STRONG DUALITY OF SPARSE FUNCTIONAL OPTIMIZATION

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

Signal processing is rich in inherently continuous applications, such as radar, MRI, and source localization, in which sparsity priors play a key role in obtaining state-of-the-art results. To cope with the infinite dimensionality and non-convexity of these estimation problems, they are typically discretized and solved by means of convex relaxations, e.g., using atomic norms. Although successful, this approach is not without issues. Discretization often leads to high dimensional, potentially ill-conditioned optimization problems. Moreover, due to grid mismatch and other coherence issues, a sparse signal in the continuous domain may no longer be sparse when discretized. Finally, performance guarantees for atomic norm relaxations hold under assumptions that may be hard to meet in practice. We address these issues by directly tackling the continuous problem cast as a sparse functional optimization program. We prove that these problems have no duality gap and show that they can be solved efficiently using duality and a stochastic gradient ascent-type algorithm. We illustrate the performance of this new approach on a line spectral estimation problem.

Index Terms— Functional optimization, sparsity, strong duality, line spectral estimation.

1. INTRODUCTION

Signal processing is rich in inherently continuous[1] problems, such as spectral or delay estimation, image recovery, source localization, radar, and array processing [1–3]. These problems are difficult to tackle directly due to their infinite dimensionality. We therefore rely on sampling results to show that in some cases operating on the original signal or a finite (or countably infinite) representation is (approximately) equivalent. For instance, we can filter bandlimited functions and process finite rate of innovation or union of subspace signals using only a discrete set of samples [4–6]. Alternatively, if the functions lie in a reproducible kernel Hilbert space, we can leverage the representer theorem to perform computations using finite descriptions, a technique sometimes known as the "kernel trick" [7].

Still, discrete problems are not necessarily easier problems. In particular, although discretization leads to finite problems, these problems often remain underdetermined: for a fine discretization, the number of parameters to estimate exceeds the number of measurements. Sparsity priors then play an important role in achieving state-of-the-art results [6, 8, 9]. However, albeit finite dimensional, sparse optimization problems are typically non-convex and in certain cases, NP-hard [10]. As before, this issue is typically addressed by solving a tractable nearby problem: the sparsity prior is relaxed to some atomic norm constraint (e.g., ℓ_1 norm) so that the modified problem is convex. Under certain measurement models and incoherence conditions, sparse and relaxed problems yield the same solution [8, 9].

Although the discretization/relaxation approach can be effective, it is not always the case. Sampling theorems are sensitive to the function class considered and are often asymptotic: results improve as the discretization becomes finer. This leads to high dimensional statistical problem with potentially poor numerical properties (high condition number). Moreover, discretization can lead to grid mismatch issues and even loss of sparsity: signals that are sparse in the continuous domain need not be sparse when discretized. These issues were recently addressed by the development of a continuous theory for compressive sensing [11–14]. Moreover, for the particular case of spectral estimation, it is known that the relaxed problem can be posed directly without discretizing [15–17]. Still, performance guarantees for these convex relaxations rely on assumptions that are sometimes difficult to meet in practice.

We therefore propose to forgo both discretization and relaxation and directly tackle the sparse functional program. Though we now combine the infinite dimensionality of functional programming with the non-convexity of sparsity, this turns out to be a fruitful approach. In fact, we show that sparse functional optimization problems can be solved exactly by leveraging duality. To do so, we first formulate a general sparse functional optimization problem (Sec. 2). Then, we prove that strong duality holds for these problems under mild conditions (Sec. 3). In other words, if we can solve their dual problems, we can solve sparse functional problems. Finally, by observing that the dual problem is a finite convex program, we propose an algorithm based on stochastic gradient ascent to solve sparse functional optimization problems without explicitly evaluating integrals (Sec. 4) and illustrate this method in a spectral estimation application (Sec. 5).

Notation: We use lowercase boldface letters for vectors ($\boldsymbol{x}$), uppercase boldface letters for matrices ($\boldsymbol{X}$), calligraphic letters for sets ($\mathcal{A}$), and fraktur font for measures ($\mathfrak{h}$). In particular, we denote the Lebesgue measure by $\mathfrak{m}$. We use $\mathbb{C}$ to denote the set of complex numbers, $\mathbb{R}$ for real numbers, and $\mathbb{R}_+$ for non-negative real numbers. For a complex number $z = a + jb$, $j = \sqrt{-1}$, we denote its real part $\mathbb{R}\text{e}[z] = a$ and its imaginary part $\mathbb{I}\text{m}[z] = b$. We use $\boldsymbol{z}^H$ to denote the conjugate transpose of the complex vector $\boldsymbol{z}$, $|\mathcal{A}|$ for the cardinality of $\mathcal{A}$, and $\text{supp}(X) = \{\beta \in \Omega \mid X(\beta) \neq 0\}$ for the support of $X : \Omega \to \mathbb{C}$. We define the indicator function $\mathbb{I} : \Omega \to \{0, 1\}$ as $\mathbb{I}(\beta \in \mathcal{E}) = 1$, if β belongs to the event $\mathcal{E}$, and zero otherwise.

[1] Throughout this work, we use the term "continuous" *only* in contrast to "discrete" and not to refer to a smoothness property.

2. PROBLEM FORMULATION

Let $(\Omega, \mathcal{B})$ be a measurable space in which $\mathcal{B}$ are the Borel sets of Ω, a compact set of the real line. We define the sparse functional optimization problem as

$$\begin{aligned} \underset{X\in L_2}{\text{minimize}} \quad & \int_\Omega \left[\mathbb{I}(X(\beta)\neq 0) + \lambda|X(\beta)|^2\right] d\beta \\ \text{subject to} \quad & \left\| \boldsymbol{y} - \int_\Omega \boldsymbol{h}(\beta)X(\beta)d\beta \right\|_2^2 \leq w, \end{aligned} \tag{PI}$$

where $\boldsymbol{y} \in \mathbb{C}^m$; $\boldsymbol{h} : \Omega \to \mathbb{C}^m$ is a vector-valued function whose elements $h_i \in L_2$ are linearly independent, measurable functions for $i = 1, \dots, m$; $w > 0$ is a fit parameter; and $\lambda > 0$ is a regularization parameter that controls shrinkage. Unless noted otherwise, all integrals are taken with respect to the Lebesgue measure. In this work, however, we will mostly focus on its reformulation

$$\begin{aligned} \underset{X\in L_2}{\text{minimize}} \quad & \int_\Omega \left[\mathbb{I}(X(\beta)\neq 0) + \lambda|X(\beta)|^2\right] d\beta \\ \text{subject to} \quad & \|\boldsymbol{y} - \boldsymbol{\Theta}\|_2^2 \leq w \\ & \boldsymbol{\Theta} = \int_\Omega \boldsymbol{h}(\beta)X(\beta)d\beta \end{aligned} \tag{PI$'$}$$

in which $\boldsymbol{\Theta} \in \mathbb{C}^m$ is a dummy vector. Although (PI) and (PI$'$) are equivalent, the dual of (PI$'$) separates across β and is therefore easier to solve (Sec. 4). We therefore refer to them interchangeably, but derive all of our results for (PI$'$).

We refer to (PI) as a sparse functional problem because it seeks the functional linear model with smallest support that fits the measurements $\boldsymbol{y}$. Indeed, observe that the objective of (PI) is the measure of the support of X: $\int_\Omega \mathbb{I}(X(\beta) \neq 0)d\beta = \mathfrak{m}\left[\text{supp}(X)\right]$. Bear in mind that although we seek a sparse solution $X^\star$ in (PI), nowhere do we assume the true parameter is sparse.

We illustrate the use of (PI) in a line spectral estimation problem, in which we are given noisy measurements from a superposition of n complex sinusoids and wish to estimate their frequency, amplitude, and phase. These measurements can be written in functional form as

$$y_i = \int_0^1 \exp(j2\pi\beta t_i)X^o(\beta)d\beta + v_i, \quad \text{for } i = 1, \dots, m, \tag{1}$$

where the t_i are the sampling times, v_i are zero-mean random variable representing measurement noise, and X^o defines the active frequencies. If sources and sensors are ideal, then $X^o(\beta) = \sum_{k=1}^n z_k\delta(\beta - f_k)$, where δ denotes the Dirac delta and $z_k \in \mathbb{C}$ determines the amplitude and phase of the component with frequency $f_k \in [0, 1]$. Phase noise and modulation errors can be accounted for by using kernels instead of impulses as in $X^o(\beta) = \sum_{k=1}^n z_k\kappa(\beta - f_k)$ for $\kappa(u) = e^{-u^2/2\sigma^2}$, where σ determines the bandwidth of the kernel. It is ready that the integral in (1) has the same form as that in the constraint of (PI) with $\Omega = [0, 1]$ and $h_i(\beta) = \exp(j2\pi\beta t_i)$. Note that the h_i take on the values of complex exponentials with different frequencies β but at fixed times t_i and are therefore square integrable.

Though useful, (PI) is both infinite dimensional and non-convex. Even more so, its discrete version is known to be NP-hard [10]. Here, making the problem discrete turns out to make it intractable. Thus, we turn to its dual problem. Indeed, the dual of its equivalent problem (PI$'$) has dimensionality on the order of the number of measurements m. Moreover, we know from duality theory that dual problems are always convex programs [18]. The dual of (PI$'$) can therefore be solved efficiently (e.g., using Alg. 1). We defer its derivation and solution to Sec. 4 and first focus on the strong duality of (PI$'$), i.e., whether its dual problem is even worth solving. Indeed, though semi-infinite convex programs are often solved using duality [15–17, 19], (PI) is not convex and is therefore not necessarily strongly dual.

3. STRONG DUALITY OF NON-CONVEX FUNCTIONAL PROGRAMS

We have argued that the dual problem of (PI$'$) is a finite dimensional convex program that can be solved efficiently. However, we are ultimately interested in the solution of (PI$'$) and since it is non-convex, there is no reason to expect that the optimal value of its dual is anything more than a lower bound on the optimal value of (PI$'$) [18]. The question therefore remains as to whether the duality approach is worth pursing. The main result of this section tackles this limitation by showing that we can solve (PI$'$) using its dual.

Theorem 1. *Suppose that $\boldsymbol{h}$ has no point masses (Dirac deltas) and that Slater's condition holds for* (PI$'$)*. Then, strong duality holds for* (PI$'$)*, i.e., if P is the optimal value of* (PI$'$) *and D is the optimal value of its dual, then $P = D$.*

Theorem 1 states that though (PI$'$) is a non-convex functional program, it is strongly dual. In Sec. 4, we show that this implies (PI$'$) can be solved exactly and efficiently using duality. A noteworthy feature of this approach is that it precludes discretization by tackling (PI$'$) directly. Discretizing (PI$'$) not only results in an NP-hard problem, but it also leads to large dimensional, potentially ill-conditioned estimation problems. It is also worth noting that Theorem 1 is a *non-parametric* result in the sense that it makes no assumption on the existence or validity of a true measurement model. More to the point, it does not assume that $\boldsymbol{y}$ arises from a specific model for which the true parameter is sparse: it simply provides an efficient method for solving (PI$'$). Hence, we can determine the sparsest linear model that fits $\boldsymbol{y}$ regardless of whether they arise from sparse linear measurements. This is important because there are arguments for obtaining sparse solutions that are not epistemological, such as reducing computational or measurement costs.

Before proceeding with the proof of Theorem 1, it is worth noting that finding a strictly feasible point for (PI$'$) is trivial, so that Slater's condition always holds [18].

Proposition 1. *There exists a measurable function $X^\ddagger \in L_2$ such that $\int_\Omega \boldsymbol{h}(\beta)X^\ddagger(\beta)d\beta = \boldsymbol{y}$.*

Proof. See extended version in [20]. ■

Proof of Thm. 1. This proof relies on a well-known result from perturbation theory connecting strong duality to the convexity of the perturbation function [21, 22]. Formally, define the perturbed version of (PI$'$) for some perturbation $\epsilon \in \mathbb{R}$ as

$$\begin{aligned} \underset{X\in L_2}{\text{minimize}} \quad & s \\ \text{subject to} \quad & f_0(X) \leq s + \epsilon \\ & \|\boldsymbol{y} - \boldsymbol{\Theta}\|_2^2 \leq w \\ & \boldsymbol{\Theta} = \int_\Omega \boldsymbol{h}(\beta)X(\beta)d\beta \end{aligned} \tag{$\widetilde{\text{PI}}$}$$

where $f_0(X) = \int_\Omega \mathbb{I}(X(\beta) \neq 0) + \lambda|X(\beta)|^2 d\beta$. Note that we use the epigraph trick to linearize the objective [18]. Let the perturbation function $P(\epsilon) = s^\star(epsilon)$ be the optimal value of ($\widetilde{\text{PI}}$) for the perturbation ϵ. Notice that since ($\widetilde{\text{PI}}$) is equivalent to (PI$'$) for $\epsilon = 0$, it holds that $P(0) = P$, the optimal value of the original problem.

Proposition 2. *If (i)* (PI$'$) *satisfies Slater's condition and (ii) the perturbation function $P(\epsilon)$ is convex, then strong duality holds for* (PI$'$).

Proof. See, e.g., [21, Cor. 30.2.2] or [23, Thm. 4.1.1]. ■

Condition (i) of Proposition 2 is satisfied by the hypotheses of Theorem 1. Suffices then to show that the perturbation function is convex [(ii)], i.e., that for every ϵ, ϵ', and $\theta \in [0, 1]$,

$$P\left[\theta\epsilon + (1-\theta)\epsilon'\right] \leq \theta P\left(\epsilon\right) + (1-\theta)P\left(\epsilon'\right). \tag{2}$$

We can do so using the following lemma whose proof relies on Lyapunov's convexity theorem [24] and can be found in [20]:

Lemma 1. *The range of the constraints of* ($\widetilde{\text{PI}}$) *given by*

$$\mathcal{C} = \big\{c : \exists X \in L_2 \text{ s.t. } c = f_0(X),\ \|\boldsymbol{y} - \boldsymbol{\Theta}\|_2^2 \leq w, \\ \text{and } \boldsymbol{\Theta} = \int_\Omega \boldsymbol{h}(\beta)X(\beta)d\beta\big\}. \tag{3}$$

is a convex set.

Suppose now that $P(\epsilon)$ and $P(\epsilon')$ are achieved for the functions X and X', respectively:

$$f_0(X) \leq P(\epsilon) + \epsilon \quad \text{and} \quad f_0(X') \leq P(\epsilon') + \epsilon'. \tag{4}$$

Since X and X' are solutions of ($\widetilde{\text{PI}}$), they are (PI$'$)-feasible [satisfy the second and third constraints of ($\widetilde{\text{PI}}$)] and $f_0(X), f_0(X') \in \mathcal{C}$. From Lemma 1, we can obtain another (PI$'$)-feasible function X_θ such that

$$f_0(X_\theta) = \theta f_0(X) + (1-\theta)f_0(X'). \tag{5}$$

Combining (4) and (5) yields

$$f_0(X_\theta) \leq \theta P_2(\epsilon) + (1-\theta)P_2(\epsilon') + \left[\theta\epsilon + (1-\theta)\epsilon'\right].$$

Hence, X_θ is ($\widetilde{\text{PI}}$)-feasible for the perturbation $\theta\epsilon + (1-\theta)\epsilon'$ and its value is $s_\theta = \theta P(\epsilon) + (1-\theta)P(\epsilon')$. However, optimality of the perturbation function implies that $P\left[\theta\epsilon + (1-\theta)\epsilon'\right] \leq s_\theta$, giving (2). Lemma 1 therefore implies the perturbation function of (PI$'$) is convex [(ii) in Proposition 2], which concludes our proof. ■

4. SOLVING THE DUAL FUNCTIONAL PROBLEM

Having established duality as a fruitful approach to solving the sparse functional program (PI$'$), we now derive its dual problem and an algorithm to solve it.

Start by noting that due to the complex-valued equality, the Lagrangian of (PI$'$) actually has three dual variables: $\nu \in \mathbb{R}_+$ corresponds to the inequality constraint, $\boldsymbol{\mu}_R \in \mathbb{R}^m$ corresponds to the real part of the equality constraint, and $\boldsymbol{\mu}_I \in \mathbb{R}^m$ corresponds to its imaginary part. We can however combine these last two into a single complex-valued dual variable by noticing that for any $\boldsymbol{z} \in \mathbb{C}^m$ we have $\boldsymbol{\mu}_R^T \operatorname{Re}[\boldsymbol{z}] + \boldsymbol{\mu}_I^T \operatorname{Im}[\boldsymbol{z}] = \operatorname{Re}\left[\boldsymbol{\mu}^H \boldsymbol{z}\right]$ with $\boldsymbol{\mu} = \boldsymbol{\mu}_R + j\boldsymbol{\mu}_I$. Hence, the Lagrangian of (PI$'$) is defined as

$$\mathcal{L}(X, \boldsymbol{\Theta}, \boldsymbol{\mu}, \nu) = \int_\Omega \mathbb{I}(X(\beta) \neq 0) + \lambda|X(\beta)|^2 d\beta \\ + \operatorname{Re}\left[\boldsymbol{\mu}^H \left(\int_\Omega \boldsymbol{h}(\beta)X(\beta)d\beta - \boldsymbol{\Theta}\right)\right] \\ + \nu\left[\|\boldsymbol{y} - \boldsymbol{\Theta}\|_2^2 - w\right], \tag{6}$$

for the dual variables $\boldsymbol{\mu} \in \mathbb{C}^m$ and $\nu \in \mathbb{R}_+$. Its dual function can then be written as

$$d(\boldsymbol{\mu}, \nu) = \min_{X \in L_2, \boldsymbol{\Theta}} \mathcal{L}(X, \boldsymbol{\Theta}, \boldsymbol{\mu}, \nu), \tag{7}$$

so that the dual problem of (PI$'$) is defined as

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

The minimization in (7) actually has a straightforward solution. First, split the joint minimization to get

$$d(\boldsymbol{\mu}, \nu) = \min_{X \in L_2} \int_\Omega F(\beta, X(\beta))d\beta \\ + \min_{\boldsymbol{\Theta}} \nu \|\boldsymbol{y} - \boldsymbol{\Theta}\|_2^2 - \operatorname{Re}[\boldsymbol{\mu}^H \boldsymbol{\Theta}] - \nu w, \tag{8}$$

with $F(\beta, x) = \mathbb{I}(x \neq 0) + \lambda|x|^2 + \operatorname{Re}\left[\boldsymbol{\mu}^H \boldsymbol{h}(\beta)x\right]$. Minimizing over $\boldsymbol{\Theta}$ is a simple quadratic program since $\nu \geq 0$. Its solution can be written explicitly as

$$\frac{\partial}{\partial \boldsymbol{\Theta}} \left(\nu \|\boldsymbol{y} - \boldsymbol{\Theta}\|_2^2 - \operatorname{Re}[\boldsymbol{\mu}^H \boldsymbol{\Theta}]\right) = 0 \Rightarrow \boldsymbol{\Theta}_d(\boldsymbol{\mu}, \nu) = \boldsymbol{y} + \frac{\boldsymbol{\mu}}{2\nu}. \tag{9}$$

Despite its non-convex nature, the minimization over X also has a closed form because we can separate the optimization across β.

Lemma 2. *Let F be defined as in* (8). *Then,*

$$\inf_{X \in L_2} \int_\Omega F\left[\beta, X(\beta)\right] d\beta = \int_\Omega \inf_{x \in \mathbb{C}} F(\beta, x)d\beta. \tag{10}$$

Proof. See [25, Thm. 3A]. ■

We can therefore solve individually for each β

$$\min_{X(\beta)} \mathbb{I}(X(\beta) \neq 0) + \lambda|X(\beta)|^2 + \operatorname{Re}\left[\boldsymbol{\mu}^H \boldsymbol{h}(\beta)X(\beta)\right],$$

which leads to

$$X_d(\beta, \boldsymbol{\mu}) = \begin{cases} -\frac{1}{2\lambda}\boldsymbol{h}(\beta)^H \boldsymbol{\mu}, & \left|\boldsymbol{\mu}^H \boldsymbol{h}(\beta)\right|^2 > 4\lambda \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Notice that the phase of X_d is the complement of the phase of $\boldsymbol{\mu}^H \boldsymbol{h}(\beta)$, since this maximizes the magnitude of their product by making it a purely real number. With (9) and (11) in hand, we can evaluate the objective of (DI) explicitly to get

$$d(\boldsymbol{\mu}, \nu) = \mathfrak{m}(\mathcal{S}) - \frac{1}{4\lambda}\boldsymbol{\mu}^H \boldsymbol{H} \boldsymbol{\mu} - \operatorname{Re}\left[\boldsymbol{\mu}^H \boldsymbol{y}\right] - \frac{\|\boldsymbol{\mu}\|_2^2}{4\nu} - \nu w, \tag{12}$$

with $\boldsymbol{H} = \int_\mathcal{S} \boldsymbol{h}(\beta)\boldsymbol{h}(\beta)^H d\beta$ and $\mathcal{S} = \{\beta \mid \left|\boldsymbol{\mu}^H \boldsymbol{h}(\beta)\right|^2 > 4\lambda\}$.

Solving (DI) yields the optimal dual variables $\boldsymbol{\mu}^\star$ and $\nu^\star$, from which we can recover $X^\star$, the solution of (PI$'$), using Theorem 1. Indeed, the strong duality of (PI$'$) implies that $(X^\star, \boldsymbol{\Theta}^\star) \in \operatorname{argmin}_{X, \boldsymbol{\Theta}} \mathcal{L}(X, \boldsymbol{\Theta}, \boldsymbol{\mu}^\star, \nu^\star)$ for the Lagrangian in (6) [18]. Since (9) and (11) yield unique minimizers for each dual variable pair $(\boldsymbol{\mu}, \nu)$, this is a singleton set and we obtain that

$$X^\star(\beta) = X_d(\beta, \boldsymbol{\mu}^\star)$$

for X_d in (11). All that remains is to derive a procedure to solve (DI).

Algorithm 1 Stochastic dual ascent for functional optimization

```
$\boldsymbol{\mu}_0 = \mathbf{0}$, $\nu_0 = 1$
for $t = 1, \dots, T$
    Draw $\beta_j$ uniformly at random from $\Omega$
    $\bar{\boldsymbol{H}} = \frac{1}{p} \sum_{j=1}^{p} \boldsymbol{h}(\beta_j) \boldsymbol{h}(\beta_j)^H \, \mathbb{I}\left( \left| \boldsymbol{\mu}^H \boldsymbol{h}(\beta_j) \right|^2 > 4\lambda \right)$
    $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} - \eta_t \left[ \frac{1}{2\lambda\nu_{t-1}} \left( \nu_{t-1} \bar{\boldsymbol{H}} + \lambda \boldsymbol{I} \right) \boldsymbol{\mu}_{t-1} + \boldsymbol{y} \right]$
    $\nu_t = \left[ \nu_{t-1} + \eta_k \left( \frac{\|\boldsymbol{\mu}_{t-1}\|_2^2}{4\nu_{t-1}^2} - w \right) \right]_+$
    $X_t(\beta) = \begin{cases} -\frac{1}{2\lambda} \boldsymbol{h}(\beta)^H \boldsymbol{\mu}_t, & \left| {\boldsymbol{\mu}_t}^H \boldsymbol{h}(\beta) \right|^2 > 4\lambda \\ 0, & \text{otherwise} \end{cases}$
end
$X_f(\beta) = \frac{1}{T} \sum_{t=1}^{T} X_t(\beta)$
```

To do so, we use the fact that the dual function is concave and perform gradient ascent [18]. Recall that the gradient of d with respect to the dual variables is given by the constraint slacks in (6), i.e.,

$$\nabla_{\boldsymbol{\mu}} d = -\frac{1}{4\lambda} \boldsymbol{\mu}^H \boldsymbol{H} - \frac{1}{2} \boldsymbol{y}^H - \frac{1}{4\nu} \boldsymbol{\mu}^H \tag{13a}$$

$$\nabla_{\nu} d = \frac{\|\boldsymbol{\mu}\|_2^2}{4\nu^2} - w \tag{13b}$$

where we used (9), (11), and the definition of $\boldsymbol{H}$ from (12). Though (13a) involves computing an integral to evaluate $\boldsymbol{H}$, notice that the integrand $\boldsymbol{h}$ is a problem constant. Only the domain $\mathcal{S}$ depends on $\boldsymbol{\mu}$. Hence, a closed form for the Gram matrix $\boldsymbol{H}$ could be obtained for certain applications.

It may happen, however, that explicit expressions for $\boldsymbol{H}$ are not available or too cumbersome to be useful in practice. In these cases, we can leverage ideas from stochastic gradient descent and solve (DI)/(PI′) using Alg. 1. This procedure is obtained by approximating $\boldsymbol{H}$ using Monte Carlo integration, i.e., by drawing a set of β_j independently and uniformly at random from Ω and taking

$$\bar{\boldsymbol{H}} = \frac{1}{p} \sum_{j=1}^{p} \boldsymbol{h}(\beta_j) \boldsymbol{h}(\beta_j)^H \, \mathbb{I}\left(\left| \boldsymbol{\mu}^H \boldsymbol{h}(\beta_j) \right|^2 > 4\lambda \right). \tag{14}$$

Since Monte Carlo integration yields an unbiased estimator of the integral, replacing $\boldsymbol{H}$ by $\bar{\boldsymbol{H}}$ in (13a) gives an unbiased estimator of the gradient. In fact, Alg. 1 for $p = 1$ can be interpreted as performing stochastic gradient ascent on d. For $p > 1$, it becomes a mini-batch type algorithm. Hence, typical convergence guarantees can also be obtained for Alg. 1 [26]. We leave these results for future work.

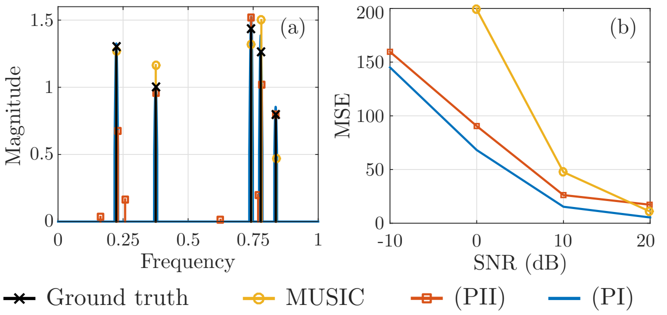

Fig. 1. Line spectral estimation: (a) amplitude estimates (SNR = 10 dB); (b) MSE for different SNRs.

5. APPLICATION: LINE SPECTRAL ESTIMATION

In this section, we illustrate the previous results in the line spectral estimation application from Sec. 2. We compare the result of three methods: MUSIC, atomic norm relaxation, and Alg. 1. In all cases, we use the methods to identify the support and recover the amplitudes and phases using least squares. MUSIC is a classical solution to line spectral estimation based on the eigendecomposition of the empirical autocorrelation matrix of the measurements $\boldsymbol{y}$ [1]. When the signal is sampled regularly, MUSIC can be used with a single snapshot—see [1] for details. It requires that the number n of sinusoids be known *a priori*. The second approach uses a convex relaxation to approximate the sparse estimation problem [15, 16]. This relaxation can be written as the semidefinite program

$$\begin{aligned} \underset{\boldsymbol{u}, \boldsymbol{x}, t}{\text{minimize}} \quad & \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{\tau}{2}(t + u_1) \\ \text{subject to} \quad & \begin{bmatrix} T(\boldsymbol{u}) & \boldsymbol{x} \\ \boldsymbol{x}^H & t \end{bmatrix} \succeq 0 \end{aligned} \tag{PII}$$

where $T(\boldsymbol{a})$ is a Hermitian Toeplitz matrix with entries from $\boldsymbol{a}$ and $\tau > 0$ is a regularization parameter. The support is estimated from the maxima of a polynomial obtained from $\boldsymbol{x}^\star$, the solution of (PII). Finally, we show results from Alg. 1. Note that though X^o contains atoms (see Sec. 2), $X^\star$ does not [$X \in L_2$ for (PI)]. We instead get a mass accumulation around each active frequency (Fig. 1a).

Take the t_i in (1) to be integers in $[-50, 50]$ and let the $v_i \in \mathbb{C}$ be independent zero-mean circular Gaussian random variables with variance σ_v^2. Let X^o be a sum of $n = 5$ Dirac deltas randomly placed in $[0, 1]$ and with amplitudes ($[0.5, 1.5]$) and phases ($[0, 2\pi]$) drawn uniformly at random. For MUSIC, we use the actual number of spectral lines n. For PII, we take τ to be the optimal regularizer from [15], which depends on σ_v^2. For Alg. 1, we use $p = 100$, $\lambda = 1$, $\eta_k = 0.09/(1 + 4k)$, and $w = \sigma_v^2$.

The estimation MSE for different levels of noise are shown in Fig. 1b. In high SNR, all methods have similar performance and correctly recover the active frequencies, amplitudes, and phases. As the SNR decreases, MUSIC's performance degrades considerably, whereas the MSE of (PI) and (PII) remain comparable. Notice, however, that the support estimated by the convex relaxation has errors already in high SNR (Fig. 1a).

6. CONCLUSION

We proposed to tackle continuous problems with sparsity priors directly by solving a sparse functional optimization problem. To do so, we showed that this problem has no duality gap and can therefore be solved through its dual. This allows us to bypass the infinite dimensionality and non-convexity hurdles of the original problem and put forward a simple algorithm to solve these non-convex functional programs. We illustrated this method in a line spectral estimation application, but foresee that this technique can be applied to a wide variety of problem. Future works include extending Theorem 1 to problems involving nonlinear measurement models and improve Alg. 1 using second-order methods and variance reduction techniques [27].

7. ACKNOWLEDGEMENTS

Dr. Ribeiro's and Mr. Chamon's work is supported by the National Science Foundation CCF 1717120 and ARO W911NF1710438.

8. REFERENCES

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

[2] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE Trans. Signal Process.*, vol. 59[10], pp. 4735–4744, 2011.

[3] O. Bar-Ilan and Y. C. Eldar, "Sub-Nyquist radar via Doppler focusing," *IEEE Trans. Signal Process.*, vol. 62[7], pp. 1796–1811, 2014.

[4] M. Unser, "Sampling—50 years after Shannon," *Proc. IEEE*, vol. 88[4], pp. 569–587, 2000.

[5] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Process.*, vol. 50[6], pp. 1417–1428, 2002.

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

[7] C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2007.

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

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

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

[20] L.F.O. Chamon, Y. C. Eldar, and A. Ribeiro, "Strong duality of sparse functional optimization," 2017, `http://bit.ly/2zVHJLy`.

[21] R. T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.

[22] A. Ribeiro, "Optimal resource allocation in wireless communication and networking," *EURASIP J. on Wireless Commun. and Network.*, vol. 2012[1], 2012.

[23] A. Shapiro, "Duality, optimality conditions, and perturbation analysis," in *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, H. Wolkowicz, R. Saigal, and L. Vandenberghe, Eds., pp. 67–91. Springer, 2000.

[24] J. Diestel and J. J. Uhl, Jr., *Vector measures*, AMS, 1977.

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

[26] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," 2016, arXiv:1606.04838.

[27] A. Mokhtari, H. Daneshmand, A. Lucchi, T. Hofmann, and A. Ribeiro, "Adaptive newton method for empirical risk minimization to statistical accuracy," in *NIPS*, 2016, pp. 4062–4070.

A. PROOF OF PROPOSITION 1

Proof. Consider a measurable partition of Ω of size p, i.e., $\mathcal{O}_k \in \mathcal{B}$, $\bigcup_{k=1}^{p} \mathcal{O}_k = \Omega$, and $\mathcal{O}_j \cap \mathcal{O}_k = \emptyset$ for all $j \neq k$. We proceed by constructing a simple function $X^\ddagger$ that takes at most p different values and whose value over each set $\mathcal{O}_k$ is denoted by x_k. Hence, we can write

$$\int_\Omega \boldsymbol{h}(\beta) X^\ddagger(\beta) d\beta = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{A} = \left[\int_{\mathcal{O}_1} \boldsymbol{h}(\beta) d\beta \cdots \int_{\mathcal{O}_p} \boldsymbol{h}(\beta) d\beta\right]$ is an $m \times p$ matrix and $[\boldsymbol{x}]_k = x_k$ is a $p \times 1$ vector. Since the h_i are linearly independent and $h_i \in L_2(\Omega) \subset L_1(\Omega)$ for compact $\Omega \subset \mathbb{R}$, there exists a $p > m$ and a partition $\{\mathcal{O}_k\}$ such that $\boldsymbol{A}$ is a full row-rank matrix with finite elements. From any solution of the system of linear equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, we can construct a simple, bounded function $X^\ddagger$ that satisfies the conditions of the proposition. Since the $\mathcal{O}_k$ are measurable, so is $X^\ddagger$. ■

B. PROOF OF LEMMA 1

Proof. Let $c, c' \in \mathcal{C}$. Then from (3), there exist (PI$'$)-feasible $X, X' \in L_2$ such that $c = f_0(X)$ and $c' = f_0(X')$. To show that $\mathcal{C}$ is convex, we need to determine a feasible $X_\theta \in L_2$ for which

$$\theta c + (1-\theta)c' = \theta f_0(X) + (1-\theta) f_0(X') = f_0(X_\theta), \qquad (15)$$

in other words, we must show that $\theta c + (1-\theta)c' \in \mathcal{C}$.

To do so, we define a $2(m+1) \times 1$ vector measure $\mathfrak{p}$ over $(\Omega, \mathcal{B})$ such that for every set $\mathcal{Z} \in \mathcal{B}$ [24]

$$\mathfrak{p}(\mathcal{Z}) = \begin{bmatrix} \int_{\mathcal{Z}} \boldsymbol{h}(\beta) X(\beta) d\beta \\ \int_{\mathcal{Z}} \boldsymbol{h}(\beta) X'(\beta) d\beta \\ \int_{\mathcal{Z}} \mathbb{I}(X(\beta) \neq 0) + \lambda |X(\beta)|^2 d\beta \\ \int_{\mathcal{Z}} \mathbb{I}(X'(\beta) \neq 0) + \lambda |X'(\beta)|^2 d\beta \end{bmatrix}. \qquad (16)$$

We care about the value of $\mathfrak{p}$ only on two different sets: the empty set and Ω. Since $\mathfrak{p}$ is a proper measure, it immediately holds that $\mathfrak{p}(\emptyset) = 0$. For Ω, the integrals in (16) match those in (PI) and we write

$$\mathfrak{p}(\Omega) = \begin{bmatrix} \int_\Omega \boldsymbol{h}(\beta) X(\beta) \mathfrak{m}(d\beta) \\ \int_\Omega \boldsymbol{h}(\beta) X'(\beta) \mathfrak{m}(d\beta) \\ f_0(X) \\ f_0(X') \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{\Theta}' \\ c \\ c' \end{bmatrix}, \qquad (17)$$

where $\boldsymbol{\Theta}, \boldsymbol{\Theta}' \in \mathbb{C}^m$. Recall that since X and X' are (PI$'$)-feasible, we have that $\|\boldsymbol{y} - \boldsymbol{\Theta}\|_2^2 \leq w$ and $\|\boldsymbol{y} - \boldsymbol{\Theta}'\|_2^2 \leq w$.

Given that $\boldsymbol{h}$ has no point masses, the measure induced by $\boldsymbol{h}$ is non-atomic. Consequently, so is $\mathfrak{p}$. We can therefore apply Lyapunov's convexity theorem to obtain that the range of $\mathfrak{p}$ is convex [24]. Hence, for every $\theta \in [0,1]$, there exists a set $\mathcal{T}_\theta \in \mathcal{B}$ such that

$$\mathfrak{p}(\mathcal{T}_\theta) = \theta \mathfrak{p}(\Omega) + (1-\theta)\mathfrak{p}(\emptyset) = \theta \mathfrak{p}(\Omega). \qquad (18)$$

Since $\mathcal{B}$ is a σ-algebra, it holds that $\Omega \setminus \mathcal{T}_\theta \in \mathcal{B}$ and by the additivity of measures we get

$$\mathfrak{p}(\Omega \setminus \mathcal{T}_\theta) = (1-\theta)\mathfrak{p}(\Omega). \qquad (19)$$

Using (18) and (19), we now construct X_θ as

$$X_\theta(\beta) = \begin{cases} X(\beta), & \text{for } \beta \in \mathcal{T}_\theta \\ X'(\beta), & \text{for } \beta \in \Omega \setminus \mathcal{T}_\theta \end{cases} \qquad (20)$$

and claim that it is (PI$'$)-feasible and satisfies (15).

To prove this is indeed the case, we start by showing that X_θ is feasible. Using (20), we obtain

$$\begin{aligned} \boldsymbol{\Theta}_\theta = \int_\Omega \boldsymbol{h}(\beta) X_\theta(\beta) d\beta = \int_{\mathcal{T}_\theta} &\boldsymbol{h}(\beta) X(\beta) d\beta \\ &+ \int_{\Omega \setminus \mathcal{T}_\theta} \boldsymbol{h}(\beta) X'(\beta) d\beta, \end{aligned}$$

which from (16) is equivalent to

$$\boldsymbol{\Theta}_\theta = \int_\Omega \boldsymbol{h}(\beta) X_\theta(\beta) d\beta = \boldsymbol{C}_1 \mathfrak{p}(\mathcal{T}_\theta) + \boldsymbol{C}_2 \mathfrak{p}(\Omega \setminus \mathcal{T}_\theta), \qquad (21)$$

where $\boldsymbol{C}_1, \boldsymbol{C}_2$ are selection matrices containing the rows 1 through m and $m+1$ through $2m$ of a $2(m+1)$-dimensional identity matrix. In other words, $\boldsymbol{C}_1$ selects the first m rows of $\mathfrak{p}$ and $\boldsymbol{C}_2$ selects the next m. Then, (18) and (19) yield

$$\begin{aligned} \boldsymbol{\Theta}_\theta = \int_\Omega \boldsymbol{h}(\beta) X_\theta(\beta) d\beta &= \theta \boldsymbol{C}_1 \mathfrak{p}(\Omega) + (1-\theta) \boldsymbol{C}_2 \mathfrak{p}(\Omega) \\ &= \theta \boldsymbol{b} + (1-\theta)\boldsymbol{b}', \end{aligned}$$

from which the feasibility of X_θ is established using the convexity of the ℓ_2 norm. Explicitly,

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{\Theta}_\theta\|_2^2 &= \left\|\boldsymbol{y} - \theta \boldsymbol{b} - (1-\theta)\boldsymbol{b}'\right\|_2^2 \\ &\leq \theta \|\boldsymbol{y} - \boldsymbol{b}\|_2^2 + (1-\theta)\left\|\boldsymbol{y} - \boldsymbol{b}'\right\|_2^2 \leq w. \end{aligned}$$

A similar machinery is used to show that X_θ satisfies (15):

$$\begin{aligned} f_0(X_\theta) &= \int_{\mathcal{T}_\theta} \mathbb{I}(X(\beta) \neq 0) + \lambda |X(\beta)|^2 d\beta \\ &\quad + \int_{\Omega \setminus \mathcal{T}_\theta} \mathbb{I}(X'(\beta) \neq 0) + \lambda \left|X'(\beta)\right|^2 d\beta \\ &= \boldsymbol{e}_{2m+1}^T \mathfrak{p}(\mathcal{T}_\theta) + \boldsymbol{e}_{2m+2}^T \mathfrak{p}(\Omega \setminus \mathcal{T}_\theta) \\ &= \theta \boldsymbol{e}_1^T \mathfrak{p}(\Omega) + (1-\theta)\boldsymbol{e}_2^T \mathfrak{p}(\Omega) = \theta c + (1-\theta)c', \end{aligned}$$

where $\boldsymbol{e}_i$ is a column vector of zeros except in the i-th position. The existence of such an $X_\theta \in L_2$ for any θ, c, and c' concludes the proof that $\mathcal{C}$ is convex. ■