Atomic norm denoising with applications to line spectral estimation*

Badri Narayan Bhaskar†, Gongguo Tang†, and Benjamin Recht♯
†Department of Electrical and Computer Engineering
♯Department of Computer Sciences
University of Wisconsin-Madison

April 2012

Abstract

The sub-Nyquist estimation of line spectra is a classical problem in signal processing, but currently popular subspace-based techniques have few guarantees in the presence of noise and rely on *a priori* knowledge about system model order. Motivated by recent work on atomic norms in inverse problems, we propose a new approach to line spectral estimation that provides theoretical guarantees for the mean-squared-error performance in the presence of noise and without advance knowledge of the model order. We propose an abstract theory of denoising with atomic norms and specialize this theory to provide a convex optimization problem for estimating the frequencies and phases of a mixture of complex exponentials with guaranteed bounds on the mean-squared error. We show that the associated convex optimization problem, called *Atomic norm Soft Thresholding* (AST), can be solved in polynomial time via semidefinite programming. For very large scale problems we provide an alternative, efficient algorithm, called *Discretized Atomic norm Soft Thresholding* (DAST), based on the Fast Fourier Transform that achieves nearly the same error rate as that guaranteed by the semidefinite programming approach. We compare both AST and DAST with Cadzow's canonical alternating projection algorithm and demonstrate that AST outperforms DAST which outperforms Cadzow in terms of mean-square reconstruction error over a wide range of signal-to-noise ratios. For very large problems DAST is considerably faster than both AST and Cadzow.

1 Introduction

Extracting the frequencies and relative phases of a superposition of complex exponentials from a small number of noisy time samples is a foundational problem in statistical signal processing. These *line spectral estimation* problems arise in a variety of applications, including the direction of arrival estimation in radar target identification [1], sensor array signal processing [2] and imaging systems [3]. Line spectral estimation also underlies techniques in ultra wideband channel estimation [4], spectroscopy [5], and power electronics [6].

While superresolution techniques based on polynomial interpolation can estimate the frequency content of a signal *exactly* from as few as $2k$ samples if there are k frequencies, these methods are very sensitive to noise. Several variants of such interpolation ideas have been proposed [7–9] for high resolution frequency estimation (for an extensive bibliography on the subject, see [10]). However, these techniques do not yield satisfactory denoising performance when the signal-to-noise ratio (SNR) is low and require a priori knowledge of model order. Motivated by recent work on atomic

*A preliminary version of this work appeared in the Proceedings of the 49th Annual Allerton Conference in 2011.

norms [11], we propose a convex relaxation approach to denoise a mixture of complex exponentials, overcoming many of the shortcomings of previous subspace-based approaches.

Our first contribution is an abstract theory of denoising with atomic norms. Atomic norms provide a natural convex penalty function for discouraging specialized notions of complexity. These norms generalize the ℓ_1 norm for sparse vector estimation [12] and the nuclear norm for low-rank matrix reconstruction [13, 14]. Here, we provide a unified approach to denoising with the atomic norm that provides a standard approach to computing low mean-squared-error estimates. Our approach is based upon a generalization of Basis Pursuit Denoising [15], and we show how certain Gaussian statistics and geometrical quantities of particular atomic norms are sufficient to bound estimation rates with these penalty functions.

Specializing these denoising results to the line spectral estimation problem, we provide mean-squared-error estimates for denoising line spectra with the atomic norm. The denoising algorithm amounts to soft thresholding the noise corrupted measurements in the atomic norm and we thus refer to the problem as *Atomic norm Soft Thresholding* (AST). We show, via an appeal to the theory of positive polynomials, that AST can be solved using semidefinite programming [16], and we provide a reasonably fast method for solving this SDP via the Alternating Direction Method of Multipliers (ADMM) [17, 18]. Our ADMM implementation enables the solution of instances with a thousand observations in a few minutes. For very large instances, we show that basis pursuit denoising on a dense oversampled grid of frequencies approximates the solution of the atomic norm minimization problem to a resolution sufficiently high to guarantee excellent mean-squared error. By leveraging the Fast Fourier Transform, this *Discretized Atomic norm Soft Thresholding* (DAST) formulation can be solved with freely available software such as SpaRSA [19]. A DAST problem with thousands of observations can be solved in under a second in Matlab.

We compare and contrast our algorithm, AST and DAST with Cadzow's iterative alternating projections approach which has empirically been shown to be an effective denoising technique at low SNR. Our experiments indicate that both AST and DAST outperform Cadzow's method in low SNR even when we provide the exact model order to Cadzow's method. Moreover, AST has the same complexity as Cadzow, alternating between a least-squares step and an eigenvalue thresholding step. DAST has even lower computational complexity, consisting of iterations based upon the Fast Fourier Transform and simple linear time soft-thresholding.

1.1 Outline and Summary of Results

The denoising problem is obtaining an estimate $\hat{x}$ of the signal $x^\star$ from $y = x^\star + w^\star$, where $w^\star$ is additive noise. For the general sparse approximation problem, where $x^\star$ is a sparse non-negative combination of points from an arbitrary set $\mathcal{A}$, we analyze the performance of regularization with the atomic norm penalty [11]. The atomic norm $\|\cdot\|_{\mathcal{A}}$ is a penalty function specially catered to the structure of $\mathcal{A}$ as we shall examine in depth in next section, and is defined as:

$$\|x\|_{\mathcal{A}} = \inf\left\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\right\}.$$

The corresponding dual norm, $\|\cdot\|_{\mathcal{A}}^*$, is given by

$$\|z\|_{\mathcal{A}}^* = \sup_{a \in \mathcal{A}} \langle a, z\rangle.$$

In Section 2, we characterize the performance of the estimate $\hat{x}$ obtained by solving

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\|x - y\|_2^2 + \tau\|x\|_{\mathcal{A}}. \tag{1.1}$$

We provide an upper bound on the mean-squared error (MSE) when the noise statistics is known:

Theorem 1. Suppose, we observe the signal $y = x^\star + w^\star$ where $x^\star$ is a sparse nonnegative combination of points in $\mathcal{A}$, the estimate $\hat{x}$ of $x^\star$ given by the solution of the atomic soft thresholding problem (1.1) has the expected mean-squared error

$$\mathbb{E}\|\hat{x} - x^\star\|_2^2 \leq \tau \|x^\star\|_{\mathcal{A}}$$

whenever $\tau \geq \mathbb{E}\|w^\star\|_{\mathcal{A}}^*$.

This theorem states that if we select a sufficiently large value of the regularization parameter, the mean-squared error of the denoised estimate is well-controlled. Our lower bound on τ is in terms of the expected dual norm of the noise process $w^\star$, equal to

$$\mathbb{E}[\sup_{a \in \mathcal{A}} \langle a, w^\star \rangle].$$

That is, the optimal τ and achievable mean-squared error can be estimated by studying the extremal values of the stochastic process indexed by the atomic set $\mathcal{A}$.

After establishing the abstract theory, we specialize the results of the abstract denoising problem to line spectral estimation in Section 3. Consider the signal x with a line spectrum composed of k unknown frequencies $f_1^\star, \ldots, f_k^\star$ bandlimited to $[-W, W]$. Then the Nyquist samples of the signal are given by

$$x_m^\star := x\left(\tfrac{m}{2W}\right) = \sum_{l=1}^{k} c_l^\star e^{i2\pi m u_l^\star} \tag{1.2}$$

where $c_1^\star, \ldots, c_k^\star$ are unknown *complex* coefficients and $u_l^\star = \frac{f_l^\star}{2W}$ for $l = 1, \ldots, k$ are the normalized frequencies. So, the vector $x^\star = [x_0^\star \ \cdots \ x_{n-1}^\star]^T \in \mathbb{C}^n$ of n time samples can be written as a non-negative linear combination of k points from the infinite set

$$\mathcal{A} = \left\{ [e^{i\phi} \ \cdots e^{i(2\pi(n-1)t+\phi)}]^T,\, t \in [0,1], \phi \in [0, 2\pi] \right\}.$$

When the number of observations $n \gg k$, $x^\star$ is k-sparse and thus line spectral estimation in the presence of noise can be thought of as instance of a sparse approximation problem. The choice of the regularization parameter for the strongest guarantee in Theorem 1 is given in terms of the expected dual norm of the noise and can be explicitly computed for many noise models. For example, when the noise model is Gaussian, we have the following theorem for the mean-squared error:

Theorem 2. Suppose $x^\star \in \mathbb{C}^n$ is given by $x_m^\star = \sum_{l=1}^k c_l^\star e^{i2\pi m u_l^\star}$ for some unknown complex numbers $c_1^\star, \ldots, c_k^\star$, unknown normalized frequencies $u_1^\star, \ldots, u_k^\star \in [0,1]$ and $w^\star \in \mathcal{N}(0, \sigma^2 I_n)$, the estimate $\hat{x}$ of $x^\star$ obtained from $y = x^\star + w^\star$ given by the solution of atomic soft thresholding problem (1.1) with $\tau = \sigma\sqrt{n \log(n)}$ has the asymptotic mean-squared-error rate

$$\frac{1}{n}\,\mathbb{E}\|\hat{x} - x^\star\|_2^2 \lesssim \sigma \sqrt{\frac{\log(n)}{n}} \sum_{l=1}^{k} |c_l^\star|.$$

Note that the number of samples we need for robust estimation is only a function of the number of frequencies present and not the bandwidth. Thus, atomic norm soft thresholding provides a robust superresolution scheme with theoretical guarantees for error rates.

We show in Section 3.1 that (1.1) for line spectral estimation can be reformulated as a semidefinite program and can be solved on moderately sized problems via semidefinite programming. We also show that we get the same performance by discretizing the problem and solving basis pursuit denoising on a grid of a large number of points. This can be solved efficiently using standard ℓ_1 minimization software, and its performance is robust to the density of the frequency grid. Our discretization results justify the success of basis pursuit denoising for superresolution problems, even though many of the common tools for compressed sensing do not apply in this case. Our measurement matrix does not obey RIP or incoherence bounds that are commonly used. Nonetheless, we are able to derive estimates on the mean-squared error and obtain excellent denoising in practice.

The canonical algorithm for denoising line spectra is Cadzow's alternating projection algorithm [20]. Our experiments in Section 5 demonstrate that our proposed estimation algorithms outperform Cadzow's technique. Both algorithms obtain lower mean-squared error, and our discretized algorithm is much faster on large problems.

2 Abstract Denoising with Atomic Norms

The foundation of our technique consists of extending some of the recent work on *atomic norms* in linear inverse problems in [11]. In this work, the authors describe how to reconstruct models which can be expressed as sparse linear combinations of *atoms* from some basic set $\mathcal{A}$. The set $\mathcal{A}$ can be very general and not assumed to be discrete. For example, if the signal is known to be a low rank matrix, $\mathcal{A}$ could be the set of all rank-1 matrices. As we will return to in the sequel, $\mathcal{A}$ could also consist of all atomic moment sequences.

We show how to use an atomic norm penalty to denoise a signal known to be a sparse nonnegative combination of atoms from a set $\mathcal{A}$. We compute the MSE for the estimate we thus obtain and propose an efficient computational method.

Definition 3 (Atomic Norm)**.** The atomic norm $\|\cdot\|_{\mathcal{A}}$ of $\mathcal{A}$ is the Minkowski functional (or the gauge function) associated with $\operatorname{conv}(\mathcal{A})$ (the convex hull of $\mathcal{A}$) and is defined by:

$$\|x\|_{\mathcal{A}} = \inf\left\{t > 0 \mid x \in t \operatorname{conv}(\mathcal{A})\right\}. \tag{2.1}$$

The gauge function is a norm if $\operatorname{conv}(\mathcal{A})$ is compact, centrally symmetric, and contains a ball of radius ϵ around the origin for some $\epsilon > 0$. Our results do not depend on $\|\cdot\|_{\mathcal{A}}$ being a norm, although it is typically a norm in many applications. When $\mathcal{A}$ is the set of unit norm 1-sparse elements in $\mathbb{C}^n$, the atomic norm $\|\cdot\|_{\mathcal{A}}$ is the ℓ_1 norm [12]. Similarly, when $\mathcal{A}$ is the set of unit norm rank-1 matrices, the atomic norm is the nuclear norm [13]. In [11], the authors showed that minimizing the atomic norm subject to equality constraints provided exact solutions of a variety of linear inverse problems with nearly optimal bounds on the number of measurements required.

To set up the atomic norm denoising problem, suppose we observe a signal $y = x^\star + w^\star$ and that we know *a priori* that $x^\star$ can be written as a linear combinations of a few atoms from $\mathcal{A}$. One way to estimate $x^\star$ from these observations would be to search over all short linear combinations from $\mathcal{A}$, and to select the one which minimizes $\|y - x\|_2$. However, this could be ly formidable: even if

the set of atoms was a discrete collection of unit vectors, this problem is the NP-hard SPARSEST VECTOR problem [21].

On the other hand, the problem (1.1) is convex, and reduces to many familiar denoising strategies for particular $\mathcal{A}$. The mapping from y to the optimal solution of the above is called the Moreau-Yosida proximal operator of the atomic norm applied to y, and can be thought of as a soft thresholded version of y. Indeed, when $\mathcal{A}$ is the set of 1-sparse atoms, the atomic norm is the ℓ_1-norm, and the proximity operator corresponds to *soft-thresholding* y by element-wise shrinking towards zero [22]. Similarly, when $\mathcal{A}$ is the set of rank-1 matrices, the atomic norm is the nuclear norm and the proximity operator shrinks the singular values of the input matrix towards zero.

We now establish some universal properties about the problem (1.1). First, we collect a simple consequence of the optimality conditions in a lemma:

Lemma 4 (Optimality Conditions)**.** $\hat{x}$ is the solution of (1.1) if and only if

1. $\|y - \hat{x}\|_{\mathcal{A}}^* \leq \tau$.
2. $\langle y - \hat{x}, \hat{x} \rangle = \tau \|\hat{x}\|_{\mathcal{A}}$

The dual atomic norm is given by the variational expression

$$\|z\|_{\mathcal{A}}^* = \sup_{\|x\|_{\mathcal{A}} \leq 1} \langle x, z \rangle. \tag{2.2}$$

which satisfies

$$\langle x, z \rangle \leq \|x\|_{\mathcal{A}} \|z\|_{\mathcal{A}}^*. \tag{2.3}$$

The supremum in (2.2) is achieved and for any x, there is a z that achieves equality. Moreover, we are guaranteed that the optimal solution will actually lie in the set $\mathcal{A}$:

$$\|z\|_{\mathcal{A}}^* = \sup_{a \in \mathcal{A}} \langle a, z \rangle. \tag{2.4}$$

The dual norm will play a critical role throughout, as our asymptotic error rates will be in terms of the dual atomic norm of noise processes. The dual atomic norm also appears in the dual problem of (1.1)

Lemma 5 (Dual Problem)**.** The dual problem of (1.1) is given by

$$\begin{aligned} &\underset{z}{\text{maximize}} \; \frac{1}{2} \left(\|y\|_2^2 - \|y - z\|_2^2 \right) \\ &\text{subject to } \|z\|_{\mathcal{A}}^* \leq \tau. \end{aligned}$$

The dual problem admits a unique solution $\hat{z}$. The primal solution $\hat{x}$ and the dual solution $\hat{z}$ are specified by the optimality conditions and there is no duality gap.

1. $y = \hat{x} + \hat{z}$,
2. $\|\hat{z}\|_{\mathcal{A}}^* \leq \tau$,
3. $\langle \hat{z}, \hat{x} \rangle = \tau \|\hat{x}\|_{\mathcal{A}}$.

The proofs of Lemma 4 and Lemma 5 are provided in Appendix A.1. The optimality conditions shows that $\hat{x}$ can be regarded as the unique solution of another problem:

Corollary 6. The solution $\hat{x}$ of (1.1) is also the unique solution of

$$\begin{aligned} &\text{minimize } \|x\|_2 \\ &\text{subject to } \|y - x\|_{\mathcal{A}}^* \leq \tau. \end{aligned}$$

We are now ready to state a proposition which gives an upper bound on the MSE with the optimal choice of the regularization parameter.

Proposition 7 (MSE and Regularization Parameter)**.** If the regularization parameter $\tau > \|w^\star\|_{\mathcal{A}}^*$, the optimal solution $\hat{x}$ of (1.1) has the mean-squared error

$$\frac{1}{n}\|\hat{x} - x^\star\|_2^2 \leq \frac{1}{n}\left(\tau\|x^\star\|_{\mathcal{A}} - \langle x^\star, w^\star\rangle\right) \leq \frac{2\tau}{n}\|x^\star\|_{\mathcal{A}}. \tag{2.5}$$

Proof.

$$\begin{aligned} \|\hat{x} - x^\star\|_2^2 &= \langle \hat{x} - x^\star, w^\star - (y - \hat{x})\rangle && (2.6)\\ &\leq \tau\|x^\star\|_{\mathcal{A}} - \langle x^\star, w^\star\rangle + (\|w^\star\|_{\mathcal{A}}^* - \tau)\|\hat{x}\|_{\mathcal{A}} && (2.7)\\ &\leq (\tau + \|w^\star\|_{\mathcal{A}}^*)\|x^\star\|_{\mathcal{A}} + (\|w^\star\|_{\mathcal{A}}^* - \tau)\|\hat{x}\|_{\mathcal{A}} && (2.8) \end{aligned}$$

The theorem now follows from (2.7) and (2.8) after substituting $\tau > \|w^\star\|_{\mathcal{A}}^*$. The value of the regularization parameter τ to ensure the MSE is upper bounded thus, is $\|w^\star\|_{\mathcal{A}}^*$. □

Example: Sparse Model Selection We can specialize our stability guarantee to Lasso [23] and recover known results. Let $\Phi \in n \times p$ be a design matrix with unit norm columns, and suppose we observe $y = x^\star + w$, where w is an additive noise term, and $x^\star = \Phi c^\star$ is an unknown k sparse combination of columns of Φ. In this case, the atomic set is the collection of columns of Φ and $-\Phi$, and the atomic norm $\|x\|_{\mathcal{A}}$ coincides with the ℓ_1 norm of the coefficients given by $\|c^\star\|_1$. Therefore, the proposed optimization problem (1.1) coincides with the Lasso estimator, also called Basis Pursuit Denoising [15]. If we assume that w is gaussian vector with a variance σ^2 for its entries, the expected dual atomic norm of the noise term, $\|w\|_{\mathcal{A}}^* = \|\Phi^* w\|_\infty$ is simply the expected maximum of p gaussian random variables. Using the well known result on the maximum of gaussian random variables [24], we have $\mathbb{E}\|w\|_{\mathcal{A}}^* \leq \sigma\sqrt{2\log(p)}$. If $\hat{x}$ is the denoised signal, we have from Theorem 1 that if $\tau = \mathbb{E}\|w\|_{\mathcal{A}}^* = \sigma\sqrt{2\log(p)}$,

$$\frac{1}{n}\,\mathbb{E}\|\hat{x} - x^\star\|_2^2 \leq \sigma\frac{\sqrt{2\log(p)}}{n}\|c^\star\|_1,$$

which is the stability result for Lasso reported in [25] assuming no conditions on Φ.

2.1 Accelerated Convergence Rates

In this section, we provide conditions under which a faster convergence rate can be obtained for atomic norm soft thresholding. We summarize our results in the following

Proposition 8 (Fast Rates)**.** Suppose the set of atoms $\mathcal{A}$ is centrosymmetric and $\|w\|_{\mathcal{A}}^*$ concentrates about its expectation so that $P(\|w\|_{\mathcal{A}}^* \geq \mathbb{E}[\|w\|_{\mathcal{A}}^*] + t) < \delta(t)$. Define the cone

$$C_\gamma(x^\star, \mathcal{A}) = \text{cone}(\{z : \exists\alpha > 0 \text{ with } \|x^\star + \alpha z\|_{\mathcal{A}} \leq \|x^\star\|_{\mathcal{A}} + \alpha\gamma\|z\|_{\mathcal{A}}\}). \tag{2.9}$$

Suppose

$$\phi_\gamma(x^\star, \mathcal{A}) := \inf\left\{ \frac{\|z\|_2}{\|z\|_\mathcal{A}} : z \in C_\gamma(x^\star, \mathcal{A}) \right\} \tag{2.10}$$

is strictly greater than zero for some $\gamma > \mathbb{E}\|w\|^*_\mathcal{A}/\tau$. Then

$$\|\hat{x} - x^\star\|_2^2 \leq \frac{(1+\gamma)^2\tau^2}{\gamma^2\phi_\gamma(x^\star, \mathcal{A})^2} \tag{2.11}$$

with probability at least $1 - \delta(\gamma\tau - \mathbb{E}\|w\|^*_\mathcal{A})$.

Having the ratio of norms bounded below is a generalization of the Weak Compatibility criterion used to quantify when fast rates are achievable for the Lasso [26]. One slight difference is that we define the corresponding cone C_γ where ϕ_γ must be controlled in parallel with the *tangent cones* studied in [11]. There, the authors showed that the mean width of the cone $C_0(x^\star, \mathcal{A})$ determined the number of random linear measurements required to recover $x^\star$ using atomic norm minimization. In our case, γ is greater than zero, and represents a "widening" of the tangent cone. When $\gamma = 1$, the cone is all of $\mathbb{R}^n$ (via the triangle inequality), hence τ must be a bit larger than the expectation to enable our proposition to hold.

Proof. We proceed by first showing $\hat{x} - x^\star \in C_\gamma(x^\star, \mathcal{A})$. Since $\hat{x}$ is optimal, we have,

$$\frac{1}{2}\|y - \hat{x}\|_2^2 + \tau\|\hat{x}\|_\mathcal{A} \leq \frac{1}{2}\|y - x^\star\|_2^2 + \tau\|x^\star\|_\mathcal{A}$$

Rearranging:

$$\tau\|\hat{x}\|_\mathcal{A} \leq \tau\|x^\star\|_\mathcal{A} + \|w\|^*_\mathcal{A}\|\hat{x} - x^\star\|_\mathcal{A}$$

as desired. Since $\|w\|^*_\mathcal{A}$ concentrates about its expectation, with probability $> 1 - \delta$ we get that $\hat{x} - x^\star \in C_\gamma(x^\star, \mathcal{A})$.

Using (2.6), if $\tau > \|w\|^*_\mathcal{A}$, we have,

$$\begin{aligned}\|\hat{x} - x^\star\|_2^2 &\leq (\tau + \|w\|^*_\mathcal{A})\|\hat{x} - x^\star\|_\mathcal{A} \\ &\leq \frac{(1+\gamma)\tau}{\gamma\phi_\gamma(x^\star, \mathcal{A})}\|\hat{x} - x^\star\|_2\end{aligned}$$

So, with probability at least $1 - \delta(\gamma\tau - \mathbb{E}\|w\|^*_\mathcal{A})$:

$$\|\hat{x} - x^\star\|_2^2 \leq \frac{(1+\gamma)^2\tau^2}{\gamma^2\phi_\gamma(x^\star, \mathcal{A})^2} \tag{2.12}$$

as desired. □

The main difference between (2.12) and (2.5) is that the mean-squared error is controlled by τ^2 rather than $\tau\|x^*\|_\mathcal{A}$. As we will now see (2.12) provides minimax optimal rates for the examples of sparse vectors and low-rank matrices.

Example: Sparse Vectors in Noise Let $\mathcal{A} = \{\pm e_1, \ldots, \pm e_n\}$, be the set of signed canonical unit vectors in $\mathbb{R}^n$ (e_i is a unit vector with the only nonzero value along the ith coordinate). In this case, $\text{conv}(\mathcal{A})$ is the unit cross polytope and the atomic norm $\|\cdot\|_{\mathcal{A}}$, coincides with the ℓ_1 norm, and the dual atomic norm is the ℓ_∞ norm. Suppose $x^\star \in \mathbb{R}^n$ and $T := \text{supp}(x^\star)$ has cardinality k. Consider the problem of estimating $x^\star$ from $y = x^\star + w^\star$ where $w^\star \sim \mathcal{N}(0, \sigma^2 I_n)$ is Gaussian noise.

We show in the appendix that in this case $\phi_\gamma(x^\star, \mathcal{A}) > \frac{(1-\gamma)}{2\sqrt{k}}$. We also have $\tau_0 = \mathbb{E}\|w\|_\infty \geq \sigma\sqrt{2\log(n)}$. Pick $\tau > \gamma^{-1}\tau_0$ for some $\gamma > 1$. Then, using our lower bound for ϕ_γ in (2.12), we get a rate of

$$\frac{1}{n}\|\hat{x} - x^\star\|_2^2 = O\left(\frac{\sigma^2 k \log(n)}{n}\right) \tag{2.13}$$

for the AST estimate with high probability. Note that this bound coincides with the minimax optimal rate derived by Donoho and Johnstone [27]. Note that if we had used (2.5) instead, our mean-squared error would have instead been $O\left(\frac{\sqrt{\sigma^2 k \log n}\|x^\star\|_2}{n}\right)$, which depends on the norm of the input signal $x^\star$.

Example: Low Rank Matrix in Noise Let $\mathcal{A}$ be the manifold of unit norm rank-1 matrices in $\mathbb{C}^{n\times n}$. In this case, the atomic norm $\|\cdot\|_{\mathcal{A}}$, coincides with the nuclear norm $\|\cdot\|_*$, and the corresponding dual atomic norm is the spectral norm of the matrix. Suppose $X^\star \in \mathbb{C}^{n\times n}$ has rank r, so it can be constructed as a combination of r atoms, and we are interested in estimating $X^\star$ from $Y = X^\star + W^\star$ where $W^\star$ has independent $\mathcal{N}(0, \sigma^2)$ entries.

We prove in the appendix that $\phi_\gamma(X^\star, \mathcal{A}) \geq \frac{1-\gamma}{2\sqrt{2r}}$. To obtain an estimate for τ, we note that the dual norm of the nuclear norm is the operator norm, and $\|W\| \leq 2\sqrt{n}$ with high probability [28]. Substituting these estimates for τ and ϕ_γ in (2.12), we get

$$\|X - \hat{X}\|_F^2 = O\left(\sigma^2 nr\right)$$

which is minimax optimal.

2.2 Expected Mean-Squared Error for Approximated Atomic Norms

We close this section by noting that it may sometimes be easier to solve (1.1) on a different set $\widetilde{\mathcal{A}}$ (say, an ϵ-net of $\mathcal{A}$) instead of $\mathcal{A}$. If for some $M > 0$,

$$M^{-1}\|x\|_{\widetilde{\mathcal{A}}} \leq \|x\|_{\mathcal{A}} \leq \|x\|_{\widetilde{\mathcal{A}}}$$

holds for every x, then Theorem 1 still applies with a constant factor M. We will need the following lemma.

Lemma 9.

$$\|z\|_{\mathcal{A}}^* \leq M\|z\|_{\widetilde{\mathcal{A}}}^* \text{ for every } z \text{ iff } M^{-1}\|x\|_{\widetilde{\mathcal{A}}} \leq \|x\|_{\mathcal{A}} \text{ for every } z.$$

Proof. We will show the forward implication – the converse will follow since the dual of the dual norm is again the primal norm. By tightness of the dual norm inequality (2.3), for any x, there

exists a z with $\|z\|_{\widetilde{\mathcal{A}}}^* \leq 1$ and $\langle x, z\rangle = \|x\|_{\widetilde{\mathcal{A}}}$. So,

$$\begin{aligned} M^{-1}\|x\|_{\widetilde{\mathcal{A}}} &= M^{-1}\langle x, z\rangle \\ &\leq M^{-1}\|z\|_{\mathcal{A}}^*\|x\|_{\mathcal{A}} && \text{by (2.3)} \\ &\leq \|x\|_{\mathcal{A}} && \text{by the assumption.} \end{aligned}$$

□

Now, we can state the sufficient condition for the following proposition in terms of either the primal or the dual norm:

Proposition. Suppose

$$\|z\|_{\widetilde{\mathcal{A}}}^* \leq \|z\|_{\mathcal{A}}^* \leq M\|z\|_{\widetilde{\mathcal{A}}}^* \text{ for every } z, \tag{2.14}$$

or equivalently

$$M^{-1}\|x\|_{\widetilde{\mathcal{A}}} \leq \|x\|_{\mathcal{A}} \leq \|x\|_{\widetilde{\mathcal{A}}} \text{ for every } x, \tag{2.15}$$

then under the same conditions as in Theorem 1,

$$\frac{1}{n}\,\mathbb{E}\|\tilde{x} - x^\star\|_2^2 \leq \frac{M\tau}{n}\|x^\star\|_{\mathcal{A}}$$

where $\tilde{x}$ is the optimal solution for (1.1) with $\widetilde{\mathcal{A}}$ substituted for $\mathcal{A}$.

Proof. By assumption, $\mathbb{E}\left(\|w^\star\|_{\mathcal{A}}^*\right) \leq \tau$. Now, (2.14) implies $\mathbb{E}\left(\|w^\star\|_{\widetilde{\mathcal{A}}}^*\right) \leq \tau$. Applying Theorem 1, and using (2.15), we get

$$\frac{1}{n}\,\mathbb{E}\|\tilde{x} - x^\star\|_2^2 \leq \frac{\tau}{n}\|x^\star\|_{\widetilde{\mathcal{A}}} \leq \frac{M\tau}{n}\|x^\star\|_{\mathcal{A}}.$$

□

3 Application to Line Spectral Estimation

Let us now return to the line spectral estimation problem, where we are trying to denoise a linear combination of complex sinusoids. The atomic set in this case should be the samples of individual sinusoids, $a_{t,\phi} \in \mathbb{C}^n$, given by

$$a_{t,\phi} = \left[e^{i\phi}\ e^{i(2\pi t+\phi)}\ \cdots\ e^{i(2\pi(n-1)t+\phi)}\right]^T. \tag{3.1}$$

The uncountably infinite collection $\mathcal{A} = \{a_{t,\phi} \mid t \in [0,1], \phi \in [0, 2\pi]\}$ forms an appropriate collection of atoms for $x^\star$, since $x^\star$ in (1.2) can be written as a sparse nonnegative combination of atoms in $\mathcal{A}$. In fact, $x^\star = \sum_{l=1}^k c_l^\star a_{u_l^\star,0} = \sum_{l=1}^k |c_l^\star| a_{u_l^\star,\phi_l}$, where $c_l^\star = |c_l^\star|e^{i\phi_l}$. The choice of the regularization parameter is dictated by the noise model and we show the optimal choice for white gaussian noise samples in our analysis of the mean-squared error in the next section.

The dual norm induced by the set $\mathcal{A}$ here takes a very intuitive form:

$$\begin{aligned} \|v\|_{\mathcal{A}}^* &= \sup_{a_{t,\phi}\in\mathcal{A}} \langle a_{t,\phi}, v\rangle = \sup_{t\in[0,1]} \sup_{\phi\in[0,2\pi]} e^{i\phi}\sum_{k=0}^{n-1} v_k e^{2\pi i k t} \\ &= \sup_{|\omega|\leq 1} \left|\sum_{k=0}^{n-1} v_k \omega^k\right|. \end{aligned} \tag{3.2}$$

In other words, $\|v\|_{\mathcal{A}}^*$ is the maximum absolute value attained on the unit circle by the polynomial $z \mapsto \sum_{k=0}^{n-1} v_k z^k$. A simple lower bound for the expected dual norm occurs when we consider the maximum value of n uniformly spaced points in the unit circle. Using the result of [24], the lower bound whenever $n \geq 5$ is

$$\sigma\sqrt{n\log(n) - \tfrac{n}{2}\log(4\pi\log(n))}\,.$$

Using a theorem of Bernstein and standard results on the extreme value statistics of Gaussian distribution, we can also obtain a non-asymptotic upper bound on the expected dual norm of noise for $n > 3$:

$$\sigma\left(1 + \frac{1}{\log(n)}\right)\sqrt{n\log(n) + n\log(4\pi\log(n))}$$

(See Appendix A.4 for a derivation of both the lower and upper bound). If we set the regularization parameter τ equal to an upper bound on the expected dual atomic norm, i.e.,

$$\tau = \sigma\left(1 + \frac{1}{\log(n)}\right)\sqrt{n\log(n) + n\log(4\pi\log(n))}. \tag{3.3}$$

an application of Theorem 1 yields the asymptotic result in Theorem 2.

3.1 Semidefinite Programming for Atomic Soft Thresholding

In this section, we present a semidefinite characterization of the atomic norm associated with the line spectral atomic set $\mathcal{A} = \{a_{t,\phi} | t \in [0,1], \phi \in [0, 2\pi]\}$. This characterization allows us to rewrite the atomic denoising problem (1.1) as an equivalent semidefinite programming problem.

Recall from (3.2) that the dual atomic norm of a vector $v \in \mathbb{C}^n$ is the maximum absolute value of a complex trigonometric polynomial $V(t) = \sum_{l=0}^{n-1} v_k e^{2\pi i l t}, t \in [0,1]$. As a consequence, a constraint on the size of the dual atomic norm is equivalent to a bound on the maganitude of $V(t)$:

$$\|v\|_{\mathcal{A}}^* \leq \tau \Leftrightarrow |V(t)|^2 \leq \tau^2, \forall t \in [0,1].$$

The function $q(t) = \tau^2 - |V(t)|^2$ is a trigonometric polynomial (that is, a polynomial in the variables z and $\bar{z}$ with $|z| = 1$). A necessary and sufficient condition for $q(t)$ to be nonnegative is that it can be written as a sum of squares of polynomials in the variables z and $\bar{z}$ [16]. Testing if q is a sum of squares can be achieved via semidefinite programming. To state the associated semidefinite program, define the map $T : \mathbb{C}^n \to \mathbb{C}^{n \times n}$ which creates a Hermitian Toeplitz matrix out of its input. That is

$$T(x) = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_n \\ \bar{x}_2 & x_1 & x_2 & \dots & x_{n-1} \\ \bar{x}_3 & \bar{x}_2 & x_1 & \dots & x_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{x}_n & \bar{x}_{n-1} & \bar{x}_{n-2} & \dots & x_1 \end{bmatrix}$$

Let T^* denote the adjoint of the map T. Then we have the following succinct characterization

Lemma 10. [29, Theorem 4.24] For any given causal trigonometric polynomial $V(t) = \sum_{l=0}^{n-1} v_l e^{2\pi i l t}$, $|V(t)| \leq \tau$ if and only if there exists complex Hermitian matrix Q such that

$$T^*(Q) = \tau^2 e_1 \quad \text{and} \quad \begin{bmatrix} Q & v \\ v^* & 1 \end{bmatrix} \succeq 0.$$

Here, e_1 is the first canonical basis vector with a one at the first component and zeros elsewhere and v^* denotes the Hermitian adjoint (conjugate transpose) of v.

Using Lemma 10, we rewrite the atomic norm $\|x\|_{\mathcal{A}} = \sup_{\|v\|_{\mathcal{A}}^* \leq 1} \langle x, v \rangle$ as the following semidefinite program:

$$\begin{array}{ll} \text{maximize}_{v,\ Q} & \langle x, v \rangle \\ \text{subject to} & T^*(Q) = e_1 \\ & \begin{bmatrix} Q & v \\ v^* & 1 \end{bmatrix} \succeq 0. \end{array} \tag{3.4}$$

The dual problem of (3.4) (after a trivial rescaling) is then equal to the atomic norm of x:

$$\begin{array}{rll} \|x\|_{\mathcal{A}} = & \min_{t,u} & \frac{1}{2}(t + u_1) \\ & \text{subject to} & \begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \succeq 0. \end{array}$$

Therefore, the atomic denoising problem (1.1) for the set of trigonometric atoms is equivalent to

$$\begin{array}{ll} \text{minimize}_{t,u,x} & \frac{1}{2}\|x - y\|_2^2 + \frac{\tau}{2}(t + u_1) \\ \text{subject to} & \begin{bmatrix} T(u) & x \\ x^* & t \end{bmatrix} \succeq 0. \end{array} \tag{3.5}$$

To gain intuition for the form of this semidefinite program, consider the atom

$$a_{\theta,\omega} = \begin{bmatrix} e^{i\omega} & e^{i(2\pi\theta+\omega)} & \cdots & e^{i(2\pi(n-1)\theta+\omega)} \end{bmatrix}^T .$$

Now, one can easily check that $a_{\theta,\omega} a_{\theta,\omega}^*$ is a Toeplitz matrix, and it is positive definite because it is an outer product. Thus,

$$\begin{bmatrix} a_{\theta,\omega} a_{\theta,\omega}^* & a_{\theta,\omega} \\ a_{\theta,\omega}^* & 1 \end{bmatrix} = \begin{bmatrix} a_{\theta,0} a_{\theta,0}^* & a_{\theta,\omega} \\ a_{\theta,\omega}^* & 1 \end{bmatrix} \tag{3.6}$$

is feasible for (3.5) and the corresponding u and t satisfy $(u_1 + t)/2 = 1$. All positive combinations of matrices of the form (3.6) are feasible for (3.5), and we will get an upper bound on the atomic norm of the final column by the sum of the coefficients in the corresponding positive combination.

The semidefinite program (3.5) can be solved by off-the-shelf solvers such as SeDuMi [30] and SDPT3 [31]. However, these solvers tend to be slow for relatively large problems. We now describe a more efficient algorithm based upon the Alternating Direction Method of Multipliers (ADMM) [18].

3.1.1 Application of the Alternating Direction Method of Multipliers

A thorough survey of the ADMM algorithm is given in [18]. We only present the details essential to the implementation of atomic norm soft thresholding. To put our problem in an appropriate form for ADMM, rewrite (3.5) as

$$\begin{array}{ll} \text{minimize}_{t,u,x,Z} & \frac{1}{2}\|x - y\|_2^2 + \frac{\tau}{2}(t + u_1) \\ \text{subject to} & Z = \begin{bmatrix} Tu & x \\ x^* & t \end{bmatrix} \\ & Z \succeq 0. \end{array}$$

and dualize the equality constraint via an Augmented Lagrangian:

$$\mathcal{L}_\rho(t,u,x,Z,\Lambda) = \frac{1}{2}\|x-y\|_2^2 + \frac{\tau}{2}(t+u_1) + \left\langle \Lambda, Z - \begin{bmatrix} Tu & x \\ x^* & t \end{bmatrix} \right\rangle + \frac{\rho}{2}\left\| Z - \begin{bmatrix} Tu & x \\ x^* & t \end{bmatrix} \right\|_F^2$$

ADMM then consists of the update steps:

$$\begin{aligned}
(t^{k+1}, u^{k+1}, x^{k+1}) &:= \arg\min_{t,u,x} \mathcal{L}_\rho(t,u,x,Z^k,\Lambda^k) \\
Z^{k+1} &:= \arg\min_{Z\succeq 0} \mathcal{L}_\rho(t^{k+1},u^{k+1},x^{k+1},Z,\Lambda^k) \\
\Lambda^{k+1} &:= \Lambda^k + \rho\left(Z^{k+1} - \begin{bmatrix} Tu^{k+1} & x^{k+1} \\ x^{k+1*} & t^{k+1} \end{bmatrix}\right).
\end{aligned}$$

The updates with respect to t, x, and u can be computed in closed form:

$$\begin{aligned}
x &= \frac{1}{2\rho+1}(y + 2\rho z_1^k + 2\lambda_1^k) \\
q &= W\left(T^*(Z_0^k + \Lambda_0^k/\rho) - \frac{\tau}{2\rho}e_1\right) \\
t &= Z_{n+1,n+1}^k + \left(\Lambda_{n+1,n+1}^k - \frac{\tau}{2}\right)/\rho
\end{aligned}$$

Here W is the diagonal matrix with entries

$$W_{ii} = \begin{cases} \frac{1}{n} & i = 1 \\ \frac{1}{2(n-i+1)} & i > 1 \end{cases}$$

and we introduced the partitions:

$$Z^k = \begin{bmatrix} Z_0^k & z_1^k \\ z_1^{k*} & Z_{n+1,n+1}^k \end{bmatrix} \quad \text{and} \quad \Lambda^k = \begin{bmatrix} \Lambda_0^k & \lambda_1^k \\ \lambda_1^{k*} & \Lambda_{n+1,n+1}^k \end{bmatrix}.$$

.

The Z update is simply the projection onto the positive definite cone

$$Z^{k+1} := \arg\min_{Z\succeq 0} \left\| Z - \begin{bmatrix} Tu^{k+1} & x^{k+1} \\ x^{k+1*} & t^{k+1} \end{bmatrix} + \Lambda^k/\rho \right\|_F^2. \tag{3.7}$$

Projecting a matrix Q onto the positive definite cone is accomplished by forming an eigenvalue decomposition of Q and setting all negative eigenvalues to zero.

To summarize, the update for (t,u,x) requires averaging the diagonals of a matrix (which is equivalent to projecting a matrix onto the space of Toeplitz matrices), and then operations that are $O(n)$. The update for Z requires projecting onto the positive definite cone. The update for Λ is simply addition of symmetric matrices.

3.2 Discretized Atomic Soft Thresholding (DAST)

When the number of samples is larger than a few hundred, the running time of our ADMM method is dominated by the eigenvalue computation (3.7). For very large problems, we now propose a basis pursuit method as an alternative to the semidefinite program (3.5). To proceed, pick a uniform grid of N frequencies and form $\mathcal{A}_N = \left\{ a_{m/N,\phi} \;\middle|\; 0 \le m \le N-1 \right\} \subset \mathcal{A}$ and solve (1.1) on this grid. i.e., we solve the problem

$$\text{minimize } \frac{1}{2}\|x-y\|_2^2 + \tau \|x\|_{\mathcal{A}_N}. \tag{3.8}$$

To see why this is to our advantage, define Φ be the $n \times N$ Fourier matrix with mth column $a_{m/N,0}$. Then any $x \in \text{conv}(\mathcal{A}_N)$ can be written as Φc for $c \in \mathbb{C}^N$, with $\|x\|_{\mathcal{A}_N} = \|c\|_1$. So, we solve

$$\text{minimize } \frac{1}{2}\|\Phi c - y\|_2^2 + \tau \|c\|_1. \tag{3.9}$$

for the optimal point $\hat{c}$ and set $\hat{x}_N = \Phi \hat{c}$ or the first n terms of the N term discrete Fourier transform (DFT) of $\hat{c}$. Furthermore, $\Phi^* z$ is simply the N term inverse DFT of $z \in \mathbb{C}^n$. This observation coupled with Fast Fourier Transform (FFT) algorithm for efficiently computing DFTs gives a fast method to solve (3.8), using standard compressed sensing software for $\ell_2 - \ell_1$ minimization, for example, SparSA [19]. We call this computational method *Discretized Atomic norm Soft Thresholding* (DAST).

Because of the relatively simple structure of the atomic set, the optimal solution $\hat{x}$ for (3.8) can be made arbitrarily close to (3.5) by picking N a constant factor larger than n. In fact, we show that the atomic norms on $\mathcal{A}$ and $\mathcal{A}_N$ are equivalent (See Appendix A.3) and using Theorem 1 and (3.3), we conclude

$$\frac{1}{n}\,\mathbb{E}\|\hat{x}_N - x^\star\|_2^2 \le \left(1 - \frac{2\pi n}{N}\right)^{-1} \sigma \left(1 + \frac{1}{\log(n)}\right) \sqrt{n\log(n) + n\log(4\pi\log(n))}\|x^\star\|_{\mathcal{A}}.$$

Due to the efficiency of the Fast Fourier Transform, DAST has a much lower algorithmic complexity than either Cadzow's alternating projections method or the ADMM method described in Section 3.1, which each require computing a singular value or eigenvalue decomposition at each iteration. Indeed, fast solvers for (3.9) converge to an ϵ optimal solution in no more than $1/\sqrt{\epsilon}$ iterations. Each iteration requires a multiplication by Φ and a simple "shrinkage" step. Multiplication by Φ or Φ^* requires $O(N \log N)$ time and the shrinkage operation can be performed in time $O(N)$.

As we discuss below, this fast form of basis pursuit has been proposed by several authors. However, analyzing this method with tools from compressed sensing has proven daunting because the matrix Φ is nowhere near a restricted isometry. Indeed, as N tends to infinity, the columns become more and more coherent. However, common sense says that a larger grid should give better performance! Indeed, by appealing to the atomic norm framework, we are able to show exactly this point: the larger one makes N, the closer one approximates the desired atomic norm soft thresholding problem. Moreover, we do not have to choose N to be too large in order to achieve nearly the same performance as the AST.

4 Prony's technique and Prior Art

Our method for denoising line spectra stands in contrast to the very classical techniques based on recurrence relations. To review, the sequence in (1.2) must satisfy a kth degree recurrence

relation whose auxiliary polynomial must have its k roots precisely at $\exp(-i2\pi u_l^\star)$ for $l = 1, \ldots, k$. The corresponding frequencies $\{u_l\}_{l=1}^k$ can be found by solving a linear system or a generalized eigenvalue problem. This technique is attributed to Prony in the eighteenth century. Prony's technique and its variants are called Linear Prediction methods, since they convert the nonlinear problem of estimating frequencies into a linear problem by estimating the coefficients of the auxiliary polynomial of the recurrence relation from the data. A survey of these methods can be found in [20] and an extensive list of references is given in [10].

One of the major drawbacks of Linear Prediction methods is that the number of sinusoids, k, must be known to implement the root finding procedure. When there is no noise, k can be determined as the rank of an appropriate Toeplitz matrix formed from $x^\star$, but this procedure breaks down in the presence of noise, since the Toeplitz matrix of moments has full rank with high probability, which makes the determination of model order very difficult. Our AST and DAST methods, on the other hand, require no *a priori* knowledge of the model order.

Even if k is known, the technique is sensitive to perturbations of the coefficients of the auxiliary polynomial [32], and Prony's technique, without sufficient preprocessing, is known to produce inconsistent estimates [33]. In order to robustify Linear Projection methods, Cadzow [34] proposed alternately projecting the observed Toeplitz matrix of moments onto the space of rank k matrices and the space of Toeplitz matrices to preserve the desired low-rank Toeplitz structure. This has been identified as a very fruitful preprocessing step [20] and is closer in spirit to our technique of encouraging sparsity and preserving structure. However, Cadzow still requires knowledge of k. Note that Cadzow has the same iteration complexity as our ADMM solver of Section 3.1. Moreover, as we will show in the experiments, Cadzow produces much worse mean-squared-error estimates than either AST or DAST.

Our algorithm DAST justifies denoising the original moment sequence using Basis Pursuit denoising on a sufficiently large grid. There is some recent work [35] on the recovery of frequencies and amplitudes from a line spectrum by assuming that the frequencies lie on some uniform grid of N points, and attempting reconstruction from the grid. In other words, a slightly mismatched basis is assumed for the time samples, but as pointed out by [36], the performance of the reconstruction technique can degrade considerably due to this basis mismatch. Moreover, these results need to carefully control the incoherence of their linear maps to apply off-the-shelf tools from compressed sensing. It is important to note that the performance of our DAST algorithm improves as the grid size increases. This seems to contradict conventional wisdom in compressed sensing because our design matrix Φ becomes more and more coherent. We note that an added feature of our abstract denoising analysis is an ability to step away from such notions as coherence, and focus on the geometry of the atomic set as the more critical feature for superresolution and related denoising algorithms.

5 Experiments

We compared the MSE performance of AST and DAST, described respectively in Section 3.1 and 3.2, with Cadzow's method. For our experiments, we generated k normalized frequencies $u_1^\star, \ldots, u_k^\star$ both uniformly randomly, and equally spaced in $[0, 1]$. For a given signal amplitude level η, the signal $x^\star \in \mathbb{C}^n$ is generated according to (1.2) with all the amplitudes $c_1^\star = \ldots = c_k^\star = \eta\tau$ where τ is chosen according to (3.3). This amplitude guaranteed that the signal is detectable above the noise floor. All of our sinusoids were then assigned a random phase (which is equivalent to multiplying

$c_k^\star$ by a random unit norm complex number). Now, the observation vector y is generated by adding complex unit variance white gaussian noise $w^\star$. For both the equispaced and the random case, we compare the average MSE of the three algorithms in 10 trials for various values of number of observations ($n = 100, 200, 400, 800$), number of frequencies ($k = 5, 10, 15$) and signal amplitude levels ($\eta = 2, 4, 8, 16$).

First, we implemented AST using the ADMM method described in Section 3.1. We used the stopping criteria described in [18] and set $\rho = 2$ for all experiments. From the optimal u, we additionally *debiased* the solution by solving a least squares solution on the estimated support. Precisely, the optimal u corresponds to a positive moment sequence whose poles and amplitudes are the same as those in $\hat{x}$. We ran Prony's technique on u to extract these frequencies ω_ℓ, and then ran the least squares problem

$$\text{minimize}_\alpha \|U\alpha - y\|^2$$

where $U_{k\ell} = \exp(2\pi i k \omega_\ell)$. After computing the optimal solution α_{opt}, we returned the prediction $\hat{x} = U\alpha_{\text{opt}}$.

Second, we implemented DAST, obtaining an estimate $\hat{x}$ of $x^\star$ from y by solving the optimization problem (3.8) with debiasing. We use the algorithm described in Section 3.2 with grid of $N = 2^{16}$ points. Once we found the optimal c_{opt}, we ran a debiasing step which solves the least squares problem

$$\text{minimize}_\beta \|\Phi_S \beta - y\|^2$$

where Φ_S is the submatrix of Φ whose columns correspond to the support of c_{opt}. We return the estimate $\hat{x} = \Phi_S \beta_{\text{opt}}$. We used the freely downloadable implementation of SpaRSA which implements the debiasing step as a subroutine. We used a stopping parameter of 10^{-4}, but otherwise use the default parameters.

Third, we implemented Cadzow's alternating projection algorithm as described by the pseudocode in [20]. As discussed above, Cadzow needs knowledge of the number of sinusoids. Rather than implementing a heuristic to estimate k, *we fed the true k to our solver*. This provides a huge advantage to the Cadzow algorithm. Neither AST or DAST are provided the true value of k.

We compare the performance of the three algorithms using performance profiles. Performance profiles provide a good visual indicator of the relative performance of many algorithms under a variety of experimental conditions [37]. Let $\mathcal{P}$ be the set of experiments and let $\text{MSE}_s(p)$ be the mean-squared error of experiment $p \in \mathcal{P}$ using the algorithm s. Then the ordinate $P_s(\beta)$ of the graph at β specifies the fraction of experiments where the ratio of the MSE of the algorithm s to the minimum MSE for the given experiment is less than β, i.e.,

$$P_s(\beta) = \frac{\#\left\{p \in \mathcal{P} \ : \ \text{MSE}_s(p) \leq \beta \min_s \text{MSE}_s(p)\right\}}{\#(\mathcal{P})}$$

From the performance profile in Figure 1, we can see that AST is the best performing algorithm over all, with DAST coming in second. Cadzow is consistently worse than both competitors even though it is fed the true number of sinusoids. When Cadzow is fed an incorrect k, even off by 1, the performance degrades drastically, and never provides adequate mean-squared error. Figure 2 shows that the DAST is quite robust to the choice of N. For $N \geq 2^{13}$, all DAST methods perform about as well as one another in accuracy when $n < 1000$.

We benchmarked Cadzow, AST, and DAST on the same Dell R510 server (2x Intel Xeon X54650 cores, 128 GB of RAM). The results of our benchmarking appear in Table 3. AST is the slowest

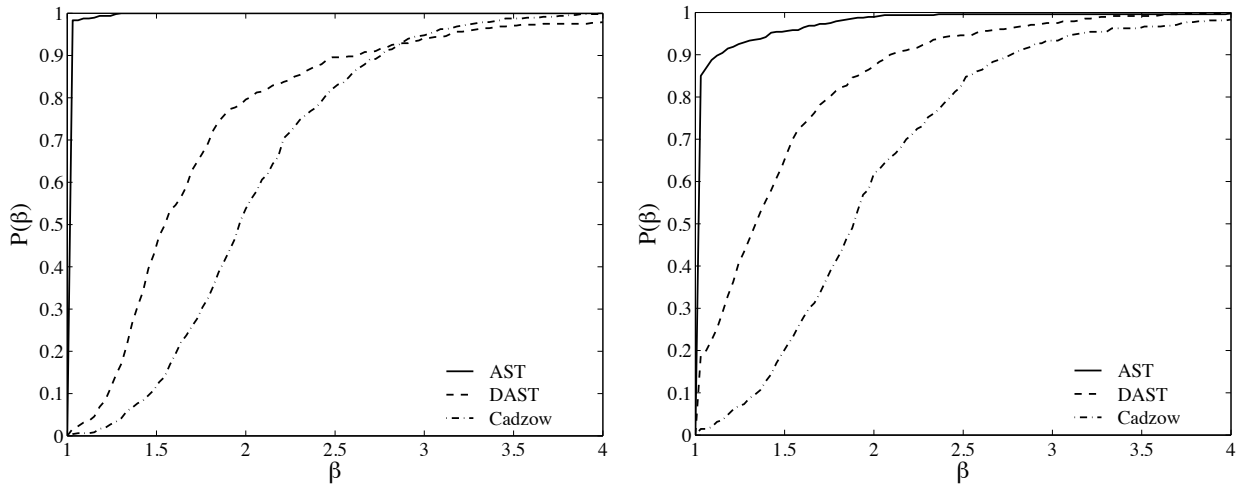

Figure 1: Performance Profiles: These graphs show the performance profiles comparing AST, DAST, and Cadzow's method for uniformly random frequencies (left) and equispaced frequencies (right).

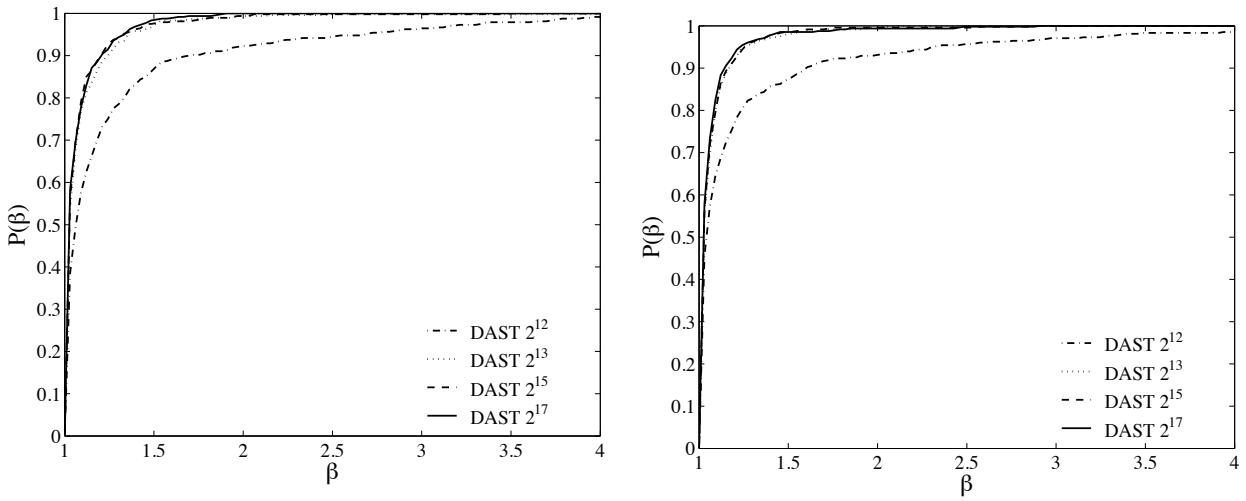

Figure 2: Sensitivity to grid size: Performance profiles for DAST with different grid sizes. All grid sizes are equally good provided they are large enough.

		AST		DAST		Cadzow	
	n	MSE	time (s)	MSE	time (s)	MSE	time (s)
Equispaced	200	0.76	2.78	0.71	0.01	1.90	0.16
	400	0.47	16.26	0.64	0.02	0.95	0.44
	800	0.28	112.57	0.30	0.05	0.39	1.50
	1600	-	-	0.25	0.12	0.28	6.44
	3200	-	-	0.08	0.34	0.15	48.6
Random	200	1.13	2.87	1.32	0.01	1.83	0.22
	400	0.78	12.32	0.57	0.03	1.53	0.88
	800	0.32	112.67	0.41	0.07	0.51	1.76
	1600	-	-	0.16	0.11	0.29	6.31
	3200	-	-	0.09	0.25	0.14	48.1

Figure 3: Benchmark comparisons of the different algorithms. In all experiments, a signal with 15 sinusoids waves with random phase and unit amplitude were added together and then Gaussian noise with variance 10 was added to each time sample. This table gives the mean-square-error and timing performance for AST DAST, and Cadzow. DAST is run with N being the smallest power of 2 that is greater than $5n$. AST always returns the lowest mean-squared error, but it the most computationally intensive. DAST always returns better mean-squared error than Cadzow, and is much faster.

algorithm, and future work will be devoted to improving its convergence rate. The dominant computation cost is the eigenvalue computation. However, AST consistently returns the lowest mean-squared error. DAST is the fastest algorithm. In these experiments, we chose the number of grid points N to be the smallest power of 2 that exceeded $5n$, where n was the number of samples. With this setting, we were able to achieve considerably lower error than Cadzow's method in a fraction of the time

6 Conclusion and Future Work

Appealing directly to the natural atomic norm formulation of line spectral estimation provided several advantages over prior approaches. By performing the analysis in the continuous domain we were able to derive simple closed form rates using fairly straightforward techniques. We were able to achieve these rates via a convex program whose iterations resemble that of prior heuristics, but where we are guaranteed convergence to a unique optimal solution. Even when we pursued a discretization, we only grid the unit circle at the very end of our analysis and determine the loss incurred from discretization. This approach allowed us to circumvent some of the more complicated theoretical arguments that arise when using concepts from compressed sensing or random matrix theory.

This work provides several interesting possible future directions, both in line spectral estimation and in signal procession in general. We conclude with a short outline of some of the possibilities.

Fast Rates Determining checkable conditions on the cones in Section 2.1 for the atomic norm problem is a major open problem. Our experiments suggest that when the frequencies are spread out, DAST performs much better with a slightly larger regularization parameter. This suggests the

fast rate developed in Section 2.1 may be active for some signals and noise regimes. Determining concrete conditions on the signal $x^\star$ that ensure this fast rate require techniques for estimating the parameter ϕ in (2.10). Such an investigation should be accompanied by a determination of the minimax rates for line spectral estimation. Such minimax rates would shed further light on the rates achievable for line spectral estimation.

Moments Supported Inside the Disk Our work also naturally extends to moment problems where the atomic measures are supported on the unit disk in the complex plane. These problems arise naturally in controls and systems theory and include model order reduction, system identification, and control design. Applying the standard program developed in Section 2 provides a new look at these classic operator theory problems in control theory. It would be of significant importance to develop specialized atomic-norm denoising algorithms for control theoretic problems. Such an approach could yield novel statistical bounds for estimation of rational functions and $\mathcal{H}_\infty$-norm approximations.

Other Denoising Models Our abstract denoising results in Section 2 apply to any atomic models and it is worth investigating their applicability for other models in statistical signal processing. For instance, it might be possible to pose a scheme for denoising a signal corrupted by multipath reflections. Here, the atoms might be all time and frequency shifted versions of some known signal. It remains to be seen what new insights in statistical signal processing can be gleaned from our unified approach to denoising.

Acknowledgements

The authors would like to thank Vivek Goyal, Parikshit Shah, and Joel Tropp for many helpful conversations and suggestions on improving this manuscript. This work was supported in part by NSF Award CCF-1139953 and ONR Award N00014-11-1-0723.

References

[1] R. Carriere and R. Moses, "High resolution radar target modeling using a modified Prony estimator," *Antennas and Propagation, IEEE Transactions on*, vol. 40, no. 1, pp. 13–18, 1992.

[2] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *Signal Processing Magazine, IEEE*, vol. 13, no. 4, pp. 67–94, 1996.

[3] L. Borcea, G. Papanicolaou, C. Tsogka, and J. Berryman, "Imaging and time reversal in random media," *Inverse Problems*, vol. 18, p. 1247, 2002.

[4] I. Maravic, J. Kusuma, and M. Vetterli, "Low-sampling rate UWB channel characterization and synchronization," *Journal of Communications and Networks*, vol. 5, no. 4, pp. 319–327, 2003.

[5] V. Viti, C. Petrucci, and P. Barone, "Prony methods in NMR spectroscopy," *International Journal of Imaging Systems and Technology*, vol. 8, no. 6, pp. 565–571, 1997.

[6] Z. Leonowicz, T. Lobos, and J. Rezmer, "Advanced spectrum estimation methods for signal analysis in power electronics," *Industrial Electronics, IEEE Transactions on*, vol. 50, pp. 514 – 519, june 2003.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, 1986.

[8] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 7, pp. 984–995, 1989.

[9] Y. Hua and T. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 5, pp. 814–824, 2002.

[10] P. Stoica, "List of references on spectral line analysis," *Signal Processing*, vol. 31, no. 3, pp. 329–340, 1993.

[11] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, "The convex geometry of linear inverse problems," *Arxiv preprint arXiv:1012.0621*, December 2010.

[12] E. Candès, J. Rombergh, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[13] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[14] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[15] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, pp. 129–159, 2001.

[16] A. Megretski, "Positivity of trigonometric polynomials," in *Proceedings of the 42nd IEEE Conference on Decision and Control*, vol. 4, pp. 3814–3817, 2003.

[17] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1997.

[18] S. Boyd, N. Parikh, B. P. E. Chu, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, December 2011.

[19] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2479–2493, 2009.

[20] T. Blu, P. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, "Sparse sampling of signal innovations," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 31–40, 2008.

[21] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal of Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[22] D. Donoho, "De-noising by soft-thresholding," *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 613–627, 1995.

[23] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[24] T. Lai and H. Robbins, "Maximally dependent random variables," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 73, no. 2, p. 286, 1976.

[25] E. Greenshtein and Y. Ritov, "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization," *Bernoulli*, vol. 10, no. 6, pp. 971–988, 2004.

[26] S. van de Geer and P. Bühlmann, "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[27] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[28] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and Banach spaces," in *Handbook on the Geometry of Banach spaces* (W. B. Johnson and J. Lindenstrauss, eds.), pp. 317–366, Elsevier Scientific, 2001.

[29] B. A. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*. Netherlands: Springer, 2007.

[30] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999.

[31] K. C. Toh, M. Todd, and R. H. Tütüncü, *SDPT3: A MATLAB software package for semidefinite-quadratic-linear programming*. Available from `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`.

[32] S. Kay and S. Marple Jr, "Spectrum analysis—a modern perspective," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1380–1419, 1981.

[33] M. Kahn, M. Mackisack, M. Osborne, and G. Smyth, "On the consistency of Prony's method and related algorithms," *Journal of Computational and Graphical Statistics*, vol. 1, no. 4, pp. 329–349, 1992.

[34] J. Cadzow, "Signal enhancement-a composite property mapping algorithm," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 1, pp. 49–62, 2002.

[35] M. Duarte and R. Baraniuk, "Spectral compressive sensing," Available from `http://dsp.rice.edu/cs`, Preprint, 2010.

[36] Y. Chi, A. Pezeshki, L. Scharf, and R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 3930–3933, IEEE, 2009.

[37] E. Dolan and J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, no. 2, pp. 201–213, 2002.

[38] A. Schaeffer, "Inequalities of A. Markoff and S. Bernstein for polynomials and related functions," *Bull. Amer. Math. Soc*, vol. 47, pp. 565–579, 1941.

A Proofs

A.1 Optimality Conditions

A.1.1 Proof of Lemma 4

Proof. The function $f(x) = \frac{1}{2}\|y - x\|_2^2 + \tau\|x\|_{\mathcal{A}}$ is minimized at $\hat{x}$, if for all $\alpha \in (0, 1)$ and all x,

$$\begin{aligned} & f(\hat{x} + \alpha(x - \hat{x})) && \geq f(\hat{x}) \\ \iff \quad & \alpha^{-1}\tau\left(\|\hat{x} + \alpha(x - \hat{x})\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}}\right) && \geq \langle y - \hat{x}, x - \hat{x}\rangle - \frac{1}{2}\alpha\|x - \hat{x}\|_2^2 \end{aligned} \tag{A.1}$$

Since $\|\cdot\|_{\mathcal{A}}$ is convex, we have

$$\|x\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}} \geq \alpha^{-1}\left(\|\hat{x} + \alpha(x - \hat{x})\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}}\right),$$

for all x and for all $\alpha \in (0, 1)$. Thus, by letting $\alpha \to 0$ in (A.1), we note that $\hat{x}$ minimizes $f(x)$ only if, for all x,

$$\tau\left(\|x\|_{\mathcal{A}} - \|\hat{x}\|_{\mathcal{A}}\right) \geq \langle y - \hat{x}, x - \hat{x}\rangle. \tag{A.2}$$

However if (A.2) holds, then, for all x

$$\frac{1}{2}\|y-x\|_2^2+\tau\|x\|_{\mathcal{A}} \geq \frac{1}{2}\|y-\hat{x}+(\hat{x}-x)\|_2^2+\langle y-\hat{x}, x-\hat{x}\rangle+\tau\|\hat{x}\|_{\mathcal{A}}$$
$$\implies f(x) \geq f(\hat{x}).$$

Thus, (A.2) is necessary and sufficient for $\hat{x}$ to minimize $f(x)$.

Note. The condition (A.2) simply says that $\tau^{-1}(y-\hat{x})$ is in the subgradient of $\|\cdot\|_{\mathcal{A}}$ at $\hat{x}$ or equivalently that $0 \in \partial f(\hat{x})$.

We can rewrite (A.2) as

$$\tau\|\hat{x}\|_{\mathcal{A}}-\langle y-\hat{x}, \hat{x}\rangle \leq \inf_x \{\tau\|x\|_{\mathcal{A}}-\langle y-\hat{x}, x\rangle\} \tag{A.3}$$

But by definition of the dual atomic norm,

$$\sup_x \{\langle z, x\rangle - \|x\|_A\} = I_{\{w:\|w\|_{\mathcal{A}}^* \leq 1\}}(z) = \begin{cases} 0 & \|z\|_{\mathcal{A}}^* \leq 1 \\ \infty & \text{otherwise.} \end{cases} \tag{A.4}$$

where $I_A(\cdot)$ is the convex indicator function. Using this in (A.3), we find that $\hat{x}$ is a minimizer if and only if $\|y-\hat{x}\|_{\mathcal{A}}^* \leq \tau$ and $\langle y-\hat{x}, \hat{x}\rangle \geq \tau\|\hat{x}\|_{\mathcal{A}}$. This proves the theorem. □

A.1.2 Proof of Lemma 5

Proof. We can rewrite the primal problem (1.1) as a constrained optimization problem:

$$\begin{aligned} &\underset{x,u}{\text{minimize}}\ \frac{1}{2}\|y-x\|_2^2+\|u\|_{\mathcal{A}} \\ &\text{subject to } u = x. \end{aligned}$$

Now, we can introduce the Lagrangian function

$$L(x,u,z) = \frac{1}{2}\|y-x\|_2^2+\|u\|_{\mathcal{A}}+\langle z, x-u\rangle.$$

so that the dual function is given by

$$\begin{aligned} g(z) &= \inf_{x,u} L(x,u,z) \\ &= \inf_x \left(\frac{1}{2}\|y-x\|_2^2+\langle z, x\rangle\right) + \inf_u (\tau\|u\|_{\mathcal{A}}-\langle z, u\rangle) \\ &= \frac{1}{2}\left(\|y\|_2^2-\|y-z\|_2^2\right) - I_{\{w:\|w\|_{\mathcal{A}}^* \leq \tau\}}(z). \end{aligned}$$

where the first infimum follows by completing the squares and the second infimum follows from (A.4). Thus the dual problem of maximizing $g(z)$ can be written as in (5).

The solution to the dual problem is the unique projection $\hat{z}$ of y on to the closed convex set $C = \{z : \|z\|_{\mathcal{A}}^* \leq \tau\}$. By projection theorem for closed convex sets, $\hat{z}$ is a projection of y onto C if and only if $\hat{z} \in C$ and $\langle z-\hat{z}, y-\hat{z}\rangle \leq 0$ for all $z \in C$, or equivalently if $\langle \hat{z}, y-\hat{z}\rangle \geq \sup_z \langle z, y-\hat{z}\rangle = \tau\|y-\hat{z}\|_{\mathcal{A}}$. These conditions are satisfied for $\hat{z} = y-\hat{x}$ where $\hat{x}$ minimizes $f(x)$ by Lemma 4. Now

the proof follows by the substitution $\hat{z} = y - \hat{x}$ in the previous lemma. The absence of duality gap can be obtained by noting that the primal objective function at $\hat{x}$,

$$f(\hat{x}) = \frac{1}{2}\|y - \hat{x}\|_2^2 + \langle \hat{z}, \hat{x} \rangle = \frac{1}{2}\|\hat{z}\|_2^2 + \langle \hat{z}, \hat{x} \rangle = g(\hat{z}).$$

□

A.2 Fast Rate Calculations

We first prove the following

Proposition 11. Let $\mathcal{A} = \{\pm e_1, \ldots, \pm e_n\}$, be the set of signed canonical unit vectors in $\mathbb{R}^n$. Suppose $x^\star \in \mathbb{R}^n$ has k nonzeros. Then $\phi_\gamma(x^\star, \mathcal{A}) \geq \frac{(1-\gamma)}{2\sqrt{k}}$.

Proof. Let $z \in C_\gamma(x^\star, \mathcal{A})$. For some $\alpha > 0$ we have,

$$\|x^\star + \alpha z\|_1 \leq \|x^\star\|_1 + \gamma \|\alpha z\|_1$$

In the above inequality, set $z = z_T + z_{T^c}$ where z_T are the components on the support of T and z_{T^c} are the components on the complement of T. Since $x^\star + z_T$ and z_{T^c} have disjoint supports, we have,

$$\begin{aligned} \|x^\star + \alpha z_T\|_1 + \alpha \|z_{T^c}\|_1 &\leq \|x^\star\|_1 + \gamma \|\alpha z_T\|_1 + \gamma \|\alpha z_{T^c}\|_1 \\ \Rightarrow \|z_{T^c}\|_1 &\leq \frac{1+\gamma}{1-\gamma} \|z_T\|_1 \end{aligned}$$

i.e., z satisfies the null space property with a constant of $\frac{1+\gamma}{1-\gamma}$. Thus,

$$\|z\|_1 \leq \frac{2}{1-\gamma}\|z_T\|_1 \leq \frac{2\sqrt{k}}{1-\gamma}\|z\|_2$$

This gives the desired lower bound. □

Now we can turn to the case of low rank matrices.

Proposition 12. Let $\mathcal{A}$ be the manifold of unit norm rank-1 matrices in $\mathbb{C}^{n \times n}$. Suppose $X^\star \in \mathbb{C}^{n \times n}$ has rank r. Then $\phi_\gamma(X^\star, \mathcal{A}) \geq \frac{1-\gamma}{2\sqrt{2r}}$.

Proof. Let $U \Sigma V^H$ be a singular value decomposition of $X^\star$ with $U \in \mathbb{C}^{n \times r}$, $V \in \mathbb{C}^{n \times r}$ and $\Sigma \in \mathbb{C}^{r \times r}$. Define the subspaces

$$\begin{aligned} T &= \{UX + YV^H \ : \ X, Y \in \mathbb{C}^{n \times r}\} \\ T_0 &= \{UMV^H \ : \ M \in \mathbb{C}^{r \times r}\} \end{aligned}$$

and let $\mathcal{P}_{T_0}$, $\mathcal{P}_T$, and $\mathcal{P}_{T^\perp}$ be projection operators that respectively map onto the subspaces T_0, T, and the orthogonal complement of T. Now, if $Z \in C_\gamma(X^\star, \mathcal{A})$, then for some $\alpha > 0$, we have

$$\|X^\star + \alpha Z\|_* \leq \|X^\star\|_* + \gamma\alpha\|Z\|_* \leq \|X^\star\|_* + \gamma\alpha\|\mathcal{P}_T(Z)\|_* + \gamma\alpha\|\mathcal{P}_{T^\perp}(Z)\|_*. \tag{A.5}$$

Now note that we have

$$\|X^\star + \alpha Z\|_* \geq \|X^\star + \alpha \mathcal{P}_{T_0}(Z)\|_* + \alpha \|\mathcal{P}_{T^\perp}(Z)\|_*$$

Substituting this in (A.5), we have,

$$\|X^\star + \alpha \mathcal{P}_{T_0}(Z)\|_* + \alpha \|\mathcal{P}_{T^\perp}(Z)\|_* \leq \|X^\star\|_* + \gamma\alpha \|\mathcal{P}_T(Z)\|_* + \gamma\alpha \|\mathcal{P}_{T^\perp}(Z)\|_*.$$

Since $\|\mathcal{P}_{T_0}(Z)\|_* \leq \|\mathcal{P}_T(Z)\|_*$, we have

$$\|\mathcal{P}_{T^\perp}(Z)\|_* \leq \frac{1+\gamma}{1-\gamma} \|\mathcal{P}_T(Z)\|_*.$$

Putting these computations together gives the estimate

$$\begin{aligned}
\|Z\|_* &\leq \|\mathcal{P}_T(Z)\|_* + \|\mathcal{P}_{T^\perp}(Z)\|_* \\
&\leq \frac{2}{1-\gamma} \|\mathcal{P}_T(Z)\| \\
&\leq \frac{2\sqrt{2r}}{1-\gamma} \|\mathcal{P}_T(Z)\|_F \\
&\leq \frac{2\sqrt{2r}}{1-\gamma} \|Z\|_F \,.
\end{aligned}$$

That is, we have $\phi_\gamma(X^\star, \mathcal{A}) \geq \frac{1-\gamma}{2\sqrt{2r}}$ as desired. □

A.3 Approximation of the Dual Atomic Norm

This section furnishes the proof that the atomic norms induced by $\mathcal{A}$ and $\mathcal{A}_N$ are equivalent. Note that the dual atomic norm of w given by

$$\|w\|_{\mathcal{A}}^* = \sqrt{n} \sup_{t \in [0,1]} \left| W_n(e^{i2\pi t}) \right|. \tag{A.6}$$

i.e., the maximum modulus of the polynomial W_n defined by

$$W_n(e^{i2\pi t}) = \frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} w_m e^{-i2\pi m t}. \tag{A.7}$$

Treating W_n as a function of t, with a slight abuse of notation, define

$$\|W_n\|_\infty := \sup_{t \in [0,1]} \left| W_n(e^{i2\pi t}) \right|.$$

We show that we can approximate the maximum modulus by evaluating W_n in a uniform grid of N points on the unit circle. To show that as N becomes large, the approximation is close to the true value, we bound the derivative of W_n using Bernstein's inequality for polynomials.

Theorem 13 (Bernstein, See, for example [38])**.** Let p_n be any polynomial of degree n with complex coefficients. Then,

$$\sup_{|z|\le 1} |p'(z)| \le n \sup_{|z|\le 1} |p(z)|.$$

Note that for any $t, s \in [0,1]$, we have

$$\begin{aligned}
\left|W_n(e^{i2\pi t})\right| - \left|W_n(e^{i2\pi s})\right| &\le \left|e^{i2\pi t} - e^{i2\pi s}\right| \|W_n'\|_\infty \\
&= 2|\sin(2\pi(t-s))| \|W_n'\|_\infty \\
&\le 4\pi(t-s)\|W_n'\|_\infty. \\
&\le 4\pi n(t-s)\|W_n\|_\infty \qquad \text{(by Bernstein's theorem).}
\end{aligned}$$

Letting s take any of the N values $0, 1/N, \ldots, (N-1)/N$, we see that,

$$\|W_n\|_\infty \le \max_{m=0,\ldots,N-1} \left|W_n\left(e^{i2\pi m/N}\right)\right| + \frac{2\pi n}{N}\|W_n\|_\infty.$$

Since the maximum on the grid is a lower bound for maximum modulus of W_n, we have

$$\begin{aligned}
\max_{m=0,\ldots,n-1} \left|W_n\left(e^{i2\pi m/N}\right)\right| \le \|W_n\|_\infty &\le \left(1 - \frac{2\pi n}{N}\right)^{-1} \max_{m=0,\ldots,N-1} \left|W_n\left(e^{i2\pi m/N}\right)\right| \\
&\le \left(1 + \frac{4\pi n}{N}\right) \max_{m=0,\ldots,N-1} \left|W_n\left(e^{i2\pi m/N}\right)\right|. \qquad \text{(A.8)}
\end{aligned}$$

Thus, for every w,

$$\|w\|_{\mathcal{A}_N}^* \le \|w\|_{\mathcal{A}}^* \le \left(1 - \frac{2\pi n}{N}\right)^{-1} \|w\|_{\mathcal{A}_N}^* \qquad \text{(A.9)}$$

or equivalently, for every x,

$$\left(1 - \frac{2\pi n}{N}\right) \|x\|_{\mathcal{A}_N} \le \|x\|_{\mathcal{A}} \le \|x\|_{\mathcal{A}_N} \qquad \text{(A.10)}$$

A.4 Dual Atomic Norm Bounds

This section derives non asymptotic upper and lower bounds for the expected dual norm of gaussian noise vectors, which are asymptotically tight unto log log factors. Recall that the dual atomic norm of w is given by $\sqrt{n}\sup_{t\in[0,1]}|W_t|$ where

$$W_t = \frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} w_m e^{-i2\pi m t}.$$

The covariance function of W_t is

$$\mathbb{E}\left[W_t W_s^*\right] = \frac{1}{n}\sum_{m=0}^{n-1} \exp(2\pi m(t-s)) = e^{\pi(n-1)(t-s)} \frac{\sin(n\pi(t-s))}{n\sin(\pi(t-s))}.$$

Thus, the n samples $\left\{W_{m/n}\right\}_{m=0}^{n-1}$ are uncorrelated and thus independent because of their joint gaussianity. This gives a simple non-asymptotic lower bound using the known result for maximum value of n independent gaussian random variables [24] whenever $n > 5$:

$$\mathbb{E}\left[\sup_{t\in T}|W_t|\right] \geq \mathbb{E}\left[\max_{m=0,\ldots,n-1}\mathrm{Re}\left(W_{m/n}\right)\right] = \sqrt{\log(n) - \tfrac{\log\log(n)+\log(4\pi)}{2}}.$$

We will show that the lower bound is asymptotically tight neglecting $\log\log$ terms. Since the dual norm induced by $\mathcal{A}_N$ approximates the dual norm induced by $\mathcal{A}$, (See A.3), it is sufficient to compute an upper bound for $\|w\|^*_{\mathcal{A}_N}$. Note that $|W_t|^2$ has a chi-square distribution since W_t is a Gaussian process. We establish a simple lemma about the maximum of chi-square distributed random variables.

Lemma 14. Let $x_1, \ldots, x_N$ be complex gaussians with unit variance. Then,

$$\mathbb{E}\left[\max_{1\leq i\leq N}|x_i|\right] \leq \sqrt{\log(N)+1}.$$

Proof. Let $x_1, \ldots, x_N$ be complex Gaussians with unit variance: $\mathbb{E}[|x_i|^2] = 1$. Note that $2|x_i|^2$ is a chi-squared random variable with two degrees of freedom. Using Jensen's inequality, also observe that

$$\mathbb{E}\left[\max_{1\leq i\leq N}|x_i|\right] \leq \mathbb{E}\left[\max_{1\leq i\leq N}|x_i|^2\right]^{1/2} \leq \tfrac{1}{\sqrt{2}}\,\mathbb{E}\left[\max_{1\leq i\leq N}2|x_i|^2\right]^{1/2} \tag{A.11}$$

Now let $z_1, \ldots, z_n$ be chi-squared random variables with 2 degrees of freedom. Then we have

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq i\leq N} z_i\right] &= \int_0^\infty P\left[\max_{1\leq i\leq N} z_i \geq t\right] dt \\
&\leq \delta + \int_\delta^\infty P\left[\max_{1\leq i\leq N} z_i \geq t\right] dt \\
&\leq \delta + N\int_\delta^\infty P\left[z_1 \geq t\right] dt \\
&= \delta + N\int_\delta^\infty \exp(-t/2)dt \\
&= \delta + 2N\exp(-\delta/2)
\end{aligned}$$

Setting $\delta = 2\log(N)$ gives

$$\mathbb{E}\left[\max_{1\leq i\leq N} z_i\right] \leq 2\log N + 2\,.$$

Plugging this estimate into (A.11) gives

$$\mathbb{E}\left[\max_{1\leq i\leq N}|x_i|\right] \leq \sqrt{\log N + 1}$$

□

Using Lemma 14, we can compute

$$\|w\|^*_{\mathcal{A}_N} = \sqrt{n}\max_{m=0,\ldots,N-1}\left|W_n\left(e^{i2\pi m/N}\right)\right| \leq \sigma\sqrt{n\left(\log N + 1\right)}$$

Plugging in $N = 4\pi n\log(n)$ and using (A.6) and (A.8) establishes a tight upper bound.