SPARSE RECOVERY OVER NONLINEAR DICTIONARIES

Luiz F. O. Chamon[*], *Yonina C. Eldar*[†], *and Alejandro Ribeiro*[*]

[*] Electrical and Systems Engineering, University of Pennsylvania
[†] Department of Electrical Engineering, Technion—Israel Institute of Technology
e-mail: luizf@seas.upenn.edu, yonina@ee.technion.ac.il, aribeiro@seas.upenn.edu

ABSTRACT

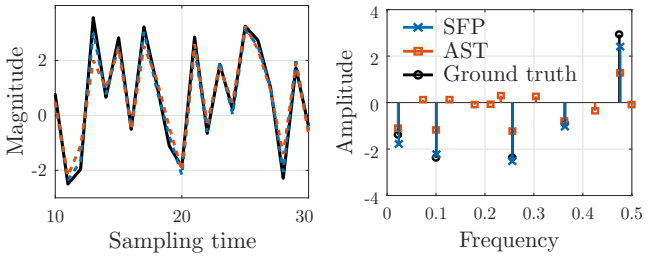
Sparse modeling seeks to represent signals as a linear combination of a small number of atoms from an overparametrized dictionary. Despite the success of these linear models, they can be too restrictive for applications involving nonlinear measurements. Using nonlinear atoms, however, poses an additional obstacle to the sparse recovery problem, since it remains non-convex even after relaxing the sparsity objective (e.g., using atomic norms). We address this issue in the context of continuous dictionaries by posing nonlinear sparse recovery as a sparse functional program that explicitly minimizes the functional equivalent of the "ℓ_0-norm," i.e., the function support measure. By proving that strong duality holds for these optimization problems, we show that nonlinear sparse recovery over continuous dictionaries precludes relaxations since it may be solved efficiently using duality. This result is non-parametric, in that it does not assume the data follows the measurement model, and does not require incoherence assumptions, such as the restricted isometry/eigenvalue property. We also use strong duality to derive a relation between minimizing the support of a function and minimizing its L_1-norm, although this does not imply that the latter leads to sparse solutions. We illustrate this new approach in a nonlinear line spectrum estimation problem.

Index Terms— Sparsity, sparse recovery, nonlinear compressive sensing, functional optimization, strong duality.

1. INTRODUCTION

Sparse modeling plays a fundamental role in contemporary signal processing and has found applications from communications to biology (e.g., [1–3]). Explicitly, it seeks to represent data or measurements as the linear combination of a small number of atoms from an overparametrized dictionary. This dictionary can be either learned from the data or be given *a priori* by the application [4]. Fitting such models is a combinatorial problem known to be NP-hard in general [5]. This issue is typically addressed using a convex relaxation, such as atomic norms (e.g., ℓ_1-norm), or some flavor of greedy algorithm, such as orthogonal matching pursuit (OMP) or iterative hard thresholding (IHT). These approaches have proven effective in practice and theoretical performance guarantees have been derived for dictionaries with incoherence properties (e.g., restricted isometry/eigenvalue property, RIP/REP) [6–8]. Functional versions of this problem for continuous dictionaries have also been studied [9–13].

Despite their success, linear models are often too restrictive for certain practical problems in which atoms and/or measurement models are nonlinear (e.g., magnetic resonance fingerprinting [1], spectrum cartography [14], and manifold data sparse coding [15]). To provide a concrete example, consider the problem of estimating the frequencies and amplitudes of a small number of sinusoidal sources whose signals saturate due to hardware limitations. Notice that the signal saturates at the sources, so that it is not possible to determine where signals saturated from their superposition (see solid line in Fig. 1a). In this case, classical linear methods (e.g., MUSIC [16]) and atomic norm relaxations (e.g., atomic soft thresholding, AST [9–11]) still provide good estimates of the components frequencies, but severely underestimate their true amplitudes (see Fig. 1b). Though sparse recovery can sometimes be solved using "linear in the parameters" models, such as splines or kernel methods (e.g., for spectrum cartography [14]), this approach is not applicable in general. Indeed, the number of kernels needed to represent a generic nonlinear atom may be so large that the dictionary model is no longer sparse. However, directly accounting for nonlinearities in sparse coding can be difficult since the problem remains non-convex even after relaxing the sparsity objective. This is evidences by the weaker guarantees existing for ℓ_1-norm relaxations in the nonlinear case [17, 18].



(a) Reconstructed signal (b) Estimated parameters

Fig. 1. Nonlinear line spectral estimation.

In this paper, we show that when posed in functional terms, sparse recovery turns out to be tractable over (nonlinear) continuous dictionaries (Theorem 1). This result extends the one from [13] to a wide class of nonlinear measurement models. In other words, nonlinear sparse models can be estimated exactly and efficiently in a myriad continuous applications, including nonlinear spectral estimation, spectrum cartography, and super-resolution imaging. This approach forgoes the use of discretizations and convex relaxations by relying instead on duality and therefore bypasses issues of grid mismatch, dictionary coherence, and ill-conditioning [19–23].

We study this approach by posing nonlinear sparse recovery over continuous dictionaries as a functional optimization problem (Section 2). We then derive its Lagrangian dual (Section 2.2) and prove it has null duality gap (Section 3) despite the non-convexity (sparsity) and nonlinearity (dictionary atoms) of the original problem. By exploiting separability, we show that nonlinear sparse recovery can be solved exactly and efficiently over continuous dictionaries. Moreover, we use this strong duality result to relate sparse functional programs and L_1-norm minimization by showing that their optimal values are (essentially) the same, though L_1-norm optimization admits solutions that are not sparse (Section 3.1). Finally, we illustrate these results in a nonlinear spectral estimation application (Section 4).

2. NONLINEAR SPARSE RECOVERY AND SFPs

2.1. Problem formulation

Nonlinear sparse recovery (sparse coding) seeks to represent a signal or data point $\boldsymbol{y} \in \mathbb{R}^p$ using as few atoms as possible from a nonlinear dictionary $\mathcal{D} = \{\boldsymbol{F}(\cdot, \beta) : \mathbb{R} \to \mathbb{R}^p \mid \beta \in \Omega\}$, where Ω is a compact set of the real line. Explicitly, we wish to find

$$\hat{\boldsymbol{y}} = \sum_{i=1}^{k} \boldsymbol{F}(x_i, \beta_i) \tag{1}$$

close to $\boldsymbol{y}$ for some small k. In this work, we take Ω to be uncountable so that we choose from a continuum of atoms as opposed to the discrete, finite case. Furthermore, we assume that the elements of the vector-valued functions $\boldsymbol{F}$ are *normal integrands* with $\boldsymbol{F}(0, \cdot) \equiv \boldsymbol{0}$. A function $f(x, \beta)$ is a normal integrand if it continuous in x for all fixed β and measurable in β for all fixed x [24]. Note that $\boldsymbol{F}$ need not be linear or convex. For instance, the nonlinear line spectrum estimation problem can be posed by taking

$$\boldsymbol{F}(x, \beta) = \rho\left[\cos(\pi \boldsymbol{t} \beta) x\right], \tag{2}$$

where $\boldsymbol{t} \in \mathbb{R}^p$ collects the sampling times and $\rho(x) = x$ for $|x| < 1$ and $\rho(x) = 1$ otherwise, applies element-wise to vectors and represents the signal saturation. In this example, $\beta \in [0, 1]$ and x represent the frequency and amplitude of each sinusoidal component respectively. The linear case is recovered by taking $\boldsymbol{F}(x, \beta) = \boldsymbol{h}(\beta) x$ for some vector-valued function $\boldsymbol{h}$ [12, 13, 20, 21].

We propose to determine the (x_i, β_i) (or more precisely $\hat{\boldsymbol{y}}$) by solving a sparse functional program (SFP). SFPs are variational problems that seek sparsest functions, i.e., functions with minimum support measure. Formally, define the L_0-norm[1] that, similar to the discrete case, measures the support of a function, i.e.,

$$\|X\|_{L_0} = \int_{\Omega} \mathbb{I}\left[X(\beta) \neq 0\right] d\beta, \tag{3}$$

where the indicator function $\mathbb{I}$ is defined as $\mathbb{I}(\beta \in \mathcal{E}) = 1$, if $\beta \in \mathcal{E}$, and zero otherwise. Unless otherwise specified, all integrals are taken with respect to the Lebesgue measure over the measurable space $(\Omega, \mathcal{B})$, where $\mathcal{B}$ are the Borel sets of Ω. SFPs explicitly minimize the L_0-norm in (3). The relation to nonlinear sparse recovery follows from the following observation:

Proposition 1. *Let* $X_B(\beta) = \sum_{i=1}^{k} x_i \mathbb{I}\left[\beta \in \mathcal{B}_i\right]$ *with* $\mathcal{B}_i = [\beta_i - B^{-1}, \beta_i + B^{-1}]$. *Then, as* $B \to \infty$,

$$\|X_B\|_{L_0} \to 0 \quad \textit{and} \quad \int_{\Omega} B\boldsymbol{F}\left(X_B(\beta), \beta\right) d\beta \to \sum_{i=1}^{k} \boldsymbol{F}(x_i, \beta_i).$$

Proof. The result follows by noting that $\int_{\Omega} B\boldsymbol{F}\left(X_B(\beta), \beta\right) d\beta = \int_{\Omega} B\,\mathbb{I}\left[\beta \in \mathcal{B}_i\right] \boldsymbol{F}\left(x_i, \beta\right) d\beta$ and from the fact that $B\,\mathbb{I}\left[\beta \in \mathcal{B}_i\right]$ converges weakly to $\delta(\beta - \beta_i)$ as $B \to \infty$, where δ is the Dirac's delta [25]. ∎

Nonlinear sparse recovery can then be posed as the SFP

$$\begin{aligned} \underset{X \in \mathcal{X}}{\text{minimize}} \quad & \lambda \|X\|_{L_0} + \int_{\Omega} F_0\left(X(\beta), \beta\right) d\beta \\ \text{subject to} \quad & \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 \leq \epsilon \\ & \hat{\boldsymbol{y}} = \int_{\Omega} B\boldsymbol{F}\left(X(\beta), \beta\right) d\beta \end{aligned} \tag{PI}$$

[1] As in the discrete case, the "L_0-norm" is not a norm. We however omit the quotation marks so as not to burden the text.

where $\lambda > 0$ and $B > 0$ are a parameters that controls the sparsity and approximation error of the solution, $F_0 : \mathbb{R} \times \Omega \to \mathbb{R}$ is an optional regularization term with $F_0(0, \cdot) \equiv 0$ (e.g., take $F_0(x, \beta) = x^2$ for shrinkage), and $\mathcal{X}$ is a *composable* function space, i.e., if $X, X' \in \mathcal{X}$, then for any $\mathcal{Z} \in \mathcal{B}$ it holds that $\bar{X} \in \mathcal{X}$ for

$$\bar{X}(\beta) = \begin{cases} X(\beta), & \beta \in \mathcal{Z} \\ X'(\beta), & \beta \notin \mathcal{Z} \end{cases}.$$

In the sequel, we will take $\mathcal{X} = L_2$, although all Lebesgue spaces or function spaces with pointwise constraints (e.g., $\mathcal{X} = \{X \in \mathcal{B} \mid X \leq \Gamma \text{ a.e.}\}$) are also composable. Since B is finite and $\mathcal{X}$ is a functional space, solutions $X^\star$ of (PI) cannot contain point masses. Instead, they will be a combination of bump functions centered around β_i. From Proposition 1, we can therefore obtain the parameters of (1) by taking β_i to be the centers of the bump functions and $x_i = B \int_{\Omega} X^\star(\beta) d\beta$.

Nevertheless, solving (PI) is challenging since it is both infinite dimensional and non-convex. Moreover, discretizing (PI) can lead to NP-hard problems [5] and even if the L_0-norm was relaxed to the L_1-norm, as in the discrete case, (PI) would remain non-convex due to the nonlinear equality. In this work, we propose to solve (PI) using duality. Though this approach is often used to solve semi-infinite convex programs [9–11, 26], SFPs are not convex optimization problems. To address this issue, we first derive the dual problem of (PI) in the next section. Since it is both finite dimensional and convex, the dual of (PI) can be solved using convex optimization methods such as (stochastic) (sub)gradient ascent. We then show that we can obtain a solution of (PI) from a solution of its dual by proving that it has null duality gap (Section 3).

2.2. The Lagrangian dual of (PI)

To formulate the dual problem of (PI), introduce the Lagrange multipliers $\boldsymbol{\mu} \in \mathbb{R}^p$, corresponding to the equality constraint, and the nonnegative $\nu \in \mathbb{R}_+$, corresponding to the inequality constraint. Then, the Lagrangian dual of (PI) is defined as

$$\begin{aligned} \mathcal{L}(X, \hat{\boldsymbol{y}}, \boldsymbol{\mu}, \nu) = \lambda \|X\|_{L_0} &+ \int_{\Omega} F_0\left(X(\beta), \beta\right) d\beta \\ &+ \nu\left(\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 - \epsilon\right) \\ &+ \boldsymbol{\mu}^T \left(\int_{\Omega} \boldsymbol{F}\left(X(\beta), \beta\right) d\beta - \hat{\boldsymbol{y}}\right). \end{aligned} \tag{4}$$

Its dual function is therefore

$$d(\boldsymbol{\mu}, \nu) = \min_{X \in \mathcal{X}, \hat{\boldsymbol{y}}} \mathcal{L}(X, \hat{\boldsymbol{y}}, \boldsymbol{\mu}, \nu), \tag{5}$$

so that the dual problem of (PI) is given by

$$\underset{\boldsymbol{\mu},\, \nu \geq 0}{\text{maximize}} \quad d(\boldsymbol{\mu}, \nu). \tag{DI}$$

By definition, (DI) is a convex program whose dimensionality is equal to the number of constraints [27]—in this case, on the order of the number of measurements p. It is therefore tractable as long as we can evaluate the dual function d. Indeed, solving (DI) is necessarily at least as hard as solving the minimization in (5). The dual function of SFPs, however, can be computed efficiently. To see why this is the case, note that the joint minimization in (5) separates into $d(\boldsymbol{\mu}, \nu) = d_X(\boldsymbol{\mu}) + d_{\hat{\boldsymbol{y}}}(\boldsymbol{\mu}, \nu) - \nu\epsilon$ with

$$\begin{aligned} d_X(\boldsymbol{\mu}) = \min_{X \in \mathcal{X}} \int_{\Omega} \Big[& F_0\left(X(\beta), \beta\right) + \lambda\, \mathbb{I}\left[X(\beta) \neq 0\right] \\ & + \boldsymbol{\mu}^T \boldsymbol{F}\left(X(\beta), \beta\right) \Big] d\beta \end{aligned} \tag{6}$$

and $d_{\hat{\boldsymbol{y}}}(\boldsymbol{\mu},\nu) = \min_{\hat{\boldsymbol{y}}} \nu \|\boldsymbol{y}-\hat{\boldsymbol{y}}\|_2^2 - \boldsymbol{\mu}^T\hat{\boldsymbol{y}} = -\|\boldsymbol{\mu}\|^2/(4\nu) - \boldsymbol{\mu}^T y$. The quadratic program $d_{\hat{\boldsymbol{y}}}$ can be evaluated in closed-form, whereas d_X is again infinite dimensional and non-convex. Yet, separability of the objective across β can be exploited to reduce this problem to a scalar optimization problem followed by thresholding.

Proposition 2. *Consider the functional optimization problem in* (6) *and define*

$$\gamma^o(\boldsymbol{\mu},\beta) = \min_{x\in\mathbb{R}} F_0(x,\beta) + \boldsymbol{\mu}^T\boldsymbol{F}(x,\beta). \tag{7}$$

Then, $d_X(\boldsymbol{\mu}) = \int_{\mathcal{S}} [\lambda + \gamma^o(\boldsymbol{\mu},\beta)]\, d\beta$ *for* $\mathcal{S} = \{\beta \in \Omega \mid \gamma^o(\boldsymbol{\mu},\beta) < -\lambda\}$.

Proof. We start by separating the objective of (6) using the following lemma:

Lemma 1. *Let* $G(x,\beta)$ *be a normal integrand. Then,*

$$\inf_{X\in\mathcal{X}} \int_\Omega G\left(X(\beta),\beta\right) d\beta = \int_\Omega \inf_{x\in\mathbb{C}} G(x,\beta) d\beta. \tag{8}$$

Proof. See [24, Thm. 3A]. ∎

We can therefore restrict ourselves to solving individually for each β the problem $G(\beta) \triangleq \min_{x\in\mathbb{R}} F_0[x,\beta] + \lambda\,\mathbb{I}(x\neq 0) + \boldsymbol{\mu}^T\boldsymbol{F}(x,\beta)$. Despite the non-convexity of the indicator function, this is a scalar minimization whose solution involves a simple thresholding scheme. Indeed, only two conditions need to be checked: (i) if $x = 0$, then $G(\beta)$ vanishes; (ii) if $x \neq 0$, the indicator function is one and $G(\beta) = \lambda + \gamma^o(\boldsymbol{\mu},\beta)$. The value of $G(\beta)$ is the minimum of these two cases. Using (8) then yields the desired result. ∎

Proposition 2 provides a practical way to evaluate (5), although it relies on being able to solve (7). Regardless of whether F_0 and/or $\boldsymbol{F}$ are non-convex functions, (7) is a scalar problem that can typically be solved efficiently using global optimization techniques [28] or through local search procedures, as in the nonlinear line spectrum estimation application in Section 4. Thus, the dual function can be evaluated as

$$d(\boldsymbol{\mu},\nu) = \int_{\mathcal{S}} [\lambda + \gamma^o(\boldsymbol{\mu},\beta)]\, d\beta - \frac{\|\boldsymbol{\mu}\|^2}{4\nu} - \boldsymbol{\mu}^T y - \nu\epsilon \tag{9}$$

for γ^o as in (7) and $\mathcal{S} = \{\beta\in\Omega \mid \gamma^o(\boldsymbol{\mu},\beta) < -\lambda\}$, and (DI) can be solved using any convex optimization algorithm [27].

Having established that we can solve the dual problem (DI), all that remains is showing how it may be used to obtain a solution for (PI). Since SFPs are not convex programs, there is no reason to expect that the optimal value of (DI) is anything more than a lower bound on the optimal value of (PI) [27]. In the sequel, we show that this is not the case by proving that (PI) has zero duality gap.

3. STRONG DUALITY OF SFPs

The main result of this section is stated below.

Theorem 1. *Suppose that* F_0 *and* $\boldsymbol{F}$ *have no point masses (Dirac deltas) and that Slater's condition holds for* (PI)*. Then, strong duality holds for* (PI)*, i.e.,* $P = D$ *for* P*, the optimal value of* (PI)*, and* D*, the optimal value of* (DI)*.*

Proof. See [29]. ∎

Theorem 1 states that although (PI) is a non-convex functional program, it has null duality gap. Because the dual problem is always convex, (PI) can be solved exactly and efficiently using convex programming. Indeed, if $\boldsymbol{\mu}^\star, \nu^\star$ are minimizers of (DI) and $X^\star, \hat{\boldsymbol{y}}^\star$ are solutions of (PI), it holds that

$$(X^\star, \hat{\boldsymbol{y}}^\star) \in \operatorname*{argmin}_{X\in\mathcal{X},\hat{\boldsymbol{y}}} \mathcal{L}(X,\hat{\boldsymbol{y}},\boldsymbol{\mu}^\star,\nu^\star). \tag{10}$$

If $\mathcal{L}$ is strongly convex (e.g., when regularizing with shrinkage), the set on the right-hand side of (10) is a singleton and the set membership becomes equality. It is worth noting that Theorem 1 is a *non-parametric* result in the sense that it makes no assumption on the existence or validity of the dictionary model (1). More to the point, it does not require the signals or data to arise from the dictionary $\mathcal{D}$ or even to be composed of few atoms. This implies, for instance, that the most parsimonious description of a signal in some dictionary can be determined regardless of whether such signal follows (1) or the dictionary actually describes the signal. In practice, this is of utmost importance given that there are arguments for obtaining sparse descriptions that are not epistemological, such as reducing computational costs, and that it is often unrealistic to assume that the dictionary was used to generate the data, especially when it is learned from samples.

A corollary of Theorem 1 is that SFPs are closely related to L_1-norm minimization problems. In a sense, it turns out these two problems are equivalent despite the fact that L_1-norm optimization does not always yield sparse solutions. We explore these conclusions in the sequel.

3.1. Relation between L_0- and L_1-norm optimization

Similar to the discrete case, there is a close relation between L_0- and L_1-norm minimization. Formally, consider

$$\begin{aligned} \underset{|X|\leq\Gamma \text{ a.e.}}{\text{minimize}} \quad & \|X\|_{L_q} \\ \text{subject to} \quad & \|\boldsymbol{y}-\hat{\boldsymbol{y}}\| \leq \epsilon \\ & \hat{\boldsymbol{y}} = \int_\Omega \boldsymbol{F}\left(X(\beta),\beta\right) d\beta \end{aligned} \tag{P_q}$$

Problem (P_0) [i.e., (P_q) with $q=0$] is an instance of (PI) without regularization ($F_0 \equiv 0$) in which $\mathcal{X}$ is the set of measurable functions bounded by $\Gamma > 0$. On the other hand, (P_1) [(P_q) for $q=1$] is a functional version of the classical ℓ_1-norm minimization problem. The following proposition shows that for a wide class of dictionaries, the optimal values of (P_0) and (P_1) are the same (up to a constant).

Proposition 3. *Let* $x^o(\boldsymbol{\mu},\beta) = \operatorname{argmin}_{|x|\leq\Gamma} |x| - \boldsymbol{\mu}^T\boldsymbol{F}(x,\beta)$ *saturate, i.e.,* $x^o(\boldsymbol{\mu},\beta) \neq 0 \Rightarrow |x^o(\boldsymbol{\mu},\beta)| = \Gamma$ *for all* $\boldsymbol{\mu}\in\mathbb{R}^p$ *and* $\beta\in\Omega$*. If* P_0 *(*P_1*) is the optimal value of* (P_q) *when* $q=0$ *(*$q=1$*) and Slater's condition holds, then* $P_0 = P_1/\Gamma$*.*

Proof. The proof follows by relating the dual values of (P_q) for $q = 0, 1$. First, define its Lagrangian as

$$\begin{aligned} \mathcal{L}(X,\boldsymbol{z},\boldsymbol{\mu},\nu) = \|X\|_{L_q} + \nu\left(\|\boldsymbol{y}-\hat{\boldsymbol{y}}\| - \epsilon\right) \\ + \boldsymbol{\mu}^T\left(\hat{\boldsymbol{y}} - \int_\Omega \boldsymbol{h}(\beta)X(\beta)d\beta\right). \end{aligned} \tag{11}$$

Then, for $q=0$, we can leverage Lemma 1 to obtain the dual function

$$d_0(\boldsymbol{\mu},\nu) = \int_\Omega I(\boldsymbol{\mu},\beta)d\beta - w(\mu,\nu), \tag{12}$$

where $I(\boldsymbol{\mu},\beta) = \min_{|x|\leq\Gamma} \mathbb{I}(x \neq 0) - \boldsymbol{\mu}^T \boldsymbol{F}(x,\beta)$ and $w(\mu,\nu) = -\|\boldsymbol{\mu}\|^2/(4\nu) - \boldsymbol{\mu}^T y - \nu\epsilon$. Notice that w is homogeneous as in $w(\alpha\boldsymbol{\mu},\alpha\nu) = \alpha w(\boldsymbol{\mu},\nu)$ for $\alpha > 0$. Since the integrand in (12) is non-zero only over the set $\mathcal{S}_0(\boldsymbol{\mu}) = \{\beta \in \Omega \mid \max_{|x|\leq\Gamma} \boldsymbol{\mu}^T \boldsymbol{F}(x,\beta) > 1\}$, we can rewrite (12) as

$$d_0(\boldsymbol{\mu},\nu) = \int_{\mathcal{S}_0(\boldsymbol{\mu})} \left[1 - \max_{|x|\leq\Gamma} \boldsymbol{\mu}^T \boldsymbol{F}(x,\beta)\right] d\beta - w(\mu,\nu), \tag{13}$$

Proceeding similarly, the dual function of (P_1) reads

$$d_1(\boldsymbol{\mu},\nu) = \int_{\mathcal{S}_1(\boldsymbol{\mu})} \left[\Gamma - \max_{|x|\leq\Gamma} \boldsymbol{\mu}^T \boldsymbol{F}(x,\beta)\right] d\beta - w(\mu,\nu), \tag{14}$$

where we used the saturation hypothesis to obtain that the integrand is non-trivial only on $\mathcal{S}_1(\boldsymbol{\mu}) = \{\beta \in \Omega \mid \max_{|x|\leq\Gamma} \boldsymbol{\mu}^T \boldsymbol{F}(x,\beta) > \Gamma\}$.

To conclude, observe from (13) and (14) that $d_0(\boldsymbol{\mu},\nu) = d_1(\Gamma\boldsymbol{\mu},\Gamma\nu)/\Gamma$ by recalling that w is homogeneous and $\mathcal{S}_1(\Gamma\boldsymbol{\mu}) = \mathcal{S}_0(\boldsymbol{\mu})$. Immediately, it therefore holds that if $(\boldsymbol{\mu}^o,\nu^o)$ is a maximum of d_0, then $(\Gamma\boldsymbol{\mu}^o,\Gamma\nu^o)$ is a maximum of d_1: suffices it to note that $\nabla d_0(\boldsymbol{\mu}^o,\nu^o) = \mathbf{0} \Leftrightarrow \nabla d_1(\Gamma\boldsymbol{\mu}^o,\Gamma\nu^o) = \mathbf{0}$. Since d_1 is a concave function, $(\Gamma\boldsymbol{\mu}^o,\Gamma\nu^o)$ is a global maximum. Given that (P_q) has zero duality gap for $q = 0$ (Theorem 1) and $q = 1$ (convex program), it holds that $P_0 = \max_{\boldsymbol{\mu},\nu\geq 0} d_0(\boldsymbol{\mu},\nu) = d_0(\boldsymbol{\mu}^\star,\nu^\star) = d_1(\Gamma\boldsymbol{\mu}^\star,\Gamma\nu^\star)/\Gamma = \max_{\boldsymbol{\mu},\nu\geq 0} d_1(\boldsymbol{\mu},\nu)/\Gamma = P_1/\Gamma$. ■

Proposition 3 shows that the L_0- and L_1-norm minimization problems found in nonlinear sparse recovery are equivalent in the sense that their optimal values are (essentially) the same. It is worth noting that establishing this relation requires virtually no assumptions: the saturation hypothesis is met by a wide class of dictionaries, most notably linear ones. This is in contrast to the discrete case, where such relations exist only for incoherent, linear dictionaries. Nevertheless, Proposition 3 does not imply that the solution of the L_0- and L_1-norm problems are the same. In fact, although they have the same optimal value, (P_1) admits solutions with larger support (see Remark 1). Although conditions exist for which the L_1-norm minimization problem with linear dictionaries yields minimum support solutions [12, 20, 21], Theorem 1 precludes the use of this relaxation for continuous dictionaries, both linear and nonlinear.

Remark 1. Proposition 3 gives an equivalence between L_0- and L_1-norm minimization problems in terms of their optimal values, but not their solutions. We illustrate this point with the following example: let $\epsilon = 0$, $\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix}^T$ with $|y_1|, |y_2| < \Gamma/2$, and $\boldsymbol{F}(x,\beta) = \boldsymbol{h}(\beta)x$, where $\boldsymbol{h}(\beta) = \begin{bmatrix} h'(\beta) & 1 - h'(\beta) \end{bmatrix}^T$ with $h'(\beta) = \mathbb{I}(\beta \in [0, 1/2])$. It is ready that the optimal value of (P_1) is $P_1 = |y_1| + |y_2|$.

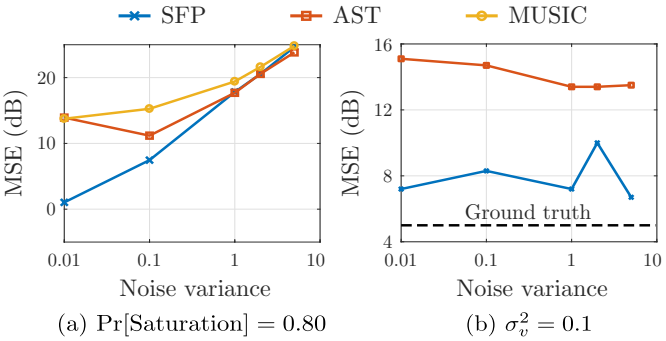
Now consider the family of functions indexed by $0 < a \leq \Gamma$, $X_a(\beta) = a\,\mathrm{sign}(y_1)\,\mathbb{I}(\beta \in \mathcal{A}_1) + a\,\mathrm{sign}(y_2)\,\mathbb{I}(\beta \in \mathcal{A}_2)$, where $\mathcal{A}_1 \subseteq [0, 1/2]$ with $|\mathcal{A}_1| = |y_1|/a$ and $\mathcal{A}_2 \subseteq [1/2, 1]$ with $|\mathcal{A}_2| = |y_2|/a$. For all a, X_a is a solution of (P_1) (it is (P_q)-feasible with value P_1) with support $\|X_a\|_{L_0} = (|y_1| + |y_2|)/a$. Thus, ($\mathrm{P}_1$) admits solutions that do not have minimum support, whereas only X_Γ is a solution of (P_0).

4. NONLINEAR LINE SPECTRAL ESTIMATION

In this section, we illustrate the use of (PI) to estimate the superposition of a small number of clipped (saturated) sinusoids. To do so, we use a dictionary composed of the atoms in (2) and add a shrinkage term to the objective by taking $\mathcal{X} = L_2$ and $F_0(x,\beta) = x^2$ for all $\beta \in \Omega$. We obtain a solution of (PI) by solving its dual problem (DI) and applying (10). Using Proposition 2, we can obtain the dual function d in the objective of (DI) by solving (7). Explicitly, we must evaluate $\gamma^o(\boldsymbol{\mu},\beta) = \min_{x\in\mathbb{R}} x^2 + \boldsymbol{\mu}^T \rho\,[\boldsymbol{h}(\beta)x]$ with $\boldsymbol{h}(\beta) = \cos(\pi \boldsymbol{t}\beta)$. Solving this non-convex problem actually reduces to finding the minimum of the values of p quadratic problems. Namely, assume that $\boldsymbol{h}$ is sorted such that $h_1 \leq \cdots \leq h_p$ and define $\boldsymbol{w}_i(x) = [h_1 x \;\cdots\; h_i x \; 1 \;\cdots\; 1]^T$. For conciseness, we omit the dependence on β. Then, $\gamma^o(\boldsymbol{\mu},\beta) = \min_{1\leq i\leq p} \gamma_i^o(\boldsymbol{\mu},\beta)$ for

$$\gamma_i^o = \min_{1/|h_{i+1}|\leq|x|\leq 1/|h_i|} x^2 + \boldsymbol{\mu}^T \boldsymbol{w}_i(x), \quad i = 1,\ldots,p-1,$$

$$\gamma_p^o = \min_{|x|\leq 1/|h_p|} x^2 + \boldsymbol{\mu}^T \boldsymbol{h} x.$$



(a) $\Pr[\text{Saturation}] = 0.80$ (b) $\sigma_v^2 = 0.1$

Fig. 2. Reconstruction MSE and average estimated support for nonlinear line spectral estimation.

We compared the SFP approach to two linear approaches: MUSIC [16], an eigendecomposition-based method, and AST [9–11], an L_1-norm convex relaxation of (PI). The resulting reconstruction mean-square errors (MSEs) with the k most significant components are shown in Fig. 2a and the average estimated support size (across 10 realizations) are shown in Fig. 2b. In these experiments, the measurement are given by $\boldsymbol{y} = \sum_{j=1}^5 \rho\,[x_j \cos(\pi \boldsymbol{t}\beta_j)] + \boldsymbol{v}$, where $\boldsymbol{t} = [t_i]$ with t_i integers in $[-30, 30]$ ($p = 61$), the frequencies β_j are drawn randomly from $[0, 0.5]$ with a minimum spacing of $4/p$ thus guaranteeing that AST can discriminate the components [11], and $\boldsymbol{v}$ is a vector collecting independent zero-mean Gaussian random variables with variance σ_v^2. The amplitudes A_j were drawn uniformly at random between 0.5 and 3, so that the probability of saturation is 80%. For MUSIC, we use the actual number of spectral lines $k = 5$. For AST, we the optimal regularizer from [10] which depends on σ_v^2. Both methods estimate the frequencies β_j and determine the amplitudes x_j using least squares. For (PI), we use $\epsilon = p\sigma_v^2$ and $\lambda = 100$ for all noise levels except $\sigma_v^2 = (2, 5)$ for which we used $\lambda = 80$. We then computed the reconstruction MSE by evaluating $\hat{\boldsymbol{y}}$ as in (PI).

5. CONCLUSION

We tackled nonlinear sparse recovery over continuous dictionaries by formulating this problem as an SFP. We then showed that these optimization problems have no duality gap and can therefore be solved efficiently using duality. This approach bypasses issues of grid mismatch and dictionary coherence found in discrete versions of sparse recovery by forgoing the use of convex relaxations. Moreover, we showed that, as in the discrete case, there is a close relation between L_0- and L_1-norm optimization, even though the latter need not yield sparse solutions. We illustrated this method by estimating a superposition of sinusoids from saturated signals, but foresee that this technique can be applied to a wide variety of problems.

6. REFERENCES

[1] D. Ma, V. Gulani, N. Seiberlich, K. Liu, J.L. Sunshine, J.L. Duerk, and M.A. Griswold, "Magnetic resonance fingerprinting," *Nature*, vol. 495[7440], pp. 187–192, 2013.

[2] L. Zhu, W. Zhang, D. Elnatan, and B. Huang, "Faster storm using compressed sensing," *Nature methods*, vol. 9, pp. 721–726, 2012.

[3] M. Mishali, Y. C. Eldar, and A. J. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Trans. Signal Process.*, vol. 59[10], pp. 4719–4734, 2011.

[4] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. of the IEEE*, vol. 98[6], pp. 1045–1057, 2010.

[5] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24[2], pp. 227–234, 1995.

[6] Y. C. Eldar and G. Kutyniok, Eds., *Compressed Sensing: Theory and Applications*, Cambridge, 2012.

[7] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Birhaüser, 2013.

[8] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*, Cambridge, 2015.

[9] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. Inf. Theory*, vol. 59[11], pp. 7465–7490, 2013.

[10] B.N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Trans. Signal Process.*, vol. 61[23], pp. 5987–5999, 2013.

[11] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67[6], pp. 906–956, 2014.

[12] G. Puy, M. E. Davies, and R. Gribonval, "Recipes for stable linear embeddings from hilbert spaces to $\mathbb{R}^m$," *IEEE Trans. Inf. Theory*, vol. 63[4], pp. 2171–2187, 2017.

[13] L.F.O. Chamon, Y. C. Eldar, and A. Ribeiro, "Strong duality of sparse functional optimization," in *ICASSP*, 2017, pp. 4739–4743.

[14] J.A. Bazerque, G. Mateos, and G.B. Giannakis, "Group-lasso on splines for spectrum cartography," *IEEE Trans. Signal Process.*, vol. 59[10], pp. 4648–4663, 2011.

[15] Y. Xie, J. Ho, and B. Vemuri, "On a nonlinear generalization of sparse coding and dictionary learning," in *ICML*, 2013, pp. III–1480–III–1488.

[16] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*, Prentice-Hall, 2005.

[17] A. Beck and Y.C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," *SIAM Journal on Optimization*, vol. 23[3], pp. 1480–1509, 2013.

[18] Z. Yang, Z. Wang, H. Liu, Y.C. Eldar, and T. Zhang, "Sparse nonlinear regression: Parameter estimation under nonconvexity," in *ICML*, 2016, pp. 2472–2481.

[19] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Process.*, vol. 59[5], pp. 2182–2195, 2011.

[20] B. Adcock and A. C. Hansen, "Generalized sampling and infinite-dimensional compressed sensing," *Foundations of Computational Mathematics*, vol. 16[5], pp. 1263–1323, 2016.

[21] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: A new theory for compressed sensing," *Forum of Mathematics, Sigma*, vol. 5, 2017.

[22] V. Duval and G. Peyré, "Sparse regularization on thin grids I: the Lasso," *Inverse Problems*, vol. 33[5], pp. 055008, 2017.

[23] V. Duval and G. Peyré, "Sparse spikes super-resolution on thin grids II: the continuous basis pursuit," *Inverse Problems*, vol. 33[9], pp. 095008, 2017.

[24] R. T. Rockafellar, *Integral functionals, normal integrands and measurable selections*, Springer, 1976.

[25] W. Rudin, *Functional Analysis*, McGraw-Hill, 1991.

[26] A. Shapiro, "On duality theory of convex semi-infinite programming," *Optimization*, vol. 54[6], pp. 535–543, 2006.

[27] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[28] E.M.T. Hendrix and B. G.-Tóth, *Introduction to Nonlinear and Global Optimization*, Springer, 2010.

[29] L.F.O. Chamon, Y. C. Eldar, and A. Ribeiro, "Functional optimization for nonlinear sparse problems," 2018, `https://arxiv.org/abs/1811.00577`.