Latent Variable Bayesian Models for Promoting Sparsity

David P. Wipf, Bhaskar D. Rao, *Fellow, IEEE*, and Srikantan Nagarajan

***Abstract*—Many practical methods for finding maximally sparse coefficient expansions involve solving a regression problem using a particular class of concave penalty functions. From a Bayesian perspective, this process is equivalent to maximum *a posteriori* (MAP) estimation using a sparsity-inducing prior distribution (Type I estimation). Using variational techniques, this distribution can always be conveniently expressed as a maximization over scaled Gaussian distributions modulated by a set of latent variables. Alternative Bayesian algorithms, which operate in latent variable space leveraging this variational representation, lead to sparse estimators reflecting posterior information beyond the mode (Type II estimation). Currently, it is unclear how the underlying cost functions of Type I and Type II relate, nor what relevant theoretical properties exist, especially with regard to Type II. Herein a common set of auxiliary functions is used to conveniently express both Type I and Type II cost functions in either coefficient or latent variable space facilitating direct comparisons. In coefficient space, the analysis reveals that Type II is exactly equivalent to performing standard MAP estimation using a particular class of dictionary- and noise-dependent, *nonfactorial* coefficient priors. One prior (at least) from this class maintains several desirable advantages over all possible Type I methods and utilizes a novel, nonconvex approximation to the ℓ_0 norm with most, and in certain quantifiable conditions all, local minima smoothed away. Importantly, the global minimum is always left unaltered unlike standard ℓ_1-norm relaxations. This ensures that any appropriate descent method is guaranteed to locate the maximally sparse solution.**

***Index Terms*—Bayesian learning, compressive sensing, latent variable models, source localization, sparse priors, sparse representations, underdetermined inverse problems.**

I. Introduction

HERE we will be concerned with the generative model

$$\boldsymbol{y} = \Phi \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{1}$$

where $\Phi \in \mathbb{R}^{n \times m}$ is a dictionary of unit ℓ_2-norm basis vectors or features, $\boldsymbol{x}$ is a vector of unknown coefficients we would like to estimate, $\boldsymbol{y}$ is the observed signal, and $\boldsymbol{\varepsilon}$ represents noise or modeling errors often assumed to be Gaussian. In many practical situations where large numbers of features are present relative to the signal dimension, implying $m > n$, the problem of estimating $\boldsymbol{x}$ is fundamentally underdetermined.

A typical remedy for this indeterminacy is to apply a penalty term into the estimation process that reflects prior, disambiguating assumptions about $\boldsymbol{x}$. This leads to the canonical regularized regression problem

$$\boldsymbol{x}_{(I)} \triangleq \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi \boldsymbol{x}\|_2^2 + \lambda \sum_i g(x_i) \tag{2}$$

which produces what is often called a *Type I* estimator denoted $\boldsymbol{x}_{(I)}$.[1] The first term in (2) enforces data fit (consistent with a Gaussian noise model), while $g(x_i)$ is a fixed penalty on individual coefficients and λ is a tradeoff parameter. For example, if we would like to penalize the ℓ_2 norm of $\boldsymbol{x}$, favoring minimum energy solutions, then we can choose $g(z) = z^2$.

Recently, there has been a growing interest in finding some $\hat{\boldsymbol{x}}$ characterized by a bi-partitioning of coefficients, meaning most elements equal zero (or are very small), and a few large unrestricted values, i.e., we are assuming the generative $\boldsymbol{x}$ is a sparse vector. Such solutions can be obtained by using

$$g(z) = h\left(z^2\right) \tag{3}$$

with h concave and nondecreasing on $[0, \infty)$ [27], [28]. Roughly speaking, the "more concave" h, the more sparse we expect global solutions of (2) to be. For example, with $h(z) = z$, we recover the ℓ_2 norm penalty, which is not sparse at all, while $h(z) = \sqrt{z}$ gives an ℓ_1 norm penalty, which under many circumstances is well-known to produce a $\hat{\boldsymbol{x}}$ with numerous elements (at least $m - n$) equal to exactly zero [8], [30]. In arguably the most extreme case, *maximally sparse* solutions are said to occur using $h(z) = \mathcal{I}_{(z \neq 0)}[z]$, which penalizes any deviation from zero uniformly, so once any deviation from zero exists, no additional penalty is incurred (Section I-B will discuss this penalty in more detail). Other common selections include $g(z) = |z|^p, p \in (0, 2]$ [7], [22], [28] and $g(z) = \log(|z| + \epsilon), \epsilon \geq 0$ [6], [14], [15], [19].

If we define the *a prior* distribution $p(\boldsymbol{x})$ and likelihood function $p(\boldsymbol{y}|\boldsymbol{x})$ via

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\sum_i g(x_i)\right] \text{ and}$$
$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left[-\frac{1}{2\lambda}\|\boldsymbol{y} - \Phi \boldsymbol{x}\|_2^2\right] \tag{4}$$

[1]While typically the global solution of (2) is unique, in certain cases it is possible to have multiple minimizers. We then adopt the convention that $\boldsymbol{x}_{(I)}$ is an arbitrary element of this nontrivial solution set (likewise for other $\arg\min$ or $\arg\max$ expressions in this paper).

Manuscript received November 15, 2009; revised July 12, 2010; accepted September 02, 2010. Date of current version August 31, 2011. The work of S. Nagarajan was supported in part by the NIH by Grants R01DC04855 and R01DC006435. The work of D. Wipf was supported by the NIH postdoctoral fellowship F32NS061395. The work of B. Rao was supported by the NSF by Grant CCF-0830612. The material in this paper was presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hawaii, April 2007.

D. P. Wipf is with the Visual Computing Group, Microsoft Research Asia, Beijing, 100080 China (e-mail: davidwipf@gmail.com).

B. D. Rao is with the Digital Signal Processing Lab, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: brao@ucsd.edu).

S. Nagarajan is with the Biomagnetic Imaging Lab, University of California, San Francisco, CA 94143 USA (e-mail: sri@mrsc.ucsf.edu).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2011.2162174

then from a Bayesian perspective (2) is equivalent (via Bayes rule [1]) to solving the *maximum aposteriori* (MAP) estimation problem

$$\boldsymbol{x}_{(I)} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}). \tag{5}$$

At this point, the Bayesian viewpoint has essentially offered nothing new, since the posterior mode (or maximum) equals the same estimator $\boldsymbol{x}_{(I)}$ we had before. However, what if we consider alternative estimators based on $p(\boldsymbol{x}|\boldsymbol{y})$ but sensitive to posterior information beyond the mode? Using variational methods [21], we will demonstrate that it is possible to develop a broader class of *Type II* estimators that is particularly well-suited to finding maximally sparse coefficients and includes (2), and therefore (5), as a special case. We should stress at the outset that, while Bayesian methodology forms the starting point and inspiration for many of the ideas forthcoming in this paper, ultimate justification of Type II estimation techniques will be completely independent of any Bayesian formalism. Instead, our strategy is to extract the underlying cost functions that emerge from this formalism, and then analyze them abstractly in the same manner that many others have analyzed (2). This is not unlike the situation surrounding the widespread use of the ℓ_1 norm for solving underdetermined inverse problems where sparse solutions are desired. While the associated Type I algorithm can be interpreted as performing MAP estimation using a Laplacian prior, the rich theory quantifying performance guarantees is completely independent of any putative association with the Laplacian distribution. We will return to this topic in more detail in Section VI.

A. Type II Bayesian Estimation

The starting point for creating the Type II estimator involves reexpressing the prior $p(\boldsymbol{x})$ in terms of a collection of nonnegative latent variables $\boldsymbol{\gamma} \triangleq [\gamma_1, \ldots, \gamma_m]^T \in \mathbb{R}^m_+$. The latent variables dictate the structure of the prior via

$$p(\boldsymbol{x}) = \prod_{i=1}^{m} p(x_i),\ p(x_i) = \max_{\gamma_i \geq 0} \mathcal{N}(x_i; 0, \gamma_i)\varphi(\gamma_i) \tag{6}$$

where $\varphi(\gamma_i)$ is a nonnegative function[2] and $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \Sigma)$ henceforth denotes a Gaussian over $\boldsymbol{z}$ with mean $\boldsymbol{\mu}$ and covariance Σ. In a machine learning context, (6) is commonly referred to as a variational representation whose form is rooted in convex analysis and duality theory [21], and when the maximization is dropped, provides a family of strict lower bounds on $p(\boldsymbol{x})$ parameterized by $\boldsymbol{\gamma}$ [27]. Note that any prior $p(\boldsymbol{x})$, constructed via $g(x_i) = h\left(x_i^2\right)$ as in (4), with h concave and nondecreasing on $[0, \infty)$, is expressable using (6) given the appropriate φ [27]. Consequently, virtually all sparse priors (based on sparse penalties) of interest can be decomposed in this manner, including the popular Laplacian, Jeffreys, Student's t, and generalized Gaussian priors.[3]

The utility of (6) comes in forming approximations to the posterior $p(\boldsymbol{x}|\boldsymbol{y})$, or for practical reasons the joint distribution $p(\boldsymbol{x}, \boldsymbol{y}) \propto p(\boldsymbol{x}|\boldsymbol{y})$, which in turn can lead to alternative sparse estimators. For example, while computing the posterior mean of $p(\boldsymbol{x}|\boldsymbol{y})$ is intractable, given an appropriate approximation, the required integrals lead to analytic solutions. One practical option is to form a Gaussian approximation using (6) as follows.

For a fixed $\boldsymbol{\gamma}$, we obtain the approximate (unnormalized) prior

$$\hat{p}_\gamma(\boldsymbol{x}) = \prod_i \mathcal{N}(x_i; 0, \gamma_i)\varphi(\gamma_i) \tag{7}$$

which leads to the approximate (normalized) posterior

$$\hat{p}_\gamma(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})\hat{p}_\gamma(\boldsymbol{x})}{\int p(\boldsymbol{y}|\boldsymbol{x})\hat{p}_\gamma(\boldsymbol{x})d\boldsymbol{x}} = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \Sigma_x) \tag{8}$$

with

$$\begin{aligned} \boldsymbol{\mu}_x &= \Gamma\Phi^T\left(\lambda I + \Phi\Gamma\Phi^T\right)^{-1}\boldsymbol{y} \\ \Sigma_x &= \Gamma - \Gamma\Phi^T\left(\lambda I + \Phi\Gamma\Phi^T\right)^{-1}\Phi \end{aligned} \tag{9}$$

where $\Gamma \triangleq \mathrm{diag}[\boldsymbol{\gamma}]$. The key task then is to choose values for the latent variables $\boldsymbol{\gamma}$ such that, to the extent possible, $p(\boldsymbol{x}|\boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \Sigma_x)$. One useful criterion that leverages the variational represention involves solving

$$\begin{aligned} \boldsymbol{\gamma}_{(II)} &\triangleq \arg\min_{\boldsymbol{\gamma}} \int p(\boldsymbol{y}|\boldsymbol{x})\left|p(\boldsymbol{x}) - \hat{p}_\gamma(\boldsymbol{x})\right| d\boldsymbol{x} \\ &= \arg\max_{\boldsymbol{\gamma}} \int p(\boldsymbol{y}|\boldsymbol{x}) \prod_i \mathcal{N}(x_i; 0, \gamma_i)\varphi(\gamma_i) dx_i \end{aligned} \tag{10}$$

where the absolute value can be conveniently removed by virtue of the variational lower bound ($\boldsymbol{\gamma}$-independent terms are omitted). The idea behind (10) is that we would like to minimize the sum of the misaligned mass between the true prior $p(\boldsymbol{x})$ and the approximate one $\hat{p}_\gamma(\boldsymbol{x})$, but only in regions where the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ is significant. If $p(\boldsymbol{y}|\boldsymbol{x}) \approx 0$, then we do not really care if the prior approximation is poor, since the ultimate contribution of this error to the posterior distribution will be minimal (see [34, Ch. IV] for more details).

Once $\boldsymbol{\gamma}_{(II)}$ is obtained, a commonly accepted point estimate for $\boldsymbol{x}$ is the posterior mean $\boldsymbol{\mu}_x$ with $\boldsymbol{\gamma}$ set to $\boldsymbol{\gamma}_{(II)}$

$$\boldsymbol{x}_{(II)} \triangleq \Gamma_{(II)}\Phi^T\left(\lambda I + \Phi\Gamma_{(II)}\Phi^T\right)^{-1}\boldsymbol{y}. \tag{11}$$

Note that if $\boldsymbol{\gamma}_{(II)}$ is sparse, the corresponding coefficient estimate $\boldsymbol{x}_{(II)}$ will be sparse as well, consistent with our modeling assumptions. Type II is sometimes referred to as *empirical Bayes*, since we are (somewhat counterintuitively) using

[2]Here we are assuming continuity for simplicity, and so (6) will have a maximum; otherwise we require a supremum operator instead.

[3]The function $\varphi(\gamma_i)$ can either be chosen constructively to produce some prior $p(x_i)$, or alternatively, for a given sparse $p(x_i)$, the associated value of $\varphi(\gamma_i)$ can be computed using convexity results [27]. However, technically there is some ambiguity involved here in that $\varphi(\gamma_i)$ need not be unique. For example, consider a prior $p(x_i)$ composed as a maximization over two zero-mean Gaussian kernels with variances σ_1^2 and σ_2^2. In this situation, the value of $\varphi(\gamma_i)$ need only be rigidly specificed at $\varphi(\sigma_1^2)$ and $\varphi(\sigma_2^2)$; at all other points its value is constrained but need not be unique. Regardless, a natural, unique selection for $\varphi(\gamma_i)$ does exist based on the concave conjugate of h from (3). We will accept this convention for $\varphi(\gamma_i)$ and discuss how it may be computed below.

the data to empirically "learn" a prior on $\boldsymbol{x}$ [1]. Relevant Type II examples include sparse Bayesian learning (SBL) [31], [34], automatic relevance determination (ARD) [26], [35], evidence maximization [29], and methods for learning overcomplete dictionaries [17].

B. Preliminary Definitions and Problem Statement

To begin, the ℓ_0 norm is defined as

$$\|\boldsymbol{x}\|_0 \triangleq \sum_{i=1}^{m} \mathcal{I}_{(x_i \neq 0)}[x_i] \tag{12}$$

where the indicator function $\mathcal{I}_{(x_i \neq 0)}$ takes a value of 0 if $x_i = 0$ and 1 otherwise.[4] With regard to the dictionary Φ, the *spark* represents the smallest number of linearly dependent columns [13]. By definition then, $2 \leq \mathrm{spark}(\Phi) \leq n+1$. As a special case, the condition spark $(\Phi) = n+1$ is equivalent to the unique representation property from [19], which states that every subset of n columns is linearly independent. Finally, we say that Φ is *overcomplete* if $m > n$.

Turning to the problem of obtaining sparse point estimates $\hat{\boldsymbol{x}}$, we start with the most straightforward case where $\boldsymbol{\varepsilon}$ from (1) is zero. If Φ is overcomplete, then we are presented with an underdetermined inverse problem unless further assumptions are made. For example, if a vector of unknown, generating coefficients $\boldsymbol{x}_{\mathrm{gen}}$ satisfies

$$\|\boldsymbol{x}_{\mathrm{gen}}\|_0 < \mathrm{spark}(\Phi)/2 \tag{13}$$

then no other solution $\boldsymbol{x}$ can exist such that $\boldsymbol{y} = \Phi\boldsymbol{x}$ and $\|\boldsymbol{x}\|_0 \leq \|\boldsymbol{x}_{\mathrm{gen}}\|_0$ [13], [18]. Furthermore, if we assume suitable randomness on the nonzero entries of $\boldsymbol{x}_{\mathrm{gen}}$, then this result also holds almost surely under the alternative inequality

$$\|\boldsymbol{x}_{\mathrm{gen}}\|_0 < \mathrm{spark}(\Phi) - 1 \tag{14}$$

which follows from [34, Lemma 2]. Given that (13) and/or (14) hold, then recovering $\boldsymbol{x}_{\mathrm{gen}}$ is tantamount to solving

$$\boldsymbol{x}_{\mathrm{gen}} = \boldsymbol{x}_0 \triangleq \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0, \quad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}. \tag{15}$$

This cost function encourages feasible solutions $\boldsymbol{x}$ with the largest possible number of elements identically equal to zero and a few unrestricted coefficients; such solutions are often referred to as *maximally sparse*. While ideal in spirit for many applications that require exact sparsity, finding the global minimum is combinatorial (NP-hard [25]) and therefore often difficult to obtain in general. Fortunately, many Type I and Type II methods represent viable surrogates that provide tractable approximations that solve (15) in many practical situations. In Sections III and IV, we will examine the solution of (15) in much further detail. For the remainder of this paper, whenever $\boldsymbol{\varepsilon} = \boldsymbol{0}$, we will assume that $\boldsymbol{x}_{\mathrm{gen}}$ satisfies (13) or (14), and so $\boldsymbol{x}_0$ and $\boldsymbol{x}_{\mathrm{gen}}$ can be used interchangeably.

Although not the primary focus of our analysis herein, when $\boldsymbol{\varepsilon} \neq \boldsymbol{0}$, things are decidedly more nebulous. Because noise is present, we typically do not expect to represent $\boldsymbol{y}$ exactly, suggesting the relaxed optimization problem

$$\boldsymbol{x}_0(\lambda) \triangleq \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_0 \tag{16}$$

where λ is a nonnegative tradeoff parameter balancing estimation quality with sparsity, noting that in the limit as $\lambda \to 0$, the problems (15) and (16) are equivalent (the limit must be taken outside of the minimization). Unfortunately, solving (16) is also NP-hard, nor is it clear how to select λ so as to best approximate $\boldsymbol{x}_{\mathrm{gen}}$.

In this paper, we will consider the application of general Type I and Type II methods to the solution of (15) and/or (16) for the purpose of estimating $\boldsymbol{x}_{\mathrm{gen}}$. On the surface, the above developments suggest that Type I methods are much more closely related to canonical sparse recovery problems; however, we will demonstrate that Type II is quite suitable, if not advantageous, as well. In general, most of the analytical results will address solutions to (15), which lends itself more directly to theoretical inquiries. Regardless, the underlying ideas still carry over to the case where noise is present.

C. Overview

In applying the many existing variants of Type I and Type II in practice, the performance recovering sparse generative coefficients can be highly varied because of convergence issues and properties of global and local minima. Moreover, the relationship between Type I methods, which involve transparently optimizing a cost function directly in $\boldsymbol{x}$-space, and Type II approaches, which effectively operate less intuitively in $\boldsymbol{\gamma}$-space, is very ambiguous. Additionally, it is not clear with Type II how to implement extensions for handling alternative noise models or constraints such as nonnegativity, etc., because the required integrals, e.g., (8) and (10), become intractable. To address all of these issues, this paper will investigate the cost functions that emerge from latent variable characterizations of sparse priors, with a particular emphasis on special cases of Type II that perform exceedingly well on sparse estimation problems.

Starting in Section II we will demonstrate a fundamental duality between Type I and Type II sparse estimation methods, showing that both can be expressed in *either* $\boldsymbol{x}$-space or $\boldsymbol{\gamma}$-space with a common underlying set of objective functions uniting all possible methods. This perspective facilitates direct comparisons and demonstrates that, for all methods, optimization or additional/alternative solution constraints can be implemented in either space depending on the application. Perhaps surprisingly, the analysis also reveals that Type I is a special limiting case of Type II, suggesting that the broader Type II may offer an avenue for improvement.

Because Type I has been thoroughly analyzed by others in a variety of contexts, we focus the next two sections on properties of Type II with respect to finding maximally sparse solutions. Working in coefficient space, Type II is shown to be exactly equivalent to standard MAP estimation using a large class of potentially dictionary- and noise-dependent, *nonfactorial* coefficient priors (meaning a prior which cannot be expressed in the

[4]Note that $\|\boldsymbol{x}\|_0$, because it does not satisfy the required axioms, is not technically a norm.

factored form $p(\boldsymbol{x}) = \prod_i p(x_i)$). This is unlike Type I, which is always restricted to factorial priors independent of λ and Φ. In Section III we demonstrate that one prior (at least) from this class maintains several desirable advantages over all possible Type I methods in finding maximally sparse solutions. In particular, it utilizes a novel, nonconvex approximation to the ℓ_0 norm with most local minima smoothed away; importantly, the global minimum is left unaltered. This prior can be viewed in some sense as a dual form of sparse Bayesian learning (SBL) [31] or automatic relevance determination (ARD) [26].

Necessary conditions for local minima are derived and depicted geometrically in Section IV providing insight into the best- and worst-case performance. Additionally, we describe how the distribution of nonzero generating coefficients affects the sparse recovery problem, defining a limited regime whereby Type II is unequivocally superior to any possible Type I approach and guaranteed to find maximally sparse solutions using a simple iterative algorithm.

Section V contains empirical experiments comparing an iterative reweighted ℓ_2-norm implementation of SBL (Type II) with basis pursuit (BP) and orthogonal matching pursuit (OMP) (Type I) recovering sparse coefficients as the dictionary size, sparsity level, and coefficient distribution are varied. In all cases, Type II is significantly more successful than Type I, even in the worst-case regime for Type II. Finally, Section VI has concluding remarks and provides an abstract perspective on the success of Type II that deviates somewhat from the underlying Bayesian model. All proofs are contained in the Appendix so as not to disrupt the flow of the main text.

Overall, Type I methods, especially when viewed as forms of sparse penalized regression, are much more prevalent in the statistics and signal processing community in the context of sparse linear inverse problems. By demonstrating a fundamental duality with Type II methods as well as some of the advantages of the associated broader class of underlying cost functions, we hope to inspire alternative means of estimating sparse solutions. Portions of this work have previously appeared in conference proceedings [35], [38], [39].

II. Duality and Unification

Previously we have described how Type I methods minimize a cost function in $\boldsymbol{x}$-space while Type II approaches operate in $\boldsymbol{\gamma}$-space. This distinction presently makes direct comparisons difficult. However, this section will demonstrate a fundamental duality between Type I and Type II. In particular, we will show how the cost functions associated with both approaches can be expressed either in $\boldsymbol{x}$-space or in $\boldsymbol{\gamma}$-space. This duality has several important consequences. First, it facilitates straightforward comparisons of the underlying cost functions and elucidates actual differences with respect to sparse estimation problems. Ultimately it will contribute substantial clarity regarding exactly how the less transparent Type II operates, leading to a variety of theoretical results linking Type I and Type II.

Secondly, it naturally allows us to impose constraints in either $\boldsymbol{\gamma}$-space or $\boldsymbol{x}$-space, depending on the application. For example, in nonnegative sparse coding applications, we require the $\boldsymbol{x}$-space constraint $\boldsymbol{x} \geq 0$ [4]. In contrast, to implement certain iterative reweighting optimization schemes designed to avoid local minima, or to allow for soft bounds on $\boldsymbol{x}$, we can include various variance constraints in $\boldsymbol{\gamma}$-space, e.g., $\boldsymbol{\gamma} \geq \epsilon$, as described in [7] and [36]. Finally, this duality suggests alternative means of constructing cost functions and algorithms for promoting sparsity. Other benefits, such as learning tradeoff parameters and quantifying sparsity with alternative data-fit terms, are discussed in [37].

To begin, we will first reexpress the Type I objective from (5) in an equivalent $\boldsymbol{\gamma}$-space representation in Section II-A. A byproduct of this analysis will be the demonstration that the Type I cost function is a special limiting case of Type II. Later, Section II-B will recast the Type II cost from (10) in $\boldsymbol{x}$-space.

A. Cost Functions in $\boldsymbol{\gamma}$-Space

Computing the integral from (10), which is a standard convolution of Gaussians for which analytic solutions exist, and then applying a $-2\log(\cdot)$ transformation gives the Type II cost function in $\boldsymbol{\gamma}$-space

$$\begin{aligned}\mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma}) &\triangleq -2\log\int p(\boldsymbol{y}|\boldsymbol{x})\prod_i \mathcal{N}(x_i;0,\gamma_i)\varphi(\gamma_i)dx_i \\ &\equiv \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \log|\Sigma_y| + \sum_{i=1}^{m} f(\boldsymbol{\gamma}_i) \end{aligned} \tag{17}$$

where

$$f(\gamma_i) \triangleq -2\log\varphi(\gamma_i). \tag{18}$$

and

$$\Sigma_y \triangleq \lambda I + \Phi\Gamma\Phi^T. \tag{19}$$

Here Σ_y represents the covariance of the data $\boldsymbol{y}$ conditioned on the latent variables $\boldsymbol{\gamma}$ (sometimes referred to as hyperparameters) after the unknown coefficients $\boldsymbol{x}$ have been integrated out. The function is then minimized to find some $\boldsymbol{\gamma}_{(II)}$ and the point estimate for $\boldsymbol{x}_{(II)}$ is subsequently obtained via (11). Note that the data-dependent term in (17) can be shown to be convex in $\boldsymbol{\gamma}$, while the log-det term is concave in $\boldsymbol{\gamma}$, and so in general $\mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})$ may have multiple unconnected local minima.

In contrast, Type I coefficient estimates $\boldsymbol{x}_{(I)}$ are obtained by minimizing

$$\begin{aligned}\mathcal{L}^{x}_{(I)}(\boldsymbol{x}) &\triangleq -2\log p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \\ &\equiv \|\boldsymbol{y}-\Phi\boldsymbol{x}\|_2^2 + \lambda\sum_i g(x_i)\end{aligned} \tag{20}$$

with g defined via (3). These estimates can be obtained from an analogous optimization procedure in $\boldsymbol{\gamma}$-space as follows:

Theorem 1: Define the $\boldsymbol{\gamma}$-space cost function

$$\mathcal{L}^{\gamma}_{(I)}(\boldsymbol{\gamma}) \triangleq \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \sum_{i=1}^{m} f_{(I)}(\gamma_i) \tag{21}$$

with $f_{(I)}(\gamma_i) \triangleq \log\gamma_i + f(\gamma_i)$ and $\boldsymbol{\gamma} \geq 0$. Then $\boldsymbol{\gamma}_{(I)}$ is a global minimum of (21) iff

$$\begin{aligned}\boldsymbol{x}_{(I)} &= \Gamma_{(I)}\Phi^T\left(\lambda I + \Phi\Gamma_{(I)}\Phi^T\right)^{-1}\boldsymbol{y},\\ \Gamma_{(I)} &= \operatorname{diag}\left[\boldsymbol{\gamma}_{(I)}\right]\end{aligned} \tag{22}$$

is a global minimum of (20). The correspondence extends to local minima as well: $\boldsymbol{\gamma}_*$ is a local minimum of (21) iff $\boldsymbol{x}_* = \Gamma_*\Phi^T\left(\lambda I + \Phi\Gamma_*\Phi^T\right)^{-1}\boldsymbol{y}$ is a local minimum of (20).

So Type I methods can always be interpreted as minimizing the Type II-like cost function $\mathcal{L}^{\gamma}_{(I)}(\boldsymbol{\gamma})$ in $\boldsymbol{\gamma}$-space, albeit without the log-det term in (17), and with a particular selection for f, i.e., $f_{(I)}$.[5]

Several points are worth mentioning with respect to this result. First, if g is known, as opposed to f or φ directly, then $f_{(I)}$ can be computed using the concave conjugate [2, Sect. 3.3] of $h(z) = g(\sqrt{z}), z \geq 0$. When composed with the reciprocal function γ_i^{-1}, this gives

$$-f_{(I)}(\gamma_i) = \min_{z\geq 0}\frac{z}{\gamma_i} - h(z). \tag{23}$$

For example, using $g(z) = |z|^p$ gives the ℓ_p-quasi-norm penalized minimization problem

$$\boldsymbol{x}_{(I)} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_p^p, p \in (0,2). \tag{24}$$

The analogous problem in $\boldsymbol{\gamma}$-space, using (23) to compute $f_{(I)}(\gamma_i)$, becomes

$$\boldsymbol{\gamma}_{(I)} = \arg\min_{\boldsymbol{\gamma}} \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \frac{(2-p)}{p}\left(\frac{p}{2}\right)^{\frac{2}{2-p}}\sum_{i=1}^{m}\gamma_i^{\frac{p}{2-p}}. \tag{25}$$

Secondly, when viewed in $\boldsymbol{\gamma}$-space, it is straightforward to add variance constraints to any Type I objective where appropriate, e.g., minimize $\mathcal{L}^{\gamma}_{(I)}(\boldsymbol{\gamma})$ with $\gamma_i \in [\epsilon, \infty)$ for all i. If ϵ is gradually reduced during optimization, we have observed that local minima to $\mathcal{L}^{\gamma}_{(I)}(\boldsymbol{\gamma})$ can often be avoided. This notion is very similar in spirit to the algorithm from [7] yet more straightforward when viewed in $\boldsymbol{\gamma}$-space. In general, convergence proofs, complementary analyses, and alternative optimization strategies are possible using this perspective. It also provides a particularly useful route for estimating the tradeoff parameter λ, which as a noise variance, is more naturally handled in the $\boldsymbol{\gamma}$-space of variances [37].

Finally, the $\boldsymbol{\gamma}$-space cost function $\mathcal{L}^{\gamma}_{(I)}(\boldsymbol{\gamma})$ can be interpreted as a special (limiting) case of the Type II cost function (17), which leads to the following.

Corollary 1: Let $\boldsymbol{x}_{(I)}$ denote Type I coefficients obtained by minimizing (20) or (21) with λ and f set to some arbitrary $\bar{\lambda}$ and $\bar{f}$. Additionally, let $\boldsymbol{x}^{\alpha}_{(II)}$ denote coefficients obtained by implementing the Type II procedure with $\lambda := \alpha^{-1}\bar{\lambda}$ and $f(\cdot) := \alpha\log[\alpha(\cdot)] + \alpha\bar{f}[\alpha(\cdot)]$. Then $\boldsymbol{x}_{(I)} = \lim_{\alpha\longrightarrow\infty}\boldsymbol{x}^{\alpha}_{(II)}$.

In conclusion then, by choosing the appropriate sparse prior, and therefore the function f, any Type I solution can be viewed as a limiting case of Type II. This also implies that the less commonly adopted Type II framework offers a wider variety of potential cost functions, relative to Type I, for tackling sparse estimation problems. Consequently, as we will argue in later sections, a selection from this larger set may possibly lead to improved performance.

B. Cost Functions in $\boldsymbol{x}$-space

Borrowing ideas from the previous section, we now demonstrate a simple means of computing $\boldsymbol{x}_{(II)}$ directly in $\boldsymbol{x}$-space.

Theorem 2: Define the $\boldsymbol{x}$-space cost function

$$\mathcal{L}^{x}_{(II)}(\boldsymbol{x}) \triangleq \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda\mathfrak{g}_{(II)}(\boldsymbol{x}) \tag{26}$$

with penalty

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) \triangleq \min_{\boldsymbol{\gamma}\geq 0}\sum_i\frac{x_i^2}{\gamma_i} + \log|\Sigma_y| + \sum_i f(\gamma_i). \tag{27}$$

Then

$$\boldsymbol{x}_{(II)} = \Gamma_{(II)}\Phi^T\left(\lambda I + \Phi\Gamma_{(II)}\Phi^T\right)^{-1}\boldsymbol{y},\ \Gamma_{(I)} = \operatorname{diag}\left[\boldsymbol{\gamma}_{(II)}\right] \tag{28}$$

is a global minimum of (26) iff $\boldsymbol{\gamma}_{(II)}$ is a global minimum of (17). Moreover, if $f[\exp(\cdot)]$ is convex, then (27) can be computed with a convex program and the correspondence extends to local minima as well: $\boldsymbol{x}_* = \Gamma_*\Phi^T\left(\lambda I + \Phi\Gamma_*\Phi^T\right)^{-1}\boldsymbol{y}$ is a local minimum of (26) iff $\boldsymbol{\gamma}_*$ is a local minimum of (17).

Consequently, Type II solutions can be obtained by minimizing a penalized regression problem similar in form to Type I. Note that the sufficient (but possibly not necessary) convexity condition on $f[\exp(\cdot)]$ for local minima correspondence is satisfied in a wide variety of cases. For example, when $g(z) = |z|^p$, meaning $p(\boldsymbol{x})$ is a generalized Gaussian, then with $\beta_i \triangleq \log\gamma_i$ we have $f[\exp(\beta_i)] = [\exp(\beta_i)]^{p/(2-p)} - \log[\exp(\beta_i)] = \exp[p\beta_i/(2-p)] - \beta_i$.[6] This expression is clearly convex in β_i.

Additionally, a natural noiseless reduction of (26) exists leading to a constrained optimization problem, analogous to Type I methods. When $\lambda \longrightarrow 0$, then Type II equivalently solves

$$\boldsymbol{x}_{(II)} = \lim_{\lambda\longrightarrow 0}\arg\min_{\boldsymbol{x}}\mathfrak{g}_{(II)}(\boldsymbol{x}), \quad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}. \tag{29}$$

The only reason we retain the limit, as opposed to merely setting $\lambda = 0$ in $\mathfrak{g}_{(II)}(\boldsymbol{x})$, is because solutions with $\|\boldsymbol{x}\|_0 < n$ will effectively involve taking the log of zero when minimizing over $\boldsymbol{\gamma}$, which is undefined. Using the limit in this manner (outside of the minimization) avoids this complication, although practical implementations for solving (29) are well-behaved and stable with $\lambda = 0$ [36].

From an optimization standpoint, the cost functions from both (26) and (29) can easily be supplemented with additional constraints, e.g., $\boldsymbol{x} \geq 0$, facilitating the extension of Type II-like methods to a much wider range of applications (e.g., see [36]). Additionally, when viewed in $\boldsymbol{x}$-space, it is very natural to consider using different values for λ, e.g., λ_1 and λ_2, given the two instances that appear in (26) and implicitly in (29). In other

[5]Alternatively, it can equivalently be viewed as minimizing a Type II-like cost function with $\log|\Gamma| = \sum_i\log\gamma_i$ replacing $\log|\Sigma_y|$.

[6]Here we have used (23) to compute $f_{(I)}$ and therefore f while ignoring constant terms.

words, the value of λ multiplying $\mathfrak{g}_{(II)}(\boldsymbol{x})$ could be set to some arbitrary λ_1 while the value embedded in Σ_y could be set to λ_2. For example, in (29) where it is assumed that $\lambda_1 \rightarrow 0$, we could easily allow a nonzero λ_2 (replacing $\lambda \rightarrow 0$ with λ_2 inside of Σ_y). While beyond the scope of this paper, when using iterative reweighted ℓ_1 minimization algorithms to solve (26) or (29), adjusting this λ_2 can potentially improve performance substantially [36], similar to the ϵ factor in the reweighting method of Candès *et al.* [6]. Note that it is only when we analyze Type II in $\boldsymbol{x}$-space (as a standard form of penalized regression) that manipulating λ in this way makes any sense; in the original hierarchical Bayesian model it is counterintuitive to maintain two values of λ. This also opens the door to using a different dictionary for constructing $\mathfrak{g}_{(II)}(\boldsymbol{x})$. This issue will be taken up again in Section VI.

III. Analysis of the Type II Cost Function in $\boldsymbol{x}$-Space

The distinguishing factor of Type II methods is the log-det term in (17) and (27); the other regularization term based on $f(\gamma_i)$ is effectively present in Type I as well (see Section II-A) and, when mapped into $\boldsymbol{x}$-space, has been analyzed extensively in this context [5], [7], [10], [13], [28], [33]. Consequently, we will concentrate our attention here on the simple case where $f(\gamma_i) = 0$ and flesh out the corresponding characteristics of the underlying Type II cost function in $\boldsymbol{x}$-space and examine the relationship with popular Type I methods. Additionally, local minimum analyses in Section IV suggest that the choice $f(\gamma_i) = 0$ is particularly useful when maximal sparsity is concerned. Alternative choices for $f(\gamma_i)$ in the context of sparse recovery are examined in [38], further justifying the selection $f(\gamma_i) = 0$.

A. General Properties of the Type II Penalty $\mathfrak{g}_{(ii)}(\boldsymbol{x})$

It is well-know that concave, nondecreasing functions of the coefficient magnitudes favor sparse solutions [28]. We now demonstrate that $\mathfrak{g}_{(II)}(\boldsymbol{x})$ is such a penalty, meaning $\mathfrak{g}_{(II)}(\boldsymbol{x}) = \mathfrak{h}(|\boldsymbol{x}|)$, where $|\boldsymbol{x}| \triangleq [|x_1|, \ldots, |x_m|]^T$ and $\mathfrak{h}$ is a concave, nondecreasing function of $|\boldsymbol{x}|$.

Theorem 3: When $f(\gamma_i) = 0$, $\mathfrak{g}_{(II)}(\boldsymbol{x})$ is a concave, nondecreasing function of $|\boldsymbol{x}|$. Additionally, every local minimum of (26) or (29) can be achieved at a solution with at most n nonzero elements, regardless of λ.

In the noiseless case, such solutions $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_0 \leq n$ are referred to as *basic feasible solutions* (BFS). The second point in Theorem 3 has also been shown for the analogous Type II cost function directly in $\boldsymbol{\gamma}$-space [34], meaning local minima can be achieved with at most n nonzero elements of $\boldsymbol{\gamma}$, but the result is much less transparent. Theorem 3 also holds for any $f(\gamma_i)$ that is concave and nondecreasing. As an aside, it also implies that globally convergent, reweighted ℓ_1 minimization is possible for optimizing $\mathcal{L}^x_{(II)}(\boldsymbol{x})$ [36], assuming again that $f(\gamma_i)$ that is concave and nondecreasing.

Regarding global minima we have the following result.

Theorem 4: Given $\mathrm{spark}(\Phi) = n+1$, assume that there exists at least one feasible solution to $\boldsymbol{y} = \Phi\boldsymbol{x}$ with $\|\boldsymbol{x}\|_0 < n$. Then the set of coefficient vectors that globally minimize (29) with $f(\gamma_i) = 0$ also globally minimize (15).

Consequently a global minimum of (29) will always correspond with a global minimum of (15). (Theorem 4 actually holds for any f that is bounded.)

Thus far we have not provided any reason why the Type II penalty $\mathfrak{g}_{(II)}(\boldsymbol{x})$ has any direct advantage over Type I. In fact, both Theorems 3 and 4 are also trivially satisfied by replacing $\mathfrak{g}_{(II)}(\boldsymbol{x})$ with the canonical sparse penalty $\|\boldsymbol{x}\|_0$, which is a special case of Type I. However, several factors distinguish $\mathfrak{g}_{(II)}(\boldsymbol{x})$ in the context of sparse approximation.

First, $\mathfrak{g}_{(II)}(\boldsymbol{x})$ is *nonseparable* , meaning $\mathfrak{g}_{(II)}(\boldsymbol{x}) \neq \sum_i g_{(II)}(x_i)$ for some function $g_{(II)}$. Equivalently, the implicit prior distribution on $\boldsymbol{x}$ given by $p_{(II)}(\boldsymbol{x}) \propto \exp[-\frac{1}{2}\mathfrak{g}_{(II)}(\boldsymbol{x})]$, is *nonfactorial*, implying dependencies between elements of $\boldsymbol{x}$. Additionally, unlike traditional Type I procedures (e.g., Lasso, ridge regression, etc.), this penalty is explicitly dependent on both the dictionary Φ and potentially the regularization parameter λ (assuming we only use a single λ as discussed above). The only exception occurs when $\Phi^T\Phi = I$; here $\mathfrak{g}_{(II)}(\boldsymbol{x})$ separates and can be expressed in closed form independently of Φ, although λ-dependency remains.

In general, the ℓ_1 norm is the optimal or tightest *convex* relaxation of the ℓ_0 norm, and therefore it is commonly used leading to the Lasso and related ℓ_1 penalty algorithms [30]. However, the ℓ_1 norm need not be the best relaxation in general. In Sections III-B and III-C we will demonstrate that the nonseparable, λ-dependent $\mathfrak{g}_{(II)}(\boldsymbol{x})$ provides a tighter, albeit *nonconvex*, approximation that promotes greater sparsity than $\|\boldsymbol{x}\|_1$ while conveniently producing many fewer local minima than when using $\|\boldsymbol{x}\|_0$ directly. We also show that, in certain settings, no separable, λ-independent regularization term can achieve similar results. Consequently, the widely used family of ℓ_p quasi-norms, i.e., $\|\boldsymbol{x}\|_p^p = \sum_i |x_i|^p, p \leq 1$ [9], or the Gaussian entropy measure $\sum_i \log|x_i|$ based on the Jeffreys prior [15] provably fail in this regard.

Finally, at a superficial level, the Φ-dependency of $\mathfrak{g}_{(II)}(\boldsymbol{x})$ leads to scale-invariant solutions in the following sense. If we rescale Φ with a diagonal matrix D, i.e., $\Phi \rightarrow \Phi D$, then the optimal solution becomes $\boldsymbol{x}_{(II)} \rightarrow D\boldsymbol{x}_{(II)}$. In contrast, when minimizing the ℓ_1 norm, such a rescaling leads to a completely different solution which requires solving an entirely new convex program; there is no simple linear relationship between the solutions.

B. Benefits of a Nonseparable Penalty

The benefits of the nonseparable nature of $\mathfrak{g}_{(II)}(\boldsymbol{x})$ are most pronounced in the overcomplete case, meaning there are more dictionary columns than dimensions of the signal $\boldsymbol{y}$. In a noiseless setting (with $\lambda \rightarrow 0$), we can explicitly quantify the potential of this property of $\mathfrak{g}_{(II)}(\boldsymbol{x})$. As discussed previously, the global minimum of (29) will equal $\boldsymbol{x}_0$, the maximally sparse solution to (15), assuming the latter is unique. The real distinction then is regarding the number of local minimum. In this capacity $\mathfrak{g}_{(II)}(\boldsymbol{x})$ is superior to any possible separable variant:

Theorem 5: In the limit as $\lambda \rightarrow 0$, no *separable* penalty $\mathfrak{g}(\boldsymbol{x}) = \sum_i g(x_i)$ exists such that, for all $\boldsymbol{y}$ and Φ with $\mathrm{spark}(\Phi) = n+1$, the corresponding Type I optimization problem

$$\min_{\boldsymbol{x}} \sum_i g(x_i) \quad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x} \tag{30}$$

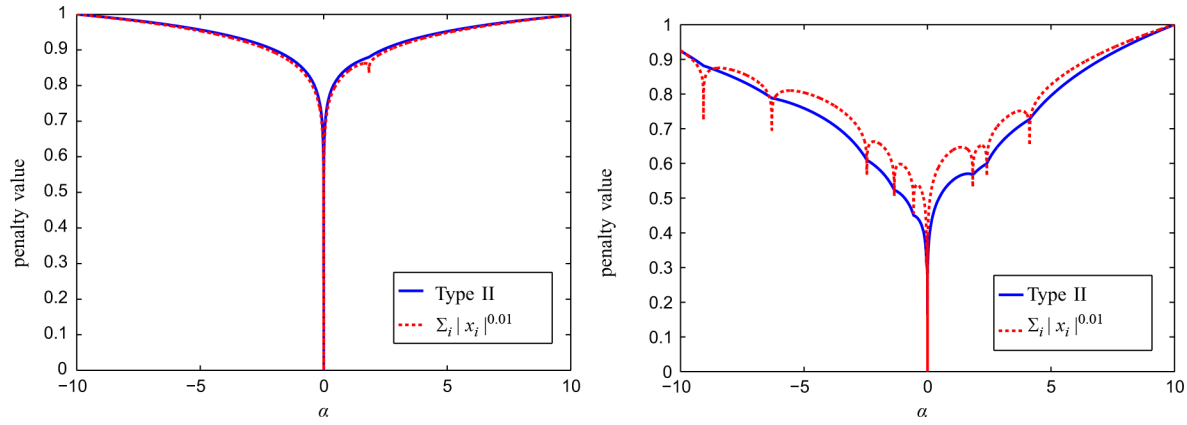

Fig. 1. Plots of the Type II penalty (normalized) across the feasible region as parameterized by α. A separable penalty given by $\mathfrak{g}(\boldsymbol{x}) \propto \sum_i |x_i|^{0.01} \approx \|\boldsymbol{x}\|_0$ is included for comparison. Both approximations to the ℓ_0 norm retain the correct global minimum, but only the Type II penalty smooths out local minima. *Left*: $\|\boldsymbol{x}_0\|_0 = 1$ (simple case). *Right*: $\|\boldsymbol{x}_0\|_0 = 9$ (hard case).

is: (i) Globally minimized only by solutions that minimize (15); and (ii) Ever has fewer local minima than when solving (29).

Note that the spark condition is merely included to simplify the proof (see the Appendix); Theorem 5 can be extended with additional effort to include other spark values.[7] In general, Theorem 5 speaks directly to the potential limitations of restricting oneself to separable penalties (or equivalently factorial priors) when maximal sparsity is paramount. As aforementioned, use of the separable ℓ_1 norm has traditionally been advocated because it represents the tightest convex approximation to the ℓ_0 norm. However, a viable alternative relaxation is to replace the convexity requirement with condition (i) from above (i.e., matching global minimum) and then ask what is the smoothest approximation to the ℓ_0 norm, separable or not, consistent with this assumption. The Type II method discussed above provides very substantial smoothing at the expense of convexity, yet can still be implemented with tractable updates characterized by provable convergence to some local minimizer (possibly global) that can never be less sparse than the minimum ℓ_1 norm solution [36].

While generally difficult to visualize, in restricted situations it is possible to explicitly illustrate the type of smoothing over local minima that is possible using nonseparable penalties. For example, consider the case where $m = n + 1$ and $\text{spark}(\Phi) = m$, implying that Φ has a null-space dimension of one. Consequently, any feasible solution to $\boldsymbol{y} = \Phi\boldsymbol{x}$ can be expressed as $\boldsymbol{x} = \boldsymbol{x}_0 + \alpha\boldsymbol{v}$, where $\boldsymbol{v} \in \text{null}(\Phi)$, α is some real-valued scalar, and $\boldsymbol{x}_0$ is the maximally sparse solution. We can now plot any penalty function $\mathfrak{g}(\boldsymbol{x})$ over the 1D feasible region of $\boldsymbol{x}$-space as a function of α to view the local minima profile.

In this simplified situation, the maximum number of local minima equals $n + 1$, since removing any column from Φ produces a BFS. However, if $\|\boldsymbol{x}_0\|_0 < n$, then not all of these BFS can be unique. For example, if $\|\boldsymbol{x}_0\|_0 = 1$, then only two BFS will be unique: one solution that includes all columns of Φ not used by $\boldsymbol{x}_0$, and then the solution $\boldsymbol{x}_0$ itself. In contrast, if $\|\boldsymbol{x}_0\|_0 = n - 1$, then there will be n unique BFS (because $\boldsymbol{x}_0$ will have two zero-valued elements and removing either associated dictionary column will lead to the same BFS). Therefore, the local minima problem is exacerbated as $\|\boldsymbol{x}_0\|_0$ becomes larger, consistent with expectations. Ideally then, a nonseparable penalty will provide additional smoothing in this regime.

We demonstrate these ideas with two test cases, both of which involve the same 10×11 dictionary Φ generated with i.i.d. unit Gaussian entries. In the first case we compute $\boldsymbol{y} = \Phi\boldsymbol{x}_0$, where $\boldsymbol{x}_0$ is a sparse vector with $\|\boldsymbol{x}_0\|_0 = 1$; the single nonzero element is drawn from a unit Gaussian. Fig. 1 (left) displays the plots of two example penalties in the feasible region of $\boldsymbol{y} = \Phi\boldsymbol{x}$: (i) the nonseparable Type II penalty $\mathfrak{g}_{(II)}(\boldsymbol{x})$, and (ii) the conventional penalty $\mathfrak{g}(\boldsymbol{x}) = \sum_i |x_i|^p$, $p = 0.01$. The later is a separable penalty that converges to the canonical ℓ_0 norm when $p \to 0$. From the figure, we observe that, while both penalties peak at the maximally sparse solution $\boldsymbol{x}_0$, the Type I penalty has a second, small local minima as well located at $\alpha \approx 2$. While the Type II penalty displays a single basin of attraction, its smoothing benefits are not very pronounced in this situation.

In the second case, we repeat the above with $\|\boldsymbol{x}_0\|_0 = 9$. This is the largest number of nonzeros such that a unique, maximally sparse solution still exists [with high probability by virtue of (14)]. Hence it is the most difficult sparse recovery problem to solve, with 10 unique local minima per the discussion above. Fig. 1 (right) shows the results. Now the Type I penalty reflects all 10 local minima (9 are shown), while Type II demonstrates dramatic smoothing. While the ℓ_1 norm (which is equivalent to the assumption $p = 1$) also smooths out local minima, the global minimum may be biased away from the maximally sparse solution in many situations, unlike Type II which provides a nonconvex approximation with its global minimum anchored at $\boldsymbol{x}_0$. We will revisit this issue in much more detail in Section IV.

In general, the Achilles heel of standard, separable penalties (Type I) is that if we want to retain the global minimum of (15), we require a highly concave penalty on each x_i. However, this implies that *all* BFS will form local minima of the penalty function constrained to the feasible region (see the proof of Theorem 5 in the Appendix). This is a very undesirable property since there are on the order of $\binom{m}{n}$ unique BFS with $\|\boldsymbol{x}\|_0 = n$ (assuming $\text{spark}(\Phi) = n + 1$), which is not very sparse. In the example from Fig. 1 (right) there are 10 such solutions and hence 10 local minima to the Type I cost. We would really like to find *degenerate* BFS, where $\|\boldsymbol{x}\|_0$ is strictly less than n. Such

[7]We can also always add an arbitrarily small amount of randomness to any dictionary to satisfy the spark constraint.

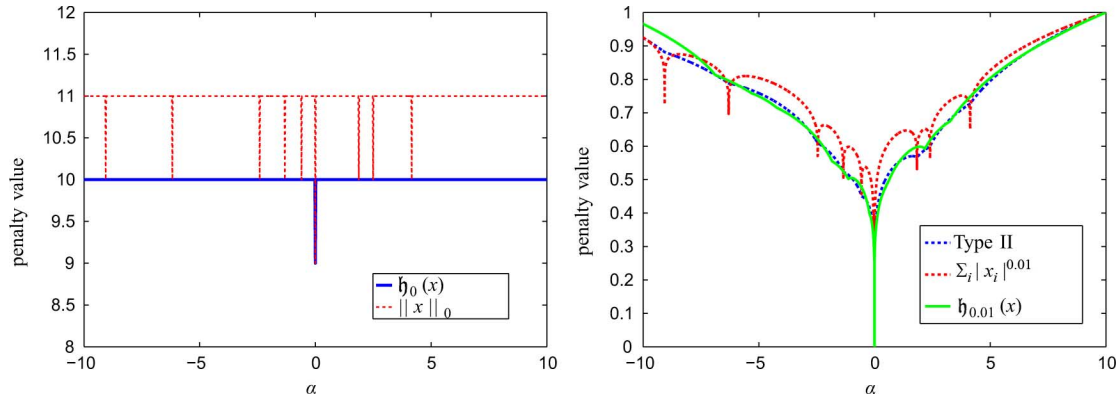

Fig. 2. Example smoothing of two hypothetical nonseparable penalties. *Left:* $\mathfrak{h}_0(\boldsymbol{x})$ and the ℓ_0 norm versus α. 10 distinct local minima are present with the ℓ_0 norm (9 are shown), but only a single degenerate BFS. However, the "truncation" of the ℓ_0 norm that characterizes $\mathfrak{h}_0(\boldsymbol{x})$ has removed all local minima; the global minimum remains unaltered. *Right:* $\mathfrak{h}_p(\boldsymbol{x})$ and $\|\boldsymbol{x}\|_p^p$ versus α (normalized), $p = 0.01$ (the Type II plot from Fig. 1 (right) is also included for comparison). This represents a more practical nonseparable approximation that retains slope information pointing towards the global solution that could, at least in principle, be used for optimization purposes.

solutions are exceedingly rare and difficult to find, yet it is these very solutions that can be favored by the proper construction of highly concave, nonseparable penalties.

A simple example serves to illustrate how a nonseparable penalty can remove nondegenerate BFS that act as local minima. Consider the penalty function $\mathfrak{h}_0(\boldsymbol{x}) \triangleq \min(\|\boldsymbol{x}\|_0, n)$, where $\mathfrak{h}_0(\boldsymbol{x})$ is equivalent to taking the ℓ_0 norm of the largest (in magnitude) n elements of $\boldsymbol{x}$; this leads to the optimization problem

$$\min_{\boldsymbol{x}} \mathfrak{h}_0(\boldsymbol{x}), \quad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}. \tag{31}$$

While the global minimum remains at $\boldsymbol{x}_0$, all local minima occurring at nondegenerate BFS have been effectively removed. In other words, at any feasible solution $\boldsymbol{x}_*$ with n nonzero entries, we can always add a small component $\alpha\boldsymbol{v} \in \text{null}(\Phi)$ and maintain feasibility without increasing $\mathfrak{h}_0(\boldsymbol{x})$, since $\mathfrak{h}_0(\boldsymbol{x})$ can never be greater than n. Therefore, we are free to move from BFS to BFS without increasing $\mathfrak{h}_0(\boldsymbol{x})$. Also, the rare degenerate BFS that do remain, even if suboptimal, are sparser by definition. Therefore, locally minimizing the new problem (31) is clearly superior to locally minimizing (15). This is possible because we have replaced the troublesome separable penalty $\|\boldsymbol{x}\|_0$ with the nonseparable surrogate $\mathfrak{h}_0(\boldsymbol{x})$.

This notion is illustrated with a simple graphic in Fig. 2 (left), which compares the ℓ_0 norm with $\mathfrak{h}_0(\boldsymbol{x})$ in a 1D feasible region parameterized by α with the same setup as in Fig. 1 (right). In this situation, all local minima are removed by the simple, nonseparable "truncated" ℓ_0 norm $\mathfrak{h}_0(\boldsymbol{x})$.

To create effective sparsity penalties in general, it may not be optimal to apply concave, sparsity-inducing functions directly to the individual coefficients (or latent variables) in an elementwise fashion (separable), which is characteristic of all Type I methods. Rather, it can be useful to map the coefficients to a lower-dimensional space first. The latter operation, which is effectively what Type II accomplishes, then necessitates that the resulting penalty be nonseparable in the original full-dimensional space. For example, $\mathfrak{h}_0(\boldsymbol{x})$ first maps to an n-dimensional space (the n largest coefficients of $\boldsymbol{x}$), before applying the ℓ_0 norm. Of course $\mathfrak{h}_0(\boldsymbol{x})$ is not viable practically since there is no gradient information or curvature, rendering minimization intractable. However, a simple alternative is $\mathfrak{h}_p(\boldsymbol{x})$, which applies the ℓ_p quasi-norm (with $0 < p < 1$) to the n largest elements of $\boldsymbol{x}$. Fig. 2 (right) compares $\mathfrak{h}_p(\boldsymbol{x})$ with direct application of $\|\boldsymbol{x}\|_p^p$, using $p = 0.01$ and the same experimental setup as before. Notice that the smoothing of local minima closely mimics that of the Type II penalty $\mathfrak{g}_{(II)}(\boldsymbol{x})$. While this may on the surface be a surprising result, analysis of Type II in $\boldsymbol{\gamma}$-space provides strong intuitive evidence for why this should be the case; however, for space considerations we defer this analysis to a future publication.

C. Benefits of λ Dependency

To briefly explore the potential benefits of λ dependency in the Type II penalty $\mathfrak{g}_{(II)}(\boldsymbol{x})$ in a noisy setting, we adopt the simplifying assumption $\Phi^T\Phi = I$. In this special case, $\mathfrak{g}_{(II)}(\boldsymbol{x})$ actually becomes separable and can be computed in closed form via

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) = \sum_i g_{(II)}(x_i) \propto \sum_i \frac{2|x_i|}{|x_i| + \sqrt{x_i^2 + 4\lambda}} + \log\left(2\lambda + x_i^2 + |x_i|\sqrt{x_i^2 + 4\lambda}\right) \tag{32}$$

which is independent of Φ. A plot of $g_{(II)}(x_i)$ is shown in Fig. 3 below. The λ dependency of (32) contributes some desirable properties to the Type II cost function. Before giving the main result, we state that $g(x)$ is a *strictly concave* function of $|x|$ if $g(x) = h(|x|)$ and $h[\alpha x + (1-\alpha)y] > \alpha h(x) + (1-\alpha)h(y)$ for all $\alpha \in (0,1)$ and $x, y \in [0, \infty)$, $x \neq y$. This leads to the following.

Theorem 6: Assuming $\Phi^T\Phi = I$, then the following hold:

1) The cost function (26) has no (nonglobal) local minima.
2) $g_{(II)}(x_i)$ is a nondecreasing and strictly concave function of $|x_i|$, and so provides a tighter approximation to $\|\boldsymbol{x}\|_0$ than $\|\boldsymbol{x}\|_1$ (see Appendix for more details).
3) No fixed, λ-independent penalty can satisfy both of the above properties for all $\lambda, \boldsymbol{y}$, and Φ.
4) Direct minimization of (16) has 2^m local minima; any other strictly concave, λ-independent penalty function can potentially have this many local minima as well, depending on Φ and $\boldsymbol{y}$.

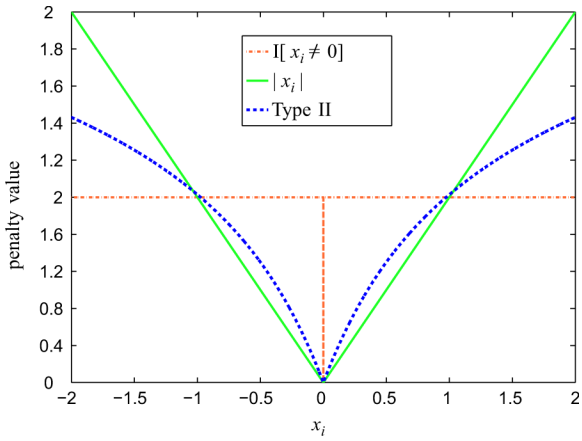

Fig. 3. 1D example of the Type II penalty (normalized) assuming $\Phi^T\Phi = I$. The ℓ_1 and ℓ_0 norms are included for comparison.

Intuitively, when λ is small, the Gaussian likelihood term (or quadratic data-fit term) is highly restrictive, constraining most of its relative mass to a very localized region of $\boldsymbol{x}$-space. Therefore, a tighter prior/penalty more closely resembling the ℓ_0 norm can be used without the risk of local minima, which occur when the spines of a sparse prior overlap nonnegligible portions of the likelihood (see Fig. 6 in [31] for a good 2D visual of a sparse prior with characteristic spines running along the coordinate axes). In the limit as $\lambda \rightarrow 0$, $\mathfrak{g}_{(II)}(\boldsymbol{x})$ converges to a scaled proxy of the ℓ_0 norm, yet no local minimum exist because the likelihood in this case only permits a single feasible solution with $\boldsymbol{x} = \Phi^T\boldsymbol{y}$. To see this, consider reexpressing (32) as

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) = \sum_i g_{(II)}(x_i) \propto \sum_i \frac{2|x_i|}{|x_i| + \sqrt{x_i^2 + 4\lambda}} + \sum_i \log\left(2\lambda + x_i^2 + |x_i|\sqrt{x_i^2 + 4\lambda}\right). \quad (33)$$

With $\lambda \rightarrow 0$, the first summation converges to $\|\boldsymbol{x}\|_0$ while the second reduces to $\sum_i \log|x_i|$, ignoring an irrelevant scale factor and a constant. Sometimes referred to as Gaussian entropy, this log-based factor can then be related to the ℓ_0 norm via $\|\boldsymbol{x}\|_0 \equiv \lim_{p\rightarrow 0}\sum_i |x_i|^p$ and $\lim_{p\rightarrow 0}\frac{1}{p}\sum_i(|x_i|^p - 1) = \sum_i \log|x_i|$.

In contrast, when λ is large, the likelihood is less constrained and a looser prior (meaning a less concave penalty function) is required to avoid local minima troubles, which will arise whenever the now relatively diffuse likelihood intersects the sharp spines of a highly sparse prior. In this situation $\mathfrak{g}_{(II)}(\boldsymbol{x})$ converges to a scaled version of the ℓ_1 norm. The Type II penalty naturally handles this transition becoming sparser as λ decreases and *vice versa*.

Of course as we alluded to previously, we can potentially treat the λ embedded in $\mathfrak{g}_{(II)}(\boldsymbol{x})$ as a separate parameter; in general there is no guarantee that keeping the two instances of λ equal is necessarily optimal. But the analysis here does motivate the point that varying the concavity of the penalty function to reflect, for example, differing noise levels can expand the utility of nonconvex approximations.

In summary, use of the ℓ_1 norm in place of $\mathfrak{g}_{(II)}(\boldsymbol{x})$ also yields no local minima; however, it is a much looser approximation of the ℓ_0 norm and penalizes coefficients linearly unlike $\mathfrak{g}_{(II)}(\boldsymbol{x})$. As a final point of comparison, the actual coefficient estimate obtained from minimizing (26) when $\Phi^T\Phi = I$ is exactly equivalent to the nonnegative garrote estimator that has been advocated for wavelet shrinkage [16], [34].

IV. Type II Local Minima Conditions

From Section III-A we know that any global minimum of the Type II cost function (whether in $\boldsymbol{x}$-space or $\boldsymbol{\gamma}$-space) coincides with a global solution to (15) when $f(\gamma_i) = 0$ and $\lambda \rightarrow 0$. Additionally, we have shown that Type II provides a way to smooth local minima created by direct use of the ℓ_0 norm (or any close, separable approximation). However, it remains unclear what determines when and where local minima will occur or conditions whereby they are all removed. From Theorem 3 we know that every local minimum can be achieved with at most n nonzero elements, i.e., a basic feasible solution (BFS). Assuming $\lambda \rightarrow 0$ (noiseless case) and $\text{spark}(\Phi) = n + 1$, this provides an easy way to bound the possible number of local minima

$$1 \leq \begin{matrix}\#\text{ of Type II} \\ \text{Local Minima}\end{matrix} \leq \begin{matrix}\#\text{ of BFS to} \\ \boldsymbol{y} = \Phi\boldsymbol{x}\end{matrix} \in \left\{\binom{m-1}{n} + 1, \binom{m}{n}\right\} \quad (34)$$

where the upper bound is from [19]. Any Type I (separable) method whose global solution always globally minimizes (15) necessarily will achieve the upper bound (see the proof of Theorem 5 in the Appendix); however, with Type II this need not be the case. In fact, most BFS will not end up being local minima [e.g., see Fig. 1 (right)]. As we will show below, in some cases it is even possible to achieve the ideal lower bound, i.e., a single minima that is globally optimal. As before, we will focus our attention to the case where $f(\gamma_i) = 0$. Local minima analyses for arbitrary $f(\gamma_i)$ are considered in [34].

A. Necessary Conditions for Local Minima

Although we cannot remove all nondegenerate local minima in all situations and still retain computational tractability, it is possible to remove many of them, providing some measure of approximation to (31). This is effectively what is accomplished using Type II as will be subsequently argued. Specifically, we will derive necessary conditions required for a nondegenerate BFS to represent a local minimum to $\mathcal{L}^x_{(II)}(\boldsymbol{x})$ (assuming $\lambda \rightarrow 0$). We will then show that these conditions are often *not* satisfied, implying that there are potentially many fewer local minima. Thus, locally minimizing $\mathcal{L}^x_{(II)}(\boldsymbol{x})$ comes closer to (locally) minimizing (31) than traditional Type I methods, which in turn, is closer to globally minimizing $\|\boldsymbol{x}\|_0$.

Suppose that we have found a nondegenerate BFS $\boldsymbol{x}_*$ and we would like to assess whether or not it is a local minimum to the Type II cost function with $\lambda \rightarrow 0$. For convenience, let $\tilde{\boldsymbol{x}}$ denote the n nonzero elements of $\boldsymbol{x}_*$ and $\tilde{\Phi}$ the associated columns of Φ (therefore, $\boldsymbol{y} = \tilde{\Phi}\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}} = \tilde{\Phi}^{-1}\boldsymbol{y}$). Intuitively, it would seem likely that if we are not at a true local minimum, then there must exist at least one additional column of Φ not in $\tilde{\Phi}$, e.g., some $\boldsymbol{u}$, that is appropriately aligned with or in some respect similar to $\boldsymbol{y}$. Moreover, the significance of this potential alignment must be assessed relative to $\tilde{\Phi}$. For example, it seems plausible (desirable) that if $\boldsymbol{u} \approx \boldsymbol{y}$ and all columns of $\tilde{\Phi}$ are not close to $\boldsymbol{y}$,

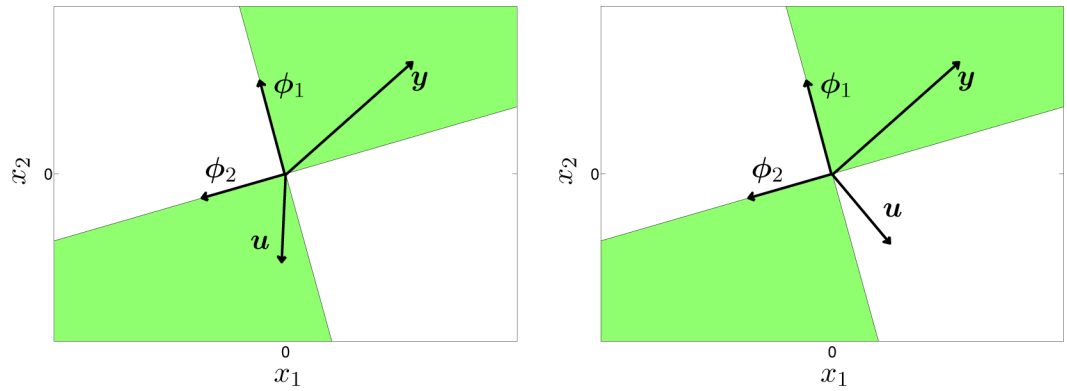

Fig. 4. 2D example with a 2×3 dictionary Φ (i.e., $n = 2$ and $m = 3$) and a basic feasible solution using the columns $\bar{\Phi} = [\boldsymbol{\phi}_1 \boldsymbol{\phi}_2]$. The shaded areas represent the cone (and its reflection about the origin) described in the main text. In this simple case, $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ divide $\boldsymbol{x}$-space into four quadrants. The shaded regions include the quadrant containing $\boldsymbol{y}$ and its reflection about zero. *Left:* In this case, $\boldsymbol{u} = \boldsymbol{\phi}_3$ penetrates the shaded region, and so we satisfy the conditions of Theorem 7, ensuring that this configuration does *not* represent a local minima of Type II. But it *does* represent a local minimum of any Type I method constrained to match the global minimum of the ℓ_0 norm. *Right:* Now $\boldsymbol{u}$ is outside of the cone (and cannot be used to form a tighter cone about $\boldsymbol{y}$), so this situation does represent a minimizing basic feasible solution for Type II.

then possibly (hopefully) we are not at a local minimum and a sparser solution can be descended upon by including $\boldsymbol{u}$.

A useful metric for comparison is realized when we decompose $\boldsymbol{u}$ with respect to $\tilde{\Phi}$, which forms a basis in $\mathbb{R}^n$ under the assumption that $\text{spark}(\Phi) = n+1$. For example, we may form the decomposition $\boldsymbol{u} = \tilde{\Phi}\tilde{\boldsymbol{v}}$, where $\tilde{\boldsymbol{v}}$ is a vector of coefficients analogous to $\tilde{\boldsymbol{x}}$. As will be shown later, the similarity required between $\boldsymbol{u}$ and $\boldsymbol{y}$ (needed for establishing the existence of a local minimum) may then be realized by comparing the respective coefficients $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{v}}$. In more familiar terms, this is analogous to suggesting that similar signals have similar Fourier expansions. Loosely, we may expect that if $\tilde{\boldsymbol{v}}$ is "close enough" to $\tilde{\boldsymbol{x}}$, then $\boldsymbol{u}$ is sufficiently close to $\boldsymbol{y}$ (relative to all other columns in $\tilde{\Phi}$) such that we are not at a local minimum. We formalize this idea via the following result:

Theorem 7: Let Φ satisfy $\text{spark}(\Phi) = n + 1$ and let $\boldsymbol{x}_*$ represent a solution vector with $\|\boldsymbol{x}_*\|_0 = n$ entries such that $\tilde{\boldsymbol{x}} = \tilde{\Phi}^{-1}\boldsymbol{y}$. Let $\mathcal{U}$ denote the set of $m - n$ columns of Φ not included in $\tilde{\Phi}$ and $\mathcal{V}$ the set of coefficients given by $\left\{\tilde{\boldsymbol{v}} : \tilde{\boldsymbol{v}} = \tilde{\Phi}^{-1}\boldsymbol{u}, \boldsymbol{u} \in \mathcal{U}\right\}$. Then $\boldsymbol{x}_*$ is *not* a local minimum of (29) if

$$\sum_{i \neq j} \frac{\tilde{v}_i \tilde{v}_j}{\tilde{x}_i \tilde{x}_j} > 0 \tag{35}$$

for some $\tilde{\boldsymbol{v}} \in \mathcal{V}$.

This theorem provides a useful picture of what is required for local minima to exist and more importantly, why many (possibly most) BFS are not local minima. Moreover, there are several convenient ways in which we can interpret this result to accommodate a more intuitive perspective.

In general, if the sign patterns of $\tilde{\boldsymbol{v}}$ and $\tilde{\boldsymbol{x}}$ tend to align, then the left-hand side of (35) will likely be positive and we cannot be at a local minimum. For illustration purposes, in the extreme instance where the sign patterns match exactly, this will necessarily be the case. This special situation can be understood geometrically as follows. Consider the convex cone constructed via the columns of the matrix $\tilde{\Phi}S$, where $S \triangleq \text{diag}\,(\text{sign}(\tilde{\boldsymbol{x}}))$. This cone is equivalent to the set vectors which can be formed as positive linear combinations of the columns of $\tilde{\Phi}S$, i.e., the set $\left\{\boldsymbol{z} : \boldsymbol{z} = \tilde{\Phi}S\boldsymbol{w}, \boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{w} \geq 0\right\}$. By definition, this cone will necessarily contain the signal $\boldsymbol{y}$. However, if this cone contains any other basis vector $\boldsymbol{u} \in \mathcal{U}$, then the sign pattern of the corresponding $\tilde{\boldsymbol{v}}$ will match $\tilde{\boldsymbol{x}}$ and we cannot be at a local minimum via (35). By symmetry arguments, the same is true for any $\boldsymbol{u}$ in the convex cone formed by $-\tilde{\Phi}S$. The simple 2D example shown in Fig. 4 helps to illustrate this point.

Alternatively, we can cast this geometric perspective in terms of relative cone sizes. For example, let C represent the convex cone, and its reflection, formed by $\tilde{\Phi}S$. Then we are not at a local minimum to $\mathcal{L}^x_{(II)}(\boldsymbol{x})$ if there exists a second convex cone C' formed from a subset of columns of Φ such that $\boldsymbol{y} \in C' \subset C$, i.e., C' is a tighter cone containing $\boldsymbol{y}$. In Fig. 4 (left), we obtain a tighter cone about $\boldsymbol{y}$ by replacing $\boldsymbol{\phi}_1$ with $\boldsymbol{u}$.

Of course we must emphasize that these geometric conditions are *much* weaker than (35), e.g., if all $\boldsymbol{u} \in \mathcal{U}$ are *not* in C, we still may not be at a local minimum. In fact, for a local minimum to occur, all $\boldsymbol{u}$ must be reasonably far from this cone such that $\sum_{i \neq j} \frac{\tilde{v}_i \tilde{v}_j}{\tilde{x}_i \tilde{x}_j} \leq 0, \forall \tilde{\boldsymbol{v}} \in \mathcal{V}$.

B. Conditions for Removing All Local Minima

This section describes conditions, based on the relative magnitudes of the nonzero elements in $\boldsymbol{x}_0$, such that all (nonglobal) local minima of (29) are removed leaving a unique global solution that equals $\boldsymbol{x}_0$. The core idea is that as these nonzero magnitudes become highly scaled, there are increasingly fewer local minima until eventually all are smoothed away. In contrast, we argue in Section IV-C that when all the nonzero coefficients have equal magnitudes, obtaining $\boldsymbol{x}_0$ is more difficult because of more local minima. However, even in this worst-case scenario we demonstrate empirically in Section V that Type II still outperforms widely used Type I algorithms.

Theorem 8: Let $x_{(i)}$ denote the ith largest coefficient magnitude of $\boldsymbol{x}$ and assume $\text{spark}(\Phi) = n+1$. Then there exists a set of $n-2$ scaling constants $\nu_i \in (0, 1]$ (i.e., strictly greater than zero) such that, for any $\boldsymbol{y} = \Phi\boldsymbol{x}'$ generated with $\|\boldsymbol{x}'\|_0 < n$ and

$$x'_{(i+1)} \leq \nu_i x'_{(i)} \quad i = 1, \ldots, n-2 \tag{36}$$

the problem (29) has a unique minimum $\boldsymbol{x}_{(II)}$ such that $\boldsymbol{x}_{(II)} = \boldsymbol{x}'$. Moreover, $\boldsymbol{x}'$ will equal $\boldsymbol{x}_0$, the unique maximally sparse solution.

This result is obviously restrictive in the sense that the dictionary-dependent constants ν_i significantly confine the class of signals $\boldsymbol{y}$ that we may represent. Moreover, we have not provided any convenient means of computing what the different scaling constants might be. But Theorem 8 nonetheless solidifies the notion that the Type II cost function is especially capable of recovering coefficients of different scales (and it must still find all nonzero elements no matter how small some of them may be). Additionally, we have specified conditions whereby we will find the unique $\boldsymbol{x}_0$ even when the sparsity is as large as $\|\boldsymbol{x}_0\| = n - 1$, provided we use an appropriate, globally-convergent algorithm such as iterative reweighted ℓ_1 minimization [36].

It is important to stress that this result specifies sufficient conditions for removing all suboptimal local minima from the Type II cost function, but these conditions are by no means necessary for removing most/all influential local minima. In practice, locally minimizing (29) performs quite well even when the coefficients are not highly scaled (see Section V). Moreover, we can always initialize at the minimum ℓ_1-norm solution (best convex approximation), and then progress from there. In fact, when optimized via an iterative reweighted ℓ_1 minimization technique, Theorem 8 can be leveraged to show that locally minimizing (29) can never do worse than the minimum ℓ_1 solution and that, for any dictionary and sparsity profile, there will always be cases where it does better (in particular, when highly scaled coefficients are present) [36]. This is true even for dictionaries with arbitrarily bad coherence properties, e.g., $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_j \approx 1$ for all $i \neq j$, where $\boldsymbol{\phi}_i$ and $\boldsymbol{\phi}_j$ are the ith and jth columns of Φ, respectively. This topic will be pursued in greater detail in a future publication examining Type II methods in the context of structured dictionaries.

In contrast, no possible Type I method satisfies a result comparable to Theorem 8.

Theorem 9: For any set of $n - 2$ nonzero scaling constants there will always exist a dictionary Φ and a set of ordered coefficients $\boldsymbol{x}'$, consistent with the stipulations of Theorem 8, such that any possible Type I cost function, given Φ and the signal $\boldsymbol{y} = \Phi \boldsymbol{x}'$, will have multiple local minima and/or a global minimum that is not maximally sparse.

At this point, it may be unclear what probability distributions are likely to produce coefficient magnitudes that satisfy the conditions of Theorem 8. It turns out that the Jeffreys prior, given by $p(x) \propto 1/x$, is appropriate for this task. This distribution has the unique property that the probability mass assigned to any given scaling is equal. More explicitly, for any $s \geq 1$

$$\text{Prob}\left(x \in \left[s^i, s^{i+1}\right]\right) \propto \log(s) \quad \forall i \in \mathbb{Z}. \tag{37}$$

For example, the probability that x is between 1 and 10 equals the probability that it lies between 10 and 100 or between 0.01 and 0.1. Because this is an improper density, we define an approximate Jeffreys prior with range parameter $a \in (0, 1)$. Specifically, we say that $x \sim J(a)$ if

$$p(x) = \frac{-1}{2\log(a)x} \quad \text{for } x \in [a, 1/a]. \tag{38}$$

With this definition in mind, we present the following result.

Theorem 10: For a given Φ that satisfies $\text{spark}(\Phi) = n + 1$, let $\boldsymbol{y}$ be generated by $\boldsymbol{y} = \Phi \boldsymbol{x}'$, where $\|\boldsymbol{x}'\|_0 < n$ with nonzero magnitudes drawn i.i.d. from $J(a)$. Then as a approaches zero, the probability that we obtain an $\boldsymbol{x}'$ such that the conditions of Theorem 8 are satisfied approaches unity.

While the proof is deferred to [34], on a conceptual level this result can be understood by considering the distribution of order statistics. For example, given $n - 1$ samples from a uniform distribution between zero and some θ, with probability approaching one, the distance between the kth and $(k+1)$th-order statistic can be made arbitrarily large as θ moves towards infinity. Likewise, with the $J(a)$ distribution, the relative scaling between order statistics can be increased without bound as a decreases towards zero, leading to the stated result.

In conclusion, we have shown that a simple, (approximate) noninformative Jeffreys prior leads to sparse inverse problems that are optimally solved via Type II with high probability. Interestingly, it is this same Jeffreys prior that forms the generating coefficient prior of Type II when $f(\gamma_i) = 0$, e.g., the prior obtained by maximizing out $\boldsymbol{\gamma}$ in (6). However, it is worth mentioning that other Jeffreys prior-based techniques, e.g., direct minimization of $-\log p(\boldsymbol{x}) \propto \prod_i \log|x_i|$ subject to $\boldsymbol{y} = \Phi \boldsymbol{x}$, do *not* provide any Type II-like guarantees. Although several algorithms do exist that can perform such a minimization task (e.g., [15] and [19]), they perform poorly with respect to (15) in our experience because of convergence to bad local minimum as shown in [34]. This is still true if the coefficients are highly scaled. Section VI will analyze this issue in more detail.

C. Worst-Case Scenario

If the best-case scenario (no local minima) occurs when the nonzero generating coefficients are all of very different scales, it is reasonable to conjecture that the most difficult sparse inverse problem may involve nonzero coefficients with equal magnitudes. If we define $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ to be the vector of d nonzero magnitudes in some generating $\boldsymbol{x}$, then this implies that $\bar{x}_1 = \bar{x}_2 = \cdots \bar{x}_d$. This notion can be formalized somewhat by considering the $\bar{\boldsymbol{x}}$ distribution that is furthest from the Jeffreys prior. First, we note that the Type II cost function is effectively independent of the overall scaling of the generating coefficients, meaning $\alpha\bar{\boldsymbol{x}}$ is functionally equivalent to $\bar{\boldsymbol{x}}$ provided α is nonzero. This invariance must be taken into account in our analysis. Therefore, we assume the coefficients are rescaled such that $\sum_i \bar{x}_i = 1$.

Given this restriction, we can easily determine the distribution of nonzero coefficient magnitudes that is most different from the Jeffreys prior. Using the standard procedure for changing the parameterization of a probability density, the joint density of the constrained variables can be computed simply as

$$p(\bar{x}_1, \ldots, \bar{x}_d) \propto \frac{1}{\prod_{i=1}^d \bar{x}_i} \quad \text{for } \sum_{i=1}^d \bar{x}_i = 1, \bar{x}_i \geq 0, \ \forall i. \tag{39}$$

From this expression, it is easily shown that $\bar{x}_1 = \bar{x}_2 = \cdots = \bar{x}_d$ achieves the global minimum. Consequently, equal coefficient magnitudes are the absolute *least* likely to occur from the

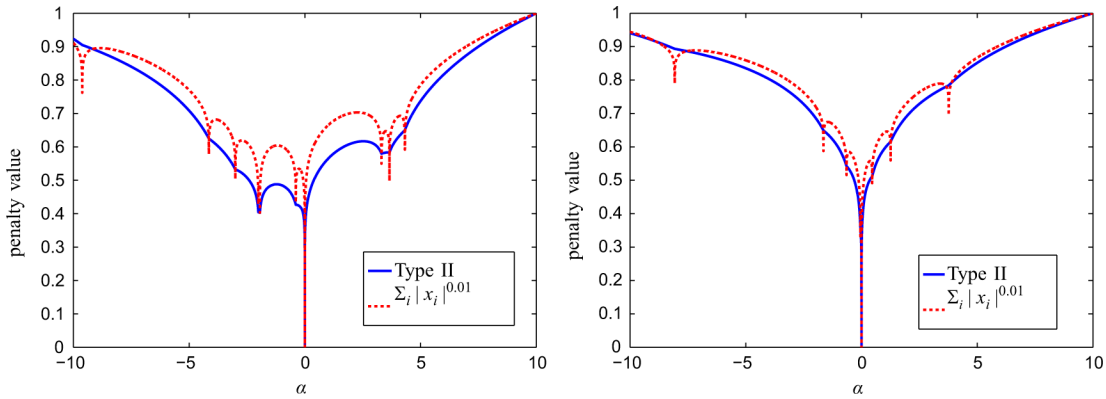

Fig. 5. Plots of the Type II penalty (normalized) across the feasible region as parameterized by α. A separable penalty given by $\mathfrak{g}(\boldsymbol{x}) \propto \sum_i |x_i|^{0.01} \approx \|\boldsymbol{x}\|_0$ is included for comparison. *Left:* Similar nonzero magnitudes (hard case). The Type II cost has 3 distinct local minima, but still many fewer than Type I. *Right:* Highly scaled nonzero magnitudes (easy case). Type II now has only a single minima at $\boldsymbol{x}_0$; the Type I example still has 10 minima (not all are shown).

Jeffreys prior. Hence, we may argue that the distribution that assigns $\bar{x}_i = 1/d$ with probability one is, in some sense, furthest from the constrained Jeffreys prior.

Nevertheless, because of the complexity of the Type II penalty, it is difficult to prove axiomatically that $\bar{\boldsymbol{x}} \sim \mathbf{1}$ is overall the most problematic distribution with respect to sparse recovery. However geometric considerations from [39] (omitted here for brevity) as well as illustrations from Section IV-D below support this conclusion. Regardless, it will be demonstrated in Section V that the worst-case performance of Type II is still better than common Type I approaches.

D. Illustration of Best- and Worse-Case Scenarios

Before proceeding to empirical results, it is insightful to observe directly the smoothing of local minima that leads to the best- and worst-case scenarios detailed in Sections IV-B and IV-C. To accomplish this, we repeat the exact same toy experiment from Section III-B, where we plotted penalty functions over a 1D feasible region parameterized by α. Using $\|\boldsymbol{x}_0\|_0 = 9$, we recreate Fig. 1 (right) with two minor alterations. First, in Fig. 5 (left), we take the square root of each nonzero coefficient magnitude, creating magnitudes with very similar scales (a more difficult situation). Second, in Fig. 5 (right), we square each nonzero magnitude, creating highly scaled coefficients (a more favorable situation). The effect then becomes very clear.

V. Empirical Results

The central purpose of this section is to present empirical evidence that supports our theoretical analysis and illustrates the improved performance afforded by Type II in solving (15) as various problem parameters are varied. We will focus our attention on the insights provided by Sections III and IV, comparing Type II (assuming $f(\gamma_i) = 0$ and $\lambda = 0$) with two standard Type I approaches, basis pursuit (BP) [8] and orthogonal matching pursuit (OMP) [32]. BP is the optimal convex approximation to (15) obtained by minimizing $\|\boldsymbol{x}\|_1$ subject to the constraint $\boldsymbol{y} = \Phi\boldsymbol{x}$; this can be solved using standard linear programming. In contrast, OMP is a greedy strategy for locally minimizing (15) that iteratively selects the basis vector most aligned with the current signal residual. At each step, a new approximant is formed by projecting $\boldsymbol{y}$ onto the range of all the selected dictionary columns. For the Type II implementation, we utilize an iterative reweighted ℓ_2 minimization technique based on convex upper bounds [36], which is equivalent to the EM implementation of sparse Bayesian learning (SBL) from [31] using $\lambda \to 0$.

Given a fixed distribution for the nonzero elements of $\boldsymbol{x}_0$, we will assess which algorithm is best (at least empirically) for most dictionaries relative to a uniform measure on the unit sphere, a metric relevant to compressive sensing.[8] To this effect, a number of Monte Carlo simulations were conducted, each consisting of the following: First, a random, overcomplete $n \times m$ dictionary Φ is created whose columns are each drawn uniformly from the surface of the unit sphere in $\mathbb{R}^n$. Next, sparse coefficient vectors $\boldsymbol{x}_0$ are randomly generated with d nonzero entries. Nonzero magnitudes $\bar{\boldsymbol{x}}_0$ are drawn i.i.d. from an experiment-dependent distribution. Signals are then computed as $\boldsymbol{y} = \Phi\boldsymbol{x}_0$. Each algorithm is presented with $\boldsymbol{y}$ and Φ and attempts to estimate $\boldsymbol{x}_0$. In all cases, we ran 1000 independent trials and compared the number of times each algorithm failed to recover $\boldsymbol{x}_0$. Under the specified conditions for the generation of Φ and $\boldsymbol{y}$, all other feasible solutions $\boldsymbol{x}$ almost surely have more nonzeros than d, so our synthetically generated $\boldsymbol{x}_0$ will be maximally sparse in practice. Moreover, Φ will almost surely satisfy $\text{spark}(\Phi) = n + 1$.

With regard to particulars, there are essentially four variables with which to experiment: (i) the distribution of $\bar{\boldsymbol{x}}_0$; (ii) the sparsity level d; (iii) the signal dimension n; and (iv) the number of dictionary columns m. In Fig. 6, we display results from an array of testing conditions. In each *row* of the figure, elements of $\bar{\boldsymbol{x}}_0$ are drawn i.i.d. from a fixed distribution; the first row uses unit nonzero coefficients, the second has elements drawn from $J(a = 0.001)$, and the third uses a unit Gaussian. In all cases, the signs of the nonzero coefficients are irrelevant due to the randomness inherent in the basis vectors.

The *columns* of Fig. 6 are organized as follows: The first column is based on the values $n = 50, d = 16$, while m is varied from n to $5n$, testing the effects of an increasing level of dictionary redundancy, m/n. The second fixes $n = 50$ and $m = 100$ while d is varied from 10 to 30, exploring the ability of each algorithm to resolve an increasing number of nonzero coefficients. Finally, the third column fixes $m/n = 2$ and $d/n \approx 0.3$ while n, m, and d are all increased proportionally. This demonstrates how performance scales with larger problem sizes.

[8] As will be explored in a future publication, however, the utility of Type II methods is more fully realized with highly structured dictionaries.

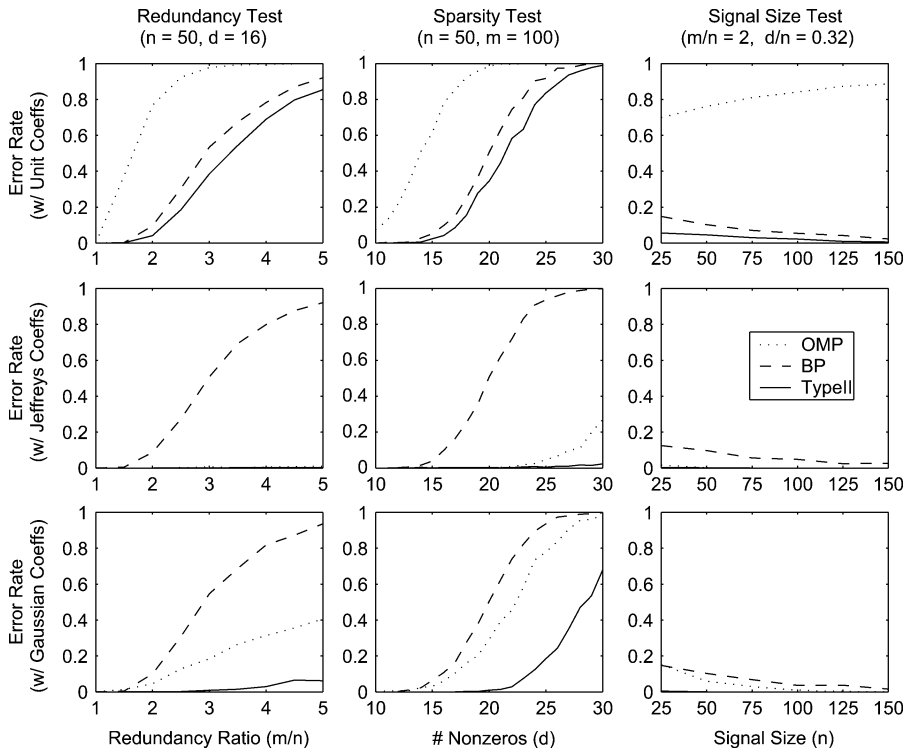

Fig. 6. Empirical results comparing the probability that OMP and BP (Type I methods), and Type II fail to find $\boldsymbol{x}_0$ under various testing conditions. Each data point is based on 1000 independent trials. The distribution of the nonzero coefficient magnitudes is labeled on the far left for each row, while the values for n, m, and d are included on the top of each column. Independent variables are labeled along the bottom of the figure. Note that in some cases the error rate is empirically indistinguishable from zero and therefore not visable.

The first row of plots essentially represents the worst-case scenario for Type II per our previous analysis, and yet performance is still consistently better than both BP and OMP. In contrast, the second row of plots approximates the best-case performance for Type II, where we see that Type II is almost infallible. The handful of failure events that do occur are because a is not sufficiently small and therefore, $J(a)$ was not sufficiently close to a true Jeffreys prior to achieve 100% success (see center plot). Finally, the last row of plots, based on Gaussian distributed coefficient amplitudes, reflects a balance between these two extremes. Nonetheless, Type II still holds a substantial advantage. In general, we observe that Type II is capable of handling more redundant dictionaries (column one) and resolving a larger number of nonzero coefficients (column two). Also, column three illustrates that it is able to recover a number of nonzero coefficients that grows linearly in the signal dimension.

By comparing row one, two and three, we observe that the performance of BP is independent of the coefficient magnitude distribution. This occurs because equivalence between the minimum ℓ_0-norm and minimum ℓ_1-norm solutions only depends on the sign pattern and sparsity profile of $\boldsymbol{x}_0$ [24]. This result suggests a potential limitation of BP, namely, it does not allow exploitation of the nonzero magnitudes (as Type II does) to increase the probability that we successfully recover $\boldsymbol{x}_0$. Moreover, BP performance is slightly below the worst-case Type II performance.

In contrast, like Type II, OMP results are highly dependent on the magnitude distribution. Unfortunately though, when the magnitude distribution is unity (top row), performance is unsatisfactory. In our experience, this appears to be a common problem with greedy methods designed to locally minimize (15). With highly scaled coefficients, OMP does considerably better than BP (middle row); however, the scale parameter a can never be adjusted such that OMP always succeeds (this can be proven using a simple toy counter-example [34]), and performance is significantly inferior to Type II. Finally, an additional weakness of OMP is that, unlike both Type II and BP, performance can potentially degrade as the problem size increases (upper right plot). Of course additional study is necessary to fully compare the relative performance of these methods on large-scale problems.

In summary, while the relative proficiency between OMP and BP is contingent on experimental particulars, Type II is uniformly superior in the cases we have tested (including examples not shown, e.g., results with other dictionary types).

VI. Conclusion

In this paper we have examined sparsity-promoting cost functions that emerge from a simple latent variable Bayesian model, emphasizing the distinction between Type I (MAP estimation) and Type II (empirical Bayes) approaches, demonstrating that the former is actually a special limiting case of the latter and that both can be equivalently expressed in either coefficient or latent variable space. This process allowed us to directly compare underlying cost functions and argue that there are many potential advantages of at least one flavor of Type II. While Bayesian considerations formed the starting point for these analyses, we should stress that the central underlying ideas regarding why Type II is so effective can be understood independently. More concretely, we do not actually believe that the unknown coefficients $\boldsymbol{x}$ are distributed as $p_{(II)}(\boldsymbol{x}) \propto \exp\left[-1/2\mathfrak{g}_{(II)}(\boldsymbol{x})\right]$ (exemplified by explicit statistical dependencies between elements) and that the validity of this assumption is the primary

reason for the success of Type II. Rather, we would argue that the Bayesian hierarchy upon which Type II is based represents a convenient fiction that happens to give rise to a useful class of sparsity-inducing cost functions. A similar point is raised in [3] regarding the performance of ℓ_1 solutions. Here success follows from desirable properties of the underlying convex cost function, not from the presumed Laplacian distribution of the unknown coefficients.

This perspective then allows us to consider alternative cost functions and manipulations of the implicit Type II sparsity penalty that may lose meaning in the context of the original Bayesian hierarchy but show promise on sparse estimation tasks. For example, we have shown that the nonseparable Type II penalty $\mathfrak{g}_{(II)}(\boldsymbol{x})$ is dependent on both the dictionary Φ and the noise variance λ. While meaningless from a Bayesian perspective, when we analyze the situation abstractly as a general form of Type I penalized regression, it becomes apparent that it could be beneficial to substitute alternative choices for λ or Φ in $\mathfrak{g}_{(II)}(\boldsymbol{x})$. In other words, if the only goal is to efficiently estimate global solutions to canonical sparse recovery problems, then it is not clear that the optimal selections are consistent with the original Bayesian model. Moreover, as we demonstrate in [36], other nonseparable penalty functions inspired by Type II but deviating from the Bayesian hierarchical derivation can be very effective as well.

All of this serves to motivate a wider class of cost functions for sparse estimation tasks and, in particular, allows us to exploit the fact that the distribution of nonzero coefficient magnitudes can drastically affect the difficulty of computing perfect sparse signal reconstructions. The popular minimum ℓ_1-norm solution (Type I) is completely blind to this distribution, and therefore exhibits performance below the worst-case regime possible via Type II. Note that neither method is given *a priori* knowledge of this distribution; rather, it is that Type II automatically operates more successfully when the distribution happens to be favorable. In general, we would argue that new sparse inverse algorithms should take these and related issues into account.

In [3] it is suggested that the distribution of nonzero coefficients is not really that important in a variety of practical situations such as image reconstruction. In its simplest form, the argument goes as follows. In some transform domain (e.g., wavelets) the coefficient distribution of many common images can be estimated and fit to a generalized Gaussian $p(\boldsymbol{x}) \propto \exp\left[-1/2\|\boldsymbol{x}\|_p^p\right]$, where the learned value of p tends to be significantly less than $p = 1$. However, when it comes to actually estimating $\boldsymbol{x}$ using a sparse recovery algorithm based on this learned value of p, i.e., solving a problem akin to (24) with $p < 1$ (Type I), performance is no better than when using $p = 1$. The authors conclude then that, given the validity of the assumption that $\boldsymbol{x}$ is sparse, the coefficient distribution of the nonzero elements is relatively inconsequential.

There are multiple reasons why this conclusion may not extend to general sparse inverse problems. For example, it is not clear that solving the Type I problem (24) is optimal since, with $p < 1$, any tractable minimization procedure will often be producing a local solution when Φ is overcomplete and performance may be no better than the standard, convex ℓ_1.[9] It is possible that an alternative procedure, potentially based on the ideas behind Type II, could do substantially better. Of course it is well-known in the Bayesian statistical literature that MAP estimation (Type I), even given the exact, generative prior, will often produce unsatisfactory results on inverse problems.

In any event, our results herein certainly suggest that the effect of coefficient distributions on performance can be important on general problems but to a degree that is highly algorithm-dependent. Two important questions are relevant in this regard:

1) Does the distribution of nonzero coefficients affect the performance of a given algorithm?
2) Assuming the true distribution of the nonzeros (or a close approximation) is known, what is the optimal sparse recovery algorithm?

The results of this paper speak directly to 1) as discussed above. Regarding 2), it is presently difficult to provide a concrete answer, although certainly we know that MAP can be suboptimal as already stated (this is an area of future research). To conclude this point then, the coefficient distribution may indeed matter in many practical situations, but only if exploited by an appropriate algorithm. Such an algorithm may or may not actually require knowledge of this distribution to succeed.

APPENDIX

This appendix contains the proofs of all results presented in this paper.

Proof of Theorem 1: From (4), (6), and (18) we have that

$$\begin{aligned} g(x_i) &= -2\log\left[\max_{\gamma_i \geq 0} \mathcal{N}(x_i; 0, \gamma_i)\varphi(\gamma_i)\right] \\ &\equiv \min_{\gamma_i \geq 0} \frac{x_i^2}{\gamma_i} + \log\gamma_i + f(\gamma_i) \end{aligned} \tag{40}$$

excluding terms independent of γ_i and x_i. Plugging this result into (20) we have

$$\begin{aligned} \mathcal{L}_{(I)}^{x}(\boldsymbol{x}) &= \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 \\ &\quad + \lambda \sum_i \left[\min_{\gamma_i \geq 0} \frac{x_i^2}{\gamma_i} + \log\gamma_i + f(\gamma_i)\right] \\ &\equiv \min_{\boldsymbol{\gamma} \geq 0} \frac{1}{\lambda}\|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} \\ &\quad + \sum_i \left[\log\gamma_i + f(\gamma_i)\right] \\ &\leq \frac{1}{\lambda}\|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} + \sum_i f_{(I)}(\gamma_i) \\ &\triangleq \mathcal{L}_{(I)}^{\gamma,x}(\boldsymbol{\gamma}, \boldsymbol{x}). \end{aligned} \tag{41}$$

[9]Note that the true distribution of *all* coefficients will not be a generalized Gaussian anyway once the zero-valued coefficients are taken into account. A more accurate description of this distribution would be a delta-function at zero and a (weighted) generalized-Gaussian distribution everywhere else. However, such a prior further exacerbates the problem of local minima. Of course if Φ is orthogonal, as is the case for standard image de-noising problems, local minima are generally not an issue.

Note that we allow $\gamma_i = 0$ when $x_i = 0$; for $x_i \neq 0$, $\gamma_i \to 0$ leads to infinity, so this value can never represent a minimizing solution. So $\mathcal{L}_{(I)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x})$ is a strict upper bound on $\mathcal{L}_{(I)}^{x}(\boldsymbol{x})$ with $\mathcal{L}_{(I)}^{x}(\boldsymbol{x}) = \min_{\boldsymbol{\gamma}\geq 0} \mathcal{L}_{(I)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x})$. With $\boldsymbol{\gamma}$ fixed, the unique value of $\boldsymbol{x}$ that minimizes $\mathcal{L}_{(I)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x})$ is given by $\boldsymbol{\mu}_x$ from (9), and from basic linear algebra manipulations we get

$$\min_{\boldsymbol{x}} \frac{1}{\lambda}\|\boldsymbol{y}-\Phi\boldsymbol{x}\|_2^2 + \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} = \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y}. \tag{42}$$

Using this expression with (41) gives

$$\mathcal{L}_{(I)}^{\gamma}(\boldsymbol{\gamma}) \triangleq \min_{\boldsymbol{x}} \mathcal{L}_{(I)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x}) = \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \sum_{i=1}^{m} f_{(I)}(\gamma_i). \tag{43}$$

Therefore, if $\boldsymbol{\gamma}_{(I)}$ minimizes (43), then by construction, it follows that

$$\boldsymbol{x}_{(I)} = \Gamma_{(I)}\Phi^T\left(\lambda I + \Phi\Gamma_{(I)}\Phi^T\right)^{-1}\boldsymbol{y} \tag{44}$$

will minimize (20). Likewise, if some $\boldsymbol{x}_{(I)}$ minimizes (20), then (44) must naturally hold for some $\boldsymbol{\gamma}_{(I)}$ that minimizes (43) (otherwise this $\boldsymbol{x}_{(I)}$ cannot be a global solution).

Additionally, the correspondence between global solutions to (20) and (43) extends to locally minimizing solutions as well in the following sense: it can be shown that $\{\boldsymbol{x}_*,\boldsymbol{\gamma}_*\}$ is a local minimum of the auxiliary function $\mathcal{L}_{(I)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x})$ iff $\boldsymbol{x}_*$ is a local minimum of (20) and $\boldsymbol{\gamma}_*$ is a local minimum of (43). This correspondence occurs because, given a fixed $\boldsymbol{x}$ (or $\boldsymbol{\gamma}$), optimization over $\boldsymbol{\gamma}$ (or $\boldsymbol{x}$) is actually convex (under the appropriate change of variables for the $\boldsymbol{\gamma}$ optimization) with a unique solution. So the local minima profile is preserved when we move from $\boldsymbol{x}$-space to $\boldsymbol{\gamma}$-space. ∎

Proof of Corollary 1: This is possible because we can always select a particular f and λ and then reparameterize things such that the $\log|\Sigma_y|$ term in (17) vanishes. Plugging $\lambda := \alpha^{-1}\bar{\lambda}$ and $f(\cdot) := \alpha\log[\alpha(\cdot)] + \alpha\bar{f}\,[\alpha(\cdot)]$ into (17), we have

$$\begin{aligned}\mathcal{L}_{(II)}^{\gamma}(\boldsymbol{\gamma}) =& \boldsymbol{y}^T\left[\alpha^{-1}\bar{\lambda}I + \Phi\Gamma\Phi^T\right]^{-1}\boldsymbol{y} \\ &+ \log\left|\alpha^{-1}\bar{\lambda}I + \Phi\Gamma\Phi^T\right| \\ &+ \sum_i \left(\alpha\log[\alpha\gamma_i] + \alpha\bar{f}\,[\alpha\gamma_i]\right) \\ \equiv& \boldsymbol{y}^T\left[\bar{\lambda}I + \alpha\Phi\Gamma\Phi^T\right]^{-1}\boldsymbol{y} \\ &+ \frac{1}{\alpha}\log\left|\bar{\lambda}I + \alpha\Phi\Gamma\Phi^T\right| \\ &+ \sum_i \left(\log[\alpha\gamma_i] + \bar{f}\,[\alpha\gamma_i]\right)\end{aligned}$$

and so as α becomes large

$$\begin{aligned}\mathcal{L}_{(II)}^{\gamma}(\boldsymbol{\gamma}) \to& \boldsymbol{y}^T\left[\bar{\lambda}I + \Phi\left(\alpha\Gamma\right)\Phi^T\right]^{-1}\boldsymbol{y} \\ &+ \sum_i \left(\log[\alpha\gamma_i] + \bar{f}\,[\alpha\gamma_i]\right).\end{aligned} \tag{45}$$

This is equivalent to (21) using $\lambda := \bar{\lambda}$ and $f := \bar{f}$ with the exception of the scaling factor of α on $\boldsymbol{\gamma}$. However, this factor is irrelevant in that the coefficient estimate obtained via (22) will be identical to that obtained from (11). ∎

Proof of Theorem 2: Using (17) and (42), we can create a strict upper bounding auxiliary function on $\mathcal{L}_{(II)}^{\gamma}(\boldsymbol{\gamma})$ given by

$$\mathcal{L}_{(II)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x}) \triangleq \frac{1}{\lambda}\|\boldsymbol{y}-\Phi\boldsymbol{x}\|_2^2 + \sum_i \frac{x_i^2}{\gamma_i} + \log|\Sigma_y| + \sum_i f(\gamma_i) \tag{46}$$

where $\mathcal{L}_{(II)}^{\gamma}(\boldsymbol{\gamma}) = \min_{\boldsymbol{x}} \mathcal{L}_{(II)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x})$ for all $\boldsymbol{\gamma} \geq 0$. When we minimize over $\boldsymbol{\gamma}$, we get

$$\mathcal{L}_{(II)}^{x}(\boldsymbol{x}) \triangleq \min_{\boldsymbol{\gamma}\geq 0} \mathcal{L}_{(II)}^{\gamma,x}(\boldsymbol{\gamma},\boldsymbol{x}) \equiv \|\boldsymbol{y}-\Phi\boldsymbol{x}\|_2^2 + \lambda\mathfrak{g}_{(II)}(\boldsymbol{x}) \tag{47}$$

with

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) \triangleq \min_{\boldsymbol{\gamma}\geq 0} \sum_i \frac{x_i^2}{\gamma_i} + \log|\Sigma_y| + \sum_i f(\gamma_i). \tag{48}$$

Therefore analogous to the proof of Theorem 1, if $\boldsymbol{x}_{(II)}$ is a minimum of (47), then there must exist some $\boldsymbol{\gamma}_{(II)}$ that minimizes $\mathcal{L}_{(II)}^{\gamma}(\boldsymbol{\gamma})$ such that

$$\boldsymbol{x}_{(II)} = \Gamma_{(II)}\Phi^T\left(\lambda I + \Phi\Gamma_{(II)}\Phi^T\right)^{-1}\boldsymbol{y}. \tag{49}$$

Likewise, if $\boldsymbol{\gamma}_{(II)}$ minimizes $\mathcal{L}_{(II)}^{\gamma}(\boldsymbol{\gamma})$, then $\boldsymbol{x}_{(II)}$ computed via (49) will minimize (47).

While $\mathfrak{g}_{(II)}(\boldsymbol{x})$ is not generally available in closed form, if $f[\exp(\cdot)]$ is convex then the optimization problem from (48) will have a single basin of attraction (meaning that all minima are connected in the rare case that multiple exist), and even convex with the reparameterization of $\boldsymbol{\gamma}$ given by $\beta_i \triangleq \log\gamma_i, \boldsymbol{\beta} \triangleq [\beta_1,\ldots,\beta_m]^T$. It can then be shown that the minimization problem

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) \equiv \min_{\boldsymbol{\beta}} \sum_i \mathrm{e}^{-\beta_i}x_i^2 + \log|\Sigma_y| + \sum_i f\left(\mathrm{e}^{\beta_i}\right) \tag{50}$$

is convex in $\boldsymbol{\beta}$, meaning that no unconnected local minima can exist (although it still need not be convex in $\boldsymbol{\gamma}$). This implies that there will be a correspondence between local minima of (47) and local minima of (17), analogous to the duality situation for Type I discussed in the previous section (the global minimum will of course correspond for a wide range of f). ∎

Proof of Theorem 3: $\log|\cdot|$ is concave in the space of positive semi-definite matrices [20]. Moreover, Σ_y is an affine function of $\boldsymbol{\gamma}$ and is positive semidefinite for any $\boldsymbol{\gamma} \geq 0$. This implies that $\log|\Sigma_y|$ is a concave, nondecreasing function of $\boldsymbol{\gamma}$, so we can express it as

$$\log|\Sigma_y| = \min_{\boldsymbol{z}\geq 0} \boldsymbol{z}^T\boldsymbol{\gamma} - h^*(\boldsymbol{z}) \tag{51}$$

where $h^*(\boldsymbol{z})$ is the concave conjugate [2] of $\log|\Sigma_y|$ given by

$$h^*(\boldsymbol{z}) \triangleq \min_{\boldsymbol{\gamma}\geq 0} \boldsymbol{z}^T\boldsymbol{\gamma} - \log|\Sigma_y|. \tag{52}$$

Therefore, we can express $\mathfrak{g}_{(II)}(\boldsymbol{x})$ as

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) = \min_{\boldsymbol{\gamma},\boldsymbol{z}\geq 0} \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} + \boldsymbol{z}^T\boldsymbol{\gamma} - h^*(\boldsymbol{z}). \tag{53}$$

Optimizing over $\boldsymbol{\gamma}$ for fixed $\boldsymbol{x}$ and $\boldsymbol{z}$, we get

$$\gamma_i^{\text{opt}} = z_i^{-1/2}|x_i|, \forall i. \tag{54}$$

Substituting this expression into (53) gives

$$\begin{aligned}\mathfrak{g}_{(II)}(\boldsymbol{x}) &= \min_{\boldsymbol{z}\geq 0}\left[\sum_{i=1}^{m}\left(\frac{x_i^2}{z_i^{-1/2}|x_i|}+z_i z_i^{-1/2}|x_i|\right)-h^*(\boldsymbol{z})\right]\\ &= \min_{\boldsymbol{z}\geq 0}\sum_{i=1}^{m}2z_i^{1/2}|x_i|-h^*(\boldsymbol{z}).\end{aligned} \tag{55}$$

This latter expression represents $\mathfrak{g}_{(II)}(\boldsymbol{x})$ as a minimum over upper-bounding hyperplanes in $|\boldsymbol{x}|$ (meaning each value of $\boldsymbol{z}$ defines a unique hyperplane with respect to $|\boldsymbol{x}|$). From basic convex analysis, any function expressable in this form is necessary concave, and since $\boldsymbol{z}^{1/2}\geq 0$, nondecreasing as well [2].

Finally, the local minima result follows directly from [28, Theorem 1], which is derived for general Type I methods but can be applied to any penalty function such as $\mathfrak{g}_{(II)}(\boldsymbol{x})$ that is a concave and nondecreasing function of each $|x_i|$. ∎

Proof of Theorem 4: In [34], we show an equivalent result using manipulations of $\mathcal{L}_{(II)}^{\gamma}(\boldsymbol{\gamma})$ in $\boldsymbol{\gamma}$-space. Here we present a much simpler, high-level proof directly in $\boldsymbol{x}$-space.

The set of global minimizers of (29) is a subset of the local minimizers, which from Theorem 3 occur at basic feasible solutions (BFS) involving at most n nonzero elements. Therefore when searching for global minima and their associated properties, we need only consider these solutions. At any BFS with $d \triangleq \|\boldsymbol{x}\|_0 = n$, $\mathfrak{g}_{(II)}(\boldsymbol{x}) \geq C_1$, where C_1 is a Φ- and $\boldsymbol{y}$-dependent (but $\boldsymbol{x}$-independent) constant given by

$$\begin{aligned}C_1 &\triangleq \min_{\boldsymbol{x}:\boldsymbol{y}=\Phi\boldsymbol{x},\|\boldsymbol{x}\|_0=n}\left(\min_{\boldsymbol{\gamma}\geq 0}\sum_i\frac{x_i^2}{\gamma_i}+\log|\Phi\Gamma\Phi^T|\right),\\ &\leq \min_{\boldsymbol{x}:\boldsymbol{y}=\Phi\boldsymbol{x},\|\boldsymbol{x}\|_0=n}\left(\min_{\boldsymbol{\gamma}\geq 0}\sum_i\frac{x_i^2}{\gamma_i}+\log|\Sigma_y|\right)\end{aligned} \tag{56}$$

where the latter inequality holds for any λ, including the limit $\lambda\rightarrow 0$. This constant C_1 will always exist and be finite given the assumption that $\text{spark}(\Phi) = n+1$. To see this, note that (56) involves a minimization over two terms with respect to $\boldsymbol{\gamma}$. The first term (convex, nonincreasing) encourages each γ_i to be large, the second (concave, nondecreasing) encourages each γ_i to be small. Whenever a given $x_i = 0$, the first term can be ignored and the associated γ_i is driven to exactly zero by the second term regardless of other $\gamma_j, j\neq i$. In contrast, for any $x_i\neq 0$, the minimizing γ_i can never be zero or the first term will be driven to infinity. This a manifestation of the fact that $\arg\min_{z\geq 0}\left[\frac{1}{z}+\log z\right] > 0$. Consequently, for any given $\boldsymbol{x}$, the associated minimizing $\boldsymbol{\gamma}$ will necessarily have a matching sparsity profile, meaning the indices of zero-valued x_i will align with zero-valued elements in $\boldsymbol{\gamma}$. Moreover there is no issue with dividing by zero in the first term and Σ_y will always be full rank in the second term (the latter because of the spark assumption and the fact that $d = n$). Therefore C_1 will always be finite for essentially the same reason that $\min_{z\geq 0}\left[\frac{1}{z}+\log z\right]$ is finite.

In contrast, at any $\boldsymbol{x}$ with $d < n$ the situation is very different. Let $\mathcal{S}$ denote the support set of $\boldsymbol{x}$, meaning the index set $\{i : x_i\neq 0\}$, and let $s_i(\boldsymbol{\gamma})$ indicate the ith nonzero eigenvalue of $\Phi\Gamma\Phi^T$. The spark assumption, coupled with the analysis above, guarantees that there will be d such nonzero eigenvalues at any minimizing $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ such that

$$\log|\Sigma_y| = \sum_{i=1}^{d}\log[s_i(\boldsymbol{\gamma}^*)+\lambda]+(n-d)\log\lambda. \tag{57}$$

We can now rewrite $\mathfrak{g}_{(II)}(\boldsymbol{x})$ as

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) = \sum_{i\in\mathcal{S}}\frac{x_i^2}{\gamma_i^*}+\sum_{i=1}^{d}\log[s_i(\boldsymbol{\gamma}^*)+\lambda]+(n-d)\log\lambda. \tag{58}$$

From this expression it follows that whenever $d < n$ we have

$$\mathfrak{g}_{(II)}(\boldsymbol{x}) \leq C_2 + (n-d)\log\lambda. \tag{59}$$

for any $\lambda < \lambda_0$, where C_2 is also a Φ- and $\boldsymbol{y}$-dependent (but $\boldsymbol{x}$-independent) constant given by

$$C_2 \triangleq \max_{\boldsymbol{x}:\boldsymbol{y}=\Phi\boldsymbol{x},\|\boldsymbol{x}\|_0<n}\left(\min_{\boldsymbol{\gamma}\geq 0}\sum_i\frac{x_i^2}{\gamma_i}+\sum_{i=1}^{\|\boldsymbol{x}\|_0}\log[s_i(\boldsymbol{\gamma})+\lambda_0]\right). \tag{60}$$

So in summary then, we know that any global minimum must occur at a BFS such that either $\mathfrak{g}_{(II)}(\boldsymbol{x}) \geq C_1$ if $d = n$ or $\mathfrak{g}_{(II)}(\boldsymbol{x}) \leq C_2+(n-d)\log\lambda$ if $d < n$. While the relative sizes of C_1 and C_2 are unknown, they are both fixed, finite constants and so as $\lambda\rightarrow 0$, as stipulated by (29), the global minimum must occur when $d < n$. In fact, using a similar process it can also be shown that $\mathfrak{g}_{(II)}(\boldsymbol{x}) \geq C_3 + (n-d)\log\lambda$ as well for come constant C_3, which then enforces that the global minimum can only occur when d is smallest. Therefore minimizing $\mathfrak{g}_{(II)}(\boldsymbol{x})$ in these conditions is tantamount to minimizing d, and so any global solution to (29) will be a global solution to (15). ∎

Proof of Theorem 5: We begin by assuming that $g(x_i)$ is a concave, nondecreasing function of $|x_i|$.[10] With some additional effort, can be shown that the theorem holds in the general case as well, consistent with intuition. We will also assume, without loss of generality, that $g(0) = 0$ and $g(1) = 1$ (we can always rescale and add a constant such that this is the case). A simple 3D example then serves to show that conditions (i) and (ii) cannot be satisfied simultaneously.

Assume we have a 3×5 dictionary Φ where the first two columns are given by $\boldsymbol{\phi}_1 \propto [1\,\alpha\,0]^T$ and $\boldsymbol{\phi}_2 \propto [-1\,\alpha\,0]^T$, with $\alpha > 0$ arbitrarily small (we use a proportionality here to avoid the irrelevant, cumbersome factor required for ℓ_2 column normalization). Now assume a coefficient vector $\boldsymbol{x}^{(1)} \triangleq [1\,1\,0\,0\,0]^T$, giving $\boldsymbol{y} = \Phi\boldsymbol{x}^{(1)} = [0\,2\alpha\,0]^T$, and that the remaining three basis vectors $\boldsymbol{\phi}_3, \boldsymbol{\phi}_4, \boldsymbol{\phi}_5$, are radially symmetric about the signal $\boldsymbol{y}$, with an equal, arbitrarily small angular distance from $\boldsymbol{y}$.[11] Then a second feasible solution $\boldsymbol{x}^{(2)} \triangleq [0\,0\,\epsilon\,\epsilon\,\epsilon]^T$ exists with ϵ, a function of an arbitrarily small α, also arbitrarily small.

[10]Any penalty arising from (6) will be concave, nondecreasing function of x_i^2, but not necessarily of $|x_i|$.

[11]By radially symmetric about $\boldsymbol{y}$, we mean they are all unique, equidistant from $\boldsymbol{y}$, and equidistant from one another; loosely they can be visualized as forming a tight equilateral triangle around $\boldsymbol{y}$.

Under these circumstances, $\boldsymbol{x}^{(1)}$ equals $\boldsymbol{x}_0$, the unique global solution to (15). To satisfy condition (i), it is therefore necessary that

$$\sum_i g\left(x_i^{(1)}\right) = 2g(1) < \sum_i g\left(x_i^{(2)}\right) = 3g(\epsilon) \tag{61}$$

or equivalently, that $g(\epsilon) > 2/3$, $\forall \epsilon > 0$. We now show that any g that satisfies this restriction cannot have fewer local minimum than when solving (29). So if we satisfy condition (i), we cannot simultaneously satisfy condition (ii).

A basic feasible solution $\boldsymbol{x}^*$ is a local minimizer of (30) if for every vector $\boldsymbol{v} \in \text{null}(\Phi)$, there is a $\delta > 0$ such that

$$d(\varepsilon) \triangleq \sum_i g(x_i^* + \varepsilon v_i) - \sum_i g(x_i^*) > 0, \ \forall \varepsilon \in (0, \delta]. \tag{62}$$

Based on the concavity of g with respect to $|x_i|$, we know that local minima are always achieved at basic feasible solutions with at least $m - n$ elements equal to zero. Consequently, we can express $d(\varepsilon)$ as

$$\begin{aligned} d(\varepsilon) &= \sum_{i \in \mathcal{Z}} [g(\varepsilon v_i) - g(0)] + \sum_{i \notin \mathcal{Z}} [g(x_i^* + \varepsilon v_i) - g(x_i^*)] \\ &= \sum_{i \in \mathcal{Z}} g(\varepsilon v_i) + \sum_{i \notin \mathcal{Z}} [g(x_i^* + \varepsilon v_i) - g(x_i^*)] \end{aligned} \tag{63}$$

where $\mathcal{Z}$ is the set of all indeces associated with zero-valued elements in $\boldsymbol{x}^*$. As a direct consequence of the assumption $\text{spark}(\Phi) = n + 1$, any $\boldsymbol{v} \in \text{null}(\Phi)$ must have a nonzero element corresponding to a zero element in $\boldsymbol{x}^*$, meaning at least one $v_i, i \in \mathcal{Z}$ must be nonzero. Therefore, the first term in (63) cannot be smaller than $2/3$ or we violate condition (i) as discussed above. Moreover, because g is concave on $[0, \infty)$, it must be continuous on $(0, \infty)$. Consequently, the second term in (63) can be made arbitrarily small in magnitude for any $\varepsilon \in (0, \delta]$ when δ is sufficiently small, implying that $d(\varepsilon)$ will always be positive. Thus $\boldsymbol{x}^*$ must be a local minimizer of (30).

In conclusion then, any g which satisfies condition (i) will have a local minimum at every basic feasible solution. Moreover, from Theorem 3, the number of distinct basic feasible solutions forms an upper bound to the number of distinct local minima to (29). Of course with the exception of very contrived situations, the number of Type II local minima will be considerably less as discussed in Sections III and IV. ■

Proof of Theorem 6: By virture of Theorem 2, the unimodality of (26) is revealed by examining the dual cost function (17) in $\boldsymbol{\gamma}$-space, which conveniently decouples because of the orthogonality assumption. This produces the element-wise cost

$$\mathcal{L}_{(II)}^{\gamma_i}(\gamma_i) \triangleq \log(\lambda + \gamma_i) + \frac{a_i^2}{\lambda + \gamma_i}, \ \forall i \tag{64}$$

where $a_i \triangleq \boldsymbol{\phi}_i^T \boldsymbol{y}$ and $\boldsymbol{\phi}_i$ is the ith column of Φ. This expression is readily shown to be unimodal in each γ_i, implying unimodality over $\boldsymbol{\gamma}$.

The second property follows by taking the gradient of (32) with respect to x_i and noting that it is positive and decreasing for all $x_i \in (0, \infty)$. We also note that any penalty $\mathfrak{g}(\boldsymbol{x})$ that is a nondecreasing and strictly concave function of $|\boldsymbol{x}|$, will both promote sparsity [28] and provide a tighter approximation to $\|\boldsymbol{x}\|_0$ than $\|\boldsymbol{x}\|_1$ in the following sense: There will always exist some positive constant $R < \infty$ such that, for any sphere $\mathcal{S}_r$ in $\mathbb{R}^m$ centered at zero with radius $r > R$, we have that

$$\begin{aligned} \min_{a,b} \int_{x \in \mathcal{S}_r} &\left| \|\boldsymbol{x}\|_0 - (a\mathfrak{g}(\boldsymbol{x}) + b) \right| d\boldsymbol{x} \\ &< \min_{a,b} \int_{x \in \mathcal{S}_r} \left| \|\boldsymbol{x}\|_0 - (a\|\boldsymbol{x}\|_1 + b) \right| d\boldsymbol{x}. \end{aligned} \tag{65}$$

In other words, the approximation error will always be smaller, assuming we adjust the slope and offset appropriately using a and b, as long as we average over a large enough region. This follows directly from the definition of strict concavity (and the implicit assumption that $\mathfrak{g}(\boldsymbol{0})$ is finite).

The third property can be shown by contradiction. Assume that $\mathfrak{g}_{(II)}(\boldsymbol{x})$ is a nondecreasing and strictly concave function of $|\boldsymbol{x}|$, but is fixed and independent of λ as in Type I methods. We will show that multiple minima are always possible for some choice of λ, Φ, and $\boldsymbol{y}$. Given the orthogonality assumption, $\mathcal{L}_{(I)}^x(\boldsymbol{x})$ decouples and we can consider each coordinate separately with the reduced cost function

$$\mathcal{L}_{(I)}^{x_i}(x_i) \triangleq x_i^2 - 2x_i a_i + \lambda g(x_i). \tag{66}$$

For simplicity, we will assume that g is differentiable, but the more general case follows with a little additional effort. We will also assume, without loss of generality that $a_i \geq 0, \forall i$. Now consider

$$\mathcal{L}_{(I)}^{x_i}{}'(x_i) \triangleq \frac{d\mathcal{L}_{(I)}^x(x_i)}{dx_i} = 2x_i - 2a_i + \lambda g'(x_i) \tag{67}$$

with $g'(x_i) \triangleq dg(x_i)/dx_i$. If $\mathcal{L}_{(I)}^{x_i}{}'(x_i)$ is positive as $x_i \longrightarrow 0^+$, then there will necessarily be one local minimum at $x_i = 0$. A second local minimum will also occur if $\mathcal{L}_{(I)}^{x_i}{}'(x_i) < 0$ for some $x_i > 0$. This is because $\mathcal{L}_{(I)}^{x_i}{}'(x_i)$ must be greater than zero for some x_i sufficiently large due to the concavity of $g(x_i)$ with respect to $|x_i|$, and so a negative gradient for smaller values of x_i implies a local minimum must exist in the middle somewhere. Therefore we only need show that both minima are possible simultaneously.

To have a minimum at $x_i = 0$, it is sufficient based on the positive gradient requirement that $a_i = \lambda g'(0)/2 - \epsilon$, where $g'(0) \triangleq \lim_{x_i \longrightarrow 0^+} g'(x_i)$, and $\epsilon > 0$ is a small constant such that a_i is positive. A second minimum will occur if

$$\mathcal{L}_{(I)}^{x_i}{}'(x_i) = 2x_i - 2\left[\lambda g'(0)/2 - \epsilon\right] + \lambda g'(x_i) < 0 \tag{68}$$

for some $x_i > 0$. We can always satisfy this inequality for some λ sufficiently large since $g'(0) > g'(x_i)$ by definition of strict concavity. Consequently, two local minima are always possible for each i, giving 2^m total local minima as an upper bound, which is trivially achieved when the ℓ_0 norm is used (forth property). ■

Proof of Theorem 7: If $\boldsymbol{x}_*$ is a nondegenerate locally minimizing solution to (29), then there is an associated $\boldsymbol{\gamma}_*$, with

matching sparsity profile, that locally minimizes $\mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})$ with $\lambda = 0$. For this to be true, the following necessary condition must hold for all $\boldsymbol{u} \in \mathcal{U}$:

$$\left.\frac{\partial \mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})}{\partial \gamma_u}\right|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_*} \geq 0 \tag{69}$$

where γ_u denotes the latent variable corresponding to the basis vector $\boldsymbol{u}$. In other words, we cannot reduce $\mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})$ along a positive gradient because this would push γ_u below zero; a negative gradient would imply that γ_u can be increased to further reduce $\mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})$, meaning a local minima is impossible. Using the matrix inversion lemma, a determinant identity, and some algebraic manipulations, we arrive at the expression

$$\left.\frac{\partial \mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})}{\partial \gamma_u}\right|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_*} = \frac{\boldsymbol{u}^T B \boldsymbol{u}}{1+\gamma_u^* \boldsymbol{u}^T B \boldsymbol{u}} - \left(\frac{\boldsymbol{y}^T B \boldsymbol{u}}{1+\gamma_u^* \boldsymbol{u}^T B \boldsymbol{u}}\right)^2 \tag{70}$$

where $B \triangleq (\tilde{\Phi}\tilde{\Gamma}\tilde{\Phi}^T)^{-1}$ and $\tilde{\Gamma}$ is the diagonal matrix of latent variables associated with $\tilde{\Phi}$. Since we have assumed that we are at a local minimum, it is straightforward to show[12] that $\tilde{\Gamma} = \mathrm{diag}(\tilde{\boldsymbol{x}})^2$ leading to the expression

$$B = \tilde{\Phi}^{-T} \mathrm{diag}(\tilde{\boldsymbol{x}})^{-2} \tilde{\Phi}^{-1}. \tag{71}$$

Substituting this expression into (70) and evaluating at the point $\gamma_u^* = 0$, the above gradient reduces to

$$\left.\frac{\partial \mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})}{\partial \gamma_u}\right|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_*} = \tilde{\boldsymbol{v}}^T \left(\mathrm{diag}(\tilde{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{x}}^{-T}) - \tilde{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{x}}^{-T}\right)\tilde{\boldsymbol{v}} \tag{72}$$

where $\tilde{\boldsymbol{x}}^{-1} \triangleq [\tilde{x}_1^{-1}, \ldots, \tilde{x}_n^{-1}]^T$. This implies that we will be at a local minimum only if

$$\sum_{i \neq j} \frac{\tilde{v}_i \tilde{v}_j}{\tilde{x}_i \tilde{x}_j} \leq 0 \quad \forall \tilde{\boldsymbol{v}} \in \mathcal{V} \tag{73}$$

which leads directly to the stated theorem. ∎

Proof of Theorem 8: Every local minimum of (29) is achieved at a basic feasible solution (BFS) (see Theorem 3). Interestingly, the converse is not true; that is, not every BFS need correspond with a minimum of (29) as shown via Theorem 7. In fact, for a suitable selection of scaling constants ν_i, we will show that this reduced set of minima naturally leads to a proof of Theorem 8. In the most general setting, the constants ν_i would all be as large as possible, leading to the largest set of allowable coefficients. However, for the proof it is sufficient to assume that $\nu_1 = \nu_2 = \cdots = \nu_{n-2} = \epsilon$, where ϵ is a constant in the interval $(0, 1]$.

We begin with an arbitrary coefficient vector $\boldsymbol{x}'$ such that $x'_{(i+1)} \leq \epsilon x'_{(i)}$ and $||\boldsymbol{x}'||_0 = d \in \{1, \ldots, n-1\}$. For convenience, we will also assume that $x'_{(i)} = |x'_i|$. In other words, the first element of $\boldsymbol{x}'$ has the largest magnitude, the second element has the second largest magnitude, and so on. To avoid any loss of generality, we incorporate an $m \times m$ permutation matrix P into our generative model, giving us the signal $\boldsymbol{y} = \Phi P \boldsymbol{x}' = \Phi' \boldsymbol{x}'$. Because $\Phi' \triangleq \Phi P$ is nothing more than Φ with reordered columns, it will necessarily satisfy the spark constraint for all P given that Φ does.

We now examine the properties of an arbitrary BFS with nonzero coefficients defined as $\bar{\boldsymbol{x}}$ (so the length of $\bar{\boldsymbol{x}}$ is less than or equal to n by definition of a BFS), and associated dictionary columns $\bar{\Phi}$ ordered as in Φ', i.e., $\boldsymbol{y} = \bar{\Phi}\bar{\boldsymbol{x}}$. There exist two possibilities for a candidate BFS:

- Case I: The columns of Φ' associated with the nonzero $d < n$ nonzero coefficients of $\boldsymbol{x}'$ are contained in $\bar{\Phi}$. By virtue of the spark assumption, no other basis vectors will be present, so we may conclude that $\bar{\Phi} = [\boldsymbol{\phi}'_1, \boldsymbol{\phi}'_2, \ldots, \boldsymbol{\phi}'_d] = \Phi'$.
- Case II: At least one of the columns associated with the d nonzero coefficients is missing from $\bar{\Phi}$.

Given this distinction, we would like to determine when a candidate BFS, particularly a Case II BFS of which there are many, is a local minimum.

To accomplish this, we let $r \in \{1, \ldots, d\}$ denote the index of the largest coefficient magnitude for which the respective dictionary column, $\boldsymbol{\phi}'_r$ is *not* in $\bar{\Phi}$. Therefore, we may assume that the first $r-1$ columns of $\bar{\Phi}$ equal $[\boldsymbol{\phi}'_1, \boldsymbol{\phi}'_2, \ldots, \boldsymbol{\phi}'_{r-1}]$. The remaining columns of $\bar{\Phi}$ are arbitrary (provided of course that $\boldsymbol{\phi}'_r$ is not included). This allows us to express any Case II BFS as

$$\bar{\boldsymbol{x}} = \bar{\Phi}^{-1}\boldsymbol{y} = \bar{\Phi}^{-1}\Phi'\boldsymbol{x}' = \sum_{k=1}^{r-1} x'_k \boldsymbol{e}_k + \bar{\Phi}^{-1} \sum_{k=r}^{d} x'_k \boldsymbol{\phi}'_k \tag{74}$$

where $\boldsymbol{e}_k$ is a zero vector with a one in the kth element and we have assumed that every Case II BFS utilizes exactly n columns of Φ' (i.e., $\bar{\Phi}$ is $n \times n$ and therefore invertible via the spark requirement). This assumption is not restrictive provided we allow for zero-padding of BFS with less than n nonzero coefficients (this implies that some elements of $\bar{\boldsymbol{x}}$ will be equal to zero if we have to add dummy columns to $\bar{\Phi}$).

Without loss of generality, we will assume that $x'_r = 1$ (the overall scaling is irrelevant). We also define $\tilde{\boldsymbol{v}} \triangleq \bar{\Phi}^{-1}\boldsymbol{\phi}'_r$, giving us

$$\bar{\boldsymbol{x}} = \bar{\Phi}^{-1}\boldsymbol{y} = \sum_{k=1}^{r-1} x'_k \boldsymbol{e}_k + \tilde{\boldsymbol{v}} + \bar{\Phi}^{-1} \sum_{k=r+1}^{d} x'_k \boldsymbol{\phi}'_k. \tag{75}$$

By virtue of the stipulated ϵ-dependent coefficient scaling, we know that

$$\bar{\Phi}^{-1} \sum_{k=r+1}^{d} x'_k \boldsymbol{\phi}'_k = \sum_{k=r+1}^{d} \mathrm{O}_n\left(\epsilon^{k-r}\right) = \mathrm{O}_n\left(\epsilon\right) \tag{76}$$

Here we adopt the notation $f(x) = \mathrm{O}(h(\epsilon))$ to indicate that $|f(x)| < C|h(\epsilon)|$ for some constant C independent of x or ϵ. $\mathrm{O}_n(h(\epsilon))$ then refers to an n-dimensional vector with all elements of order $\mathrm{O}(h(\epsilon))$. Combining (75) and (76), we can express the ith element of $\bar{\boldsymbol{x}}$ as

$$\bar{x}_i = x'_i \mathrm{I}\left[i < r\right] + \tilde{v}_i + \mathrm{O}\left(\epsilon\right). \tag{77}$$

Provided ϵ is chosen suitably small, we can ensure that all $\bar{x}_i$ are necessarily nonzero (so in fact no zero-padding is ever necessary). When $i \geq r$, this occurs because all elements of $\tilde{\boldsymbol{v}}$ must be strictly nonzero or we violate the spark assumption. For the

[12]At any local minimum, $\mathcal{L}^{\gamma}_{(II)}(\boldsymbol{\gamma})$ must be minimized with respect to $\tilde{\Gamma}$, assuming all other elements of $\boldsymbol{\gamma}$ are equal to zero. Given the stipulated conditions, this is a simple matter since the cost function conveniently decouples into n separate functions, giving optimal values $\bar{\gamma}_i = \bar{x}_i^2$.

$i < r$ case, a sufficiently small ϵ means that the x_i' term (which is of order $\mathrm{O}\left(1/\epsilon^{r-i}\right)$ by virtue of (36)) will dominate, leading to a nonzero $\bar{x}_i$. This allows us to apply Theorem 7, from which we can conclude that a candidate BFS with n nonzero coefficients will not represent a local minimum if

$$\sum_{i\neq j} \frac{\tilde{v}_i \tilde{v}_j}{\bar{x}_i \bar{x}_j} > 0. \tag{78}$$

Substituting (77) into this criterion, we obtain

$$\sum_{i\neq j} \left(\frac{\tilde{v}_i}{x_i' \mathbf{I}[i<r] + \tilde{v}_i + \mathrm{O}(\epsilon)}\right)\left(\frac{\tilde{v}_j}{x_j' \mathbf{I}[j<r] + \tilde{v}_j + \mathrm{O}(\epsilon)}\right) = \mathrm{O}(\epsilon) + \sum_{i\neq j; i,j\geq r} \left(\frac{\tilde{v}_i}{\tilde{v}_i + \mathrm{O}(\epsilon)}\right)\left(\frac{\tilde{v}_j}{\tilde{v}_j + \mathrm{O}(\epsilon)}\right). \tag{79}$$

Since $d < n$, then $r < n$ by definition and so there will always be at least one set of indices i and j that satisfy the above summation constraints (since both i and j run from 1 to n). This then implies that

$$\sum_{i\neq j} \frac{\tilde{v}_i \tilde{v}_j}{\bar{x}_i \bar{x}_j} \approx \sum_{i\neq j; i,j\geq r} 1 > 0 \tag{80}$$

since each $\tilde{v}_i$ is a nonzero constant independent of ϵ. So (78) holds and we are not at a local minimum.

In summary, we have shown that, provided ϵ is small enough, an arbitrary Case II BFS cannot be a local minimum to (29). The exact value of this ϵ will depend on the particular BFS and permutation matrix P. However, if we choose the smallest ϵ across all possibilities, it follows that no Case II BFS can be a local minimum. The unique minimum that remains is the Case I BFS which will satisfy $d = \|\boldsymbol{x}_0\|_0$, so $\boldsymbol{x}' = \boldsymbol{x}_0$. ∎

Proof of Theorem 9: We assume $g(x_i)$ is a nondecreasing, concave function of $|x_i|$; with other allowable choices it can be shown that the global minimum will not generally equal $\boldsymbol{x}_0$. For the special case where $g(x_i) = |x_i|$, $\forall i$, the cost function is convex; however, regardless of nonzero coefficient scalings, any global minimum need not be maximally sparse under the stated conditions. This directly follows from [24, Theorem 6], from which we can infer that the success of the minimum ℓ_1-norm solution only depends on the sparsity profile and sign pattern of $\boldsymbol{x}_0$; it is independent of the nonzero magnitudes. Since the minimum ℓ_1-norm solution cannot always recover $\boldsymbol{x}_0$ given only the spark and sparsity level assumptions of the theorem, the restriction on the magnitudes will not help, and so the unique global minimum will not always equal $\boldsymbol{x}_0$.

Now assume that $g(x_i)$ is a strictly concave function of $|x_i|$; the more general case (concave and nonlinear but not necessarily strictly concave) easily follows. If $\lim_{\epsilon \to 0} [g(\epsilon) - g(0)]/\epsilon = \infty$, then based on the proof of Theorem 5, there will exist a local minimum at every basic feasible solution; this result is independent of nonzero coefficient magnitudes. The more ambiguous case is when $\lim_{\epsilon \to 0} [g(\epsilon) - g(0)]/\epsilon = C < \infty$. In this situation, a simple 2D counter example suffices to show that local minima are still always possible. Let

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & \frac{1}{\sqrt{1+\alpha^2}} & \frac{1}{\sqrt{1+\alpha^2}} \\ 0 & \frac{\alpha}{\sqrt{1+\alpha^2}} & -\frac{\alpha}{\sqrt{1+\alpha^2}} \end{bmatrix}$$
$$\boldsymbol{y} = \Phi \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{81}$$

where $\alpha > 0$ is a small constant and the $\sqrt{1+\alpha^2}$ factor is included only for normalization purposes. Here $\boldsymbol{x}_0$ and Φ satisfy the conditions of Theorem 8. Now consider the alternative solution $\boldsymbol{x}' = \left[0 \ \frac{\sqrt{1+\alpha^2}}{2} \ \frac{\sqrt{1+\alpha^2}}{2}\right]^T$. For α sufficiently small, this solution will always be a local minimum for any strictly concave Type I method. To see this, consider the following. The dictionary Φ has a 1D null-space spanned by the vector $\boldsymbol{v} \triangleq \left[1 \ \frac{-\sqrt{1+\alpha^2}}{2} \ \frac{-\sqrt{1+\alpha^2}}{2}\right]^T$ since $\Phi \boldsymbol{v} = 0$, and so any feasible solution can be expressed as $\boldsymbol{x}' + \epsilon \boldsymbol{v}$ for some constant ϵ; to move towards $\boldsymbol{x}_0$ requires $\epsilon > 0$. By taking the gradient of $\sum_i g(x_i)$ with respect to ϵ evaluated at $\epsilon \to 0^+$, we can determine if $\boldsymbol{x}'$ is a local minimum; namely, a local minima occurs if this gradient is positive when the limit is approached from the right. With $g'(x_i) \triangleq dg(x_i)/dx_i$, the relevant limit is

$$\begin{aligned} &\lim_{\epsilon \to 0^+} \frac{d}{d\epsilon} \sum_i g(x_i' + \epsilon v_i) \\ &= \lim_{\epsilon \to 0^+} \frac{d}{d\epsilon}\left(g(\epsilon) + 2g\left[\frac{\sqrt{1+\alpha^2}}{2}(1-\epsilon)\right]\right) \\ &= \lim_{\epsilon \to 0^+}\left(g'(\epsilon) - \sqrt{1+\alpha^2} g'\left[\frac{\sqrt{1+\alpha^2}}{2}(1-\epsilon)\right]\right) \\ &= g'(\epsilon)|_{\epsilon \to 0^+} - \sqrt{1+\alpha^2} g'\left[\frac{\sqrt{1+\alpha^2}}{2}\right]. \end{aligned} \tag{82}$$

By definition of strict concavity, this expression will be positive for some α sufficiently small, implying that $\boldsymbol{x}'$ is a local minimum. ∎

REFERENCES

[1] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag, 1985.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[3] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

[4] A. Bruckstein, M. Elad, and M. Zibulevsky, "A nonnegative and sparse enough solution of an underdetermined linear system of equations is unique," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4813–4820, Nov. 2008.

[5] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[6] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[7] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. Int. Conf. Accoust., Speech, Signal Process.*, 2008.

[8] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientif. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.
[9] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Apr. 2005.
[10] M. Davies and R. Gribonval, "Restricted isometry constants where ℓ_p sparse recovery can fail for $0 < p \leq 1$," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2203–2214, May 2009.
[11] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure and Appl. Math.*, vol. 59, no. 6, pp. 797–829, Jun. 2006.
[12] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[13] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.
[14] C. Févotte and S. Godsill, "Blind separation of sparse sources using Jeffreys inverse prior and the EM algorithm," in *Proc. 6th Int. Conf. Independ. Compon. Anal. Blind Source Separat.*, Mar. 2006.
[15] M. Figueiredo, "Adaptive sparseness using Jeffreys prior," *Adv. Neural Inf. Process. Syst.*, vol. 14, pp. 697–704, 2002.
[16] H. Gao, "Wavelet shrinkage denoising using the nonnegative garrote," *J. Computat. Graph. Statist.*, vol. 7, no. 4, pp. 469–488, 1998.
[17] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computat.*, vol. 13, no. 11, pp. 2517–2532, 2001.
[18] I. Gorodnitsky, J. George, and B. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm," *J. Electroencephalogr. Clin. Neurophysiol.*, vol. 95, no. 4, pp. 231–251, Oct. 1995.
[19] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.
[20] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge , U.K.: Cambridge Univ. Press, 1990.
[21] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.
[22] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T.-Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computat.*, vol. 15, no. 2, pp. 349–396, Feb. 2003.
[23] D. MacKay, "Bayesian interpolation," *Neural Computat.*, vol. 4, no. 3, pp. 415–447, 1992.
[24] D. Malioutov, M. Çetin, and A. Willsky, "Optimal sparse representations in general overcomplete bases," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, vol. 2, pp. II-793–II-796.
[25] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, Apr. 1995.
[26] R. Neal, *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.
[27] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," *Adv. Neural Inf. Process. Syst.*, vol. 18, pp. 1059–1066, 2006.
[28] B. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 760–770, Mar. 2003.
[29] M. Sahani and J. Linden, "Evidence optimization techniques for estimating stimulus-response functions," *Adv. Neural Inf. Process. Syst.*, vol. 15, pp. 301–308, 2003.
[30] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.
[31] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
[32] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2231–2242, Oct. 2004.
[33] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[34] D. Wipf, "Bayesian methods for finding sparse representations," Ph. D., Univ. Calif., San Diego, 2006.
[35] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," *Adv. Neural Inf. Process. Syst.*, vol. 20, 2008.
[36] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *J. Sel. Top. Signal Process. , Special Iss. Compress. Sens.*, vol. 4, no. 2, Apr. 2010.
[37] D. Wipf and S. Nagarajan, "Dual-Space Observations From the Sparse Linear Model," UCSF Tech. Rep., June 2010..
[38] D. Wipf, J. Palmer, B. Rao, and K. Kreutz-Delgado, "Performance analysis of latent variable models with sparse priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007.
[39] D. Wipf and B. Rao, "Comparing the effects of different weight distributions on finding sparse representations," *Adv. Neural Inf. Process. Syst.*, vol. 18, 2006.

David P. Wipf received the B.S. degree in electrical engineering from the University of Virginia, Charlottesville, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, San Diego. He was later an NIH Postdoctoral Fellow in the Biomagnetic Imaging Lab at the University of California, San Francisco where his research involved the development and analysis of Bayesian learning algorithms for functional brain imaging and sparse coding. Currently he is with the Visual Computing Group at Microsoft Research Asia.

Bhaskar D. Rao (F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, in 1979, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively.

Since 1983, he has been with the University of California at San Diego (UCSD), La Jolla, where he is currently a Professor with the Electrical and Computer Engineering Department. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human–computer interactions.

Dr. Rao holds the Ericsson endowed chair in Wireless Access Networks and is the Director of the Center for Wireless Communications. His research group has received several paper awards. His paper received the Best Paper Award at the 2000 Speech Coding Workshop and his students have received student paper awards at both the 2005 and 2006 International Conference on Acoustics, Speech and Signal Processing Conference, as well as the Best Student Paper Award at NIPS 2006. A paper he coauthored with B. Song and R. Cruz received the 2008 Stephen O. Rice Prize Paper Award in the Field of Communications Systems, as well as a paper he coauthored with S. Shivappa and M. Trivedi which received the Best Paper Award at AVSS 2008. He also received the Graduate Teaching Award from the graduate students in the Electrical Engineering Department, UCSD, in 1998. He was elected an IEEE Fellow in 2000 for his contributions in high-resolution spectral estimation. He has been a member of the Statistical Signal and Array Processing Technical Committee, the Signal Processing Theory and Methods Technical Committee, and the Communications Technical Committee of the IEEE Signal Processing Society. He has also served on the editorial board of the EURASIP *Signal Processing Journal*.

Srikantan Nagarajan received the M.S. and Ph.D. degrees in biomedical engineering from Case Western Reserve University.

He completed a postdoctoral fellowship with the Keck Center for Integrative Neuroscience, University of California, San Francisco (UCSF). Currently, he is a Professor in the Departments of Radiology and Biomedical Imaging and Bioengineering and Therapeutic Sciences, UCSF, and a faculty member with the UCSF/UCB Joint Graduate Program in Bioengineering. His research interests, in the areas of neural engineering and machine learning, are to better understand neural mechanisms of sensorimotor learning and speech motor control and to develop algorithms for improved functional brain imaging, biomedical signal processing, and brain computer interfaces.