$\begin{aligned}
\epsilon_s &= \epsilon_0/2 + o\left(\frac{1 - \alpha\eta_0^2\tilde{\eta}_i^2)^s \|z_0 - z_s\|}{\alpha} + \frac{\epsilon_0}{\alpha}\right) \\
&= \epsilon_0/2 + o\left(\epsilon_0\right) + o\left(\left(1 - \alpha\eta_0^2\right)^s \|z_0 - z_s\|^2\right) \\
&\leq \epsilon_0 + \alpha \cdot o\left((1 - \alpha\eta_0)^s \|z_0 - z_s\|^2\right)
\end{aligned} \tag{1.8}$$

Hence, by Theorem 2.1.3,

$$\|A_i^{s+1} - A^*\|^2 \leq (1 - \alpha\eta_0)^{s+1}\| + \epsilon_0/\alpha \tag{1.9}$$

for all columns i. Since $\delta^{s+1} = \max_{i \in [m]} \|A_i^{s+1} - A^*\|$, Equation 1.9 both shows that the invariant in Equation 1.7 is maintained, and gives the desired bound on $\|A_i^{s+1} - A^*\|$. Plugging in $\epsilon_0/\alpha = 64\zeta^2/\alpha_{\min}^2$ concludes the proof. □

Proof of Claim A.1.1. If g is nearness well conditioned, then

$$\begin{aligned}\|A^* - A'\| &= \|(A^* - A) - (A^* - A)\cdot \mathrm{diag}(\alpha_i\eta_i q_i) - A\mathrm{diag}(\eta_i\beta q_i) - \tilde{A}\| \\ &\le \|(A^* - A) - (A^* - A)\cdot \mathrm{diag}(\alpha_i\eta_i q_i)\| + \|A\|\max_i(\beta_i q_i\eta_i) + \|\tilde{A}\| \\ &\le \|(A^* - A) - (A^* - A)\cdot \mathrm{diag}(\alpha_i\eta_i q_i)\| + o(\|A\| + \|A^*\|) \\ &= \|(A^* - A) - (A^* - A)\cdot \mathrm{diag}(\alpha_i\eta_i q_i)\| + o(\|A^*\|) \\ &= \|(A^* - A)\cdot \mathrm{diag}(1 - \alpha_i\eta_i q_i)\| + o(\|A^*\|)\end{aligned}$$

where we use the fact that $\|A\| \le O(\|A^*\|)$ by the $(\delta, 2)$-nearness, and that the assumptions of the Lemma require $\eta_i \mathrm{Pr}(i \in S) = O(1)$. To wrap up

$$\|(A^* - A)\cdot \mathrm{diag}(1 - \alpha_i\eta_i q_i)\| \le \|A^* - A\|\max_i(1 - \alpha_i\eta_i q_i) \le 2\|A^*\|(1 - \Omega(1))$$

Combining the two displays proves the lemma. □

Proof of Lemma 3.1.4. The result follows easily once we establish the following supporting claim:

Claim A.1.2. $A_i^T y = x_i + \mathrm{O}^*\left((\delta + \mu\sqrt{k}/n)\log n\right)$

Proof. The proof of this lemma slightly generalizes and sharpens Lemma 23 in Arora et al. (2015), and control each separately $A_i^T y = (A^*)^T y + (A^* - A_i)^T y$. For the first term, we expand

$$(A^*)^T y = (A^*)^T A_S^* x = x_i + \sum_{j\neq i\in S} A_i^{*T} A_j^* x_j \tag{1.10}$$

Since x_j are $O(1)$ subgaussian, $A_i^{*T}A_j^* x_j$ are $O(|A_i^{*T}A_j^*|^2) = O(\mu^2/n)$ - subgaussian. As $|S| \le K$ and the x_j are independent conditioned on S, it holds that $\sum_{j\neq i\in S} A_i^{*T}A_j^* x_j$ is $O(k\mu^2/n)$-Subgaussian by Proposition C.1.2. By Proposition C.1.1 and the fact that x_j are all mean zero, it follows that $\sum_{j\neq i\in S} A_i^{*T}A_j^* x_j = \mathrm{O}\left(\mu\sqrt{k}\sqrt{\log(1/\delta)}/n\right)$ with probability δ. Taking $\delta = cn^{-\log n} = n^{-\omega(1)}$ for a suitably small constance c shows that $\sum_{j\neq i\in S} A_i^{*T}A_j^* x_j = \mathrm{O}^*\left(\mu\sqrt{k}\log n/n\right)$ with probability $n^{-\omega(1)}$. For the second quantity, let $\Delta = A^* - A_i$. We have that $(A^* - A_i)^T y = \sum_{j\in S}\Delta^T A_j^* x_j$. Hence, $(A^* - A_i)^T y$ is $\sum_i(\Delta^T A_j^*)^2 = \Delta^T A_S^*\Delta \le \delta^2\|A_S^*\|$- subgaussian. By the Gershgorin circle theorem, $\|A^*S\| = O(1)$, so applying Proposition C.1.1, we see that $(A^* - A_i)^T y$ is $\mathrm{O}^*(\delta\log n)$ with high probability. □

□

A.1.3 Auxillary Claims for Decomposable Rules

In this section, we fill in the details of the proofs of Claim A.1.4 and Lemma A.1.5 used in the proof of Proposition 3.2.10. The ideas are roughly the sames as those presented in the proof of Proposition 3.2.6, making extensive use of the Anti-Concentration Corollary 3.2.7. We will assume that the sparse coefficients are x_i ρ-smooth, rather than (C, ρ)-smooth. The

generalization to the (C, ρ)-smooth case follows from a similar argument as presented in the proof of Proposition 3.2.6.Throughout this subsection, we define $Y_j := x_j \text{sign}_\tau(A_i^T y)$ and $\mathbb{1}_r := \mathbb{1}(r \in \hat{S}$. Before proving our main results, we establish a simple claim that generalizes the arguments in Proposition 3.2.6:

Claim A.1.3.

$$|\mathbb{E}_S[Y_j]| \lesssim \rho \left|A_i^T A^*{}_j\right| \tag{1.11}$$

Moreover generally, if $f(\cdot)$ is an even, measurable function, then

$$|\mathbb{E}_S[Y_j f(x_j)]| \leq \rho \left|A_i^T A^*{}_j\right| \mathbb{E}[|x_j f(x_j)|] \tag{1.12}$$

Proof. Let's prove the more general point first. We begin by decomposing:

$$\begin{aligned} \mathbb{E}_S[Y_j f(x_j)] &= \mathbb{E}_S[x_j f(x_j) \text{sign}_\tau(A_i^T y_{-j})] && (1.13)\\ &= \mathbb{E}_S[x_j f(x_j)[\text{sign}_\tau(A_i^T y) - \text{sign}_\tau(A_i^T y_{-j})] + \mathbb{E}_S[x_j f(x_j) \mathbb{1}(|A_i^T y_{-j}| > \tau)] && (1.14) \end{aligned}$$

The second term in the last line is 0, since $\mathbb{E}_S[x_j f(x_j)] = 0$ as f is even, and $x_j f(x_j)$ is independent of $A_i^T y_{-j}$. To control the first term, we apply Corrolary 3.2.7 with random variables $Z_r = x_r$ and weights $a_r = A_i^T A^*{}_r$ to get

$$\begin{aligned} |\mathbb{E}_S[x_j f(x_j)[\text{sign}_\tau(A_i^T y) - \text{sign}_\tau(A_i^T y_{-j})]| &\leq \mathbb{E}_S[|x_j f(x_j)| \cdot \left|\text{sign}_\tau(A_i^T y) - \text{sign}_\tau(A_i^T y_{-j})\right|] \\ &\lesssim \frac{\rho |A_i^T A^*{}_j|}{|A_i^T A^*{}_i|} \mathbb{E}_S[|x_j^2 f(x_j)|] \\ &\lesssim \rho |A_i^T A^*{}_j| \mathbb{E}_S[|x_j^2 f(x_j)|] \end{aligned}$$

since $A_i^T A^*{}_i = 1 - o(1)$. □

Claim A.1.4. *If $S \supset \{i, j, r\}$, then*

$$\mathbb{E}_S[Y_j \mathbb{1}_r^c] \lesssim \rho \left(|A_r^T A^*{}_j| + |A_i^T A^*{}_j|\right) \lesssim \rho(\delta + \mu/\sqrt{n}) \tag{1.15}$$

Proof. First, we break up $\mathbb{E}_S[Y_j \mathbb{1}_r^c]$ into easier-to-handle components

$$\begin{aligned} \mathbb{E}_S[Y_j \mathbb{1}_r^c] &= \mathbb{E}_{i,j,r}[x_j \text{sign}(A_i^T y) \mathbb{1}(|A_r^T y| \leq \tau)] \\ &= \mathbb{E}_S[x_j \left[\text{sign}(A_i^T y) - \text{sign}_\tau(A_i^T y_{-j}\right] \mathbb{1}(|A_r^T y| \leq \tau)] \\ &+ \mathbb{E}_S[x_j \text{sign}_\tau(A_i^T y_{-j} \left[\mathbb{1}(|A_r^T y| \leq \tau) - \mathbb{1}(|A_r^T y_{-j}| \leq \tau)\right]] \\ &+ \mathbb{E}_S\left[x_j \text{sign}_\tau(A_i^T y_{-j} \mathbb{1}(|A_r^T y_{-j}| \leq \tau)\right] \end{aligned}$$

The last line has expectation 0, so it suffices to control the second- and third-to-last terms. Using a similar line of argument in Claim A.1.3, the the non-correlation Corrolary 3.2.7 yields

$$\begin{aligned} & |\mathbb{E}_S[x_j \text{sign}_\tau(A_i^T y_{-j} \left[\mathbb{1}(|A_r^T y| \leq \tau) - \mathbb{1}(|A_r^T y_{-j}| \leq \tau)\right]]| \\ \leq\ & \mathbb{E}_S[|x_j| \left|\mathbb{1}(|A_r^T y| \leq \tau) - \mathbb{1}(|A_r^T y_{-j}| \leq \tau)\right|] \\ \lesssim\ & \rho |A_r^T A^*{}_j| \end{aligned}$$

Intuitively, the factor of ρ captures the degree of smoothness, and the factor of $|A_r^T A^*{}_j|$ results from the fact that $A_r^T y$ and $A_r^T y_{-j}$ differ by a term of size $|A_r^T A^*{}_j x_j|$. Similarly,

$$|\mathbb{E}_S[x_j \left[\text{sign}(A_i^T y) - \text{sign}_\tau(A_i^T y_{-j}\right] \mathbb{1}(|A_r^T y| \le \tau)]| \lesssim \rho |A_i^T A^*{}_j| \tag{1.16}$$

Putting these bounds together completes the claim. □

The proof of the following lemma is slightly more involved:

Lemma A.1.5. *The following two bounds hold*

$$|\mathbb{E}_{i,j}[Y_j \mathbb{1}_j]| \lesssim \rho |A_i^T A^*{}_j| + \tau\rho |A_j^T A^*{}_i| + \rho^2\tau^2 \|A^*\| \sqrt{\frac{k}{m}} + \gamma \tag{1.17}$$

and

$$\left|\mathbb{E}_{i,j}\left[Y_j \mathbb{1}_j^c\right]\right| \lesssim \tau\rho A_j^T A^*{}_i + \rho^2\tau^2 \|A^*\| \sqrt{\frac{k}{m}} \tag{1.18}$$

Proof. We prove the first claim; the second will follows from essentially the same argument. Let S be a set which contains i, j, and set $Z_{i,j} = A_i^T \sum_{r \in S - \{i,j\}} A_r^* x_r$. Write $\mathbb{1}_j = \mathbb{1}_j \mathbb{1}(x_j > 2\tau) + \mathbb{1}_j \mathbb{1}(x_j \le 2\tau)$. Since $j \in \hat{S}$ when $\mathbb{1}(x > 2\tau)$ with probability $1 - n^{-\omega(1)}$, we have that by Claim A.1.3 that

$$\begin{aligned}
|\mathbb{E}_S\left[Y_j \mathbb{1}_j \mathbb{1}(|x_j| > 2\tau)\right]| &\le |\mathbb{E}\left[Y_j \mathbb{1}(|x_j| > 2\tau)\right]| + \gamma \\
&\lesssim \rho |A_i^T A^*{}_j| \mathbb{E}_S[|x_j^2 \mathbb{1}(|x_j| > 2\tau)|] \\
&\le \rho |A_i^T A^*{}_j| \mathbb{E}_S[|x_j^2|] \\
&\lesssim \rho |A_i^T A^*{}_j|
\end{aligned}$$

The term $\mathbb{E}_S\left[Y_j \mathbb{1}_j \mathbb{1}(|x_j| \le 2\tau)\right]$ is controlled by Claim A.1.6 by

$$|\mathbb{E}_S\left[Y_j \mathbb{1}_j \mathbb{1}(|x_j| \le 2\tau)\right]| \lesssim \tau\rho A_j^T A^*{}_i + \min(\rho, \rho^2\tau^3) |A_i^T A^*{}_j| + \rho^2\tau^2 \mathbb{E}_S[|Z_{i,j}|] \tag{1.19}$$

Taking the expectation over all sets $S \supset \{i, j\}$ give

$$|\mathbb{E}_{i,j}\left[Y_j \mathbb{1}_j \mathbb{1}(|x_j| \le 2\tau)\right]| \lesssim \tau\rho A_j^T A^*{}_i + \rho^2\tau^2 \left(\tau |A_i^T A^*{}_j x_j| + \mathbb{E}_{S: S \supset \{i,j\}}[\mathbb{E}_S[|Z_{i,j}|])\right]$$

It now suffices to control the term $\mathbb{E}_{S: S \supset \{i,j\}}[\mathbb{E}_S[|Z_{i,j}|]]$. By Jensen's inequality it holds that

$$\begin{aligned}
\mathbb{E}_{S: S \supset \{i,j\}}[\mathbb{E}_S[|Z_{i,j}|]] &\le \sqrt{\mathbb{E}_{S: S \supset \{i,j\}}[\mathbb{E}_S[|Z_{i,j}|^2]]} \\
&\le \sqrt{\mathbb{E}_{i,j}[|Z_{i,j}|^2]]} \\
&\le \sqrt{\mathbb{E}_{i,j}[A_i^T A^*{}_{-\{i,j\}} x_{-i,j} x_{-i,j}^T] A^{*T}{}_{-\{i,j\}} A_i} \\
&= \sqrt{A_i^T A^*{}_{-\{i,j\}} \mathbb{E}_{i,j}[x_{-i,j} x_{-i,j}^T] A^{*T}{}_{-\{i,j\}} A_i} \\
&\le \|A^*\| \cdot \sqrt{\|\mathbb{E}_{i,j}[x_{-i,j} x_{-i,j}^T]\|}
\end{aligned}$$

Now, $\mathbb{E}_{i,j}[x_{-i,j} x_{-i,j}^T] = \text{diag}(\mathbb{E}_{i,j}[x_r^2])_{r \ne i,j}$, which has spectral norm $\lesssim k/m$. Hence, the above expression is bounded by above by $\sqrt{\|A^*\|^2 \cdot k/m} \lesssim \sqrt{k/n}$, as needed. □

Claim A.1.6. *Let* $Z_{i,j} = A_i^T \sum_{r \neq i,j} A^*{}_r x_r$. *Then*

$$|\mathbb{E}_S [Y_j \mathbb{1}_j \mathbb{1}(|x_j| \leq 2\tau)]| \lesssim \tau\rho |A_j^T A^*{}_i| + \min(\rho, \rho^2\tau^3) \cdot |A_i^T A^*{}_j| + \rho^2\tau^2 \mathbb{E}_S[|Z_{i,j}|] \tag{1.20}$$

Similarly,

$$\left|\mathbb{E}_S \left[Y_j \mathbb{1}_j^c\right]\right| \lesssim \tau\rho |A_j^T A^*{}_i| + \min(\rho, \rho^2\tau^3) \cdot |A_i^T A^*{}_j| + \rho^2\tau^2 \cdot \mathbb{E}_S[|Z_{i,j}|]) \tag{1.21}$$

Proof. We begin by proving Equation 1.20. Recall the definition $Y_j := x_j \mathrm{sign}_\tau(A_i^T y)$ and $\mathbb{1}_j := \mathbb{1}(j \in \hat{S}) = \mathbb{1}(|A_j^T y| > \tau)$. Define $\mathbb{1}_{j,i} := \mathbb{1}(|A_j^T y_{-i}| > \tau)$, so that by linearity of expectations we have

$$\mathbb{E}_S [Y_j \mathbb{1}_j \mathbb{1}(|x_j| \leq 2\tau)] \;=\; \mathbb{E}_S \left[x_j \mathbb{1}(|x_j| \leq 2\tau) \mathrm{sign}_\tau(A_i^T y) \mathbb{1}_j\right] = E_1 + E_2 + E_3 \tag{1.22}$$

Where

$$\begin{aligned}
E_1 &= \mathbb{E}_S \left[x_j \mathbb{1}(|x_j| \leq 2\tau) \mathrm{sign}_\tau(A_i^T y) \left(\mathbb{1}_j - \mathbb{1}_{j,i}\right)\right] \\
E_2 &= \mathbb{E}_S \left[x_j \mathbb{1}(|x_j| \leq 2\tau) \mathbb{1}_{j,i} \left(\mathrm{sign}_\tau(A_i^T y) - \mathrm{sign}_\tau(A_i^T A^*{}_i x_i)\right)\right] \\
E_3 &= \mathbb{E}_S \left[x_j \mathbb{1}(|x_j| \leq 2\tau) \mathbb{1}_{j,i} \mathrm{sign}_\tau(A_i^T A^*{}_i x_i)\right]
\end{aligned}$$

E_3 is exactly zero, since $x_j \mathbb{1}(|x_j| \leq 2\tau)\mathbb{1}_{j,i}$ depend only on the nonzero entries of x other than x_i, while $\mathrm{sign}_\tau(A_i^T A^*{}_i x_i)$ has mean zero and depends only on x_i. To control $|E_1|$, we have

$$\begin{aligned}
|\mathbb{E}_S \left[x_j \mathbb{1}(|x_j| \leq 2\tau) \mathrm{sign}_\tau(A_i^T y)(\mathbb{1}_j - \mathbb{1}_{j,i})\right]| &\leq \mathbb{E}_S \left[|x_j \mathbb{1}(|x_j| \leq 2\tau) \mathrm{sign}_\tau(A_i^T y)(\mathbb{1}_j - \mathbb{1}_{j,i})|\right] \\
&\leq 2\tau \mathbb{E}_S \left[|\mathbb{1}_j - \mathbb{1}_{j,i}|\right] \\
&\lesssim \frac{\tau\rho A_j^T A^*{}_i}{A_j^T A^*{}_j} \\
&\asymp \tau\rho A_j^T A^*{}_i
\end{aligned}$$

To control the E_2 term, let $\tilde{Z} := A_i^T y_{-i}$. Then $\mathrm{sign}_\tau(A_i^T y) = \mathrm{sign}_\tau(A_i^T A^*{}_i x_i + \tilde{Z})$, so thatt

$$\begin{aligned}
& |\mathbb{E}_S \left[x_j \mathbb{1}(|x_j| \leq 2\tau) \mathbb{1}_{j,i} \left(\mathrm{sign}_\tau(A_i^T y) - \mathrm{sign}_\tau(A_i^T A_i^* x_i)\right)\right]| \\
= \;& |\mathbb{E}_S \left[x_j \mathbb{1}(|x_j| \leq 2\tau) \mathbb{1}_{j,i} \cdot \mathbb{E}\left[\left(\mathrm{sign}_\tau((A_i^T A^*{}_i x_i + \tilde{Z}) - \mathrm{sign}_\tau(A_i^T A^*{}_i x_i)\right) \left|\{x_r\}_{r \neq i}\right.\right]\right]| \\
\leq \;& \mathbb{E}_S \left[|x_j \mathbb{1}(|x_j| \leq 2\tau) \mathbb{1}_{j,i}| \cdot |\mathbb{E}\left[\left(\mathrm{sign}_\tau((A_i^T A_i^* x_i + \tilde{Z}) - \mathrm{sign}_\tau(A_i^T A_i^* x_i)\right) \left|\{x_r\}_{r \neq i}\right.\right]|\right] \\
= \;& \mathbb{E}_S \left[|x_j \mathbb{1}(|x_j| \leq 2\tau) \mathbb{1}_{j,i}| \cdot |\mathbb{E}\left[\left(\mathrm{sign}_\tau((A_i^T A_i^* x_i + \tilde{Z}) - \mathrm{sign}_\tau(A_i^T A_i^* x_i)\right) \left|\tilde{Z}\right.\right]|\right] \\
\lesssim \;& \frac{\rho}{A_i^{*T} A_i} \mathbb{E}[|\tilde{Z}| x_j \mathbb{1}(|x_j| \leq 2\tau)] \\
\lesssim \;& \rho \mathbb{E}_S[|\tilde{Z}| x_j \mathbb{1}(|x_j| \leq 2\tau)]
\end{aligned}$$

where the penultimate step follows from Proposition C.2.2, and last step follows from the fact that $A_i^{*T} A_i = 1 \pm \mathrm{o}(1)$. To conlude, set $Z_{i,j} := A_i^T y_{-i,j}$, so that $|\tilde{Z}| \leq |A_i^T A^*{}_j x_j| + |Z_{i,j}|$. Using Proposition C.2.3 to bound $\mathbb{E}_S[x_j^2 \mathbb{1}(|x_j| \leq 2\tau)]$ by $O(\rho\tau^3)$, we have

$$\begin{aligned}
|E_2| &\lesssim |A_i^T A^*{}_j| \rho \mathbb{E}_S[x_j^2 \mathbb{1}(|x_j| \leq 2\tau)] + \rho \mathbb{E}_S[|Z_{i,j}| |x_j| \mathbb{1}(|x_j| \leq 2\tau)] \\
&\lesssim |A_i^T A^*{}_j| \min(\rho, \rho^2\tau^3) + \rho^2\tau^2 \mathbb{E}_S[|Z_{i,j}|] \\
&= \min(\rho, \rho^2\tau^3)\tau |A_i^T A^*{}_j| + \rho^2\tau^2 \mathbb{E}_S[|Z_{i,j}|]
\end{aligned}$$

Summing up the bounds on $|E_1|$ and $|E_2|$ yields the first display in the claim. To get the second claim, note that $j \in \hat{S}$ with very high probability as long as $|x_j| \geq 2\tau$. Hence, we have that

$$\left|\mathbb{E}_S\left[Y_j \mathbb{1}_j^c\right]\right| \leq \gamma + \left|\mathbb{E}_S\left[Y_j \mathbb{1}_j^c \mathbb{1}_j(|x_j| \leq 2\tau)\right]\right| \tag{1.23}$$

The second term on the right can now be controlled is the same fashion as $|\mathbb{E}_S\left[Y_j \mathbb{1}_j \mathbb{1}(|x_j| \leq 2\tau)\right]|$ to get the desired bound. □

A.2 Sample Complexity Analysis

In this section, we derive the needed sample complexity bounded to prove Theorem 3.2.1, Theorem 3.2.2], and Theorem 4.1.1. Section A.2.1 establish that empirical the gradients are concentrated around their mean (possibly after rescaling), and Section A.2.2 shows that the empirical gradients remain nearness-well-conditioned.

Proof of Theorem 3.2.1 and Theorem 3.2.2. The gradients $\hat{g}$ are nearness well-conditioned by Lemma A.2.8 as long as $p = \tilde{\Omega}(mk)$ samples in the Toy Rule, and $p = \tilde{\Omega}(mk^2)$-samples in the neural rule. For the Toy Rule, Theorem 3.2.3 ensures that $g_s^s \mathbb{E}[\hat{g}_i^s]$ is $(\Theta(1), 0, k/n)$-true, and so Lemma A.2.1 and the fact that rescalings preserve (α, δ, ζ)-trueness (see Lemma 3.1.2), it follows that $\hat{g}_i^s = cg_i^s + o(1/\sqrt{t})$ is $(\Theta(1), 0, k/n + 1/\sqrt{t})$-true with $p = \tilde{\Omega}(mt)$ samples. The desired convergences is ensured by Theorem 3.1.3. Theorem 3.2.2 follows similarly. □

We now prove Theorem 4.1.1

Proof of Theorem 4.1.1. The proof is similar to the proofs of the toy and neural rules, exact we don't care about maintaining nearness, and the gradient concentration follows from Lemma A.2.2 □

A.2.1 Concentration of the Gradients

Lemma A.2.1. *Suppose that A^s is $(\delta_s, 2)$-near to A^*, that Assumptions 1-5 hold, and that the threshold τ is (δ, C)-suitable. For an arbitrary projector matrix M_i , let $\hat{g}_i^s$ be as defined in the Update Step of Algorithm 3, and let $g_i^s = \mathbb{E}[\hat{g}_i^s]$ be defined as in Equation 1.7. Then with $M = X_i$ chosen as in the Toy Rule and $p = Omega(mt)$ samples, then with very high probability, it holds that*

$$\|\hat{g}_i^s - cg_i^s\| = o(1/t) \tag{2.24}$$

where c lies in the interval $[\frac{1}{2}, 2]$ with high probability. Furthermore, if τ is chosen so that $\tau = \tilde{O}(\mu\sqrt{k}/n + \delta)$, then

1. *With $M_i = I - \sum_{j \in \hat{S}} n_j A_j A_j^T$ is chosen as in the Neural Update Rule and with $p = \tilde{\Omega}(mk^2)$, then with high probability*

$$\|\hat{g}_i^s - cg_i^s\| = \mathrm{o}\left(\frac{\mu}{\sqrt{n}} + \delta\right) \tag{2.25}$$

2. *With $M_i = M_i^{\mathrm{prj}}(\hat{S})$ as in the Projection rule and with $p = \tilde{O}(mk^{3/2}t)$, then with high probability*

$$\|\hat{g}_i^s - cg_i^s\| = \mathrm{o}\left(\delta + \frac{\tau}{kt} + \frac{\sqrt{\rho\mu^3}}{n^{3/4}t^{-1/2}}\right) \tag{2.26}$$

where c lies in the interval $[\frac{1}{2}, 2]$ with high probability.

Remark. In fact, we can define c explicitly by $c = (pq_i)^{-1} \cdot \#\{$ samples for which $i \in S\}$.

Proof of Lemma A.2.1. Let $W = \{j : i \in S\}$, $c = \frac{|W|}{|p|}$ and define

$$G_i^s = \mathbb{E}_i[M_i y^{(j)} \mathrm{sign}_\tau(A_i^T y^{(j)}) \big| i \in S] \tag{2.27}$$

Then,

$$\begin{aligned}
\hat{g}_i^{(s)} &= \frac{1}{p}\sum_{j=1}^{p} M_i y^{(j)} \mathrm{sign}_\tau(A_i^T y^{(j)}) \\
&= c\ \Pr(i \in S)\frac{1}{|W|}\sum_{j \in W} M_i y^{(j)} \mathrm{sign}_\tau(A_i^T y^{(j)})
\end{aligned}$$

for the random constant $c := \frac{|W|}{p}\Pr(i \in S)^{-1}$. Since $p \cdot \Pr(i \in S) = \tilde{\Omega}(1)$, the Multiplicative Chernoff Bound in Proposition C.1.6 ensures that we $\frac{1}{2}p\Pr(i \in S) \le |W| \le 2p\Pr(i \in S)$ with very high probability; that is, $c \in [1/2, 2]$. As a consequence,

$$\begin{aligned}
\hat{g}_i^{(s)} - cg_i &= \hat{g}_i^{(s)} - c\ \Pr(i \in S)G_i^s + \gamma && (2.28)\\
&= \hat{g}_i^{(s)} - c\ \Pr(i \in S)G_i^s + \gamma && (2.29)\\
&= \Pr(i \in S) \cdot \frac{1}{|W|}\sum_{j \in |W|} (My^{(j)} \mathrm{sign}_\tau(A_i^T y^{(j)}) - G_i^s && (2.30)\\
&= \frac{\Pr(i \in S)}{|W|} \cdot \sum_{j \in |W|} Z^{(j)} - G_i^s && (2.31)\\
& && (2.32)
\end{aligned}$$

where $Z^{(j)} := My^{(j)} \mathrm{sign}_\tau(A_i^T y^{(j)})$. Since $\frac{p}{2} \le |W| \le 2p$ with very high probability, we should observe at least $|W| = \tilde{\Omega}(pk/m)$. More precisely, we will assume that $l \ge \log^C(n)kp'/m$ for some suitably large constant C, where $p' = mt$ for the toy rule, mk^2 in the neural rule, and $mk^{3/2}t$ for the projection rule.

Relabel the samples $Z^{(j)}$ for $j \in W$ by $\{Z^{(r)}\}_{1 \le r \le l}$, and note that $Z^{(r)} \overset{iid}{\sim} My\mathrm{sign}_\tau(y^T A_i) \big| i \in S$. Hence $G_i^s = \mathbb{E}[Z^{(r)}]$. As long as $l = \mathrm{poly}(n)$, then the Truncated Bernstein Inequality (Lemma C.1.5) gives:

$$\begin{aligned}
\|\frac{1}{r}\sum_{r=1}^{l} Z^{(r)} - G_i^s\| &= \|\frac{1}{l}\sum_{r=1}^{l} Z^{(r)} - \mathbb{E}[Z^{(r)}]\| \\
&\le \tilde{O}\left(\frac{R}{l} + \sqrt{\frac{\sigma^2}{l}}\right) + \gamma
\end{aligned}$$

for any R such that $\|Z^{(r)}\| \leq R$ with probability $1 - n^{-\omega(1)}$, and $\sigma^2 \geq \mathbb{E}[\|Z^{(r)}\|^2]$. The result now follows from substituting in the values of R and σ^2 specified in Lemma A.2.3. For completeness, we walk through the exact substitutions in the case of the neural update rule. Taking $p' = mk^2$, gives $l = mk^3 \log^C(n)$ samples for which $i \in S$. Plugging in Lemma A.2.3

$$\|g_i^s - \frac{1}{r}\sum_{r=1}^{l} Z^{(r)}\| \quad \leq \quad o\left(\frac{\mu}{\sqrt{n}} + k^{-3}\tau + \delta + \rho^{1/2}\tau^{3/2}k^{-1}\right) \tag{2.33}$$

where we can take the exponent C to be large enough to kill off all log factors that would arise from Bernstein's inequality. By increasing C if necessary, we can also kill off the log factors in $\tau = \tilde{O}(\mu\sqrt{k}/n + \delta) \leq (\mu\sqrt{k}/n + \delta)\log^c(n)$ to give

$$\begin{aligned}
\|\frac{1}{r}\sum_{r=1}^{l} Z^{(r)}\| \quad &\leq \quad o\left(\frac{\mu}{\sqrt{n}} + k^{-2.5}\mu/\sqrt{n} + \delta + \rho^{1/2}n^{-3/4}k^{-1/4}\right) && (2.34)\\
&\leq \quad o\left(\frac{\mu}{\sqrt{n}} + k\delta + \mu/\sqrt{n}(\rho^{1/2}(kn)^{-1/4}/\mu\right) && (2.35)\\
&\leq \quad o\left(\frac{\mu}{\sqrt{n}} + k\delta\right) && (2.36)
\end{aligned}$$

by the scaling assumptions specified in Assumption 5. □

Remark. We remark here that the dependence between the complement projector matrix M_i and the sames increases the sample complexity by a factor of roughly k compaired to the perfect-thresholding setting with C-lower bounded distributions. Perhaps this factor can be reduced slightly, but we did not attempt to do so in this work.

Lemma A.2.2. *Suppose that A^s is $(\delta_s, 2)$-near to A^*, that Assumptions 1-5 hold, and that τ is (δ, C)-suitable. For an arbitrary projector matrix M_i , let $\hat{g}_i^s$ be as defined in the Update Step of Algorithm 3, and let $g_i^s = \mathbb{E}[\hat{g}_i^s]$ be defined as in Equation 1.7. Then, if Assumptions 1-5 hold, and there is an oracle which ensure that $\hat{S} = S$ with high probability, then*

1. *With $M_i = I - \sum_{j\in\hat{S}} n_j A_j A_j^T$ is chosen as in the Neural Update Rule, and with $p = \tilde{\Omega}(mk)$ samples, then with very high probability*

$$\|\hat{g}_i^s - cg_i^s\| = o\left(\frac{\mu}{\sqrt{n}} + \delta\right) \tag{2.37}$$

2. *With $M_i = M_i^{\mathrm{prj}}(\hat{S})$ is chosen as in the Neural Update Rule, and with $p = \tilde{O}(m)$ samples, then with very high probability*

$$\|\hat{g}_i^s - cg_i^s\| = o(\delta) \tag{2.38}$$

where c lies in the interval $[\frac{1}{2}, 2]$ with high probability.

Proof. The proof follows the same steps as Lemma A.2.1, and we omit it here for the sake of brevity. □

We now prove the bounds on R and σ^2 needed to apply Bernstein's inequality in the proofs of Lemmae A.2.1 and A.2.2:

Lemma A.2.3. *Suppose that Assumptions 1-5 holds, and let* $Z \sim M_i y \mathrm{sign}_\tau(A_i^T y) \big| i \in \hat{S}$. *Then* Z *satisfies*

$$\|Z\| \le R \quad \text{with probability } 1 - n^{-\omega(1)} \quad \text{and} \quad \mathbb{E}[\|Z\|^2] \le \sigma^2 \tag{2.39}$$

for the following constants R *and* σ^2 *in each of the update rules:*

1. *In the toy rule,* $R = \tilde{O}(\sqrt{k})$ *and* $\sigma = k$.
2. *Under the neural update rule,* $R = \tilde{O}(\mu k^{3/2}/\sqrt{n} + k\delta + k^{3/2}\delta^2 + \tau)$, *and* $\sigma = O(\mu k^{3/2}/\sqrt{n} + k\delta + k^{3/2}\delta^2 + \sqrt{k}\rho^{1/2}\tau^{3/2}$
3. *Under the projection rule,* $R = \tilde{O}(k\delta + \tau)$ *and* $\sigma = O(\sqrt{k}\rho^{1/2}\tau^{3/2} + k\delta)$

Furthermore, if $\hat{S} = S$ *with probability* $1 - n^{-\omega(1)}$, *then we can take*

1. $R = \tilde{O}(\mu k/\sqrt{n} + \sqrt{k}\delta + k\delta^2)$, *and* $\sigma = O(\mu k/\sqrt{n} + \sqrt{k}\delta + k\delta^2)$ *in the Neural Rule*
2. $R = \tilde{O}(\sqrt{k}\delta)$, *and* $\sigma = O(\sqrt{k}\delta)$ *in the Projection Rule.*

Proof. In each of the following, it suffices to control $\|My_i\|$ and $\mathbb{E}[\|My_i\|^2]$. For the toy rule, we have that $M = I - \frac{1}{n_i} A_i A_i^T$, so that

$$\|M\| \le 1 + \|A_i\| \le 2 + \delta = O(1) \tag{2.40}$$

and $\|A^*{}_S\|_2 \le O(1)$ by the Gershgorin circle theorem. Hence $\|My\| \le \|MA^*{}_S x\| \lesssim \|x\|$. Since x is k-sparse and $O(1)$ subgaussian, we have that $\mathbb{E}[\|x\|^2] = k$and $\|x\| = \tilde{O}(\sqrt{k})$ with very high probability. For the next two rules, decompose: $y = y_1 + y_2$, where

$$y_1 \;=\; \sum_{j \in S} \mathbb{1}(j \in \hat{S}) A_j^* x_j \quad \text{and} \quad y_2 = \sum_{j \in S} A^*{}_j \mathbb{1}(j \notin \hat{S}) x_j \tag{2.41}$$

Now, let $Z_1 = M_i y_1 \big| i \in S$ and let $Z_2 = M_i y_2 \big| i \in S$. The proof of the imprecise thresholding part of the claim now follows by combining Claims A.2.4 and A.2.5, and canceling out terms of lower order. The proof of the results under very high probability support recoery follows from Claim A.2.4, whilst noting that if $\hat{S} = S$ with very high probability, then $\|Z_2\| = 0$ with very high probability, and hence $\mathbb{E}[\|Z_2\|^2] = n^{-\omega(1)}$ by Lemma C.1.4. □

We start off by controlling Z_1, which captures all of the entries $j \in \mathrm{supp}(x)$ which are correctly included in $\hat{S}$.

Claim A.2.4. *Let* Z_1 *be as in the proof of Lemma A.2.3. Under the neural update rule, then*

$$\|Z_1\| \le \sqrt{k} \cdot \tilde{O}(\mu k/\sqrt{n} + \sqrt{k}\delta + k\delta^2) \tag{2.42}$$

For the projection rule, $\|Z_1\| \lesssim \tilde{O}(\delta)$. *Moreover,* $\sqrt{[\|Z_1\|^2]} = \sqrt{k} \cdot O(\mu k/\sqrt{n} + \sqrt{k}\delta + k\delta^2)$ *in the neural update rule and* $\tilde{O}(k\delta)$ *in the projection rule. If sign thresholding occurs perfectly, then we can improve all bounds by removing a factor of* $\sqrt{k}$

Proof. It suffices to control the term $M_i y = M_i A_S^* x_S$. The idea is to prove a bound on $\|MA^*_{\hat{S}}\|_F$. If sign thresholding determined all of S accurately with very high probability, we would have $MA^*_{\hat{S}} x_{\hat{S}} = M_i A_S^* x_S$ with very high probability, and we could control $\|M_i A_S^* x_S\|$ using the Hanson-Wright inequality stated in Lemma C.1.8.

Unfortunately, the entries of $x_{\hat{S}}$ are not independent in our case, due to correlation that results from innacurate thresholding. Consequently, we will rely on the more naive bound $\|M_i A^*_{\hat{S}} x_{\hat{S}}\| \le \|M_i A^*_{\hat{S}}\|_2 \|x_{\hat{S}}\| \le \|M_i A^*_{\hat{S}}\|_F \|x_{\hat{S}}\|$, which gives bounds of $O(\sqrt{k}\|M_i A^*_{\hat{S}}\|_F)$ and $\tilde{O}(\sqrt{k}\|M_i A^*_{\hat{S}}\|_F)$ in expectation and with high probability, respectively. We could imagine applying similar techniques as in expectation computations to derive tighter concentration inequalities and expectation bound for $\|M_i A^*_{\hat{S}} x_{\hat{S}}\|^2$, but for the sake of brevity and clarity, we do not undertake such efforts here. Indeed, the factor of $\sqrt{k}$ lost only effects the sample complexity of our algorithm, but not the bias.

Under the Neural Rule, a simple modification of Arora et al. (2015) Claim 47 (to account for the renomralizing by $\frac{1}{n_i}$) establishes the bound $\|MA^*_{\hat{S}}\|_F \le \mu k/\sqrt{n} + \sqrt{k}\delta + k\delta^2$. For the second point, let $M_i = P_1 + M_1$ where P_1 is the projection term onto the orthogonal complement of $A_{\hat{S}}$, and $M_1 = A_i A_i^T \left(\frac{1}{\|A_i\|^2} - \frac{1}{\|A_i\|}\right)$. Then

$$\|M_1\|_F = \|A_i A_i^T\|_F^2 |n_i^2 - n_i| = |\|A_i\| - 1| \tag{2.43}$$

But $\|A_i\| = \|A_i^*\| \pm \delta = 1 \pm \delta$, so that $\|M_1\|_F = O(\delta)$. Hence,

$$\|M_1 A^*_{\hat{S}}\|_F \le \|M_1\|_F \|A^*_{\hat{S}}\| = O(\delta) \tag{2.44}$$

where we use the fact that $\|UV\|_F \le \|U\|_F \|V\|$, and that $\|A^*_{\hat{S}}\| \le 2$ by the Gershgorin circle theorem. On the other hand,

$$\begin{aligned} \|P_1 A^*_{\hat{S}}\|_F \;=\; \|P_1(A^*_{\hat{S}} - A_{\hat{S}})\|_F \;\le\; \|P_1\|\|(A^*_{\hat{S}} - A_{\hat{S}})\|_F \;\le\; \sqrt{\sum_{i\in\hat{S}} \|A_i^* - A_i\|^2} \\ \le\; \delta\sqrt{|\hat{S}|} \le \delta\sqrt{k} \end{aligned}$$

□

We now control Z_2, which captures all of the entries $j \in \mathrm{supp}(x)$ which we omitted from $\hat{S}$.

Claim A.2.5. *Let Z_2 be as in the proof of Lemma A.2.3. Then, for the neural update rule, $\|Z_2\| \le \tilde{O}(\tau(1 + \delta\sqrt{k})$, and for the projection rule $\|Z_2\| \le \tilde{O}(\tau)$.*

Proof. Using the same arguments as in Claim A.2.4, we can bound $\|M_i\| \le (1 + \delta\sqrt{k})$. In the projection rule, we have $\|M_i\| = O(1)$. To conclude, we bound $\|y_2\|$ with the following subclaim:

Claim A.2.6. *Let $y_2 = \sum_{j\in S} A^*{}_j \mathbb{1}(j \notin \hat{S}) x_j | i \in \hat{S}$. Then with very high probability, if $\Pr(|x_j| \le 2\tau) \le 1/2$, then $\|y_2\| \le 2\tau \log^2 n$.*

Proof. It would be tempting to apply a Bernstein inequality to the mean zero random variables $A_j^* x_j$, but we cannot do this. While the x_j are independent, $\mathbb{1}(j \notin \hat{S})$ are not.

With high probability, $j \in \hat{S}$ only when $|x_j| \le 2\tau$. Hence, with high probability,

$$\begin{aligned}
\|y_2\| &= \|\sum_{j\in S} A_j^* \mathbb{1}(j \notin \hat{S}) x_j\| \\
&\le \|\sup_{w:\|w\|_\infty \le 2\tau} \sum_{j\in S} A_j^* \mathbb{1}(j \notin \hat{S}) w_j\| \\
&= \|\sup_{w:\|w\|_\infty \le 2\tau} w^T A^*_{\hat{S}}\| \\
&= \|\sup_{w:\|w\|_2 \le 2\tau\sqrt{|\hat{S}|}} w^T A^*_{\hat{S}}\| \\
&\le 2\tau\sqrt{|\hat{S}|}\|A^*_{\hat{S}}\| \\
&\lesssim 2\tau\sqrt{|\hat{S}|}
\end{aligned}$$

since $\|A^*{}_{\hat{S}}\| = O(1)$ by the Gershgorin circle theorem. Now, by the assumption that τ is (δ, C)-suitable, where C is given by assumption 5, we have that $\Pr(|x_j| \le 2\tau) \le 1/2$. Hence,

$$\begin{aligned}
\Pr(\hat{S} \ge t) &\le \gamma + \Pr(|\{j : |x_j| \le 2\tau\}| \ge t) &(2.45)\\
&\le \sum_{i\ge t}(1/2)^t = (1/2)^{t-1} &(2.46)
\end{aligned}$$

Hence, it holds with probability $n^{-\log n}$ that $|\hat{S}| \le \log^2 n$. Thus, with high probability, $\|y_2\| \le 2\tau \log^2 n$. □

□

Claim A.2.7 (Variance Calculation)**.** *We have that*

$$\mathbb{E}[\|Z_2\|^2] \lesssim k \cdot \max\{1, \rho\tau^3\} + \gamma \tag{2.47}$$

for the neural update rule and

$$\sqrt{\mathbb{E}[\|Z_2\|^2]} \lesssim k^2\delta + k \cdot \max\{1, \rho\tau^3\} + \gamma \tag{2.48}$$

in the projection based rule.

Proof. Suppose $\|M_i A^*{}_{\hat{S}}\|_2 \le R_2$ with probability $1 - n^{-\omega(1)}$. Since $\hat{S} \subset S$ with high probability, we have

$$\begin{aligned}
\mathbb{E}[\|Z_2\|^2] &\le \mathbb{E}[\|M_i A^*_{\hat{S}}\|^2 \|x_{\tilde{\hat{S}}}\|^2] &(2.49)\\
&\le R_2^2 \mathbb{E}[\|x_{\hat{S}}\|^2] + n^{-\omega(1)} &(2.50)\\
&= R_2^2 \sum_{j\in[m]} \mathbb{E}[x_j^2 \mathbb{1}(j \in \hat{S})] + n^{-\omega(1)} &(2.51)\\
&= R_2^2 \sum_{j\in S} \mathbb{E}[x_j^2 \mathbb{1}(j \in \hat{S})] + n^{-\omega(1)} &(2.52)\\
&\le R_2^2 \sum_{j\in S} \mathbb{E}[x_j^2 \mathbb{1}(|x_j| \le 2\tau)] + n^{-\omega(1)} &(2.53)\\
&\lesssim R_2^2 \max\{1, k\rho\tau^3\} + n^{-\omega(1)} &(2.54)
\end{aligned}$$

where the last step follows from Proposition C.2.3. In the case of the projection rule, $R_2 = O(1)$ with high probability, and in the case of the neurall update rule, $\|M_i A^*_{\tilde{S}}\| \le 2\|M_i\| \le 2\|\mathrm{diag}(n_i)A_{\hat{S}}\| \le 2(1+\delta)\|A_{\hat{S}}\|^2$, which is less than $4(1+\delta)\|A^*_{\hat{S}}\|^2 + 4(1+\delta)\|A^*_{\hat{S}} - \hat{A}\|^2 \lesssim 1 + k\delta^2$. Hence,

$$\begin{aligned}\mathbb{E}[\|Z_2\|^2] &\lesssim k\rho\tau^3 + k\delta^2\mathbb{E}[\|x_{\tilde{S}}\|^2] + \gamma &(2.55)\\ &\lesssim k\rho\tau^3 + k^2\delta^2 + \gamma &(2.56)\end{aligned}$$

in the Neural Update rule, while

$$\mathbb{E}[\|Z\|^2] \lesssim k\rho\tau^3 + \gamma \tag{2.57}$$

in the projection based rule. □

A.2.2 Maintaining Nearness

Lemma A.2.8. *Suppose that A^s is $(\delta^s, 2)$-near to A^*. Then, under the toy rule update rule, it holds with high probability that $\hat{g}$ is nearness well conditioned, as long as $p = \tilde{\Omega}(mk)$-samples are used. For the neural update rule, the necessary sample sample size jumps to $p = \tilde{\Omega}(mk^2)$.*

Proof. For simplicity, we drop the dependence on s. Let $\tilde{G}$ denote the matrix whose columns are the toy rule updates $X_i\ y \cdot \mathrm{sign}_\tau(A_i^T y)$. We can express $\tilde{G}$ in matrix notation as $\tilde{G} = Z_1 + Z_2$, where

$$Z_1 = y\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))^T \quad \text{and} \quad Z_2 := A_{\hat{S}}\mathrm{diag}\left(A_i^T y \mathrm{sign}_\tau(A_i^T y n_i)\right) \tag{2.58}$$

Let's control each term separately. For Z_1,

Claim A.2.9. *$\|Z_1\| \le \tilde{O}(k)$ with high probability, $\|\mathbb{E}[Z_1 Z_1^T]\| \le \tilde{O}(k^2\|A^*\|^2/m)$, and $\|\mathbb{E}[Z_1 Z_1^T]\| \le k^3/m$*

To control Z_2, we have

Claim A.2.10. *$\|Z_2\| \le \tilde{O}(1 + \sqrt{k}\delta)$ with high probability and*

$$\max(\|\mathbb{E}[Z_2 Z_2^T]\|, \|\mathbb{E}[Z_2 Z_2^T]\|) \le \tilde{\mathrm{O}}\left(\|A^*\|^2 k(1+\sqrt{k}\delta)^2/m\right) = \tilde{\mathrm{O}}\left(\|A^*\|^2 k^2/m\right) \tag{2.59}$$

Hence, applying Matrix Bernstein inequality with $p = \tilde{\Omega}\left(mk\right)$ samples gives

$$\hat{g} - g = \frac{1}{p}\sum_i \mathcal{G}^{(i)} - \mathbb{E}[\mathcal{G}^{(i)}] = \mathrm{o}\left(\frac{k}{m}\|A^*\|\right) \tag{2.60}$$

Since g is nearness well conditioned, so is $\hat{g}$ (see Remark 3.1.2). The result for the neural rule follows using similar arguments, and we omit here in the interest of brevity. As in the gradient concentration, the imperfect thresholding forces us to pick up an extra factor of k in the sample complexity required to ensure $\hat{g} - g = o(\frac{k}{m}\|A^*\|)$, as compared to Lemma 42 in Arora et al. (2015). □

Proof of Claim A.2.9. With high probability, $\mathrm{sign}_\tau(A_i^T y)$ is non zero for at most k entries, while $\|y\| = \|A_S^* x\| \le \tilde{O}(\sqrt{k})$, so that

$$\|Z_1\| \le \tilde{O}(k) \tag{2.61}$$

Furthermore, with high probability $\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))^T \mathrm{vec}(\mathrm{sign}_\tau(A_i^T y)) = \mathrm{supp}(\hat{S}) \le k$, so that with high probability

$$y\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))^T \mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))y^T \preceq kyy^T \tag{2.62}$$

and consequently

$$\begin{aligned} &\|\mathbb{E}[y\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))^T \mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))y^T]\| \\ \le\ & \gamma + k\mathbb{E}[yy^T] = kA^*\mathbb{E}[xx^T]A^{*T} = \|A^*\|^2 O(k^2/m) + \gamma \end{aligned}$$

Moreover, since $\|y\|^2 = \tilde{O}(k)$ with high probability,

$$\mathrm{sign}_\tau(A_i^T y))\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))^T \preceq \mathrm{vec}(\mathbb{1}(i \in S))\mathrm{vec}(\mathbb{1}(i \in S))^T \tag{2.63}$$

whenever $\hat{S} \subset S$, which also occurs with high probability. Noting that $\|y\|^2 \le \|x\|^2 \|A_{\hat{S}}^*\|^2 = O(\|x\|^2)$ by the Gershgorin circle theorem, and that $\|x\|^2 = \tilde{O}(k)$ by Lemma C.1.7. Hence, an applying of Lemma D.2.2 gives

$$\|\mathbb{E}[\vec{(}\mathrm{sign}_\tau(A_i^T y))^T y^T y\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))]\| \quad \le \quad \tilde{O}(k)\|\mathbb{E}[\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))^T \mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))]\| + \gamma$$

Let $Q := \mathbb{E}[\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))^T \mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))$. The diagonal components of Q are bounded above by k/m in magnitude, and its off diagonals by k^2/m^2. Hence $\|Q\| \le k^2/m$, which bounds $\mathbb{E}[\vec{(}\mathrm{sign}_\tau(A_i^T y))^T y^T y\mathrm{vec}(\mathrm{sign}_\tau(A_i^T y))]$ above by k^3/m. Note that, in the C-lower bounded case, we would have the bound $\|Q\| \le k/m + \gamma$, since $\mathrm{sign}_\tau(A_i^T y)$ and $\mathrm{sign}_\tau(A_j^T y)$ would be essentially uncorrelated. □

Proof of Claim A.2.10. For the term Z_2, we have

$$\|Z_2\| \quad \le \quad \|A_{\hat{S}}\mathrm{diag}(A_i^T y\mathrm{sign}_\tau(A_i^T y)n_i)\| \le \|A_{\hat{S}}\|\|A_i^T y\| \max_i \mathrm{diag}(n_i) \tag{2.64}$$

Reusing the same arguments seen before, we have that $\|A_{\hat{S}}\| \lesssim 1+\sqrt{k}\delta$ with high probability, that $\|A_i^T y\| = \tilde{O}(1)$ with high probability, and $\max_i n_i = O(1)$. Putting these bounds together shows that $\|Z_2\| = \tilde{O}((1 + \delta\sqrt{k})$ with high probability. Next up, we compute

$$\|\mathbb{E}[Z_2^T Z_2]\| := \mathbb{E}[\mathrm{diag}(A_i^T y\mathrm{sign}_\tau(A_i^T y)n_i)A_{\hat{S}}^T A_{\hat{S}}^T \mathrm{diag}(A_i^T y\mathrm{sign}_\tau(A_i^T y)n_i)] \tag{2.65}$$

Since $A_{\hat{S}}^T A_{\hat{S}} \preceq \|A_{\hat{S}}\|^2 I \preceq (1 + \sqrt{k}\delta)^2 I$ with high probability, we have

$$\mathbb{E}[Z_2^T Z_2] \preceq \tilde{\mathrm{O}}\left((1 + \sqrt{k}\delta)^2\right)\mathbb{E}[\mathrm{diag}\left(A_i^T y n_i \mathrm{sign}_\tau(A_i^T y)\right)^2] \preceq \tilde{\mathrm{O}}\left(\frac{k}{m}(1 + \sqrt{k}\delta)^2\right) + \gamma \tag{2.66}$$

Hence $\|\mathbb{E}[Z_2^T Z_2]\| = \tilde{\mathrm{O}}\left(\frac{k^2}{m}\right)$. On the other hand, since $A_i^T y n_i = \tilde{O}(1)$ with high probability,

$$\begin{aligned}
\|\mathbb{E}[Z_2 Z_2^T]\| &= \|\mathbb{E}[A_{\hat{S}}\mathrm{diag}\left(A_i^T y \mathrm{sign}_\tau(A_i^T y) n_i\right)^2\| && (2.67)\\
&= \|AA^T\|\|\mathbb{E}[\mathrm{diag}\left(\mathbb{1}(i \in \hat{S}) A_i^T y n_i^2\right)]\| && (2.68)\\
&\leq \|A\|^2 \max_i \mathbb{E}[\mathbb{1}(i \in \hat{S}) A_i^T y n_i^2] && (2.69)\\
&\lesssim \frac{\|A\|^2 k}{m} && (2.70)
\end{aligned}$$

□

Appendix B

Nonnegative Dictionary Learning

B.1 Proof of Proposition B.1.1

In this section, we prove the more precise form of Proposition 5.1.5, Proposition B.1.1. This version makes explicit the constant factors in Proposition B.1.1, and relates them to a notion of the "condition number" of a NOID, as laid out in the following definition:

Definition B.1.1 ((κ, τ)-well conditioned)**.** Let B be nonnegative offset incoherent dictionary B with decomposition $B = A + vc^T$, where $\|v\| = 1$, and $a \in R^m$ has entries $a_i = \|A_i\|$. We say that B is (κ, τ)-well conditioned with respect to the decomposition $B = A + cv^T$ if there exists some $\kappa \geq 1$ and $\tau \geq 1$ for which $\kappa^{-1} \leq |c_i/a_i| \leq \kappa$ and

$$\tau^{-1} \leq \left(\frac{\|a\|_1}{m}\right)^2, \left(\frac{\|c\|_1}{m}\right)^2, \left(\frac{\|a\|_2^2}{m}\right) \leq \tau \tag{1.1}$$

We are now ready to state the main proposition:

Proposition B.1.1. *Suppose that B is a nonnegative offset incoherent dictionary with decomposition $B = A + vc^T$ and incoherence parameter $\mu/\sqrt{n}$, which is also (κ, τ)-well conditioned. Define*

$$\hat{v} = \sum_i B_i \quad \text{and} \quad P := Proj_{\hat{v}^\perp} \tag{1.2}$$

Then, if $z = \frac{3\mu}{\sqrt{n}} + \frac{1}{m}$ and $6z \max(\kappa^2\sqrt{\tau}, 4\kappa\tau^2) \leq \frac{1}{10}$, we have

$$\cos(PB_i, PB_j) \leq 10\tau^2\kappa^2 z \tag{1.3}$$

Moreover,

$$\|PB_i\|^2 \geq (9/10)a_i^2 \tag{1.4}$$

Proof of Proposition B.1.1. By Lemma B.1.6, we have

$$\|PB_i\|^2 \geq a_i^2\left(1 - 6z\max(\kappa^2\sqrt{\tau}, 4\kappa\tau^2)\right) \geq \left(1 - \frac{1}{10}\right)a_i^2 \geq (9/10)a_i^2 \tag{1.5}$$

and hence, again invoking Lemma B.1.6, we have

$$\begin{aligned}\cos(PB_i, PB_j) &\leq \frac{\langle PB_i, PB_j\rangle}{\|PB_i\|\|PB_j\|} \leq \frac{1}{.9a_ia_j}\langle PB_i, PB_j\rangle \\ &\leq \frac{1}{.9a_ia_j}\cdot 8\tau^2\kappa^2 a_ia_j \leq 10\tau^2\kappa^2\end{aligned} \tag{1.6}$$

□

B.1.1 Supporting Result for Proposition B.1.1

Throughout, we assume that B is a nonnegative offset incoherent dictionary with decomposition $B = A + vc^T$ and incoherence parameter $\mu/\sqrt{n}$. We let a be the vectors whose entries are $a_i = \|A_i\|$, and assume $\|v\| = 1$. We let $\hat{v} = \frac{1}{m}\sum_i B_i$, and $P = \mathrm{Proj}_{\hat{v}^\perp}$. At this stage, we do not yet assume that B is (κ, τ) well conditioned; instead, we will establish more flexible, granular controls on the normalized inner products $\cos(PA_i, PA_j)$, and then substitute the well conditioning assumptions at the end. Finally, introduce the notation

$$\bar{a} := \frac{1}{m}\sum_i a_i \quad \bar{c} := \tfrac{1}{m}\textstyle\sum_i c_i \quad C_a := \frac{1}{m}\sum_i a_i^2 \tag{1.7}$$

We have the following Lemma:

Lemma B.1.2. *Define*

$$E := \|\hat{v}\|^2 - c^2 - C_a/m \quad \textit{and} \quad E_i := \hat{v}^T B_i - c_i\bar{c} - \frac{a_i}{m} \tag{1.8}$$

Then,

$$|E| \leq (2\bar{c}\bar{a} + \bar{a}^2)\cdot \mu/\sqrt{n} \quad \textit{and} \quad |E_i| \leq \frac{\mu(a_i\bar{a} + a_i\bar{c} + c_i\bar{a})}{\sqrt{n}} \tag{1.9}$$

Proof. For the first point, we have

$$\|\hat{v}\|^2 \;=\; \bar{c}^2 + 2\langle v\bar{c}, \frac{1}{m}\sum_i A_i\rangle + \frac{1}{m^2}\sum_i \|A_i\|^2 + \sum_{i\neq j}\langle A_i, A_j\rangle = \bar{c}^2 + E$$

where E is as given by

$$\begin{aligned}E \;&:=\; \frac{1}{m^2}\sum_i \|A_i\|^2 + 2\langle v\bar{c}, \frac{1}{m}\sum_i A_i\rangle + \frac{1}{m^2}\sum_{i\neq j}\langle A_i, A_j\rangle \\ &:=\; \frac{1}{m^2}\sum_i a_i^2 + 2\frac{\bar{c}}{m}\sum_i a_i\cos(A_i, v) + \frac{1}{m^2}\sum_{i\neq j} a_ia_j\cos(A_i, A_j)\end{aligned}$$

and hence

$$|E| \;\leq\; C_a/m + \left(2\bar{c}\bar{a} + \bar{a}^2\right)\mu/\sqrt{n}$$

For the second point,

$$\hat{v}^T B_i \;=\; \sum_j \langle \frac{1}{m} A_j + c_j v, A_i + c_i v\rangle = c_i \bar{c} + E_i$$

where

$$\begin{aligned} E_i &:= \frac{1}{m}\left(\sum_j \langle A_j, A_i\rangle + \sum_j \langle c_j v, A_i\rangle + \sum_j \langle A_j, c_i v\rangle\right) \\ &:= \frac{1}{m}\left(a_i \sum_j a_j \cos(A_i, A_j) + \cos(A_i, v)\sum_j a_i c_j + \sum_j a_j c_i \cos(A_j, v)\right) \end{aligned}$$

and hence

$$|E_i| \;\le\; \frac{a_i}{m} + \frac{a_i \sum_{j\neq i} a_j}{m}\cdot\frac{\mu}{\sqrt{n}} + \bar{c}\frac{\mu}{\sqrt{n}} + c_i\frac{\mu}{\sqrt{n}} \le \frac{\mu(a_i\bar{a} + a_i\bar{c} + c_i\bar{a})}{\sqrt{n}} + \frac{a_i}{m}$$

□

We now need two other small technical results. Set $Q = Proj_{\hat{v}} = \frac{1}{\|\hat{v}\|^2}\hat{v}\hat{v}^T$and $\tilde{Q} = \frac{1}{\bar{c}^2}\hat{v}\hat{v}^T$. Note that $P = I - Q$. We have the following two facts:

Claim B.1.3. $\left\|\tilde{Q} - Q\right\| \le \frac{|E|}{\bar{c}^2}$

Proof. We see that $\tilde{Q} - Q = (1 - \frac{\|\hat{v}\|^2}{\bar{c}^2})Q$ and $\|Q\| = 1$, so $\|\tilde{Q} - Q\| \le \frac{|E|}{\bar{c}^2}$ □

The triangle inequality gives the last claim:

Claim B.1.4. $\|B_i\| \le a_i + c_i$

Before proving our next lemma, we establish yet more notation:

$$E_i' = E_i + \frac{a_i}{m} \quad \text{and} \quad \tau_i := \inf\{t : a_i + mE_i \le t a_i\} = \inf\{t : mE_i' \le t\} \tag{1.10}$$

and

$$\tau_0 := \inf\{t \ge: E \le t(\bar{c}^2 + C_a/m)\} \quad \text{and} \quad \tau := \frac{1}{1-\tau_0} \tag{1.11}$$

In general $\tau_0 = o(1)$, since $E \le O(\mu/\sqrt{n})$, and hence $\tau \approx 1$. With this new notation, we can put together Lemma B.1.2 and Lemma B.1.3 to conclude:

Lemma B.1.5.

$$\begin{aligned} \|PB_i\|^2 &\ge a_i^2 + (1-\tau)c_i^2 - \tau\left(\frac{\tau_i^2 a_i^2}{C_a m} + \frac{\tau_i c_i a_i}{\sqrt{C_a m}}\right) \\ |\langle PB_i, PB_j\rangle| &\le (a_i a_j + c_i a_i + c_j a_j)\frac{\mu}{\sqrt{n}} \\ &+ \left|\frac{(c_i\tau_j a_j + c_j\tau_i a_i) + a_i a_j \tau_i\tau_j \frac{1}{m\bar{c}}}{m\bar{c}}\right| \\ &+ |E| c_i c_j + \frac{|E|\left(c_i\tau_j a_i + c_j\tau_i a_j\right)/m}{\bar{c}} + \frac{\tau_i\tau_j a_i a_j |E|}{m^2\bar{c}^2} \end{aligned}$$

Proof. Using the fact that $P + Q = I$ and P and Q are orthogonal projections and then the triangle inequality, we can write

$$\begin{aligned}
\|PB_i\|^2 &= \|B_i\|^2 - \|QB_i\|^2 \\
&\geq a_i^2 + c_i^2 - \frac{\mu}{\sqrt{n}} a_i c_i - \frac{(c_i \bar{c} + \frac{a_i \tau_i}{m})^2}{\bar{c}^2 + C_a/m - E} \\
&= a_i^2 + c_i^2 - \frac{\mu}{\sqrt{n}} a_i c_i - \frac{c_i^2 \bar{c}^2 + \frac{\tau_i a_i}{m} c_i \bar{c} + (\tau_i^2 a_i^2/m^2)}{\bar{c}^2 + C_a/m - E} \\
&\geq a_i^2 + c_i^2 - \frac{\mu}{\sqrt{n}} a_i c_i - \tau \frac{c_i^2 \bar{c}^2 + \frac{\tau_i a_i}{m} c_i \bar{c} + (\tau_i^2 a_i^2/m^2)}{\bar{c}^2 + C_a/m}
\end{aligned}$$

Now, we have

$$\frac{c_i^2 \bar{c}^2}{\bar{c}^2 + C_a/m} \leq c_i^2 \qquad \text{and} \qquad \frac{\tau_i^2 \frac{a_i^2}{m^2}}{\bar{c}^2 + C_a/m} \leq \frac{\tau_i^2 a_i^2}{C_a m}$$

Thus, using the indentity $a^2 + b^2 \geq 2ab$, it holds

$$\frac{2(\tau_i \left(a_i c_i \bar{c}/m\right)}{\bar{c}^2 + C_a/m} \leq \frac{2 a_i c_i \bar{c}/m}{2\sqrt{C_a}\sqrt{m}\bar{c})} = \frac{\tau_i c_i a_i}{\sqrt{C_a m}} \tag{1.12}$$

Hence, replacing $\frac{1}{1-\tau_0}$ with τ, we have

$$\|PB_i\|^2 \geq a_i^2 + (1-\tau) c_i^2 - \tau \left(\frac{\tau_i^2 a_i^2}{C_a m} + \frac{\tau_i c_i a_i}{\sqrt{C_a m}} \right)$$

For the next point, we have

$$\begin{aligned}
\langle PB_i, PB_j \rangle &= B_i^T P^2 B_j = B_i^T P B_j = B_i^T B_j - B_i^T Q B_j \\
&= B_i^T B_j - B_i^T \tilde{Q} B_j + B_i^T \left(Q - \tilde{Q} \right) B_j^T \\
&= B_i^T B_j - \frac{1}{\bar{c}^2} (\hat{v}^T B_i)(\hat{v}^T B_j) + B_i^T \left(Q - \tilde{Q} \right) B_j^T \\
&= B_i^T B_j - (c_i - \frac{1}{\bar{c}} E_j') \left(c_j - \frac{1}{\bar{c}} E_j \right) + B_i^T \left(Q - \tilde{Q} \right) B_j^T \\
&= B_i^T B_j - c_i c_j - \frac{c_i E_j' + c_j E_i' + \left(E_i' E_j' \right) \bar{c}^{-1}}{\bar{c}} + B_i^T \left(Q - \tilde{Q} \right) B_j^T
\end{aligned} \tag{1.13}$$

Now,

$$\begin{aligned}
\left| B_i^T B_j - c_i c_j \right| &= \left| c_i c_j \|v\|^2 + \left(c_j A_i + c_i A_j \right)^T v + \langle A_i, A_j \rangle - c_i c_j \right| \\
&= \left| \left(c_j A_i + c_i A_j \right)^T v + a_i a_j \cos\left(A_i, A_j \right) \right| \\
&= \left(a_i a_j + c_i a_i + c_j a_j \right) \frac{\sqrt{\mu}}{\sqrt{n}}
\end{aligned} \tag{1.14}$$

and finally

$$\begin{aligned}\left\|B_i^T(Q-\tilde{Q})B_j^T\right\| &\le \frac{|E|}{\bar{c}^2}\left(\frac{a_i}{m}+\bar{c}c_i+E_i\right)\left(\frac{a_j}{m}+\bar{c}c_j+E_i\right)\\ &= |E|\,c_ic_j+\frac{|E|\left(c_i\frac{\tau_ja_j}{m}+c_j\frac{\tau_ia_i}{m}\right)}{\bar{c}}+\frac{\tau_i\tau_ja_ia_j\,|E|}{m^2\bar{c}^2}\end{aligned} \tag{1.15}$$

Replacing $E_i' = \tau_i a_i/m$ into Equation 1.13 concludes the lemma. □

Now we make use of the regularity assumptions in Proposition B.1.1, by applying them to Lemma B.1.2:

Corollary. *Let $z = 3\mu/\sqrt{n}$ be as given by Proposition B.1.1. Furhtermore, assume that B is (κ,τ) well conditioned: that is, there is some $\kappa > 0$ for which $\kappa^{-1} \le a_i \le \kappa$and $\kappa^{-1} \le c_i \le \kappa$, and that $C_a, \bar{a}^2, \bar{c}^2 \in [\tau^{-1},\tau]$. Then*

$$|E| \le \tau z \quad \textit{and} \quad |E_i| \le a_i\kappa\sqrt{\tau}z \tag{1.16}$$

Proof. We have

$$|E| \le \left(2\bar{c}\bar{a}+\bar{a}^2\right)\cdot\mu/\sqrt{n} \quad \text{and} \quad |E_i| \le \frac{\mu(a_i\bar{a}+a_i\bar{c}+c_i\bar{a})}{\sqrt{n}}$$

By our assumptions, $\bar{c}\bar{a} \le \kappa^2$, so that $|E| \le \left(2\bar{c}\bar{a}+\bar{a}^2\right)\cdot\mu/\sqrt{n} \le 3\tau\mu/\sqrt{n} = \tau z$. On the other hand, $(a_i\bar{a}+a_i\bar{c}+c_i\bar{a}) \le a_i\left(\bar{a}+\bar{c}+(c_i/a_i)\bar{a}\right) \le (2+\kappa)\sqrt{\tau} \le 3\kappa\sqrt{\tau}$. This gives the second bound. □

Putting together Corollary B.1.1 with Lemma B.1.5 yields:

Lemma B.1.6.

$$\|PB_i\|^2 \ge a_i^2\left(1-6z\max(\kappa^2\sqrt{\tau},4\kappa\tau^2)\right) \quad \textit{and} \quad |\langle PB_i, PB_j\rangle| \le 8\tau^2\kappa^2a_ia_j \tag{1.17}$$

Proof of Lemma B.1.6. Let E_i and E be as in Lemma B.1.2. Then

$$\begin{aligned}\|PB_i\|^2 &\ge a_i^2 - 3\kappa a_i\left(|E_i|+2|E|a_i\kappa\tau\right)\\ &\ge a_i^2 - \left(3\kappa^2a_i^2\sqrt{\tau}+2a_i^2\kappa\tau^2\right)\left(\frac{3\mu}{\sqrt{n}}+\frac{1}{m}\right)\\ &\ge a_i^2(1-2\max(3\kappa^2\sqrt{\tau},2\kappa\tau^2)z\\ &\ge a_i^2\left(1-6z\max(\kappa^2\sqrt{\tau},4\kappa\tau^2)\right)\end{aligned}$$

On the other hand,

$$\begin{aligned}\|\langle PB_i, PB_j\rangle\| &\le (a_ia_j+c_ia_j+c_ja_i)\sqrt{\mu}\sqrt{n}+\left|\frac{c_iE_j+c_jE_i+(E_iE_j)\bar{c}^{-1}}{\bar{c}}\right|+\frac{(a_i+c_i)(a_j+c_j)|E|}{\bar{c}^2}\\ &\le 3\kappa^2a_ia_j\sqrt{\mu}\sqrt{n}+(a_i\kappa E_j+a_j\kappa E_i)\sqrt{\tau}+\tau E_iE_j+(1+\kappa)^2a_ia_j\frac{|E|}{\bar{c}^2}\\ &\le 3\kappa^2a_ia_j\sqrt{\mu}\sqrt{n}+2a_ia_j\kappa^2\tau z+z^2\tau^2a_{i,j}+4\kappa^2a_ia_j\tau^2z\\ &\le \kappa^2a_ia_j\left(3\frac{\sqrt{\mu}}{\sqrt{n}}+7\tau^2z\right)\\ &\le 8\tau^2\kappa^2a_ia_j\end{aligned}$$

□

B.2 Proof of Lemma 5.1.9

Proof of Lemma 5.1.9. Let $\mu^* = \mathbb{E}[x]$, and let $\mu = \sum_j x^{(j)}$. Then, if we define

$$\Sigma := \frac{1}{N}\sum_j (x^{(j)} - \mu^*)(x^{(j)} - \mu^*)^T \tag{2.18}$$

then we have

$$\begin{aligned}\tilde{\Sigma} = \Sigma^* + (\Sigma - \Sigma^*) + \frac{1}{N}\sum_j (x^{(j)} - \mu^*)(\mu - \mu^*)^T \\ + \frac{1}{N}\sum_j (\mu - \mu^*)(x^{(j)} - \mu^*)^T + (\mu - \mu^*)(\mu - \mu^*)^T\end{aligned} \tag{2.19}$$

First, we see that the singular values of $\mathbb{E}[\Sigma] = \Sigma^*$ are all $\Theta(k/m)$, it suffices to show that , once show that the remaining terms in the above $o(k/m)$ display Equation 2.19 are $o(k/m)$ in norm.

First, since the vectors $x^{(j)} - \mu^*$ are $O(1)$-subgaussian, the satisfy $\|x^{(j)} - \mu^*\| \le \tilde{O}(\sqrt{k}$ with very high probability, we have by Corollary C.1.1 that

$$|\Sigma - \Sigma^*| \le \|\Sigma^*\|^{1/2}\tilde{O}(\frac{\sqrt{k}}{\sqrt{N}}) + \frac{k}{N} \tag{2.20}$$

Hence, taking $N = \tilde{\Omega}(m)$, we can ensure that $|\Sigma - \mathbb{E}[\Sigma]| = o(k/m)$. For the next term, we have

$$\|\frac{1}{N}\sum_j (x^{(j)} - \mu^*)(\mu - \mu^*)^T\| \quad \le \quad \|\mu - \mu^*\|\|\sqrt{\|\Sigma^*\|}\| = O(\|\mu - \mu^*\|\|\Sigma^*\|) = O(\|\mu - \mu^*\| \cdot \sqrt{k/m}) \tag{2.21}$$

with high probability. The term after that is bounded similarly.

Finally, the last term in Equation 2.19 has norm no more than $\|\mu - \mu^*\|^2$. Hence, it suffices to ensure that $\mu - \mu^*\| \le o(\sqrt{k/m})$, and we can write

$$\mu - \mu^* = \frac{1}{N}\sum_j x^{(j)} - \mathbb{E}[x] \tag{2.22}$$

A routine application of truncated Bernstein bounds shows that for $N = \tilde{O}(m)$, $\sum_j x^{(j)} - \mathbb{E}[x] = o(\sqrt{k}/m)$. □

B.3 Concentration Results

We establish some concentration results of (C, k)-sparse distributions.

Claim B.3.1. *If x is the coefficient vector from a (C, k)-sparse distribution, then $\|x\| \lesssim \sqrt{k} + k\log n$ with with high probability*

Proof. Write $x = \mathbb{E}[x] + (x - \mathbb{E}[x])$. The first has has norm $\mathrm{O}\left(\sqrt{k}\right)$, and the second has norm $O(\sqrt{k} + \sqrt{\log(1/\delta)}$ with probability $1 - \delta$ by Lemma C.1.7. Taking $\delta = \log^2 n$ proves the claim. □

As a consequence,

Claim B.3.2. *Let* $Z \sim M_i^{\mathrm{prj}}\ y\mathrm{sign}_\tau(A_i^T y)\big|i \in S$. *Then,* $\|Z\| \le \sqrt{k}\delta\|x\| \lesssim (k + \sqrt{k}\log n)\delta$ *with high probability*

Proof. Write $Z = \sum_{j\in S} M_i^{\mathrm{prj}} A_j x_j \mathrm{sign}_\tau(A_i^T y)$. For $j \in S$ Let B be the matrix whose columsn are $M_i^{\mathrm{prj}} A_j = O(\delta)$ and let $\tilde{x}$ be the vector whose entries are $x_j \mathrm{sign}_\tau(A_i^T y)$. Then $\|Z\| = \|B\|\|\tilde{x}\| \le \sqrt{k}\delta\|\tilde{x}\|$. The claim now follows from Claim B.3.1 □

B.4 Proof of Theorem 5.2.1

The convergence rate in Theorem 5.2.1 follows from combining the following lemma with, Theorem 3.1.3, and the fact that (by a multiplicative Chernoff Bound) $\tilde{Omega}(mk)$-samples suffice to get $p = \tilde{\Omega}\left(k^2\right)$ samples for the update rule in Algorithm 8. The signed-closeness follows from the analysis of Algorithm 10 in Proposition 5.2.11

Lemma B.4.1. *Suppose that* A^s *is* $(\delta_s, 2)$*-near to* A^*, *that* $z = y^{(1)} - y^{(2)}$ *comes from a* (C, k)*-favorable nonnegative, and that the estimated support* $\hat{S}$ *is equal to* S *with high probability. Let* $M_i = M_i^{\mathrm{prj}}$, *let* $\hat{g}_i^s$ *be as defined in the Update Step of Algorithm 8, and let* $g_i^s = \mathbb{E}[\hat{g}_i^s]$. *Then, given* $p = \tilde{\Omega}\left(k^2\right)$*-samples as chosen in Algorithm 8, it holds that*

$$q_i \hat{g}_i^s \quad \textit{is } (\Theta, \delta, n^{-\omega(1)}) \textit{ true} \tag{4.23}$$

Proof. Let $Z^{(r)}$ be iid copies of $Z \sim M_i^{\mathrm{prj}}\ z\mathrm{sign}_\tau(A_i^T y)\big|i \in \mathrm{supp}(x^{(1)}) \cap \mathrm{supp}(x^{(2)})$, and note that $\mathbb{E}[Z^{(r)}] = G_i$, where $G_i := \mathbb{E}_i[M_i^{\mathrm{prj}} z\mathrm{sign}(A_i^T z)]$. Bounding $\|Z\|$ and $\|\mathbb{E}[\|Z\|^2]\|$ by $\tilde{\mathrm{O}}\left(k\delta\right)$ (whp) and $\tilde{\mathrm{O}}\left(k^2\delta^2\right)$ Claim B.3.2 and a trunctated Bernstein inequality, we have that

$$\frac{1}{p}(\sum_{r=1}^{p} Z^{(r)} - \mathbb{E}[Z^{(r)}]) \le \mathrm{o}\left(\delta\right) + n^{-\omega(1)} \tag{4.24}$$

with high probability. Now, let $\tilde{z}^{(r)}$ have the distribution of the samples chosen for g_i by Algorithm 8. As Proposition 5.2.7, we define the event $E^{(r)}$ that all samples $\tilde{z}^{(r)}$ up to sample r where chosen from the first $2r$ samples $y = A^* x$ for which $i \in \mathrm{supp}(x)$. Since $E^{(r)}$ occurs with high probability, and $Z^{(r)}, \tilde{z}^{(r)}$ have the same distribution under E, the result now follows from the fact that, as show in the proof of Theorem 5.2.7, $q_i G_i$ is $(\Theta(1), \delta, n^{-\omega(1)})$-true. □

B.5 Sign Thresholding

Proof of Lemma 5.2.5. Let $S = \mathrm{supp}(x)$

$$A^{*T}{}_i y \ge x_i \mathbb{1}(i \in S) + \sum_{j\neq i} x_j |\langle A^{*T}{}_i A^*{}_j|\rangle| \tag{5.25}$$

$$\ge x_i \mathbb{1}(i \in S) - \tau \sum_{j\neq i \in \mathrm{supp}(x)} x_j \tag{5.26}$$

$$\ge x_i \mathbb{1}(i \in S) - O((k + \sqrt{k}\log n)\tau) \tag{5.27}$$

with high probability using the subguassian concentration in Lemmas C.1.1 and C.1.2, since

$$\sum_{j \neq i \in S} x_j^{(2)} = \mathbb{E}[\sum_{j \neq i \in S} x_j] + \sum_{j \neq i \in S} [x_j - \mathbb{E}[x_j] \tag{5.28}$$

$$\lesssim k + \sqrt{k} \log n \tag{5.29}$$

$$\tag{5.30}$$

with high probability. For the second point, let $w = A^*{}_S v$, and note that $\|w\| \leq 2\|v\|$ by the Gergorin circle Theorem. Hence,

$$v^T y = w^T x_S \leq \|w\| \|x_s\| \lesssim \|v\|(\sqrt{k} + \log n) \tag{5.31}$$

$$\tag{5.32}$$

by Lemma B.3.1. □

Appendix C

Concentration and Anti-Concentration

C.1 Concentration of Measure

Definition C.1.1 (Subgaussian Random Variable)**.** We say that a random variable Z is σ^2-sub-Gaussian with variance proxy σ^2 if, for all $t \in \mathbb{R}$, it satisfies the following bound on its generating function:

$$\mathbb{E}[\exp(t(Z - \mathbb{E}(Z)))] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right) \tag{1.1}$$

Proposition C.1.1 (Tail and Moment Bounds for sub-Gaussian Random Variables)**.** *Let Z be a σ^2 sub-Gaussian random variable. Then*

$$\Pr(Z - \mathbb{E}[Z] > \sigma t) \leq \exp(-\frac{t^2}{2}) \quad \textit{and} \quad Pr(Z - \mathbb{E}[Z] < -\sigma t) \leq \exp(-\frac{t^2}{2}) \tag{1.2}$$

Moreover, for any $k \in \mathbb{N}$

$$\mathbb{E}[|Z - \mathbb{E}[Z]|^] \leq \sigma \tag{1.3}$$

In particular, $\mathbb{E}[|Z - \mathbb{E}[Z]|] \leq 2\sigma$

Proof. The first point is a standard result in concentration; see Chapter 1 in Rigollet (2014) for an in depth discussion. The second bound follows from the fact that □

Proof.

$$\mathbb{E}[|Z - \mathbb{E}[Z]|] \;=\; \int_{t=0}^{\infty} t\Pr(|Z - \mathbb{E}[Z]| > t) \;\leq\; 2\int_{t=0}^{\infty} \sigma t e^{-\frac{t^2}{2}} dt \;=\; 2\sigma \tag{1.4}$$

□

The next proposition lists a few additional properties of subgaussian random variables. Again, we direct the curious reader to Chapter 1 in Rigollet (2014) for further details

Proposition C.1.2. *Let $Z_1, \dots, Z_n$ be a collection of independent σ_i^2 random variables. Then*

1. *$\sum_{i=1}^n Z_i$ is a $\sum_i \sigma_i^2$-subgaussian random variable.*
2. *Moreover generally, $\sum_{i=1}^n a_i Z_i$ is a $\sum_i a_i^2 \sigma_i^2$-subgaussian random variable*
3. *$\sup_{1 \le i \le n} Z_i/\sigma_i \le \sqrt{2\log(n/\delta)}$ with probability $1-\delta$*

The next proposition is indispensible when control the deviation of sums of bounded, vector-valued random variables from their mean:

Proposition C.1.3 (Matrix Bernstein Inequality, Theorem 1.6 in Tropp (2012))**.** *Let $Z_1, \dots, Z_p$ be independent $m \times n$ matrices such that $\|Z_i\| \le R$ almost surely and $\max\{\|\mathbb{E}[ZZ^T]\|, \|\mathbb{E}[Z^TZ]\|\} \le \sigma^2$. Then, with probability $1-\delta$, it holds that*

$$\|\frac{1}{p}\sum_{i=1}^p (Z_i - \mathbb{E}[Z_i])\| \lesssim \frac{R}{p}\log((n+m)/\delta) + \sqrt{\frac{\sigma^2 \log((n+m)/\delta)}{p}} \tag{1.5}$$

If Z_i are vectors instead of matrices, it suffices to choose $\sigma^2 \ge \mathbb{E}[\|Z\|^2]$.

As outlined in Section[Gamma Notation], we shall often need to control the expectations of the tails of random variables with super-polynomial decay. To this end, we introduce the following proposition:

Proposition C.1.4 (Tail Expectation Bound)**.** *Suppose that Z is a random variable such that $\Pr(\|Z\| \ge R(\log(1/\delta))^C] \le 1-\delta$ for some constant $c > 0$, where $R \le n^C$ for some $C > 0$. Then,*

1. *Let $p \le n^{O(')}$, and consider p iid copies $Z_1, \dots, Z_p$ of Z. Then, with probability $1 - n^{-\log(n)}$, it holds independently for each $Z_i \lesssim R(\log n)^{2C}$.*
2. *If $\tilde{R} \ge R\log^{2c}(n)$, then*

$$\|\mathbb{E}[Z\mathbb{1}(\|Z\| \ge \tilde{R})]\| \le \mathbb{E}[\|Z\| Ii(\|Z\| \ge \tilde{R})] = n^{-\omega(1)} \tag{1.6}$$

3. *If A is an event that occurs with probability $n^{-\omega}(1)$, then*

$$\|\mathbb{E}[Z\mathbb{1}_A]\| \le \mathbb{E}[\|Z\|\mathbb{1}_A] = n^{-\omega(1)} \tag{1.7}$$

4. *More generally, if X_1 and X_2 are random variables such that $\|X_1\|, \|X_2\| \le n^{C'}$ for some $C' > 0$, then*

$$\|\mathbb{E}[Z \cdot X_1] - \mathbb{E}[Z \cdot X_2]\| \le n^{-\omega(1)} \tag{1.8}$$

Proof. The first and second point are a more precise statement of Lemma 45 in Arora et al. (2015). For the third point, set $\tilde{R} = R\log^{2c}(n)$. To prove the third point, let

$$\begin{aligned}
\|\mathbb{E}[Z\mathbb{1}_A]\| &\le \mathbb{E}[\|Z\|\mathbb{1}_A] && (1.9)\\
&= \mathbb{E}[\|Z\|\mathbb{1}(\|Z\| \le \tilde{R})\mathbb{1}_A] + \mathbb{E}[\|Z\|(\|Z\| \ge \tilde{R})] && (1.10)\\
&\le \tilde{R}\Pr(\mathbb{1}_A) + \mathbb{E}[\|Z\|\mathbb{1}(\|Z\| \ge \tilde{R})] && (1.11)\\
&= n^{-\omega(1)} && (1.12)
\end{aligned}$$

The last point follows by letting A denote the event that $X_1 \neq X_2$, and noting that $\|\mathbb{E}[Z(X_1 - X_2)]\| \leq n^{C'}\|\mathbb{E}[Z\mathbb{1}_A]\| \leq n^{C'-\omega(1)} = n^{-\omega(1)}$. □

Lemma C.1.5 (Truncated Berstein Inequality)**.** *Let $Z_1, \dots, Z_p$ be independent $m \times n$ matrices such that $\|Z_i\|^2 \leq R$ with high probability and and $\max\{\|\mathbb{E}[ZZ^T]\|, \|\mathbb{E}[Z^TZ]\|\} \leq \sigma^2$. Suppose as well that there is a $C > 0$ for which $\Pr(\|Z\| \geq R(\log(1/\delta))^C] \leq 1 - \delta$, where $R = n^{O(1)}$ (say, the entries of Z are $O(1)$-subgaussian). Then with probability $1 - \delta$,*

$$\|\frac{1}{p}\sum_{i=1}^{p}(Z_i - \mathbb{E}[Z_i])\| \lesssim \frac{R}{p}\log((m+n)/\delta) + \sqrt{\frac{\sigma^2\log((m+n)/\delta)}{p}} + n^{\omega(1)} \tag{1.13}$$

If Z_i are vectors instead of matrices, it suffices to choose $\sigma^2 \geq \mathbb{E}[\|Z\|^2]$.

Proof. Write

$$\frac{1}{p}\sum_{i=1}^{p}(Z_i\mathbb{1}(|Z_i| \leq R) - \mathbb{E}[Z_i\mathbb{1}(|Z_i| \leq R)]) + \frac{1}{p}\sum_{i=1}^{p}(\mathbb{E}[Z_i\mathbb{1}(|Z_i| \leq R)] - \mathbb{E}[Z_i]) \tag{1.14}$$

Applying Bernstein's inequality to $\frac{1}{p}\sum_{i=1}^{p}(Z_i\mathbb{1}(|Z_i| \leq R) - \mathbb{E}[Z_i\mathbb{1}(|Z_i| \leq R)])$ gives

$$\|\frac{1}{p}\sum_{i=1}^{p}(Z_i\mathbb{1}(|Z_i| \leq R) - \mathbb{E}[Z_i\mathbb{1}(|Z_i| \leq R)])\| \lesssim \frac{R}{p}\log(1/\delta) + \sqrt{\frac{\sigma^2\log(1/\delta)}{p}} \tag{1.15}$$

with high probability, while Part 3 of Lemma C.1.4 gives $\|\frac{1}{p}\sum_{i=1}^{p}(\mathbb{E}[Z_i\mathbb{1}(|Z_i| \leq R)] - \mathbb{E}[Z_i])\| \leq n^{-\omega(1)}$. □

C.1.1 Additional Bounds

To control the frequence of events, we use the following Proposition:

Proposition C.1.6 (Multiplicative Chernoff Bound for Bernoulli Random Variables)**.** *Let $Z_1, \dots, Z_n$ be independent random variables taking values in $\{0, 1\}$ for which $\mathbb{E}[\sum_i Z_i] = \mu$. Then*

$$\Pr(\sum_i Z_i \geq (1+\delta)\mu) \leq e^{-\delta^2\mu/2} \quad \text{and} \quad \Pr(\sum_i Z_i \leq (1-\delta)\mu) \leq e^{-\delta^2\mu/3} \tag{1.16}$$

Finally, the following lemma controls the norms of random subgaussian vectors:

Lemma C.1.7 (Theorem 1.15 in Rigollet (2014))**.** *Suppose that $Z = (Z_1, \dots, Z_k)$ is a random vector whose entries are iid, have variance $\sigma^2 = \Theta(1)$, and are σ-subgaussian. Then*

$$\|Z\| \leq 4\sigma\sqrt{k} + 2\sigma\sqrt{2\log(1/\delta)} \tag{1.17}$$

The previous lemma controls the quadratic form Z^TIZ. To control general qudratic forms, we have the Hanson-Wright inequality.

Lemma C.1.8 (Hanson-Wright, Theorem 2.1 in Rudelson and Vershynin)**.** *Let A be a fixed $m \times n$ matrix, and let $Z = (Z_1, \dots, Z_n)$ be a random with iid, mean-zero $O(1)$-subgaussian entries. Then with high probability,*

$$\|AZ\|_2 = \tilde{\mathrm{O}}\left(\|A\|_F\right) \tag{1.18}$$

For a more precise statement and a proof, see Theorem 2.1 in Rudelson and Vershynin.Finally, we have the following bound on the difference between random covariance matrices and their expectations

Lemma C.1.9 (Theorem 5.44 in Vershynin (2010))**.** *Let A be an $N \times n$ matrix whose rows are independent random vectors in $\mathbb{R}^n$ with the common moment matrix $\Sigma = \mathbb{E}[A_i A_i^T]$, and let R be such that $\|A_i\|_2 \le R$ almost surely. Then for every $t \ge 0$, the following inequality holds with probability at least $1 - n\exp(-ct^2)$:*

$$\|\frac{1}{N}A^T A - \Sigma\| \le \max\{\} \tag{1.19}$$

where $\delta = t\sqrt{\frac{R}{N}}$.

Using the sample truncation trick as in Lemma C.1.5, we have

Corollary. *Let A be an $N \times n$ matrix whose rows are independent random vectors in $\mathbb{R}^n$ with the common moment matrix $\Sigma = \mathbb{E}[A_i A_i^T]$ and super-polynomial tails (i.e. $O(n)$-subgaussian entries). Moreover, R be such that $\|A_i\|_2 \le R$ with high probability. Then*

$$\|\frac{1}{N}A^T A - \Sigma\| = \tilde{\mathrm{O}}\left(\|\Sigma\|^{1/2}\delta, \delta^2\right) \tag{1.20}$$

where $\delta = t\sqrt{\frac{R}{N}}$.

C.1.2 Incoherence of Random Matrices

Proposition C.1.10. *Suppose that $B \in \mathbb{R}^{n\times m}$ has entries uniform, iid entries which are have variance $\sigma^2 = \Omega(1)$, and are $K^2 = O(1)$-subgaussian. Then, $A := B - \mathbb{E}[B]$ is $\frac{\sigma\log(nm)}{\sqrt{n}}$-incoherent with probability $1 - (mn)^{-\omega(1)}$.*

Proof. First, let's bound $|A_i^T A_j/\|A_j\||$ with high probability whenenver $A_j \neq 0$. To this end, let v be the random variable whose distribution is $A_j/\|A_j\|\|A_j\| \neq 0$. Since $v, A_j \perp A_i$, it holds that and note that

$$\begin{aligned}
\Pr(|A_i^T A_j/\|A_j\| > t\big|\|A_j\| \neq 0) &= \Pr(|A_i^T v|\big|v \neq 0) && (1.21)\\
&= \mathbb{E}_v[\Pr(|A_i^T v| > t\big|v)] && (1.22)\\
&\le \mathbb{E}_v[\exp(-\frac{t^2}{2K^2})] && (1.23)\\
&\le \exp(-\frac{t^2}{2K^2}) && (1.24)
\end{aligned}$$

Letting $t = \sqrt{2}K\sqrt{\log(1/\delta)}$ whenever $A_j \neq 0$ with probability $1 - \delta$.

Now we can wrap up. Using a vector Berstein bound, one can show that $\|A_i\|, \|A_j\| \geq \sqrt{n\sigma/2}$ with probability $n^{-\omega(1)}$, and so in fact $|\cos(A_i, A_j)| \leq \frac{2K\sqrt{\log(1/\delta)}}{\sigma\sqrt{n}}$ with probability $1-\delta-n^{-\omega(1)}$. Setting $\delta = (mn)^{-\log(mn)}$ and taking a union bound over all pairs of columns concludes the proof. □

Remark. By sacrificing log factors in the incoherence, we place weaker assumptions on the tails of the entries of B. Indeed, we only really require that there is some $R = O(1)$ and $C = O(1)$ for which $\Pr(B_{i,j} \geq R\log(1/\delta)^C) \leq \delta$.

C.2 Anti-Concetration of Measure

In what follows, we show how the ρ-smoothness anti-concentration assumption can be turned into an "non-correlation" result about quantities which depend on certain ρ-smooth variables more heavily than others. In what follows, let $Z_1, \ldots, Z_n$ be independent real valued random variables for which there exists a constant $\rho > 0$, such that for any set $S \subset \mathbb{R}$, $\Pr(Z_i \in S) \leq \rho \cdot \text{vol}(S)$ for all $i \in [n]$. We will also fix a vector $a \in \mathbb{R}^n$, and define the random variables $Y_i = a_i Z_i$.

Here, the notation $\text{vol}(S)$ denotes the Lebesque measure of the set S. For any $u \in \mathbb{R}^n$, we will use the notation $S - u = \{w \in \mathbb{R} : w + u \in S\}$ and $uS = \{w \cdot u : w \in S\}$. Note that $\text{vol}(S) = \text{vol}(S - u)$ for all $u \in \mathbb{R}^n$, and $\text{vol}(uS) = u\text{vol}(S)$. These insights give the following Lemma

Lemma C.2.1. *Let $\{Z_i\}_{1\leq i\leq n}$ be independent,real random variables. Then*

1. *If* any *Z_i is ρ_i-smoothly distributed, then $\sum_i Z_i$ is ρ_i-smoothly distributed for all $i \in [n]$.*
2. *If each Z_i is ρ-smoothly distributed, then each $a_i Z_i$ is $\rho_i/|a_i|$ smoothly distributed.*
3. *If each Z_i is ρ-smoothly distributed, then $\sum_{i=1}^n a_i Z_i$ is $\rho_i/|a_i|$ for any $i \in [m]$*

Proof. For the first point, it suffices to prove that $\sum_i Z_i$ is ρ_1-smoothly distributed:

$$\Pr(\sum_{i=1}^n Z_i \in S) \quad = \quad \Pr(Z_1 \in S - \sum_{i\geq 2} Z_i) \leq \rho_1 \text{vol}(S - \sum_{i\geq 2} Z_i)) \leq \rho_1 \text{vol}(S)$$

For the second point,

$$\Pr(a_i Z_i \in S) = \Pr(Z_i \in \frac{1}{a_i} \cdot S) \leq \rho \text{vol}(\frac{1}{a_i} \cdot S) = \frac{\rho}{|a_i|}$$

The third point follows by combining the two. □

As a consequence, we have our first non-correlation result about the difference between the sign-threshold of the sum of two ρ-smooth random variables $Y_1 + Y_2$, and the sign-threshold of only Y_1 alone:

Proposition C.2.2. *Suppose that Y_1, Y_2 are two, independent real valued random variables and that there is a constant ρ such that $\Pr(Y_1 \in S) \leq \rho vol(S)$ for all Borel sets $S \subset \mathbb{R}^n$. Then,*

$$\mathbb{E}\left[|\text{sign}_\tau(Y_1) - \text{sign}_\tau(Y_1 + Y_2)| \,\middle|\, Y_2\right] \leq 4\rho|Y_2| \tag{2.25}$$

where the factor of 4 can be replaced with 2 if $|Y_1| \leq 2\tau$ almost surely. Similarly,

$$\mathbb{E}\left[|\text{thres}_\tau(Y_1) - \text{thres}_\tau(Y_1 + Y_2)| \,\middle|\, Y_2\right] \leq 2\rho|Y_2| \tag{2.26}$$

Proof. We prove the first point; the second follows similarly. The key idea is to define the sets

$$U_\tau^+(t) := [-\tau - |t|, \tau] \cup [\tau - |t|, \tau] \quad \text{and} \quad U_\tau^-(t) := [-\tau, -\tau + |t|] \cup [\tau, \tau + |t|] \tag{2.27}$$

We now state the two crucial, though rather self evident properties of U_τ^+ and $U^\tau -$

1. $\text{vol}(U_\tau^+(t)) = \text{vol}(U_\tau^-(t)) = 2|t|$
2. Given two random variables Y_1, Y_2, it holds that $\text{sign}_\tau(Y_1 + Y_2) \neq \text{sign}_\tau(Y_1)$ only when $Y_2 > 0 \wedge Y_1 \in U_\tau^+(Y_2)$ or $Y_2 < 0 \wedge Y_1 \in U_\tau^-(Y_2)$.

Stated otherwise, $\text{sign}_\tau(Y_1+Y_2) \neq \text{sign}_\tau(Y_1)$ only when Y_1 lies in a set of volume proportional to Y_2. Indeed, since $|\text{sign}_\tau(Y_1) - \text{sign}_\tau(Y_1 + Y_2)| \leq 2$ (less than 1 if $Y_1 \leq 2\tau$ almost surely), we have that

$$|\text{sign}_\tau(Y_1) - \text{sign}_\tau(Y_1 + Y_2)| \quad \leq \quad 2\mathbb{1}\left(Y_2 > 0 \vee Y_1 \in U_\tau^+(Y_2) \middle| Y_2\right) + 2\mathbb{1}\left(Y_2 < 0 \vee Y_1 \in U_\tau^-(Y_2) \middle| Y_2\right)$$

Thus

$$\begin{aligned}
&\mathbb{E}\left[|\text{sign}_\tau(Y_1) - \text{sign}_\tau(Y_1 + Y_2)| \,\middle|\, Y_2\right] \\
&\leq 2\Pr\left(Y_2 > 0 \vee Y_1 \in U_\tau^+(Y_2) \middle| Y_2\right) + 2\Pr\left(Y_2 < 0 \vee Y_1 \in U_\tau^-(Y_2) \middle| Y_2\right) \\
&= 2\mathbb{1}\left(Y_2 > 0\right)\Pr\left(Y_1 \in U^+(Y_2) \middle| Y_2\right) + 2\mathbb{1}\left(Y_2 < 0\right)\Pr\left(Y_1 \in U^-(Y_2) \middle| Y_2\right) \\
&\leq 2\rho\left\{\mathbb{1}(Y_2 > 0)\text{vol}(U^+(Y_2)) + \mathbb{1}(Y_2 < 0)\text{vol}(U^-(Y_2))\right\} \\
&= 4\rho|Y_2|
\end{aligned} \tag{2.28}$$

If $Y_1 \leq 2\tau$ almost surely, then $|\text{sign}_\tau(Y_1) - \text{sign}_\tau(Y_1 + Y_2)| \leq 1$ almost surely, which would yield a factor of 4 instead of 2 in the last line of Equation 2.28 □

We can state the previous proposition in a slightly easier-to-use form:

Corollary. *Let $Z_1, \ldots, Z_k$ be real random variables such that $Z_1 \perp (Z_2, \ldots, Z_k)$, and Z_1 is ρ-smoothly distributed. Then, for any measurable function $f(\cdot)$ and any $\tau \in \mathbb{R}$, and any vector $a \in \mathbb{R}^k$, it holds that*

$$\mathbb{E}\left[|f(Z_1)|\left|\text{sign}_\tau(\sum_{i\neq 2}^{n} a_r Z_r) - \text{sign}_\tau(\sum_{i=1}^{k} a_i Z_i)\right|\right] \leq 4\rho\left|\frac{a_2}{a_1}\right|\mathbb{E}\left[|Z_2 f(Z_2)|\right] \tag{2.29}$$

If in addition $a_1 Z_1 \leq 2\tau$ almost surely, or if the sign_τ functions are replaced with thres_τ, then the factor of 4 can be replaced by a factor of 2 in the above display.

Proof. Let v be the random vector $(Z_3, \dots, Z_k)$. Now, let $Y_1 = a_1 Z_1 + \sum_{i>2} a_i Z_i$ and $Y_2 = a_2 Z_2$. Note that, conditioned on v, $Y_1 = a_1 Z_1 + c(v)$, where $c(v) = \sum_{i>2} a_i Z_i$ is constant. Hence Y_1 is $a_1\rho$-smoothly distributed by Lemma Lemma C.2.1, and $Y_1, Y_2 \big| v$. Hence,

$$\begin{aligned}\mathbb{E}\left[|\mathrm{sign}_\tau(Y_1) - \mathrm{sign}_\tau(Y_1 + Y_2)| \,\middle|\, Z_2, Z_3, \dots, Z_k\right] &= \mathbb{E}\left[|\mathrm{sign}_\tau(Y_1) - \mathrm{sign}_\tau(Y_1 + Y_2)| \,\middle|\, v\right] \\ &\leq 4\rho|Y_1|/|a_1| = 4\rho|a_2/a_1||Z_2|\big|Z_2, v =\end{aligned}$$

Hence, taking the expectation over $Z_2, \dots, Z_k$ gives

$$\begin{aligned}&\mathbb{E}\left[f(Y_2)\,|\mathrm{sign}_\tau(Y_1) - \mathrm{sign}_\tau(Y_1 + Y_2)|\right] \\ =\ &\mathbb{E}_{Z_2,\dots,Z_k}\left[|\mathrm{sign}_\tau(Y_1) - \mathrm{sign}_\tau(Y_1 + Y_2)| \,\middle|\, Z_2, Z_3, \dots, Z_k\right] \\ \leq\ &\mathbb{E}_{Z_2,\dots,Z_k}\left[4\rho|a_2/a_1||Z_2|\right] \\ =\ &4\rho|a_2/a_1|f(Z_2)|Z_2|\end{aligned}$$

□

We conclude this section with a self-evident proposition which controls the expectation of ρ-smooth quanties over sets of small volumes:

Proposition C.2.3 (Expectation of ρ-Smooth quantities over sets of small volume)**.** *Suppose that Z is (C, ρ)-smooth and that $\tau \leq C$. Then,*

$$|\mathbb{E}[f(Z)\mathbb{1}(Z \leq \tau)]| \leq 2\tau\rho \max_{t\in[-\tau,\tau]} |f(Z)| \tag{2.30}$$

In particular,

$$\mathbb{E}[|Z|^k \mathbb{1}(Z \leq \tau)] \leq 2\rho\tau^{k+1} \tag{2.31}$$

Appendix D

Linear Algebra

D.1 Projections onto Span of Incoherent Vectors

In what follows, let $A \in \mathbb{R}^{m \times n}$ be a μ-incoherent matrix with unit norm columns. For a subset $S \subset [m]$, let A_S denote the matrix whose columns are the columns of A indexed by the elements of S. Let Q_S denote the projection onto $\mathcal{V}_S$, and let $P_S = I - Q_S$ denote the projection on $\mathcal{V}_S^{\perp}$. The following Lemma shows that, if the columns of A are sufficiently incoherent, then $A_S A_S^T$ looks like an orthogonal projection:

Lemma D.1.1. *If $|S| = k$, then*

$$\left(1 - \frac{\mu(k-1)}{\sqrt{n}}\right) Q_S \preceq A_S A_S^T \preceq \left(1 + \frac{\mu(k-1)}{\sqrt{n}}\right) Q_S \tag{1.1}$$

and

$$\left\| A_S^{\dagger} \right\|^2 = \left\| \left(A_S^T A_S\right)^{\dagger} \right\| \leq \left(1 - \frac{\mu(|S|-1)}{\sqrt{n}}\right)^{-1} \tag{1.2}$$

More generally, if $\sigma_{\min}(A_S) \geq c$, then $Q_S \preceq c^{-2} A_S A_S^T$.

Proof. Since $|S| = k$, we may write $A_S A_S^T = U \Sigma U^T$ where $\Sigma \in \mathbb{R}^{k \times k}$ is diagonal and non-negative, and $U \in \mathbb{R}^{n \times k}$ is orthogonal. Then $(\min_i \Sigma_{ii}) UU^T \preceq A_S A_S^T \preceq (\max_i \Sigma_{ii}) UU^T$. This proves the more general result when $\|\sigma_{\min}(A_S)\| \geq c$.

For the specific case, we have that the same column space and U is orthogonal, it follows that $UU^T = Q_S$. On the other hand, $\|A_S^{\dagger}\|^2 = \|(A_S^T A_S)^{\dagger}\| = (\min_i \Sigma_{ii})^{-1}$. Hence, both Equation 1.1 and Eqution 1.2 will follow once we show that the entries of Σ, or equivalently, the squares of the top k singular values of A_S, lie in the interval

$$\left[1 - \frac{\mu(k-1)}{\sqrt{n}}, 1 + \frac{\mu(k-1)}{\sqrt{n}}\right] \tag{1.3}$$

It suffices to show that the eigenvalues of $M := A_S^T A_S$ lie in the above interval. To this end, we apply the Gershgorin circle theorem to M. $M_{ii} = \langle A_i, A_i \rangle = 1$, let $|M_{i,j}| = |\langle A_i, A_j \rangle| \leq \mu/\sqrt{n}$, and thus $\sum_j |M_{i,j}| \leq (k-1)\mu/\sqrt{n}$. By the Gershgorin Circle Theorem, it holds that the eigenvalues of $A_S^T A_S$ lie in complex disks of radius no more than $\frac{(k-1)\mu}{\sqrt{n}}$ around $\|A_i\|^2$ (which is 1 in the simple case, and lie in $[c, C]$ in the general case]. But $A_S^T A_S$ is self adjoint, so its eigenvalues are real hence each lies in the interval given by Equation 1.3 $\square$

Remark. The incoherence assumption in Lemma D.1.1 is absolutely indispensible. Indeed, consider the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \tag{1.4}$$

for some $\epsilon > 0$. Then A is nonsingular, and consequently Proj_A is just the identity. However, one can check that top singular value of A is at least $\sqrt{2}$, while the bottom singular value of A is no more than ϵ. This has an algorithmic implication: unless if a dictionary A is an estimate of a true dictionary A^*, then it is more numerically stable to use AA^T over $Proj_A$ in projection pursuit algorithms.

Next, we establish some elementary properties of the projection matrices P_S and Q_S

Lemma D.1.2. *The following hold:*

1. *If $V \subset S$, then P_S commutes with P_V and Q_S commutes with Q_V*
2. *If $V \subset S$, then $P_S P_V = P_V P_S = P_S$*
3. *$Q_V - Q_S = P_V(Q_V - Q_S)P_V$*
4. *$A_V^T P_V = 0$ and $Q_V P_V = 0$*

Proof. For the first point, $V \subset S \implies \mathcal{V}_V \subset \mathcal{V}_S$, and $\mathcal{V}_S^\perp \subset \mathcal{V}_V^\perp$. The result now follows form that fact that if $\mathcal{V} \subset \mathcal{W}$ are two subspaces, then $\mathrm{Proj}_{\mathcal{V}}$ and $\mathrm{Proj}_{\mathcal{W}}$ commute. The second point follows from the fact that if $\mathcal{V} \subset \mathcal{W}$ are two subspaces, then $\mathrm{Proj}_{\mathcal{V}}\mathrm{Proj}_{\mathcal{W}} = \mathrm{Proj}_{\mathcal{V}}$, and the commutativity established in the first point. From the first result, we may write $P_S - P_V = Q_V - Q_S$, while also, $P_S - P_V = P_V(P_S - P_V)P_V = P_V(Q_V - Q_S)P_V$ so that $Q_V - Q_S = P_V(Q_V - Q_S)P_V$. To show that $A_V^T P_V = 0$, note that for any $x \in \mathbb{R}^n$, $P_v x$ is perpendicular to all of the columns of A_V, and hence $A_V^T P_V x = \sum_{j \in V} \langle A_j, P_v x \rangle = 0$. $Q_V P_V = Q_V^T P_V = 0$ for precisely the same reason. □

We are now ready to prove Theorem 4.1.5, which we restate here:

Theorem D.1.3 (Restatement of Theorem 4.1.5)**.** *If $\sigma_{\min}(A_S) \geq c$ then*

$$Q_S - Q_V \preceq c^{-2} P_V (A_{S-V} A_{S-V}^T) P_V \tag{1.5}$$

as long as $c > 0$.

Proof of 4.1.5. Using Lemma D.1.2 parts 3 and 4, Lemma D.1.1, and then Lemma D.1.2 part 4 again, we have

$$\begin{aligned} Q_S - Q_V &= P_V(Q_S - Q_V)P_V \\ &= P_V Q_S P_V \\ &\preceq K \cdot P_V(A_S A_S^T)P_V \text{ by Lemma D.1.1} \\ &= K \cdot P_V(A_{S-V}A_{S-V}^T + A_V A_V^T)P_V \\ &= K \cdot P_V(A_{S-V}A_{S-V}^T)P_V \end{aligned} \tag{1.6}$$

□

D.2 General Purpose Bounds

Claim D.2.1. *Let $v_1, \dots, v_k \in \mathbb{R}^n$ be fixed vectors of norm no more than δ, and let $u \in \mathbb{R}^k$ be a random vector . Then $\mathbb{E}[\|\sum_i u_i v_i\|^2] \le \delta^2 \|\mathbb{E}[uu^T]\|_{l_1}$, where $\|\cdot\|_{l_1}$ is the l_1 norm of the matrix viewed as a vector in $\mathbb{R}^{k^2}$.*

Proof. For a matrix $S \in \mathbb{R}^{k\times k}$, let $\|\cdot\|_{l_\infty}$ denote its l_∞-norm viewed as a vector in $\mathbb{R}^{k^2}$. By Holders inequality $\sup_{S:\|S\|_{l_\infty}\le 1} \sum_{i,j} S_{i,j}U_{i,j} = \|U\|_{l_1}$ for any $U \in \mathbb{R}^{k\times k}$.Now, write $\mathbb{E}[\|\sum_i u_i v_i\|^2] = \mathbb{E}[\langle \sum_i u_i v_i, \sum_j u_j v_j\rangle] = \mathbb{E}[\sum_{i,j} u_i u_j \langle v_i, v_j\rangle] = \sum_{i,j} \mathbb{E}[u_i u_j]\langle v_i, v_j\rangle$. Since $|\langle v_i, v_j\rangle| \le \|v_i\|\|v_j\| \le \delta^2$, we have

$$\begin{aligned}
\mathbb{E}[\|\sum_i u_i v_i\|^2] &\le \sup_{s_{i,j}\in[-\delta^2,\delta^2]} \sum_{i,j} s_{i,j}\mathbb{E}[u_i u_j] \\
&= \delta^2 \sup_{S\in\mathbb{R}^{k\times k}:\|S\|_{l_\infty}\le 1} \sum_{i,j} S_{i,j}\mathbb{E}[u_i u_j] \\
&= \delta^2 \|\mathbb{E}[uu^T]\|_{l_1}
\end{aligned}$$

□

Lemma D.2.2. *Let M, N be two symmetric random matrices such that $N \succeq 0$ and $-N \preceq M \preceq N$ almost surely. Then, for any unit vector v and random variables Z, Z' such that $Z' \ge |Z|$ almost surely, it holds that*

$$\|\mathbb{E}[Z \cdot Mv]\| \le \|\mathbb{E}[Z \cdot Mv]\| \le \left\|\mathbb{E}\left[Z' \cdot N\right]\right\| \tag{2.7}$$

More generally, if M, N, Z', Z have exponential tails and $-N \preceq M \preceq N$ with probability $1-\gamma$, then

$$\|\mathbb{E}[Z \cdot Mv]\| \le \|\mathbb{E}[Z \cdot Mv]\| \le \left\|\mathbb{E}\left[Z' \cdot N\right]\right\| \tag{2.8}$$

Proof. We prove the case when $-N \preceq M \preceq N$; the general case follows from an application of Lemma C.1.4:

$$\begin{aligned}
\|\mathbb{E}[Z \cdot Mv]\| &\le \sup_{w\in\S^{d-1}} \|\mathbb{E}[Z \cdot Mw]\| && (2.9)\\
&\le \sup_{w,w_1\in\S^{d-1}} w_1^T \mathbb{E}[Z \cdot M]\, w && (2.10)\\
&\le \sup_{w\in\S^{d-1}} \left|w^T \mathbb{E}[Z \cdot M]\, w\right| && (2.11)\\
&= \|\mathbb{E}[Z \cdot M]\| && (2.12)
\end{aligned}$$

where the last step follows since $\mathbb{E}[Z \cdot M]$ is symmetric. Let w^* be the vector which attains the supremum in the above equation. By our assumptions on N and M, it holds that

$$-|Z|N \preceq ZM \preceq |Z|N \tag{2.13}$$

Hence, $(w^*)^T\mathbb{E}[Z\cdot M]w^* = \mathbb{E}[Z(w^*)^T M(w^*)]$ lies in the closed interval between $\mathbb{E}[-|Z|(w^*)^T N w^*] = -\mathbb{E}[|Z|(w^*)^T N w^*]$ and $\mathbb{E}[|Z|(w^*)^T N w^*]$, and so, for any $\zeta \ge |Z|$ almost surely:

$$\begin{aligned}
\|\mathbb{E}[Z \cdot Mv]\| &\le \left|(w^*)^T\mathbb{E}[Z \cdot M](w^*)\right| \\
&\le \left|\mathbb{E}[|Z|(w^*)^T N w^*]\right| = \mathbb{E}[|Z|(w^*)^T N w^*] \\
&\le \mathbb{E}[Z'(w^*)^T N w^*] = (w^*)^T\mathbb{E}[Z'N]w^* \le \|\mathbb{E}[Z'N]\|
\end{aligned}$$

□

Lemma D.2.3. *Let D_1, D_2 be diagonal matrices where D_1 has nonnegative entries and the entries of D_2 are uniformly bounded by d. Then, for any matrix X, $XD_1D_2X^T \preceq dX_1D_1X^T$.*

Lemma D.2.4. *Let v and w be orthgonal vectors. Then $\|vw^T + wv^T\| = \|v\|\|w\|$*

Proof. Because $vw^T + wv^T$ is symmetric, we have

$$\|vw^T + wv^T\| = \sup_{a:\|a\|=1} a^T(vw^T + wv^T)a \tag{2.14}$$

Let $a = a_v + a_w + a_0$ be a decomposition of a into the span of v, w, and the orthogonal complement of the span of v and w respectivly. Then $a^T(vw^T + wv^T)a = \|v\|\|w\|(a_va_w + a_wa_v) = 2\|v\|\|w\|(a_va_w)$. It is easy to see $a_v = a_w = 1/\sqrt{2}$ maximizes this expression under the $\|a\| = 1$ constraint. □