Provable Dictionary Learning via Column Signatures

Sanjeev Arora* Aditya Bhaskara† Rong Ge‡ Tengyu Ma§

April 17, 2014

Abstract

In *dictionary learning*, also known as *sparse coding*, we are given samples of the form $y = Ax$ where $x \in \mathbb{R}^m$ is an unknown random sparse vector and A is an unknown dictionary matrix in $\mathbb{R}^{n\times m}$ (usually $m > n$, which is the *overcomplete* case). The goal is to learn A and x. This problem has been studied in neuroscience, machine learning, vision, and image processing. In practice it is solved by heuristic algorithms and provable algorithms seemed hard to find. Recently, provable algorithms were found that work if the unknown feature vector x is $\sqrt{n}$-sparse or even sparser. [SWW12] did this for dictionaries where $m = n$; [AGM13] gave an algorithm for overcomplete ($m > n$) and incoherent matrices A; and [AAN13] handled a similar case but with somewhat weaker guarantees.

This raised the problem of designing provable algorithms that allow sparsity $\gg \sqrt{n}$ in the hidden vector x. The current paper designs algorithms that allow sparsity up to $n/poly(\log n)$. It works for a class of matrices where features are *individually recoverable*, a new notion identified in this paper that may motivate further work. The algorithm runs in quasipolynomial time because it uses limited enumeration.

*Princeton University, Computer Science Department and Center for Computational Intractability. Email: arora@cs.princeton.edu. This work is supported by the NSF grants CCF-0832797, CCF-1117309, CCF-1302518, DMS-1317308, and Simons Investigator Grant.

†Google Research NYC. Email: bhaskaraaditya@google.com. Part of this work was done while the author was a Postdoc at EPFL, Switzerland.

‡Microsoft Research. Email: rongge@microsoft.com. Part of this work was done while the author was a graduate student at Princeton University and was supported in part by NSF grants CCF-0832797, CCF-1117309, CCF-1302518, DMS-1317308, and Simons Investigator Grant.

§Princeton University, Computer Science Department and Center for Computational Intractability. Email: tengyu@cs.princeton.edu. This work is supported by the NSF grants CCF-0832797, CCF-1117309, CCF-1302518, DMS-1317308, and Simons Investigator Grant.

1 Introduction

Dictionary learning, also known as *sparse coding* in neuroscience, tries to understand the structure of observed samples y by representing them as sparse linear combinations of "dictionary" elements. More precisely, there is an unknown dictionary matrix $A \in \mathbb{R}^{n\times m}$ (usually $m > n$, which is the *overcomplete* case), and the algorithm is given samples $y = Ax$ where x is an unknown sparse vector. (We say a vector is *k-sparse* if it has at most k nonzero coordinates.) The goal is to learn A and x. Such sparse representation was first studied in neuroscience, where Olshausen and Field [OF97] suggested that dictionaries fitted to real-life images have similar properties as the receptive fields of neurons in the first layer of visual cortex. Inspired by this neural analog, dictionary learning is widely used in machine learning for *feature selection* ([AEP06]). More recently the idea of sparse coding has also influenced deep learning ([BC$^+$07]). In image processing, learned dictionaries have been successfully applied to image denoising ([EA06]), edge detection ([MLB$^+$08]) and super-resolution ([YWHM08]). For exposition purposes we refer to the coordinates of the hidden vector x as *features*, and those of the visible vector $y = Ax$ as *pixels*, even though the discussion below applies more broadly than computer vision.

Provable guarantees for dictionary learning have seemed difficult because the obvious mathematical programming formulation is nonconvex: both A and the x's are unknown. Even when the dictionary A is known, it is in general NP-hard to get the sparse combination x given *worst-case* y ([DMA97]). This problem of decoding x given Ax with full knowledge of A is called *sparse recovery* or *sparse regression*, and is closely related to *compressed sensing*. For dictionaries with special structure, sparse recovery was shown to be tractable even on worst-case y; eg this was shown for *incoherent* matrices by [DH01], who required x to be $\sqrt{n}$-sparse. Then Candes et al. [CRT06] showed how to do sparse recovery even when the sparsity is $\Omega(n)$, assuming A satisfies the *restricted isometry property* (RIP) (which random matrices do).

But dictionary learning itself (recovering A given samples y) has proved much harder and heuristic algorithms are widely used. The method in [LS00] was the first, followed by the method of optimal directions (MOD) ([EAHH99]) and K-SVD ([AEB06]). See [Aha06] for more references. However, until recently no algorithms were known that provably recover the correct dictionary. Recently Spielman et al. [SWW12] gave such an algorithm for the full rank case (i.e., $m = n$) and the unknown feature vector x is $\sqrt{n}$-sparse. However, in practice overcomplete dictionaries ($m > n$) are preferred. Arora et al. [AGM13] gave the first provable learning algorithm for overcomplete dictionaries that runs in polynomial-time; they required x to be $n^{1/2-\epsilon}$-sparse (roughly speaking) and A to be incoherent. Independently, Agarwal et al. [AAN13] gave a weaker algorithm that also assumes A is incoherent and allows x to be $n^{1/4}$-sparse. All three papers are inherently unable to handle sparsity more than $\sqrt{n}$: they require two random x, x' to intersect in no more than $O(1)$ coordinates with high probability, which is false whp when sparsity $\gg \sqrt{n}$. Since sparse recovery (where A is known) is possible even up to sparsity $\Omega(n)$, this raised the question whether dictionary learning is possible in that regime. In this paper we will refer to feature vectors with sparsity $n/poly(\log n)$ as *slightly-sparse*. The recent paper on learning deep neural networks ([ABGM13], Section 7) shows how to solve dictionary learning in the slightly-sparse regime for dictionaries which are adjacency matrices of random weighted sparse bipartite graphs.

Our result in ([ABGM13]) may seem natural enough since dictionaries corresponding to weighted sparse random graphs were earlier shown to allow compressed sensing [Ind08, JXHC09, BGI$^+$08]). But one should beware of defining the goals of dictionary learning –where x is random and the unknown A corresponds to naturally-occuring features– by analogy to *compressed sensing*, where x is worst-case and A is synthesized by the algorithm designer. Thus the natural question is whether dictionary learning is possible in the slightly sparse case for classes of dictionaries other than sparse random graphs (which seems a strong assumption about "nature").

In this work, we give quasipolynomial-time algorithms (i.e., $n^{poly(\log n)}$ time) for the slightly-sparse case for an interesting class of dictionaries where features are **individually recoverable**, a new notion we now introduce.

Some of our discussion below refers to nonnegative dictionary learning, which constrains matrices A and hidden vector x to have nonnegative entries. This is a popular variant proposed by Hoyer [Hoy02], motivated again partly by the neural analogy. Algorithms like NN-K-SVD ([AEB05]) were then applied to image classification tasks. This version is also related to *nonnegative matrix factorization* ([LS99]), which has been observed to lead to factorizations that are usually sparser and more local than traditional methods like SVD.

1.1 Interesting cases for dictionary learning

Individual recoverability of features means, roughly speaking, that to an observer who knows the dictionary, the presence of a particular feature should not be confusable with the effects produced by the usual distribution of other features: x_iA_i should not be confusable with $\sum_{j\neq i} x_jA_j$. In fact, should even be able to detect the presence of the ith feature by looking only at the pixels that are nonzero in A_i. All the above-mentioned works on dictionary learning involve dictionaries with this property; (see Appendix A for details) but RIP matrices don't necessarily.

A precise analysis of the above intuitive notion seems difficult (though we hope this will stimulate further work analogous to Dasgupta's formalization of separability for gaussian mixtures [Das99]). Instead we identify other natural definitions that imply it, i.e. are stronger. Roughly speaking, these require that each column has *many* significant entries, and that most pairs of columns do not "overlap much" (though we do not require this for *every* pair, as we will see).

1.2 Our assumptions

Let us start with some basic notation. The dictionary matrix is denoted $A \in \mathbb{R}^{n\times m}$; its jth column is denoted by A_j, the ith row by $A^{(i)}$, and the i,j entry of A by $A_j^{(i)}$. We are given N i.i.d samples $y^1, \ldots, y^N$ generated as $y^i = Ax^i$, where each $x^i \in \mathbb{R}^m$ is chosen from the same distribution. Coordinates of x are called *features* and those of y are *pixels*.

Assumptions about x In most of the paper, the unknown vector is assumed for ease of exposition to be drawn iid as a Bernoulli variable with probability ρ of being 1. Our proofs actually only require the coordinates to be pairwise independent, and $\sum_i w_ix_i$ (for any fixed weights w_i's) to have have tails that drop fast (a la Bernstein bounds). Nonzero entries of x can also be in $[1, c]$ instead of being exactly 1, for some constant c which will then come up in the running time. We can also handle the case in which x_i can be 1 with unequal (and unknown) probability, as long as it is within a constant of ρ.

Let G_τ denote the support of the entries in A that are at least τ in magnitude. We think of it as a bipartite graph with features on one side and pixels on the other.

Motivation for many large entries. A key assumption for us is that the matrix A has many *significant* entries. The rough motivation for this is as follows: suppose we define two dictionaries

A and $\hat{A}$ to be "ϵ-equivalent", if for a random vector x drawn from our distribution, Ax and $\hat{A}x$ are entry-wise ϵ-close with high probability. Then from a practical perspective, it does not matter if we recovered A or $\hat{A}$ (if ϵ is small enough). To this end, we note that the entries of A that are $< \epsilon^2/\log n$ are not "interesting". I.e., we can see (by easy Chernoff bounds) that their effect on each pixel has deviation lower than $\epsilon/2$ w.h.p., thus for purposes of an ϵ-approximation such entries can be zeroed out and their effect mimicked by adding a suitable constant to each entry in the corresponding rows!

Conditions for Nonnegative case.

By simple scaling one can assume that the expected value of each pixel is 1. Assume σ is a constant > 0. We now present our conditions formally. The parameters are fixed later in Section 2.

Assumption 1: (Every feature has significant effect on pixels) Each column of A has at least d entries of magnitude $\geq \sigma$. I.e., the degree of each feature in G_σ is $\geq d$.

Assumption 2': (Low pairwise intersections among features) In G_τ (for a τ which will be $< \sigma$), the intersection of the neighborhoods of any two features is less than κ, where $\tau = O(1/\log n)$ and $\kappa = O(d/\log^2 n)$, as explained below.

In the latter assumption, note that the intersection we allow is much larger than that in a random graph of degree d (in that case the intersection is d^2/n, which is much smaller than κ). This assumption still seems strong for real life, where one may wish to allow feature pairs to have constant fraction overlap. Our algorithms work with a weaker assumption that does allow such overlaps, provided it does not happen for too many pairs of features.

Assumption 2: In G_τ (for a τ which will be $< \sigma$) the neighborhood of each feature (that is, $\{i \in [n] : A_j^{(i)} \geq \tau\}$) has intersection up to $d/10$ (with total weight $< d\sigma/10$) with each of at most $o(1/\sqrt{\rho})$ other features, and intersection at most κ with the neighborhood of each remaining features. Here τ is $O(1/\log n)$ and $\kappa = O(d/\log^2 n)$.

For ease of exposition, our proofs will work with Assumption 2', and at the end of Section 3, we will outline how to work with Assumption 2 instead.

Remark: Assumption 2 is *essentially* the best possible. For instance, if we allow $poly(1/\rho)$ features to intersect the neighborhood of feature j using edges of total weight $\Omega(\ell_1\text{-norm of } A_j)$ then feature j is could no longer be individually recoverable: its effect can be duplicated w.h.p. by combinations of these other features. But a more precise characterization of individual recoverability would be nice, as well as a matching algorithm.

Conditions for case with positive and negative entries

Now the natural normalization is the one that makes the variances of y_i's equal to 1. We assume that the magnitude of edge weights are at most Λ, and that features do not overlap a lot as before. We need one additional assumption to bound the variance contributed by the small entries. Formally, the assumptions are:

Assumption G1: The degree in G_σ of every feature is larger than $2d$.

Assumption G2': In G_τ, (for a $\tau < \sigma$), the intersection of the neighborhoods of any two features j, k is less than κ,where $\tau = O_\theta(1/\log n)$ and $\kappa = O_\theta(d/\log^2 n)$.

Assumption G3: (small entries of A do not cause large deviations) $\rho||A^{(i)}_{\le\tau}||_2^2 \le \gamma$, where $A^{(i)}_{\le\delta}$ denotes the vector consisting of the entries of $A^{(i)}$ that are at most δ, and $\gamma = \sigma^4/2\Delta\Lambda^2\log n$ where Δ is a large enough constant.

Note that Assumption G1 differs from Assumption 1 by a constant factor 2 just to simplify some notation later. Assumption G2' is the same as before.

The assumption G3 intuitively says that for each $y_i = \sum_k A^{(i)}_k x_k$, the smaller $A^{(i)}_k$'s should not contribute too much to the variance of y_i. This is automatically satisfied for nonnegative dictionaries in our setting. Notice that this assumption is talking about rows of matrix A (corresponding to pixels), whereas the earlier assumptions talk about columns of A (corresponding to features).

Comparing with incoherence. Our combinatorial assumptions, though similar in spirit, are not directly related to incoherence. On the one hand, we require the column entries to be "chunky", i.e., have reasonably many significant entries (as motivated below), a seemingly strong assumption. On the other hand, our requirement on the dot products is quite weak. We only require, roughly, that $|\langle A_i, A_j\rangle|/\|A_i\|\|A_j\| < 1/\log^2 n$,[1] which is much weaker than $n^{-1/2+\epsilon}$ required in typical results that solely assume incoherence.

2 Main Results

We will think of $\sigma \le 1$ as a small constant as in assumption 1. The largest magnitude of edge weights is $\Lambda \ge 1$, a constant, and Δ will be a sufficiently large constant that controls the error guarantee and the other parameters. For convenience, let $\theta = (\sigma, \Lambda, \Delta)$. We use the notation $O_\theta(\cdot)$ to hide the dependencies of σ, Λ, Δ. Also, we think of ρ as $< 1/\text{poly}(\log n)$, and d being $\ll n$. The normalization assumption (essentially) implies that $md\rho \in [n/\Lambda, n/\tau]$.

Precisely, for our algorithms to work, we need $d \ge \Delta\Lambda\log^2 n/\sigma^2$, $\tau = O(\sigma^4/\Delta\Lambda^2\log n) = O_\theta(1/\log n)$, $\kappa = O(\sigma^8 d/\log^2 n\Delta^2\Lambda^6) = O_\theta(d/\log^2 n)$ (recall Δ is a large constant), and the density $\rho = o(\sigma^5/\Lambda^{6.5}\log^{2.5} n) = o_\theta(1/\log^{2.5} n)$. The main theorem is now the following:

Theorem 1 (Non-negative Case). *Under Assumptions 1 and 2, when $\rho = o_\theta(1/\log^{2.5} n)$, Algorithm 2 runs in $n^{O_\theta(\log^2 n)}$ time, uses poly(n) samples and outputs a matrix $\hat{A}$ that is entry-wise $o(\rho)$-close to the true dictionary A.*[2] *Furthermore, under Assumptions 1 and 2' the same algorithm returns an $\hat{A}$ that is entry-wise n^{-C}-close to the true dictionary, using n^{4C+3} samples, where C is a large constant controlled by Δ.*

Now we move to the general case. In term of parameters, we still need $d \ge \Delta\Lambda\log^2 n/\sigma^2$, and κ, τ as before, and Δ to be a large enough constant.

Theorem 2 (General Case). *Under Assumptions G1, G2' and G3, when $\rho = o(\sigma^5/\Lambda^{6.5}\log^{2.5} n) = o_\theta(1/\log^{2.5} n)$ there is an algorithm that runs in $n^{O(\Delta\Lambda\log^2 n/\sigma^2)}$ time, uses $n^{4C+5}m$ samples and outputs $\hat{A}$ that is entry-wise n^{-C}-close to the true dictionary A, where C is a constant depending on Δ.*

[1] Even here, we allow exceptions.

[2] It also turns out to be "$o(\rho)$-equivalent" to A, as defined in the motivation earlier.

Section 3 presents the algorithm for nonnegative dictionaries and serves to illustrate our algorithmic ideas. The case of general dictionaries (Theorem 2) is then sketched in Section 4. Details are in the appendix.

A few months after the preliminary version of this paper, Barak et al. have informed us (personal communication) that they have improved upon the results here using semidefinite programming.

3 Nonnegative Dictionary Learning

Let us start with a rough outline of the algorithm. Recall that the difficulty in dictionary learning is that both A and x are unknown. To get around this problem, previous works (e.g. [AGM13]) try to extract information about the assignment x without first learning A (but assuming nice properties of A). After finding x, recovering A becomes easy. In [AGM13] the unknown x's were recovered via an overlapping clustering procedure. The procedure relies on incoherence of A, as when A is incoherent it is possible to test whether the support of x^1, x^2 intersect. This idea fails when x is only slightly sparse, because in this setting the supports of x^1, x^2 always have a large intersection.

3.1 Outline of our algorithm

Our algorithm instead relies on correlations among *pixels*. The key observation is as follows: if the jth bit in x is 1, then $Ax = A_j + \sum_{k\neq j} A_k x_k$. Pixels with high values in A_j tend to be *elevated* above their mean values (recall A is nonnegative). At first it is unclear how this elevation can be spotted, since A_j is unknown and these elevations/correlations among pixels are much smaller than the standard deviation of individual pixels. Therefore we look for *subsets* of pixels that are *jointly* elevated (in terms of the sum of pixel values). For every feature, we define *signature sets* (Definition 1), that help us identify when $x_j = 1$. That is, if the pixles in a signature set are jointly elevated, then with good probability, it must be because feature x_j is present in the image.

Our assumptions will imply the existence of signature sets of polylogarithmic size. Thus in quasi-polynomial time we can afford to enumerate all sets of that size, and check if the pixels in these sets are likely to be elevated together. However there can be many sets – called *correlated sets* below – that could be elevated together, but not all of them are signature sets for some feature. The challenge is thus to separate signature sets from other correlated sets.

This leads to the next idea: try to *expand* a correlated set. If it happens to be a signature set for some feature x_j, then we obtain a set of size d (which is $\gg \text{polylog}(n)$, and hence could not have been found by exhaustive guessing) that still *behaves like* a signature set. The key to the analysis is to prove that expanded sets look different for signature sets as compared to other correlated sets.

Once we find expanded sets with this 'signature like' property, and we can get a rough estimate for the matrix A. Then, using the individually recoverable properties of the features, we can refine the solution to be inverse polynomially close to the true dictionary.

The overall algorithm is described at the end of Section 3; the concepts such as correlated sets, and empirical bias are defined below. It has three main steps. Section 3.2 explains how to test for correlated sets and expand a set (steps 1-2 in the algorithm); Section 3.3 shows how to identify expansions of signature sets and use it to roughly estimate A (steps 3-6); finally Section 3.4 shows how to refine the solution and get $\hat{A}$ that is inverse polynomially close to A (steps 7-10).

3.2 Correlated Sets, Signature Sets and Expanded Sets

We consider a set of pixels T of size $t = \Omega(\mathrm{poly}\log n)$ (to be specified later), and denote by β_T the random variable representing the sum of all pixels in T, i.e., $\beta_T = \sum_{i\in T} y_i$. We can expand β_T as

$$\beta_T = \sum_{i\in T} y_i = \sum_{i\in T}\left(\sum_{j=1}^{m} A_j^{(i)} x_j\right) = \sum_{j=1}^{m}\left(\sum_{i\in T} A_j^{(i)}\right) x_j.$$

Let $\beta_{j,T} = \left(\sum_{i\in T} A_j^{(i)}\right)$ be the contribution of x_j to the sum β_T, then β_T is just

$$\beta_T = \sum_{j=1}^{m} \beta_{j,T} x_j \tag{1}$$

Note that by the normalization of $\mathbb{E}[y_i]$, we have $\mathbb{E}[\beta_T] = \sum_{i\in T}\mathbb{E}[y_i] = t$. Intuitively, if for all j, $\beta_{j,T}$'s are relatively small, β_T should concentrate around its mean. On the other hand, if there is some j whose coefficient $\beta_{j,T}$ is significantly larger than other $\beta_{k,T}$, then β_T will be *elevated* by $\beta_{j,T}$ precisely when $x_j = 1$. That is, with probability roughly ρ (corresponding to when $x_j = 1$), we should observe β_T to be roughly $\beta_{j,T}$ larger than its expectation.

This motivates the definition of *signature sets*, which have only one large value $\beta_{k,T}$.

Definition 1 (Signature Set)**.** *A set of pixels T of size t is a* signature set *for the feature x_j, if $\beta_{j,T} \geq \sigma t$, and for all $k \neq j$, the contribution $\beta_{k,T} \leq \sigma^2 t/_{\Delta \log n}$. Here Δ is a large enough constant.*

The following lemma formalizes the earlier intuition that if T is a signature set for x_j, then a large β_T is highly correlated with the event $x_j = 1$.

Lemma 3. *Suppose T of size t is a signature set for feature x_j with $t = \omega(\sqrt{\log n})$. Let E_1 be the event that $x_j = 1$ (feature is on) and E_2 be the event that $\beta_T \geq \mathbb{E}[\beta_T] + 0.9\sigma t$ (signature is observed). Then for large constant C (depending on the Δ in Definition 1)*

1. $\Pr[E_1] + n^{-2C} \geq \Pr[E_2] \geq \Pr[E_1] - n^{-2C}$.
2. $\Pr[E_2|E_1] \geq 1 - n^{-2C}$, *and* $\Pr[E_2|E_1^c] \leq n^{-2C}$.
3. $\Pr[E_1|E_2] \geq 1 - n^{-C}$.

Proof. (Sketch) We can write β_T as $\beta_T = \beta_{j,T} x_j + \sum_{k\neq j} \beta_{k,T} x_k$. The idea is that the summation in the RHS above is highly concentrated around its mean, which is roughly t. Note that it is a sum of independent variables with maximum value $M = \sigma^2 t/(\Delta \log n)$, by the definition of signature sets. Thus the variance $\rho \sum_{k\neq j} \beta_{k,T}^2$ is bounded by $M\rho \sum_{k\neq j} \beta_{k,T} \leq Mt$. Then by Bernstein's inequality this sum is $0.1\sigma t$-close to its mean with high probability. Therefore since $\beta_{j,T} > \sigma t$, we know $\beta_T > t + 0.9\sigma t$ essentially iff $x_j = 1$. A formal proof can be found in Appendix B.1. □

Thus if we can find a signature set for x_j, we would roughly know the samples in which $x_j = 1$. The following lemma shows that assuming a low pairwise overlap among features, there *exists* a signature set for every feature x_j.

Lemma 4. *Suppose A satisfies Assumptions 1 and 2, let $t = \Omega(\Lambda\Delta \log^2 n/\sigma^2)$, then for any feature $j \in [n]$, there exists a signature set of size t for x_j.*

The proof is by the probabilistic method, noting that each x_j has at least d neighbors in G_σ, and then using Assumption 2'. See Appendix B.2 for the full proof.

Although signature sets exist for all x_j, it is difficult to find them; even if we enumerate all subsets of size t, it is not clear how to know when we found a signature set. Thus we first look for "correlated" sets, which are defined as follows:

Definition 2 (Correlated Set)**.** *A set of pixels T of size t is called* correlated*, if with probability at least $\rho - 1/n^2$ over the choice of x's, $\beta_T \geq \mathbb{E}[\beta_T] + 0.9\sigma t = t + 0.9\sigma t$.*

It follows easily (Lemma 3) that signature sets must be correlated sets. However, the other direction is far from true. There can be many correlated sets that are not signature sets . A simple counterexample would be that there are j and j' such that both $\beta_{j,T}$ and $\beta_{j',T}$ are larger than σt. This kind of counterexample seems inevitable for any test on a set T of polylogarithmic size.

To overcome this, we *expand* the correlated set T into $\tilde{T}$ of size d. Using these larger sets, we will see how to find signature sets. Algorithm 1 and Definition 3 show how to expand T to $\tilde{T}$. The quantity $\hat{\mathbb{E}}[f(y)]$ denotes the "empirical expectation" of $f(y)$, i.e.,$\frac{1}{N}\sum_{i=1}^N f(y^i)$.

```
Algorithm 1 $\tilde{T}$ = expand($T$, threshold)
Input: Correlated set $T$, $d$, and $N$ samples $\{y^1, \dots, y^N\}$
Output: vector $\tilde{A}_T \in \mathbb{R}^n$ and expanded set $\tilde{T}$ of size $d$
1: Let $L$ be the set of samples whose $\beta_T$ values are larger than $\hat{\mathbb{E}}[\beta_T] + threshold$
       $L = \left\{y^k \mid \beta_T^k \geq \hat{\mathbb{E}}[\beta_T] + threshold\right\}$   (here $\beta_T^k$ is the value of $\beta_T$ in sample $y^k$)
2: Estimation Step: Compute empirical mean in $L$, and obtain $\tilde{A}_T$
       $\hat{\mathbb{E}}_L[y] = \frac{1}{|L|}\left(\sum_{y^k \in L} y^k\right)$, and $\tilde{A}_T(i) = \max\{0, (\hat{\mathbb{E}}_L[y_i] - \hat{\mathbb{E}}[y_i])/(1-\rho)\}$
3: Expansion Step: $\tilde{T}$ = {$d$ largest coordinates of $\tilde{A}_T$}
```

Definition 3. *For any set of pixels T of size t, the expanded set $\tilde{T}$ for T is defined as the one output by the procedure expand$(T, 0.9\sigma t)$ (Algorithm 1). The estimation $\tilde{A}_T$ is the output at step 2.*

When T is a signature set for x_j, it turns out that $\tilde{A}_T$ is close to the true A_j, and the expanded set $\tilde{T}$ is essentially the set of largest entries of A_j.

Lemma 5. *If T is a signature set for x_j and the number of samples $N = \Omega(n^{2C}/\rho^3)$, where δ is any positive constant, then with high probability $||\tilde{A}_T - A_j||_\infty \leq 1/n^C$.*

The proof uses Lemma 3 to conclude that L will almost precisely be the samples in which $x_j = 1$, which then implies the bound. See appendix B.3 for the details.

3.3 Using Expanded Sets

We will now see the advantage that the expanded sets $\tilde{T}$ provide. If T happens to be a signature set, the expanded set $\tilde{T}$ for T also has similar property (defined below). But now $\tilde{T}$ is a much larger set (size d as opposed to polylog(n)), so Assumption 2' will imply that if we see a large elevation it is much more likely to be caused by a single feature.

Definition 4 (Weak signature property)**.** *An expanded set $\tilde{T}$ is said to have a weak signature property for x_j if $\beta_{j,\tilde{T}} \geq 0.7\sigma d$ and for all $k \neq j$, $\beta_{k,\tilde{T}} \leq 0.3\sigma d$.*

Note that the weak signature property only requires a constant gap between largest $\beta_{j,T}$ and the second largest one, as opposed to logarithmic in the definition of signature sets. We now have the following:

Lemma 6. *If T is a signature set for x_j, then the expanded set $\tilde{T}$ has the weak signature property for x_j. In fact, we have $\beta_{j,\tilde{T}} \geq 0.9\sigma d$.*

The proof proceeds by combining Lemma 5, which implies that $\tilde{T}$ contains almost all the d large neighbors of x_j (thus $\beta_{j,\tilde{T}}$ is large), and Assumption 2', which gives that for $k \neq j$, $\beta_{k,\tilde{T}}$ should be small. See Section B.4 for the details. We now introduce a key notion that helps us identify signature sets. It is a precise way to measure simultaneous elevation of coordinates in $\tilde{T}$.

Definition 5 (Empirical Bias)**.** *The empirical bias $\hat{B}_{\tilde{T}}$ of an expanded set $\tilde{T}$ is defined to be the largest B that satisfies*

$$\left|\left\{k \in [N] : \beta^k_{\tilde{T}} \geq \hat{\mathbb{E}}[\beta_{\tilde{T}}] + B\right\}\right| \geq \rho N/2.$$

In other words, $\hat{B}_{\tilde{T}}$ is the difference between the $\rho N/2$-th largest $\beta^k_{\tilde{T}}$ in the samples and $\hat{\mathbb{E}}[\beta_{\tilde{T}}]$.

It will turn out that the bias of $\tilde{T}$ is roughly equal to the largest $\beta_{j,\tilde{T}}$. The key lemma is now that the expanded set with largest empirical bias must have the weak signature property for some x_j. This immediately lets us estimate the column A_j to a good accuracy, which we then "subtract off" and iteratively estimate columns.

Lemma 7. *Let $\tilde{T}^*$ be the set with largest empirical bias $\hat{B}_{\tilde{T}^*}$ among all the expanded sets $\tilde{T}$. Then $\tilde{T}^*$ has the weak signature property for some x_j.*

The lemma is proved in multiple steps. The first is to show that the bias of $\tilde{T}$ is roughly equal to the largest $\beta_{j,\tilde{T}}$. If $\beta_{\tilde{T}}$ contains a large term $\beta_{j,\tilde{T}}x_j$, then certainly this term will contribute to the bias $\hat{B}_{\tilde{T}}$; on the other hand, suppose for instance that $\beta_{\tilde{T}}$ has precisely two non-zero terms $\beta_{j,\tilde{T}}x_j + \beta_{k,\tilde{T}}x_k$. Then they cannot contribute more than $\max\{\beta_{j,\tilde{T}}, \beta_{k,\tilde{T}}\}$ to the bias, because otherwise both x_k and x_j have to be 1 to make the sum larger than $\max\{\beta_{j,\tilde{T}}, \beta_{k,\tilde{T}}\}$, and this only happens with probability $\rho^2 \ll \rho/2$.

The intuitive argument above is not far from true: basically we will show that (a) there are very few large coefficients $\beta_{k,\tilde{T}}$ (see Claim 8 for the precise statement), and (b) the sum of the terms with small $\beta_{k,\tilde{T}}$ concentrates around its mean, thus will not contribute much to the bias.

After relating the bias of $\tilde{T}$ to the largest coefficients $\max_j \beta_{j,\tilde{T}}$, we will argue that taking the set $\tilde{T}^*$ with largest bias among all the $\tilde{T}$, we not only see a large coefficient $\beta_{j,\tilde{T}}$, but we also see a gap between the the top $\beta_{j,\tilde{T}}$ and other $\beta_{k,T}$, thus establishing the weak signature property for x_j.

We make the arguments above precise by the following claims. First, we shall show there cannot be too many large coefficients $\beta_{j,D}$ for any set D of size d

Claim 8. *For any set of pixels $\tilde{T}$ of size d, the number of features k such that $\beta_{k,\tilde{T}}$ is larger than $d\sigma^4/\Delta\Lambda^2 \log n$ is at most $O(\Delta\Lambda^3 \log n/\sigma^4)$.*

Each large $\beta_{k,\tilde{T}}$ implies that the neighborhood of x_k has a large intersection with $\tilde{T}$. Therefore many such large $\beta_{k,\tilde{T}}$ implies there are k, k' whose neighborhoods have a large intersection, which contradicts assumption 2. See the formal proof in Appendix B.6.

Let us define $k^* = \arg\max_k \beta_{k,\tilde{T}}$. The next claim shows that the empirical bias $\hat{B}_{\tilde{T}}$ is a good estimate of $\beta_{k^*,\tilde{T}}$ when $\beta_{k^*,\tilde{T}}$ is large.

Claim 9. *For any expanded $\tilde{T}$ of size d, with high probability over the choices of all the N samples, the empirical bias $\hat{B}_{\tilde{T}}$ is within $0.1d\sigma^2/\Lambda$ to $\beta_{k^*,\tilde{T}} = \max_k \beta_{k,\tilde{T}}$ when $\beta_{k^*,\tilde{T}}$ is at least $0.5d\sigma$.*

The idea is to show that the empirical bias is determined by the samples in which $x_{k^*} = 1$. Roughly speaking, this is because the *small* $\beta_{k,T}$ do not contribute to the bias (because of concentration bounds), and there are only a few large $\beta_{k,T}$ by the earlier claim, so the probability that *two* such x_k are 1 is $\ll \rho/2$. See Appendix B.5 for a formal proof.

Now we are ready to prove Lemma 7.

Proof of Lemma 7. By Claim 9 and Lemma 6, we know the maximum bias is at least $0.8\sigma d$. Apply Claim 9 again, we know for the set $\tilde{T}^*$ that has largest bias, there must be a feature j with $\beta_{j,\tilde{T}^*} \geq 0.7\sigma d$.

For the sake of contradiction, let us assume that this $\tilde{T}^*$ does not have the weak signature property. Then there must be some $k \neq j$ where $\beta_{k,\tilde{T}^*} \geq 0.3\sigma d$. Let Q_j and Q_k be the set of nodes in $\tilde{T}^*$ that are connected to j and k in G_τ (these are the same Q's as in the proof of Claim 8). We know $|Q_j \cap Q_k| \leq \kappa$ by assumption, and $|Q_k| \geq 0.3\sigma d/\Lambda$. This implies that $|Q_j| \leq d - 0.3\sigma d/\Lambda + \kappa$.

Now let T' be a signature set for x_j, and let $\tilde{T}'$ be its expanded set. From Lemma 5 we know $\beta_{j,\tilde{T}'}$ is almost equal to the sum of the d largest entries in A_j, which is at least $0.2\sigma^2 d/\Lambda$ larger than $\beta_{j,\tilde{T}^*}$, since $|Q_j| \leq d - 0.2\sigma d/\Lambda$. By Claim 9 we know $\hat{B}(\tilde{T'}) \geq \beta_{j,\tilde{T}'} - 0.1\sigma^2 d/\Lambda > \beta_{j,\tilde{T}^*} + 0.1\sigma^2 d/\Lambda \geq \hat{B}(\tilde{T})$, which contradicts the assumption that $\tilde{T}^*$ is the set with max bias. □

We now show that having a set $\tilde{T}$ with a weak signature property for x_j allows us to obtain a good estimate for the column A_j using the procedure expand() (Algorithm 1).

Lemma 10. *Suppose $\tilde{T}$ has the weak signature property for x_j, and let $\tilde{A}_{\tilde{T}}$ be the column output by expand($\tilde{T}, 0.6\sigma d$), then with high probability $\|\tilde{A}_{\tilde{T}} - A_j\|_\infty \leq O(\rho(\Lambda^3 \log n/\sigma^2)^2\sqrt{\Lambda \log n}) = o(\sigma)$.*

The lemma is very similar to the Lemma 5 for signature sets, and is proved using a similar idea (see Appendix B.7). However, the main advantage is that it helps us recover all the *significant* entries in the column A_j! We will now use this to iteratively find other columns.

Suppose we estimated k columns. For simplicity, assume the estimates we obtained are for the first k columns (call the estimates $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_k$). Since they are close to $A_1, A_2, \ldots, A_k$ respectively, for any expanded set $\tilde{T}$, we can estimate $\hat{\beta}_{j,\tilde{T}}$ up to an additive $o(\sigma d)$, for any $1 \leq j \leq k$. We now have the following:

Lemma 11. *Suppose we estimated the first k columns as $\tilde{A}_i$, each entry correct up to an additive $o(\sigma)$ error. Let $\tilde{T}$ be the set with largest empirical bias among the expanded sets that have $\hat{\beta}_{j,\tilde{T}} < 0.2\sigma d$ for all $j \leq k$. Then $\tilde{T}$ has the weak signature property for some x_j with $j > k$.*

The proof is almost identical to that of Lemma 7 (see Appendix B.9).

3.4 Refining an Approximate Dictionary

Repeatedly finding columns as above, we obtain estimates $\tilde{A}_j$ of all the columns A_j that are entry-wise $o(\sigma)$ close. We will now see how to refine this solution to obtain dictionaries that are entry-wise ϵ-close for very small ϵ. The key is to look at *all* the large entries in the column A_j, and use them to identify whether feature x_j is 1 or 0.

Lemma 12. *Let S_j be the set of all entries larger than $\sigma/2$ in $\tilde{A}_j$, then $|S_j| \ge d$, $\beta_{j,S_j} \ge (0.5 - o(1))\,|S_j|\,\sigma$, and for all $k \ne j$ $\beta_{k,S_j} \le \sigma^2\,|S_j|\,/\Delta \log n$ where Δ is a large enough constant.*

This follows directly from the assumptions (see Appendix B.8).

We can now prove the main theorem.

Proof of Theorem 1. Since S_j has a unique large coefficient β_{j,S_j}, and the rest of the coefficients are much smaller, when Δ is large enough, and $N \ge n^{4C+\delta}/\rho^3$ we have that $\hat{A}_j$ is entry-wise $n^{-2C}/\log n$ close to A_j (this is using the same argument as in Lemma 5), and this completes the proof of the theorem. □

This completes the description of the algorithm, and the proof of (the second part of) Theorem 1. We formally write down the algorithm below (Algorithm 2).

Algorithm 2 Nonnegative Dictionary Learning

```
Input: $N$ samples $\{y^1, \dots, y^N\}$ generated by $y^i = Ax^i$. Unknown dictionary $A$ satisfies Assumptions 1 and 2.
Output: $\hat{A}$ that is $n^{-C}$ close to $A$
Enumerate all sets of size $t = O(\Lambda \log^2 n/\sigma^4)$, keep the sets that are correlated.
Expand all correlated sets $T$, $\tilde{T} = Expand(T, 0.9\sigma t)$.
for $j = 1$ TO $m$ do
    Let $\tilde{T}_j$ be the set with largest empirical bias, and $\forall k < j$, $\hat{\beta}_{k,\tilde{T}} = \sum_{i \in T} \tilde{A}_{\tilde{T}_k}(i) \le 2d\sigma$.
    Let $\tilde{A}_{\tilde{T}_k}$ be the result of estimation step in $Expand(\tilde{T}, 0.6\sigma d)$.
end for
for $j = 1$ TO $m$ do
    Let $S_j$ be the set of entries that are larger than $\sigma/2$ in $\tilde{A}_{\tilde{T}_j}$
    Let $\hat{A}_i$ be the result of estimation step in $Expand(S_j, 0.4\sigma\,|S_j|)$
end for
```

3.5 Working with Assumption 2

In order to assume Assumption 2 instead of 2', we need to change the definition of signature sets to allow $o(1/\sqrt{\rho})$ "moderately large" $(\sigma t/10)$ entries. This makes the definition look similar to the weak signature property. Such signature sets still exist by similar probabilistic argument as in Lemma 4. Lemma 6 and Claims 8 and 9 can also be adapted.

Finally, in the proof of Theorem 1, the guarantee will be weaker (there can be $o(1/\sqrt{\rho})$ moderately large coefficients). The algorithm will only estimate x_j incorrectly if at least 6 such coefficients are "on" (has the corresponding x_j being 1), which happens with less than $o(\rho^3)$ probability. By argument similar to Lemma 5, we get the first part of Theorem 1.

4 General Case

We show that with minor modifications, our algorithm and its analysis can be adapted to the general case in which the matrix A can have both positive and negative entries (Theorem 2).

We follow the outline from the non-negative case, and look at sets T of size t. The quantities β_T and $\beta_{j,T}$ are defined exactly the same as in Section 3.2. Additionally, let ν_T be the standard deviation of β_T, and let $\nu_{-j,T}$ be the standard deviation of $\beta_T - \beta_{j,T}x_j$. That is,

$$\nu_{-j,T}^2 = \mathbb{V}[\beta_T - \beta_{j,T}x_j] = \rho \sum_{k\neq j} \beta_{k,T}^2.$$

The definition of signature sets requires an additional condition to take into account the standard deviations.

Definition 6 ((General) Signature Set)**.** *A set T of size t is a* signature set *for x_j, if for some large constant Δ, we have: (a) $|\beta_{j,T}| \geq \sigma t$, (b) for all $k \neq j$, the contribution $|\beta_{k,T}| \leq \sigma^2 t/(\Delta \log n)$, and additionally, (c) $\nu_{-j,T} \leq \sigma t/\sqrt{\Delta \log n}$.*

In the nonnegative case the additional condition $\nu_{-j,T} \leq \sigma t/\sqrt{\Delta \log n}$ was automatically implied by nonnegativity and scaling. Now we use Assumption G3 to show there exist T in which (c) is true along with the other properties. To do that, we prove a simple lemma which lets us bound the variance (the same lemma is also used in other places).

Lemma 13. *Let T be a set of size t and S be an arbitrary subset of features, and consider the sum $\beta_{S,T} = \sum_{j\in S} \beta_{j,T}x_j$. Suppose for each $j \in S$, the number of edges from j to T in graph G_τ is bounded by W. Then the variance of $\beta_{S,T}$ is bounded by $2tW + 2t^2\gamma$.*

Proof. (Sketch) The idea is to split the weights $A_j^{(i)}$ into the *big* and *small* ones (threshold being τ). Intuitively, on one hand, the contribution to the variance from large weights is bounded above because the number of such large edges in bounded by W. On the other hand, by assumption (3), the total variance of small weights is less than γ, which implies that the contribution of small weight to the variance is also bounded. A full proof can be found in Section D.1. □

Lemma 14. *Suppose A satisfies our assumptions for general dictionaries, and let $t = \Omega(\Lambda\Delta \log^2 n/\sigma^2)$. Then for any $j \in [n]$, there exists a general signature set of size t for node x_j (as in Definition 6).*

The proof is very similar to that of Lemma 4, which uses probabilistic method. We defer the proof to Appendix D.2.

The proof of Lemma 3 now follows in the general case (here we will use the variance bound (c) in the general definition of signature sets), except that we need to redefine event E_2 to handle the negative case. For completeness, we state the general version of Lemma 3 in Appendix C. As before, signature sets give a very good idea of whether $x_j = 1$.

Let us now define correlated sets: here we need to consider both positive and negative bias

Definition 7 ((General) Correlated Set)**.** *A set T of size t is* correlated*, if either with probability at least $\rho - 1/n^2$ over the choice of x's, $\beta_T \geq \mathbb{E}[\beta_T] + 0.8\sigma t$, or with probability at least $\rho - 1/n^2$, $\beta_T \leq \mathbb{E}[\beta_T] - 0.8\sigma t$.*

Starting with a correlated set (a potential signature set), we expand it similar to (Definition 3), except that we find $\tilde{T}$ as follows:

$$\tilde{T}_{temp} = \{2d \text{ coordinates of largest } \textit{magnitude} \text{ in } \hat{A}_T\}, \tilde{T}_1 = \{i \in \tilde{T}_{temp} : \hat{A}_T \geq 0\}$$
$$\tilde{T} = \begin{cases} \tilde{T}_1 & \text{if } |T_1| \geq d \\ \tilde{T}_{temp} \setminus \tilde{T}_1 & \text{otherwise} \end{cases}$$

Our earlier definitions of the weak signature property and bias can also be adapted naturally:

Definition 8 ((General) Weak Signature Property)**.** *An expanded set $\tilde{T}$ is said to have the weak signature property for x_j if $|\beta_{j,\tilde{T}}| \geq 0.7\sigma d$ and for all $k \neq j$, $|\beta_{k,\tilde{T}}| \leq 0.3\sigma d$.*

Since Lemma 5 still holds, Lemma 6 is straightforward. That is, there always exists an expanded set $\tilde{T}$ with the general weak signature property, that is produced by a set T of size $t = O_\theta(\log n^2)$. (We use the fact that G_σ has degree at least $2d$)

Definition 9 ((General) Empirical Bias)**.** *The empirical bias $\hat{B}_{\tilde{T}}$ of an expanded set $\tilde{T}$ of size d is defined to be the largest B that satisfies*

$$\left|\left\{k \in [p] : \ \left|\beta^k_{\tilde{T}} - \hat{\mathbb{E}}[\beta_{\tilde{T}}]\right| \geq B\right\}\right| \geq \rho N/2.$$

In other words, $\hat{B}_{\tilde{T}}$ is the difference between the $\rho N/2$-th largest $\beta^k_{\tilde{T}}$ in the samples and $\hat{\mathbb{E}}[\beta_{\tilde{T}}]$.

Let us now intuitively describe why the analog of Lemma 7 holds in the general case. The formal statement and the proof can be found in Appendix C

1. The first step, Claim 8 is a statement purely about the magnitudes of the edges (in fact, cancellations in $\beta_{k,\tilde{T}}$ for $k \neq j$ only help our case).
2. The second step, Claim 9 essentially argues that the small $\beta_{k,\tilde{T}}$ do not contribute much to the bias (a concentration bound, which still holds due to Lemma 13), and that the probability of *two* "large" features j, j' being on simultaneously is very small. The latter holds even if the $\beta_{j,\tilde{T}}$ have different signs.
3. The final step in the proof of Lemma 7 is an argument which uses the assumption on the overlap between features to contradict the maximality of bias, when $\beta_{j,\tilde{T}}$ and $\beta_{j',\tilde{T}}$ are both "large". This only uses the magnitudes of the entries in A, and thus also follows.

Recovering an approximate dictionary. The main lemma in the nonnegative case, which shows that Algorithm 1 *roughly* recovers a column, is Lemma 10. The proof uses the property that expanded sets with the weak signature property are elevated "almost iff" the $x_j = 1$ to conclude that we get a good approximation to one of the columns. We have seen that this also holds in the general case, and since the rest of the argument deals only with the magnitudes of the entries, we conclude that we can roughly recover a column also in the general case. Let us state this formally.

Lemma 15. *Suppose an expanded set $\tilde{T}$ has the weak signature property for x_j, and $\tilde{A}_{\tilde{T}}$ is the corresponding column output by Algorithm 1. Then with high probability,*

$$\|\tilde{A}_{\tilde{T}} - A_j\|_\infty \leq O(\rho(\Lambda^3 \log n/\sigma^2)^2 \sqrt{\Lambda \log n}) = o(\sigma).$$

Proof of Theorem 2. Once we have all the entries which are $> \sigma/2$ in magnitude, we can use the refinement trick of Lemma 12 to conclude that we can recover the entries up to a much higher precision. The argument is very similar to Lemma 5. □

References

[AAN13] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *CoRR*, abs/1309.1952, 2013.

[ABGM13] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. *CoRR*, abs/1310.6343, 2013.

[AEB05] Michal Aharon, Michael Elad, and Alfred M Bruckstein. K-svd and its non-negative variant for dictionary design. In *Optics & Photonics 2005*, pages 591411–591411. International Society for Optics and Photonics, 2005.

[AEB06] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[AEP06] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NIPS*, pages 41–48, 2006.

[AGM13] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *CoRR*, abs/1308.6273, 2013.

[Aha06] Michal Aharon. *Overcomplete Dictionaries for Sparse Representation of Signals*. PhD thesis, Technion - Israel Institute of Technology, 2006.

[BC$^+$07] Y-lan Boureau, Yann L Cun, et al. Sparse feature learning for deep belief networks. In *Advances in neural information processing systems*, pages 1185–1192, 2007.

[Ben62] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):pp. 33–45, 1962.

[Ber27] S. Bernstein. *Theory of Probability*, 1927.

[BGI$^+$08] R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss. Combining geometry and combinatorics: a unified approach to sparse signal recovery. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 798–805, 2008.

[CRT06] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In *FOCS*, pages 634–644. IEEE Computer Society, 1999.

[DH01] David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845–2862, 2001.

[DMA97] Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[EA06] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

[EAHH99] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.

[Hoy02] Patrik O Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE, 2002.

[Ind08] Piotr Indyk. Explicit constructions for compressed sensing of sparse signals. In Shang-Hua Teng, editor, *SODA*, pages 30–33. SIAM, 2008.

[JXHC09] Sina Jafarpour, Weiyu Xu, Babak Hassibi, and A. Robert Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Transactions on Information Theory*, 55(9):4299–4308, 2009.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LS00] Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

[MLB$^+$08] Julien Mairal, Marius Leordeanu, Francis Bach, Martial Hebert, and Jean Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *Computer Vision–ECCV 2008*, pages 43–56. Springer, 2008.

[OF97] 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.

[SWW12] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. *Journal of Machine Learning Research - Proceedings Track*, 23:37.1–37.18, 2012.

[YWHM08] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

A Individual Recoverability of features

All the recent works on dictionary learning mentioned in the introduction use assumptions that imply individual recoverability. The papers of [AGM13] and [AAN13] assume that the columns of A are *incoherent*, i.e., they have pairwise inner product at most $\mu/\sqrt{n}$ where μ is small (about poly$(\log n)$). In such matrices, the features are individually recoverable since $A^T A \approx I$, so given Ax one can take its inner product with the ith column A_i to roughly determine the extent to which feature i is present. So also, dictionaries corresponding to sparse random graphs with random edges weights in $[-1, 1]$ (handled in [ABGM13]) also have sparse, individually recoverable features. However, the learning algorithm there requires strong random-like assumptions on the graph.

B Full Proofs from Section 3

In this section we give the full proofs of lemmas and theorems from Section 3.

B.1 Proof of Lemma 3

We can write β_T as

$$\beta_T = \beta_{j,T}x_j + \sum_{k\neq j}\beta_{k,T}x_k \tag{2}$$

Formally, observe that $\mathbb{E}[\beta_{j,T}x_j] = \rho\beta_{j,T} \leq \rho\Lambda t = o(\sigma t)$, and recall that $\mathbb{E}[\beta_T] = t$, we have $\mathbb{E}[\sum_{k\neq j}\beta_{k,T}x_k] = (1-o(\sigma))t$. Let $M = \sigma^2 t/(\Delta \log n)$ be the upper bound for $\beta_{k,T}$, and then the variance of the sum $\sum_{k\neq j}\beta_{k,T}x_k$ is bounded by $\rho M\sum_{k\neq j}\beta_{k,T} \leq Mt$. Then by calling Bernstein inequality (see Theorem 20, but note that σ there is the standard deviation), we have

$$\Pr\left[\left|\sum_{k\neq j}\beta_{k,T}x_k - \mathbb{E}[\sum_{k\neq j}\beta_{k,T}x_k]\right| > \sigma t/20\right] \quad \leq \quad 2\exp(-\frac{\sigma^2 t^2/400}{2Mt + \frac{2}{3}\frac{M}{\sqrt{Mt}}\sigma t/20}) \leq n^{-2C}.$$

where C is a large constant depending Δ.

Part (2) immediately follows: if $x_j = 1$, then $\beta_T < t + 0.9\sigma t$ iff the sum deviates from its expectation by more than $\sigma t/20$, which happens with probability $< n^{-2C}$. So also if $x_j = 0$, E_2 occurs with probability $< n^{-2C}$.

This then implies part (1), since the probability of E_1 is precisely ρ.

Combining the (1) and (2), and using Bayes' rule $\Pr[E_1|E_2] = \Pr[E_2|E_1]\Pr[E_1]/\Pr[E_2]$, we obtain (3).

B.2 Proof of Lemma 4

We show the existence by probabilistic method. By Assumption 1, node x_j has at least d neighbors in G_σ. Let T be a uniformly random set of t neighbors of x_j in G_σ. Now by the definition of G_σ we have $\beta_{j,T} \geq \sigma t$.

Using a bound on intersection size (Assumption 2') followed by Chernoff bound, we show that T is a signature set with good probability. For $k \neq j$, let $f_{k,T}$ be the number of edges from x_k to T in graph G_τ. Then we can upperbound $\beta_{k,T}$ by $t\tau + f_{k,T}\Lambda$ since all edge weights are at most Λ and there are at most $f_{j,T}$ edges with weights larger than τ. Using simple Chernoff bound and union bound, we know that with probability at least $1 - 1/n$, for all $k \neq j$, $f_{k,T} \leq 4\log n$. Therefore $\beta_{k,T} \leq t\tau + f_{k,T}\Lambda \leq \sigma^2 t/(\Delta\log n)$ for $t \geq \Omega(\Lambda\Delta\log^2 n/\sigma^2)$, and $\tau = O(\sigma^2/_{\Delta\log n})$.

B.3 Proof of Lemma 5

Let us first consider $\mathbb{E}[\tilde{A}_T] \triangleq (\mathbb{E}[y|E_2] - \mathbf{1})/(1-\rho)$ where E_2 is the event that $\beta_T \geq t + 0.9\sigma t$ defined in Lemma 3. Recall that because of normalization, we know for any j, $\sum_{i\in[n]} A_j^{(i)} = 1/\rho$, so in particular $y_i \leq 1/\rho$. By Lemma 3 and some calculations (see Lemma 18), we have that $|\mathbb{E}[y|E_2] - \mathbb{E}[y|E_1]|_\infty \leq n^{-C}$. Note that $\mathbb{E}[y|E_1] = \mathbf{1} + (1-\rho)A_j$. Therefore we have that $|\mathbb{E}[\tilde{A}_T] - A_j|_\infty \leq n^{-C}$.

B.4 Proof of Lemma 6

Since we know there are at least d weights $A_j^{(i)}$ bigger than σ for any column A_j, by Lemma 5 we know $\beta_{j,\tilde{T}} \geq \sigma d - o(1)d \geq 0.9\sigma d$.

Furthermore, Lemma 5 says x_j connects to every node in $\tilde{T}$ with weights larger than 0.9σ (since by Assumption 1 there are more than d edges of weight at least σ from node j). By Assumption 2 on the graph, for any other $k \neq j$, the number of y_i's that are connected to both k and j in G_τ is bounded by κ. In particular, the number of edges from k to $\tilde{T}$ with weights more than τ is bounded by κ. Therefore the coefficient $\beta_{k,\tilde{T}} = \sum_{(i,k)\in G_\tau} A_k^{(i)} + \sum_{(i,k)\notin G_\tau} A_k^{(i)}$ is bounded by $\Lambda\kappa + |\tilde{T}|\tau = o(d) \leq 0.3d$. (Recall $\tau = o(1)$ and $\kappa = o(d)$)

B.5 Proof of Claim 9

Let $K'_{large} = K_{large} \setminus \{k^*\}$ [3], and $\beta_{small,\tilde{T}} = \sum_{k\notin K_{large}} \beta_{k,\tilde{T}} x_k$, and $\beta_{large,\tilde{T}} = \sum_{k\in K'_{large}} \beta_{k,\tilde{T}} x_k$.

First of all, the variance of $\beta_{small,\tilde{T}}$ is bounded by $\rho \sum_{k\notin K_{large}} \beta^2_{k,\tilde{T}} \leq d\sigma^4/\Delta\Lambda^2 \log n \cdot \left(\rho \sum_{k\notin K_{large}} \beta_{k,\tilde{T}}\right) \leq d^2\sigma^4/\Delta\Lambda^2 \log n$. By Bernstein's inequality, for sufficiently large Δ, with probability at most $1/n^2$ over the choice of x, the value $|\beta_{small,\tilde{T}} - \mathbb{E}[\beta_{small,\tilde{T}}]|$ is larger than $0.05d\sigma^2/\Lambda$, that is, $\beta_{small,\tilde{T}}$ nicely concentrates around its mean. Secondly, with probability at most ρ we have $x_{k^*} = 1$, and then $\beta_{k^*,\tilde{T}} x_{k^*}$ is elevated above its mean by roughly $\beta_{k^*,\tilde{T}}$. Thirdly, the mean of $\beta_{large,\tilde{T}}$ is at most $\rho \sum_{k\in K'_{large}} \beta_{k,\tilde{T}} \leq \rho|K|d$, which is $o(\sigma d)$ by Claim 8. These three points altogether imply that with probability at least $\rho - n^{-2}$, $\beta_{\tilde{T}}$ is above its mean by $\beta_{k^*,\tilde{T}} - 0.1\sigma^2 d/\Lambda$. Also note that the empirical mean $\hat{\mathbb{E}}[\beta_{\tilde{T}}]$ is sufficiently close to the $\beta_{\tilde{T}}$ with probability $1-\exp(-\Omega(n))$ over the choices of N samples, when $N = poly(n)$. Therefore with probability $1 - \exp(-\Omega(n))$ over the choices of N samples, $\hat{B}_{\tilde{T}} > \beta_{k^*,\tilde{T}} - 0.1\sigma^2 d/\Lambda$.

It remains to prove the other side of the inequality, that is, $\hat{B}_{\tilde{T}} \leq \beta_{k^*,\tilde{T}} + 0.1\sigma^2 d/\Lambda$.

Note that $|K_{large}| = O(\log n)$, thus with probability at least $1 - 2\rho^2|K|^2$, at most one of the $x_k, (k \in K_{large})$ is equal to 1. Then with probability at least $1 - 2\rho^2|K|^2$ over the choices of x, $\beta_{large,\tilde{T}} + \beta_{k^*,\tilde{T}}$ is elevated above its mean by at most $\beta_{k^*,\tilde{T}}$. Also with probability $1 - n^{-2}$ over the choices of x, $\beta_{small,\tilde{T}}$ is above its mean by at most $0.1\sigma^2 d/\Lambda$. Therefore with probability at least $1 - 3\rho^2|K|^2$ over the choices of x, $\beta_{\tilde{T}}$ is above its mean by at most $\beta_{k^*,\tilde{T}} + 0.1\sigma d/\Lambda$. Hence when $3\rho^2|K|^2 \leq \rho/3$, with probability at least $1 - \exp(-\Omega(n))$ over the choice of the N samples, $\hat{B}_{\tilde{T}} \leq \beta_{k^*,\tilde{T}} + 0.1\sigma^2 d/\Lambda$. The condition is satisfied when $\rho \leq c/\log^2 n$ for a small enough constant c.

B.6 Proof of Claim 8

For the ease of exposition, we define $K_{large} = \{k : \beta_{k,\tilde{T}} \geq d\sigma^4/\Delta\Lambda^2 \log n\}$. Hence the goal is to prove that $|K_{large}| \leq O(\Delta\Lambda^3 \log n/\sigma^4)$. Recall that $\beta_{k,\tilde{T}} = \sum_{i\in\tilde{T}} A_k^{(i)}$. Let $Q_k = \{i \in \tilde{T} : A_k^{(i)} \geq \tau\}$ be the subset of nodes in $\tilde{T}$ that connect to k with weights larger than τ. We have that $\beta_{k,\tilde{T}} = \sum_{i\notin Q_k} A_k^{(i)} + \sum_{i\in Q_k} A_k^{(i)}$. The first sum is upper bounded by $d\tau \leq d\sigma^4/2\Delta\Lambda^2 \log n$. Therefore for $k \in K_{large}$, the second sum is lower bounded by $d\sigma^4/2\Delta\Lambda^2 \log n$. Since $A_k^{(i)} \leq \Lambda$, we

[3] K_{large} is defined in proof of Claim 8

have $|Q_k| \geq \sigma^4 d/2\Delta\Lambda^3 \log n$. We will use this, along with a bound on $|Q_k \cap Q_{k'}|$ to bound the size of K_{large}.

By Assumption 2' we know in graph G_τ, any two features cannot share too many pixels: for any k and k', $|Q_k \cap Q_{k'}| \leq \kappa$. Also note that by definition, $Q_j \subset \tilde{T}$, which implies that $|\cup_{k \in K_{large}} Q_k| \leq |\tilde{T}| = d$. By inclusion-exclusion we have

$$d \geq |\bigcup_{k \in K_{large}} Q_k| \geq \sum_{k \in K_{large}} |Q_k| - \sum_{k,k' \in K_{large}} |Q_k \cap Q_{k'}| \geq |K_{large}| \sigma^4 d/2\Delta\Lambda^3 \log n - |K_{large}|^2/2 \cdot \kappa \quad (3)$$

This implies that $|K_{large}| \leq O(\Delta\Lambda^3 \log n/\sigma^4)$, when $\kappa = O(\sigma^8 d/\Delta^2\Lambda^6 \log^2 n)$. [4]

B.7 Proof of Lemma 10

Define E_1 to be the event that $x_j = 1$, and E_2 to be the event that $\beta_{\tilde{T}} \geq 0.6 d\sigma$.

When E_1 happens, event E_2 always happen unless $\beta_{\tilde{T},small}$ is far from its expectation. In the proof of Claim 9 we've already shown the number of such samples is at most n with very high probability.

Suppose E_2 happens, and E_1 does not happen. Then either $\beta_{\tilde{T},small}$ is far from its expectation, or at least two x_j's with large coefficients $\beta j, \tilde{T}$'s are on. Recall by Claim 8 the number of x_j's with large coefficients is $|K| \leq O(\Lambda^3 \log n/\sigma^2)$, so the probability that at least two large coefficient is "on" (with $x_j = 1$) is bounded by $O(\rho^2 \cdot |K|^2) = \rho \cdot O(\rho\Lambda^6 \log^2 n/\sigma^4) = \rho \cdot o(\sigma/\sqrt{\Lambda \log n})$. With very high probability the number of such samples is bounded by $\rho N \cdot o(\sigma/\sqrt{\Lambda \log n})$.

Combining the two parts, we know the number of samples that is in $E_1 \oplus E_2$ (the symmetric difference between E_1 and E_2) is bounded by $\rho N \cdot o(\sigma/\sqrt{\Lambda \log n})$. Also, with high probability $(1 - n^{-C})$ all the samples have entries bounded by $O(\sqrt{\Lambda \log n})$ by Bernstein's inequality (variance of y_i is bounded by $\sum_j \rho(A_j^{(i)})^2 \leq \max_j A_j^{(i)} \sum_j \rho A_j^{(i)} \leq \Lambda$). Notice that this is a statement of the entire sample independent of the set T, so we do not need to apply union bound over all expanded signature sets.

Therefore by Lemma 18

$$\|\tilde{A}_{\tilde{T}} - A_j\|_\infty \leq o(\sigma/\sqrt{\Lambda \log n}) \cdot O(\sqrt{\Lambda \log n}) = o(\sigma).$$

B.8 Proof of Lemma 12

This follows directly from the assumptions. By Assumption 1, there are at least d entries in A_j that are larger than σ, all these entries will be at least $(1 - o(1))\sigma$ in $\tilde{A}_{\tilde{T}_j}$, so $|S_j| \geq d$.

Also, since for all $i \in S_j$, $\tilde{A}_{\tilde{T}_j}(i) \geq 0.5\sigma$, we know $A_j(i) \geq 0.5\sigma - o(\sigma)$, hence $\beta_{j,S_j} \geq (0.5 - o(1))\,|S_j|\,\sigma$.

By Assumption 2, for any $k \neq j$, the number of edges in G_τ between k and S_j is bounded by κ, so $\beta_{k,S_j} \leq \tau\,|S_j| + \kappa\Lambda \leq \sigma^2\,|S_j|\,/\Delta \log n$.

[4] Note that any subset of K_{large} also satisfies equation (3), thus we don't have to worry about the other range of the solution of (3)

B.9 Proof of Lemma 11

First, if T is a signature set of x_j where $j > k$, then by Lemma 6 $\tilde{T}$ must satisfy $\hat{\beta}_{j,\tilde{T}} < 0.2\sigma d$, so it will compete for the set with largest empirical bias.

Also, since $\hat{\beta}_{j,\tilde{T}} < 0.2\sigma d$, we know the coefficients in $\beta_{j,\tilde{T}}$ must have $j > k$. Leveraging this observation in the proof of Lemma 7 gives the result.

C Detailed Lemmas from Section 4

Lemma 16 (General Version of Lemma 3)**.** *Suppose T of size t is a general signature set for x_j with $t = \omega(\sqrt{\log n})$. Let E_1 be the event that $x_j = 1$ and E_2 be the event that $\beta_T \geq \mathbb{E}[\beta_T] + 0.9\sigma t$ if $\beta_{j,T} \geq \sigma t$, and the event $\beta_T \leq \mathbb{E}[\beta_T] - 0.9\sigma t$ if $\beta_{j,T} \leq -\sigma t$. Then for large constant C (depending on Δ)*

1. $\Pr[E_1] + n^{-2C} \geq \Pr[E_2] \geq \Pr[E_1] - n^{-2C}$.
2. $\Pr[E_2|E_1] \geq 1 - n^{-2C}$, *and* $\Pr[E_2|E_1^c] \leq n^{-2C}$.
3. $\Pr[E_1|E_2] \geq 1 - n^{-C}$.

Proof. It is a straightforward modification of the proof of Lemma 3. First of all, $|\mathbb{E}[\beta_{j,T}x_j]| = o(\sigma t)$,and thus mean of $\sum_{k\neq j} \beta_{k,T}x_k$ only differs from that of β_T by at most $o(\sigma t)$. Secondly, Bernstein inequality requires the largest coefficients and the total variance be bounded, which correspond to exactly property (b) and (c) of a general signature set. The rest of the proof follows as in that for Lemma 3. □

Lemma 17 (General version of Lemma 7)**.** *Let $\tilde{T}^*$ be the set with largest* general *empirical bias $\hat{B}_{\tilde{T}^*}$ among all the expanded sets $\tilde{T}$. The set $\tilde{T}^*$ has the weak signature property for some x_j.*

Proof. We first prove an analog of Claim 8. Let $W = \sigma^4 d/2\Delta\Lambda^3 \log n$. Let's redefine $K_{large} := \{k \in [m] : \left|\{i \in \tilde{T} : |A_k^{(i)}| \geq \tau\}\right| \geq W\}$ be the subset of nodes in $[m]$ which connect to at least W nodes in $\tilde{T}$ in the subgraph G_τ.Note that this implies that if $k \notin K_{large}$, then $|\beta_{k,\tilde{T}}| \leq d\tau + W\Lambda \leq d\sigma^4/\Delta\Lambda^2 \log n$. Let $Q_k = \{i \in \tilde{T} : |A_k^{(i)}| \geq \tau\}$. By definition, we have for $k \in K_{large}$, $|Q_K| \geq W$. Then similarly as in the proof of Claim 8, using the fact that $|Q_k \cap Q_{k'}| \leq \kappa$, and inclusion-exclusion, we have that $|K_{large}| \leq O(\Delta\Lambda^3 \log n/\sigma^4)$.

Then we prove an analog of Claim 9. Let $\beta_{small,\tilde{T}}$ and $\beta_{large,\tilde{T}}$ be defined as in the proof of Claim 9 (with the new definition of K_{large}). By Lemma 13, the variance of $\beta_{small,\tilde{T}}$ is bounded by $2dW + 2d^2\gamma \leq 2d^2\sigma^4/\Delta\Lambda^2 \log n$. Therefore by Bernstein's inequality we have that for sufficiently large Δ, with probability at least $1 - n^{-2}$ over the choice of x, $|\beta_{small,\tilde{T}} - \mathbb{E}[\beta_{small,\tilde{T}}]| \leq 0.05d\sigma^2/\Lambda$. It follows from the same argument of Claim 9 that with high probability over the choice of N samples, $|\hat{B}_T - \max_k \beta_{k,\tilde{T}}| \leq 0.1d\sigma^2/\Lambda$ holds when $\max_k \beta_{k,\tilde{T}} \geq 0.5d\sigma$.

We apply almost the same argument as in the proof of Lemma 7. We know that our algorithm must produce an expanded set of size d with bias at least $0.8\sigma d$ (the expansion of any signature set), and thus the set $\tilde{T}^*$ with largest bias must has a large coefficient j with $\beta_{j,\tilde{T}^*} \geq 0.7\sigma d$. If there is some other k such that $\beta_{k,\tilde{T}^*} \geq 0.3\sigma d$, then $|Q_k| \geq 0.3\sigma d/\Lambda$ and therefore we could remove those elements in $\tilde{T}^* - Q_j$, which has size larger than $0.3\sigma d/\Lambda - \kappa$ by Assumption G2. Then by adding some other elements which are in the neighborhood of j in G_σ into the set Q_j we get a set with bias larger than $\tilde{T}^*$, which contradicts our assumption that there exists k with $\beta_{k,\tilde{T}^*} \geq 0.3\sigma d$. Thus $\tilde{T}^*$ has the (general) weak signature property for x_j and the proof is complete.

□

D Full Proofs from Section 4

D.1 Proof of Lemma 13

As sketched, we split $\beta_{j,T}^2$ into small and big ones, and bound the variance contributed by small ones using assumption G3, while the big ones by the number of edges W and the maximum weights Λ. Formally, we have

$$\begin{aligned}
\mathbb{V}[\beta_{S,T}] = \rho \sum_{j\in S} \beta_{j,T}^2 &= \rho \sum_{j\in S} \left(\sum_{i\in T} A_j^{(i)} \right)^2 \\
&= \rho \sum_{j\in S} \left(\sum_{i:i\in T,(i,j)\in G_\tau} A_j^{(i)} + \sum_{i:i\in T,(i,j)\notin G_\tau} A_j^{(i)} \right)^2 \\
&\leq 2\rho \sum_{j\in S} \left[\left(\sum_{i:i\in T,(i,j)\in G_\tau} A_j^{(i)} \right)^2 + \left(\sum_{i:i\in T,(i,j)\notin G_\tau} A_j^{(i)} \right)^2 \right] \\
&\leq 2\rho \sum_{j\in S} \left[W \left(\sum_{i:i\in T,(i,j)\in G_\tau} \left(A_j^{(i)}\right)^2 \right) + t \left(\sum_{i:i\in T,(i,j)\notin G_\tau} \left(A_j^{(i)}\right)^2 \right) \right] \\
&= 2\rho W \sum_{i\in T} \sum_{j\in S} \left(A_j^{(i)}\right)^2 + 2\rho t \sum_{i\in T} \sum_{j:(i,j)\notin G_\tau} \left(A_j^{(i)}\right)^2 \\
&\leq 2tW + 2t^2\gamma.
\end{aligned}$$

In the fourth line we used Cauchy-Schwarz inequality and in the last step, we used Assumption G3 about the total variance due to small terms being small, as well as the normalization of the variance in each pixel.

D.2 Proof of Lemma 14

As before, we use the probabilistic method. Suppose we fix some j. By Assumption G1, in G_σ, node x_j has either at least d positive neighbors or d negative ones. W.l.o.g., let us assume there are d negative neighbors. Let T be uniformly random subset of size t of these negative neighbors. By definition of G_σ, we have $\beta_{j,T} \leq -\sigma t$.

For $k \neq j$, let $f_{k,T}$ be the number of edges from x_k to T in graph G_τ. Using the same argument as in the proof of Lemma 4, we have $f_{k,T} \leq 4\log n$ w.h.p. for all such $k \neq j$. Thus $|\beta_{k,T}| \leq t\tau + f_{k,T}\Lambda \leq \sigma^2 t/(\Delta \log n)$. Thus it remains to bound $\nu_{-j,T}$.

We could apply Lemma 13 with $W = 4\log n \geq f_{k,T}$, and $S = [m] \setminus \{j\}$ on set T: we get $\nu_{-j,T}^2 \leq 2tW + 2t^2\gamma$. Recall that $\gamma = \sigma^2/3\Delta^2 \log n$ and thus $\nu_{-j,T} \leq \sigma t/\sqrt{\Delta \log n}$.

E Probability Inequalities

Lemma 18. *Suppose X is a bounded random variable in a normed vector space with $||X|| \leq M$. If event E happens with probability $1-\delta$ for some $\delta < 1$, then $|| \mathbb{E}[X|E] - \mathbb{E}[X]|| \leq 2\delta M$*

Proof. We have $\mathbb{E}[X] = \mathbb{E}[X|E]\Pr[E] + \mathbb{E}[X|E^c]\Pr[E^c] = \mathbb{E}[X|E] + (\mathbb{E}[X|E^c] - \mathbb{E}[X|E])\Pr[E^c]$, and therefore $|| \mathbb{E}[X|E] - \mathbb{E}[X]|| \leq 2\delta M$. □

Lemma 19. *Suppose X is a bounded random variable in a normed vector space with $||X|| \leq M$. If events E_1 and E_2 have small symmetrical differences in the sense that $\Pr[E_1|E_2] \leq \delta$ and $\Pr[E_2|E_1] \leq \delta$. Then $|| \mathbb{E}[X|E_1] - \mathbb{E}[X|E_2]|| \leq 4\delta M$.*

Proof. Let $Y = X|E_2$, by Lemma 18, we have $|| \mathbb{E}[Y|E_1] - \mathbb{E}[Y]|| \leq 2\delta M$, that is, $|| \mathbb{E}[X|E_1E_2] - \mathbb{E}[X|E_2]|| \leq 2\delta M$. Similarly $|| \mathbb{E}[X|E_1E_2] - \mathbb{E}[X|E_1]|| \leq 2\delta M$, and hence $|| \mathbb{E}[X|E_1] - \mathbb{E}[X|E_2]|| \leq 4\delta M$. □

Theorem 20 (Bernstein Inequality[Ber27] cf. [Ben62])**.** *Let $x_1, \ldots, x_n$ be independent variables with finite variance $\sigma_i^2 = \mathbb{V}[x_i]$ and bounded by M so that $|x_i - \mathbb{E}[x_i]| \leq M$. Let $\sigma^2 = \sum_i \sigma_i^2$. Then we have*

$$\Pr\left[\left|\sum_{i=1}^n x_i - \mathbb{E}[\sum_{i=1}^n x_i]\right| > t\right] \leq 2\exp(-\frac{t^2}{2\sigma^2 + \frac{2}{3}Mt})$$