Structured and Sparse Signal Estimation
Fundamental Limits and Error Bounds

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Akshay Soni

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Prof. Jarvis Haupt

May, 2015

Acknowledgements

The very old Chinese proverb *"The fragrance always remains in the hand that gives roses"*, very rightly captures my thoughts while writing this thesis. I would have never seen this day without the many who constantly guided, supported, loved and respected me. I wish to thank all of them from the bottom of my heart.

I begin by thanking god for providing me the ability, "India" for allowing me to dream globally and "The United States of America" for providing me the opportunity to accomplish my dreams.

I would like to thank my parents, whose love for each other and for the family inspired me at every step of life. In particular, I would like to thank my mother Mrs. Anjana Soni for her everlasting support and constant motivation, and my father Mr. Vijay Soni for teaching me the value of being organized and honest. I would never forget the sacrifices you have made for my future, and I would try my level best to keep you happy.

I would like to thank my sisters Aditi and Bharti for their love and for allowing me to share my emotions at hard times with them. I would also like to thank my brothers Gaurav and Alish for being my best friends for all these years and for all the memories we have made together.

Thank you my late sister Rekha – I am sure if you were with us, you would have been very happy with my achievements. I miss you!

I owe a major part of my gratitude to my Ph.D. advisor, Jarvis Haupt, for believing in me, for supporting me throughout my graduate school. I would always feel proud of being your Ph.D. student and more than that, your first Ph.D. student. I would consider myself extremely lucky if I can be half as down to earth and approachable as you are. Positivity, honesty and hard work are the traits which I have learnt from you

and I am sure they helped me in shaping my overall personality. Thanks for allowing me to work and grow together with you. I would also like to acknowledge and express my gratitude for financial support from Defense Advanced Research Projects Agency, National Science Foundation, and the University of Minnesota, Twin Cities.

I would like to thank Prof. Keshab Pahri for being my academic advisor during first year of school and for introducing me to Jarvis.

I would like to thank my committee members – Prof. Georgios Giannakis, Prof. Nikos Sidiropoulos, and Prof. Arindam Banerjee – for their time and consideration, and for the many fruitful suggestions to improve my quality of work.

I would like to thank my lab mates – Swayambhoo Jain, Sirisha Rambhatla, Xingguo Li, Mojtaba Kadkhodaie Elyaderani, Di Xiao, Scott Sievert and Alex Gutierrez – for the enlightening discussions and other fun activities. I would specially like to thank Swayambhoo and Sirisha for the much needed coffee breaks and the most entertaining political discussions (and sometimes work too!).

I would like to thank Dr. Fatih Porikli (MERL) for offering me an internship while we met during ICASSP 2012. It was a great learning experience and helped me to see the potentials of my research in practical scenarios.

I would also like to thank GraphLab for getting me in touch with Dr. Debora Donato (StumbleUpon). I would like to thank Debora for allowing me to work within her group as an intern. It was a great experience and I am looking forward to work with you in future.

I would like to thank Quresh, Gaurav, and Aritra for being great roommates and friends – there are a lot of memories with you that I will cherish for life. Thank you Ashutosh and Amrita for being the awesome couple they are – your love and support is truly a factor in this milestone of mine. Thanks to all my friends – Shraddha, Shivani, Prashant, Vikas, Vishal, Saunil, Sanika and Parita – for the good times that we shared in Minneapolis.

Finally, I would like to thank my best friend and my wife Tanvi Sharma, for her love and care, for accepting me as I am, and for her constant support. I am truly blessed to have you as my life partner.

Dedication

To my mother Mrs. Anjana Soni – eagerness of whom to see me succeed, kept me motivated to do hard work.

Abstract

Over the past decade, sparsity has become one of the most prevalent themes in signal processing and Big-Data applications. In general, sparsity describes the phenomenon where high-dimensional data can be explained by only a few variables, values, or coefficients. The presence of sparsity often enables efficient algorithms for extracting relevant information from the data. This effort focuses on the theoretical treatment of specialized sensing and inference techniques that exploit sparsity and other forms of structured low-dimensional representations.

The first part of this work focuses on noisy matrix estimation and completion problems. We consider the problem of estimating matrices that adhere to a "sparse-factor model" decomposition – matrices that may be accurately described by a product of two matrices, one of which is sparse – from noisy observations, where the noise is modeled as random and may arise from any of a number of various likelihood models (e.g., Gaussian, Poisson, Laplace, and even one-bit models). Sparse-factor models can be used to describe collections of vectors that reside in a union of linear subspaces, and can be viewed as a powerful generalization the widely-used principal component analysis technique, which assumes data reside on or near a single subspace. We establish estimation error guarantees for sparse-factor matrix estimation problems (where a noisy observation of each matrix entry is observed) and matrix completion problems (where only a subset of elements is observed, each corrupted by noise), and describe an efficient algorithm for performing inference in problems of this form.

In the second part of this work, we examine and quantify the benefits of "adaptive sensing" techniques, which employ data-dependent feedback in the data acquisition process, in the context of a structured sparse inference task. This work is motivated by a desire to formally exploit the structural characteristics and dependencies present in the wavelet representations of many natural images. We devise an efficient and provably-optimal (in a minimax sense) adaptive acquisition method for estimating the locations of the significant wavelet coefficients from noisy observations. Our results demonstrate the significant improvements that can be obtained when leveraging the inherent structural dependencies in the sparse representation of the signal to be acquired and incorporating feedback in the measurement process, relative to the best possible methods that utilize either structural information or adaptivity alone. Overall, our results provide essential new insights into the virtues of adaptive data acquisition in sparse inference tasks.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past decade, sparsity has become one of the most prevalent themes in signal processing and Big-Data applications. In general sparsity describes the phenomenon where high-dimensional data can be explained by only a few variables or values. The notion of sparsity enforces a small number of degrees of freedom in the otherwise high-dimensional data – potentially leading to efficient algorithms to extract relevant information without wasting expensive and often-scare resources.

Geometrically, sparsity forces a high-dimensional signal (or vector) to lie in a low-dimensional subspace of the high-dimensional space. To be precise, let a n-dimensional signal be k-sparse (i.e., only k out of its n entries are nonzero) in the canonical (i.e., identity) basis. The canonical basis vectors corresponding to the nonzero entries of the signal defines the basis for the k-dimensional subspace in which this n-dimensional signal resides. It is then clear that a k-sparse n-dimensional signal can belong to any one of the $\binom{n}{k}$ k-dimensional linear subspaces of $\mathbb{R}^n$. Any recovery procedure will have to allocate some resources for each of these subspaces. A small k (i.e., more sparsity) then would reduce the candidate number of subspaces, thus enabling us to recover the relevant information either by using less resources or more efficiently by allocating the available resources only to a small number of subspaces.

Structured sparsity refers to sparse representations that are drawn from a restricted union of subspaces where only a subset of $\binom{n}{k}$ subspaces are allowable. The advantages are clear – if we can exploit the structure information then we can use the available resources more judiciously and no allocation of resources would be required for the subspaces which are not allowed. Several procedures have been developed which takes the advantage of the underlying sparsity structure and provides accurate recovery with either less resources or efficient algorithms or both [128–130].

Active research areas like image processing, machine learning, speech processing, compressive sensing (CS), radar signal processing, genomics, network sciences, nuclear medicine, medical imagining and wireless communications, have all received a revived interest due to advances in sparsity guided state-of-the-art algorithms and applications. The focus has been on the theoretical treatment of techniques that exploit sparsity (or structured sparsity) in problems involving high-dimensional data, along with practical data acquisition methods – be it non-adaptive sampling like in CS or adaptive sampling (sampling with feedback) like in many recent works on adaptive CS [1–4]. This work makes a leap forward by studying implications of different kinds of structure (and sparsity) assumptions that can be imposed on the data along with their intimate interplay with corresponding adaptive sampling and recovery methods.

The first part of this work focuses on noisy matrix estimation and completion problems. We consider the problem of estimating matrices that adhere to a "sparse-factor model" decomposition – matrices that may be accurately described by a product of two matrices, one of which is sparse – from noisy observations, where the noise is modeled as random and may arise from any of a number of various likelihood models (e.g., Gaussian, Poisson, Laplace, and even one-bit models). Sparse-factor models can be used to describe collections of vectors that reside in a union of linear subspaces, and can be viewed as a powerful generalization the widely-used principal component analysis technique, which assumes data reside on or near a single subspace. We establish estimation error guarantees for sparse-factor matrix estimation problems (where a noisy observation of each matrix entry is observed) and matrix completion problems (where only a subset of elements is observed, each corrupted by noise), and describe an efficient algorithm for performing inference in problems of this form.

Specifically, in Chapter-2, we provide guarantees for the denoising problem where we get Poisson distributed samples of all the entries of the matrix. We formulate sparse and structured dictionary-based Poisson denoising problem as a constrained maximum likelihood estimation problem, and establish performance bounds for their mean-square estimation error using the framework of complexity penalized maximum likelihood analyses. Our results [5] provides theoretical foundations to existing *dictionary learning* based experimental Poisson denoising procedures.

A follow-on effort, extends this problem to the case of missing-data where we only get to observe a subset of entries of the matrix of interest. We extend our Poisson denoising analyses to missing data case and provide performance bounds for a variety of noise models. This work appears as Chapter-3.

The second part of this thesis studies the problem of support recovery (locations of the nonzero elements) of tree-sparse signals from noisy linear measurements. We propose a simple structured-adaptive support recovery procedure and provide sufficient condition on the signal amplitude in order to recover the exact support of a tree-sparse signal with high probability. We further establish fundamental performance limits for the task of support recovery of tree-sparse signals from noisy measurements, in settings where measurements may be obtained either non-adaptively (using a randomized Gaussian measurement strategy motivated by initial CS investigations) or by any adaptive sensing strategy. Our main results imply that the proposed adaptive tree sensing procedure is nearly optimal, in the sense that no other sensing and estimation strategy can perform fundamentally better for identifying the support of tree-sparse signals. This work establishes that the combination of structure and adaptivity is a powerful one and for some structures (like tree) both the necessary and sufficient conditions are independent of ambient dimension and only depends on the sparsity level. This work appears as Chapter-4.

In Chapter-5, several future directions are discussed along with some concluding remarks.

Advances in Structured Matrix Recovery

This part contains two contributions to the matrix recovery problem. Our specific focus is on settings where the matrix to be estimated is well-approximated by a product of two (a priori unknown) matrices, one of which is sparse. Such structural models – referred to here as “sparse factor models” – have been widely used, for example, in subspace clustering applications, as well as in contemporary sparse modeling and dictionary learning tasks.

Chapter-2 is a reprint of our IEEE International Symposium on Information Theory paper [5]. This work provides a theoretical foundation for the experimentally studied Poisson denoising tasks using dictionary learning approach [6, 7], where underlying structural assumption on data is of a sparse factor model. Specifically, we formulate sparse and structured dictionary-based Poisson denoising methods as constrained maximum likelihood estimation strategies, and establish performance bounds for their mean-square estimation error using the framework of complexity penalized maximum likelihood analysis.

Chapter-3 examines a general class of noisy matrix completion tasks where the goal is to estimate a matrix with sparse factor model from observations obtained at a subset of its entries, each of which is subject to random noise or corruption. Our main theoretical contributions are estimation error bounds for sparsity-regularized maximum likelihood estimators for problems of this form, which are applicable to a number of different observation noise or corruption models. This work has been submitted to the IEEE Transactions on Information Theory, and is currently under review [8].

Chapter 2

Estimation Error Guarantees for Poisson Denoising with Sparse and Structured Dictionary Models

Poisson processes are commonly used models for describing discrete arrival phenomena arising, for example, in photon-limited scenarios in low-light and infrared imaging, astronomy, and nuclear medicine applications. In this context, several recent efforts have evaluated Poisson denoising methods that utilize contemporary sparse modeling and dictionary learning techniques designed to exploit and leverage (local) shared structure in the images being estimated. This work establishes a theoretical foundation for such procedures. Specifically, we formulate sparse and structured dictionary-based Poisson denoising methods as constrained maximum likelihood estimation strategies, and establish performance bounds for their mean-square estimation error using the framework of complexity penalized maximum likelihood analyses.[1]

[1] The material in Chapter-2 is ©2014 IEEE. Reprinted, with permission, from *IEEE International Symposium on Information Theory*, "Estimation Error Guarantees for Poisson Denoising with Sparse and Structured Dictionary Models," A. Soni and J. Haupt.

2.1 Introduction

Across a broad range of engineering application domains, Poisson processes have been utilized to describe discrete event or arrival phenomena. For example, in a host of imaging applications (including infrared and thermal imaging, night vision, astronomical imaging, and nuclear medicine, to name a few) the random arrival of photons at each detector in an array may be modeled using Poisson-distributed random variables, with unknown rates or intensities. A fundamental problem in these applications is that of estimating the unknown rates associated with each of the sources, a task typically referred to as *Poisson denoising*.

We consider here a denoising task along these lines. Suppose that we are equipped with a collection of detectors, and that the arrival of photons at each individual detector may be accurately described by a Poisson process with some unknown (non-negative) rate. At each detector we acquire a single integer-valued observation, corresponding to the number of photons arriving at the detector over some fixed (but not necessarily specified) time interval that we assume to be the same across all detectors. It follows that the observation at each detector is a Poisson-distributed random variable whose parameter is the product of the underlying rate parameter of the process and the length of the time interval (see, e.g., [9]). We assume the Poisson processes giving rise to the observations at each detector are mutually independent.

Suppose that there are a total of d detectors. For each $\ell \in [d]$ where $[d]$ is shorthand for the set $\{1, 2, \ldots, d\}$, we denote the Poisson-distributed observation at the ℓ-th detector as y_ℓ and denote by x_ℓ^* its unknown parameter. Letting $\mathrm{Poi}(y_\ell|x_\ell^*) = (x_\ell^*)^{y_\ell}\exp(-x_\ell^*)/(y_\ell)!$ denote the univariate Poisson probability mass function (pmf) defined on nonnegative integers $y_\ell \in \mathbb{N}_0$, we may write the joint pmf of the d observations, defined on $\mathbb{N}_0^d$, as

$$p(\{y_\ell\}_{\ell\in[d]}|\{x_\ell^*\}_{\ell\in[d]}) = \prod_{\ell\in[d]} \mathrm{Poi}(y_\ell|x_\ell^*), \tag{2.1}$$

where the product form on the right-hand side follows from our independence assumption on the individual Poisson processes.

2.1.1 Exploiting Data Structure in Poisson Denoising Tasks

In the absence of any structural dependencies among the collection of rates $\{x_\ell^*\}_{\ell\in[d]}$, the Poisson denoising task is somewhat trivial – in this case, classical estimation theoretic analyses establish that each observation is itself the minimum variance unbiased estimator of its underlying parameter (see, e.g., [10]). More interesting approaches to the denoising task, then, seek to exploit some form of underlying structure among the individual rates. Efforts along these lines include [11, 12], which proposed and analyzed estimation strategies applicable in scenarios where the collection of rates (appropriately arranged) admits a simple representation in terms of a wavelet representation, and [13], which also examined multiresolution representations of the collection of rates. Along similar lines, the work [14] analyzed estimation procedures tailored to signals that are sparse (or nearly so) in any orthonormal basis, within the context of a compressed sensing approach to the Poisson denoising problem.

A number of related efforts have examined Poisson denoising tasks using data representations or bases that are learned from the data themselves, in contrast to the efforts described above that utilize fixed bases or representations. Such "data-driven" estimation strategies include Poisson-specific extensions of classical methods like principal component analysis and other matrix factorization methods [15, 16], as well as application of contemporary ideas from sparse dictionary learning [17–19] to Poisson-structured data [20]. We note, in particular, the recent works [7] and [6], which describe estimation tasks employing models that may be described as sparse or structured dictionary-based models; our effort here is motivated by a desire to provide theoretical justification for these dictionary-based techniques.

2.1.2 Our Approach

The sparse and structured dictionary-based models upon which our analyses are based describe underlying data structure in terms of matrix factorization models. To that end, we will find it useful here to formulate our model so that the collection of d observations are interpreted as elements of an $m \times n$ matrix (with $d = mn$) denoted by $\mathbf{Y}$, and having elements $Y_{i,j}$, where for $i \in [m]$ and $j \in [n]$, $Y_{i,j}$ is a Poisson random variable with rate $X_{i,j}^*$. Letting $\mathbf{X}^*$ be the $m \times n$ matrix with entries $X_{i,j}^*$, we overload (slightly) the notation in (2.1), and write the joint pmf of the observations in this case as

$$p(\mathbf{Y}|\mathbf{X}^*) = \prod_{i\in[m],j\in[n]} \mathrm{Poi}(Y_{i,j}|X^*_{i,j}) \triangleq \mathrm{Poi}(\mathbf{Y}|\mathbf{X}^*). \tag{2.2}$$

Our interest here is primarily on settings where the matrix $\mathbf{X}^*$ admits a dictionary-based factorization, so that $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$, where $\mathbf{D}^* \in \mathbb{R}^{m\times p}$ and $\mathbf{A}^* \in \mathbb{R}^{p\times n}$. Since such factorization models are themselves fairly general, we restrict our attention here to two specific settings – the first being when the matrix $\mathbf{A}^*$ is sparse so that only a small fraction of its elements are nonzero (along the lines of models employed in dictionary learning efforts), and the second when p, the number of columns of $\mathbf{D}^*$ and rows of $\mathbf{A}^*$, is small relative to m and n (in which case $\mathbf{X}^*$ admits a *low-rank* decomposition). That said, the analytical approach we develop here is fairly general, and thus may readily be extended to other factorization models (e.g., non-negative matrix factorization, structured dictionary models, etc.).

The estimation approaches we analyze here amount to constrained maximum likelihood estimation procedures. Abstractly, we consider a set $\mathcal{X}$ of candidate estimates $\mathbf{X}$ for $\mathbf{X}^*$, each of which admits a factorization of the form $\mathbf{X} = \mathbf{D}\mathbf{A}$. The elements of the factors $\mathbf{D}$ and $\mathbf{A}$ may themselves be constrained to enforce the type of structure that we assume present in $\mathbf{X}^*$. Formally, we construct sets C_{D} and C_{A} and a set

$$\mathcal{X} \triangleq \left\{\mathbf{X} = \mathbf{D}\mathbf{A} \ : \mathbf{D} \in C_{\mathrm{D}},\ \mathbf{A} \in C_{\mathrm{A}},\ \max_{i,j} |X_{i,j}| \leq \mathrm{X}_{\max}\right\}$$

where $0 < \mathrm{X}_{\max} < \infty$ is a constant that describes the maximum rate of the underlying processes (and whose specific role will become evident in our analysis), and we consider estimates $\widehat{\mathbf{X}}$ of $\mathbf{X}^*$ constructed according to

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathcal{X}} -\log p(\mathbf{Y}|\mathbf{X}) + \lambda\ \mathrm{pen}(\mathbf{X}), \tag{2.3}$$

where $\mathrm{pen}(\mathbf{X})$ is a non-negative penalty that quantifies the inherent "complexity" of each estimate $\mathbf{X} \in \mathcal{X}$, and $\lambda > 0$ is a user-specified regularization parameter. For both the low-rank and the sparse dictionary based models we consider here, we describe the construction of suitable sets C_{D} and C_{A}, cast each corresponding estimation procedure in terms of an optimization of the form (2.3) (with appropriately constructed penalties), and derive mean-square estimation error rates using analysis techniques motivated by those employed in [13, 14, 21–26].

2.1.3 Related Efforts in Poisson Restoration

While our focus here is on Poisson denoising, we briefly note several related efforts that examine restoration and deblurring methods for Poisson-distributed data [27–31]. These works employ regularized maximum likelihood estimation strategies similar in form to those we analyze in this effort. More recently, [32] proposed a dictionary-based approach to the Poisson deblurring task.

2.1.4 Organization and Notation

The remainder of this paper is organized as follows. We present our main theoretical results, stated in terms of the estimation procedures proposed in [6, 7], in Section 2.2, and provide proofs in Section 2.3. In Section 2.4 we briefly discuss how our analytical approach overcomes somewhat limiting minimum rate assumptions inherent in several prior works that use penalized maximum likelihood methods for Poisson denoising. In Section 2.5, we conclude with a discussion of potential extensions of our analysis.

A brief note on notation employed in the sequel – for a matrix $\mathbf{A}$, we denote its number of nonzero elements by $\|\mathbf{A}\|_0$, the sum of absolute values of its elements by $\|\mathbf{A}\|_1$, and its dimension (the product of its row and column dimensions) by $\dim(\mathbf{A})$. For an integer $m \in \mathbb{N}$, the notation $\mathbf{1}_m$ denotes an all-ones length m column vector.

2.2 Main Results

As noted above, our analyses here are motivated by recent efforts ([6, 7]) that examine Poisson denoising tasks arising in imaging problems and provide empirical evaluations of procedures that exploit local shared structure in the rates being estimated. These prior works each utilize "patch-level" structural models for the underlying image, in which the shared structure arises in terms of factorizations of matrices comprised of vectorized versions of small image patches.

The first procedure proposed in [7] is a non-local variant of a principal component analysis (PCA) method. That approach uses an initial clustering step designed to identify collections of similar patches, then obtains estimates of the underlying rate functions of the image by performing low-rank factorizations of patch-level matrix representations of each data cluster. In terms of our model here, the approximation step inherent to this approach may be described by assuming the true matrix of rates $\mathbf{X}^* \in \mathbb{R}^{m\times n}$ giving rise to independent Poisson-distributed observations $\mathbf{Y}$ in each data cluster admits a decomposition of the form $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$, where $\mathbf{D}^* \in \mathbb{R}^{m\times p}$ and $\mathbf{A}^* \in \mathbb{R}^{p\times n}$ for some $p \leq \min(m,n)$.

Both [7] and [6] also examine sparse dictionary-based denoising methods along the lines of recent efforts in the dictionary learning literature (see, e.g. [19]), which seek to model the image patches as sparse linear combinations of columns of a learned dictionary matrix. Here, this model assumes that the true rate matrix $\mathbf{X}^*$ admits a decomposition of the form $\mathbf{D}^*\mathbf{A}^*$ where $\mathbf{A}^*$ is sparse (e.g., having fewer than some $k_{\max}$ non zeros per column). Sparse dictionary-based models may be interpreted as a natural extension of low-rank models; the latter essentially fits the data to a single low-dimensional linear subspace, while the former utilizes a union of linear subspaces.

Our main results establish mean square error guarantees for estimates for these tasks that are obtained via penalized maximum likelihood estimation strategies. In order to state our results, we need to first construct a set $\mathcal{X}$ of candidate reconstructions, with appropriate penalties. To that end, we fix parameters $\mathrm{A}_{\max} > 0$, and $\mathrm{X}_{\max} > 0$, and $\lambda' > 1$, let q be a positive integer satisfying

$$q \geq \max\left\{4, 3 + \log\left(\frac{18\mathrm{A}_{\max}}{\lambda'\log(2)}\right), 1 + \log\left(\frac{36\mathrm{A}_{\max}}{\mathrm{X}_{\max}}\right)\right\}, \tag{2.4}$$

and let L be the smallest integer exceeding $(\max(m,n))^q$. Now, for any positive integer $p \leq \min(m,n)$ we let $\mathcal{X}$ be the set of candidate reconstructions of the form $\mathbf{X} = \mathbf{DA}$ satisfying $\max_{i,j}|X_{i,j}| \leq X_{\max}$, where $\mathbf{D} \in \mathcal{D}$ are in $\mathbb{R}^{m\times(p+1)}$ and $\mathbf{A} \in \mathcal{A}$ are in $\mathbb{R}^{(p+1)\times n}$, so that each entry of $\mathbf{D}$ takes values on one of L uniformly-spaced quantization levels in the range $[-1,1]$ and each element of $\mathbf{A}$ takes on one of L possible uniformly spaced quantization levels in the range $[-\mathrm{A}_{\max},\ \mathrm{A}_{\max}]$.

Our first result, stated here as a theorem, pertains to sparse dictionary-based models.

Theorem 2.2.1. *Let the true rate matrix* $\mathbf{X}^*$ *be* $m \times n$*, where* $\max(m,n) \geq 3$*. Suppose* $\mathbf{X}^*$ *satisfies the constraint* $\max_{i,j} X^*_{i,j} < \mathrm{X}_{\max}/2$*, and admits a dictionary-based decomposition of the form* $\mathbf{D}^*\mathbf{A}^*$*, where the dictionary* $\mathbf{D}^*$ *is* $m \times p$ *for* $p < n$ *with entries bounded in magnitude by* 1*, and the coefficient matrix* $\mathbf{A}^*$ *is* $p \times n$ *whose elements are* *bounded in magnitude by* $\mathrm{A}_{\max}$. *Let observations* $\mathbf{Y}$ *of* $\mathbf{X}^*$ *be acquired according to the model* (2.2).

Form the set $\mathcal{X}$ *as above, and let* $\mathrm{pen}(\mathbf{X}) = [q{\cdot}\mathrm{dim}(\mathbf{D}){+}(q{+}2){\cdot}\|\mathbf{A}\|_0]{\cdot}\log(\max(m,n))$. *The estimate* $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}(\mathbf{Y}) = \widehat{\mathbf{D}}\widehat{\mathbf{A}}$ *formed using the solution of the penalized maximum likelihood problem*

$$\{\widehat{\mathbf{D}}, \widehat{\mathbf{A}}\} = \arg \min_{\mathbf{D}\in C_{\mathrm{D}}, \mathbf{A}\in C_{\mathrm{A}}:\mathbf{DA}\in\mathcal{X}} -\log p(\mathbf{Y}|\mathbf{DA}) + \lambda\|\mathbf{A}\|_0, \tag{2.5}$$

with $\lambda = \lambda'{\cdot}(q{+}2){\cdot}\log(\max(m,n))\log(2)$ *(and where* λ' *is as specified in the construction of* $\mathcal{X}$*) satisfies*

$$\begin{aligned}&\frac{\mathbb{E}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{mn} \\ &\qquad \preceq \lambda' \mathrm{X}_{\max}\left(\frac{m(p+1)}{mn} + \frac{\|\mathbf{A}^*\|_0 + n}{mn}\right)\log(\max(m,n)).\end{aligned}$$

Here, the expectation is with respect to the distribution of $\mathbf{Y}$ *parameterized by the true rate matrix* $\mathbf{X}^*$, *and the notation* $\preceq$ *suppresses leading (finite, positive) constants.*

The salient take-away point here is that the average per-element estimation error is upper bounded by a term that decays essentially in proportion to the number of "degrees of freedom" in the model divided by the number of observations. In other words, our result here establishes that the estimation error exhibits characteristics of the well-known parametric rate.

The result of Thm. 2.2.1 also provides guidance on when dictionary-based estimation procedures are viable. Consider, for example, a setting where the true matrix $\mathbf{A}^*$ in the dictionary-based decomposition of $\mathbf{X}^*$ has some $k_{\max}$ nonzero elements per column. Here, Theorem 2.2.1 establishes that the mean-square estimation error for estimating $\mathbf{X}^*$ decays in proportion to $(p+1)/n + (k_{\max}+1)/m$, ignoring leading constants and logarithmic factors. This result implies natural conditions on the estimation task – that accurate estimation is possible when the number of columns of $\mathbf{X}^*$ exceeds (by a multiplicative constant times a factor logarithmic in the dimension) the number of true dictionary elements p, and the number of rows of $\mathbf{X}^*$ exceeds (by a multiplicative constant times a factor logarithmic in the dimension) the number of non zeros in the sparse representation of each column. This latter condition is reminiscent of conditions arising in compressive sensing (see, e.g., [26, 33, 34]).

We obtain an analogous result for the case where the true rate matrix $\mathbf{X}^*$ admits a low-rank decomposition. We state the result here as a corollary of Theorem 2.2.1.

Corollary 2.2.1. *Suppose that* $\max(m,n) \geq 3$*, and that the true rate matrix* $\mathbf{X}^* \in \mathbb{R}^{m\times n}$ *admits a low-rank decomposition, so that it may be written as* $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$*, where* $\mathbf{D}^*$ *is* $m \times p$ *and* $\mathbf{A}^*$ *is* $p \times n$ *with* $p \leq \min(m,n)$*, and such that* $X^*_{i,j} \leq \mathrm{X}_{\max}/2,\ \forall i,j$*. Let observations* $\mathbf{Y}$ *be acquired via the model* (2.2)*. Form the set* $\mathcal{X}$ *as above, and let* $\mathrm{pen}(\mathbf{X}) = [q\cdot \dim(\mathbf{D}) + (q+2)\cdot\dim(\mathbf{A})]\cdot\log(\max(m,n))$.

The estimate $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}(\mathbf{Y}) = \widehat{\mathbf{D}}\widehat{\mathbf{A}}$ *formed using the solution of the following penalized maximum likelihood problem*

$$\{\widehat{\mathbf{D}},\widehat{\mathbf{A}}\} = \arg\min_{\mathbf{D}\in C_{\mathrm{D}},\mathbf{A}\in C_{\mathrm{A}}:\mathbf{DA}\in\mathcal{X}} -\log p(\mathbf{Y}|\mathbf{DA}), \tag{2.6}$$

satisfies

$$\frac{\mathbb{E}\left[\|\mathbf{X}^*-\widehat{\mathbf{X}}\|_F^2\right]}{mn} \preceq \lambda'\mathrm{X}_{\max}\left(\frac{(p+1)(m+n)}{mn}\right)\log(\max(m,n)),$$

where as above the expectation is with respect to the distribution of $\mathbf{Y}$ *parameterized by the true rate matrix* $\mathbf{X}^*$*, and the notation* $\preceq$ *suppresses leading (finite) constants.*

Note that in this case the penalty $\mathrm{pen}(\mathbf{X})$ is actually the same for all $\mathbf{X}\in\mathcal{X}$, as it depends only on the dimensions of the two factors, which are the same for all candidates by construction of $\mathcal{X}$. Thus, the estimation approach here reduces to just a maximum likelihood estimation over constrained sets. As above, the estimation error rate exhibits characteristics of the parametric rate, as the low-rank model here has $\mathcal{O}(p(m+n))$ degrees of freedom.

2.3 Proofs of Main Results

We write $p_{X_{i,j}}(\cdot)$ as shorthand for the scalar Poisson pmf with rate $X_{i,j}$, and we denote the multivariate Poisson pmf $p(\cdot|\mathbf{X})$ defined in (2.2) (parameterized by the collection of rates $\{X_{i,j}\}_{i,j}$) by $p_{\mathbf{X}}(\cdot)$.

Central to our analysis will be the aforementioned countable sets $\mathcal{X}$ of candidate reconstructions of the unknown (non-negative) rate matrix $\mathbf{X}^*$. We consider sets $\mathcal{X}$ constructed as above, and assign to each $\mathbf{X}\in\mathcal{X}$ a non-negative "penalty" quantity denoted by pen($\mathbf{X}$) (which here will quantify the "complexity" of the corresponding estimate), so that the collection of penalties satisfies the summability condition $\sum_{\mathbf{X}\in\mathcal{X}} 2^{-\mathrm{pen}(\mathbf{X})} \leq 1$. Note that this condition is just the Kraft-McMillan inequality; in constructing penalties for elements of $\mathcal{X}$ we will employ the well-known fact that the Kraft-McMillan inequality is satisfied provided we may construct a *uniquely decodable code* for the elements $\mathbf{X} \in \mathcal{X}$; see [35]. With this, we begin by establishing a fundamental result, from which our results follow.

Lemma 2.3.1. *Suppose that the elements of the unknown non-negative rate matrix* $\mathbf{X}^*$ *are bounded in amplitude, so that for some fixed* $\mathrm{X}_{\max} > 0$*, we have* $0 \leq X^*_{i,j} \leq \mathrm{X}_{\max}/2$ *for all* $i \in [m]$ *and* $j \in [n]$*. Let* $\mathcal{X}$ *be a countable set of candidate solutions* $\mathbf{X}$ *satisfying the uniform bound* $\max_{i\in[m],j\in[n]}|X_{i,j}| \leq X_{\max}$*, with associated non-negative penalties* $\{\mathrm{pen}(\mathbf{X})\}_{\mathbf{X}\in\mathcal{X}}$ *satisfying the Kraft-McMillan inequality as stated above. Collect a total of* mn *independent Poisson measurements* $\mathbf{Y} = \{Y_{i,j}\}_{i\in[m],j\in[n]}$*, parameterized by* $\mathbf{X}^*$*, according to the model* (2.2)*. If there exists* $\mathbf{X}^+ \in \mathcal{X}$ *such that* $X^+_{i,j} - X^*_{i,j} \geq 0$ *for all* $i \in [m]$ *and* $j \in [n]$*, then for any choice of* $\lambda' > 1$*, the complexity penalized maximum likelihood estimate*

$$\widehat{\mathbf{X}} = \arg \min_{\mathbf{X}\in\mathcal{X}} \{-\log p\left(\mathbf{Y}|\mathbf{X}\right) + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X})\}, \tag{2.7}$$

satisfies,

$$\frac{\mathbb{E}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{mn} \\ \leq \frac{4\mathrm{X}_{\max}}{mn}\left[\|\mathbf{X}^* - \mathbf{X}^+\|_1 + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X}^+)\right], \tag{2.8}$$

where the expectation is taken with respect to the distribution of $\mathbf{Y} \sim p_{\mathbf{X}^*}$*.*

Proof. Our proof utilizes a straight-forward extension of a result stated and utilized in [13,14] (based on the essential ideas of [24,36]), which we provide here without proof: for any $\lambda' > 1$ the complexity regularized maximum likelihood solution $\widehat{\mathbf{X}}$ of the form (2.7), obtained by optimizing over any countable set $\mathcal{X}$ of candidates having penalties $\{\mathrm{pen}(\mathbf{X})\}_{\mathbf{X}\in\mathcal{X}}$ satisfying the Kraft-McMillan inequality, satisfies

$$-2\mathbb{E} \log \mathrm{A}(p_{\mathbf{X}^*}, p_{\widehat{\mathbf{X}}}) \leq \min_{\mathbf{X}\in\mathcal{X}} \left[\mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}}) + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X})\right], \tag{2.9}$$

where the expectation is with respect to the distribution of $\mathbf{Y} \sim p_{\mathbf{X}^*}$. Here,

$$\mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}}) \triangleq \sum_{\mathbf{Y} \in \mathbb{N}_0^{m \times n}} \log\left(\frac{p(\mathbf{Y}|\mathbf{X}^*)}{p(\mathbf{Y}|\mathbf{X})}\right) p(\mathbf{Y}|\mathbf{X}^*)$$

denotes the *Kullback-Leibler divergence* (or KL divergence)[2] of $p_{\mathbf{X}}$ from $p_{\mathbf{X}^*}$, and the quantity

$$\mathrm{A}(p_{\mathbf{X}^*}, p_{\widehat{\mathbf{X}}}) \triangleq \sum_{\mathbf{Y} \in \mathbb{N}_0^{m \times n}} \sqrt{p(\mathbf{Y}|\mathbf{X}^*) \cdot p(\mathbf{Y}|\widehat{\mathbf{X}})}$$

is the *Hellinger Affinity* between $p_{\mathbf{X}^*}$ and $p_{\widehat{\mathbf{X}}}$. Now, since the upper bound in (2.9) holds for $\mathbf{X} \in \mathcal{X}$ which achieves the minimum, it holds for all $\mathbf{X} \in \mathcal{X}$. Considering, specifically, the estimator $\mathbf{X}^+ \in \mathcal{X}$, we have

$$-2\mathbb{E} \log \mathrm{A}(p_{\mathbf{X}^*}, p_{\widehat{\mathbf{X}}}) \leq \mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}^+}) + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X}^+). \tag{2.10}$$

Specializing to the Poisson case, we use the results of Lemmas 2.3.2 and 2.3.3 (in Section 2.3.3) to obtain, respectively, a lower bound for the left-hand side and an upper bound for the right-hand side of (2.10). The result follows. □

Our main results of Section 2.2 follow from specializing this result to each of the two structural models. We establish first a proof of the sparse dictionary-based inference estimation procedure; the analogous result for estimation in low-rank models follows as a simple corollary.

2.3.1 Proof of Theorem 2.2.1

The proof of our first main result follows directly from Lemma 2.3.1 above. First, note that each candidate estimate $\mathbf{X} = \mathbf{DA} \in \mathcal{X}$ may be described via a code, in which each element of $\mathbf{D}$ is encoded using $\log(L) = q \log(\max(m, n))$ bits and each nonzero element of $\mathbf{A}$ is encoded using $\log(\dim(\mathbf{A}))$ bits to denote its location, and $\log(L)$ bits for its amplitude. Thus, a total of $q \cdot \dim(\mathbf{D}) \cdot \log(\max(m, n))$ bits suffice to encode $\mathbf{D}$, and since $\log(\dim(\mathbf{A})) < \log(\max(m, n)^2)$, matrices $\mathbf{A}$ having $\|\mathbf{A}\|_0$ nonzero entries

[2] Note that the KL divergence is only well-defined here for non-negative $\mathbf{X}$, when the corresponding Poisson pmf $p(\mathbf{Y}|\mathbf{X})$ is well-defined. We make no specific constraint here that each $\mathbf{X} \in \mathcal{X}$ be non-negative, but without loss of generality we may take $\mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}})$ to be infinite also when $\mathbf{X}$ has any non-negative entries. Further, note the KL divergence is infinite here if for any i, j, $X_{i,j} = 0$ but $X^*_{i,j} \neq 0$ (i.e., when the distribution $p_{\mathbf{X}*}$ is not absolutely continuous with respect to $p_{\mathbf{X}}$).

can be described using no more than $\|\mathbf{A}\|_0 \cdot (q+2) \cdot \log(\max(m,n))$ bits. Overall, this implies we may choose $\text{pen}(\mathbf{X}) = q \cdot \dim(\mathbf{D}) \cdot \log(\max(m,n)) + \|\mathbf{A}\|_0 \cdot (q+2) \cdot \log(\max(m,n))$. Note that while constructing the codes we did not care about the uniform bounded condition (i.e., that each entry should be bounded by $\mathrm{X}_{\max}$); in effect, we formed uniquely decodable codes for a bigger set $\mathcal{X}'$ such that $\mathcal{X} \subseteq \mathcal{X}'$, so we always have $\sum_{\mathbf{X}\in\mathcal{X}} 2^{-\text{pen}(\mathbf{X})} \leq \sum_{\mathbf{X}\in\mathcal{X}'} 2^{-\text{pen}(\mathbf{X})} \leq 1$.

Now, consider a candidate reconstruction of the form $\mathbf{X}_Q = \mathbf{D}_Q\mathbf{A}_Q + \mathbf{1}_m(\alpha\mathbf{1}_n^T) \triangleq \tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q$, where $\mathbf{D}_Q$ and $\mathbf{A}_Q$ are the closest quantized surrogates of the true parameters $\mathbf{D}^*$ and $\mathbf{A}^*$, and $0 \leq \alpha \leq A_{\max}$ is a quantity to be specified. Denote $\mathbf{D}_Q = \mathbf{D}^* + \triangle_{\mathbf{D}}$ and $\mathbf{A}_Q = \mathbf{A}^* + \triangle_{\mathbf{A}}$, where $\triangle_{\mathbf{D}}$ and $\triangle_{\mathbf{A}}$ are the quantization error matrices. Then, it is easy to see that

$$\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^* = \mathbf{1}_m(\alpha\mathbf{1}_n^T) + \mathbf{D}^*\triangle_{\mathbf{A}} + \triangle_{\mathbf{D}}\mathbf{A}^* + \triangle_{\mathbf{D}}\triangle_{\mathbf{A}}. \tag{2.11}$$

To satisfy the conditions of Lemma 2.3.1, we must have that $\mathbf{X}_Q$ overestimates (element-wise) the true rate matrix, and that the right-hand side of (2.11) be no larger than $\mathrm{X}_{\max}/2$. To that end, our aim is to choose α so that the right side of (2.11) becomes element-wise nonnegative, but no larger than $\mathrm{X}_{\max}/2$. It is straightforward to see that each entry of the matrices $\mathbf{D}^*\triangle_{\mathbf{A}}$ and $\triangle_{\mathbf{D}}\mathbf{A}^*$ is bounded in magnitude by $2p\mathrm{A}_{\max}/L$. Also, the elements of the matrix $\triangle_{\mathbf{D}}\triangle_{\mathbf{A}}$ are bounded in magnitude by $4p\mathrm{A}_{\max}/L^2 \leq 4p\mathrm{A}_{\max}/L$. Thus, it suffices to choose α as the smallest quantization level exceeding $8p\mathrm{A}_{\max}/L$ to ensure the each element of the matrix on the right-hand side of (2.11) is nonnegative. Since we choose α to be the higher quantization level of $8p\mathrm{A}_{\max}/L$, and the quantization levels for elements of $\mathbf{A}$ are of size $2\mathrm{A}_{\max}/L$, we have that $\alpha \leq (8p+2)\mathrm{A}_{\max}/L$. In order for α to be a valid entry of $\mathbf{A}$, it must be bounded by $\mathrm{A}_{\max}$, which is true whenever $L \geq (8p+2)$.

We can now bound each entry of $\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^*$ as follows

$$\begin{aligned}
&(\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^*)_{i,j} \\
&\quad = \quad (\mathbf{1}_m(\alpha\mathbf{1}_n)^T + \mathbf{D}^*\triangle_{\mathbf{A}} + \triangle_{\mathbf{D}}\mathbf{A}^* + \triangle_{\mathbf{D}}\triangle_{\mathbf{A}})_{i,j} \\
&\quad \leq \quad \frac{(8p+2)\mathrm{A}_{\max}}{L} + \frac{2p\mathrm{A}_{\max}}{L} + \frac{2p\mathrm{A}_{\max}}{L} + \frac{4p\mathrm{A}_{\max}}{L} \\
&\quad = \quad \frac{16p\mathrm{A}_{\max}}{L} + \frac{2\mathrm{A}_{\max}}{L} \leq \frac{18p\mathrm{A}_{\max}}{L},
\end{aligned}$$

where the second inequality follows from bounds on the entries of each matrix mentioned above and the last inequality is valid for $p \geq 1$. This quantity is no larger than $\mathrm{X}_{\max}/2$ whenever $L \geq 36p\mathrm{A}_{\max}/\mathrm{X}_{\max}$, and in this case, we ensure that $\mathbf{X}_Q \in \mathcal{X}$.

Now, note that $\|\mathbf{X}^* - \mathbf{X}_Q\|_1 = \sum_{i\in[m],j\in[n]}(\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^*)_{i,j} \leq 18p\cdot(mn)\cdot\mathrm{A}_{\max}/L$, and if we now evaluate the oracle bound (2.8) from Lemma 2.3.1 at the candidate $\mathbf{X}_Q$ which overestimates $\mathbf{X}^*$ (entry-wise), we have

$$\begin{aligned}
&\frac{\mathbb{E}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{mn} \\
&\qquad\leq \frac{4\mathrm{X}_{\max}}{mn}\left[\|\mathbf{X}^* - \mathbf{X}_Q\|_1 + \lambda'\log(2)\cdot\mathrm{pen}(\mathbf{X}_Q)\right] \\
&\qquad\leq \frac{72pX_{\max}\mathrm{A}_{\max}}{L} + \lambda'\cdot 4\log(2)\mathrm{X}_{\max}\cdot\frac{\mathrm{pen}(\mathbf{X}_Q)}{mn} \\
&\qquad\leq \lambda'\cdot 8\log(2)\mathrm{X}_{\max}\cdot\frac{\mathrm{pen}(\mathbf{X}_Q)}{mn},
\end{aligned}$$

where the last line follows whenever $L \geq \frac{18\mathrm{A}_{\max}mnp}{\lambda'\log(2)}$ (since $\mathrm{pen}(\mathbf{X}_Q)$ corresponds to a binary code having length greater than 0, we have $\mathrm{pen}(\mathbf{X}_Q) \geq 1$).

Overall, then, the result follows since by construction, we have $\dim(\widetilde{\mathbf{D}}_Q) \leq mp + m$, and $\|\widetilde{\mathbf{A}}_Q\|_0 \leq \|\mathbf{A}^*\|_0 + n$, and the assumption (2.4) implies

$$L \geq \max\left\{8p+2, \frac{18\mathrm{A}_{\max}mnp}{\lambda'\log(2)}, \frac{36p\mathrm{A}_{\max}}{\mathrm{X}_{\max}}\right\}.$$

2.3.2 Proof of Corollary 2.2.1

The proof of Corollary 2.2.1 follows directly from the proof of Theorem 2.2.1 – in particular, by substituting $\|\mathbf{A}^*\|_0 = pn$.

2.3.3 Useful Lemmata

The following lemmata are used in the proof of Lemma 2.3.1.

Lemma 2.3.2 (From [14])**.** *For any two (non-negative) Poisson rate matrices* $\mathbf{X}^a$ *and* $\mathbf{X}^b$*, having entries uniformly bounded above by* $\mathrm{X}_{\max}$*, we have*

$$\frac{1}{4\mathrm{X}_{\max}}\|\mathbf{X}^a - \mathbf{X}^b\|_F^2 \leq -2\cdot\log\mathrm{A}(p_{\mathbf{X}^a}, p_{\mathbf{X}^b}).$$

Lemma 2.3.3. *For non-negative Poisson rate matrices* $\mathbf{X}^a$ *and* $\mathbf{X}^b$ *such that* $\mathbf{X}^b$ *overestimates* $\mathbf{X}^a$ *element-wise i.e.,* $X^b_{i,j} - X^a_{i,j} \geq 0$ *for all* $i \in [m]$ *and* $j \in [n]$*, we have* $\mathrm{K}(p_{\mathbf{X}^a}, p_{\mathbf{X}^b}) \leq \|\mathbf{X}^b - \mathbf{X}^a\|_1$.

Proof. By independence and the definition of the KL divergence,

$$\begin{aligned}\mathrm{K}(p_{\mathbf{X}^a}, p_{\mathbf{X}^b}) &= \sum_{i\in[m],j\in[n]} \left[X^a_{i,j} \log \frac{X^a_{i,j}}{X^b_{i,j}} + X^b_{i,j} - X^a_{i,j} \right] \\ &\leq \sum_{i\in[m],j\in[n]} \left[X^b_{i,j} - X^a_{i,j} \right] = \|\mathbf{X}^a - \mathbf{X}^b\|_1,\end{aligned}$$

where the inequality follows from the fact that $X^a_{i,j} \log \frac{X^a_{i,j}}{X^b_{i,j}} \leq 0$ since $X^b_{i,j} \geq X^a_{i,j}$ (and following standard convention that $a \log(a/0) = \infty, 0 \log(0/a) = 0$ for $a > 0$). □

2.4 Discussion

It is worthwhile to explicitly point out a unique point in our analysis – introducing the additional dimension in the model to ensure that our class of candidate solutions contains an element that always overestimates, element-wise, the rates in the true parameter matrix $\mathbf{X}^*$ – enables us to obtain estimation error rates without making any assumptions on the *minimum* rate of the underlying Poisson processes. This is a significant contrast with prior efforts employing penalized maximum likelihood analyses (but with different structural models) on Poisson-distributed data [13,14], each of which prescribe adopting an assumption that the rates associated with each Poisson-distributed observation be strictly bounded away from 0.

Our extension here is an important advance, especially in the context of extremely photon-limited scenarios. Indeed, in these settings it is somewhat counter-intuitive (or at least, restrictive) to assume that the rates be bounded away from zero, as it is precisely in these scenarios when one might be most interested in estimating rates that are very near zero. Further, classical analyses suggest that there may be no *fundamental* reason why zero or nearly-zero rates become more difficult to estimate. For instance, in the scalar Poisson rate estimation problem, the Cramer-Rao lower bound for estimating a Poisson rate parameter from n iid $\mathrm{Poi}(\cdot|\theta)$ observations (achievable with the sample average estimator) is θ/n, suggesting that the estimation problem actually becomes easier as the rate decreases. The analytical framework we develop here facilitates analysis of these important low-rate cases under sparse and structured data model assumptions.

Finally, we note that Poisson models also find utility other application domains beyond imaging. In networking tasks, for example, Poisson processes are a natural choice to model arrival events, such as packets arriving at each of a number of network routers our flows across network links (see, e.g., [37]). Our techniques and analysis here would extend directly to other application domains, as well.

2.5 Conclusions

In this work, we described a framework for quantifying the mean-square error of constrained maximum likelihood Poisson denoising strategies, in settings where the collection of underlying rates (appropriately arranged) admits a low-rank or sparse dictionary-based decomposition. We established that, in these cases, the mean-square estimation error exhibits characteristics of the familiar parametric rate, in that the error essentially takes the form of "degrees of freedom" divided by "number of observations." In analogy to related analyses in [14, 26], our analysis can also be used to obtain error rates for data adhering to models that are not exactly sparse, but instead are characterized by coefficients whose ordered amplitudes decay (e.g., at a polynomial rate). Finally, while our analysis here was formulated in terms of matrix-structured data and factorization models, these methods may be extended straightforwardly to encompass also sparse and low-rank models for higher-order tensor structure data. We defer in-depth investigations of these extensions to a future effort.

Chapter 3

Noisy Matrix Completion under Sparse Factor Models

This work [8] examines a general class of noisy matrix completion tasks where the goal is to estimate a matrix from observations obtained at a subset of its entries, each of which is subject to random noise or corruption. Our specific focus is on settings where the matrix to be estimated is well-approximated by a product of two (a priori unknown) matrices, one of which is sparse. Such structural models – referred to here as “sparse factor models” – have been widely used, for example, in subspace clustering applications, as well as in contemporary sparse modeling and dictionary learning tasks. Our main theoretical contributions are estimation error bounds for sparsity-regularized maximum likelihood estimators for problems of this form, which are applicable to a number of different observation noise or corruption models. Several specific implications are examined, including scenarios where observations are corrupted by additive Gaussian noise or additive heavier-tailed (Laplace) noise, Poisson-distributed observations, and highly-quantized (e.g., one-bit) observations. We also propose a simple algorithmic approach based on the alternating direction method of multipliers for these tasks, and provide experimental evidence to support our error analyses.

3.1 Introduction

In recent years, there has been significant research activity aimed at the analysis and development of efficient *matrix completion* methods, which seek to "impute" missing elements of a matrix given possibly noisy or corrupted observations collected at a subset of its locations. Let $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ denote a matrix whose elements we wish to estimate, and suppose that we observe $\mathbf{X}^*$ at only a subset $\mathcal{S} \subset [n_1] \times [n_2]$ of its locations, where $[n_1] = \{1, 2, \ldots, n_1\}$ is the set of all positive integers less or equal to n_1 (and similarly for n_2), obtaining at each $(i, j) \in \mathcal{S}$ a noisy, corrupted, or inexact measurement denoted by $Y_{i,j}$. The overall aim is to estimate $\mathbf{X}^*$ given $\mathcal{S}$ and the observations $\{Y_{i,j}\}_{(i,j)\in\mathcal{S}}$. Of course, such estimation problems may be ill-posed without further assumptions, since the values of $\mathbf{X}^*$ at the unobserved locations could in general be arbitrary. A common approach is to augment the inference method with an assumption that the underlying matrix to be estimated exhibits some form of intrinsic *low-dimensional* structure.

One application where such techniques have been successfully utilized is collaborative filtering (e.g., as in the well-known *Netflix Prize* competition [38]). There, the matrix to be estimated corresponds to an array of users' preferences or ratings for a collection of items (which could be quantized, e.g., to one of a number of levels); accurately inferring missing entries of the underlying matrix is a useful initial step in *recommending* items (here, movies or shows) to users deemed likely to rate them favorably. A popular approach to this problem utilizes a low-rank modeling assumption, which implicitly assumes that individual ratings depend on some unknown but nominally small number (say r) of features, so that each element of $\mathbf{X}^*$ may be described as an inner product between two length-r vectors – one quantifying how well each of the features are embodied or represented by a given item, and the other describing a user's affinity for each of the features. Recent works examining the efficacy of low-rank models for matrix completion include [39–45].

Several other applications where analogous ideas have been employed, but which leverage different structural modeling assumptions, include:

- **Sparse Coding for Image Inpainting and Demosaicing**: Suppose that the underlying data to be estimated takes the form of an $n_1 \times n_2$ color image, which may be interpreted as an $n_1 \times n_2 \times 3$ array (the three levels correspond to values in three color planes). The image inpainting task amounts to estimating the image from a collection of (possibly noisy) observations obtained at individual pixel locations (so that at each pixel, either all or none of the color planes are observed), and the demosaicing task entails estimating the image from noisy measurements corresponding to only one of the 3 possible color planes at each pixel. The recent work [46] proposed estimation approaches for these tasks that leverage *local shared structure* manifesting at the patch level. Specifically, in that work, the overall image to be estimated is viewed equivalently as a matrix comprised of vectorized versions of its small (e.g., $5 \times 5 \times 3$ or $8 \times 8 \times 3$) blocks, and the missing values are imputed using a structural assumption that this patch-based matrix be well-approximated by a product of two matrices, one of which is sparse.

- **Sparse Models for Learning and Content Analytics**: A recent work [47] investigated a matrix completion approach to machine-based learning analytics. There, the elements of the $n_1 \times n_2$ matrix to be estimated, say $\mathbf{X}^*$, are related to the probability with which one of n_1 questions will be answered correctly by one of n_2 "learners" through a *link function* $\Phi : \mathbb{R} \rightarrow [0,1]$, so that the value $\Phi(X^*_{i,j})$ denotes the *probability* with which question i will be correctly answered by learner j. The observed data are a collection of some $m < n_1 n_2$ binary values, which may be interpreted as (random) Bernoulli($\Phi(X^*_{i,j})$) variables. The approach proposed in [47] entails maximum-likelihood estimation of the unknown latent factors of $\mathbf{X}^*$, under an assumption that $\mathbf{X}^*$ be well-approximated by a sum of two matrices, the first being product of a sparse non-negative matrix (relating questions to some latent "concepts") and a matrix relating a learner's knowledge to the concepts, and the second quantifying the intrinsic difficulty of each question.

- **Subspace Clustering from Missing Data**: The general subspace clustering problem entails separating a collection of data points, using an assumption that similar points are described as points lying in the same subspace, so that the overall collection of data are represented as points belonging generally to a union of (ostensibly, low-dimensional) subspaces. This general task finds application in image processing, computer vision, and disease detection, to name a few (see, e.g., [48–53], and the references therein). One direct way to perform clustering in such applications entails approximating the underlying matrix $\mathbf{X}^*$ whose columns comprise the (uncorrupted) data points by a product of two matrices, the second of which is sparse, so that the support (the set of locations of the nonzero elements) of each column of the sparse matrix factor identifies the subspace to which the corresponding column of $\mathbf{X}^*$ belongs.

While these examples all seem qualitatively similar in scope, their algorithmic and analytical tractability can vary significantly depending on the type of structural model adopted. In the collaborative filtering application, for example, a desirable aspect of adopting low-rank models is that the associated inference (imputation) procedures can be relaxed to efficient convex methods that are amenable to precise performance analyses. Indeed, the statistical performance of convex methods for low-rank matrix completion are now well-understood in noise-free settings (see, e.g., [39–42]), in settings where observations are corrupted by some form of additive uncertainty [54–58], and even in settings where the observations may be interpreted as nonlinear (e.g., highly-quantized) functions of the underlying matrix entries [59–61]. In contrast, the aforementioned inference methods based on general bilinear (and sparse) factor models are difficult to solve to global optimality, and are instead replaced by tractable alternating minimization methods. More fundamentally, the statistical performance of inference methods based on these more general bilinear models, in scenarios where the observations could arise from general (perhaps nonlinear) corruption models or could even be multi-modal in nature, has not (to our knowledge) been fully characterized.

This work provides some initial results in this direction. We establish a general-purpose estimation error guarantee for matrix completion problems characterized by any of a number of structural data models and observation noise/corruption models. For concreteness, we instantiate our main result here for the special case where the matrix to be estimated adheres to a *sparse factor model*, meaning that it is well-approximated by the product of two matrices, one of which is sparse (or approximately so). Sparse factor models are inherent in the modeling assumptions adopted in the aforementioned works on image denoising/demosaicing, content analytics, and subspace clustering, and are also at the heart of recent related efforts in dictionary learning [17–19]. Sparse factor models may also serve as a well-motivated extension to the low-rank models often utilized in collaborative filtering tasks. There, while it is reasonable to assume that users' preferences will depend on a small number of abstract features, it may be that any particular user's preference relies heavily on only a subset of the features, and that the features that are most influential in forming a rating may vary from user to user. Low rank models alone are insufficient for capturing this "higher order" structure on the latent factors, while this behavior may be well-described using the sparse factor models we consider here.

3.1.1 Our Contributions

We address general problems of matrix completion under sparse factor modeling assumptions using the machinery of *complexity-regularized maximum likelihood* estimation. Our main contributions come in the form of estimation error bounds that are applicable in settings where the available data correspond to an incomplete collection of noisy observations of elements of the underlying matrix (obtained at random locations), and under general (random) noise/corruption models. We examine several specific implications of our main result, including for scenarios characterized by additive Gaussian noise or additive heavier-tailed (Laplace) noise, Poisson-distributed observations, and highly-quantized (e.g., one-bit) observations. Where possible, we draw direct comparisons with existing results in the low-rank matrix completion literature, to illustrate the potential benefit of leveraging additional structure in the latent factors. We also propose an efficient unified algorithmic approach based on the alternating direction method of multipliers [62] for obtaining a local solution to the (non-convex) optimizations prescribed by our analysis, and provide experimental evidence to support our error results.

3.1.2 Connections with Existing Works

As alluded above, our theoretical analyses here are based on the framework of complexity regularized maximum likelihood estimation [23,24], which has been utilized in a number of works to establish error bounds for Poisson estimation problems using multi scale models [13, 63], transform domain sparsity models [14], and dictionary-based matrix factorization models [5]. Here, our analysis extends that framework to the "missing data" scenarios inherent in matrix completion tasks (and also provides a missing-data extension of our own prior work on dictionary learning from 1-bit data [64]).

Our proposed algorithmic approach is based on the alternating direction method of multipliers (ADMM) [62]. ADMM-based methods for related tasks in *dictionary learning* (DL) were described recently in [65], and while our algorithmic approach here is qualitatively similar to that work, we consider missing data scenarios as well as more general loss functions that arise as negative log-likelihoods for our various probabilistic corruption models (thus generalizing these techniques beyond common squared error losses). In addition, our algorithmic framework also allows for direct incorporation of constraints not only on estimates of the matrix factors, but also on the estimate of $\mathbf{X}^*$ itself to account for entry-wise structural constraints that could arise naturally in many matrix completion scenarios. Several other recent efforts in the DL literature have proposed algorithmic procedures for coping with missing data [66, 67], and a survey of algorithmic approaches to generalized low-rank modeling tasks is given in the recent work [68].

Our inference tasks here essentially entail learning two factors in a bilinear model. With a few notable exceptions (e.g., low-rank matrices, and certain non-negative matrices [69–72]), the joint non-convexity of these problems can complicate their analysis. Recently, several efforts in the dictionary learning literature have established theoretical guarantees on identifiability, as well as local correctness of a number of factorization methods [73–78], including in noisy settings [79]. Our efforts here may be seen as a complement to those works, providing additional insight into the achievable *statistical* performance of similar methods under somewhat general noise models.

The factor models we employ here essentially enforce that each column of $\mathbf{X}^*$ lie in a union of linear subspaces. In this sense our efforts here are also closely related to problems in sparse principal component analysis [80], which seek to decompose the (sample) covariance matrix of a collection of data points as a sum of rank-one factors expressible as outer products of sparse vectors. Several efforts have examined algorithmic approaches to the sparse PCA problem based on greedy methods [81] or convex relaxations [82–84], and very recently several efforts have examined the statistical performance of cardinality- (or ℓ_0-) constrained methods for identifying the first sparse principal component [85, 86]. These latter approaches are related to our effort here, as our analysis below pertains to the performance of matrix completion methods utilizing an ℓ_0 penalty on one of the matrix factors.

Finally, we note that problems of subspace clustering from missing or noisy data have received considerable attention in recent years. Algorithmic approaches to subspace clustering with missing data were proposed in [53, 87, 88], and several recent works have identified sufficient conditions under which tractable algorithms will provably recover the unknown subspaces in missing data (but noise-free) scenarios [89]. Robustness of subspace clustering methods to missing data, additive noise, and potentially large-valued outliers were examined recently in [52, 53, 90].

3.1.3 Outline

The remainder of this paper is organized as follows. Following a brief discussion of several preliminaries (below), we formalize our problem in Section 3.2 and present our main result establishing estimation error guarantees for a general class of estimation problems characterized by incomplete and noisy observations. In Section 3.3 we discuss implications of this result for several specific noise models. In Section 3.4 we discuss a unified algorithmic approach to problems of this form, based on the alternating direction method of multipliers, and provide a brief experimental investigation that partially validates our theoretical analyses. We conclude with a brief discussion in Section 3.5. Auxiliary material and detailed proofs are relegated to the appendix.

3.1.4 Preliminaries

To set the stage for the statement of our main result, we remind the reader of a few key concepts. First, recall that for $p \leq 1$ a vector $\mathbf{x} \in \mathbb{R}^n$ is said to belong to a weak-ℓ_p ball of radius $R > 0$, denoted $\mathbf{x} \in w\ell_p(R)$, if its ordered elements $|x_{(1)}| \geq |x_{(2)}| \geq \cdots \geq |x_{(n)}|$ satisfy

$$|x_{(i)}| \leq R i^{-1/p} \quad \text{for all } i \in \{1, 2, \ldots, n\}, \tag{3.1}$$

see e.g., [91]. Vectors in weak-ℓ_p balls may be viewed as approximately sparse; indeed, it is well-known (and easy to show, using standard results for bounding sums by integrals) that for a vector $\mathbf{x} \in w\ell_p(R)$, the ℓ_q error associated with approximating $\mathbf{x}$ by its best k-term approximation obtained by retaining its k largest entries in amplitude (denoted here by $\mathbf{x}^{(k)}$) satisfies

$$\|\mathbf{x} - \mathbf{x}^{(k)}\|_q \triangleq \left(\sum_{i=1}^{n} |x_i - x_i^{(k)}|^q \right)^{1/q} \leq R \; C_{p,q} \; k^{1/q - 1/p}, \tag{3.2}$$

for any $q > p$, where $C_{p,q}$ is given by

$$C_{p,q} = \left(\frac{p}{q-p} \right)^{1/q}. \tag{3.3}$$

For the special case $q \geq 2p$, we have $C_{p,q} \leq 1$, and so

$$\|\mathbf{x} - \mathbf{x}^{(k)}\|_q \leq R \; k^{1/q - 1/p}. \tag{3.4}$$

We also recall several information-theoretic preliminaries. When $p(Y)$ and $q(Y)$ denote the pdf (or pmf) of a real-valued random variable Y, the Kullback-Leibler divergence (or KL divergence) of q from p is denoted $\mathrm{D}(p\|q)$ and given by

$$\mathrm{D}(p\|q) = \mathbb{E}_p \left[\log \frac{p(Y)}{q(Y)} \right]$$

where the logarithm is taken to be the natural log. By definition, $\mathrm{D}(p\|q)$ is finite only if the support of p is contained in the support of q. Further, the KL divergence satisfies $\mathrm{D}(p\|q) \geq 0$ and $\mathrm{D}(p\|q) = 0$ when $p(Y) = q(Y)$. We also use the Hellinger affinity denoted by $\mathrm{A}(p, q)$ and given by

$$\mathrm{A}(p, q) = \mathbb{E}_p \left[\sqrt{\frac{q(Y)}{p(Y)}} \right] = \mathbb{E}_q \left[\sqrt{\frac{p(Y)}{q(Y)}} \right]$$

Note that $\mathrm{A}(p, q) \geq 0$ essentially by definition, and a simple application of the Cauchy-Schwarz inequality gives that $\mathrm{A}(p, q) \leq 1$, implying overall that $0 \leq \mathrm{A}(p, q) \leq 1$. When p and q are parameterized by elements $X_{i,j}$ and $\widetilde{X}_{i,j}$ of matrices $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, respectively, so that $p(Y_{i,j}) = p_{X_{i,j}}(Y_{i,j})$ and $q(Y_{i,j}) = q_{\widetilde{X}_{i,j}}(Y_{i,j})$, we use the shorthand notation $\mathrm{D}(p_{\mathbf{X}} \| q_{\widetilde{\mathbf{X}}}) \triangleq \sum_{i,j} \mathrm{D}(p_{X_{i,j}} \| q_{\widetilde{X}_{i,j}})$ and $\mathrm{A}(p_{\mathbf{X}}, q_{\widetilde{\mathbf{X}}}) \triangleq \prod_{i,j} \mathrm{A}(p_{X_{i,j}}, q_{\widetilde{X}_{i,j}})$.

Finally, for a matrix $\mathbf{M}$ we denote by $\|\mathbf{M}\|_0$ its number of nonzero elements, $\|\mathbf{M}\|_1$ the sum of absolute values of its elements, $\|\mathbf{M}\|_{\max}$ the magnitude of its largest element (in absolute value), and $\|\mathbf{M}\|_*$ its nuclear norm (sum of singular values).

3.2 Problem Statement, Approach, and a General Recovery Result

As above, we let $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ denote the unknown matrix whose entries we seek to estimate. Our focus is on cases where the unknown matrix $\mathbf{X}^*$ admits a factorization of the form

$$\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*, \tag{3.5}$$

where for some integer $r \leq n_2$, $\mathbf{D}^* \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{A}^* \in \mathbb{R}^{r \times n_2}$ are *a priori unknown* factors. For pragmatic reasons, we assume that the elements of $\mathbf{D}^*$, $\mathbf{A}^*$, and $\mathbf{X}^*$ are bounded, in the sense that

$$\|\mathbf{D}^*\|_{\max} \leq 1, \quad \|\mathbf{A}^*\|_{\max} \leq \mathrm{A}_{\max}, \quad \text{and} \quad \|\mathbf{X}^*\|_{\max} \leq \mathrm{X}_{\max}/2 \tag{3.6}$$

for some constants $0 < \mathrm{A}_{\max} \leq (n_1 \vee n_2) = \max\{n_1, n_2\}$ and $\mathrm{X}_{\max} \geq 1$. Bounds on the amplitudes of the elements of the matrix to be estimated often arise naturally in practice[1] , while our assumption that the entries of the factor matrices be bounded is essentially to fix scaling ambiguities associated with the bilinear model. Our particular focus here will be on cases where (in addition to the entry-wise bounds) the matrix $\mathbf{A}^*$ is *sparse* (having no more than $k < rn_2$ nonzero elements), or *approximately sparse*, in the sense that for some $p \leq 1$, all of its columns lie in a weak-ℓ_p ball of radius $\mathrm{A}_{\max}$.

Rather than acquire all of the elements of $\mathbf{X}^*$ directly, we assume here that we only observe $\mathbf{X}^*$ at a known *subset* of its locations, obtaining for each observation a noisy or corrupted version of the underlying matrix entry. Here, we will interpret the notion of "noise" somewhat generally in an effort to make our analysis amenable to any of a number of different corruption models; in what follows, we will model each entry-wise observation as a random quantity (either continuous or discrete-valued) whose probability density (or mass) function is parameterized by the true underlying matrix entry. We denote by $\mathcal{S} \subseteq [n_1] \times [n_2]$ the set of locations at which observations are collected, and assume that the sampling locations are random in the sense that for an integer m satisfying $4 \leq m \leq n_1 n_2$ and $\gamma = m(n_1 n_2)^{-1}$, $\mathcal{S}$ is generated according to the independent Bernoulli(γ) model so that each $(i, j) \in [n_1] \times [n_2]$ is included in

[1] Here, the factor of $1/2$ in the bound on $\|\mathbf{X}^*\|_{\max}$ is somewhat arbitrary – any factor in $(0, 1)$ would suffice – and is chosen to facilitate our subsequent analysis.

$\mathcal{S}$ independently with probability γ. Then, given $\mathcal{S}$, we model the collection of $|\mathcal{S}|$ measurements of $\mathbf{X}^*$ in terms of a collection $\{Y_{i,j}\}_{(i,j)\in\mathcal{S}} \triangleq \mathbf{Y}_{\mathcal{S}}$ of conditionally (on $\mathcal{S}$) independent random quantities. Formally, we write the joint pdf (or pmf) of the observations as

$$p_{\mathbf{X}^*_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) \triangleq \prod_{(i,j)\in\mathcal{S}} p_{X^*_{i,j}}(Y_{i,j}), \tag{3.7}$$

where $p_{X^*_{i,j}}(Y_{i,j})$ denotes the corresponding scalar pdf (or pmf), and we use the shorthand $\mathbf{X}^*_{\mathcal{S}}$ to denote the collection of elements of $\mathbf{X}^*$ indexed by $(i,j)\in\mathcal{S}$. In terms of this model, our task may be described concisely as follows: given $\mathcal{S}$ and corresponding noisy observations $\mathbf{Y}_{\mathcal{S}}$ of $\mathbf{X}^*$ distributed according to (3.7), our goal is to estimate $\mathbf{X}^*$ under the assumption that it admits a sparse factor model decomposition.

Our approach will be to estimate $\mathbf{X}^*$ via sparsity-penalized maximum likelihood methods; we consider estimates of the form

$$\widehat{\mathbf{X}} = \arg \min_{\mathbf{X}=\mathbf{D}\mathbf{A}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) + \lambda\cdot\|\mathbf{A}\|_0\right\}, \tag{3.8}$$

where $\lambda > 0$ is a user-specified regularization parameter, $\mathbf{X}_{\mathcal{S}}$ is shorthand for the collection $\{X_{i,j}\}_{(i,j)\in\mathcal{S}}$ of entries of $\mathbf{X}$ indexed by $\mathcal{S}$, and $\mathcal{X}$ is an appropriately constructed class of candidate estimates. To facilitate our analysis here, we take $\mathcal{X}$ to be a countable class of estimates constructed as follows: first, for a specified $\beta \geq 1$, we set $L_{\text{lev}} = 2^{\lceil \log_2 (n_1\vee n_2)^{\beta}\rceil}$ and construct $\mathcal{D}$ to be the set of all matrices $\mathbf{D}\in\mathbb{R}^{n_1\times r}$ whose elements are discretized to one of L_{lev} uniformly-spaced levels in the range $[-1,1]$ and $\mathcal{A}$ to be the set of all matrices $\mathbf{A}\in\mathbb{R}^{r\times n_2}$ whose elements either take the value zero, or are discretized to one of L_{lev} uniformly-spaced levels in the range $[-\mathrm{A}_{\max}, \mathrm{A}_{\max}]$. Then, we let

$$\mathcal{X}' \triangleq \left\{\mathbf{X}=\mathbf{D}\mathbf{A} \;:\; \mathbf{D}\in\mathcal{D},\ \mathbf{A}\in\mathcal{A},\ \|\mathbf{X}\|_{\max}\leq \mathrm{X}_{\max}\right\}, \tag{3.9}$$

and take $\mathcal{X}$ to be any subset of $\mathcal{X}'$. This general formulation will allow us to easily and directly handle additional constraints (e.g., non-negativity constraints on the elements of $\mathbf{X}$, as arise in our treatment of the Poisson-distributed observation model), within the same unified analytical framework.

Our first main result establishes error bounds for sparse factor model matrix completion problems under general noise or corruption models, where the corruption is described by any generic likelihood model. We state the result here as a theorem; its proof appears in Appendix 3.6.1 and utilizes a key lemma that extends a main result of [24] to "missing data" scenarios inherent in completion tasks.

Theorem 3.2.1. *Let the sample set $\mathcal{S}$ be drawn from the independent Bernoulli model with $\gamma = m(n_1 n_2)^{-1}$ as described above, and let $\mathbf{Y}_{\mathcal{S}}$ be described by* (3.7)*. If C_{D} is any constant satisfying*

$$C_{\mathrm{D}} \geq \max_{\mathbf{X}\in\mathcal{X}} \max_{i,j} \ D(p_{X^*_{i,j}} \| p_{X_{i,j}}), \tag{3.10}$$

where $\mathcal{X}$ is as above for some $\beta \geq 1$, then for any

$$\lambda \geq 2 \cdot (\beta + 2) \cdot \left(1 + \frac{2C_{\mathrm{D}}}{3}\right) \cdot \log(n_1 \vee n_2), \tag{3.11}$$

the complexity penalized maximum likelihood estimator (3.8) *satisfies the (normalized, per-element) error bound*

$$\begin{aligned} \frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq{} & \frac{8C_{\mathrm{D}}\log m}{m} \\ + \quad 3 \cdot \min_{\mathbf{X}\in\mathcal{X}} & \left\{ \frac{\mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}})}{n_1 n_2} + \left(\lambda + \frac{4C_{\mathrm{D}}(\beta+2)\log(n_1 \vee n_2)}{3}\right)\left(\frac{n_1 p + \|\mathbf{A}\|_0}{m}\right)\right\}. \end{aligned} \tag{3.12}$$

In the next section we consider several specific instances of this result, but we first note a few salient points about this result in its general form. First, as alluded above, our result is not specific to any one observation model; thus, our general result will allow us to analyze the error performance of sparse factor matrix completion methods under a variety of different noise or corruption models. Specialization to a given noise model requires us to only compute (or appropriately bound) the KL divergences and negative log Hellinger affinities of the corresponding probability densities or probability mass functions. Second, our error bound is a kind of *oracle* bound, in that it is specified in terms of a minimum over $\mathbf{X} \in \mathcal{X}$. In practice, we may evaluate this oracle term for *any* $\mathbf{X} \in \mathcal{X}$ and still obtain a valid upper bound (since our guarantee is in terms of the minimum). In our analyses that follows we will impose assumptions on β and $\mathbf{X}^*$ that ensure $\mathbf{X}^*$ be sufficiently "close" to some element $\mathbf{X}$ of $\mathcal{X}$. This will enable us to obtain non-trivial bounds on the first term in the oracle expression, and to subsequently quantify the corresponding normalized, per-element error (as described in terms of the corresponding negative log Hellinger affinity) by judiciously "balancing" the terms in the oracle expression. This approach will be illustrated in the following section.

Finally, it is worth noting that the estimation strategies prescribed by our analysis are not computationally tractable. Indeed, as written, formation of our estimators would require solving a combinatorial optimization, because of the ℓ_0 penalty, as well as the optimization over the discrete set $\mathcal{X}$. However, it is worth noting that inference in the bilinear models we consider here is fundamentally challenging on account of the fact that these inference problems cannot directly be cast as (jointly) convex optimizations in the matrix factors. In that sense, our results here may be interpreted as quantifying the performance of one (benchmark) estimation approach for sparse factor matrix completion under various corruption models. (We discuss several extensions, including potential avenues for convexification, in Section 3.5.)

3.3 Implications for Specific Noise Models

In this section we consider the implications of Theorem 3.2.1 in four unique scenarios, characterized by additive Gaussian noise, additive heavier-tailed (Laplace) noise, Poisson-distributed observations, and quantized (one-bit) observations. In each case, our aim is to identify the scaling behavior of the estimation error as a function of the key problem parameters. To that end, we consider for each case the fixed choice

$$\beta = \max\left\{1, 1 + \frac{\log(8r\mathrm{A}_{\max}/\mathrm{X}_{\max})}{\log(n_1 \vee n_2)}\right\} \tag{3.13}$$

for describing the number of discretization levels in the elements of each of the matrix factors. Then, for each scenario (characterized by its own unique likelihood model) we consider a specific choice of $\mathcal{X}$, and an estimate obtained according to (3.8) with the specific choice

$$\lambda = 2\left(1 + \frac{2C_{\mathrm{D}}}{3}\right)(\beta + 2) \cdot \log(n_1 \vee n_2), \tag{3.14}$$

(where C_{D} depends on the particular likelihood model), and simplify the resulting oracle bounds for both sparse and approximately sparse factors. In what follows, we will make use of the fact that our assumption $\mathrm{X}_{\max} \geq 1$ implies $\beta = \mathcal{O}\left(\log(r \vee \mathrm{A}_{\max})/\log(n_1 \vee n_2)\right)$, and so $(\beta + 2)\log(n_1 \vee n_2) = \mathcal{O}\left(\log(n_1 \vee n_2)\right)$, on account of the fact that $r < n_2$ and $\mathrm{A}_{\max} < (n_1 \vee n_2)$ by assumption.

3.3.1 Additive Gaussian Noise

We first examine the implications of Theorem 3.2.1 in a setting where observations are corrupted by independent additive zero-mean Gaussian noise with known variance. In this case, the observations $\mathbf{Y}_{\mathcal{S}}$ are distributed according to a multivariate Gaussian density of dimension $|\mathcal{S}|$ whose mean corresponds to the collection of matrix parameters at the sample locations, and with covariance matrix $\sigma^2\mathbf{I}_{|\mathcal{S}|}$, where $\mathbf{I}_{|\mathcal{S}|}$ is the identity matrix of dimension $|\mathcal{S}|$, so

$$p_{\mathbf{X}^*_S}(\mathbf{Y}_S) = \frac{1}{(2\pi\sigma^2)^{|S|/2}} \exp\left(-\frac{1}{2\sigma^2}\ \|\mathbf{Y}_S - \mathbf{X}^*_S\|_F^2\right), \tag{3.15}$$

where we have used the representative shorthand notation $\|\mathbf{Y}_S - \mathbf{X}^*_S\|_F^2 \triangleq \sum_{(i,j)\in S}(Y_{i,j} - X^*_{i,j})^2$. In this setting we have the following result; its proof appears in Appendix 3.6.2.

Corollary 3.3.1 (Sparse Factor Matrix Completion with Gaussian Noise)**.** *Let* β *be as in* (3.13)*, let* λ *be as in* (3.14) *with* $C_{\mathrm{D}} = 2\mathrm{X}_{\max}^2/\sigma^2$*, and let* $\mathcal{X} = \mathcal{X}'$*. The estimate* $\widehat{\mathbf{X}}$ *obtained via* (3.8) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} = \mathcal{O}\left((\sigma^2 + \mathrm{X}_{\max}^2)\left(\frac{n_1 r + \|\mathbf{A}^*\|_0}{m}\right)\log(n_1 \vee n_2)\right) \tag{3.16}$$

when $\mathbf{A}^*$ *is exactly sparse, having* $\|\mathbf{A}^*\|_0$ *nonzero elements. If, instead, the columns of* $\mathbf{A}^*$ *are approximately sparse in the sense that for some* $p \leq 1$ *each belongs to a weak-*ℓ_p *ball of radius* $\mathrm{A}_{\max}$*, then the estimate* $\widehat{\mathbf{X}}$ *obtained via* (3.8) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} =$$
$$\mathcal{O}\left(\mathrm{A}_{\max}^2\left(\frac{n_2}{m}\right)^{\frac{2\alpha}{2\alpha+1}} + (\sigma^2 + \mathrm{X}_{\max}^2)\left(\frac{n_1 r}{m} + \left(\frac{n_2}{m}\right)^{\frac{2\alpha}{2\alpha+1}}\right)\log(n_1 \vee n_2)\right), \tag{3.17}$$

where $\alpha = 1/p - 1/2$*.*

Remark 3.3.1. *We utilize Big-Oh notation to suppress leading constants for clarity of exposition, and to illustrate the dependence of the bounds on the key problem parameters. Our proofs for of each of the specific results provides the explicit constants.*

A few comments are in order regarding these error guarantees. First, we note that our analysis provides some useful (and intuitive) understanding of how the estimation error decreases as a function of the number of measurements obtained, as well as the dimension and sparsity parameters associated with the matrix to be estimated. Consider, for instance, the case when $\mathbf{A}^*$ is sparse and where $\log m < n_1 r + \|\mathbf{A}^*\|_0$ (which should often be the case, since $\log(m) \leq \log(n_1 n_2)$). In this setting, our error bound shows that the dependence of the estimation error on the dimension (n_1, n_2, r) and sparsity $(\|\mathbf{A}^*\|_0)$ parameters, as well as the (nominal) number of measurements m is

$$\frac{n_1 r + \|\mathbf{A}^*\|_0}{m} \log(n_1 \vee n_2). \tag{3.18}$$

We may interpret the quantity $n_1 r + \|\mathbf{A}^*\|_0$ as the number of *degrees of freedom* in the matrix $\mathbf{X}^*$ to be estimated, and in this sense we see that the error rate of the penalized maximum likelihood estimator exhibits characteristics of the well-known parametric rate (modulo the logarithmic factor). Along related lines, note that in the case where columns of $\mathbf{A}^*$ are approximately sparse, the $(n_2/m)^{\frac{2\alpha}{2\alpha+1}}$ term that arises in the error rate is reminiscent of error rates that arise when estimating approximately sparse vectors in noisy compressive sensing (e.g., see [26, 92]). Indeed, since $(n_2/m)^{\frac{2\alpha}{2\alpha+1}} \leq n_2 m^{-\frac{2\alpha}{2\alpha+1}}$, we see that the overall matrix estimation error may be interpreted as being comprised of errors associated with approximating the n_2 nearly-sparse columns of $\mathbf{A}^*$ in this noisy setting, each of which would contribute a (normalized) error on the order of $m^{-\frac{2\alpha}{2\alpha+1}}$.

Next, our error bounds provide some guidelines for identifying in which scenarios accurate estimation may be possible. Consider a *full sampling* scenario where the matrix $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$ has a coefficient matrix with no more than k nonzero elements per column (thus, $\|\mathbf{A}^*\|_0 \leq n_2 k$). Now, to ensure that

$$\frac{n_1 r + \|\mathbf{A}^*\|_0}{n_1 n_2} \log(n_1 \vee n_2) \preceq 1, \tag{3.19}$$

(where the notation $\preceq$ suppresses leading constants) it is sufficient to have $n_1 n_2 \succeq n_1 r \log(n_1 \vee n_2)$ and $n_1 n_2 \succeq \|\mathbf{A}^*\|_0 \log(n_1 \vee n_2)$. Simplifying a bit, we see that the first sufficient condition is satisfied when $n_2 \succeq 2r \log(n_1 \vee n_2)$, or when the number of columns of the matrix $\mathbf{X}^*$ exceeds (by a multiplicative constant and logarithmic factor) the number of columns of its dictionary factor $\mathbf{D}^*$. Further, the second sufficient condition holds when $n_1 \succeq k \log(n_1 \vee n_2)$, or when the number of measurements of each column exceeds (again, by a multiplicative constant and logarithmic factor) the number of nonzeros in the sparse representation of each column. This latter condition is reminiscent of the sufficient conditions arising in sparse inference problems inherent in noisy compressive sensing (see, e.g., [92, 93]). Analogous insights may be derived from our results for the subsampled regimes that comprise our main focus here (i.e., when $m < n_1 n_2$).

Further, we comment on the presence of the $\mathrm{X}_{\max}^2$ term present in the error bounds for both the sparse and nearly-sparse settings. Readers familiar with the literature on matrix completion under low rank assumptions will recall that various forms of "incoherence" assumptions have been utilized to date as a means to ensure identifiability under various sampling models, and that the form of the resulting error bounds depend on the particular type of assumption employed. For example, the authors of [56] consider an additive noise model similar to here but employ incoherence assumptions that essentially enforce that the row and column spaces of the matrix to be estimated not be overly aligned with the canonical bases (reminiscent of initial works on noise-free matrix completion [39]) and obtain estimation error bounds that do not depend on max-norm bounds of the matrix to be estimated (though the necessary conditions on the number of samples obtained do depend on the incoherence parameters). The work [58] also examines matrix completion problems with additive noises but utilizes a different form of incoherence assumption formulated in terms of the "spikiness" of the matrix to be estimated (and quantified in terms of the ratio between the max norm and Frobenius norm). There, the estimation approach entails optimization over a set of candidates that each satisfy a "spikiness" constraint, and the bounds so obtained scale in proportion to the max-norm of the matrix to be estimated (similar to here). Incoherence assumptions manifesting as an assumed bound on the largest matrix element also arise in [60, 61].

One direct point of comparison to our result here is [57], which considers matrix completion problems characterized by entry-wise observations obtained at locations chosen uniformly at random (with replacement), each of which may be modeled as corrupted by independent additive noise, and estimates obtained by nuclear norm penalized estimators. Casting the results of that work (specifically, [57, Corollary 2]) to the setting we consider here, we observe that those results imply rank-r matrices may be accurately estimated in the sense that

$$\frac{\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2}{n_1 n_2} \leq c\ (\sigma \vee \mathrm{X}_{\max})^2 \left(\frac{(n_1 \vee n_2)\ r}{m}\right) \log(n_1 + n_2) \tag{3.20}$$

$$\leq c'\ (\sigma^2 + \mathrm{X}_{\max}^2) \left(\frac{(n_1 + n_2) r}{m}\right) \log(n_1 \vee n_2) \tag{3.21}$$

with high probability, where c, c' are positive constants. Comparing this last result with our result (3.16), we see that our guarantees exhibit the same effective scaling with the max-norm bound $\mathrm{X}_{\max}$, but can have an (perhaps significantly) improved error performance in the case where $\|\mathbf{A}^*\|_0 \ll n_2 r$ – precisely what we sought to identify by considering sparse factor models in our analyses. The two bounds roughly coincide in the case where $\mathbf{A}^*$ is not sparse, in which case we may take $\|\mathbf{A}^*\|_0 = n_2 r$ in our error bounds.

3.3.2 Additive Laplace Noise

As another example, suppose that the observations $\mathbf{Y}_{\mathcal{S}}$ are corrupted by independent additive heavier-tailed noises, each of which we model using a Laplace distribution with parameter $\tau > 0$. In this scenario, we have that

$$p_{\mathbf{X}^*_S}(\mathbf{Y}_S) = \left(\frac{\tau}{2}\right)^{|S|} \exp\left(-\tau\ \|\mathbf{Y}_S - \mathbf{X}^*_S\|_1\right), \tag{3.22}$$

where we use $\|\mathbf{Y}_S - \mathbf{X}^*_S\|_1 \triangleq \sum_{(i,j)\in S} |Y_{i,j} - X^*_{i,j}|$ for shorthand. The following result holds; its proof appears in Appendix 3.6.3.

Corollary 3.3.2 (Sparse Factor Matrix Completion with Laplace Noise)**.** *Let β be as in* (3.13)*, let λ be as in* (3.14) *with $C_{\mathrm{D}} = 2\tau\mathrm{X}_{\max}$, and let $\mathcal{X} = \mathcal{X}'$. The estimate $\widehat{\mathbf{X}}$ obtained via* (3.8) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} = \mathcal{O}\left(\left(\frac{1}{\tau} + \mathrm{X}_{\max}\right)^2 \tau\mathrm{X}_{\max}\ \left(\frac{n_1 r + \|\mathbf{A}^*\|_0}{m}\right) \log(n_1 \vee n_2)\right), \tag{3.23}$$

when $\mathbf{A}^*$ is exactly sparse, having $\|\mathbf{A}^*\|_0$ nonzero elements. If, instead, for some $p \leq 1/2$ the columns of $\mathbf{A}^*$ belong to a weak-ℓ_p ball of radius $\mathrm{A}_{\max}$, then the estimate $\widehat{\mathbf{X}}$ obtained *via* (3.8) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} = \mathcal{O}\left(\left(\frac{1}{\tau} + \mathrm{X}_{\max}\right)^2 \tau \mathrm{A}_{\max}\left(\frac{n_2}{m}\right)^{\frac{\alpha'}{\alpha'+1}} + \left(\frac{1}{\tau} + \mathrm{X}_{\max}\right)^2 \tau \mathrm{X}_{\max}\left(\frac{n_1 r}{m} + \left(\frac{n_2}{m}\right)^{\frac{\alpha'}{\alpha'+1}}\right)\log(n_1 \vee n_2)\right),$$

where $\alpha' = 1/p - 1$.

A few comments are in order regarding these results. First, recall that our main theorem naturally provides error guarantees in terms of KL divergences and negative log Hellinger affinities. However, here we state our bounds in terms of the average per element squared error, and draw comparisons with the previous case (and, perhaps, to make the results more amenable to interpretation). To achieve this we employed a series of bounds – quadratic (in the parameter difference) lower bounds on the negative log Hellinger affinities, and upper bounds on the KL divergences that are proportional to the absolute deviations between the parameters (see the proof for details). It is interesting to note that this bounding approach, the error performance that we obtain for the case where $\mathbf{A}^*$ is sparse again exhibits characteristics of the parametric rate, while we do obtain different error behavior as compared to the Gaussian noise case for the case where $\mathbf{A}^*$ is nearly sparse. As one specific example, consider the case where the coefficients of $\mathbf{A}^*$ exhibit the ordered decay with $p = 1/3$ (a parameter that is valid for both Corollaries 3.3.1 and 3.3.2). The error rate for the Gaussian noise setting in this case contains a term that decays on the order of $(m/n_2)^{-5/6}$, while here when the noise is heavy-tailed, the analogous term decays at a slower rate, like $(m/n_2)^{-2/3}$. Overall, casting the error bounds all in terms of the same loss metric (here, ℓ_2) makes our results directly amenable to such comparisons.

Along related lines, it is interesting to note that the estimation error bound here is slightly "inflated" relative to the Gaussian-noise counterparts (albeit with constants suppressed in each case). Recall that the variance of a Laplace(τ) random variable is $2/\tau^2$; thus, the leading term $(1/\tau + \mathrm{X}_{\max})^2 = \mathcal{O}(2/\tau^2 + \mathrm{X}_{\max}^2)$ here is somewhat analogous to the $(\sigma^2 + \mathrm{X}_{\max}^2)$ factor arising in the Gaussian-noise error bounds. In this sense, we see that the factor of $\tau\mathrm{X}_{\max}$ in the Laplace-noise case appears to be "extra." Here, this factor is effectively introduced by our attempt to cast the "natural" error guarantees arising from our analysis (which manifest in terms of negative log Hellinger affinities) into more interpretable squared-error bounds.

3.3.3 Poisson-distributed Observations

We now consider an example motivated by applications where the observed data may correspond to discrete "counts" (e.g., in imaging applications). Suppose that the entries of the matrix $\mathbf{X}^*$ are all non-negative and that our observation at each location $(i,j) \in \mathcal{S}$ is a Poisson random variable with rate $X^*_{i,j}$. In this setting, our matrix completion problem amounts to a kind of Poisson denoising task; we have that $\mathbf{Y}_\mathcal{S} \in \mathbb{N}^{|\mathcal{S}|}$ and

$$p_{\mathbf{X}^*_S}(\mathbf{Y}_S) = \prod_{(i,j)\in S} \frac{(X^*_{i,j})^{Y_{i,j}} e^{-X^*_{i,j}}}{(Y_{i,j})!}. \tag{3.24}$$

In this case, we employ Theorem 3.2.1 to obtain the following result; a sketch of the proof is provided in Appendix 3.6.4.

Corollary 3.3.3 (Sparse Factor Matrix Completion with Poisson Noise)**.** *Suppose that the elements of the matrix* $\mathbf{X}^*$ *to be estimated satisfy* $\min_{i,j} |X^*_{i,j}| \geq \mathrm{X}_{\min}$ *for some constant* $\mathrm{X}_{\min} > 0$*. Let* β *be as in* (3.13)*, let* λ *be as in* (3.14) *with* $C_\mathrm{D} = 4\mathrm{X}^2_{\max}/\mathrm{X}_{\min}$*, and let* $\mathcal{X}$ *be the subset of* $\mathcal{X}'$ *comprised of all candidate estimates having non-negative entries. The estimate* $\widehat{\mathbf{X}}$ *obtained via* (3.8) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} = \mathcal{O}\left(\left(\mathrm{X}_{\max} + \frac{\mathrm{X}_{\max}}{\mathrm{X}_{\min}} \cdot \mathrm{X}^2_{\max}\right)\left(\frac{n_1 r + \|\mathbf{A}^*\|_0}{m}\right)\ \log(n_1 \vee n_2)\right),$$

when $\mathbf{A}^*$ *is exactly sparse, having* $\|\mathbf{A}^*\|_0$ *nonzero elements. If, instead, for some* $p \leq 1$ *the columns of* $\mathbf{A}^*$ *belong to a weak-*ℓ_p *ball of radius* $\mathrm{A}_{\max}$*, then the estimate* $\widehat{\mathbf{X}}$ *obtained via* (3.8) *satisfies*

$$\begin{aligned}\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} &= \mathcal{O}\Bigg(\mathrm{A}^2_{\max}\left(\frac{\mathrm{X}_{\max}}{\mathrm{X}_{\min}}\right)\left(\frac{n_2}{m}\right)^{\frac{2\alpha}{2\alpha+1}} + \\ &\left(\mathrm{X}_{\max} + \frac{\mathrm{X}_{\max}}{\mathrm{X}_{\min}} \cdot \mathrm{X}^2_{\max}\right)\left(\frac{n_1 r}{m} + \left(\frac{n_2}{m}\right)^{\frac{2\alpha}{2\alpha+1}}\right)\log(n_1 \vee n_2)\Bigg),\end{aligned} \tag{3.25}$$

where $\alpha = 1/p - 1/2$*.*

As in the previous case, our analysis approach here entails bounding (appropriately) the KL divergence and negative log Hellinger affinities each in terms of squared Frobenius norms; similar bounding methods were employed in [14], which analyzed a compressive sensing sparse vector reconstruction task under a Poisson observation model. Overall, we observe an interesting behavior relative to the preceding two cases. Recall that the bounds for the setting where $\mathbf{A}^*$ is exactly sparse, for each of the previous two cases, exhibited a leading factor that was essentially the sum of the variance and $\mathrm{X}_{\max}^2$. In each of those cases, the per-observation noise variances were independent of the underlying matrix entry; in contrast, Poisson-distributed observations exhibit a variance equal to the underlying rate parameter. So, in this sense, we might interpret the $(\mathrm{X}_{\max} + (\mathrm{X}_{\max}/\mathrm{X}_{\min})\mathrm{X}_{\max}^2)$ term as roughly corresponding to a "worst-case" variance plus $\mathrm{X}_{\max}^2$. Indeed, when $\mathrm{X}_{\max}/\mathrm{X}_{\min}$ is upper-bounded by a (small) constant; then, this leading factor is $\mathcal{O}(\mathrm{X}_{\max} + \mathrm{X}_{\max}^2)$, somewhat analogously to the leading factor arising in the Laplace-noise and Gaussian-noise bounds. More generally, that the error behavior in Poisson denoising tasks be similar to the Gaussian case is perhaps not surprising. Indeed, a widely used approach in Poisson inference tasks is to employ a *variance stabilizing transformations*, such as the Anscombe transform [94], so that the transformed data distribution be "approximately" Gaussian.

It is worth commenting a bit further on our *minimum* rate assumption on the elements of $\mathbf{X}^*$, that each be no smaller than some constant $\mathrm{X}_{\min} > 0$. Similar assumptions were employed in [14], as well as other works that examine Poisson denoising tasks using the penalized ML analysis framework (e.g., [13]). Here, this $\mathrm{X}_{\min}$ parameter shows up in the denominator of a leading factor in our bound, suggesting that the bounds become more loose as the estimation task transitions closer to scenarios characterized by "low-rate" Poisson sources. Indeed, closer inspection of our error bounds as stated above shows that they diverge (tend to $+\infty$) as $\mathrm{X}_{\min}$ tends to zero, suggesting that the estimation task becomes more difficult in "low rate" settings. Contrast this with classical analyses of scalar Poisson rate estimation problems show that the Cramer-Rao lower bound associated with estimating the rate parameter θ of a Poisson random variable using n iid observations is θ/n, and this error is achievable with the sample average estimator. This suggests that the estimation problem actually becomes *easier* as the rate decreases, at least for the scalar estimation problem. On this note, we briefly mention several recent works that rectify this apparent discrepancy, for matrix estimation tasks as here [5] , and for sparse vector estimation from Poisson-distributed compressive observations [95]. The ideas underlying those works might have applicability for the completion problems we consider here, but this extension may require imposing different (or even much stronger) forms of incoherence assumptions on the matrix to be estimated as compared to the bounded-entry condition we adopt here. We do not pursue those extensions here, opting instead to state our result as a direct instantiation of our main result in Theorem 3.2.1.

3.3.4 Quantized (One-bit) Observation Models

We may also utilzie our main result to assess the estimation performance in scenarios where entry-wise observations of the matrix are quantized to few bits, or even a *single bit*, each. Such quantized observations are natural in collaborative filtering applications such as the aforementioned *Netflix* problem, where users' ratings are quantized to fixed levels. One may also envision applications in distributed estimation tasks where one seeks to estimate some underlying matrix from highly-quantized observations of a subset of its entries; here, the quantization could serve as a mechanism for enforcing global communication rate constraints (e.g., when the data is transmitted to a centralized location for inference). Our general framework would facilitate analysis of observations quantized to any of a number of levels; here, for concreteness, we consider a one-bit observation model.

Formally, given a sampling set $\mathcal{S}$ we suppose that our observations are conditionally (on $\mathcal{S}$) independent random variables described by

$$Y_{i,j} = \mathbf{1}_{\{Z_{i,j} \geq 0\}}, \quad (i,j) \in \mathcal{S}, \tag{3.26}$$

where

$$Z_{i,j} = X^*_{i,j} - W_{i,j}, \tag{3.27}$$

the $\{W_{i,j}\}_{i\in[m],j\in[n]}$ are some iid continuous zero-mean real scalar "noises" having probability density function and cumulative distribution function $f(w)$ and $F(w)$, respectively, for $w \in \mathbb{R}$, and $\mathbf{1}_{\{\mathcal{E}\}}$ denotes the indicator of the event $\mathcal{E}$ that takes the value

1 when $\mathcal{E}$ occurs and zero otherwise. Note that in this model, we assume that the individual noise realizations $\{W_{i,j}\}_{(i,j)\in\mathcal{S}}$ are unknown (but we assume that the noise *distribution* is known). Stated another way, we may interpret the observations modeled as above essentially as quantized noisy versions of the true matrix parameters (the minus sign on the $W_{i,j}$'s is merely a modeling convenience here, and is intended to simplify the exposition). Under this model, it is easy to see that each $Y_{i,j}$ is a Bernoulli random variable whose parameter is related to the true parameter through the cumulative distribution function. Specifically, note that for any fixed $(i,j) \in \mathcal{S}$, we have that $\Pr(Y_{i,j} = 1) = \Pr(W_{i,j} \leq X^*_{i,j}) = F(X^*_{i,j})$. Thus, in this scenario, we have that $\mathbf{Y}_\mathcal{S} \in \{0,1\}^{|\mathcal{S}|}$ and

$$p_{\mathbf{X}^*_\mathcal{S}}(\mathbf{Y}_\mathcal{S}) = \prod_{(i,j)\in\mathcal{S}} \left[F(X^*_{i,j})\right]^{Y_{i,j}} \left[1 - F(X^*_{i,j})\right]^{1-Y_{i,j}} \tag{3.28}$$

We will also assume here that $\mathrm{X}_{\max}$ and $F(\cdot)$ are such that $F(\mathrm{X}_{\max}) < 1$ and $F(-\mathrm{X}_{\max}) > 0$; it follows that the true Bernoulli parameters (as well as the Bernoulli parameters associated with candidate estimates $\mathbf{X} \in \mathcal{X}$) are bounded away from 0 and 1; these assumptions will allow us to avoid some pathological scenarios in our analysis.

Given the above model and assumptions, we may establish the following result; the proof is provided in Appendix 3.6.5.

Corollary 3.3.4 (Sparse Factor Matrix Completion from One-bit Observations)**.** *Let β be as in* (3.13)*, let $\mathcal{X} = \mathcal{X}'$, and let $p_{\mathbf{X}^*_\mathcal{S}}$ be of the form in* (3.28) *with $F(\mathrm{X}_{\max}) < 1$ and $F(-\mathrm{X}_{\max}) > 0$. Define*

$$c_{F,\mathrm{X}_{\max}} \triangleq \left(\sup_{|t|\leq \mathrm{X}_{\max}} \frac{1}{F(t)(1-F(t))}\right) \cdot \left(\sup_{|t|\leq \mathrm{X}_{\max}} f^2(t)\right), \tag{3.29}$$

and

$$c'_{F,\mathrm{X}_{\max}} \triangleq \inf_{|t|\leq \mathrm{X}_{\max}} \frac{f^2(t)}{F(t)(1-F(t))}, \tag{3.30}$$

and let λ be as in (3.14) *with $C_\mathrm{D} = 2c_{F,\mathrm{X}_{\max}}\mathrm{X}^2_{\max}$. The estimate $\widehat{\mathbf{X}}$ obtained via* (3.8) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} = \\ \mathcal{O}\left(\left(\frac{c_{F,\mathrm{X}_{\max}}}{c'_{F,\mathrm{X}_{\max}}}\right)\left(\frac{1}{c_{F,\mathrm{X}_{\max}}} + \mathrm{X}^2_{\max}\right)\left(\frac{n_1 r + \|\mathbf{A}^*\|_0}{m}\right)\log(n_1 \vee n_2)\right), \tag{3.31}$$

when $\mathbf{A}^*$ *is exactly sparse, having* $\|\mathbf{A}^*\|_0$ *nonzero elements. If, instead, for any* $p \leq 1$ *the columns of* $\mathbf{A}^*$ *belong to a weak-*ℓ_p *ball of radius* $\mathrm{A}_{\max}$*, then the estimate* $\widehat{\mathbf{X}}$ *obtained via* (3.8) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} = \mathcal{O}\left(\left(\frac{c_{F,\mathrm{X}_{\max}}}{c'_{F,\mathrm{X}_{\max}}}\right)\mathrm{A}_{\max}^2\left(\frac{n_2}{m}\right)^{\frac{2\alpha}{2\alpha+1}} + \right.$$
$$\left.\left(\frac{c_{F,\mathrm{X}_{\max}}}{c'_{F,\mathrm{X}_{\max}}}\right)\left(\frac{1}{c_{F,\mathrm{X}_{\max}}} + \mathrm{X}_{\max}^2\right)\left(\frac{n_1 r}{m} + \left(\frac{n_2}{m}\right)^{\frac{2\alpha}{2\alpha+1}}\right)\log(n_1 \vee n_2)\right). \tag{3.32}$$

where $\alpha = 1/p - 1/2$.

It is interesting to compare the results of (3.31) and (3.32) with the analogous results (3.16) and (3.17). Specifically, we see that our estimation error guarantees for each case exhibit the same fundamental dependence on the dimension, sparsity, and (nominal) number of measurements, with the primary difference overall arising in the form of the leading factors (that in the one-bit case depend on the specific distribution of the $W_{i,j}$ terms). That the estimation errors for rate-constrained tasks approximately mimic that of their Gaussian-corrupted counterparts was observed in earlier works on rate-constrained parameter estimation (see, e.g., [96, 97]), and more recently in [60], which considered low-rank matrix completion from one-bit measurements, using a generative model analogous to the model we consider here.

It is also worth noting that the cdf $F(\cdot)$ that we specify here could be replaced by any of a number of commonly-used *link functions*. For example, choosing $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ to be the cdf of a standard Gaussian random variable gives rise to the well-known probit model, while taking $F(x)$ to be the logistic function, $F(x) = \frac{1}{1+e^{-x}}$, leads to the logit regression model. In this sense, our results are related to classical methods on inference in generalized linear models (see, e.g., [98]); a key distinction here is that we assume both of the factors in the bilinear form to be unknown.

Finally, we briefly compare our results with the results of [60] for low-rank matrix completion from one-bit observations. In that work, the authors consider maximum-likelihood optimizations over a (convex) set of max-norm and nuclear-norm constrained matrices, and show that the estimates so-obtained satisfy

$$\frac{\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2}{n_1 n_2} = \mathcal{O}\left(C_{F,\mathrm{X}_{\max}} \mathrm{X}_{\max} \sqrt{\frac{(n_1+n_2)r}{m}} \right) \tag{3.33}$$

with high probability, where $C_{F,\mathrm{X}_{\max}}$ is a parameter that depends on the max-norm constraint cdf $F(\cdot)$ and pdf $f(\cdot)$, somewhat analogously to the leading factor of $(c_{F,\mathrm{X}_{\max}}/ c'_{F,\mathrm{X}_{\max}})$ in our bounds. It is interesting to note a main qualitative difference between that result and ours. For concreteness, let us consider the case where $\mathbf{A}^*$ is not sparse, so that we may set $\|\mathbf{A}^*\|_0 = n_2 r$ in (3.31). In that case, it is easy to see that the overall estimation error behavior predicted by our bound (3.31) scales in proportion to ratio between the number of degrees of freedom ($(n_1+n_2)r$) and the nominal number of measurements m, while the bound in [60] scales according to the square root of that ratio. The authors of [60] proceed to show that the estimation error rate they obtain is minimax optimal over their set of candidate estimates; on the other hand, our bound appears (at least up to leading factors) to be tighter for this case where $\mathbf{X}^*$ is exactly low-rank (e.g., setting $\|\mathbf{A}^*\|_0 = n_2 r$ in our bound) and $m \geq c(n_1+n_2)r$ for a constant $c > 1$. That said, our approach also enjoys the benefit of having the rank or an upper bound for it be known (and being combinatorial in nature!), while the procedure in [60] assumes only a bound on the nuclear norm of the unknown matrix. Whether our bounds here exhibit minimax-optimal estimation error rates for matrix completion under sparse factor models and for the several various likelihood models we consider here is still an open question since (to our knowledge) lower bounds for these problems have not yet been established (but are a topic of our ongoing efforts).

3.4 Experimental Evaluation

In this section we provide experimental evidence to validate the error rates established by our theoretical results. Recall (as noted above) that the original problem (3.8) we aim to solve has multiple sources of non-convexity, including the bilinear matrix factor model (i.e., $\mathbf{X} = \mathbf{DA}$), the presence of the ℓ_0 penalty, and the discretized sets $\mathcal{D}$ and $\mathcal{A}$. In what follows, we undertake a slight relaxation of (3.8) replacing the sets $\mathcal{D}$ and $\mathcal{A}$ by their convex hulls (and with slight overloading of notation in what follows, we refer to these new sets also as $\mathcal{D}$ and $\mathcal{A}$). With this relaxation, the set $\mathcal{X}$ becomes a set of all matrices $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ with bounded entries. Note that with these simplifying relaxations, the sets $\mathcal{X}$, $\mathcal{D}$ and $\mathcal{A}$ are convex.

Now, for each likelihood model, our aim is to solve a constrained maximum likelihood problem of the form

$$\begin{array}{cl} \displaystyle\min_{\mathbf{D}\in\mathbb{R}^{n_1\times r},\mathbf{A}\in\mathbb{R}^{r\times n_2}} & \displaystyle\sum_{i,j} s_{i,j}\ell(Y_{i,j}, X_{i,j}) + I_{\mathcal{X}}(\mathbf{X}) + I_{\mathcal{D}}(\mathbf{D}) + I_{\mathcal{A}}(\mathbf{A}) + \lambda\|\mathbf{A}\|_0 \\ \text{s.t.} & \mathbf{X} = \mathbf{DA}. \end{array} \tag{3.34}$$

where $\ell(Y_{i,j}, X_{i,j}) = -\log(p_{X_{i,j}}(Y_{i,j}))$ is the negative log-likelihood for the corresponding noise model, $s_{i,j}$ is a selector taking the value 1 when $(i,j) \in \mathcal{S}$ and 0 otherwise, $\lambda \geq 0$ is a regularization parameter, and each of $I_{\mathcal{X}}(\cdot)$, $I_{\mathcal{D}}(\cdot)$, and $I_{\mathcal{A}}(\cdot)$ are the indicator functions of the sets $\mathcal{X}$, $\mathcal{D}$ and $\mathcal{A}$ respectively[2] . Here, we have that each of the indicator functions is separable in the individual entries of its argument, e.g. $I_{\mathcal{X}}(\mathbf{X}) = \sum_{i,j} I_{\mathcal{X}_{i,j}}(X_{i,j})$, and similarly for the indicator functions of $\mathcal{D}$ and $\mathcal{A}$.

We propose a solution approach based on the Alternating Direction Method of Multipliers (ADMM) [62]. First we write the augmented Lagrangian of (3.34) as

$$\begin{aligned} \mathcal{L}(\mathbf{D}, \mathbf{A}, \mathbf{X}, \mathbf{\Lambda}) = \sum_{i,j} s_{i,j}\ell(Y_{i,j}, X_{i,j}) + I_{\mathcal{X}}(\mathbf{X}) + I_{\mathcal{D}}(\mathbf{D}) + I_{\mathcal{A}}(\mathbf{A}) + \\ \lambda\|\mathbf{A}\|_0 + \operatorname{tr}(\mathbf{\Lambda}(\mathbf{X} - \mathbf{DA})) + \frac{\rho}{2}\|\mathbf{X} - \mathbf{DA}\|_{\mathrm{F}}^2, \end{aligned} \tag{3.35}$$

where $\mathbf{\Lambda}$ is a matrix of Lagrange multiplier parameters and $\rho > 0$ is a parameter. Then, starting with some feasible $\mathbf{A}^{(0)}, \mathbf{D}^{(0)}, \mathbf{\Lambda}^{(0)}$ we iteratively update $\mathbf{X}$, $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{\Lambda}$ according to

$$(\mathbf{S1}:)\ \mathbf{X}^{(k+1)} := \arg\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \mathcal{L}(\mathbf{D}^{(k)}, \mathbf{A}^{(k)}, \mathbf{X}, \mathbf{\Lambda}^{(k)}) \tag{3.36}$$

$$(\mathbf{S2}:)\ \mathbf{A}^{(k+1)} := \arg\min_{\mathbf{A}\in\mathbb{R}^{r\times n_2}} \mathcal{L}(\mathbf{D}^{(k)}, \mathbf{A}, \mathbf{X}^{(k+1)}, \mathbf{\Lambda}^{(k)}) \tag{3.37}$$

$$(\mathbf{S3}:)\ \mathbf{D}^{(k+1)} := \arg\min_{\mathbf{D}\in\mathbb{R}^{n_1\times r}} \mathcal{L}(\mathbf{D}, \mathbf{A}^{(k+1)}, \mathbf{X}^{(k+1)}, \mathbf{\Lambda}^{(k)}) \tag{3.38}$$

$$(\mathbf{S4}:)\ \mathbf{\Lambda}^{(k+1)} = \mathbf{\Lambda}^{(k)} + \rho(\mathbf{X}^{(k+1)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)}), \tag{3.39}$$

[2] Recall that the indicator function is defined as a function that takes values 0 or ∞ depending on whether its argument is an element of the set described as the subscript.

until convergence, which here is quantified in terms of when norms of primal and dual residuals become sufficiently small (along the lines of the criteria described in [62]). Next we describe how to solve each of these steps.

Solving **S1** involves the following optimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \sum_{i,j} s_{i,j}\ell(Y_{i,j},X_{i,j}) + I_{\mathcal{X}}(\mathbf{X}) + \mathrm{tr}\left(\mathbf{\Lambda}^{(\mathrm{k})}(\mathbf{X}-\mathbf{D}^{(\mathrm{k})}\mathbf{A}^{(\mathrm{k})})\right) + \frac{\rho}{2}\|\mathbf{X}-\mathbf{D}^{(\mathrm{k})}\mathbf{A}^{(\mathrm{k})}\|_{\mathrm{F}}^2,$$

which after completing the square and ignoring constant terms is equivalent to

$$\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \quad \sum_{i,j} s_{i,j}\ell(Y_{i,j},X_{i,j}) + I_{\mathcal{X}}(\mathbf{X}) + \frac{\rho}{2}\left\|\mathbf{X}-\mathbf{D}^{(k)}\mathbf{A}^{(k)} + \frac{\mathbf{\Lambda}^{(k)}}{\rho}\right\|_F^2. \quad (3.40)$$

Due to the assumed separability of the indicator function, the above problem is separable in each entry $X_{i,j}$ and the entries can be updated in parallel by solving the following scalar convex optimization problem for each entry. When $\ell(y,x)$ is a convex function of x, the solution is given by

$$\begin{aligned} X_{i,j}^{(k+1)} &= \mathrm{Proj}_{\mathcal{X}_{i,j}}\left[\mathrm{prox}_{s_{i,j},\ell}\left((\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} - \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho};\rho,Y_{i,j}\right)\right], \\ &= \begin{cases} \mathrm{Proj}_{\mathcal{X}_{i,j}}\left[\mathrm{prox}_{\ell}\left((\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} - \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho};\rho,Y_{i,j}\right)\right], & \text{if } s_{i,j}=1 \\ \mathrm{Proj}_{\mathcal{X}_{i,j}}\left[(\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} - \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho}\right], & \text{otherwise} \end{cases} \quad (3.41) \end{aligned}$$

where $\mathrm{prox}_\ell(z;\rho,y) = \arg\min_{x\in\mathbb{R}} \ \ell(y,x) + \frac{\rho}{2}(x-z)^2$ is the *proximal operator* of the loss function $\ell(y,\cdot)$ and $\mathrm{Proj}_{\mathcal{X}_{i,j}}(x)$ is the projection[3] of the scalar x onto the set $\mathcal{X}_{i,j}$. For several of the loss functions we consider, the proximal operator can be computed in closed form; for the one-bit settings we can use Newton's second order method (or gradient descent) to numerically evaluate it as described later. A table of proximal operators for the various losses we consider here is provided in Table 3.1.

Completing the square and ignoring the constant terms the subproblem **S2** is equivalent to

$$\mathbf{A}^{(k+1)} = \arg\min_{\mathbf{A}\in\mathbb{R}^{r\times n_2}} \quad I_{\mathcal{A}}(\mathbf{A}) + \lambda\|\mathbf{A}\|_0 + \frac{\rho}{2}\left\|\mathbf{X}^{(k+1)} - \mathbf{D}^{(k)}\mathbf{A} + \frac{\mathbf{\Lambda}^{(k)}}{\rho}\right\|_F^2 .(3.42)$$

[3] Here, this set is just an interval and the projection operator returns x if $x \in \mathcal{X}_{i,j}$ or the nearest endpoint of the interval $\mathcal{X}_{i,j}$ otherwise.

	$\ell(y,x)$	$\mathrm{prox}_\ell(z;\rho,y)$
Gaussian	$\frac{(y-x)^2}{2\sigma^2}$	$\frac{y+\sigma^2\rho z}{1+\sigma^2\mu}$
Poisson	$x - y\log(x)$	$\frac{\rho z-1+\sqrt{(\rho z-1)^2+4\rho y}}{2\rho}$
Laplace	$\lambda\lvert y-x\rvert$	$y - \mathrm{Soft}(z-y,\lambda/\rho)$
One-bit	$-y\log(F(x)) - (1-y)\log(1-F(x))$	$(\textit{Newton's method})$

Table 3.1: Expressions for $\mathrm{prox}_\ell(z;\rho,y) = \arg\min_{x\in\mathbb{R}}\ \ell(y,x) + \frac{\rho}{2}(x-z)^2$ for different $\ell(y,x)$, corresponding to negative log-likelihoods for the models we examine. Here, for $\lambda > 0$, $\mathrm{soft}(x,\lambda) = \mathrm{sgn}(x)\max\{|x|-\lambda, 0\}$

In order to solve this problem we adopt the constrained iterative hard thresholding approach from [99], as outlined in Algorithm 1. Finally, after completing the square and ignoring the constant terms, we see that the subproblem **S3** is equivalent to

$$\mathbf{D}^{(k+1)} = \arg\min_{\mathbf{D}\in\mathbb{R}^{n_1\times r}} \qquad I_{\mathcal{D}}(\mathbf{D}) + \frac{\rho}{2}\left\|\mathbf{X}^{(k+1)} - \mathbf{D}\mathbf{A}^{(k+1)} + \frac{\mathbf{\Lambda}^{(k)}}{\rho}\right\|_F^2, \tag{3.43}$$

which we solve here by projected Newton gradient descent algorithm, described in Algorithm 2. Our overall algorithmic approach is summarized in Algorithm 3.

Algorithm 1 A_IHT($\mathbf{X}$, $\mathbf{D}$, $\mathbf{Z}$, ϵ) – For solving $\min_{\mathbf{A}\in\mathbb{R}^{n_1\times p}} I_{\mathcal{A}}(\mathbf{A}) + \lambda\|\mathbf{A}\|_0 + \frac{\rho}{2}\|\mathbf{Z}-\mathbf{D}\mathbf{A}\|_F^2$

```
Inputs: $\mathbf{X}, \mathbf{D}, \mathbf{Z}, \epsilon, \rho$
Initialize: $\mathbf{A}^{(0)} = \mathbf{0}$
  repeat
    $\mathbf{Y}^{(k+1)} = \mathbf{A}^{(k)} - \mathbf{D}^T(\mathbf{D}\mathbf{A}^{(k)} - \mathbf{Z})/\|\mathbf{D}\|_2^2$
    Update: $Y_{i,j}^{(k+1)} = 0$ if $\left|Y_{i,j}^{(k+1)}\right| \leq \sqrt{\frac{2\lambda}{\rho\|\mathbf{D}\|_2^2}}$.
      if $Y_{i,j}^{(k+1)} \in \mathcal{A}_{i,j}$ :
          $A_{i,j}^{(k+1)} = Y_{i,j}^{(k+1)}$;
      else:
          $A_{i,j}^{(k+1)} = \arg\min_{x\in\mathcal{A}_{i,j}} \left\{x^2 - 2x\left[A_{i,j}^{(k)} - \frac{((\mathbf{D}^T\mathbf{D}\mathbf{A}^{(k)})_{i,j} - (\mathbf{D}^T\mathbf{Z})_{i,j})}{\|\mathbf{D}\|_2^2}\right]\right\}$
  until $\frac{\|\mathbf{A}^{(k+1)}-\mathbf{A}^{(k)}\|_F}{\|\mathbf{A}^{(k)}\|_F} \leq \epsilon$
Output: $\mathbf{A} = \mathbf{A}^{(k+1)}$
```

Algorithm 2 D_Newton($\mathbf{X}$, $\mathbf{A}$, $\mathbf{Z}$, ϵ) – For solving $\min_{\mathbf{D}\in\mathbb{R}^{n_1\times p}} I_{\mathcal{D}}(\mathbf{D}) + \frac{\rho}{2}\left\|\mathbf{Z}-\mathbf{DA}\right\|_F^2$

```
Inputs: $\mathbf{X}, \mathbf{A}, \mathbf{Z}, \epsilon, \rho$
Initialize: $\mathbf{D}^{(0)} = \mathbf{0}$
  repeat
    $\mathbf{D}^{(k+1)} = \mathrm{Proj}_{\mathcal{D}}\left[\mathbf{D}^{(k)} - \rho\left(\mathbf{D}^{(k)}\mathbf{A} - \mathbf{Z}\right)\mathbf{A}^T\left(\rho\mathbf{A}\mathbf{A}^T + \delta\mathbf{I}\right)^{-1}\right]$
  until $\frac{\|\mathbf{D}^{(k+1)} - \mathbf{D}^{(k)}\|_F}{\|\mathbf{D}^{(k)}\|_F} \leq \epsilon$
Output: $\mathbf{D} = \mathbf{D}^{(k+1)}$
```

Algorithm 3 ADMM algorithm for solving problem (3.34)

```
Inputs: $\epsilon_1, \epsilon_2,\ \Delta_1,\ \Delta_2,\ \Delta_1^{\text{stop}},\ \Delta_2^{\text{stop}},\ \eta,\ \rho^{(0)} > 0$
Initialize: $\mathbf{D}^{(0)} \in \mathcal{D}$ , $\mathbf{A}^{(0)} \in \mathcal{A}$, $\mathbf{\Lambda}^{(0)}$.
  repeat
    $\mathbf{X}_{i,j}^{(k+1)} = \mathrm{Proj}_{\mathcal{X}}\left[\mathrm{prox}_{s_{i,j}\ell}\left((\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} - \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho^{(k)}}; \rho^{(k)}, Y_{i,j}\right)\right]$
    $\mathbf{A}^{(k+1)} := \mathrm{A\_IHT}\left(\mathbf{X}_{(k+1)}, \mathbf{D}^{(k)}, \mathbf{X}^{(k+1)} + \mathbf{\Lambda}^{(k)}/\rho^{(k)}, \epsilon_1\right)$
    $\mathbf{D}^{(k+1)} := \mathrm{D\_Newton}\left(\mathbf{X}_{(k+1)}, \mathbf{A}^{(k+1)}, \mathbf{X}^{(k+1)} + \mathbf{\Lambda}^{(k)}/\rho^{(k)}, \epsilon_2\right)$
    $\mathbf{\Lambda}^{(k+1)} = \mathbf{\Lambda}^{(k)} + \rho^{(k)}(\mathbf{X}^{(k+1)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)})$
    Set $\Delta_1 = \|\mathbf{X}^{(k+1)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)}\|_F$ and $\Delta_2 = \rho^{(k)} \cdot \|\mathbf{D}^{(k)}\mathbf{A}^{(k)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)}\|_F$
    $\rho^{(k+1)} = \begin{cases} \eta \cdot \rho^{(k)}, & \text{if } \Delta_1 \geq 10 \cdot \Delta_2 \\ \rho^{(k)}/\eta, & \text{if } \Delta_2 \geq 10 \cdot \Delta_1 \\ \rho^{(k)}, & \text{otherwise} \end{cases}$
  until $\Delta_1 \leq \Delta_1^{\text{stop}}$ and $\Delta_2 \leq \Delta_2^{\text{stop}}$
Output: $\mathbf{D} = \mathbf{D}^{(k+1)}$ and $\mathbf{A} = \mathbf{A}^{(k+1)}$
```

3.4.1 Experiments

We perform experimental validation of our theoretical results on synthetic data for two different scenarios, corresponding to when the columns of the matrix $\mathbf{A}^*$ are k-sparse, and when each belongs to a weak-l_p ball. For each scenario we construct the true data matrices $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$ by individually constructing the matrices $\mathbf{D}^*$ and $\mathbf{A}^*$ (as described below), where the entries of the true matrices $\mathbf{X}^*$, $\mathbf{D}^*$, and $\mathbf{A}^*$ are bounded in $[\mathrm{X}^*_{\min},\ \mathrm{X}^*_{\max}]$, $[\mathrm{D}^*_{\min},\ \mathrm{D}^*_{\max}]$ and $[\mathrm{A}^*_{\min},\ \mathrm{A}^*_{\max}]$ respectively.

We generate the $\mathbf{D}^*$ matrix by first generating a Gaussian random matrix of size $n_1 \times r$ whose entries are distributed as $\mathcal{N}(0,1)$, then multiplying each element by $(\mathrm{D}^*_{\max} - \mathrm{D}^*_{\min})$ to avoid pathological scaling issues. Finally, we project the resulting scaled matrix onto the set $\mathcal{D}$, which here is done by truncating all the entries bigger than $\mathrm{D}^*_{\max}$ to $\mathrm{D}^*_{\max}$ and truncating all the entries smaller than $\mathrm{D}^*_{\min}$ to $\mathrm{D}^*_{\min}$. We construct sparse $\mathbf{A}^*$ by generating a Gaussian random matrix of size $r \times n_2$, multiplying it by $(\mathrm{A}^*_{\max} - \mathrm{A}^*_{\min})/3$, and projecting it onto the set $\mathcal{A}$. Then we randomly select $r - k$ locations from each column of the resulting matrix and set the corresponding entries to 0. For the approximately sparse $\mathbf{A}^*$, we generate each column to be a randomly permuted version of $\{\mathrm{A}^*_{\max} \cdot i^{-1/p}\}_{i=1}^{r}$ with random signs (except for the Poisson likelihood case, where each column of $\mathbf{A}^*$ has nonnegative elements).

We define the set $\mathcal{X}$ such that each entry of $\mathbf{X}$ is bounded in the range $[\mathrm{X}_{\min},\ \mathrm{X}_{\max}]$, $\mathcal{D}$ and $\mathcal{A}$ are the set of all matrices $\mathbf{D} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times n_2}$ whose entries are bounded in the range $[\mathrm{D}_{\min},\ \mathrm{D}_{\max}]$ and $[\mathrm{A}_{\min},\ \mathrm{A}_{\max}]$ respectively. It is important to note that in general the actual bounds on the magnitude of the entries of true matrices (for e.g., $\mathrm{D}^*_{\min}$, $\mathrm{A}^*_{\max}$ etc.) are unknown, and therefore during optimization we might have to use their approximations (which here are denoted as $\mathrm{D}_{\min}$, $\mathrm{A}_{\max}$ etc.) to define the feasible sets $\mathcal{X}$, $\mathcal{D}$ and $\mathcal{A}$. Our specific choices of parameters for the four different likelihoods considered in this paper are summarized in Table 3.2.

Now, our experimental approach is as follows. For sparse and nearly-sparse (with columns belonging to a weak ℓ_p ball with $p = 1/3$) coefficient matrices $\mathbf{A}^*$ we generate a corresponding matrix $\mathbf{X}^*$ as above. Then, for each of a number of regularization parameters $\lambda > 0$ and and sampling rates $\gamma \in (0, 1]$ we perform 20 trials of the following experiment: we generate $\mathcal{S}$ according to the independent Bernoulli(γ) model, obtain noisy observations of $\mathbf{X}^*$ according to the (3.7), use Algorithm 3 to obtain[4] an estimate $\widehat{\mathbf{X}} = \widehat{\mathbf{D}}\widehat{\mathbf{A}}$, and compute its approximation error $\frac{\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2}{n_1 n_2}$. We then compute the empirical average of the errors over the 20 trials for each setting. Fig. 3.1 shows the results of this experiment for the Gaussian, Laplace and Poisson likelihood models. The plots depict the empirical average (over 20 trials) per-element error as a function of sampling rate on a log-log scale; the curves shown are corresponding to the best (lowest) errors achieved over all of the regularization parameters λ we examined. The first row corresponds to exactly sparse $\mathbf{A}^*$ matrices and the plots in the second row corresponds to settings where $\mathbf{A}^*$ is approximately sparse. The three columns correspond to three different regimes for the Gaussian and Laplace settings (we chose the parameters σ and

[4] For Algorithm 3 we set $\epsilon_1 = \epsilon_2 = 10^{-7}$, $\Delta_1^{\text{stop}} = \Delta_2^{\text{stop}} = 10$, $\eta = 1.05$ and $\rho^{(0)} = 0.001$.

Parameters \ Likelihood	**Gaussian**	**Laplace**
$n_1 \times n_2$	100×1000	100×1000
$r,\ k,\ p$	$20,\ 8,\ 1/3$	$20,\ 8,\ 1/3$
$[\mathrm{D}^*_{\mathrm{min}},\ \mathrm{D}^*_{\mathrm{max}}]$	$[-1,\ 1]$	$[-1,\ 1]$
$[\mathrm{A}^*_{\mathrm{min}},\ \mathrm{A}^*_{\mathrm{max}}]$	$[-20,\ 20]$	$[-20,\ 20]$
$[\mathrm{D}_{\mathrm{min}},\ \mathrm{D}_{\mathrm{max}}]$	$[-2,\ 2]$	$[-2,\ 2]$
$[\mathrm{A}_{\mathrm{min}},\ \mathrm{A}_{\mathrm{max}}]$	$[-40,\ 40]$	$[-40,\ 40]$
$[\mathrm{X}_{\mathrm{min}},\ \mathrm{X}_{\mathrm{max}}]$	$[-2 \cdot \mathrm{X}^*_{\mathrm{min}},\ 2 \cdot \mathrm{X}^*_{\mathrm{max}}]$	$[-2 \cdot \mathrm{X}^*_{\mathrm{min}},\ 2 \cdot \mathrm{X}^*_{\mathrm{max}}]$
Parameters \ Likelihood	**Poisson**	**One-bit**
$n_1 \times n_2$	100×1000	1000×1000
$r,\ k,\ p$	$20,\ 8,\ 1/3$	$5,\ 2,\ 1/3$
$[\mathrm{D}^*_{\mathrm{min}},\ \mathrm{D}^*_{\mathrm{max}}]$	$[0.1,\ 1]$	$[-1,\ 1]$
$[\mathrm{A}^*_{\mathrm{min}},\ \mathrm{A}^*_{\mathrm{max}}]$	$[0,\ 40]$	$[-20,\ 20]$
$[\mathrm{D}_{\mathrm{min}},\ \mathrm{D}_{\mathrm{max}}]$	$[-2,\ 2]$	$[-2,\ 2]$
$[\mathrm{A}_{\mathrm{min}},\ \mathrm{A}_{\mathrm{max}}]$	$[-80,\ 80]$	$[-40,\ 40]$
$[\mathrm{X}_{\mathrm{min}},\ \mathrm{X}_{\mathrm{max}}]$	$[0,\ 2 \cdot \mathrm{X}^*_{\mathrm{max}}]$	$[-2 \cdot \mathrm{X}^*_{\mathrm{min}},\ 2 \cdot \mathrm{X}^*_{\mathrm{max}}]$

Table 3.2: Experimental parameters for different likelihood models we examine. Here $\mathrm{X}^*_{\mathrm{min}} = \min_{i,j} X^*_{i,j}$ and $\mathrm{X}^*_{\mathrm{max}} = \|\mathbf{X}^*\|_{\mathrm{max}}$.

τ for the Gaussian and Laplace settings, respectively, to yield identical variances; the first column corresponds to $\sigma = 0.5$ and $\tau = \sqrt{8}$, the second column corresponds to $\sigma = 1$ and $\tau = \sqrt{2}$, and the third column corresponds to $\sigma = 2$ and $\tau = 1/\sqrt{2}$).

A few interesting points are worth noting here. First, for the case where $\mathbf{A}^*$ is exactly sparse, our theoretical results predict the error decay be inversely proportional to the nominal sampling rate γ; viewed on a log-log scale, this would correspond to the error decay having slope -1. Our experimental results provide some evidence to validate our analysis, at least in the settings where the sampling rate $\gamma > 0.4$ – there, the slopes of the error decays for each of the likelihood models is indeed approximately -1. For the settings where the columns of $\mathbf{A}^*$ belong to a weak-ℓ_p (with $p = 1/3$) our theory predicts that the slope of the error decay (on a log-log scale) be at least $-5/6$ for the Gaussian-noise and Poisson-distributed cases, and at least $(-2/3)$ for the Laplace-noise case. For our experiments here, it appears that the error decay in these approximately-sparse settings is actually a bit faster than predicted by the theory, as the error appears to

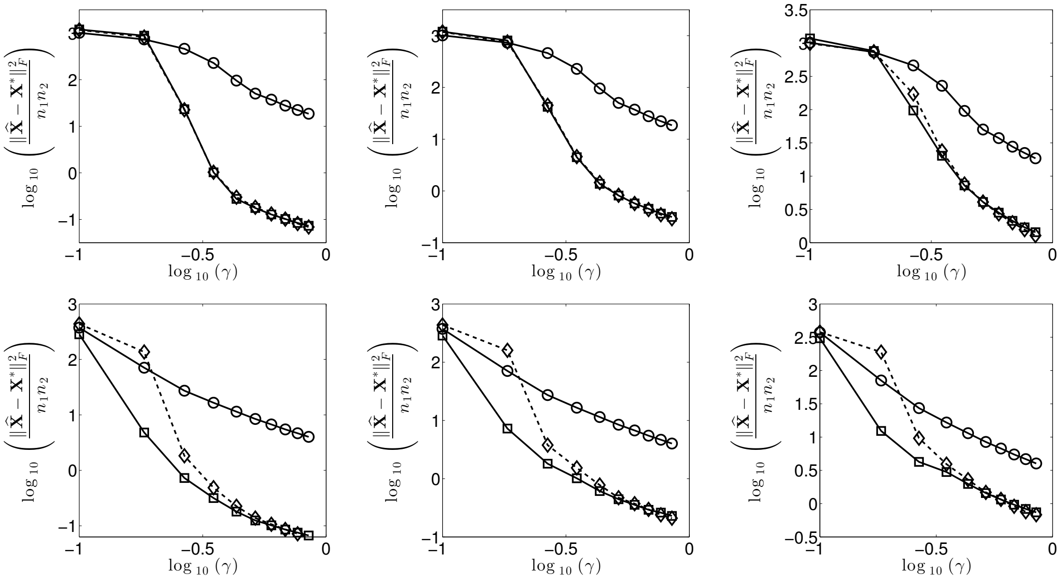

Figure 3.1: Results of synthetic experiments for matrix completion with Gaussian, Laplace and Poisson likelihoods: —-□—- is Gaussian, – – ◇ – – is Laplace, and —-○—- is Poisson. Top row corresponds to sparse-factor model with $k = 8$ while the bottom row corresponds to weak-l_p model with $p = 1/3$. Column 1 corresponds to $\sigma^2 = (0.5)^2$ (for Laplace $\tau = \sqrt{8}$), column 2 corresponds to $\sigma^2 = (1)^2$ (for Laplace $\tau = \sqrt{2}$) and column 3 corresponds to $\sigma^2 = (2)^2$ (for Laplace $\tau = 1/\sqrt{2}$). Here $n_1 = 100$, $n_2 = 1000$ and $r = 20$.

decay with a slope of approximately -1. That said, it is worth noting that our predicted rate in these cases was obtained essentially by a (squared) bias-variance tradeoff, so quantify a kind of worst-case behavior that may not always be observed in practice.

We also evaluated the performance in this setting for a one-bit observation model, using an analogous experimental setting as above. Here, we used the logistic cumulative distribution function as the link function, i.e., $F(x) = \frac{1}{1+e^{-x/s}}$ where $s = \frac{\sqrt{3}\cdot\sigma}{\pi}$ and σ is a parameter that could be viewed as additive noise standard deviation[5] , for the

[5] For this link function the proximal operator is $\text{prox}_\ell(z;\rho,y) = \arg\min_{x\in\mathbb{R}} \; -y\log(F(x)) - (1-y)\log(1-F(x)) + \frac{\rho}{2}(x-z)^2$, which in general is not solvable in closed form. Here, we resort to Newton's gradient descent algorithm – rewriting the problem as $\text{prox}_\ell(z;\rho,y) = \arg\min_{x\in\mathbb{R}} \; G(x)$, where $G(x) = -y\log(F(x)) - (1-y)\log(1-F(x)) + \frac{\rho}{2}(x-z)^2$, it is easy to show that the gradient is $\nabla G(x) = -\frac{y}{s} + \frac{F(x)}{s} + \rho(x-z)$ and the Hessian is $\nabla^2 G(x) = \frac{F(x)(1-F(x))}{s^2} + \rho$. We can then iteratively solve for $\text{prox}_\ell(z;\rho,y)$ by Newton steps (starting from a random $x^{(0)}$) of the form $x^{(k+1)} = x^{(k)} - \frac{\nabla G(x)}{\nabla^2 G(x)}$ until convergence (here, until $\|x^{(k+1)} - x^{(k)}\| \leq 10^{-7}$).

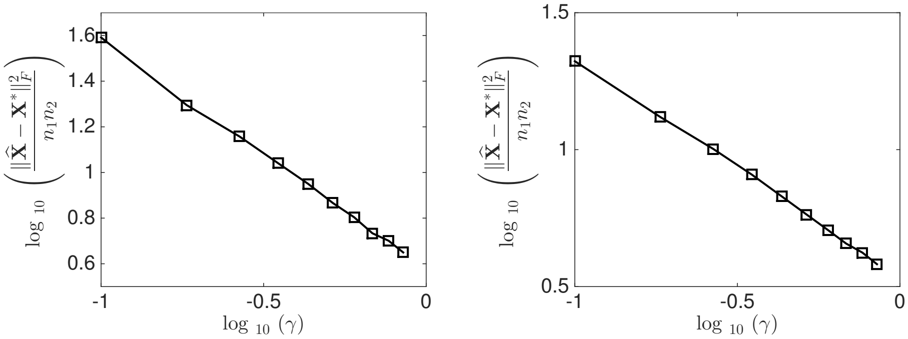

Figure 3.2: Results of synthetic experiments for one-bit matrix completion under sparse factor models, using the logistic link function. The left panel corresponds to the case where the $\mathbf{A}^*$ matrix is exactly sparse; the right when columns of $\mathbf{A}^*$ lie in a weak-l_p ball with $p = 1/3$.

specific choice $\sigma = 0.1$. Fig. 3.2 shows the error results for this case, with the first plot corresponds to sparse $\mathbf{A}^*$ and the second to when each column of $\mathbf{A}^*$ lies in a weak-ℓ_p ball with $p = 1/3$. As in the previous experiments, it appears here that the slope of the error decay is approximately -1 in each case. Note that we adapted the experimental setting here to be more amenable to this more difficult estimation regime (specifically, we consider slightly larger matrices but having smaller rank and fewer nonzeros per column of the factor $\mathbf{A}^*$, as outlined in Table 3.2, so that the number of observations per parameter to be estimated is larger than in the previous three experimental settings).

3.5 Discussion and Conclusions

We conclude with a brief discussion of our results and potentially interesting future directions.

3.5.1 Extensions to Other Data Models

Each of our theoretical results above follow essentially from the specialization of a more general result (appearing below as Lemma 3.6.1) to the case of sparse factor models. It is interesting to note that this lemma may also be specialized (in a straightforward manner) to any of a number of other interesting factor models (e.g., non-negative matrix factorizations, factorizations where each factor may be sparse, etc.) under the same general observation models we consider here. Further, while we provide Lemma 3.6.1 specifically for the case of matrix completion, the essential analysis extends (simply) to higher-order structures (i.e., tensors) as well.

3.5.2 Convexification?

As discussed in several points in the preceding sections, the optimization associated with the estimators we consider here is non-convex on account of several factors, including the presence of the ℓ_0 term in the objective, our optimization over a discretized set, and more fundamentally, the fact that we perform inference in a general bilinear model, where both factors are unknown. Resolving ourselves, then, to seek only local optima of the corresponding optimizations allows us to bring to bear alternating direction method of multipliers techniques, in which the ℓ_0-based optimization *subproblems* may be solved efficiently. Interestingly, within this framework we may also directly incorporate the constraints that the matrix factor elements each come from a discretized set (indeed, this would correspond to choosing set indicator functions that take the value ∞ outside of the discretized sets over which we seek to optimize). We did not pursue this latter condition in our simulations, assuming instead that the discretization of each of the elements be "sufficiently fine" so that we may solve the optimization numerically at machine precision (and replace the discretized sets for the candidate matrix factors by their convex hulls).

The fact that we can (locally) handle the ℓ_0 constraints within the ADMM framework notwithstanding, it is interesting to consider whether there is any benefit to relaxing this constraint to a convex surrogate (e.g., replacing the ℓ_0 penalty with an ℓ_1 penalty). The resulting procedure would still be jointly non-convex in the matrix factors, but could be addressed within a similar algorithmic framework to the one we propose above. Analytically, methods that prescribe optimization over a convex set comprised of the Cartesian product of a set $\mathcal{D}$ of matrices whose elements satisfy a max-norm constraint and a set $\mathcal{A}$ of matrices whose columns satisfy an ℓ_1-constraint may be amenable to analysis using entropy-based methods that can be employed to analyze estimation error performance by bounding suprema of empirical processes indexed by elements of the feasible set of candidate estimates – see, e.g., [100–102]. It would be interesting to see whether analyses along these yield substantially different results than our analysis here;

analyses along these lines are a subject of our ongoing work and will be reported in a subsequent effort.

In the meantime, it is interesting to examine (albeit, empirically) whether our algorithmic approach yields significantly different performance if we replace the ℓ_0 regularization term by an ℓ_1 term. To provide some insight into this, we consider a problem of completing a 50×500 matrix $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$, where $\mathbf{D}^*$ is 50×10 and $\mathbf{A}^*$ is 10×500 and sparse, having 4 nonzero elements per column. We consider Gaussian noise-corrupted observations obtained at a subset of locations of $\mathbf{X}^*$ (generated according to the independent Bernoulli model), and three different reconstruction approaches: the first is the algorithmic approach described in the previous section, the second is a slight variation of our proposed approach where we replace the ℓ_0 penalty by an ℓ_1 penalty (and replace the corresponding inference step with an accelerated first-order method as in [103]), and the the third method is a more standard low-rank recovery obtained via *nuclear-norm* regularization, as $\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{Y}_{\mathcal{S}} - \mathbf{X}_{\mathcal{S}}\|_F^2 + \lambda\|\mathbf{X}\|_*$. For each method, we examined a range of possible values for the regularization parameter, and selected the reconstruction corresponding (clairvoyantly) to the best choice for each method. The results, provided in Figure 3.3, show that the best-performing ℓ_0 and ℓ_1 regularized sparse factor completion methods perform comparably, while both achieve (slightly) lower error than the best nuclear norm regularized completion estimate. Of course, as noted above, our algorithmic approach identifies (at best) a local minimum of the overall non-convex problem we aim to solve, but even at that, it is encouraging to see that the ADMM-based optimization(s) identify good-quality estimates.

It is also interesting to consider an alternative, more essential, convexification of our problem of interest here, using the machinery of *atomic norms* as introduced in [104]. Specifically, one may view matrices adhering to the sparse factor models we investigate here as sums of rank-one matrices formed as outer products between a (non-sparse) $n_1 \times 1$ vector and a (sparse) $n_2 \times 1$ vector. Following [104], one can consider the convex hull of the set of all such rank-one atoms having unit (Frobenius) norm as the unit-ball for a norm that serves as a regularizer for matrices representable by weighted sums of only a few atoms. A very recent work [105] has begun to identify properties of atomic norms so-formed, and extensions to the cases where both of the vectors may be sparse, and have established some estimation guarantees for recovering simple matrices

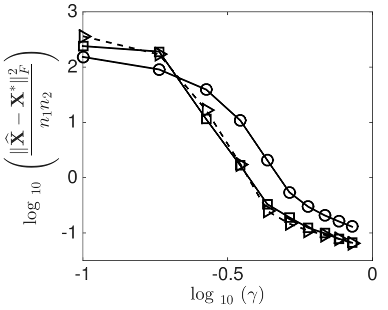

Figure 3.3: Comparison between sparse-factor and nuclear-norm-regularized matrix completion methods. The curves are: our proposed procedure with ℓ_0 regularizer ($\square$), the ℓ_1 regularized variant of our approach ($\triangleright$), and nuclear norm regularized low-rank matrix completion ($\circ$). The sparse factor completion methods perform similarly, and both achieve a lower error than the best nuclear-norm regularized estimate for sampling rates $\gamma \geq 10^{-0.5} \approx 30\%$.

(comprised of a single rank-one outer product of sparse vectors) from a collection of Gaussian measurements. Interestingly, the authors of [105] note that resulting inference procedures using their so-called (k, q)-norm (formed from atoms that are rank-one outer products between k-sparse and q-sparse vectors), while convex, may still be computationally intractable (even NP-hard)! At any rate, it would be quite interesting to extend this approach to the entry-wise sampling models and various likelihood models we consider here, and we defer investigations along these lines to a future work.

3.5.3 Lower Bounds

Our error bounds here provide some insight into the performance of sparsity-penalized maximum likelihood estimation approaches to sparse factor matrix completion tasks. To the best of our knowledge, lower bounds on the achievable mean-square estimation error for these tasks have not been established, but would be a valuable complement to place our results here into a broader context. Efforts along these lines are ongoing, and will be reported in a future work.

3.6 Appendix

3.6.1 Proof of Theorem 3.2.1

Our proof of Theorem 3.2.1 is based on an application of the following general lemma, which we prove in Appendix 3.6.6.

Lemma 3.6.1. *Let* $\mathbf{X}^*$ *be an* $n_1 \times n_2$ *matrix whose elements we aim to estimate, and let* $\mathcal{X}$ *be a countable collection of candidate reconstructions* $\mathbf{X}$ *of* $\mathbf{X}^*$*, each with corresponding penalty* $\text{pen}(\mathbf{X}) \geq 1$*, so that the collection of penalties satisfies the summability condition* $\sum_{\mathbf{X}\in\mathcal{X}} 2^{-\text{pen}(\mathbf{X})} \leq 1$*.*

Fix an integer m *with* $4 \leq m \leq n_1 n_2$*, let* $\gamma = m(n_1 n_2)^{-1}$*, generate a sampling set* $\mathcal{S}$ *according to the independent Bernoulli(*γ*) model so that each* $(i,j) \in [n_1] \times [n_2]$ *is included in* $\mathcal{S}$ *independently with probability* γ*, and obtain corresponding observations* $\mathbf{Y}_{\mathcal{S}} \sim p_{\mathbf{X}^*_{\mathcal{S}}} = \prod_{(i,j)\in\mathcal{S}} p_{X^*_{i,j}}$*, which are assumed to be conditionally independent given* $\mathcal{S}$*. Then, if* C_{D} *is any constant satisfying*

$$C_{\text{D}} \geq \max_{\mathbf{X}\in\mathcal{X}} \max_{(i,j)\in[n_1]\times[n_2]} \text{D}(p_{X^*_{i,j}} \| p_{X_{i,j}}), \tag{3.44}$$

we have that for any

$$\xi \geq \left(1 + \frac{2C_{\text{D}}}{3}\right) \cdot 2\log 2, \tag{3.45}$$

the complexity penalized maximum likelihood estimator

$$\widehat{\mathbf{X}}^{\xi} = \widehat{\mathbf{X}}^{\xi}(\mathcal{S}, \mathbf{Y}_{\mathcal{S}}) = \arg\min_{\mathbf{X}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) + \xi \cdot \text{pen}(\mathbf{X})\right\}, \tag{3.46}$$

satisfies the (normalized, per-element) error bound

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log \text{A}(p_{\widehat{\mathbf{X}}^{\xi}}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \\ 3 \cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{\frac{\text{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}})}{n_1 n_2} + \left(\xi + \frac{4C_{\text{D}}\log 2}{3}\right)\frac{\text{pen}(\mathbf{X})}{m}\right\} + \frac{8C_{\text{D}}\log m}{m}, \tag{3.47}$$

where, as denoted, the expectation is with respect to the joint distribution of $\mathcal{S}$ *and* $\mathbf{Y}_{\mathcal{S}}$*.*

In order to use this result here, we need to define penalties $\text{pen}(\mathbf{X}) \geq 1$ on candidate reconstructions $\mathbf{X}$ of $\mathbf{X}^*$, so that for every subset $\mathcal{X}$ of the set $\mathcal{X}'$ specified in the conditions of Theorem 3.2.1 the summability condition $\sum_{\mathbf{X}\in\mathcal{X}} 2^{-\text{pen}(\mathbf{X})} \leq 1$ holds. To this end, we will use the fact that for any $\mathcal{X} \subseteq \mathcal{X}'$ we always have $\sum_{\mathbf{X}\in\mathcal{X}} 2^{-\mathrm{pen}(\mathbf{X})} \leq \sum_{\mathbf{X}\in\mathcal{X}'} 2^{-\mathrm{pen}(\mathbf{X})}$; thus, it suffices for us to show that for the specific set $\mathcal{X}'$ described in Section 3.2,

$$\sum_{\mathbf{X}\in\mathcal{X}'} 2^{-\mathrm{pen}(\mathbf{X})} \leq 1. \tag{3.48}$$

Note that the condition (3.48) is the well-known Kraft-McMillan Inequality for coding elements of $\mathcal{X}'$ with an alphabet of size 2, which is satisfied automatically if we choose the penalties to be *code lengths* for some uniquely decodable binary code for the elements $\mathbf{X} \in \mathcal{X}'$; see [35]. This interpretation will provide us with a *constructive* approach to designing penalties, as we will see below.

Now, consider any discretized matrix factors $\mathbf{D} \in \mathcal{D}$ and $\mathbf{A} \in \mathcal{A}$, as described in Section 3.2. Let us fix an ordering of the indices of elements of $\mathbf{D}$ and encode the amplitude of each element using $\log_2 L_{\mathrm{lev}}$ bits, and for $L_{\mathrm{loc}} \triangleq 2^{\lceil \log_2 rn_2 \rceil}$ we encode each *nonzero* element of $\mathbf{A}$ using $\log_2 L_{\mathrm{loc}}$ bits to denote its location and $\log_2 L_{\mathrm{lev}}$ bits for its amplitude. With this strategy, a total of $n_1 r \log_2 L_{\mathrm{lev}}$ bits are used to encode $\mathbf{D}$ and matrices $\mathbf{A}$ having $\|\mathbf{A}\|_0$ nonzero entries are encoded using $\|\mathbf{A}\|_0(\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}})$ bits. Now, we let $\mathcal{X}''$ be the set of all such $\mathbf{X} = \mathbf{DA}$, and let the code for each $\mathbf{X}$ be the concatenation of the (fixed-length) code for $\mathbf{D}$ followed by the (variable-length) code for $\mathbf{A}$. It follows that we may assign penalties $\mathrm{pen}(\mathbf{X})$ to all $\mathbf{X} \in \mathcal{X}''$ whose lengths satisfy

$$\mathrm{pen}(\mathbf{X}) = n_1 r \log_2 L_{\mathrm{lev}} + \|\mathbf{A}\|_0(\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}}). \tag{3.49}$$

It is easy to see that such codes are (by construction) uniquely decodable, so we have that $\sum_{\mathbf{X}\in\mathcal{X}''} 2^{-\mathrm{pen}(\mathbf{X})} \leq 1$. Now, the set $\mathcal{X}'$ specified in the theorem is a subset of $\mathcal{X}''$ (or perhaps $\mathcal{X}''$ itself, if all elements satisfy the max norm bound condition $\|\mathbf{X}\|_{\max} \leq \mathrm{X}_{\max}$), so (3.48) holds for $\mathcal{X}'$ as specified in the theorem.

Now let $\mathcal{X}$ be any subset of $\mathcal{X}'$. By the above argument the summability condition holds for $\mathcal{X}$, so we may apply the results of Lemma 3.6.1. For randomly subsampled and noisy observations $\mathbf{Y}_{\mathcal{S}}$ our estimates take the form

$$\begin{aligned} \widehat{\mathbf{X}}^{\xi} &= \arg\min_{\mathbf{X}=\mathbf{DA}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) + \xi \cdot \mathrm{pen}(\mathbf{X})\right\} \\ &= \arg\min_{\mathbf{X}=\mathbf{DA}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) + \xi \cdot (\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}}) \cdot \|\mathbf{A}\|_0\right\}. \end{aligned} \tag{3.50}$$

where the last line follows by disregarding additive constants in the optimization arising from terms that do not depend on $\mathbf{X}$ (or more specifically, on $\mathbf{D}$ or $\mathbf{A}$) in the penalty.

Further, when ξ satisfies (3.45), we have

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^{\xi}}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \frac{8C_{\mathrm{D}}\log m}{m} + \\ 3\cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \frac{\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}})}{n_1 n_2} + \left(\xi + \frac{4C_{\mathrm{D}}\log 2}{3}\right)(\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}})\left(\frac{n_1 r + \|\mathbf{A}\|_0}{m}\right)\right\},$$

Finally, letting

$$\lambda \quad = \quad \xi \cdot (\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}}) \tag{3.51}$$

and using the fact that

$$\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}} \leq (\beta+2)\cdot\log(n_1 \vee n_2)\cdot 2\log 2 \tag{3.52}$$

which follows by our selection of L_{lev} and L_{loc} and the fact that $r < n_2$, it follows (after some straightforward simplification) that for

$$\lambda \geq 2(\beta+2)\left(1+\frac{2C_{\mathrm{D}}}{3}\right)\log(n_1 \vee n_2) \tag{3.53}$$

the estimate

$$\widehat{\mathbf{X}}^{\lambda} = \arg\min_{\mathbf{X}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) + \lambda\cdot\|\mathbf{A}\|_0\right\}$$

satisfies

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^{\lambda}}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \frac{8C_{\mathrm{D}}\log m}{m} + \\ 3\cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \frac{\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}})}{n_1 n_2} + \left(\lambda + \frac{4C_{\mathrm{D}}(\beta+2)\log(n_1 \vee n_2)}{3}\right)\left(\frac{n_1 r + \|\mathbf{A}\|_0}{m}\right)\right\},$$

as claimed.

3.6.2 Proof of Corollary 3.3.1

We first establish a general error bound, which we then specialize to the case stated in the corollary. Note that for $\mathbf{X}^*$ as specified and any $\mathbf{X}\in\mathcal{X}$, using the model (3.15) we have

$$\mathrm{D}(p_{X^*_{i,j}}\|p_{X_{i,j}}) = \frac{(X^*_{i,j} - X_{i,j})^2}{2\sigma^2} \tag{3.54}$$

for any fixed $(i,j)\in S$. It follows that $\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}}) = \|\mathbf{X}^* - \mathbf{X}\|_F^2/2\sigma^2$, and using the fact that the amplitudes of entries of $\mathbf{X}^*$ and all $\mathbf{X}\in\mathcal{X}$ are no larger than $\mathrm{X}_{\max}$, it is clear that we may choose $C_{\mathrm{D}} = 2\mathrm{X}_{\max}^2/\sigma^2$. Further, for any $\mathbf{X} \in \mathcal{X}$ and any fixed $(i,j) \in \mathcal{S}$ it is easy to show that in this case

$$-2\log \mathrm{A}(p_{X_{i,j}}, p_{X^*_{i,j}}) = \frac{(X^*_{i,j} - X_{i,j})^2}{4\sigma^2}, \tag{3.55}$$

so that $-2\log \mathrm{A}(p_{\mathbf{X}}, p_{\mathbf{X}^*}) = \|\mathbf{X}^* - \mathbf{X}\|_F^2/4\sigma^2$. It follows that

$$\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}}, p_{\mathbf{X}^*})\right] = \frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{4\sigma^2}. \tag{3.56}$$

Incorporating this into Theorem 3.2.1, we obtain that for any

$$\lambda \geq \left(1 + \frac{4\mathrm{X}_{\max}^2}{3\sigma^2}\right) \cdot 2(\beta+2)\cdot \log(n_1 \vee n_2), \tag{3.57}$$

the sparsity penalized ML estimate satisfies the per-element mean-square error bound

$$\begin{aligned} \frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} &\leq \frac{64\mathrm{X}_{\max}^2 \log m}{m} \\ &+ \ 6\cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \frac{\|\mathbf{X}^* - \mathbf{X}\|_F^2}{n_1 n_2} + \right. \\ &\left. \left(2\sigma^2\lambda + \frac{16\mathrm{X}_{\max}^2 (\log 2)^2 (\beta+1)\log(n_1\vee n_2)}{3}\right)\left(\frac{n_1 p + \|\mathbf{A}\|_0}{m}\right)\right\}. \end{aligned} \tag{3.58}$$

We now establish the error bound for the case where the coefficient matrix $\mathbf{A}^*$ is exactly sparse and λ is fixed to the value specified in (3.14). Consider a candidate reconstruction of the form $\mathbf{X}_Q^* = \mathbf{D}_Q^*\mathbf{A}_Q^*$, where the elements of $\mathbf{D}_Q^*$ are the closest discretized surrogates of the entries of $\mathbf{D}^*$, and the entries of and $\mathbf{A}_Q^*$ are the closest discretized surrogates of the *nonzero* entries of $\mathbf{A}^*$ (and zero otherwise). Denote $\mathbf{D}_Q^* = \mathbf{D}^* + \triangle_{\mathbf{D}^*}$ and $\mathbf{A}_Q^* = \mathbf{A}^* + \triangle_{\mathbf{A}^*}$. Then it is easy to see that

$$\mathbf{D}_Q^*\mathbf{A}_Q^* - \mathbf{D}^*\mathbf{A}^* = \mathbf{D}^*\triangle_{\mathbf{A}^*} + \triangle_{\mathbf{D}^*}\mathbf{A}^* + \triangle_{\mathbf{D}^*}\triangle_{\mathbf{A}^*}. \tag{3.59}$$

Given the range limits on allowable $\mathbf{D}$ and $\mathbf{A}$ and that each range is quantized to L_{lev} levels, we have that $\|\triangle_{\mathbf{D}^*}\|_{\max} \leq 1/(L_{\mathrm{lev}}-1)$ and $\|\triangle_{\mathbf{A}^*}\|_{\max} \leq \mathrm{A}_{\max}/(L_{\mathrm{lev}}-1)$. Now, we can obtain a bound on the magnitudes of the elements of $\mathbf{D}_Q^*\mathbf{A}_Q^* - \mathbf{D}^*\mathbf{A}^*$ that hold uniformly over all i, j, as follows

$$\begin{aligned}
\|\mathbf{D}_Q^*\mathbf{A}_Q^* - \mathbf{D}^*\mathbf{A}^*\|_{\max} &= \max_{i,j} |(\mathbf{D}^*\triangle_{\mathbf{A}^*} + \triangle_{\mathbf{D}^*}\mathbf{A}^* + \triangle_{\mathbf{D}^*}\triangle_{\mathbf{A}^*})_{i,j}| \\
&\leq \max_{i,j} |(\mathbf{D}^*\triangle_{\mathbf{A}^*})_{i,j}| + |(\triangle_{\mathbf{D}^*}\mathbf{A}^*)_{i,j}| + |(\triangle_{\mathbf{D}^*}\triangle_{\mathbf{A}})_{i,j}| \\
&\leq \frac{r\mathrm{A}_{\max}}{L_{\mathrm{lev}} - 1} + \frac{r\mathrm{A}_{\max}}{L_{\mathrm{lev}} - 1} + \frac{2r\mathrm{A}_{\max}}{(L_{\mathrm{lev}} - 1)^2} \\
&\leq \frac{8r\mathrm{A}_{\max}}{L_{\mathrm{lev}}}, \qquad (3.60)
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second from the bounds on $\|\triangle_{\mathbf{D}^*}\|_{\max}$ and $\|\triangle_{\mathbf{A}^*}\|_{\max}$ and the entry-wise bounds on elements of allowable $\mathbf{D}$ and $\mathbf{A}$, and the last because $L_{\mathrm{lev}} \geq 2$. Now, it is straight-forward to show that our choice of β in (3.13) implies $L_{\mathrm{lev}} \geq 16r\mathrm{A}_{\max}/\mathrm{X}_{\max}$, so each entry of $\mathbf{D}_Q^*\mathbf{A}_Q^* - \mathbf{D}^*\mathbf{A}^*$ is bounded in magnitude by $\mathrm{X}_{\max}/2$. It follows that each element of the candidate $\mathbf{X}_Q^*$ constructed above is bounded in magnitude by $\mathrm{X}_{\max}$, so $\mathbf{X}_Q^*$ is indeed a valid element of the set $\mathcal{X}$.

Further, the approximation error analysis above also implies directly that

$$\begin{aligned}
\frac{\|\mathbf{X}^* - \mathbf{X}_Q^*\|_F^2}{n_1 n_2} &= \frac{1}{n_1 n_2} \sum_{i \in [n_1], j \in [n_2]} (\mathbf{D}_Q^*\mathbf{A}_Q^* - \mathbf{D}^*\mathbf{A}^*)_{i,j}^2 \\
&\leq \frac{64p^2\mathrm{A}_{\max}^2}{L_{\mathrm{lev}}^2} \\
&\leq \frac{\mathrm{X}_{\max}^2}{m}, \qquad (3.61)
\end{aligned}$$

where the last line follows from the fact that our specific choice of β in (3.13) also implies $L_{\mathrm{lev}} \geq 8r\sqrt{m}\mathrm{A}_{\max}/\mathrm{X}_{\max}$. Now, evaluating the oracle term at the candidate $\mathbf{X}_Q^* = \mathbf{D}_Q^*\mathbf{A}_Q^*$, and using the fact that $\|\mathbf{A}_Q^*\|_0 = \|\mathbf{A}^*\|_0$, we have

$$\begin{aligned}
&\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} \leq \\
&\qquad \frac{70\mathrm{X}_{\max}^2 \log m}{m} + 8(3\sigma^2 + 8\mathrm{X}_{\max}^2)(\beta + 2)\log(n_1 \vee n_2)\left(\frac{n_1 r + \|\mathbf{A}^*\|_0}{m}\right).
\end{aligned}$$

Finally, we establish the error bound for the case where columns of $\mathbf{A}^*$ are in a weak ℓ_p ball of radius $\mathrm{A}_{\max}$, for $p \leq 1$. To that end, let us denote the columns of $\mathbf{A}^*$ by $\mathbf{a}_j^*$ for $j \in [n_2]$, and for any $k \in [r]$, we let $\mathbf{a}_j^{*,(k)}$ denote the best k-term approximation of $\mathbf{a}_j^*$, formed by retaining the largest (in magnitude) elements and setting the rest to zero.

For shorthand, we denote by $\mathbf{A}^{*,(k)}$ the matrix with columns $\mathbf{a}_j^{*,(k)}$ for $j \in [n_2]$. Now, the approximation error incurred may be bounded as

$$
\begin{aligned}
\|\mathbf{X}^* - \mathbf{X}^{*,(k)}\|_F^2 &= \sum_{i,j} (\mathbf{D}^*(\mathbf{A}^* - \mathbf{A}^{*,(k)}))_{i,j}^2 \\
&\leq \sum_{i,j} \|\mathbf{a}_j^* - \mathbf{a}_j^{*,(k)}\|_2^2, \qquad (3.62)
\end{aligned}
$$

where the inequality follows from the fact that each $(\mathbf{D}^*(\mathbf{A}^* - \mathbf{A}^{*,(k)}))_{i,j}$ may be expressed as an inner product between the i-th row of $\mathbf{D}^*$ (whose elements are no larger than 1 in magnitude) and the j-th column of $\mathbf{A}^* - \mathbf{A}^{*,(k)}$. To simplify further, we use the fact that $p \leq 1$ (and $q \geq 2p$), and the approximation behavior of vectors in weak ℓ_p balls (discussed in the preliminaries) to obtain that $\|\mathbf{a}_j^* - \mathbf{a}_j^{*,(k)}\|_2^2 \leq \mathrm{A}_{\max}^2 k^{-2(1/p-1/2)}$. Letting $\alpha = 1/p - 1/2$, we have that the approximation error associated with approximating $\mathbf{A}^*$ by its best k-term approximation satisfies $\|\mathbf{X}^* - \mathbf{X}^{*,(k)}\|_F^2 \leq n_1 n_2 \mathrm{A}_{\max}^2 k^{-2\alpha}$.

Now, we consider a candidate reconstruction of the form $\mathbf{X}_Q^{*,(k)} = \mathbf{D}_Q^* \mathbf{A}_Q^{*,(k)}$ where $\mathbf{D}_Q^*$ is as above and where the nonzero elements of $\mathbf{A}_Q^{*,(k)}$ are taken to be the closest quantized surrogates of the corresponding nonzero elements of $\mathbf{A}^{*,(k)}$. Using the fact that

$$
\begin{aligned}
\frac{\|\mathbf{X}^* - \mathbf{X}_Q^{*,(k)}\|_F^2}{n_1 n_2} &\leq \frac{4\left(\|\mathbf{X}^* - \mathbf{X}^{*,(k)}\|_F^2 + \|\mathbf{X}^{*,(k)} - \mathbf{X}_Q^{*,(k)}\|_F^2\right)}{n_1 n_2} \\
&\leq 4\mathrm{A}_{\max}^2 k^{-2\alpha} + \frac{4\mathrm{X}_{\max}^2}{m}, \qquad (3.63)
\end{aligned}
$$

where the first term on the bottom results from the approximation error analysis above and the second from our analysis of the first result of the corollary, we evaluate the oracle bound at the candidate $\mathbf{X}_Q^{(k)}$ to obtain

$$
\begin{aligned}
&\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} \qquad (3.64) \\
&\leq \frac{88\mathrm{X}_{\max}^2 \log m}{m} + \\
&\quad \min_{k \geq 1} \left\{ 24\mathrm{A}_{\max}^2 k^{-2\alpha} + 8(3\sigma^2 + 8\mathrm{X}_{\max}^2)(\beta+2)\log(n_1 \vee n_2)\left(\frac{n_1 r + k n_2}{m}\right) \right\}.
\end{aligned}
$$

Finally, we choose $k = (m/n_2)^{1/(1+2\alpha)}$ to balance the decay rates on the $k^{-2\alpha}$ and

kn_2/m terms, and thus obtain

$$\begin{aligned}
&\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} \leq \\
&\quad \frac{88\mathrm{X}_{\max}^2 \log m}{m} + 8(3\sigma^2 + 8\mathrm{X}_{\max}^2)(\beta+2)\log(n_1 \vee n_2)\frac{n_1 r}{m} \\
&\quad + \left[24\mathrm{A}_{\max}^2 + 8(3\sigma^2 + 8\mathrm{X}_{\max}^2)(\beta+2)\log(n_1 \vee n_2)\right]\left(\frac{n_2}{m}\right)^{\frac{2\alpha}{2\alpha+1}}. \qquad (3.65)
\end{aligned}$$

The stated bounds in each case follow from some straight-forward bounding, as well as the fact mentioned in Section 3.3, that under our assumptions, $(\beta+2)\log(n_1 \vee n_2) = \mathcal{O}(\log(n_1 \vee n_2))$.

3.6.3 Proof of Corollary 3.3.2

We follow a similar approach as in the proof of Corollary 3.3.1, and first establish the general error bound. For $\mathbf{X}^*$ as specified and any fixed $\mathbf{X} \in \mathcal{X}$. We have by (relatively) straight-forward calculation that for any fixed $(i,j) \in S$,

$$\begin{aligned}
\mathrm{D}(p_{X^*_{i,j}} \| p_{X_{i,j}}) &= \tau\ |X^*_{i,j} - X_{i,j}| - (1 - e^{-\tau\ |X^*_{i,j} - X_{i,j}|}) \\
&\leq \tau\ |X^*_{i,j} - X_{i,j}| \qquad (3.66)
\end{aligned}$$

where the inequality follows from the fact that $(1 - e^{-\tau\ |X^*_{i,j} - X_{i,j}|}) \geq 0$, and

$$\begin{aligned}
-2\log \mathrm{A}(p_{X_{i,j}}, p_{X^*_{i,j}}) &= \tau\ |X^*_{i,j} - X_{i,j}| - 2\log\left(1 + \tau\ \frac{|X^*_{i,j} - X_{i,j}|}{2}\right) \\
&\geq \frac{\tau^2}{4(\tau\mathrm{X}_{\max}+1)^2}(X^*_{i,j} - X_{i,j})^2, \qquad (3.67)
\end{aligned}$$

where the inequality follows from the convexity of the negative log Hellinger affinity along with an application of Taylor's theorem[6] . It follows from this that $\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}}) \leq$

[6] Formally, letting $x \triangleq X^*_{i,j} - X_{i,j}$ and $f(x) = \tau|x| - 2\log(1 + \tau|x|/2)$ we have

$$f'(x) = \frac{\tau^2}{2}\left(\frac{x}{1+\tau|x|/2}\right) \quad \text{and} \quad f''(x) = \frac{\tau^2}{2(1+\tau|x|/2)^2}.$$

Thus, $f(x)$ is twice differentiable (everywhere). The result follows from the fact that $f(0) = f'(0) = 0$ and

$$f''(x) \geq \frac{\tau^2}{2(1+\tau\mathrm{X}_{\max})^2}$$

for all x of the specified form, given the assumptions on $\mathbf{X}^*$ and $\mathbf{X}$.

$\tau\ \|\mathbf{X}^* - \mathbf{X}\|_1$, and

$$\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}}, p_{\mathbf{X}^*})\right] \geq \frac{\tau^2}{4(\tau \mathrm{X}_{\max}+1)^2}\ \mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]. \tag{3.68}$$

Further, we may choose $C_{\mathrm{D}} = 2\tau\mathrm{X}_{\max}$. Incorporating this into Theorem 3.2.1, we have that for any

$$\lambda \geq 2(\beta+2)\left(1+\frac{4\tau\mathrm{X}_{\max}}{3}\right)\log(n_1 \vee n_2), \tag{3.69}$$

the sparsity-penalized ML estimate satisfies

$$\begin{aligned}\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1n_2} &\leq \frac{1}{\tau}\cdot\frac{64(\tau\mathrm{X}_{\max}+1)^2\mathrm{X}_{\max}\log m}{m} + \frac{12(\tau\mathrm{X}_{\max}+1)^2}{\tau}\cdot \\ &\min_{\mathbf{X}\in\mathcal{X}}\left\{\frac{\|\mathbf{X}^*-\mathbf{X}\|_1}{n_1n_2} + \left(\frac{\lambda}{\tau}+\frac{8\mathrm{X}_{\max}(\beta+2)\log(n_1\vee n_2)}{3}\right)\left(\frac{n_1p+\|\mathbf{A}\|_0}{m}\right)\right\}.\end{aligned} \tag{3.70}$$

We now establish the error bound for the case where the coefficient matrix $\mathbf{A}^*$ is sparse and λ is fixed to the value (3.14). We again consider a candidate reconstruction of the form $\mathbf{X}_Q^* = \mathbf{D}_Q^*\mathbf{A}_Q^*$, where the elements of $\mathbf{D}_Q^*$ are the closest discretized surrogates of the entries of $\mathbf{D}^*$, and the entries of and $\mathbf{A}_Q^*$ are the closest discretized surrogates of the nonzero entries of $\mathbf{A}^*$ (and zero otherwise). Now, since β is the same as in the proof of Corollary 3.3.1, we can directly apply the bound of (3.60) (and use the fact that $L_{\mathrm{lev}} \geq 16r\mathrm{A}_{\max}/\mathrm{X}_{\max}\}$) to conclude that

$$\frac{\|\mathbf{X}^*-\mathbf{X}_Q^*\|_1}{n_1n_2} \leq \frac{\mathrm{X}_{\max}}{2n_1n_2} \leq \frac{\mathrm{X}_{\max}}{m}. \tag{3.71}$$

Now, evaluating the oracle term at the candidate $\mathbf{X}_Q^* = \mathbf{D}_Q^*\mathbf{A}_Q^*$, and using the fact that $\|\mathbf{A}_Q^*\|_0 = \|\mathbf{A}^*\|_0$, we have

$$\begin{aligned}\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1n_2} &\leq \frac{76(\tau\mathrm{X}_{\max}+1)^2}{\tau^2}\cdot\frac{\tau\mathrm{X}_{\max}\log m}{m} \\ &+\frac{12(\tau\mathrm{X}_{\max}+1)^2}{\tau^2}\left(2+\frac{16\tau\mathrm{X}_{\max}}{3}\right)(\beta+2)\log(n_1\vee n_2)\left(\frac{n_1p+\|\mathbf{A}^*\|_0}{m}\right).\end{aligned}$$

Finally, we establish the error bound for the case where the columns of $\mathbf{A}^*$ are vectors in a weak ℓ_p ball for $p \leq 1/2$. By a similar analysis as above, we conclude that $\|\mathbf{X}^* - \mathbf{X}^{*,(k)}\|_1 \leq n_1n_2\mathrm{A}_{\max}k^{-\alpha'}$, where $\alpha = 1/p - 1$. Now, we consider a candidate reconstruction of the form $\mathbf{X}_Q^{*,(k)} = \mathbf{D}_Q^*\mathbf{A}_Q^{*,(k)}$ where $\mathbf{D}_Q^*$ is as above and where the nonzero elements of $\mathbf{A}_Q^{*,(k)}$ are taken to be the closest quantized surrogates of the corresponding nonzero elements of $\mathbf{A}^{*,(k)}$. Using the fact that

$$\begin{aligned}\frac{\|\mathbf{X}^* - \mathbf{X}_Q^{*,(k)}\|_1}{n_1 n_2} &\leq \frac{\|\mathbf{X}^* - \mathbf{X}^{*,(k)}\|_1 + \|\mathbf{X}^{*,(k)} - \mathbf{X}_Q^{*,(k)}\|_1}{n_1 n_2} \\ &\leq \mathrm{A}_{\max} k^{-\alpha'} + \frac{\mathrm{X}_{\max}}{m}, \end{aligned} \tag{3.72}$$

where the first term on the bottom results from the approximation error analysis above and the second from our analysis of the first result of the corollary, we evaluate the oracle bound at the candidate $\mathbf{X}_Q^{(k)}$ to obtain

$$\begin{aligned}\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} &\leq \frac{76(\tau \mathrm{X}_{\max}+1)^2}{\tau^2} \cdot \frac{\tau \mathrm{X}_{\max} \log m}{m} \\ &+ \frac{12(\tau \mathrm{X}_{\max}+1)^2}{\tau^2} \min_{k \geq 1} \left\{ \tau \mathrm{A}_{\max} k^{-\alpha'} + \right. \\ &\left. \left(2 + \frac{16\tau \mathrm{X}_{\max}}{3}\right)(\beta+2)\log(n_1 \vee n_2)\left(\frac{n_1 p + n_2 k}{m}\right)\right\}.\end{aligned}$$

Finally, we choose $k = (m/n_2)^{1/(1+\alpha')}$ to balance the $k^{-\alpha'}$ and $n_2 k/m$ terms, and thus obtain

$$\begin{aligned}\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} &\leq \frac{76(\tau \mathrm{X}_{\max}+1)^2}{\tau^2} \cdot \frac{\tau \mathrm{X}_{\max} \log m}{m} \\ &+ \frac{12(\tau \mathrm{X}_{\max}+1)^2}{\tau^2}\left(2 + \frac{16\tau \mathrm{X}_{\max}}{3}\right)(\beta+2)\log(n_1 \vee n_2)\left(\frac{n_1 p}{m}\right) \\ &+ \frac{12(\tau \mathrm{X}_{\max}+1)^2}{\tau^2}\left(\tau \mathrm{A}_{\max} + \left(2 + \frac{16\tau \mathrm{X}_{\max}}{3}\right)(\beta+2)\log(n_1 \vee n_2)\right)\left(\frac{n_2}{m}\right)^{\frac{\alpha'}{\alpha'+1}}.\end{aligned} \tag{3.73}$$

3.6.4 Proof of Corollary 3.3.3 (Sketch)

We follow a similar approach as for the previous proofs, by first establishing a general error bound. We make use of intermediate results from [14] to bound the KL divergences and negative log Hellinger affinities for the Poisson pmf in terms of quadratic differences. Applying those techniques to our setting, we obtain that

$$\mathrm{D}(p_{X_{i,j}^*} \| p_{X_{i,j}}) \leq \frac{(X_{i,j}^* - X_{i,j})^2}{\mathrm{X}_{\min}} \tag{3.74}$$

and

$$-2\log \mathrm{A}(p_{X^*_{i,j}}, p_{X_{i,j}}) \geq \frac{(X^*_{i,j} - X_{i,j})^2}{4\mathrm{X}_{\max}}. \tag{3.75}$$

It follows that

$$\mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) \leq \|\mathbf{X}^* - \mathbf{X}\|_F^2 / \mathrm{X}_{\min},$$

$$\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\lambda}, p_{\mathbf{X}^*})\right] \geq \mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}^\lambda\|_F^2\right] / 4\mathrm{X}_{\max},$$

and we may choose $C_\mathrm{D} = 4\mathrm{X}^2_{\max}/\mathrm{X}_{\min}$. Incorporating this into Theorem 3.2.1, we obtain that for any

$$\lambda \geq \left(1 + \frac{8\mathrm{X}^2_{\max}}{3\mathrm{X}_{\min}}\right) 2(\beta + 2) \cdot \log(n_1 \vee n_2), \tag{3.76}$$

the sparsity penalized ML estimate satisfies

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} \leq \frac{1}{\mathrm{X}_{\min}} \cdot \frac{128\mathrm{X}^3_{\max}\log m}{m} + \tag{3.77}$$
$$\frac{12\mathrm{X}_{\max}}{\mathrm{X}_{\min}} \cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \frac{\|\mathbf{X}^* - \mathbf{X}\|_F^2}{n_1 n_2} + \left(\lambda + \frac{16\mathrm{X}^2_{\max}(\beta+2)\log(n_1 \vee n_2)}{3}\right)\left(\frac{n_1 r + \|\mathbf{A}\|_0}{m}\right)\right\}.$$

Now, the approximation error term in the oracle bound is in terms of a squared Frobenius norm, so the analysis for the case where λ is fixed to the specified value proceeds in an analogous manner to that in Appendix 3.6.2 for both the sparse and approximately sparse settings. We omit the details.

3.6.5 Proof of Corollary 3.3.4

For $\mathbf{X}^*$ as above and any $\mathbf{X} \in \mathcal{X}$, and using the model (3.28), it is easy to show that

$$\mathrm{D}(p_{X^*_{i,j}} \| p_{X_{i,j}}) = F(X^*_{i,j}) \cdot \log\left(\frac{F(X^*_{i,j})}{F(X_{i,j})}\right) + (1 - F(X^*_{i,j})) \cdot \log\left(\frac{1 - F(X^*_{i,j})}{1 - F(X_{i,j})}\right) \tag{3.78}$$

for any fixed $(i,j) \in S$. Now, we make use of two results that follow directly from lemmata established in [64]. The first lemma provides quadratic bounds on the KL divergence in terms of the Bernoulli parameters; its proof relies on a straightforward application of Taylor's theorem.

Lemma 3.6.2 (from [64])**.** *Let p_π and $p_{\pi'}$ be Bernoulli pmf's with parameters $\pi, \pi' \in (0,1)$. The KL divergences satisfy*

$$\mathrm{D}(p_{\pi'} \| p_\pi), \mathrm{D}(p_\pi \| p_{\pi'}) \leq \frac{1}{2}\left(\sup_{|t|\leq \mathrm{X}_{\max}} \frac{1}{F(t)(1-F(t))}\right)(\pi - \pi')^2. \tag{3.79}$$

The second lemma we utilize establishes a bound on the squared difference between Bernoulli parameters in terms of the squared difference of the underlying matrix elements; its proof is straightforward, and essentially entails establishing the Lipschitz continuity of F.

Lemma 3.6.3 (from [64])**.** *Let $\pi = \pi(X)$ and $\pi = \pi'(X')$ be Bernoulli parameters that are related to some underlying real-valued parameters X and X' via $\pi(X) = F(X)$ and $\pi'(X') = F(X')$, where $F(\cdot)$ is the cdf of a continuous random variable with density $f(\cdot)$. If $|X|, |X'| \leq \mathrm{X}_{\max}$, then*

$$\begin{aligned}(\pi(X) - \pi'(X'))^2 &\leq \left(\sup_{|t|\leq \mathrm{X}_{\max}} f(t)\right)^2 (X - X')^2, && (3.80)\\ &= \left(\sup_{|t|\leq \mathrm{X}_{\max}} f^2(t)\right)(X - X')^2. && (3.81)\end{aligned}$$

Together, these results allow us to claim here that for

$$c_{F,\mathrm{X}_{\max}} \triangleq \left(\sup_{|t|\leq \mathrm{X}_{\max}} \frac{1}{F(t)(1-F(t))}\right) \cdot \left(\sup_{|t|\leq \mathrm{X}_{\max}} f^2(t)\right). \tag{3.82}$$

we have

$$\mathrm{D}(p_{X^*_{i,j}} \| p_{X_{i,j}}) \leq \frac{1}{2} \cdot c_{F,\mathrm{X}_{\max}} (X^*_{i,j} - X_{i,j})^2. \tag{3.83}$$

It follows that we may take $C_{\mathrm{D}} = 2c_{F,\mathrm{X}_{\max}}\mathrm{X}^2_{\max}$, and we have

$$\mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) \leq (c_{F,\mathrm{X}_{\max}}/2)\ \|\mathbf{X}^* - \mathbf{X}\|_F^2.$$

We next obtain a (quadratic) lower bound on the negative log Hellinger affinity. To that end, we introduce the squared Hellinger distance between $p_{X^*_{i,j}}$ and $p_{X_{i,j}}$, denoted here by $\mathrm{H}^2(p_{X^*_{i,j}}, p_{X_{i,j}})$ and given by

$$\mathrm{H}^2(p_{X^*_{i,j}}, p_{X_{i,j}}) = \sum_{y\in\{0,1\}} \left(\sqrt{p_{X^*_{i,j}}(y)} - \sqrt{p_{X_{i,j}}(y)}\right)^2. \tag{3.84}$$

It is straightforward to see that $\mathrm{H}^2(p_{X^*_{i,j}}, p_{X_{i,j}}) = 2(1 - \mathrm{A}(p_{X^*_{i,j}}, p_{X_{i,j}}))$. Now, recall that the Hellinger affinity is always between 0 and 1, so using the fact that $\log(x) \leq x - 1$ for $x > 0$, we see directly that

$$\mathrm{H}^2(p_{X^*_{i,j}}, p_{X_{i,j}}) \leq -2 \log \mathrm{A}(p_{X^*_{i,j}}, p_{X_{i,j}}). \tag{3.85}$$

Now, a direct application of the result of [60, Lemma 2] derived for a similar subproblem to our problem here yields that for

$$c'_{F,\mathrm{X}_{\max}} \triangleq \inf_{|t|\leq \mathrm{X}_{\max}} \frac{f^2(t)}{F(t)(1-F(t))}, \tag{3.86}$$

we have that

$$\mathrm{H}^2(p_{X^*_{i,j}}, p_{X_{i,j}}) \geq \frac{1}{8} c'_{F,\mathrm{X}_{\max}} (X^*_{i,j} - X_{i,j})^2. \tag{3.87}$$

It follows that for any fixed $\mathbf{X} \in \mathcal{X}$, we have $-2\log \mathrm{A}(p_{\mathbf{X}^*}, p_{\mathbf{X}}) \geq (c'_{F,\mathrm{X}_{\max}}/8)\ \|\mathbf{X}^* - \mathbf{X}\|_F^2$.

Incorporating all of the above into Theorem 3.2.1 with

$$\lambda \geq 2(\beta+2)\left(1 + \frac{4c_{F,\mathrm{X}_{\max}}\ \mathrm{X}^2_{\max}}{3}\right)\log(n_1 \vee n_2), \tag{3.88}$$

the sparsity penalized ML estimate satisfies the per-element mean-square error bound

$$\begin{aligned} \frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} \leq \left(\frac{c_{F,\mathrm{X}_{\max}}}{c'_{F,\mathrm{X}_{\max}}}\right) \cdot \frac{128\mathrm{X}^2_{\max}\log m}{m} + \\ 24\left(\frac{c_{F,\mathrm{X}_{\max}}}{c'_{F,\mathrm{X}_{\max}}}\right) \cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \frac{\|\mathbf{X}^* - \mathbf{X}\|_F^2}{n_1 n_2} + \right. \\ \left. \left(\frac{\lambda}{c_{F,\mathrm{X}_{\max}}} + \frac{8\mathrm{X}^2_{\max}(\beta+2)\log(n_1 \vee n_2)}{3}\right)\frac{n_1 r + \|\mathbf{A}\|_0}{m}\right\}. \end{aligned} \tag{3.89}$$

Now, the approximation error term in the oracle bound is again in terms of a squared Frobenius norm, so the analysis for the case where λ is fixed to the specified value proceeds in an analogous manner to that in Appendix 3.6.2 for both the sparse and nearly sparse settings. We again omit the details.

3.6.6 Proof of Lemma 3.6.1

Our estimation approach here is, at its essence, a constrained maximum likelihood method and our proof approach follows the general framework proposed in [24] (see also [23, 36, 106]) and utilized in [13, 14, 63]. Compared with these existing efforts, the main challenge in our analysis here arises because of the "missing data" paradigm, since we aim to establish consistency results that hold *globally* (at all locations of the unknown matrix) using observations obtained at only a subset of the locations. Our approach will be to identify conditions under which, for the purposes of our analysis, a set of sample locations is deemed "good," in a manner to be made explicit below. The primary characteristic of good sets S of sample locations that we will leverage in our analysis is that they be such that KL divergences and (negative logarithms of) Hellinger affinities evaluated only at the locations in S be representative surrogates for the corresponding quantities were we to evaluate them at all $(i,j) \in [n_1] \times [n_2]$ (i.e., even at the unmeasured locations). Clearly, such conditions will inherently rely on certain properties of the matrices that we seek to estimate, somewhat analogously to how notions of incoherence facilitate matrix completion analyses under low rank matrix models. Here, we will see these conditions manifest not as properties of the singular vectors of the unknown matrix to be estimated as in existing matrix completion works, but instead, as conditions on the magnitude of the largest matrix entry.

Our approach will be as follows. First, we describe formally the notion of "good" sets of sample locations, and we show that sets of sample locations generated randomly according to an independent Bernoulli model are "good" with high probability. Then, we establish error guarantees that hold conditionally on the event that the set of sample locations is "good." Finally, we obtain our overall result using some simple conditioning arguments.

3.6.6.1 "Good" Sample Set Characteristics

We begin by characterizing, formally, the properties of certain sets of sample locations that will be useful for our analysis here. As above $\mathbf{X}^*$ denotes the true (unknown) matrix that we aim to estimate, and $\mathcal{X}$ is a countable set of candidate estimates $\mathbf{X}$, each with corresponding penalty $\text{pen}(\mathbf{X}) \geq 1$ chosen so the inequality (3.48) is satisfied. Also, recall that $\mathrm{X}_{\max} > 0$ is a finite constant for which $\max_{i,j} |X^*_{i,j}| \leq \mathrm{X}_{\max}/2$ and $\max_{\mathbf{X}\in\mathcal{X}} \max_{i,j} |X_{i,j}| \leq \mathrm{X}_{\max}$. Finally, we let C_{A} and C_{D} be any upper bounds, respectively, on (twice) the negative log Hellinger affinities between $p_{X^*_{i,j}}$ and $p_{X_{i,j}}$, and the KL divergences of $p_{X_{i,j}}$ from $p_{X^*_{i,j}}$ that hold over all indices, and for all elements $\mathbf{X} \in \mathcal{X}$, so that

$$C_{\mathrm{A}} \geq \max_{\mathbf{X}\in\mathcal{X}} \max_{i,j} \; -2\log \mathrm{A}(p_{X^*_{i,j}}, p_{X_{i,j}}) \tag{3.90}$$

and

$$C_{\mathrm{D}} \geq \max_{\mathbf{X}\in\mathcal{X}} \max_{i,j} \; D(p_{X^*_{i,j}} \| p_{X_{i,j}}). \tag{3.91}$$

Note that the statement of Theorem 3.2.1 only prescribed a condition on $C_{\rm D}$; our introduction of an additional constant $C_{\rm A}$ here is only to simplify the subsequent analysis. In the concluding steps of the proof we will claim that upon selecting a suitable $C_{\rm D}$, one may always obtain a valid choice of $C_{\rm A}$ by taking $C_{\rm A} = C_{\rm D}$. This will enable us to eliminate the $C_{\rm A}$ terms that arise in our bound by bounding them in terms of the constant $C_{\rm D}$.

Let $m \in [n_1 n_2]$ denote a nominal number of measurements, and let $\gamma = m/n_1 n_2 \in (0, 1]$ denote the corresponding nominal fraction of observed matrix elements. For this γ and any fixed $\delta \in (0, 1)$, we define the "good" set $\mathcal{G}_{\gamma,\delta} = \mathcal{G}_{\gamma,\delta}(\mathbf{X}^*, \mathcal{X})$ of possible sample location sets as

$$\begin{aligned}\mathcal{G}_{\gamma,\delta} \triangleq \Bigg\{ & S \subseteq [n_1] \times [n_2] \ : \\ & \bigcap_{\mathbf{X}\in\mathcal{X}} D(p_{\mathbf{X}^*_S} \| p_{\mathbf{X}_S}) \leq \frac{3\gamma}{2} \mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) + 2\left(\frac{2C_{\rm D}}{3}\right) [\log(1/\delta) + \mathrm{pen}(\mathbf{X}) \log 2] \ \cap \\ & \bigcap_{\mathbf{X}\in\mathcal{X}} (-2 \log \mathrm{A}(p_{\mathbf{X}^*_S}, p_{\mathbf{X}_S})) \geq \\ & \qquad \frac{\gamma}{2} (-2 \log \mathrm{A}(p_{\mathbf{X}^*}, p_{\mathbf{X}})) - 2\left(\frac{2C_{\rm A}}{3}\right) [\log(1/\delta) + \mathrm{pen}(\mathbf{X}) \log 2] \Bigg\}.\end{aligned}$$

Directly certifying whether any fixed set S is an element of $\mathcal{G}_{\gamma,\delta}$ may be difficult in general. However, our observation model here assumes that the sample location set is generated randomly, according to an independent Bernoulli(γ) model, where each location is included in the set independently with probability $\gamma \in (0, 1]$. In this case, we have that random sample location sets $\mathcal{S}$ so generated satisfy $\mathcal{S} \in \mathcal{G}_{\gamma,\delta}$ with high probability, as shown in the following lemma.

Lemma 3.6.4. *Let $\mathcal{X}$ be any countable collection of candidate estimates $\mathbf{X}$ for $\mathbf{X}^*$, with corresponding penalties* $\mathrm{pen}(\mathbf{X})$ *satisfying* (3.48)*. For any fixed $\gamma \in (0, 1)$, let $\mathcal{S} \subseteq [n_1] \times [n_2]$ be a random sample set generated according to the independent Bernoulli(γ) model. Then, for any $\delta \in (0, 1)$ we have* $\Pr(\mathcal{S} \notin \mathcal{G}_{\gamma,\delta}) \leq 2\delta$*.*

Proof. Write $\{\mathcal{S} \in \mathcal{G}_{\kappa,\delta}\} = \mathcal{E}_u \cap \mathcal{E}_l$, where

$$\mathcal{E}_u \triangleq \left\{ \bigcap_{\mathbf{X}\in\mathcal{X}} D(p_{\mathbf{X}^*_S} \| p_{\mathbf{X}_S}) \leq \frac{3\gamma}{2} \mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) + 2\left(\frac{2C_{\rm D}}{3}\right) [\log(1/\delta) + \mathrm{pen}(\mathbf{X}) \log 2] \right\},$$

and

$$\mathcal{E}_l \triangleq \left\{ \bigcap_{\mathbf{X}\in\mathcal{X}} (-2\log \mathrm{A}(p_{\mathbf{X}^*_S}, p_{\mathbf{X}_S})) \geq \right.$$
$$\left. \frac{\gamma}{2}\,(-2\log \mathrm{A}(p_{\mathbf{X}^*}, p_{\mathbf{X}})) - 2\left(\frac{2C_{\mathrm{A}}}{3}\right)[\log(1/\delta) + \mathrm{pen}(\mathbf{X})\log 2] \right\},$$

Then, by straight-forward union bounding, $\Pr(\mathcal{S} \notin \mathcal{G}_{\gamma,\delta}(\mathcal{X})) \leq \Pr(\mathcal{E}_u^c) + \Pr(\mathcal{E}_l^c)$. The proof of the lemma entails bounding each term on the right-hand side, in turn.

We focus first on bounding the probability of the complement of $\mathcal{E}_u$. To proceed, we will find it convenient to consider an alternative (but equivalent) representation of the sampling operator described explicitly in terms of a collection $\{B_{i,j}\}_{(i,j)} \in [n_1] \times [n_2]$ of independent Bernoulli(γ) random variables, so that $\mathcal{S} = \{(i,j) : B_{i,j} = 1\}$. On account of our assumption that the observations be conditionally independent given $\mathcal{S}$, we have that for any fixed $\mathbf{X} \in \mathcal{X}$,

$$D(p_{\mathbf{X}^*_{\mathcal{S}}} \| p_{\mathbf{X}_{\mathcal{S}}}) = \sum_{(i,j)\in\mathcal{S}} \mathrm{D}(p_{X^*_{i,j}}, p_{X_{i,j}}) = \sum_{i,j} B_{i,j} \cdot \mathrm{D}(p_{X^*_{i,j}}, p_{X_{i,j}}). \tag{3.92}$$

Thus, our analysis reduces to quantifying the concentration behavior of random sums of these forms. For this, we employ a powerful version of Bernstein's Inequality established by Craig [107] that, for our purposes, may be stated as follows: let $\{U_{i,j}\}$ be a collection of independent random variables indexed by (i,j), each satisfying the moment condition that for some $h > 0$,

$$\mathbb{E}\left[|U_{i,j} - \mathbb{E}[U_{i,j}]|^k\right] \leq \frac{\mathrm{var}(U_{i,j})}{2}\, k!\, h^{k-2},$$

for all integers $k \geq 2$. Then, for any $\tau > 0$ and $0 \leq \epsilon h \leq \theta < 1$, the probability that

$$\sum_{i,j}(U_{i,j} - \mathbb{E}\left[U_{i,j}\right]) \geq \frac{\tau}{\epsilon} + \frac{\epsilon\, \sum_{i,j} \mathrm{var}\left(U_{i,j}\right)}{2(1-\theta)} \tag{3.93}$$

is no larger than $e^{-\tau}$. A useful (and easy to verify) fact is that whenever $|U_{i,j} - \mathbb{E}[U_{i,j}]| \leq \beta$, the moment condition is satisfied by the choice $h = \beta/3$.

Now, fix $\mathbf{X} \in \mathcal{X}$, and let $U_{i,j}(\mathbf{X}) = B_{i,j} \cdot \mathrm{D}(p_{X^*_{i,j}}, p_{X_{i,j}})$ and $\mathbb{E}\left[U_{i,j}(\mathbf{X})\right] = \gamma \cdot \mathrm{D}(p_{X^*_{i,j}}, p_{X_{i,j}})$. Applying Craig's version of Bernstein's inequality with $\theta = 1/4$, $h = C_{\mathrm{D}}/3$, and $\epsilon = \theta/h = 3/(4C_{\mathrm{D}})$, and using the fact that

$$\mathrm{var}(U_{i,j}(\mathbf{X})) = \gamma(1-\gamma)\left(\mathrm{D}(p_{X^*_{i,j}}, p_{X_{i,j}})\right)^2 \leq \gamma\left(\mathrm{D}(p_{X^*_{i,j}}, p_{X_{i,j}})\right)^2 \tag{3.94}$$

we obtain that for any $\tau > 0$,

$$\Pr\left(\sum_{i,j}(B_{i,j}-\gamma)\mathrm{D}(p_{X^*_{i,j}},p_{X_{i,j}}) \geq \frac{4C_{\mathrm{D}}\tau}{3} + \frac{\sum_{i,j}\gamma\left(\mathrm{D}(p_{X^*_{i,j}},p_{X_{i,j}})\right)^2}{2C_{\mathrm{D}}}\right) \leq e^{-\tau}. \quad (3.95)$$

Now, since $\mathrm{D}(p_{X^*_{i,j}},p_{X_{i,j}}) \leq C_{\mathrm{D}}$ by definition, the above result ensures that for any $\tau > 0$,

$$\Pr\left(\sum_{i,j}(B_{i,j}-\gamma)\mathrm{D}(p_{X^*_{i,j}},p_{X_{i,j}}) \geq \frac{4C_{\mathrm{D}}\tau}{3} + \frac{\gamma}{2}\mathrm{D}(p_{\mathbf{X}^*},p_{\mathbf{X}})\right) \leq e^{-\tau}. \quad (3.96)$$

Letting $\delta = e^{-\tau}$ and simplifying a bit, we obtain that for any $\delta \in (0,1)$,

$$\Pr\left(\mathrm{D}(p_{\mathbf{X}^*_{\mathcal{S}}},p_{\mathbf{X}_{\mathcal{S}}}) \geq \frac{4C_{\mathrm{D}}\log(1/\delta)}{3} + \frac{3\gamma}{2}\mathrm{D}(p_{\mathbf{X}^*},p_{\mathbf{X}})\right) \leq \delta. \quad (3.97)$$

Now, if for each $\mathbf{X} \in \mathcal{X}$ we let $\delta_{\mathbf{X}} = \delta \cdot 2^{-\mathrm{pen}(\mathbf{X})}$, we can apply the union bound to obtain that

$$\Pr\left(\bigcup_{\mathbf{X}\in\mathcal{X}}\mathrm{D}(p_{\mathbf{X}^*_{\mathcal{S}}},p_{\mathbf{X}_{\mathcal{S}}}) \geq \frac{3\gamma}{2}\mathrm{D}(p_{\mathbf{X}^*},p_{\mathbf{X}}) + 2\left(\frac{2C_{\mathrm{D}}}{3}\right)\left[\log(1/\delta) + \mathrm{pen}(\mathbf{X})\cdot\log 2\right]\right) \leq \delta. \quad (3.98)$$

Following a similar approach for the affinity terms (with $U_{i,j}(\mathbf{X}) = -B_{i,j}\cdot(-2\log \mathrm{A}(p_{X^*_{i,j}},p_{X_{i,j}}))$ for all i,j), we obtain that for any $\delta \in (0,1)$,

$$\begin{aligned}\Pr\Bigg(\bigcup_{\mathbf{X}\in\mathcal{X}}\left(-2\log \mathrm{A}(p_{\mathbf{X}^*_{\mathcal{S}}},p_{\mathbf{X}_{\mathcal{S}}})\right) &\leq \frac{\gamma}{2}\left(-2\log \mathrm{A}(p_{\mathbf{X}^*},p_{\mathbf{X}})\right) \\ &-2\left(\frac{2C_{\mathrm{A}}}{3}\right)\left[\log(1/\delta) + \mathrm{pen}(\mathbf{X})\cdot\log 2\right]\Bigg) \leq \delta. \quad (3.99)\end{aligned}$$

The overall result now follows by combining equations (3.98) and (3.99) using a union bound. □

Next, we show how the implications of a sample set being "good" can be incorporated into the analysis of [24] to provide (conditional) error guarantees for completion tasks.

3.6.6.2 A Conditional Error Guarantee

Next, we establish the consistency of complexity penalized maximum likelihood estimators, conditionally on the event that the sample set $\mathcal{S}$ is a fixed set S, such that for fixed $\gamma \in (0,1)$ and $\delta \in (0,1)$, $S \in \mathcal{G}_{\gamma,\delta}$ (i.e., S is "good" according to the criteria outlined above). Our analysis then proceeds along the lines of the approach of [24], but with several key differences that arise because of our subsampling model.

As above, $\mathcal{X}$ is a countable set of candidate estimates $\mathbf{X}$ for $\mathbf{X}^*$, with corresponding penalties $\text{pen}(\mathbf{X})$ satisfying (3.48). Now, for any choice of μ satisfying $\mu \geq 1 + 2C_{\text{A}}/3$, we form an estimate $\widehat{\mathbf{X}}^{\mu} = \widehat{\mathbf{X}}^{\mu}(\mathbf{Y}_S)$ according to

$$\begin{aligned} \widehat{\mathbf{X}}^{\mu} &= \arg\min_{\mathbf{X}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_S}(\mathbf{Y}_S) + 2\mu \cdot \text{pen}(\mathbf{X}) \log 2\right\} \\ &= \arg\max_{\mathbf{X}\in\mathcal{X}} \left\{ \sqrt{p_{\mathbf{X}_S}(\mathbf{Y}_S)} \cdot 2^{-\mu\cdot\text{pen}(\mathbf{X})} \right\}. \end{aligned} \tag{3.100}$$

By this choice, we have that for any $\mathbf{X} \in \mathcal{X}$,

$$\sqrt{p_{\widehat{\mathbf{X}}^{\mu}_S}(\mathbf{Y}_S)} \cdot 2^{-\mu\cdot\text{pen}(\widehat{\mathbf{X}}^{\mu})} \geq \sqrt{p_{\mathbf{X}_S}(\mathbf{Y}_S)} \cdot 2^{-\mu\cdot\text{pen}(\mathbf{X})}. \tag{3.101}$$

This implies that for the particular (deterministic, and μ-dependent) candidate

$$\widetilde{\mathbf{X}}^{\mu} = \arg\min_{\mathbf{X}\in\mathcal{X}} \left\{ \text{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) + \frac{2}{\gamma} \cdot \left(\mu + \frac{2C_{\text{D}}}{3}\right) \text{pen}(\mathbf{X}) \log 2 \right\}, \tag{3.102}$$

(whose specification will become clear shortly) we have

$$\frac{\sqrt{p_{\widehat{\mathbf{X}}^{\mu}_S}(\mathbf{Y}_S)} \cdot 2^{-\mu\cdot\text{pen}(\widehat{\mathbf{X}}^{\mu})}}{\sqrt{p_{\widetilde{\mathbf{X}}^{\mu}_S}(\mathbf{Y}_S)} \cdot 2^{-\mu\cdot\text{pen}(\widetilde{\mathbf{X}}^{\mu})}} \geq 1. \tag{3.103}$$

Using this, along with some straight-forward algebraic manipulations, we have

$$\begin{aligned}
-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}_S^\mu}, p_{\mathbf{X}_S^*}) &= 2\log\left(\frac{1}{\mathrm{A}(p_{\widehat{\mathbf{X}}_S^\mu}, p_{\mathbf{X}_S^*})}\right) \\
&\leq \; 2\log\left(\frac{\sqrt{p_{\widehat{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)}\cdot 2^{-\mu\cdot\mathrm{pen}(\widehat{\mathbf{X}}^\mu)}}{\sqrt{p_{\widetilde{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)}\cdot 2^{-\mu\cdot\mathrm{pen}(\widetilde{\mathbf{X}}^\mu)}}\cdot\frac{1}{\mathrm{A}(p_{\widehat{\mathbf{X}}_S^\mu}, p_{\mathbf{X}_S^*})}\right) \\
&= \; 2\log\left(\frac{\sqrt{p_{\widehat{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)}}{\sqrt{p_{\widetilde{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)}}\cdot\frac{\sqrt{p_{\mathbf{X}_S^*}(\mathbf{Y}_S)}}{\sqrt{p_{\mathbf{X}_S^*}(\mathbf{Y}_S)}}\cdot\frac{2^{-\mu\cdot\mathrm{pen}(\widehat{\mathbf{X}}^\mu)}}{2^{-\mu\cdot\mathrm{pen}(\widetilde{\mathbf{X}}^\mu)}}\cdot\frac{1}{\mathrm{A}(p_{\widehat{\mathbf{X}}_S^\mu}, p_{\mathbf{X}_S^*})}\right) \\
&= \; \log\left(\frac{p_{\mathbf{X}_S^*}(\mathbf{Y}_S)}{p_{\widetilde{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)}\right) + 2\mu\cdot\mathrm{pen}(\widetilde{\mathbf{X}}^\mu)\log 2 + \\
&\qquad 2\log\left(\frac{\sqrt{p_{\widehat{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)/p_{\mathbf{X}_S^*}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\widehat{\mathbf{X}}_S^\mu}, p_{\mathbf{X}_S^*})}\cdot 2^{-\mu\cdot\mathrm{pen}(\widehat{\mathbf{X}}^\mu)}\right). \qquad (3.104)
\end{aligned}$$

At this point, we make our first use of the implications of the “good” sample set condition. In particular, since $S \in \mathcal{G}_{\gamma,\delta}$ and $\widehat{\mathbf{X}}^\mu \in \mathcal{X}$, we have that

$$-2\log \mathrm{A}(p_{\mathbf{X}_S^*}, p_{\widehat{\mathbf{X}}_S^\mu}) \geq \frac{\gamma}{2}\left(-2\log \mathrm{A}(p_{\mathbf{X}^*}, p_{\widehat{\mathbf{X}}^\mu})\right) - 2\left(\frac{2C_{\mathrm{A}}}{3}\right)\left[\log(1/\delta) + \mathrm{pen}(\widehat{\mathbf{X}}^\mu)\log 2\right]. \qquad (3.105)$$

Incorporating this into (3.104), we have

$$\begin{aligned}
\frac{\gamma}{2}\left(-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*})\right) &\leq \log\left(\frac{p_{\mathbf{X}_S^*}(\mathbf{Y}_S)}{p_{\widetilde{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)}\right) + 2\mu\cdot\mathrm{pen}(\widetilde{\mathbf{X}}^\mu)\log 2 + 2\left(\frac{2C_{\mathrm{A}}}{3}\right)\log(1/\delta) \\
&\quad +2\log\left(\frac{\sqrt{p_{\widehat{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)/p_{\mathbf{X}_S^*}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\widehat{\mathbf{X}}_S^\mu}, p_{\mathbf{X}_S^*})}\cdot 2^{-\left(\mu-\frac{2C_{\mathrm{A}}}{3}\right)\mathrm{pen}(\widehat{\mathbf{X}}^\mu)}\right).
\end{aligned}$$

Now, we take expectations (formally, with respect to the conditional distribution of $\mathbf{Y}_{\mathcal{S}}$ given $\{\mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\}$) on both sides to obtain that

$$\begin{aligned}
&\frac{\gamma}{2}\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*}) \,\middle|\, \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right] \leq \\
&\quad \mathrm{D}(p_{\mathbf{X}_S^*}\|p_{\widetilde{\mathbf{X}}_S^\mu}) + 2\mu\cdot\mathrm{pen}(\widetilde{\mathbf{X}}^\mu)\log 2 + 2\left(\frac{2C_{\mathrm{A}}}{3}\right)\log(1/\delta) \\
&\quad +2\mathbb{E}\left[\log\left(\frac{\sqrt{p_{\widehat{\mathbf{X}}_S^\mu}(\mathbf{Y}_S)/p_{\mathbf{X}_S^*}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\widehat{\mathbf{X}}_S^\mu}, p_{\mathbf{X}_S^*})}\cdot 2^{-\left(\mu-\frac{2C_{\mathrm{A}}}{3}\right)\mathrm{pen}(\widehat{\mathbf{X}}^\mu)}\right) \,\middle|\, \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right].
\end{aligned}$$

Using again the implications of $S \in \mathcal{G}_{\gamma,\delta}$, that

$$D(p_{\mathbf{X}^*_S} \| p_{\widetilde{\mathbf{X}}^\mu_S}) \le \frac{3\gamma}{2}\mathrm{D}(p_{\mathbf{X}^*} \| p_{\widetilde{\mathbf{X}}^\mu}) + 2\left(\frac{2C_\mathrm{D}}{3}\right)\left[\log(1/\delta) + \mathrm{pen}(\widetilde{\mathbf{X}}^\mu)\log 2\right] \tag{3.106}$$

since $\widetilde{\mathbf{X}}^\mu \in \mathcal{X}$, we have that

$$\begin{aligned} &\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*}) \;\middle|\; \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right] \le \\ &\qquad 3\mathrm{D}(p_{\mathbf{X}^*} \| p_{\widetilde{\mathbf{X}}^\mu}) + \frac{4}{\gamma}\left(\mu + \frac{2C_\mathrm{D}}{3}\right)\mathrm{pen}(\widetilde{\mathbf{X}}^\mu)\log 2 + \frac{4}{\gamma}\left(\frac{2(C_\mathrm{A} + C_\mathrm{D})}{3}\right)\log(1/\delta) \\ &\qquad + \frac{4}{\gamma}\mathbb{E}\left[\log\left(\frac{\sqrt{p_{\widehat{\mathbf{X}}^\mu_S}(\mathbf{Y}_S)}/\sqrt{p_{\mathbf{X}^*_S}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\widehat{\mathbf{X}}^\mu_S}, p_{\mathbf{X}^*_S})} \cdot 2^{-\left(\mu - \frac{2C_\mathrm{A}}{3}\right)\mathrm{pen}(\widehat{\mathbf{X}}^\mu)}\right) \;\middle|\; \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right]. \end{aligned} \tag{3.107}$$

Turning our attention to the last term on the right-hand side, we have that

$$\begin{aligned} &\mathbb{E}\left[\log\left(\frac{\sqrt{p_{\widehat{\mathbf{X}}^\mu_S}(\mathbf{Y}_S)}/\sqrt{p_{\mathbf{X}^*_S}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\widehat{\mathbf{X}}^\mu_S}, p_{\mathbf{X}^*_S})} \cdot 2^{-\left(\mu - \frac{2C_\mathrm{A}}{3}\right)\mathrm{pen}(\widehat{\mathbf{X}}^\mu)}\right) \;\middle|\; \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right] \\ &\quad \overset{(a)}{\le} \log\left(\mathbb{E}\left[\frac{\sqrt{p_{\widehat{\mathbf{X}}^\mu_S}(\mathbf{Y}_S)}/\sqrt{p_{\mathbf{X}^*_S}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\widehat{\mathbf{X}}^\mu_S}, p_{\mathbf{X}^*_S})} \cdot 2^{-\left(\mu - \frac{2C_\mathrm{A}}{3}\right)\mathrm{pen}(\widehat{\mathbf{X}}^\mu)} \;\middle|\; \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right]\right) \\ &\quad \overset{(b)}{\le} \log\left(\mathbb{E}\left[\sum_{\mathbf{X} \in \mathcal{X}} \frac{\sqrt{p_{\mathbf{X}_S}(\mathbf{Y}_S)}/\sqrt{p_{\mathbf{X}^*_S}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\mathbf{X}_S}, p_{\mathbf{X}^*_S})} \cdot 2^{-\left(\mu - \frac{2C_\mathrm{A}}{3}\right)\mathrm{pen}(\mathbf{X})} \;\middle|\; \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right]\right) \\ &\quad = \log\left(\sum_{\mathbf{X} \in \mathcal{X}} 2^{-\left(\mu - \frac{2C_\mathrm{A}}{3}\right)\mathrm{pen}(\mathbf{X})}\, \mathbb{E}\left[\frac{\sqrt{p_{\mathbf{X}_S}(\mathbf{Y}_S)}/\sqrt{p_{\mathbf{X}^*_S}(\mathbf{Y}_S)}}{\mathrm{A}(p_{\mathbf{X}_S}, p_{\mathbf{X}^*_S})} \;\middle|\; \mathcal{S} = S, S \in \mathcal{G}_{\gamma,\delta}\right]\right) \\ &\quad \overset{(c)}{=} \log\left(\sum_{\mathbf{X} \in \mathcal{X}} 2^{-\left(\mu - \frac{2C_\mathrm{A}}{3}\right)\mathrm{pen}(\mathbf{X})}\right). \end{aligned} \tag{3.108}$$

In the above, (a) follows from Jensen's Inequality, (b) from the facts that $\widehat{\mathbf{X}}^\mu \in \mathcal{X}$ and each term in the sum is non-negative, and (c) from the definition of the Hellinger affinity. Now, because $\mathrm{pen}(\mathbf{X}) \ge 1$ and $\mu \ge 1 + 2C_\mathrm{A}/3$ we have that

$$\sum_{\mathbf{X} \in \mathcal{X}} 2^{-\left(\mu - \frac{2C_\mathrm{A}}{3}\right)\mathrm{pen}(\mathbf{X})} \le \sum_{\mathbf{X} \in \mathcal{X}} 2^{-\mathrm{pen}(\mathbf{X})} \le 1. \tag{3.109}$$

Thus, since the expectation term on the right-hand side of (3.107) is not positive, we can disregard it in the upper bound to obtain that

$$\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^{\mu}}, p_{\mathbf{X}^*}) \;\middle|\; \mathcal{S}=S, S\in\mathcal{G}_{p,\delta}\right] \leq$$
$$3\cdot \mathrm{D}(p_{\mathbf{X}^*}\|p_{\widetilde{\mathbf{X}}^{\mu}}) + \frac{6}{\gamma}\left(\lambda+\frac{2C_{\mathrm{D}}}{3}\right)\mathrm{pen}(\widetilde{\mathbf{X}}^{\mu})\log 2 + \frac{4}{\gamma}\left(\frac{2(C_{\mathrm{A}}+C_{\mathrm{D}})}{3}\right)\log(1/\delta),$$

where we have also inflated (slightly) the leading constant on the second term on the right-hand side to simplify subsequent analysis. Now, recalling the definition of $\widetilde{\mathbf{X}}^{\mu}$, we can state the result equivalently as an oracle bound, as

$$\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^{\mu}}, p_{\mathbf{X}^*}) \;\middle|\; \mathcal{S}=S, S\in\mathcal{G}_{\gamma,\delta}\right] \leq$$
$$3\cdot \min_{\mathbf{X}\in\mathcal{X}}\left\{\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}}) + \frac{2}{\gamma}\left(\mu+\frac{2C_{\mathrm{D}}}{3}\right)\mathrm{pen}(\mathbf{X})\log 2\right\} + \frac{4}{\gamma}\left(\frac{2(C_{\mathrm{A}}+C_{\mathrm{D}})}{3}\right)\log(1/\delta). \tag{3.110}$$

3.6.6.3 Putting the Pieces Together

The last steps of the analysis entail straightforward applications of conditioning arguments, along with the use of a well-known (and easy to verify) information inequality. First, note that

$$\begin{aligned}
&\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^{\mu}}, p_{\mathbf{X}^*}) \;\middle|\; \mathcal{S}\in\mathcal{G}_{\gamma,\delta}\right]\\
&\quad= \sum_{S\in[n_1]\times[n_2]} \mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^{\mu}}, p_{\mathbf{X}^*}) \;\middle|\; \mathcal{S}=S, S\in\mathcal{G}_{\gamma,\delta}\right]\cdot \Pr(\mathcal{S}=S|\mathcal{S}\in\mathcal{G}_{\gamma,\delta})\\
&\quad\leq 3\cdot \min_{\mathbf{X}\in\mathcal{X}}\left\{\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}}) + \frac{2}{\gamma}\left(\mu+\frac{2C_{\mathrm{D}}}{3}\right)\mathrm{pen}(\mathbf{X})\log 2\right\}\\
&\qquad\qquad + \frac{4}{\gamma}\left(\frac{2(C_{\mathrm{A}}+C_{\mathrm{D}})}{3}\right)\log(1/\delta),
\end{aligned}$$

where the last step follows from using the bound in (3.110) and bringing that term outside of the sum since it does not depend on S, and using the fact that the conditional probability mass function $\Pr(\mathcal{S} = S | \mathcal{S} \in \mathcal{G})$ sums to 1. Now, using the fact that

$$\begin{aligned}&\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*})\right] = \\ &\quad \mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*}) \;\middle|\; \mathcal{S} \in \mathcal{G}_{\gamma,\delta}\right] \cdot \Pr(\mathcal{S} \in \mathcal{G}_{\gamma,\delta}) + \\ &\qquad\qquad \mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*}) \;\middle|\; \mathcal{S} \notin \mathcal{G}_{\gamma,\delta}\right] \cdot \Pr(\mathcal{S} \notin \mathcal{G}_{\gamma,\delta}),\end{aligned}$$

where the expectation on the left-hand side is with respect to the joint distribution of $\mathbf{Y}_{\mathcal{S}}$ and $\mathcal{S}$, we obtain that

$$\begin{aligned}&\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*})\right] \leq \\ &\quad 3 \cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) + \frac{2}{\gamma}\left(\mu + \frac{2C_{\mathrm{D}}}{3}\right) \mathrm{pen}(\mathbf{X}) \log 2 \right\} + \\ &\qquad\qquad \frac{4}{\gamma}\left(\frac{2(C_{\mathrm{A}} + C_{\mathrm{D}})}{3}\right) \log(1/\delta) + 2\delta \cdot n_1 n_2 C_{\mathrm{A}},\end{aligned}$$

where we use the trivial upper bound $\mathbb{E}[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*}) \mid \mathcal{S} \notin \mathcal{G}_{\gamma,\delta}] \leq n_1 n_2 C_{\mathrm{A}}$. Now, since the result holds for any choice of $\delta \in (0,1)$, we can choose δ judiciously to "balance" the last two terms. The particular choice $\delta = m^{-1} = (\gamma n_1 n_2)^{-1}$ yields

$$\begin{aligned}&\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*})\right] \\ &\leq \; 3 \cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) + \frac{2}{\gamma}\left(\mu + \frac{2C_{\mathrm{D}}}{3}\right) \mathrm{pen}(\mathbf{X}) \log 2 \right\} + \frac{8(C_{\mathrm{A}} + C_{\mathrm{D}})\log m}{3\gamma} + \frac{2C_{\mathrm{A}}}{\gamma},\end{aligned}$$

which implies the simpler (but slightly looser) bound

$$\begin{aligned}&\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*})\right] \leq \\ &\quad 3 \cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) + \frac{2}{\gamma}\left(\mu + \frac{2C_{\mathrm{D}}}{3}\right) \mathrm{pen}(\mathbf{X}) \log 2 \right\} + \frac{4\,(C_{\mathrm{A}} + C_{\mathrm{D}})\log m}{\gamma}.\end{aligned}$$

Finally, we make use of the fact that for each i, j, we have $-2\log \mathrm{A}(p_{X^*_{i,j}}, p_{X_{i,j}}) \leq \mathrm{D}(p_{X^*_{i,j}}, p_{X_{i,j}})$, which is readily verified with one application of Jensen's inequality. It follows that upon identifying a suitable C_{D}, we may always take $C_{\mathrm{A}} = C_{\mathrm{D}}$. Thus, it is sufficient to choose $\mu > 1 + 2C_{\mathrm{D}}/3$ when forming our complexity regularized maximum likelihood estimator. We conclude that the error of any estimator formed using an appropriate regularization parameter μ satisfies

$$\begin{aligned}&\frac{\mathbb{E}\left[-2\log \mathrm{A}(p_{\widehat{\mathbf{X}}^\mu}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \\ &\quad 3 \cdot \min_{\mathbf{X}\in\mathcal{X}} \left\{ \frac{\mathrm{D}(p_{\mathbf{X}^*} \| p_{\mathbf{X}})}{n_1 n_2} + \left(\mu + \frac{2C_{\mathrm{D}}}{3}\right) \frac{\mathrm{pen}(\mathbf{X}) 2\log 2}{m} \right\} + \frac{8C_{\mathrm{D}}\log(m)}{m},\end{aligned}$$

where we have divided both sides by $n_1 n_2$ and used the fact that $m = \gamma n_1 n_2$. Finally, making the substitution $\xi = 2\mu \log 2$ yields the stated version of the result.

Advances in Structured Adaptive Sensing

In Chapter-4 we summarize our work on adaptive compressive sensing (CS) and study the benefits of tree-sparse structure assumption on the signals. It is a reprint of our journal paper [108] with some minor modifications.

Here we establish the potential benefits that can be achieved when fusing the notions of adaptive sensing and structured sparsity, and propose an adaptive algorithm to recover the exact support of tree-sparse signals from noisy linear measurements, and established that an adaptive sensing strategy specifically tailored to signals that are tree-sparse can significantly outperform adaptive and non-adaptive sensing strategies that are agnostic to the underlying structure.

The proposed algorithm and corresponding analyses, first appeared in [109]. A follow-on effort [108] established that our algorithm is nearly optimal, in the sense that no other sensing and estimation strategy can perform fundamentally better for identifying the support of tree-sparse signals.

Chapter 4

On the Fundamental Limits of Recovering Tree Sparse Vectors from Noisy Linear Measurements

Recent breakthrough results in compressive sensing (CS) have established that many high dimensional signals can be accurately recovered from a relatively small number of non-adaptive linear observations, provided that the signals possess a sparse representation in some basis. Subsequent efforts have shown that the performance of CS can be improved by exploiting additional structure in the locations of the nonzero signal coefficients during inference, or by utilizing some form of data-dependent adaptive measurement focusing during the sensing process. To our knowledge, our own previous work was the first to establish the potential benefits that can be achieved when fusing the notions of adaptive sensing and structured sparsity. In that work, we examined the task of support recovery from noisy linear measurements, and established that an adaptive sensing strategy specifically tailored to signals that are tree-sparse can significantly outperform adaptive and non-adaptive sensing strategies that are agnostic to the underlying structure. In this work we establish fundamental performance limits for the task of support recovery of tree-sparse signals from noisy measurements, in settings where measurements may be obtained either non-adaptively (using a randomized Gaussian measurement strategy motivated by initial CS investigations) or by any adaptive sensing strategy. Our main results here imply that the adaptive tree sensing procedure analyzed in our previous work is nearly optimal, in the sense that no other sensing and estimation strategy can perform fundamentally better for identifying the support of tree-sparse signals.[1]

4.1 Introduction

In recent years, the development and analysis of new sampling and inference methods that make efficient use of measurement resources has received a renewed and concentrated focus. Many of the compelling new investigations in this area share a unifying theme – they leverage the phenomenon of *sparsity* as a means for describing inherently simple (i.e., low-dimensional) structure that is often present in many signals of interest.

Consider the task of inferring a (perhaps very high-dimensional) vector $\mathbf{x} \in \mathbb{R}^n$. Compressive sensing (CS) prescribes collecting non-adaptive linear measurements of $\mathbf{x}$ by "projecting" it onto a collection of n-dimensional "measurement vectors." Formally, CS observations may be modeled as

$$y_j = \langle \mathbf{a}_j, \mathbf{x} \rangle + w_j = \mathbf{a}_j^T \mathbf{x} + w_j, \quad \text{for } j = 1, 2, \ldots, m, \tag{4.1}$$

where $\mathbf{a}_j$ is the j-th measurement vector and w_j describes the additive error associated with the j-th measurement, which may be due to modeling error or stochastic noise. Initial breakthrough results in CS established that sparse vectors $\mathbf{x}$ having no more than $k < n$ nonzero elements can be exactly recovered (in noise-free settings) or reliably estimated (in noisy settings) from a collection of only $m = O(k \log n)$ measurements of the form (4.1) using, for example, ensembles of randomly generated measurement vectors whose entries are iid realizations of certain zero-mean random variables (e.g., Gaussian) – see, for example, [92] as well as numerous CS-related efforts at `dsp.rice.edu/cs`.

While many of the initial efforts in CS focused on purely randomized measurement vector designs and considered recovery of arbitrary sparse vectors, several powerful extensions to the original CS paradigm have been investigated in the literature. One

[1] The material in Chapter-4 is ©2014 IEEE. Reprinted, with permission, from *IEEE Transactions on Information Theory*, "On the Fundamental Limits of Recovering Tree Sparse Vectors from Noisy Linear Measurements," A. Soni and J. Haupt.

such extension allows for additional flexibility in the measurement process, so that information gleaned from previous observations may be employed in the design of future measurement vectors. Formally, such *adaptive sensing* strategies are those for which the j-th measurement vector $\mathbf{a}_j$ is obtained as a (deterministic or randomized) function of previous measurement vectors and observations $\{\mathbf{a}_\ell, y_\ell\}_{\ell=1}^{j-1}$, for each $j = 2, 3, \ldots, m$. Non-adaptive sensing strategies, by contrast, are those for which each measurement vector is independent of all past (and future) observations. The randomized measurement vectors typically employed in CS settings comprise an example of a non-adaptive sensing strategy. Adaptive sensing techniques have been shown beneficial in sparse inference tasks, enabling an improved resilience to measurement noise relative to techniques based on non-adaptive measurements (see, for example, [110–124] as well as the summary article [125] and the references therein) and further reductions in the number of compressive measurements required for recovering sparse vectors in noise-free settings [126, 127].

Another powerful extension to the canonical CS framework corresponds to the exploitation of additional *structure* that may be present in the locations of the nonzeros of $\mathbf{x}$. To formalize this notion, we first define the support $\mathcal{S} = \mathcal{S}(\mathbf{x})$ of a vector $\mathbf{x} = [x_1\ x_2\ \ldots\ x_n]^T$ as

$$\mathcal{S}(\mathbf{x}) \triangleq \{i : x_i \neq 0\}, \tag{4.2}$$

and note that, in general, the support of a k-sparse n-dimensional vector corresponds to one of the $\binom{n}{k}$ distinct subsets of $\{1, 2, \ldots, n\}$ of cardinality k. The term *structured sparsity* describes a restricted class of sparse signals whose supports may occur only on a (known) subset of these $\binom{n}{k}$ distinct subsets. Generally speaking, knowledge of the particular structure present in the object being inferred can be incorporated into sparse inference procedures, and for certain types of structure this can result either in a reduction in the number of measurements required for accurate inference, or improved estimation error guarantees, or both (see, e.g., [128–130], as well as the recent survey article [131] on structured sparsity in compressive sensing).

To the best of our knowledge, our own previous work [109] was the first to identify and quantify the benefits of using adaptive sensing strategies that are tailored to certain types of structured sparsity, in noisy sparse inference tasks. Specifically, the work [109] established that a simple adaptive compressive sensing strategy for *tree-sparse* vectors could successfully identify the support of much weaker signals than what could be recovered using non-adaptive or adaptive sensing strategies that were agnostic to the structure present in the signal being acquired. Subsequent efforts by other authors have similarly identified benefits of adaptive sensing techniques tailored to other forms of structured sparsity in noisy sparse inference tasks [122, 123, 132].

The primary aim of this effort is to establish the optimality of the strategy analyzed in [109], by identifying the fundamental performance limits associated with the task of support recovery of tree-sparse signals from noisy measurements that may be obtained adaptively. For completeness, and in an effort to put these results into a broader context, we also identify here the performance limits associated with the same support recovery task in settings where measurements are obtained non-adaptively using randomized (Gaussian) measurement vector ensembles, as in the initial efforts in CS. We begin by formalizing the notion of tree-structured sparsity, and reviewing the results of [109].

4.1.1 Adaptive Sensing of Tree Sparse Signals

Tree sparsity essentially describes the phenomenon where the nonzero elements of the signal being inferred exhibit clustering along paths in some known underlying tree. For the purposes of our investigation here, we formalize the notion of tree sparsity as follows. Suppose that the set $\{1, 2, \ldots, n\}$ that indexes the elements of $\mathbf{x} \in \mathbb{R}^n$ is put into a one-to-one correspondence with the nodes of a known tree of degree $d \geq 1$ having n nodes, which we refer to as the *underlying tree*. We say that a vector $\mathbf{x}$ is *k-tree sparse* (with respect to the underlying tree) when the indices of the support set $\mathcal{S}(\mathbf{x})$ correspond, collectively, to a rooted connected subtree of the underlying tree. In the sequel we restrict our attention to n-dimensional signals that are tree sparse in a known underlying *binary* tree ($d = 2$), though our approach and main results can be extended, in a relatively straightforward manner, to underlying trees having degree $d > 2$. For illustration, Figure 4.1 depicts a graphical representation of a signal that is 4-tree sparse in an underlying complete tree of degree 2 with 7 nodes.

Tree sparsity arises naturally in the wavelet coefficients of many signals including, in particular, natural images (see, for example, [133–135]), and this fact has motivated several investigations into CS inference techniques that exploit or leverage underlying tree structure in the signals being acquired [128, 129, 136–138]. More aligned with our focus here are several prior efforts that have examined specialized *sensing* techniques,

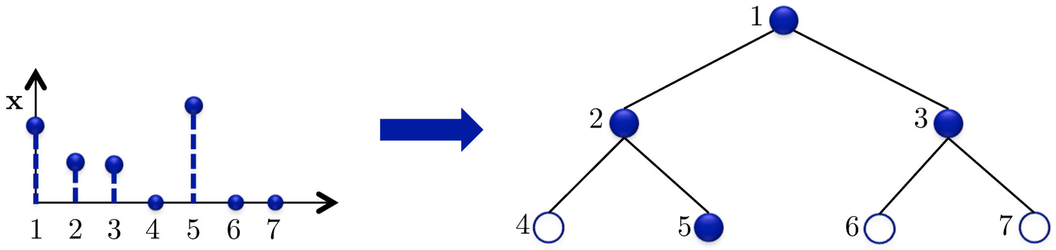

Figure 4.1: A signal $\mathbf{x} \in \mathbb{R}^7$ (left) that is 4-tree sparse in an underlying binary tree having 7 nodes (right). The support $\mathcal{S}(\mathbf{x}) = \{1, 2, 3, 5\}$ corresponds to a rooted connected subtree of the underlying tree.

designed to exploit the inherent tree-based structure present in the wavelet-domain representations of certain signals in various application domains. The work [1], for example, examined dynamic MRI applications where non-Fourier (in this case, wavelet domain) encoding is employed along one of the spatial dimensions, and proposed a sequential sensing strategy that acquires observations of the wavelet coefficients of the object being observed in a "coarse-to-fine" (i.e., top-down, in the wavelet representation) manner. The work [2] compared a coarse-to-fine direct wavelet coefficient sensing approach to a sensing approach based on Bayesian experimental design in the context of an imaging application. More recently, [3] proposed a top-down adaptive wavelet sensing strategy in the context of compressive imaging and provided an analysis of the sample complexity of such strategies in noise-free settings, but did not investigate how such procedures would perform in noisy scenarios; see also [4]. Motivated by these existing efforts, the essential aim of the authors' own prior work [109] was to assess the performance of such strategies in noisy settings; for completeness, we summarize the approach and main results of that work here.

Let us assume, for simplicity, that the signal $\mathbf{x}$ being acquired is tree sparse in the canonical (identity) basis, though extensions to signals that are tree sparse in any other orthonormal basis (e.g., a wavelet basis) are straightforward. Noisy observations of $\mathbf{x}$ are obtained according to (4.1) by projecting $\mathbf{x}$ onto a sequence of adaptively designed measurement vectors, each of which corresponds to a basis vector of the canonical basis, and we assume that each measurement vector has unit norm. Now, to simplify the description of the procedure, we introduce some slightly different notation to index the individual observations. Specifically, rather than indexing observations by the order in

Algorithm 4 Adaptive sensing procedure for acquiring signals assumed tree-sparse in a (known) underlying tree.

```
Input: Threshold $\tau \geq 0$; Support Estimate $\mathcal{S} = \emptyset$,
          Data Structure $\mathcal{Q}$ containing the index of the root of the underlying tree
  while $\mathcal{Q} \neq \emptyset$ do
    Remove an index $\ell$ from $\mathcal{Q}$
    Collect noisy observation $y_{(\ell)} = \mathbf{e}_\ell^T \mathbf{x} + \mathcal{N}(0, \sigma^2)$
    if $|y_{(\ell)}| \geq \tau$ then
      Add indices corresponding to children of $\ell$ in the underlying tree to $\mathcal{Q}$
      Update support estimate: $\widehat{\mathcal{S}} \leftarrow \widehat{\mathcal{S}} \cup \ell$
    end if
  end while
Output: Final Support Estimate $\widehat{\mathcal{S}}$
```

which they were obtained as in (4.1), we instead index each measurement according to the index of the basis vector onto which $\mathbf{x}$ is projected, or equivalently here, according to the location of $\mathbf{x}$ that was observed. To that end, let us denote by $y_{(j)}$ the measurement obtained by projecting $\mathbf{x}$ onto the vector $\mathbf{e}_j$ having a single nonzero in the j-th location for any $j \in \{1, 2, \ldots, n\}$.

Now, begin by specifying a threshold $\tau \geq 0$, and by initializing a support estimate $\widehat{\mathcal{S}} = \emptyset$ and a data structure $\mathcal{Q}$ (which could be a stack, queue, or simply a set) to contain the index corresponding to the root of the underlying tree. While the data structure $\mathcal{Q}$ is nonempty, remove an element ℓ from $\mathcal{Q}$, collect a noisy measurement $y_{(\ell)}$ by projecting $\mathbf{x}$ onto $\mathbf{e}_\ell$, and perform the following hypothesis test. If $|y_{(\ell)}| \geq \tau$, add the indices corresponding to the children of node ℓ in the underlying tree to the data structure $\mathcal{Q}$ and update the support estimate to include the index ℓ; on the other hand, if $|y_{(\ell)}| < \tau$, then keep $\mathcal{Q}$ and $\widehat{\mathcal{S}}$ unchanged. Continue in this fashion, at each step obtaining a new measurement and performing a corresponding hypothesis test to determine whether the amplitude of the coefficient measured in that step was significant. When the overall procedure terminates it outputs its final support estimate $\widehat{\mathcal{S}}$, which essentially corresponds to the set of locations of $\mathbf{x}$ for which the corresponding measurements exceeded τ in amplitude.

The main result of [109] quantifies the performance of this type of sensing strategy for acquiring tree-sparse signals in settings where each measurement is corrupted by additive white Gaussian noise; the overall approach in this context is depicted as Algorithm 4. We provide a restatement of the main result of [109][2] here as a Lemma, and provide a proof in the appendix, for completeness. It is worth noting that the choice of data structure $\mathcal{Q}$ in the procedure implicitly determines the order in which measurements are obtained; our analysis, however, is applicable regardless of which particular data structure $\mathcal{Q}$ is used.

Lemma 4.1.1. *Specify a sparsity parameter $k' \in \mathbb{N}$, intended to be an upper-bound for the true sparsity level of the signal being acquired, and choose any $\delta \in (0,1)$. Set the threshold τ in Algorithm 4 to be*

$$\tau = \sqrt{2\sigma^2 \log\left(\frac{4k'}{\delta}\right)}. \tag{4.3}$$

Now, if the signal $\mathbf{x} \in \mathbb{R}^n$ being acquired by the procedure is k-tree sparse for some $k \geq 2$, the specified sparsity parameter k' satisfies $k' \leq \beta k$ for some $\beta \geq 1$, and the nonzero components of $\mathbf{x}$ satisfy

$$|x_i| \geq \sqrt{8\left[1 + \log\left(\frac{4\beta}{\delta}\right)\right]} \cdot \sqrt{\sigma^2 \log k}, \tag{4.4}$$

for every $i \in \mathcal{S}(\mathbf{x})$, then with probability at least $1-\delta$ the following are true: the algorithm terminates after collecting $m \leq 2k+1$ measurements, and the support estimate $\widehat{S}$ produced by the procedure satisfies $\widehat{\mathcal{S}} = \mathcal{S}(\mathbf{x})$.

In words, this result ensures that when the magnitudes of the nonzero signal components are sufficiently large – satisfying the condition specified in (4.4) – the procedure depicted in Algorithm 4 will correctly identify the support of the tree sparse vector (with high probability), and will do so using no more than $2k+1$ measurements.

Now, as a simple extension, suppose that we seek to identify the support of a k-tree sparse vector, and are equipped with a budget of m measurements, where $m \geq r(2k+1)$ for some integer constant $r \geq 1$. In this setting, the procedure described above may be easily modified to obtain a total of r measurements (each with its own independent

[2] We note that we have not attempted to optimize constants in our derivation of Lemma 4.1.1, opting instead for simple expressions that better illustrate the scaling behavior with respect to the problem parameters.

additive noise) at each step. If these replicated measurements are averaged prior to performing the hypothesis test at each step, the results of Lemma 4.1.1 can be extended directly to this setting. We formalize this extension here as a corollary.

Corollary 4.1.1. *Let* $\mathbf{x}$ *be as in Lemma 4.1.1, and consider acquiring* $\mathbf{x}$ *using a variant of the adaptive tree sensing procedure described in Algorithm 4, where* $r \geq 1$ *measurements are obtained in each step and averaged to reduce the effective measurement noise prior to each hypothesis test. Choose* $\delta \in (0,1)$ *and sparsity parameter* $k' \in \mathbb{N}$*, and set the threshold* τ *as*

$$\tau = \sqrt{2\left(\frac{\sigma^2}{r}\right)\log\left(\frac{4k'}{\delta}\right)}. \tag{4.5}$$

If $\mathbf{x}$ *is* k*-tree sparse for some* $k \geq 2$*, the sparsity parameter* $k' \leq \beta k$ *for some* $\beta \geq 1$*, and the amplitudes of the nonzero components of* $\mathbf{x}$ *satisfy*

$$|x_i| \geq \sqrt{8\left[1+\log\left(\frac{4\beta}{\delta}\right)\right]} \cdot \sqrt{\left(\frac{\sigma^2}{r}\right)\log k}, \tag{4.6}$$

for every $i \in \mathcal{S}(\mathbf{x})$ *then with probability at least* $1-\delta$ *the following are true: the algorithm terminates after collecting* $m \leq r(2k+1)$ *measurements, and the support estimate* $\widehat{S}$ *produced by the procedure satisfies* $\widehat{\mathcal{S}} = \mathcal{S}(\mathbf{x})$*.*

Note that since $m \leq r(2k+1)$ we have that $1/r \leq 3k/m$ provided $k \geq 1$. It follows from the corollary that when the sparsity parameter k' does not overestimate the true sparsity level by more than a constant factor (i.e., $\beta \geq 1$ is a *constant*), then a sufficient condition to ensure that the support estimate produced by the repeated-measurements variant of the tree sensing procedure is correct with probability at least $1-\delta$, is that the nonzero components of $\mathbf{x}$ satisfy

$$|x_i| \geq \sqrt{24\left[1+\log\left(\frac{4\beta}{\delta}\right)\right]} \cdot \sqrt{\sigma^2\left(\frac{k}{m}\right)\log k}, \tag{4.7}$$

for all $i \in \mathcal{S}(\mathbf{x})$. Identifying whether any other procedure can accurately recover the support of tree-sparse signals having fundamentally weaker amplitudes is the motivation for our present effort.

4.1.2 Problem Statement

As stated above, the essential aim of this work is to establish whether the adaptive sensing procedure for tree-sparse signals analyzed by the authors in the previous work [109], and summarized above as Algorithm 4 is optimal. Our specific focus here is on establishing fundamental performance limits for the support recovery task – that of identifying the locations of the nonzeros of $\mathbf{x}$ – in settings where $\mathbf{x}$ is k tree-sparse, and when observations may be designed either non-adaptively (e.g., measurement vectors whose elements are random and iid, as in traditional CS) or adaptively based on previous observations. We formalize this problem here.

4.1.2.1 Signal Model

Let $\mathcal{T}_{n,k}$ denote the set of all unique supports for n-dimensional vectors that are k-tree sparse in the same underlying binary tree with n nodes. For technical reasons, we further assume that the underlying trees are *nearly complete*, meaning that all levels of the underlying tree are full with the possible exception of the last (i.e., the bottom) level, and all nodes in any partially full level are as far to the left as possible.

Our specific focus will be on classes of k-tree sparse signals, $2 \leq k \leq (n+1)/2$, where each k-sparse signal $\mathbf{x}$ has support $\mathcal{S}(\mathbf{x}) \in \mathcal{T}_{n,k}$, and for which the amplitudes of all nonzero signal components are greater or equal to some non-negative quantity μ. Formally, for a given underlying tree, fixed sparsity level k, and $\mathcal{T}_{n,k}$ as described above, we define the signal class

$$\mathcal{X}_{\mu;\mathcal{T}_{n,k}} \triangleq \left\{\mathbf{x} \in \mathbb{R}^n : x_i = \alpha_i \mathbf{1}_{\{i \in T\}},\ |\alpha_i| \geq \mu > 0,\ T \in \mathcal{T}_{n,k}\right\}, \tag{4.8}$$

where $\mathbf{1}_{\{\mathcal{B}\}}$ denotes the indicator function of the event $\mathcal{B}$. In the sequel, we choose to simplify the exposition by denoting the signal class $\mathcal{X}_{\mu;\mathcal{T}_{n,k}}$ using the shorthand notation $\mathcal{X}_{\mu,k}$, effectively leaving the problem dimension and specification of the underlying tree (and corresponding set of allowable k-tree sparse supports) to be implicit. As we will see, the conditions required for accurate support recovery of k-tree sparse signals as defined above are directly related to the signal amplitude parameter μ.

4.1.2.2 Sensing Strategies

We examine the support recovery task under both adaptive and non-adaptive sensing strategies. The non-adaptive sensing strategies that we examine here are motivated by initial efforts in CS, which prescribe collecting observations using ensembles of randomly generated measurement vectors. Here, when considering performance limits of non-adaptive sensing, we consider observations obtained according to the model (4.1), where each $\mathbf{a}_j$, $j = 1, 2, \ldots, m$, is an independent random vector, whose elements are iid $\mathcal{N}(0, 1/n)$ random variables. This normalization ensures that each measurement vector has norm one in expectation; that is, $\mathbb{E}\left[\|\mathbf{a}_j\|_2^2\right] = 1$ for all $j = 1, 2, \ldots, m$. Our investigation of adaptive sensing strategies focuses on observations obtained according to (4.1), using measurement vectors satisfying $\|\mathbf{a}_j\|_2^2 = 1$, for $j = 1, 2, \ldots, m$, and for which $\mathbf{a}_j$ is allowed to explicitly depend on $\{\mathbf{a}_\ell, y_\ell\}_{\ell=1}^{j-1}$ for $j = 2, 3, \ldots, m$, as described above.

Overall, as noted in [124], we can essentially view any (non-adaptive, or adaptive) sensing strategy in terms of a collection M of *conditional distributions* of measurement vectors $\mathbf{a}_j$ given $\{\mathbf{a}_\ell, y_\ell\}_{\ell=1}^{j-1}$ for $j = 2, 3, \ldots, m$. We adopt this interpretation here, denoting by $M_{m,\text{na}}$ the specific sensing strategy based on non-adaptive Gaussian random measurements described above, and by $\mathcal{M}_m$ be the collection of all adaptive (or non-adaptive) sensing strategies based on m measurements, where each measurement vector is exactly norm one (with probability one).

4.1.2.3 Observation Noise

In each case, we model the noises associated with the linear measurements as a sequence of independent $\mathcal{N}(0, \sigma^2)$ random variables. We further assume that each noise w_j is independent of the present and all past measurement vectors $\{\mathbf{a}_\ell\}_{\ell=1}^{j}$. For the non-adaptive sensing strategies we examine here noises will also be independent of future measurement vectors, though by design, future measurement vectors generally *will not* be independent of present noises when adaptive sensing strategies are employed.

4.1.2.4 The Support Estimation Task

We define a support estimator ψ to be a (measurable) function from the space of measurement vectors and associated observations to the power set of $\{1, 2, \ldots, n\}$. In other words, an estimator ψ takes as its input a collection of measurement vectors and associated observations, $\{\mathbf{a}_j, y_j\}_{j=1}^m$, denoted by $\{\mathbf{A}_m, \mathbf{y}_m\}$ in the sequel (for shorthand), and outputs a subset of the index set $\{1, 2, \ldots, n\}$. We note that any estimator can, in general, have knowledge of the sensing strategy that was employed during the measurement process, and we make that dependence explicit here. Overall, we denote a support estimate based on observations $\mathbf{A}_m, \mathbf{y}_m$ obtained using sensing strategy M by $\psi(\mathbf{A}_m, \mathbf{y}_m; M)$.

Now, under the 0/1 loss function $d(S_1, S_2) \triangleq \mathbf{1}_{\{S_1 \neq S_2\}}$ defined on elements $S_1, S_2 \subseteq \{1, 2, \ldots, n\}$, the (maximum) risk of an estimator ψ based on sensing strategy M over the set $\mathcal{X}_{\mu,k}$ is given by

$$\begin{aligned} \mathcal{R}_{\mathcal{X}_{\mu,k}}(\psi, M) &\triangleq \sup_{\mathbf{x} \in \mathcal{X}_{\mu,k}} \mathbb{E}_{\mathbf{x}} \left[d(\psi(\mathbf{A}_m, \mathbf{y}_m; M), \mathcal{S}(\mathbf{x})) \right] \\ &= \sup_{\mathbf{x} \in \mathcal{X}_{\mu,k}} \mathbb{E}_{\mathbf{x}} \left[\mathbf{1}_{\{\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x})\}} \right] \\ &= \sup_{\mathbf{x} \in \mathcal{X}_{\mu,k}} \Pr_{\mathbf{x}} \left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x}) \right), \end{aligned} \tag{4.9}$$

where $\mathbb{E}_{\mathbf{x}}$ and $\Pr_{\mathbf{x}}$ denote, respectively, expectation and probability with respect to the joint distribution $\mathbb{P}(\mathbf{A}_m, \mathbf{y}_m; \mathbf{x}) \triangleq \mathbb{P}_{\mathbf{x}}(\mathbf{A}_m, \mathbf{y}_m)$ of the quantities $\{\mathbf{A}_m, \mathbf{y}_m\}$ that is induced when $\mathbf{x}$ is the true signal being observed. In words, the (maximum) risk essentially quantifies the worst-case performance of a specified estimator ψ when estimating the "most difficult" element $\mathbf{x} \in \mathcal{X}_{\mu,k}$ (here, the element whose support is most difficult to accurately estimate) from observations obtained via sensing strategy M.

Now, we define the *minimax risk* $\mathcal{R}^*_{\mathcal{X}_{\mu,k}, \mathcal{M}}$ associated with the class of distributions $\{\mathbb{P}_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}_{\mu,k}\}$ induced by elements $\mathbf{x} \in \mathcal{X}_{\mu,k}$ and the class $\mathcal{M}$ of allowable sensing strategies as the infimum of the (maximum) risk over all estimators ψ and sensing strategies $M \in \mathcal{M}$; that is,

$$\begin{aligned} \mathcal{R}^*_{\mathcal{X}_{\mu,k}, \mathcal{M}} &\triangleq \inf_{\psi; M \in \mathcal{M}} R_{\mathcal{X}_{\mu,k}}(\psi, M) \\ &= \inf_{\psi; M \in \mathcal{M}} \sup_{\mathbf{x} \in \mathcal{X}_{\mu,k}} \Pr_{\mathbf{x}} \left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x}) \right). \end{aligned} \tag{4.10}$$

In words, the minimax risk quantifies the error incurred by the best possible estimator when estimating the support of the "most difficult" element $\mathbf{x} \in \mathcal{X}_{\mu,k}$ using observations obtained via any sensing strategy $M \in \mathcal{M}$.

Note that when the minimax risk is bounded away from zero, so that $\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}} \geq \gamma$ for some $\gamma > 0$, it follows that regardless of the particular estimator ψ and sensing strategy $M \in \mathcal{M}$ employed, there will always be at least one signal $\mathbf{x} \in \mathcal{X}_{\mu,k}$ for which $\Pr_{\mathbf{x}}\left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x})\right) \geq \gamma$. Clearly, such settings may be undesirable in practice, since in this case we can make no *uniform* guarantees regarding accurate support recovery of signals $\mathbf{x} \in \mathcal{X}_{\mu,k}$ – there will always be some worst-case scenario for which the support recovery error probability will exceed γ. Our aim here is to identify these problematic scenarios; formally, we aim to identify signal classes $\mathcal{X}_{\mu,k}$ of the form (4.8), parameterized by their corresponding signal amplitude parameters μ, for which the minimax risk will necessarily be bounded away from zero.

4.1.3 Summary of Our Contributions

Our first main result analyzes the support recovery task for tree-sparse signals in a non-adaptive sensing scenario motivated by the randomized sensing strategies typically employed in compressive sensing. We state the result here as a theorem, and provide a proof in the next section.

Theorem 4.1.1. *Let $\mathcal{X}_{\mu,k}$ be the class of k-tree sparse n-dimensional signals defined in* (4.8) *where $2 \leq k \leq (n+1)/2$, and consider acquiring m measurements of $\mathbf{x} \in \mathcal{X}_{\mu,k}$ using the non-adaptive (random, Gaussian) sensing strategy $M_{m,\text{na}}$. If*

$$\mu \leq \sqrt{\frac{1-2\gamma}{25}} \cdot \sqrt{\sigma^2 \left(\frac{n}{m}\right) \log(k)}, \tag{4.11}$$

for some $\gamma \in (0, 1/3)$ then the minimax risk $\mathcal{R}^*_{\mathcal{X}_{\mu,k},M_{m,\text{na}}}$ defined in (4.10) *obeys the bound*

$$\mathcal{R}^*_{\mathcal{X}_{\mu,k},M_{m,\text{na}}} \geq \gamma. \tag{4.12}$$

As alluded above, the direct implication of Theorem 4.1.1 is that no uniform recovery guarantees can be made for any estimation procedure for recovering the support of tree-sparse signals $\mathbf{x} \in \mathcal{X}_{\mu,k}$ when the signal amplitude parameter μ is "too small."

Our second main result concerns support recovery in scenarios where adaptive sensing strategies may be employed. We state this result as Theorem 4.1.2, and provide a proof in the next section.

Theorem 4.1.2. *Let $\mathcal{X}_{\mu,k}$ be the class of k-tree sparse n-dimensional signals defined in* (4.8) *where $2 \leq k \leq (n+1)/2$, and consider acquiring m measurements of $\mathbf{x} \in \mathcal{X}_{\mu,k}$ using any sensing strategy $M \in \mathcal{M}_m$. If*

$$\mu \leq (1-2\gamma)\sqrt{\sigma^2\left(\frac{k}{m}\right)}, \tag{4.13}$$

for some $\gamma \in (0, 1/3)$ then the minimax risk $\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}_m}$ defined in (4.10) *obeys the bound*

$$\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}_m} \geq \gamma. \tag{4.14}$$

Similar to the discussion following the statement of Theorem 4.1.1 above, here we have that that no uniform guarantees can be made regarding accurate support recovery of signals $\mathbf{x} \in \mathcal{X}_{\mu,k}$ for small μ.

Table 4.1 depicts a summary of our main results in a broader context. Overall, we compare four distinct scenarios corresponding to a taxonomy of adaptive and non-adaptive sensing strategies for recovering k-sparse signals under assumptions of unstructured sparsity and tree sparsity. For each, we identify (up to an unstated constant) a critical value of the signal amplitude parameter, say μ^*, such that for the support recovery task the minimax risk over the class $\mathcal{X}_{\mu,k}$ will necessarily be bounded away from zero when $\mu \leq \mu^*$. The conditions for support estimation of *unstructured* sparse vectors listed in Table 4.1 are a restatement of some known results, and are provided here (with references) for comparison[3] . Our main contributions here are depicted in the bottom row of the table, which correspond to the values identified in equations (4.11) and (4.13), respectively (with the leading multiplicative factors suppressed).

[3] Necessary conditions on the signal amplitude parameter required for exact support recovery from non-adaptive compressive samples (and for unstructured sparse signals) were provided in [139]; related efforts along these lines include [140–144]. Necessary conditions for exact support recovery using adaptive sensing strategies were provided in [145] for the case where the number of measurements exceeds the signal dimension ($m > n$), while to the best of our knowledge results of this flavor have not yet been established for the compressive regime (where $m < n$). Finally, we note that several related efforts have established necessary conditions for weaker metrics of *approximate* support recovery using non-adaptive sensing [146, 147] and adaptive sensing strategies [119, 124].

Table 4.1: Summary of necessary conditions for exact support recovery using non-adaptive or adaptive sensing strategies that obtain m measurements of k-sparse n-dimensional signals that are either unstructured or tree sparse in an underlying nearly complete binary tree. For each setting, we state the critical value of μ such that whenever μ is smaller than a constant times the stated quantity, the minimax risk over the class of signals $\mathcal{X}_{\mu,k}$ of the form (4.8) will be strictly bounded away from zero.

Sampling Strategy Sparsity Model	Non-adaptive Sensing	Adaptive Sensing
Unstructured Sparsity	$\sqrt{\sigma^2 \left(\frac{n}{m}\right) \log n}$ [139]; see also [140–144]	$\sqrt{\sigma^2 \left(\frac{n}{m}\right) \log k}$ [145] (when $m > n$)
Tree Sparsity	$\sqrt{\sigma^2 \left(\frac{n}{m}\right) \log k}$ Theorem 4.1.1	$\sqrt{\sigma^2 \left(\frac{k}{m}\right)}$ Theorem 4.1.2

Two salient points are worth noting when comparing the necessary conditions summarized in Table 4.1 with the sufficient condition (4.7) for the repeated-measurement variant of the adaptive tree sensing procedure of Algorithm 4. First, the results of Theorem 4.1.2, summarized in the lower-right corner of Table 4.1, address our overall question – the simple adaptive tree sensing procedure described above is indeed nearly optimal for estimating the support of k-tree sparse vectors, in the following sense: Corollary 4.1.1 describes a technique that accurately recovers (with probability at least $1-\delta$, where δ can be made arbitrarily small) the support of any k-tree sparse signal from $m \leq r(2k+1)$ measurements, provided the amplitudes of the nonzero signal components all exceed $c_\delta \cdot \sqrt{\sigma^2\,(k/m) \log k}$ for some constant c_δ. On the other hand, for any estimation strategy based on any adaptive or non-adaptive sensing method, support recovery will fail (with probability at least γ) to accurately recover the support of some signal or signals in a class comprised of k-tree sparse vectors whose nonzero components exceed $c_\gamma \cdot \sqrt{\sigma^2\,(k/m)}$ in amplitude, for a constant c_γ.

The second noteworthy point here concerns the *relative* performances of the four strategies summarized in Table 4.1. Overall, we see that techniques that *either* employ adaptive sensing strategies *or* exploit tree structure in the signal being inferred (but not both) may indeed outperform non-adaptive sensing techniques that do not exploit structure, in the sense that either may succeed in recovering signals whose nonzero components are weaker. That said, the potential improvement arises only in the logarithmic factor present in the amplitudes, implying that either of these improvements by themselves can recover signals whose amplitudes are weaker by a factor that is (at best) a constant multiple of $\sqrt{\log k/\log n}$. On the other hand, techniques that leverage *both* adaptivity *and* structure, such as the adaptive tree sensing strategy analyzed above, can provably recover signals whose nonzero component amplitudes are *significantly* weaker than those that can be recovered via any of the other strategies depicted in the table. Specifically, in this case the relative difference in amplitudes is on the order of a constant times $\sqrt{k/(n\log n)}$, which could be much more significant, especially in high-dimensional settings. The experimental evaluation in Section 4.3 provides some additional empirical evidence along these lines.

4.1.4 Relations to Existing Works

As alluded above, several recent efforts have proposed (e.g., [136–138]) and analyzed (e.g., [128, 129]) specialized techniques for estimating tree-sparse signals from non-adaptive compressive samples, each of which are designed to exploit the fundamental connectivity structure present in the underlying signal during the inference task. The work [2] was among the first to propose and experimentally evaluate a direct wavelet sensing approach for acquiring and estimating wavelet sparse signals (there, images) in the context of compressive sensing tasks, and the sample complexity of a similar procedure in noise free settings was analyzed in [3], [4]. These works served as the motivation for our initial investigation [109] into the performance of such approaches in noisy settings.

Since our work [109] appeared, several related efforts in the literature have investigated adaptive sensing strategies for structured sparse signals. The work [122], for example, examined the problem of localizing block-structured activations in matrices from noisy measurements, and established fundamental limits for this task using proof techniques based on [148]. We adopt a similar approach based on [148] below in the proof of one of our main results. A follow-on work [123] examined a more general setting, that of support recovery of signals whose supports correspond to (unions of) smaller clusters in some underlying graph. That work assumed that the clusters comprising the signal model were such that they could be organized into a (nearly balanced) hierarchical clustering having relatively few levels. While this model is quite general, we note that the class of tree sparse signals we consider here comprise a particularly difficult (in fact, nearly pathological!) scenario for the strategy of [123]; indeed, the tree-sparse case comprises one example of a problematic scenario identified in [123] where that approach "does not significantly help when distinguishing clusters that differ only by a few vertices."

It is interesting to note that different structure models can give rise to different thresholds for localization from non-adaptive measurements. We note, for example, that the thresholds identified in [122] for localizing block-sparse signals using non-adaptive compressive measurements are weaker than the corresponding threshold we identify in Theorem 4.1.1 here for localizing tree-sparse signals[4] . This difference arises as a direct result of the different signal models, and in particular, how these differences manifest themselves in the reduction strategy inherent in the proofs based on the ideas of [148]. For the analysis of block-sparse signals in [122] the reduction to hypotheses that are difficult to distinguish leads to consideration of block-sparse signals that either differ on about $k^{1/2}$ locations or do not overlap at all, while in contrast, the performance limits in our case are dictated by tree sparse signals that can differ on as few as two locations. Stated another way, the tree-sparse signal model we consider here contains subsets of signals that are necessarily more difficult to discern than does the block-sparse model analyzed in [122], and this gives rise to the higher necessary signal amplitude thresholds required for localization using non-adaptive compressive measurements for the tree-sparse model we examine here, as compared with the block-sparse model examined in [122].

[4] Specifically, the results of [122] imply (adapted to the notation we employ here) that accurate localization of block-sparse signals is impossible when the nonzero signal components have amplitudes smaller than a constant times $\sqrt{\sigma^2 \left(\frac{n}{m}\right) \max\left\{\frac{1}{k^{1/2}}, \frac{\log n}{k}\right\}}$.

We also note a recent related work which proposed a technique for sensing signals that are “almost” tree-sparse in a wavelet representation, in the sense that their supports may correspond to disconnected subtrees in some underlying tree [132]. While the sensing strategy proposed in that work was demonstrated experimentally to be effective for acquiring natural images, only a partial analysis of the procedure was provided. Specifically, [132] analyzed their procedure only for the case where the signal supports *do* correspond to connected subtrees in some underlying tree, which was effectively the case analyzed in [109]. Further, the analysis in [132] did not explicitly quantify the sufficient conditions on the signal component amplitudes for which the procedure would successfully recovery the signal support, stating instead only that $m = 2k + 1$ measurements were sufficient to recover the support provided the SNR was “sufficiently large.”

While our focus here is specifically on the support recovery task, we note that the related prior work [122] also identified fundamental limits for the task of *detecting* the presence of block-structured activations in matrices using adaptive or non-adaptive measurements, and established that signals whose nonzero components are essentially “too weak” cannot be reliably detected by any method. Analogous fundamental limits for the detection of certain *tree-sparse* signals have also been established in the literature. Specifically, in the context of our effort here, the problem examined in [149] may be viewed in terms of identifying the support of (a subset of) tree sparse signals whose nonzero elements have the same amplitude μ, from a total of $m = n$ noisy measurements, corresponding to one measurement per node of the underlying tree. Interestingly, that work established that all detection approaches (for simple trees with no branching) are unreliable when $\mu < c\sqrt{\sigma^2(n/m)} = c\sqrt{\sigma^2}$ for a specified constant $c > 0$. This threshold differs from the lower bound we establish for the support recovery task by only a logarithmic factor. This slight difference may arise from the fact that our tree-sparse model contains many more allowable supports (and therefore, more signal candidates) than the path-based model examined in [149], or it may be that, (at least for the “full-measurement” scenario where $m = n$) the support recovery task is slightly more difficult than the detection task. A full characterization of this type of detection problem for general tree-sparse signals, in settings where measurements may be compressive ($m < n$) as well as adaptive or non-adaptive, is beyond of the scope of our effort here, and remains an (as yet) open problem.

Finally, while our focus here was specifically on the adaptive tree-sensing strategy and fundamental recovery limits for tree-sparse signals, we note that previous results have established that the necessary conditions for recovery of unstructured sparse signals in the top row of Table 4.1 are essentially *tight*, in the sense that there exist sensing strategies and associated estimation procedures in each case that are capable of accurate support recovery of sparse signals whose nonzero components exceed a constant times the specified quantity – see, for example, [139,140,150,151], which consider the identification of necessary conditions for support recovery of (unstructured) sparse signals from non-adaptive measurements, and [145], which analyzes an adaptive sensing strategy for recovering (unstructured) sparse vectors in noisy settings. Support recovery of (group) structured sparse signals was also examined recently in [152–154].

4.1.5 Organization

The remainder of this paper is organized as follows. The proofs of our main results, Theorems 4.1.1 and 4.1.2, are presented in Section 4.2. In Section 4.3 we provide an experimental evaluation of the support recovery task for tree sparse signals. Specifically, where we compare the performance of the tree sensing procedure described above with an inference procedure based on non-adaptive (compressive) sensing that is designed to exploit the tree structure, as well as with adaptive and non-adaptive CS techniques that are agnostic to the underlying tree structure. We also provide experimental evidence to validate the scaling behavior predicted in (4.7) for a fixed measurement budget. We discuss some natural extensions of this effort, and provide a few concluding remarks, in Section 4.4. Several auxiliary results, as well as a proof of Lemma 4.1.1, are relegated to the Appendix.

4.2 Proofs of Main Results

Our first main result, Theorem 4.1.1, concerns the support recovery task for tree-sparse signals in a non-adaptive sensing scenario motivated by the randomized sensing strategies typically employed in compressive sensing. Our analysis here follows a similar strategy as in a recent related effort [122], which is based on the general reduction strategy described by Tsybakov [148]. Our second main result, Theorem 4.1.2, concerns support recovery for tree-sparse vectors in scenarios where adaptive sensing strategies may be employed. Our proof approach in this scenario is again based on a reduction strategy – we argue (formally) that the support recovery task in this case is at least as difficult as the task of localizing a single nonzero signal component of a vector of reduced dimension, and leverage a result of the recent work [119] which examined support recovery from non-adaptive measurements for general (unstructured) sparse signals.

Before we proceed, we first introduce some notation that will be used throughout the proofs here. For any $T \in \mathcal{T}_{n,\ell}$ with $1 \leq \ell < n$, corresponding to the support of a rooted connected subtree with ℓ nodes (in some underlying nearly complete binary tree with n nodes), we define $N(T)$ to be the set of locations in the underlying tree, such that for any $j \in N(T)$ the augmented set $T \cup j$ corresponds to a tree with $\ell + 1 \leq n$ nodes that is itself another rooted connected subtree of the same underlying tree. Formally, for $T \in \mathcal{T}_{n,\ell}$ we define

$$N(T) \triangleq \{j \in \{1, 2, \ldots, n\} \ : \ \{T \cup j\} \in \mathcal{T}_{n,\ell+1}\} . \tag{4.15}$$

With this, we are in position to proceed with the proofs of Theorems 4.1.1 and 4.1.2.

4.2.1 Proof of Theorem 4.1.1

The result of Theorem 4.1.1 quantifies the limits of support recovery for tree sparse signals using non-adaptive randomized sensing strategies. Our analysis is based on the general reduction strategy proposed by Tsybakov [148], and follows a similar approach as that in a recent, related effort that identified performance limits for estimating block-structured matrices from noisy measurements [122].

Recall the problem formulation and notation introduced in the previous section, and note that for any set $\mathcal{X}'_{\mu,k} \subseteq \mathcal{X}_{\mu,k}$, any estimator ψ, and any measurement strategy $M \in \mathcal{M}$, we have that

$$\sup_{\mathbf{x} \in \mathcal{X}_{\mu,k}} \Pr_{\mathbf{x}} \left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x})\right) \geq \sup_{\mathbf{x} \in \mathcal{X}'_{\mu,k}} \Pr_{\mathbf{x}} \left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x})\right), \tag{4.16}$$

where as described above the notation $\Pr_{\mathbf{x}}(\cdot)$ denotes probability with respect to the joint distribution $\mathbb{P}(\mathbf{A}_m, \mathbf{y}_m; \mathbf{x}) \triangleq \mathbb{P}_{\mathbf{x}}(\mathbf{A}_m, \mathbf{y}_m)$ of the quantities $\mathbf{A}_m$ and $\mathbf{y}_m$ that is induced when $\mathbf{x}$ is the true signal being observed. This implies, in particular, that

$$\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}} \geq \mathcal{R}^*_{\mathcal{X}'_{\mu,k},\mathcal{M}} \tag{4.17}$$

and it follows that we can obtain valid lower bounds on $\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}}$ by instead seeking lower bounds on the minimax risk over any restricted signal class $\mathcal{X}'_{\mu,k} \subseteq \mathcal{X}_{\mu,k}$. This is the strategy we employ here.

For technical reasons we address the cases $k = 2$ and $3 \leq k \leq (n+1)/2$ separately, but the essential approach is similar in both cases. Namely, for each k we construct a set $\mathcal{X}'_{\mu,k}$ of signals whose nonzero components have the *same* amplitude μ, and whose supports are "close" in the sense that the symmetric difference between supports of any pair of distinct signals in the class is a set of cardinality two. In each case these signal classes are of the form

$$\mathcal{X}'_{\mu,k}(T^*) \triangleq \left\{\mathbf{x} \in \mathbb{R}^n : x_i = \mu \mathbf{1}_{\{i \in T\}},\ T = T^* \cup j,\ j \in N(T^*)\right\}, \tag{4.18}$$

for some (specific) $T^* \in \mathcal{T}_{n,k-1}$ and $N(T)$ is as defined above. This allows us to reduce our problem to the consideration of a hypothesis testing problem over a countable (and finite) number of elements $\mathbf{x} \in \mathcal{X}'_{\mu,k}$.

4.2.1.1 Case 1: $k = 2$

We begin by choosing $T^* \in \mathcal{T}_{n,1}$ to be an element of $\mathcal{T}_{n,1}$ for which $|N(T^*)| = 2$, and for this T^* we form the set $\mathcal{X}_{\mu,2}$ of the form (4.18) above[5] . It follows from the definition of $N(T^*)$ that $\mathcal{X}'_{\mu,2}(T^*)$ is a set of signals whose supports are each an element of $\mathcal{T}_{n,2}$, and since each nonzero element has amplitude exactly equal to μ, it follows that every $\mathbf{x} \in \mathcal{X}'_{\mu,2}(T^*)$ is also an element of the class of signals $\mathcal{X}_{\mu,2}$ defined in (4.8) when $k = 2$. Thus, we have overall that $\mathcal{X}'_{\mu,2}(T^*) \subset \mathcal{X}_{\mu,2}$. Now, our approach is to obtain lower bounds on the minimax risk $\mathcal{R}^*_{\mathcal{X}_{\mu,2},\mathcal{M}}$ when $\mathcal{M} = \{M_{m,\text{na}}\}$ by considering the minimax risk over the set $\mathcal{X}'_{\mu,2}(T^*)$, which ultimately corresponds to assessing the error performance of a hypothesis testing problem with two simple hypotheses.

[5] Note that this is a somewhat degenerate scenario – here, T^* can be chosen to be the set that contains only the index of the root node of the underlying tree. Further, that $k \leq (n+1)/2$ implies $n \geq 3$ here, and since the underlying tree is assumed nearly complete, it follows that the root node has two descendants in the underlying tree.

Our analysis relies on a result of Tsybakov [148, Theorem 2.2], which provides lower-bounds on the minimax probability of error for a binary hypothesis testing problem. We state that result here as a lemma.

Lemma 4.2.1 (Tsybakov)**.** *Let $\mathbb{P}_0, \mathbb{P}_1$ be probability distributions (on a common measurable space) for which the Kullback-Leibler (KL) divergence of $\mathbb{P}_0$ from $\mathbb{P}_1$ satisfies $K(\mathbb{P}_1, \mathbb{P}_0) \leq \alpha < \infty$. Then, the minimax probability of error over all (measurable) tests ψ that map observations to an element of the set $\{0, 1\}$, given by*

$$p_{e,1} \triangleq \inf_{\psi} \max_{j=0,1} Pr_j\left(\psi \neq j\right), \tag{4.19}$$

where $Pr_j(\cdot)$ denotes probability with respect to the distribution $\mathbb{P}_j$ induced on the observations when hypothesis j is the correct hypothesis, obeys the bound

$$p_{e,1} \geq \max\left\{\frac{1}{4}\exp\left(-\alpha\right), \frac{1-\sqrt{\alpha/2}}{2}\right\} \geq \frac{1-\sqrt{\alpha/2}}{2}. \tag{4.20}$$

In order to apply this result in our setting, we first need to evaluate the KL divergence $K(\mathbb{P}_1, \mathbb{P}_0)$, where $\mathbb{P}_1$ and $\mathbb{P}_0$ are distributions that characterize our testing problem of identifying which of the two unique elements $\mathbf{x}_0, \mathbf{x}_1 \in \mathcal{X}'_{\mu,2}(T^*)$, respectively, was observed. Now, under the assumption here that the elements of each measurement vector are (iid) Gaussian distributed, we have that the KL divergence of $\mathbb{P}_0$ from $\mathbb{P}_1$ can be expressed in terms of the corresponding probability densities $f_1 = f_1(\{\mathbf{a}_i, y_i\}_{i=1}^m)$ and $f_0 = f_0(\{\mathbf{a}_i, y_i\}_{i=1}^m)$ as

$$K(\mathbb{P}_1, \mathbb{P}_0) = \mathbb{E}_1\left[\log\left(\frac{f_1(\mathbf{A}_m, \mathbf{y}_m)}{f_0(\mathbf{A}_m, \mathbf{y}_m)}\right)\right], \tag{4.21}$$

which is just the expectation of the log-likelihood ratio with respect to the distribution $\mathbb{P}_1$.

It follows from the assumptions of our measurement model, specifically that the measurement vectors and noises are mutually independent, that each of the densities f_p, $p \in \{0, 1\}$, can be factored in the form

$$f_p(\mathbf{A}_m, \mathbf{y}_m) = \prod_{i=1}^{m} f(\mathbf{a}_i)\ f_p(y_i|\mathbf{a}_i) \tag{4.22}$$

where each $f(\mathbf{a}_i)$ is multivariate Gaussian density and $f_p(y_i|\mathbf{a}_i)$ is a (signal-dependent) conditional density of the observation y_i given the measurement vector $\mathbf{a}_i$. Note that the conditional densities of y_i given $\mathbf{a}_i$ are also Gaussian distributed because of the additive noise modeling assumptions. Overall, the log-likelihood ratio in (4.21) can be simplified as

$$\begin{aligned}
\log\left(\frac{f_1(\mathbf{A}_m,\mathbf{y}_m)}{f_0(\mathbf{A}_m,\mathbf{y}_m)}\right) &= \sum_{i=1}^{m}\log\left(\frac{f_1(y_i|\mathbf{a}_i)}{f_0(y_i|\mathbf{a}_i)}\right)\\
&= \sum_{i=1}^{m}\frac{\left(y_i-\mathbf{a}_i^T\mathbf{x}_0\right)^2-\left(y_i-\mathbf{a}_i^T\mathbf{x}_1\right)^2}{2\sigma^2}\\
&= \sum_{i=1}^{m}\frac{\left(\mathbf{a}_i^T\mathbf{x}_0\right)^2-2y_i\mathbf{a}_i^T\mathbf{x}_0-\left(\mathbf{a}_i^T\mathbf{x}_1\right)^2+2y_i\mathbf{a}_i^T\mathbf{x}_1}{2\sigma^2}. \qquad (4.23)
\end{aligned}$$

Now, using the fact that under the distribution $\mathbb{P}_1$ we have that $y_i = \mathbf{a}_i^T\mathbf{x}_1 + w_i$ for $i = 1,\ldots,m$, and that the noise w_i is zero mean and independent of $\mathbf{a}_i$, we can simplify the expression (4.21) as

$$K(\mathbb{P}_1,\mathbb{P}_0) = \mathbb{E}_1\left[\sum_{i=1}^{m}\frac{\left(\mathbf{a}_i^T(\mathbf{x}_1-\mathbf{x}_0)\right)^2}{2\sigma^2}\right].$$

Note that by the construction of $\mathcal{X}'_{\mu,2}(T^*)$, the vector $\mathbf{x}_1-\mathbf{x}_0$ has exactly two nonzero elements, each having amplitude μ (but with different signs). It follows that $\mathbf{a}_i^T(\mathbf{x}_1 - \mathbf{x}_0) \sim \mathcal{N}(0, 2\mu^2/n)$ for each $i = 1,\ldots,m$, and thus the KL divergence can be expressed simply as

$$K(\mathbb{P}_1,\mathbb{P}_0) = \frac{m\mu^2}{n\sigma^2}. \qquad (4.24)$$

Letting $\alpha = \frac{m\mu^2}{n\sigma^2}$, it is easy to see from (4.20) that if $\alpha \leq 2(1-2\gamma)^2$, or equivalently, if

$$\begin{aligned}
\mu &\leq \sqrt{2(1-2\gamma)^2}\cdot\sqrt{\sigma^2\left(\frac{n}{m}\right)}\\
&= \sqrt{\frac{2(1-2\gamma)^2}{\log 2}}\cdot\sqrt{\sigma^2\left(\frac{n}{m}\right)\log 2}, \qquad (4.25)
\end{aligned}$$

for any $\gamma \in (0,1/2)$, then $p_{e,1} \geq \gamma$.

4.2.1.2 Case 2: $3 \leq k \leq (n+1)/2$

Analogously to the $k = 2$ case, we begin by choosing $T^* \in \mathcal{T}_{n,k-1}$ to be an element of $T_{n,k-1}$ for which $|N(T^*)| = k$ (the existence of such an element T^* is established by Lemma 4.5.2 in the appendix) and constructing the set $\mathcal{X}'_{\mu,k}(T^*)$ to be of the form (4.18). As in the previous case, it follows here that $\mathcal{X}'_{\mu,k}(T^*) \subset \mathcal{X}_{\mu,k}$, so our approach here ultimately corresponds to assessing the error performance of a multiple hypothesis testing problem with k simple hypotheses.

We again employ a result of Tsybakov [148, Proposition 2.3], which provides lower-bounds on the minimax probability of error for a hypothesis testing problem deciding among some $L+1$ hypotheses. We state that result here as a lemma.

Lemma 4.2.2 (Tsybakov)**.** *Let $\mathbb{P}_0, \ldots, \mathbb{P}_L$ be probability distributions (on a common measurable space) satisfying*

$$\frac{1}{L}\sum_{j=1}^{L} K(\mathbb{P}_j, \mathbb{P}_0) \leq \alpha \tag{4.26}$$

with $0 < \alpha < \infty$. Then, the minimax probability of error over all (measurable) tests ψ that map observations to an element of the set $\{0, 1, \ldots, L\}$, given by

$$p_{e,L} \triangleq \inf_{\psi} \max_{0 \leq j \leq L} Pr_j\left(\psi \neq j\right), \tag{4.27}$$

obeys the bound

$$p_{e,L} \geq \sup_{0<\tau<1}\left[\frac{\tau L}{1+\tau L}\left(1 + \frac{\alpha + \sqrt{\alpha/2}}{\log \tau}\right)\right]. \tag{4.28}$$

As in the previous case we again need to evaluate KL divergences, this time for pairs of distributions $\mathbb{P}_p$ and $\mathbb{P}_q$ induced by signals $\mathbf{x}_p, \mathbf{x}_q \in \mathcal{X}'_{\mu,k}(T^*)$. The computation of each KL divergence mirrors the derivation in the previous case; overall, it is straightforward to show that

$$\frac{1}{L}\sum_{j=1}^{L} K(\mathbb{P}_j, \mathbb{P}_0) = \frac{m\mu^2}{n\sigma^2}. \tag{4.29}$$

Now, note that we can lower-bound the supremum term in the minimum probability of error expression (4.28) by evaluating the right hand side for any $\tau \in (0,1)$. Since our test is over k hypotheses we let $L = k-1$ here. Further, since we consider the case $k \geq 3$ here, we have that $L \geq 2$, so we can choose $\tau = 1/\sqrt{L} \in (0,1)$ to obtain that under the conditions of Lemma 4.2.2,

$$\begin{aligned} p_{e,L} &\geq \frac{\sqrt{L}}{1+\sqrt{L}}\left(1+\frac{\alpha+\sqrt{\alpha/2}}{\log(1/\sqrt{L})}\right) \\ &\geq \frac{1}{2}\left(1-\frac{(2\alpha+\sqrt{2\alpha})}{\log L}\right) \\ &= \frac{1}{2}\left(1-\frac{(2\alpha+\sqrt{2\alpha})}{\log(k-1)}\right). \end{aligned} \tag{4.30}$$

Now, note that for any $\gamma \in (0,1/3)$, we have $p_{e,L} \geq \gamma$ whenever $2\alpha+\sqrt{2\alpha} \leq (1-2\gamma)\log(k-1)$, or equivalently, whenever α satisfies

$$0 \leq \sqrt{\alpha} \leq \frac{\sqrt{2+8(1-2\gamma)\log(k-1)}-\sqrt{2}}{4}, \tag{4.31}$$

which follows from the monotonicity of the function $2\alpha+\sqrt{2\alpha}$ and a straightforward application of the quadratic formula.

As in the previous case, we let $\alpha = \frac{m\mu^2}{n\sigma^2}$ and simplify to obtain that $p_{e,L} \geq \gamma$ whenever

$$\mu \leq \left[\sqrt{1+\frac{1}{f_{\gamma,k}}}-\sqrt{\frac{1}{f_{\gamma,k}}}\,\right]\sqrt{\frac{1-2\gamma}{2}}\cdot\sqrt{\sigma^2\frac{n}{m}\log(k-1)}, \tag{4.32}$$

where $f_{\gamma,k} = 4(1-2\gamma)\log(k-1)$. Now, for the range of k and γ values we consider here we have that $f_{\gamma,k} \geq (4/3)\log(2)$, implying (after a straightforward calculation) that the term in square brackets in (4.32) is always greater than $0.4 = 2/5$. Thus, we see that $p_{e,L} \geq \gamma$ whenever

$$\mu \leq \sqrt{\frac{2(1-2\gamma)}{25}}\cdot\sqrt{\sigma^2\left(\frac{n}{m}\right)\log(k-1)}, \tag{4.33}$$

We make one more simplification, using the fact that $\log(k)/2 < \log(k-1)$ when $k \geq 3$, to claim that if

$$\mu \leq \sqrt{\frac{1-2\gamma}{25}}\cdot\sqrt{\sigma^2\left(\frac{n}{m}\right)\log k}, \tag{4.34}$$

then $p_{e,L} \geq \gamma$.

4.2.1.3 Putting the Results Together

In order to combine the results from the previous two cases into one concise form, we first note that for $\gamma \in (0, 1/3)$,

$$\sqrt{\frac{1-2\gamma}{25}} < \sqrt{\frac{2(1-2\gamma)^2}{\log 2}}. \tag{4.35}$$

With this, we can claim overall that for any $2 \leq k < (n+1)/2$, if for some $\gamma \in (0, 1/3)$,

$$\mu \leq \sqrt{\frac{1-2\gamma}{25}} \cdot \sqrt{\sigma^2 \left(\frac{n}{m}\right) \log k}, \tag{4.36}$$

then the minimax risk over the class $\mathcal{X}_{\mu,k}$ of k-tree sparse signals defined in (4.8) satisfies $\mathcal{R}^*_{\mathcal{X}_\mu, M_{m,\mathrm{na}}} \geq \gamma$, as claimed.

4.2.2 Proof of Theorem 4.1.2

Our proof approach in this scenario leverages an essential result from recent efforts characterizing the fundamental limits of support recovery for one-sparse n-dimensional vectors [119]. In order to put the results of that work into context here, let us define a class of one-sparse n-dimensional vectors as

$$\mathcal{X}_\mu^{(1)} \triangleq \left\{\mathbf{x} \in \mathbb{R}^n : x_i = \alpha_i \mathbf{1}_{\{i \in T\}},\ |\alpha_i| \geq \mu > 0,\ T \in [n]\right\}, \tag{4.37}$$

where $[n] = \{1, 2, \ldots, n\}$. Note that we use slightly different notation for the signal class to distinguish it from the tree-sparse classes described above. In particular, signals in the class (4.37) could have their support on any element of $\{1, 2, \ldots, n\}$, while in contrast, one-sparse signals that are also tree-sparse must be such that their single nonzero occurs at the root of the underlying tree.

In terms of the definition (4.37) above, the results of [119] (see also the discussion following [155, Theorem 2]) can be summarized as a lemma.

Lemma 4.2.3. *The minimax risk*

$$\mathcal{R}^*_{\mathcal{X}_\mu^{(1)}, \mathcal{M}_m} = \inf_{\psi; M \in \mathcal{M}_m} \sup_{\mathbf{x} \in \mathcal{X}_\mu^{(1)}} Pr_{\mathbf{x}} \left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x})\right) \tag{4.38}$$

over all support estimators ψ and sensing strategies $M \in \mathcal{M}_m$ satisfies the bound

$$\mathcal{R}^*_{\mathcal{X}_\mu^{(1)}, \mathcal{M}_m} \geq \frac{1}{2}\left(1 - \sqrt{\frac{m\mu^2}{n\sigma^2}}\right). \tag{4.39}$$

It follows directly from this result that if

$$\mu \leq (1-2\gamma)\sqrt{\sigma^2\left(\frac{n}{m}\right)} \tag{4.40}$$

for some $\gamma \in (0, 1/3)$, then $\mathcal{R}^*_{\mathcal{X}_\mu^{(1)},\mathcal{M}_m} \geq \gamma$. We proceed here by showing (formally) that our problem of interest – recovering the support of a k-tree sparse n-dimensional vector using any estimator and any adaptive sensing strategy – is at least as difficult as recovering the support of a one-sparse vector in some $\widetilde{n} < n$ dimensional space using any estimator and any sensing strategy $M \in \mathcal{M}_m$. Then, we adapt the result of Lemma 4.2.3 to establish Theorem 4.1.2.

We will find it useful in the derivation that follows to introduce an alternative, but equivalent, notation to describe the support estimators and signal supports. Namely, we associate with any support estimator ψ a corresponding n-dimensional vector-valued function $\boldsymbol{\varphi} = [\varphi_1\ \varphi_2\ \ldots\ \varphi_n]^T$, such that each support estimate $\psi(\mathbf{A}_m, \mathbf{y}_m; M)$ corresponds to a vector whose elements are given by

$$\varphi_i(\mathbf{A}_m, \mathbf{y}_m; M) = \mathbf{1}_{\{i \in \psi(\mathbf{A}_m, \mathbf{y}_m; M)\}}, \tag{4.41}$$

for $i = 1, 2, \ldots, n$. Similarly, we can interpret the signal support $\mathcal{S}(\mathbf{x})$ of any vector $\mathbf{x}$ in terms of an n-dimensional binary vector $\mathbf{S}(\mathbf{x}) = [S_1(\mathbf{x})\ S_2(\mathbf{x})\ \ldots\ S_n(\mathbf{x})]^T$ with elements

$$S_i(\mathbf{x}) = \mathbf{1}_{\{i \in \mathcal{S}(\mathbf{x})\}}, \tag{4.42}$$

for $i = 1, \ldots, n$.

As in the proof of Theorem 4.1.1, for any fixed $2 \leq k \leq (n+1)/2$ we choose $T^* \in \mathcal{T}_{n,k-1}$ to be an element of $\mathcal{T}_{n,k-1}$ for which $|N(T^*)| = k$, and let $\mathcal{X}'_{\mu,k}(T^*)$ be of the form (4.18). Now, observe

$$\begin{aligned}
\sup_{\mathbf{x} \in \mathcal{X}_{\mu,k}} \Pr_{\mathbf{x}}\left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x})\right) &\geq \sup_{\mathbf{x} \in \mathcal{X}'_{\mu,k}(T^*)} \Pr_{\mathbf{x}}\left(\psi(\mathbf{A}_m, \mathbf{y}_m; M) \neq \mathcal{S}(\mathbf{x})\right) \\
&= \sup_{\mathbf{x} \in \mathcal{X}'_{\mu,k}(T^*)} \Pr_{\mathbf{x}}\left(\cup_{i=1}^n \{\varphi_i(\mathbf{A}_m, \mathbf{y}_m; M) \neq S_i(\mathbf{x})\}\right) \\
&\geq \sup_{\mathbf{x} \in \mathcal{X}'_{\mu,k}(T^*)} \Pr_{\mathbf{x}}\left(\cup_{i \in \mathcal{I}} \{\varphi_i(\mathbf{A}_m, \mathbf{y}_m; M) \neq S_i(\mathbf{x})\}\right),
\end{aligned}$$

where in the last line $\mathcal{I}$ is any subset of $\{1, 2, \ldots, n\}$. In particular, this implies that

$$\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}_m} \geq \inf_{\boldsymbol{\varphi}; M \in \mathcal{M}_m} \sup_{\mathbf{x} \in \mathcal{X}'_{\mu,k}(T^*)} \Pr_{\mathbf{x}}\left(\cup_{i \in N(T^*)} \{\mathcal{E}_i\}\right), \tag{4.43}$$

where $\mathcal{E}_i$ is the event $\{\varphi_i(\mathbf{A}_m, \mathbf{y}_m) \neq S_i(\mathbf{x})\}$. Now, since for any signal $\mathbf{x} \in \mathcal{X}'_{\mu,k}(T^*)$ the collection $\{S_i(\mathbf{x})\}_{i \in N(T^*)}$ contains exactly one '1' and $k-1$ zeros, it follows that the right hand side of (4.43) is equivalent to the minimax risk associated with the task of recovering the support of a one-sparse $|N(T^*)|$-dimensional vector whose single nonzero element has amplitude μ, in settings where measurements can be obtained via any (possibly adaptive) sensing strategy $M \in \mathcal{M}_m$. Thus, we can employ the result of Lemma 4.2.3 to conclude that

$$\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}_m} \geq \frac{1}{2}\left(1 - \sqrt{\frac{m\mu^2}{|N(T^*)|\sigma^2}}\right). \tag{4.44}$$

Finally, since $|N(T^*)| = k$, it follows that if for any $\gamma \in (0, 1/3)$ we have

$$\mu \leq (1-2\gamma)\sqrt{\sigma^2\left(\frac{k}{m}\right)}, \tag{4.45}$$

then $\mathcal{R}^*_{\mathcal{X}_{\mu,k},\mathcal{M}_m} \geq \gamma$, as claimed.

4.3 Experimental Evaluation

In this section we provide several experimental evaluations to validate our theoretical results, and to illustrate the performance improvements that can be achieved in the support recovery task using the adaptive tree sensing procedure.

In our first experiment we investigate the performance of the tree-sensing approach analyzed here, as the underlying signal dimension increases, and compare the performance of the tree sensing approach with three other strategies from the literature. Overall, we evaluate four sensing and support estimation strategies, each of which corresponds to one of the four scenarios identified in Table 4.1 (adaptive vs. non-adaptive sensing, and structured vs. unstructured sparsity). The support estimation strategies based on non-adaptive sensing that we evaluate here each utilize measurements obtained according to the model (4.1), where measurement vectors are independent Gaussian random vectors with iid $\mathcal{N}(0, 1/n)$ elements. They are

- (non-adaptive sensing, unstructured sparsity) a Lasso-based strategy that, from m non-adaptive Gaussian random measurements, first obtains an estimate $\widehat{\mathbf{x}}_{\text{Lasso}}$

according to

$$\widehat{\mathbf{x}}_{\text{Lasso},\lambda} = \arg \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}_m - \mathbf{A}_m\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{4.46}$$

for a constant $\lambda > 0$, then forms a corresponding support estimate according to $\widehat{\mathcal{S}}_{\text{Lasso},\lambda} = \mathcal{S}(\widehat{\mathbf{x}}_{\text{Lasso},\lambda})$, and

- (non-adaptive sensing, tree sparsity) a Group Lasso-based approach that first identifies an estimate $\widehat{\mathbf{x}}_{\text{GLasso}}$ according to

$$\widehat{\mathbf{x}}_{\text{GLasso},\lambda} = \arg \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}_m - \mathbf{A}_m\mathbf{x}\|_2^2 + \lambda \sum_{G\in\mathcal{G}} \|\mathbf{x}_G\|_2, \tag{4.47}$$

for a constant $\lambda > 0$, where $\mathbf{x}_G$ denotes the sub vector of $\mathbf{x}$ indexed by elements in the set $G \subseteq \{1, 2, \ldots, n\}$ and $\mathcal{G}$ is the (pre-specified) set of hierarchically overlapping groups which enforce tree structure (see, e.g., [156, 157]), then forms a support estimate according to $\widehat{\mathcal{S}}_{\text{GLasso},\lambda} = \mathcal{S}(\widehat{\mathbf{x}}_{\text{GLasso},\lambda})$.

The adaptive sensing strategies we evaluate are

- (adaptive sensing, unstructured sparsity) the near-optimal adaptive compressive sensing strategy proposed and analyzed in [158], and
- (adaptive sensing, structured sparsity) the repeated-measurement variant of the adaptive tree sensing approach in Algorithm 4 above.

We consider overall three different scenarios, corresponding to three different values of the problem dimension ($n = 2^8 - 1$, $n = 2^{10} - 1$, and $n = 2^{12} - 1$, chosen so that the underlying trees in each case are complete), and in each case we evaluate the performance of each approach over a range of signal amplitude parameters μ, as follows. In each of 100 trials we first generate a random n-dimensional tree-sparse signal with $k = 16$ nonzero components of amplitude μ. We construct the signals here so that all nonzero components are non-negative, for simplicity, and to facilitate direct comparison with the procedure analyzed in [158]. We fix $m = 4(2k + 1)$ and apply each of the procedures described above (with additive noise variance $\sigma^2 = 1$), and assess whether it correctly identifies the true support by comparing the support estimate obtained by the procedure with the true support of the tree signal. The final empirical probabilities of support recovery error for each approach (and each fixed n and μ) were obtained by averaging results over the 100 trials.

For completeness, we mention a few additional details regarding our implementations here. First, for the Lasso-based approaches based on Gaussian measurements, a new independent measurement ensemble was generated to obtain measurements in each trial, but the same measurement vectors and corresponding measurements are used for both of the approaches in a given trial. Further, since each of the Lasso-based approaches relies on specification of a tuning (regularization) parameter λ, when evaluating the performance of those approaches we swept over the range of allowable λ values, obtaining for each a support estimate as above, and we declare the approach a success if the correct support estimate is identified for *any* value of λ. We also note that due to implementation and machine precision issues, the estimates $\widehat{\mathbf{x}}$ obtained by the Lasso-based estimation strategies may not be exactly sparse; in the experiments we obtained support estimates for each of the Lasso-based estimators by identifying the sets of locations where the corresponding reconstructed signal component amplitudes exceeded $\mu/3$. The Lasso-based procedures were implemented here using the Sparse Modeling Software (SpaMS), available online at `spams-devel.gforge.inria.fr`.

Our choice of m corresponds to $r = 4$ in the repeated-measurement variant of the tree-sensing procedure of Algorithm 4. The threshold for this approach was obtained according to (4.5) using $\delta = 0.01$ and $\beta = 1$. Note that this choice of β corresponds to an instance where the true underlying sparsity level is known prior to implementing the procedure; we afford the procedure of [158] the same prior knowledge of sparsity level. Further, strictly speaking, the approach in [158] does not fit the unit-norm measurement model of (4.1), but instead imposes a global constraint on the measurement ensemble of the form $\sum_j \|\mathbf{a}_j\|_2^2 \leq m$. In this more general interpretation, the parameter m may be viewed not as the number of measurements per se, but instead as a "sensing energy" budget. Nevertheless, we note that in implementation, each measurement prescribed by the method in [158] could be synthesized either using a *collection* of measurements obtained using measurement vectors that satisfy $\|\mathbf{a}_j\|_2^2 = \epsilon$ for all j and some (small) $\epsilon > 0$, or equivalently, by appropriately adjusting the effective noise variance per measurement. We used the latter approach here when implementing the method in [158], along with one additional modification. Namely, we note that the procedure in [158]

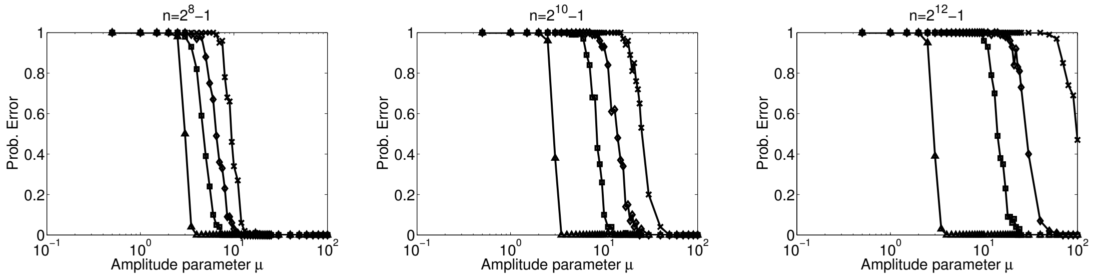

Figure 4.2: Empirical probability of support recovery error as a function of signal amplitude parameter μ in three different problem dimensions n. In each case, four different sensing and support recovery approaches – the adaptive tree sensing procedure described here ($\triangle$ markers); the adaptive compressive sensing approach of [158] ($\Box$ markers); a Group Lasso approach for recovering tree-sparse vectors ($\diamond$ markers), and a Lasso approach for recovering unstructured sparse signals ($\times$ markers) – were employed to recover the support of a tree-sparse signal with 16 nonzeros of amplitude μ. The proposed tree-sensing procedure outperforms each of the other methods, and exhibits performance that is unchanged as the problem dimension increases.

may not satisfy the sensing energy constraint with equality, in effect leaving some sensing energy unused, which may lead to sacrificed performance. Here, we account for this by explicitly rescaling (increasing) the energy allocations at each step so that the overall sensing energy constraint is satisfied with equality.

Figure 4.2 depicts the results of this experimental comparison, where for each value of n and for each method we plot the empirical probability of support recovery error, averaged over the 100 trials, as a function of the (logarithm of the) signal amplitude parameter μ. Here, the tree-sensing procedure corresponds to the curve with triangle ($\triangle$) markers, the adaptive CS approach of [158] is shown with square ($\Box$) markers, the Group Lasso approach is shown with diamond ($\diamond$) markers, and the Lasso approach is shown with the $\times$ markers.

A few interesting points are worth noting here. First, as expected, the adaptive tree-sensing procedure outperforms each of the other approaches in each of the three scenarios. Indeed, the performance of the four approaches follows a fairly intuitive ordering – the tree-sensing procedure performs best, followed by the adaptive sensing strategy of [158], then the Group-Lasso based approach that exploits tree-structure in the inference task (but uses non-adaptive sensing), and finally, the Lasso-based approach that uses non-adaptive sensing, and does not exploit tree structure. Overall, the results suggest that either utilizing adaptive sensing or exploiting tree structure (alone) can indeed result in techniques that outperform traditional CS, but even more significant improvements are possible when leveraging adaptivity and structure together, confirming our claim in the discussion in Section 4.1.

Further, it is interesting to note that the performance of the tree-sensing procedure is *unchanged* as the problem dimension increases, in agreement with the result of the result of Corollary 4.1.1, where the sufficient condition on μ that ensures accurate support recovery does not depend on the ambient dimension n. By comparison, the performance of each of the other approaches degrades as the problem dimension increases – a "curse of dimensionality" suffered by each of these other techniques. While our experimental results only compare problems across 4 orders of magnitudes, the relative performance differences will become much more significant here as the problem dimension becomes even larger[6] .

For completeness, we note that the the result of Corollary 4.1.1, with the specific parameter choices utilized in our experimental setup, ensures that accurate support recovery (with probability at least $1 - \delta = 0.99$ here) occurs when $\mu \geq 6.2$. Here, we observe that accurate support recovery occurs for the tree-sensing procedure for slightly weaker signals whose component amplitudes μ satisfy $\mu \geq 3.5$. Of course, the condition identified in Corollary 4.1.1 is only a *sufficient* condition, and as stated in the discussion in Section 4.1, we made little effort to optimize the constants associated with the sufficient conditions here, opting instead for results of a simple functional form. Nevertheless, even with the bounding we employed in our proof, the conditions we identified are fairly representative of the behavior of the procedure in practice.

Our second experimental evaluation is designed to investigate the scaling behavior predicted by the theoretical guarantees we provide in Corollary 4.1.1 – namely, that accurate support estimation is achievable provided the nonzero signal components satisfy the condition given in (4.7). To that end, we provide a *phase transition* plot for our approach that depicts, for a measurement budget $m_{\max}$, whether the tree sensing procedure results in accurate support recovery of k tree-sparse signals whose nonzero

[6] We chose these representative problem sizes here, in part, because of computational limitations associated with implementing the Lasso-based experiments on larger problem sizes. By comparison, our tree-sensing procedure executes in under 1 second in `MATLAB` on our desktop system, even for problem sizes where $n \sim 2^{27}$.

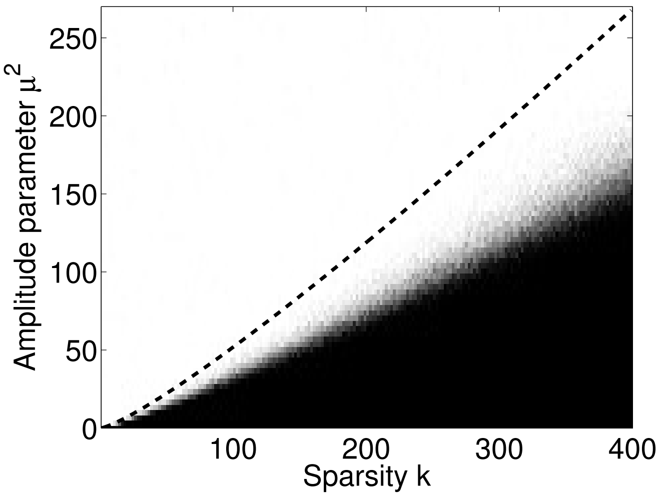

Figure 4.3: Empirical probability of successful support recovery for the tree-sensing procedure of Algorithm 4 as a function of signal sparsity k and squared signal component amplitude μ^2, for a fixed measurement budget. Here, the light and dark regions correspond to settings where the empirical probability of correct support recovery (averaged over 100 trials) is nearly one or nearly zero, respectively. The dashed line corresponds to the threshold above which our theoretical results guarantee correct support recovery with probability at least 0.99. The empirical results here appear to validate our theoretical predictions for this scenario (see text for specific simulation details).

amplitudes each have amplitude μ (for varying parameters k and μ). More specifically, for this experiment we fix the signal dimension to be $n = 2^{16} - 1$ and the noise variance $\sigma^2 = 1$, we choose $m_{\max} = 1000$. Then, for each choice of the pair (k, μ) chosen from a discretization of the space $\{1, 2, \ldots, 400\} \times [0, 17]$ we implement 100 trials of the following experiment: generate a random k-tree sparse signal having nonzero components with amplitude μ, and implement the tree-sensing strategy described in Corollary 4.1.1 with threshold τ as in (4.5), where $r = \lfloor m_{\max}/(2k+1) \rfloor$, $k' = k$, and $\delta = 0.01$. We then record, for each choice of sparsity and amplitude parameter, how many of the trials resulted in successful support recovery.

The results in Figure 4.3 depict, for a range of sparsity and amplitude parameters, the fraction of the trials in which the support was exactly identified. Here, the black regions correspond to the value 0 and white regions to the value 1; in words, the dark regions of the plot depict regimes where most or all of the trials failed to successfully identify the true support, the white regions depict regions where the support was accurately identified in a large fraction of the trials, and the grey regions depict the "transition" region, where the fraction of trials in which the support was accurately identified is between 0 and 1.

We note that, given our choice of $\delta = 0.01$, we expect that the probability of successful support recovery for the tree-sensing procedure should be at least 0.99 provided the condition (4.7) is satisfied. For comparison, we plot this critical value of signal amplitude corresponding to when the condition (4.7) is satisfied with equality for our parameter choices outlined above (which imply, in particular, that $\beta = 1$) in Figure 4.3 as a dashed line. From this, we surmise that the (k, μ^2) pairs depicted by points above the dashed line do indeed correspond to regions where nearly all of the trials resulted in successful support recovery, though as alluded in the discussion of our first experimental evaluation above, the sufficient condition of (4.7) may be a bit conservative.

The results of Figure 4.3 allow us to make one additional comparison with the behavior identified by the sufficient condition (4.7). Namely, note that for fixed m, β, and σ^2, we expect from (4.7) that the minimum signal amplitude μ above which exact support recovery is achieved (with high probability) by the tree sensing procedure should increase in proportion to $\sqrt{k \log k}$. Now, the results of Figure 4.3 depict success probability as a function of the *square* of the signal amplitude parameter μ^2 and sparsity level k, so in this case, we expect that the line in the μ^2 vs k plane above which successful support recovery occurs with probability at least 0.99 should be functionally proportional to $k \log k$. This appears to be the case here – it looks (at least visually, and for this experimental evaluation) like the line demarcating the transition from the black region to the white region does indeed grow super linearly in k, providing some additional (visual) validation of the results of Corollary 4.1.1.

4.4 Discussion and Conclusions

In this section we conclude with a few final thoughts including, in particular, some comments on the philosophical difference between the tree-sensing strategy of Algorithm 4 and many existing adaptive sensing strategies, as well as a discussion of the implications of our results here for the task of signal estimation.

4.4.1 Adaptive Sensing Strategies for Structured Sparsity

It is interesting to note that, to date, binary-search-based sensing strategies have been the essential idea behind most of the adaptive sensing procedures that have been proposed and analyzed for sparse recovery tasks in prior efforts including, for example, the aforementioned compressive binary search efforts [117, 122, 123, 155, 159]; the *distilled sensing* strategy of [113] (whose analysis provided the first performance guarantees for adaptive sensing strategies in sparse inference tasks) and its compressive sensing variants [114, 120]; and the sequential thresholding technique in [118, 145]. The essential functionality of these strategies amounts to "sequential rejection," in the sense that measurements (either compressive, or "uncompressed") are initially obtained over all signal locations, and then focused in subsequent steps onto groups or sets of locations of decreasing size, in an attempt to hone in on the true signal components.

On the other hand, we note that the tree-sensing strategy in Algorithm 4 is fundamentally different, in that it is a *constructive* approach. Indeed, the essential idea behind the procedure of Algorithm 4 is to *construct*, in subsequent steps, an subspace of increasing dimension that well-approximates the signal being acquired. This seemingly subtle difference turns out to be extremely powerful: when using the constructive approach in an adaptive sensing strategy, measurements can be focused locally onto the subspace where the signal exists essentially from the start of the procedure; in other words, no "global" measurements need be obtained. In contrast, the binary-search-based strategies necessarily must allocate measurement resources more broadly at the outset, and then only gradually focus onto the signal subspace as it becomes clear via rejection of (enough of) the subspaces or dimensions where the signal is unlikely to reside.

This fundamental difference has profound implications in the signal recovery task, especially for very high dimensional problems. Namely, we saw here that the support of k-tree sparse signals can be accurately identified provided their nonzero component amplitudes exceed a constant times $\sqrt{\sigma^2(k/m)\log k}$. This suggests that the problem becomes more "difficult" as the sparsity level k increases (as expected) but, notably, the performance is *independent of the ambient signal dimension n.* This was verified in the experimental evaluation in Section 4.3. On the other hand sensing and estimation strategies based on compressive binary search ideas and (for example) the general cluster sparse structure investigated in [123] require component amplitudes that are at least as large as a constant times $\sqrt{\sigma^2(n/m)\log\log n}$, and further, no sensing and estimation procedure for that type of cluster structure will provide uniform recovery guarantees for signals whose component amplitudes are smaller than a constant times $\sqrt{\sigma^2(n/m)}$. In other words, both the necessary and sufficient conditions for recovery of signals exhibiting the form of cluster sparsity studied in [123] grow with the ambient dimension n, implying that the support recovery task for those problems becomes inherently more difficult as the signal dimension n increases, even if the signal sparsity remains fixed.

That said, it is worth noting a key difference between the tree-sparse signal models we consider here, and the block- and graph-structured models analyzed in [122, 123], that gives rise to this distinction. Namely, in the settings we examine here we enjoy the benefit of a priori "partial localization" information, in the sense that we know at the outset one index (corresponding to the root node of the underlying tree) at which the unknown signal has a non-zero component. This knowledge, along with the strong spatial regularity imposed by the tree structure, is what enables us to accurately localize tree-sparse signals whose component strengths are independent of dimension. More generally, it is quite likely that the approaches in [122, 123], if equipped with analogous partial localization information, could also enjoy the dimension-independent localization thresholds we identify here for the tree-sparse signal models. Overall, we note that the tree-sparse model we consider here comprises but one useful form of structured sparsity for which the necessary and sufficient conditions for recovery do not suffer an inherent "curse of dimensionality;" full characterization of other forms of structured sparsity that exhibit this favorable characteristic is a fruitful path for future investigations.

4.4.2 Implications for Signal Estimation

Finally, we note that in some sparse inference tasks it may be more beneficial to assess performance in terms of achievable mean-square error (MSE), rather than by probability of accurate support identification, as here. While our focus here was specifically on the support recovery task, we conclude with a brief discussion of our results in the context of these estimation tasks.

Several recent efforts have quantified fundamental lower bounds on the achievable

MSE when estimating unstructured k-sparse signals using (adaptive, or non-adaptive) measurements obtained according to the model (4.1). Specifically, [160] established that when estimating unstructured k-sparse signals $\mathbf{x} \in \mathbb{R}^n$ using any estimator $\widehat{\mathbf{x}}$ based on any non-adaptive sensing strategies $M \in \mathcal{M}_{\text{na}}$ for which the ensemble of measurement vectors satisfies the norm constraint

$$\|\mathbf{A}_m\|_F^2 \leq m, \tag{4.48}$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm, the minimax MSE satisfies the bound

$$\inf_{\widehat{\mathbf{x}}, M \in \mathcal{M}_{\text{na}}} \sup_{\mathbf{x}:|\mathcal{S}(\mathbf{x})|=k} \mathbb{E}\left[\|\widehat{\mathbf{x}}(\mathbf{A}_m, \mathbf{y}_m; M) - \mathbf{x}\|_2^2\right] \geq c\,\sigma^2 \left(\frac{n}{m}\right) k \log n, \tag{4.49}$$

for a specified constant $c > 0$. This result established that noisy CS estimation strategies, such as the Dantzig selector [92] are essentially optimal, in the sense that there exist measurement ensembles satisfying (4.48) such that for any k-sparse signal $\mathbf{x} \in \mathbb{R}^n$, the Dantzig selector produces from $m = O(k \log n)$ measurements an estimate $\widehat{\mathbf{x}}_{\text{DS}}$ satisfying $\|\widehat{\mathbf{x}}_{\text{DS}} - \mathbf{x}\|_2^2 = O\left(\sigma^2 \left(\frac{n}{m}\right) k \log n\right)$ with high probability. The works [119] and [124] considered adaptive sensing strategies $M \in \mathcal{M}_{\text{ad}}$ satisfying norm constraints analogous to (4.48) in the context of estimating unstructured sparse signals, and showed that in this case the minimax MSE satisfies

$$\inf_{\widehat{\mathbf{x}}, M \in \mathcal{M}_{\text{ad}}} \sup_{\mathbf{x}:|\mathcal{S}(\mathbf{x})|=k} \mathbb{E}\left[\|\widehat{\mathbf{x}}(\mathbf{A}_m, \mathbf{y}_m; M) - \mathbf{x}\|_2^2\right] \geq c''\,\sigma^2 \left(\frac{n}{m}\right) k, \tag{4.50}$$

where $c'' > 0$ is another constant. Overall, the improvement that can be achieved using adaptivity when estimating unstructured sparse signals amounts to at most a constant times a logarithmic factor.

On the other hand, a simple consequence of our support recovery result implies that adaptive sensing strategies for structured sparse signals can result in significant improvements in estimation MSE, as well. Note that any accurate sparse support estimation procedure based on compressive measurements can easily be parlayed into an estimation procedure by first identifying the locations of the nonzero elements, and then in a second step, collecting direct measurements of the nonzero components (this point was also noted in [119]). Applying this idea using the adaptive tree sensing strategy described above for the support estimation task, we can establish a result of the following form.

Lemma 4.4.1. *There exists a two-stage (support recovery, followed by direct measurement) adaptive compressed sensing procedure for k-tree sparse signals that produces, from $m = O(k)$ measurements, an estimate $\widehat{\mathbf{x}}$ satisfying*

$$\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 = O\left(\sigma^2 \left(\frac{k}{m}\right) k\right) \tag{4.51}$$

with high probability, provided the nonzero signal component amplitudes exceed a constant times $\sqrt{\sigma^2 \left(\frac{k}{m}\right) \log k}$ in amplitude.

We omit the proof. Note that if this type of approach were used to acquire (and estimate) tree-sparse signals whose nonzero components have equal amplitudes, it follows that the estimate $\widehat{\mathbf{x}}$ produced would satisfy

$$\mathbb{E}\left[\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2\right] = O\left(\sigma^2 \left(\frac{k}{m}\right) k \log k\right). \tag{4.52}$$

Somewhat astonishingly, this bound is (up to constants) within a logarithmic factor of the estimation error that would be produced were an oracle to provide the *exact* locations of the nonzero components *before any measurements were obtained*! This argument suggests that accurate estimation approaches (based on adaptive sensing strategies) can be constructed for recovering a broad class of relatively weak tree-sparse signals (i.e., signals having very small individual component amplitudes), and that these strategies could be capable of producing estimators whose errors are on the order of those incurred by oracle-aided sensing strategies. We defer a more thorough investigation of the performance limits for this tree-sparse signal estimation task to a later effort.

4.5 Appendix

We first establish a few useful intermediate results that will be used in the proof of Lemma 4.1.1 as well as in the proofs of our main theorems. Recall from the discussion in Section 4.2 that for any $T \in \mathcal{T}_{n,\ell}$, corresponding to a tree with $1 \leq \ell < n$ nodes that is a rooted connected subtree of some underlying nearly complete binary tree of n nodes, we defined the set $N(T)$ to be the set of locations at which one additional node can be added to the tree described by T to yield a tree with $\ell + 1$ nodes that is itself another rooted connected subtree of the same binary tree. The following lemma provides a bound on the sizes of the sets $N(T)$.

Lemma 4.5.1. *For any $T \in \mathcal{T}_{n,k}$ with $k < n$, we have that $|N(T)| \leq k+1$.*

Proof. The proof proceeds by induction on $k \leq n$. The case $k=1$ is a trivial case where $\mathcal{T}_{n,1}$ contains only a single element corresponding to the index of the root node of the underlying tree. Now, since the underlying tree is binary we have that the number of locations at which one node can be added is less than or equal to 2.

Now, for some $k < n$ we assume that $|N(T)| \leq k+1$ for all $T \in \mathcal{T}_{n,k}$; we aim to show that $N(T') \leq (k+1)+1$ for all $T' \in \mathcal{T}_{n,k+1}$. To that end, we note that any $T' \in \mathcal{T}_{n,k+1}$ can be expressed as $T' = T \cup j$ for some $T \in \mathcal{T}_{n,k}$ and $j \in N(T)$. This implies that $N(T')$ contains all elements in the set $N(T) \setminus j$, as well as the children of the index j, of which there are at most two. Thus, it follows that for any $T' \in \mathcal{T}_{n,k+1}$ we have that $|N(T')| \leq (|N(T)|-1)+2 \leq (k+1)-1+2 = (k+1)+1$, which is what we intended to show. □

It is worth noting that results of this flavor are somewhat classical. For example, [161] establishes a related result in a setting where the size of the underlying tree is essentially unconstrained, implying that each node has exactly two children. This corresponds to a special case of the above result, where the number of locations at which one node may be added is *exactly* $k+1$. We opt to provide the above proof for completeness, but also to highlight the difference in the setting where the size of the underlying tree is constrained, which is an essential characteristic of our signal model.

Our second intermediate result identifies settings where the bounds obtained above on the cardinality of the set $N(T)$ hold with equality. For this, we make explicit use of the assumption that the underlying tree be nearly complete; even in this case, the result holds only for signals that are "sparse enough."

Lemma 4.5.2. *For every integer $2 \leq k \leq (n+1)/2$, there exists at least one $T^* \in \mathcal{T}_{n,k-1}$ for which $|N(T^*)| = k$.*

It follows directly from this lemma that there exists a subset $\mathcal{T}^* = \{T^* \cup j : j \in N(T^*)\} \subseteq \mathcal{T}_{n,k}$ of allowable supports for k-tree sparse vectors, for which $|\mathcal{T}^*| = k$, and the symmetric difference $T_i \Delta T_j \triangleq (T_i \cup T_j) \setminus (T_i \cap T_j)$ satisfies $|T_i \Delta T_j| = 2$ for any pair $T_i, T_j \in \mathcal{T}^*$ with $i \neq j$.

Proof. Recall that the underlying tree is assumed to be nearly complete, meaning that all levels of the underlying tree are full with the possible exception of the last level, and that nodes in the last level are as far to the left as possible. Our proof is constructive – for each $2 \leq k \leq (n+1)/2$ we let T^* be the set of indices that corresponds to the $k-1$ nodes in the underlying tree selected in a top-down, left-to-right manner.

First, note that when $k - 1 = 2^q - 1$ for some integer $q \in \mathbb{N}$ the set of indices in T^* will correspond to a *complete* subtree of the underlying nearly complete tree. Further, the underlying tree must contain all $2^q = k$ nodes in the level immediately below the last level filled by the indices in T^*; if not, then the total number of nodes in the tree would satisfy $n < (k-1) + k = 2k - 1$, which contradicts the condition that $k \leq (n+1)/2$. Thus, in this case the $N(T^*)$ can be taken to be the $2^q = k$ descendants of the nodes in the last full level of the subtree described by the indices in T^*.

For other values of k that cannot be written as integer powers of 2, the set T^* will not correspond to a complete subtree, but instead, a nearly complete subtree of the underlying tree. In this case, note that $d^* = d^*(k) = \lfloor \log_2(k-1) \rfloor$ is the depth of the last (partially-filled) level of the subtree corresponding to the $k - 1$ elements in T^*. It follows that the total number of indices in all of the filled layers of the subtree defined by elements of T^* is $2^{d^*} - 1$, and thus, the number of indices in the last partially-filled level is given by $v^* = (k-1) - (2^{d^*} - 1)$. Now, the set $N(T^*)$ can be constructed to contain all of the $2^{d^*} - v^*$ indices in the last partially filled level, plus $2v^*$ indices in the level immediately below that are the descendants of the indices in the last partially filled level. For this construction, note that

$$\begin{aligned} |N(T^*)| &= 2^{d^*} - v^* + 2v^* \\ &= 2^{d^*} + (k-1) - (2^{d^*} - 1) \\ &= k. \end{aligned} \tag{4.53}$$

Finally, for completeness, we note that the level immediately below the last partially filled level of the subtree described by elements of T^* must indeed contain at least $2v^*$ indices. If not, then the total number of indices in the underlying tree would be $n < (k-1)+k = 2k-1$, which again contradicts the requirement that $k \leq (n+1)/2$. □

4.5.1 Proof of Lemma 4.1.1

The proof of Lemma 4.1.1 relies on the fact that when acquiring a particular signal $\mathbf{x}$ having support $\mathcal{S}(\mathbf{x}) \in \mathcal{T}_{n,k}$, the support estimate $\widehat{\mathcal{S}}$ produced when the adaptive sensing procedure of Algorithm 4 terminates will exactly equal the true support when the event

$$\mathcal{E}_{\mathbf{x}} \triangleq \left\{ \bigcap_{i \in \mathcal{S}(\mathbf{x})} |y_{(i)}| \geq \tau \right\} \cap \left\{ \bigcap_{j \in N(\mathcal{S}(\mathbf{x}))} |y_{(j)}| < \tau \right\} \tag{4.54}$$

occurs[7] . More specifically, the event $\mathcal{E}_{\mathbf{x}}$ corresponds to a (valid) instance of the procedure where measurements of $\mathbf{x}$ are obtained at all locations $\ell \in \mathcal{S}(\mathbf{x}) \cup N(\mathcal{S}(\mathbf{x}))$ and the hypothesis test associated with each measurement produces the correct result, thus resulting in a correct final support estimate.

In other words, the above discussion establishes that $\mathcal{E}_{\mathbf{x}} \subseteq \{\widehat{\mathcal{S}} = \mathcal{S}(\mathbf{x})\}$. (Actually, the events $\mathcal{E}_{\mathbf{x}}$ and $\{\widehat{\mathcal{S}} = \mathcal{S}(\mathbf{x})\}$ can be shown to be equal, though we don't explicitly need this fact for our proof.) This implies, in turn, that $\{\widehat{\mathcal{S}} = \mathcal{S}(\mathbf{x})\}^c \subseteq \mathcal{E}_{\mathbf{x}}^c$; thus,

$$\begin{aligned}
\Pr_{\mathbf{x}}\left(\widehat{\mathcal{S}} \neq \mathcal{S}(\mathbf{x})\right) &\leq \Pr_{\mathbf{x}}\left(\left\{ \bigcup_{i \in \mathcal{S}(\mathbf{x})} |y_{(i)}| < \tau \right\} \cup \left\{ \bigcup_{j \in N(\mathcal{S}(\mathbf{x}))} |y_{(j)}| \geq \tau \right\} \right) \\
&\leq \sum_{i \in \mathcal{S}(\mathbf{x})} \Pr_{\mathbf{x}}\left(|y_{(i)}| < \tau\right) + \sum_{j \in N(\mathcal{S}(\mathbf{x}))} \Pr_{\mathbf{x}}\left(|y_{(j)}| \geq \tau\right). \qquad (4.55)
\end{aligned}$$

The proof proceeds by identifying simple upper bounds for each term in the sum on the right hand side of (4.55). To that end, note that for $j \in N(\mathcal{S}(\mathbf{x}))$ we have that $y_{(j)} \sim \mathcal{N}(0, \sigma^2)$. Thus,

$$\begin{aligned}
\Pr_{\mathbf{x}}\left(|y_{(j)}| \geq \tau\right) &= \Pr_{\mathbf{x}}\left(\{y_{(j)} \geq \tau\} \cup \{y_{(j)} \leq -\tau\}\right) \\
&= 2 \cdot \Pr_{\mathbf{x}}\left(y_{(j)} \geq \tau\right) \\
&\leq \exp\left(-\frac{\tau^2}{2\sigma^2}\right), \qquad (4.56)
\end{aligned}$$

where the second line follows by symmetry and the fact that the events are disjoint, and the third line utilizes a standard bound on the tail of the Gaussian distribution.

We now consider obtaining bounds on the probabilities of the events $\{|y_{(i)}| < \tau\}$ for $i \in \mathcal{S}(\mathbf{x})$. Note that for $i \in \mathcal{S}(\mathbf{x})$ we have $y_{(i)} = \alpha_i + w$ where $w \sim \mathcal{N}(0, \sigma^2)$ represents

[7] Note that in our proof we adopt the alternative notation used in our description of the adaptive sensing procedure, where measurements are indexed according to the location that was observed.

the additive noise for that observation. Since we placed no condition on the signs of the nonzero elements of $\mathbf{x}$ we ultimately have to consider two cases to establish our bound. Consider, first, the case where the nonzero element at location i satisfies $\alpha_i > 0$. We have

$$\{|y_{(i)}| < \tau\} = \{-\tau - \alpha_i < w < \tau - \alpha_i\} \subset \{w < \tau - \alpha_i\}, \tag{4.57}$$

implying that $\Pr_{\mathbf{x}}\left(\{|y_{(i)}| < \tau\}\right) \leq \Pr_{\mathbf{x}}\left(w < \tau - \alpha_i\right)$. Now, for $\tau < \mu \leq \alpha_i$ we can again employ a standard bound on the tail of the Gaussian distribution to claim

$$\Pr_{\mathbf{x}}\left(\{|y_{(i)}| < \tau\}\right) \leq \exp\left(-\frac{(\alpha_i - \tau)^2}{2\sigma^2}\right). \tag{4.58}$$

Using a similar approach for the case $\alpha_i < 0$ and the same τ, we obtain (after some straightforward computations) that the overall the bound

$$\begin{aligned} \Pr_{\mathbf{x}}\left(\{|y_{(i)}| < \tau\}\right) &\leq \exp\left(-\frac{(|\alpha_i| - \tau)^2}{2\sigma^2}\right) \\ &\leq \exp\left(-\frac{(\mu - \tau)^2}{2\sigma^2}\right) \end{aligned} \tag{4.59}$$

holds for any $i \in \mathcal{S}(\mathbf{x})$, where the last step follows from the fact that $|\alpha_i| \geq \mu$ for all $i \in \mathcal{S}(\mathbf{x})$. Thus

$$\Pr_{\mathbf{x}}\left(\widehat{\mathcal{S}} \neq \mathcal{S}(\mathbf{x})\right) \leq k \exp\left(-\frac{(\mu - \tau)^2}{2\sigma^2}\right) + (k+1)\exp\left(-\frac{\tau^2}{2\sigma^2}\right). \tag{4.60}$$

Note that the leading factor of $k+1$ in the second term of (4.60) is a consequence of Lemma 4.5.1.

The last step of the proof is straightforward, and amounts to showing that for any $\delta \in (0,1)$, when

$$\mu = \sqrt{8\log(4/\delta)} \cdot \sqrt{\sigma^2 \log k} \tag{4.61}$$

and

$$\tau = \sqrt{2\sigma^2 \log(4k/\delta)}, \tag{4.62}$$

each of the two terms in the sum in (4.60) can be upper bounded by $\delta/2$. Further, it is easy to verify that for these choices $\tau < \mu$ (as required by our proof) whenever $k > 1$. These steps are straightforward, so we omit the details.

4.6 Acknowledgements

The authors are grateful to the anonymous reviewers for their detailed and thorough evaluations, and in particular, for pointing out some subtle errors in the initial versions of our proofs of our main results, and for helping to clarify the potential implications of the results of [160] for achievable minimax MSE for estimation of signals exhibiting structured sparsity.

Chapter 5

Directions for Future Study

5.1 Matrix Completion

Lower Bounds for Matrix Completion with Sparse Factor Models:

The upper bounds for the per-element squared error for the matrix completion problems with sparse factor model established in Chapters 2 and 3, depends on the ratio of number of degrees of freedom and number of measurements – there we used penalized maximum likelihood formulation. It would be interesting to see if our problem formulation (and therefore error bounds) is optimal. To be precise, we would like to know if there exists any other formulation whose corresponding error bounds would be drastically better than our results.

In order to claim optimality of our procedure, it would be of interest to establish fundamental lower bounds on the per-element squared error and show that it also has the (almost) same dependence on number of degrees of freedom and number of measurements. If so, no other formulation would be able to perform fundamentally better than the proposed formulations.

Analysis of Matrix Completion Algorithms with Sparse Factor Models:

Our ADMM based optimization procedure for matrix completion with sparse factor models as explained in Section 3.4 requires further theoretical foundation and convergence analysis. Each step of our procedure is separable in matrix elements and is computationally efficient, and theoretical understanding of it would certainly vouch for its wide acceptability. While there is no clear understanding of ADMM applied to non-convex problem, our structural assumptions may still help in analysis.

Apart from ADMM based algorithms, *alternating minimization* has recently been shown to converge (to global optimal if properly initialized) for dictionary learning [162] and low-rank matrix completion problems [163]. These works provide guarantees for squared loss model (under Gaussian noise assumptions). In order to extend these results to Sparse Factor model under different likelihoods, the analyses needs to be done for each of the considered likelihood models.

Bounds for ℓ_1-norm based Matrix Completion with Sparse Factor Models:

Our penalized maximum likelihood formulation and penalty construction entails to a ℓ_0-norm regularized optimization problem. An interesting direction would be to construct penalty which gives us ℓ_1-norm regularized problems. Specifically, it would correspond to coding strategies which are signal amplitude dependent. It would be of interest to know if by going from a non-convex ℓ_0-norm to its convex relaxation ℓ_1-norm, we make the error bounds loose or not.

If no such penalty can be constructed, there is a need of alternative analyses techniques which provides error bounds for ℓ_1 regularized maximum likelihood problems – for e.g., packing set based arguments and PAC-Bayesian analyses.

5.2 Tree-sparse Signal Estimation

Lower Bounds for Tree-sparse Signal Estimation:

Several recent efforts have quantified fundamental lower bounds on the achievable MSE when estimating unstructured k-sparse signals using (adaptive, or non-adaptive) measurements obtained according to the model (4.1). Specifically, [160] established the fundamental limits for estimating unstructured k-sparse signals $\mathbf{x} \in \mathbb{R}^n$ using any estimator $\widehat{\mathbf{x}}$ based on any non-adaptive sensing strategies. This result established that noisy CS estimation strategies, such as the Dantzig selector [92] are essentially optimal.

The works [119] and [124] considered adaptive sensing strategies satisfying norm constraints analogous to (4.48) in the context of estimating unstructured sparse signals, and provided the minimax bounds. Overall, the improvement that can be achieved using adaptivity when estimating unstructured sparse signals amounts to at most a constant times a logarithmic factor.

On the other hand, a simple consequence of our support recovery result implies that adaptive sensing strategies for structured sparse signals can result in significant improvements in estimation MSE, as well, see Lemma 4.4.1. It would be of interest to establish minimax bounds for tree-sparse signal estimation with both adaptive and non-adaptive sensing schemes.

References

[1] L. P. Panych and F. A. Jolesz. A dynamically adaptive imaging algorithm for wavelet-encoded MRI. *Magnetic Resonance in Medicine*, 32(6):738–748, 1994.

[2] M. W. Seeger and H. Nickisch. Compressed sensing and Bayesian experimental design. In *Proc. ICML*, pages 912–919, 2008.

[3] S. Deutsch, A. Averbuch, and S. Dekel. Adaptive compressed image sensing based on wavelet modeling and direct sampling. In *Proc. Intl. Conf on Sampling Theory and Applications*, 2009.

[4] A. Averbuch, S. Dekel, and S. Deutsch. Adaptive compressed image sensing using dictionaries. *SIAM J Imaging Sciences*, 5(1):57–89, 2012.

[5] A. Soni and J. Haupt. Estimation error guarantees for Poisson denoising with sparse and structured dictionary models. In *Proc. International Symposium on Information Theory*, 2014.

[6] R. Giryes and M. Elad. Sparsity based Poisson denoising with dictionary learning. *arXiv preprint arXiv:1309.4306*, 2013.

[7] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett. Poisson noise reduction with non-local PCA. *Journal of Mathematical Imaging and Vision*, 48(2):279–294, 2014.

[8] A. Soni, S. Jain, J. Haupt, and S. Gonella. Noisy matrix completion under sparse factor models. In *IEEE Transactions on Information Theory (submitted)*, 2014.

[9] J. A. Gubner. *Probability and random processes for electrical and computer engineers.* Cambridge University Press, 2006.

[10] S. M. Kay. *Fundamentals of Statistical signal processing, Volume 1: Estimation Theory.* Prentice Hall PTR, 1993.

[11] R. D. Nowak and R. G. Baraniuk. Wavelet-domain filtering for photon imaging systems. *IEEE Trans. Image Processing*, 8(5):666–678, 1999.

[12] K. E. Timmermann and R. D. Nowak. Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Trans. Information Theory*, 45(3):846–862, 1999.

[13] E. D. Kolaczyk and R. D. Nowak. Multiscale likelihood analysis and complexity penalized estimation. *Ann. Statist.*, pages 500–527, 2004.

[14] M. Raginsky, R. Willett, Z. T. Harmany, and R. F. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010.

[15] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624, 2001.

[16] A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases*, pages 358–373. 2008.

[17] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[18] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Proc.*, 54(11):4311–4322, 2006.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proc. ICML*, 2009.

[20] P. Chainais. Towards dictionary learning from images with non Gaussian noise. In *IEEE Intl. Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.

[21] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, pages 561–576. Springer, 1991.

[22] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Information Theory*, 37(4):1034–1054, 1991.

[23] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[24] Q. J. Li and A. R. Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems*, volume 12, pages 279–285, 2000.

[25] T. Zhang. On the convergence of MDL density estimation. In *Learning Theory*, pages 315–330. Springer, 2004.

[26] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

[27] F.-X. Dupé, J. M. Fadili, and J.-L. Starck. A proximal iteration for deconvolving Poisson noisy images using sparse representations. *IEEE Trans. Image Processing*, 18(2):310–321, 2009.

[28] M. A. T. Figueiredo and J. M. Bioucas-Dias. Restoration of Poissonian images using alternating direction optimization. *IEEE Trans. Image Processing*, 19(12):3133–3145, 2010.

[29] S. Setzer, G. Steidl, and T. Teuber. Deblurring Poissonian images by split Bregman techniques. *Journal of Visual Communication and Image Representation*, 21(3):193–199, 2010.

[30] M. Carlavan and L. Blanc-Féraud. Sparse Poisson noisy image deblurring. *IEEE Trans. Image Processing*, 21(4):1834–1846, 2012.

[31] Z. T. Harmany, R. F. Marcia, and R. M. Willett. This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms – theory and practice. *IEEE Trans. Image Processing*, 21(3):1084–1096, 2012.

[32] L. Ma, L. Moisan, J. Yu, and T. Zeng. A dictionary learning approach for Poisson image deblurring. *IEEE Trans. Medical Imaging*, 32(7):1277–1289, 2013.

[33] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[34] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[35] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

[36] Q. J. Li. *Estimation of mixture models*. PhD thesis, Yale University, Dept. of Statistics, 1999.

[37] Y. Vardi. Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91(433):365–377, 1996.

[38] R. M. Bell and Y. Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[39] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[40] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[41] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[42] B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[43] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[44] W. Dai and O. Milenkovic. SET: An algorithm for consistent matrix completion. In *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, pages 3646–3649, 2010.

[45] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

[46] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008.

[47] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. *arXiv preprint arXiv:1303.5685*, 2013.

[48] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *ACM SIGMOD Int. Conf. Management of Data*, pages 94–105, 1998.

[49] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

[50] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

[51] M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[52] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[53] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès. Robust subspace clustering. *arXiv preprint arXiv:1301.2603*, 2013.

[54] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

[55] K. Lee and Y. Bresler. ADMiRA: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

[56] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[57] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[58] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[59] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems*, pages 1321–1328, 2004.

[60] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2012.

[61] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[62] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[63] R. M. Willett and R. D. Nowak. Multiscale Poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007.

[64] J. D. Haupt, N. D. Sidiropoulos, and G. B. Giannakis. Sparse dictionary learning from 1-bit data. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2014.

[65] A. Rakotomamonjy. Applying alternating direction method of multipliers for constrained dictionary learning. *Neurocomputing*, 106:126–136, 2013.

[66] Z. Xing, M. Zhou, A. Castrodad, G. Sapiro, and L. Carin. Dictionary learning for noisy and incomplete hyperspectral images. *SIAM Journal on Imaging Sciences*, 5(1):33–56, 2012.

[67] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin. Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Transactions on Image Processing*, 21(1):130–144, 2012.

[68] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *arXiv preprint arXiv:1410.0342*, 2014.

[69] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, 2003.

[70] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proc. ACM Symp. on Theory of Computing*, pages 145–162, 2012.

[71] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin. A convex model for nonnegative matrix factorization and dimensionality reduction on physical space. *IEEE Transactions on Image Processing*, 21(7):3239–3252, 2012.

[72] B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pages 1214–1222, 2012.

[73] M. Aharon, M. Elad, and A. M. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1):48–67, 2006.

[74] R. Gribonval and K. Schnass. Dictionary identification – Sparse matrix-factorization via $l_1$minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

[75] Q. Geng, H. Wang, and J. Wright. On the local correctness of ℓ^1 minimization for dictionary learning. *Submitted*, 2011. online at: `arxiv.org/abs/1101.5672`.

[76] D. A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3087–3090, 2013.

[77] K. Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Submitted*, 2013. online at: `arxiv.org/abs/1301.3375`.

[78] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013.

[79] R. Jenatton, R. Gribonval, and F. Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. *Submitted*, 2012. online at: `arxiv.org/abs/1210.0685`.

[80] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[81] B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, pages 915–922, 2005.

[82] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

[83] H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

[84] Y. Zhang, A. dAspremont, and L. El Ghaoui. Sparse PCA: Convex relaxations, algorithms and applications. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 915–940. Springer, 2012.

[85] V. Q. Vu and J. Lei. Minimax rates of estimation for sparse PCA in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1278–1286, 2012.

[86] K. Lounici. Sparse principal component analysis with missing observations. In *High Dimensional Probability VI*, pages 327–356. Springer, 2013.

[87] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the EM algorithm. In *Proc. Computer Vision and Pattern Recognition*, 2004.

[88] R. Vidal, R. Tron, and R. Hartley. Multiframe motion segmentation with missing data using PowerFactorization and GPCA. *International Journal of Computer Vision*, 79(1):85–105, 2008.

[89] B. Eriksson, L. Balzano, and R. Nowak. High-rank matrix completion and subspace clustering with missing data. *arXiv preprint arXiv:1112.5629*, 2011.

[90] A. Singh, A. Krishnamurthy, S. Balakrishnan, and M. Xu. Completion of high-rank ultrametric matrices using selective entries. In *Proc. IEEE International Conference on Signal Processing and Communications*, pages 1–5, 2012.

[91] R. G. Baraniuk, V. Cevher, and M. B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971, 2010.

[92] E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[93] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[94] F. J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3-4):246–254, 1948.

[95] X. Jiang, G. Raskutti, and R. Willett. Minimax optimal rates for Poisson inverse problems with physical constraints. *arXiv preprint arXiv:1403.6532*, 2014.

[96] Z.-Q. Luo. Universal decentralized estimation in a bandwidth constrained sensor network. *IEEE Transactions on Information Theory*, 51(6):2210–2219, 2005.

[97] A. Ribeiro and G. B. Giannakis. Bandwidth-constrained distributed estimation for wireless sensor networks-part i: Gaussian case. *IEEE Transactions on Signal Processing*, 54(3):1131–1143, 2006.

[98] P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman & Hall, 1989.

[99] Z. Lu. Iterative hard thresholding methods for ℓ_0 regularized convex cone programming. *Mathematical Programming*, pages 1–30, 2012.

[100] S. van de Geer. *Empirical Processes in M-estimation*, volume 105. Cambridge University Press, 2000.

[101] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033. Springer, 2011.

[102] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[103] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[104] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[105] E. Richard, G. Obozinski, and J.-P. Vert. Tight convex relaxations for sparse matrix factorization. *arXiv preprint arXiv:1407.5158*, 2014.

[106] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

[107] C. C. Craig. On the Tchebychef inequality of Bernstein. *The Annals of Mathematical Statistics*, 4(2):94–102, 1933.

[108] A. Soni and J. Haupt. On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. *IEEE Transactions on Information Theory*, 60(1):133–149, Jan 2014.

[109] A. Soni and J. Haupt. Efficient adaptive compressive sensing using sparse hierarchical learned dictionaries. In *Proc. Asilomar Conf. on Signals, Systems, and Computers*, pages 1250–1254, 2011.

[110] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Trans Signal Processing*, 56(6):2346–2356, 2008.

[111] R. M. Castro, J. Haupt, R. Nowak, and G. M. Raz. Finding needles in noisy haystacks. In *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing*, pages 5133–5136, 2008.

[112] E. Bashan, R. Raich, and A. O. Hero. Optimal two-stage search for sparse targets using convex criteria. *IEEE Trans Signal Processing*, 56(11):5389–5402, 2008.

[113] J. Haupt, R. M. Castro, and R. Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Trans. Information Theory*, 57(9):6222–6235, 2011.

[114] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Proc. Asilomar Conf. on Signals, Systems, and Computers*, pages 1551–1555, 2009.

[115] G. Newstadt, E. Bashan, and A. O. Hero. Adaptive search for sparse targets with informative priors. In *Proc. IEEE Intl Conf on Acoustics Speech and Signal Processing*, pages 3542–3545, 2010.

[116] E. Bashan, G. Newstadt, and A. O. Hero. Two-stage multiscale search for sparse targets. *IEEE Trans Signal Processing*, 59(5):2331–2341, 2011.

[117] M. Iwen and A. Tewfik. Adaptive group testing strategies for target detection and localization in noisy environments. *IEEE Trans. Signal Proc.*, 60(5):2344–2353, 2012.

[118] M. Malloy and R. Nowak. Sequential testing for sparse recovery. *Submitted*, 2012. online at `arxiv.org/abs/1212.1801`.

[119] E. Arias-Castro, E. J. Candes, and M. Davenport. On the fundamental limits of adaptive sensing. *Submitted*, 2011. online at `arxiv.org/abs/1111.4646`.

[120] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak. Sequentially designed compressed sensing. In *Proc. IEEE Statistical Signal Processing Workshop*, pages 401–404, 2012.

[121] D. Wei and A. O. Hero. Multistage adaptive estimation of sparse signals. *Submitted*, 2012. online at `arxiv.org/abs/1210.1473`.

[122] S. Balakrishnan, M. Kolar, A. Rinaldo, and A. Singh. Recovering block-structured activations using compressive measurements. *Submitted*, 2012. online at `arxiv.org/abs/1209.3431`.

[123] A. Krishnamurthy, J. Sharpnack, and A. Singh. Recovering graph-structured activations using adaptive compressive measurements. *Submitted*, 2013. online at `arxiv.org/abs/1305.0213`.

[124] R. M. Castro. Adaptive sensing performance lower bounds for sparse signal detection and support estimation. *Submitted*, 2012. online at `arxiv.org/abs/1206.0648`.

[125] J. Haupt and R. Nowak. Adaptive sensing for sparse recovery. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and applications*. Cambridge University Press, 2011.

[126] P. Indyk, E. Price, and D. P. Woodruff. On the power of adaptivity in sparse recovery. In *Proc. IEEE Foundations of Computer Science*, pages 285–294, 2011.

[127] E. Price and D. P. Woodruff. Lower bounds for adaptive sparse recovery. *Submitted*, 2012. online at `arxiv.org/abs/1205.3518`.

[128] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proc. Intl. Conf. Machine Learning*, pages 417–424, 2009.

[129] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, 2010.

[130] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. Signal Processing*, 57(8):3075–3085, 2009.

[131] M. F. Duarte and Y. C. Eldar. Structured compressed sensing: From theory to applications. *IEEE Trans Signal Proc*, 59(9):4053–4085, 2011.

[132] N. Rao and R. Nowak. Adaptive sensing with structured sparsity. In *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing*, 2013.

[133] S. Mallat. *A wavelet tour of signal processing: The sparse way.* Academic Press, 2008.

[134] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Processing*, 46(4):886–902, 1998.

[135] J. K. Romberg, H. Choi, and R. G. Baraniuk. Bayesian tree-structured image modeling using wavelet-domain hidden Markov models. *IEEE Trans. Image Processing*, 10(7):1056–1068, 2001.

[136] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk. Fast reconstruction of piecewise smooth signals from incoherent projections. In *Proc. SPARS*, 2005.

[137] C. La and M. N. Do. Signal reconstruction using sparse tree representation. In *Proc. Wavelets XI at SPIE Optics and Photonics*, 2005.

[138] S. Som and P. Schniter. Compressive imaging using approximate message passing and a Markov-tree prior. *IEEE Trans. Signal Processing*, 60(7):3439–3448, 2012.

[139] S. Aeron, V. Saligrama, and M. Zhao. Information theoretic bounds for compressed sensing. *IEEE Trans. Inform. Theory*, 56(10):5111–5130, 2010.

[140] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[141] M. Wainwright. Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12), 2009.

[142] C. Genovese, J. Jin, and L. Wasserman. Revisiting marginal regression. *Manuscript*, 2009. online at `arxiv.org/abs/0911.4080`.

[143] A. K. Fletcher, S. Rangan, and V. K. Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inform. Theory*, 55(12):5758–5772, 2009.

[144] W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inform. Theory*, 56(6):2967–2979, 2010.

[145] M. Malloy and R. Nowak. Sequential analysis in high-dimensional multiple testing and sparse recovery. In *Proc. IEEE Intl. Symp. on Information Theory*, pages 2661–2665, 2011.

[146] G. Reeves and M. Gastpar. The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Trans. Inform. Theory*, 58(5):3065–3092, 2012.

[147] G. Reeves and M. Gastpar. Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Trans. Inform. Theory*, 59(6):3451–3465, 2013.

[148] A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer, 2009.

[149] E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 36(4):1726–1757, 2008.

[150] E. J. Candès Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[151] C. Dossal, M.-L. Chabanol, G. Peyré, and J. Fadili. Sharp support recovery from noisy random measurements by ℓ_1 minimization. *Applied and Computational Harmonic Analysis*, 33(1):24–43, 2012.

[152] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proc Intl Conf on Machine Learning*, pages 433–440, 2009.

[153] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

[154] G. Obozinski, L. Jacob, and J.-P. Vert. Group lasso with overlaps: The latent group lasso approach. *Submitted*, 2011. online at `arxiv.org/abs/1110.0413`.

[155] M. A. Davenport and E. Arias-Castro. Compressive binary search. In *Proc. IEEE Intl. Symp on Information Theory*, pages 1827–1831, 2012.

[156] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

[157] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Submitted*, 2010. online at `arxiv.org/abs/1009.2139`.

[158] M. Malloy and R. Nowak. Near-optimal adaptive compressive sensing. In *Proc. Asilomar Conf. on Signals, Systems, and Computers*, 2012.

[159] J. Haupt, R. Castro, and R. Nowak. Adaptive sensing for sparse signal recovery. In *Proc. IEEE DSP Workshop and Workshop on Sig. Proc. Education*, pages 702–707, 2009.

[160] E. J. Candès and M. A. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[161] D. E. Knuth. *Art of Computer Programming Volume 1: Fundamental Algorithms.* Addison-Wesley Publishing Company, 1972.

[162] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *CoRR*, abs/1310.7991, 2013.

[163] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, New York, NY, USA, 2013. ACM.