Functional Nonlinear Sparse Models

Luiz F. O. Chamon ⓘ, *Student Member, IEEE*, Yonina C. Eldar ⓘ, *Fellow, IEEE*, and Alejandro Ribeiro ⓘ

Abstract—Signal processing is rich in inherently continuous and often nonlinear applications, such as spectral estimation, optical imaging, and super-resolution microscopy, in which sparsity plays a key role in obtaining state-of-the-art results. Coping with the infinite dimensionality and non-convexity of these problems typically involves discretization and convex relaxations, e.g., using atomic norms. Nevertheless, grid mismatch and other coherence issues often lead to discretized versions of sparse signals that are not sparse. Even if they are, recovering sparse solutions using convex relaxations requires assumptions that may be hard to meet in practice. What is more, problems involving nonlinear measurements remain non-convex even after relaxing the sparsity objective. We address these issues by directly tackling the continuous, nonlinear problem cast as a sparse functional optimization program. We prove that when these problems are non-atomic, they have no duality gap and can therefore be solved efficiently using duality and (stochastic) convex optimization methods. We illustrate the wide range of applications of this approach by formulating and solving problems from nonlinear spectral estimation and robust classification.

Index Terms—Functional optimization, sparsity, nonlinear models, strong duality, compressive sensing.

I. INTRODUCTION

THE analog and often nonlinear nature of the physical world make for two of the main challenges in signal and information processing applications. Indeed, there are many examples of inherently continuous[1] problems, such as spectral estimation, image recovery, and source localization [3]–[6], as well as nonlinear ones, e.g., magnetic resonance fingerprinting, spectrum cartography, and manifold data sparse coding [7]–[9]. These challenges are often tackled by imposing structure on the signals. For instance, bandlimited, finite rate of innovation, or union-of-subspaces signals can be processed using an appropriate discrete set of samples [10]–[13]. Functions in reproducing

kernel Hilbert spaces (RKHSs) also admit finite descriptions through variational results known as representer theorems [14], [15]. The infinite dimensionality of continuous problems is therefore often overcome by means of sampling theorems. Similarly, nonlinear functions with bounded total variation or lying in an RKHS can be written as a finite linear combination of basis functions. Under certain smoothness assumptions, nonlinearity can be addressed using "linear-in-the-parameters" methods.

Due to the limited number of measurements, however, discretization often leads to underdetermined problems. Sparsity priors then play an important role in achieving state-of-the-art results by leveraging the assumption that there exists a signal representation in terms of only a few atoms from an overparametrized dictionary [13], [16], [17]. Since fitting these models leads to non-convex (and possibly NP-hard [18]) problems, sparsity is typically replaced by a tractable relaxation based on an atomic norm (e.g., the ℓ_1-norm). For linear, incoherent dictionaries, these relaxed problems have be shown to retrieve the desired sparse solution [16], [17]. Nevertheless, discretized continuous problems rarely meet these conditions. Only in specific instances, such as line spectrum estimation, there exist guarantees for relaxations that forgo discretization [12], [19]–[24].

This discretization/relaxation approach, however, is not always effective. Indeed, discretization can lead to grid mismatch issues and even loss of sparsity: infinite dimensional sparse signals need not be sparse when discretized [25]–[27]. Also, 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 problems with potentially poor numerical properties (high condition number). In fact, ℓ_1-norm-based recovery of spikes on fine grids (essentially) finds twice the number of actual spikes and the number of support candidate points increases as the number of measurements decreases [28], [29]. Furthermore, performance guarantees for convex relaxations rely on incoherence assumptions (e.g., restricted isometry/eigenvalue properties) that may be difficult to meet in practice and are NP-hard to check [30]–[32]. Finally, these guarantees hold for linear measurements models.

Directly accounting for nonlinearities in sparse models makes a difficult problem harder, since the optimization program remains non-convex even after relaxing the sparsity objective. This is evidenced by the weaker guarantees existing for ℓ_1-norm relaxations in nonlinear compressive sensing problems [33], [34]. Though "linear-in-the-parameters" models, such as splines or kernel methods, may sometimes be used (e.g., spectrum cartography [8]), they are not applicable in general. Indeed, the number of kernels needed to represent a generic nonlinear

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

model may be so large that the solution is no longer sparse. What is more, there is no guarantee that these models meet the incoherence assumptions required for the convex relaxation to be effective [13], [16], [17].

In this work, we propose to forgo both discretization and relaxation and directly tackle the continuous problem using *sparse functional programming*. Although sparse functional programs (SFPs) combine the infinite dimensionality of functional programming with the non-convexity of sparsity and nonlinear atoms, we show that they are tractable under mild conditions (Theorem 1). To do so, this paper develops the theory of sparse functional programming by formulating a general SFP (Section II), deriving its Lagrangian dual problem (Section II-B), and proving that strong duality holds under mild conditions (Section III). This shows that SFPs can be solved exactly by means of their dual problems. Moreover, we use this result to obtain a relation between minimizing the support of a function ("L_0-norm") and its L_1-norm, even though the latter may yield non-sparse solutions (Section III-B). We then propose two algorithms to solve SFPs, based on subgradient and stochastic subgradient ascent, by leveraging different numerical integration methods (Section IV). Finally, we illustrate the expressiveness of SFPs by using them to cast different signal processing problems and provide numerical examples to showcase the effectiveness of this approach (Section V).

Throughout the paper, 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 write $\mathbb{Re}[z] = a$ for its real part and $\mathbb{Im}[z] = b$ for its imaginary part. We use $\boldsymbol{z}^H$ for the conjugate transpose of the complex vector $\boldsymbol{z}$, $[\boldsymbol{z}]_i$ to indicate its i-th component, $|\mathcal{A}|$ for the cardinality of $\mathcal{A}$, and $\text{supp}(X) = \{\boldsymbol{\beta} \in \Omega \mid X(\beta) \neq 0\}$ for the support of $X : \Omega \to \mathbb{C}$. For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we write $\boldsymbol{x} \succeq \boldsymbol{y}$ to denote that $[\boldsymbol{x}]_i \geq [\boldsymbol{y}]_i$ for all $i = 1, \ldots, n$. We define the indicator function $\mathbb{I} : \Omega \to \{0, 1\}$ as $\mathbb{I}(\boldsymbol{\beta} \in \mathcal{E}) = 1$, if $\boldsymbol{\beta}$ belongs to the event $\mathcal{E}$, and zero otherwise.

II. SPARSE FUNCTIONAL PROGRAMS AND DUALITY

A. Sparse Functional Programs

SFPs are variational problems that seek sparsest functions, i.e., functions with minimum support measure. Explicitly, let $(\Omega, \mathcal{B})$ be a measurable space in which $\mathcal{B}$ are the Borel sets of Ω, a compact set of $\mathbb{R}^n$. In a parallel with the discrete case, define the L_0-norm[2] to be the measure of the support of a function, i.e., for a measurable function $X : \Omega \to \mathbb{C}$,

$$\|X\|_{L_0} = \mathfrak{m}\left[\text{supp}(X)\right] = \int_\Omega \mathbb{I}\left[X(\boldsymbol{\beta}) \neq 0\right] d\boldsymbol{\beta}. \quad (1)$$

Note that the integral in (1) is a multivariate integral over vectors $\boldsymbol{\beta} \in \Omega$. Unless otherwise specified, all integrals are taken with respect to the Lebesgue measure $\mathfrak{m}$.

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

A general SFP is then defined as the optimization problem

$$\begin{aligned}
\underset{X \in \mathcal{X}, \boldsymbol{z} \in \mathbb{C}^p}{\text{minimize}} \quad & \int_\Omega F_0\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} + \lambda \|X\|_{L_0} \\
\text{subject to} \quad & g_i(\boldsymbol{z}) \leq 0, \quad i = 1, \ldots, m \\
& \boldsymbol{z} = \int_\Omega \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} \\
& X(\boldsymbol{\beta}) \in \mathcal{P} \text{ a.e.}
\end{aligned} \quad \text{(P-SFP)}$$

where $\lambda > 0$ is a regularization parameter that controls the sparsity of the solution; $g_i : \mathbb{C}^p \to \mathbb{R}$ are convex functions; $F_0 : \mathbb{C} \times \Omega \to \mathbb{R}$ is an optional, not necessarily convex regularization term (e.g., take $F_0(x, \boldsymbol{\beta}) = |x|^2$ for shrinkage); $\boldsymbol{F} : \mathbb{C} \times \Omega \to \mathbb{C}^p$ is a vector-valued (possibly nonlinear) function; $\mathcal{P}$ is a (possibly non-convex) set defining an almost everywhere (a.e.) pointwise constraints on X, i.e., a constraint that holds for all $\boldsymbol{\beta} \in \Omega$ except perhaps over a set of measure zero (e.g., $\mathcal{P} = \{x \in \mathbb{C} \mid |x| \leq \Gamma\}$ for some $\Gamma > 0$); and $\mathcal{X}$ is a *decomposable* 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

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

Lebesgue spaces (e.g., $\mathcal{X} = L_2$ or $\mathcal{X} = L_\infty$) or more generally Orlicz spaces are typical examples of decomposable function spaces. The spaces of constant or continuous functions, for instance, are *not* decomposable [35].

The linear, continuous sparse recovery/denoising problem is a particular case of (P-SFP). Here, we seek to represent a signal $\boldsymbol{y} \in \mathbb{R}^p$ as a linear combination of a continuum of atoms $\phi(\beta)$ indexed by $\beta \in \Omega \subset \mathbb{R}$, i.e., as $\int_\Omega X(\beta)\phi(\beta)d\beta$, using a sparse functional coefficient X. This problem can be posed as

$$\begin{aligned}
\underset{X \in L_2, \hat{\boldsymbol{y}} \in \mathbb{R}^p}{\text{minimize}} \quad & \|X\|_{L_2}^2 + \lambda \|X\|_{L_0} \\
\text{subject to} \quad & \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 \leq \epsilon \\
& \hat{\boldsymbol{y}} = \int_\Omega X(\beta)\phi(\beta)d\beta,
\end{aligned} \quad \text{(PI)}$$

where $\epsilon > 0$ is a goodness-of-fit parameter. Notice that when discretized, this problem yields the classical (NP-hard [18]) dictionary denoising problem with $\hat{\boldsymbol{y}} = \boldsymbol{\phi}^T \boldsymbol{x}$, where $\boldsymbol{\phi} = [\phi(\beta_j)]$ and $\boldsymbol{x} = [X(\beta_j)]$, for a set of $\beta_j \in \Omega$, $j = 1, \ldots, m$.

The expressiveness of SFPs comes from their ability to accommodate nonlinear measurement models (through $\boldsymbol{F}$) and non-convex objective functions. For instance, (PI) can also be posed using the nonlinear model

$$\hat{\boldsymbol{y}} = \int_\Omega \rho\left[X(\beta)\phi(\beta)\right] d\beta, \quad (2)$$

where ρ represents, for instance, a source saturation (as in, e.g., Section V). Yet, the abstract formulation in (P-SFP) certainly obfuscates the applicability of SFPs. Additionally, severe technical challenges, such as infinite dimensionality and non-convexity, appear to hinder their usefulness. We defer the issue of applicability to Section V, where we illustrate the use of SFPs in the

context of nonlinear spectral estimation and nonlinear functional data analysis. Instead, we first focus on whether problems of the form (P-SFP) can even be solved. Indeed, note that the discrete versions of certain SFPs are known to be NP-hard [18]. Hence, discretizing the functional problem in this case makes it intractable.

We propose to solve SFPs using duality. It is worth noting that duality is often used to solve semi-infinite convex programs [19]–[21], [36]. In these cases, strong duality holds under mild conditions and solving the dual problem leads to a solution of the original optimization problem of interest. However, SFPs are not convex. To address this issue, we first derive the dual problem of (P-SFP) in the next section, noting that it is both finite dimensional and convex. Then, we show that we can obtain a solution of (P-SFP) from a solution of its dual by proving that SFPs have zero duality gap under quite general conditions (Section III). Finally, we suggest different algorithms to solve the dual problem of (P-SFP) (Section IV).

B. The Lagrangian Dual of Sparse Functional Programs

To formulate the dual problem of (P-SFP), we first introduce the Lagrange multipliers $\nu_i \in \mathbb{R}_+$, corresponding to the inequalities $g_i(z) \leq 0$, and $\boldsymbol{\mu}_R, \boldsymbol{\mu}_I \in \mathbb{R}^p$, corresponding to the real and imaginary parts respectively of the complex-valued equality $\boldsymbol{z} = \int_\Omega \boldsymbol{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}]d\boldsymbol{\beta}$. To simplify the derivations, we collect the former into the vector $\boldsymbol{\nu} \in \mathbb{R}^m_+$ and combine the latter two multipliers into a single complex-valued dual variable by noticing that for any vector $\boldsymbol{x} \in \mathbb{C}^m$, it holds that $\boldsymbol{\mu}_R^T \mathbb{R}\mathrm{e}[\boldsymbol{x}] + \boldsymbol{\mu}_I^T \mathbb{I}\mathrm{m}[\boldsymbol{x}] = \mathbb{R}\mathrm{e}[\boldsymbol{\mu}^H \boldsymbol{x}]$, where $\boldsymbol{\mu} = \boldsymbol{\mu}_R + j\boldsymbol{\mu}_I$.

The Lagrangian dual of (P-SFP) is then defined as

$$\mathcal{L}(X, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \int_\Omega F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}]d\boldsymbol{\beta} + \lambda\|X\|_{L_0}$$
$$+ \mathbb{R}\mathrm{e}\left[\boldsymbol{\mu}^H\left(\int_\Omega \boldsymbol{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}]d\boldsymbol{\beta} - \boldsymbol{z}\right)\right] \quad (3)$$
$$+ \sum_{i=1}^m \nu_i g_i(\boldsymbol{z})$$

and its dual function is given by

$$d(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\substack{X \in \mathcal{X},\, \boldsymbol{z} \in \mathbb{C}^p, \\ X(\boldsymbol{\beta}) \in \mathcal{P}}} \mathcal{L}(X, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\nu}). \quad (4)$$

The fact that the pointwise constraint holds almost everywhere in Ω is omitted for conciseness. Thus, the dual problem of (P-SFP) is given by

$$\underset{\boldsymbol{\mu},\, \boldsymbol{\nu} \succeq 0}{\text{maximize}} \quad d(\boldsymbol{\mu}, \boldsymbol{\nu}). \quad \text{(D-SFP)}$$

By definition, (D-SFP) is a convex program whose dimensionality is equal to the number of constraints [37]—in this case, on the order of p. It is therefore tractable as long as we can evaluate the dual function d. Indeed, solving (D-SFP) is at least as hard as solving the minimization in (4). We next show that the dual function of SFPs is often efficiently computable.

The joint minimization in (4) separates as

$$d(\boldsymbol{\mu}, \boldsymbol{\nu}) = d_X(\boldsymbol{\mu}) + d_{\boldsymbol{z}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \quad (5)$$

with

$$d_X(\boldsymbol{\mu}) = \min_{\substack{X \in \mathcal{X}, \\ X(\boldsymbol{\beta}) \in \mathcal{P}}} \int_\Omega \Big\{ F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}] + \lambda\mathbb{I}[X(\boldsymbol{\beta}) \neq 0]$$
$$\qquad\qquad\qquad (6)$$
$$+ \mathbb{R}\mathrm{e}\left[\boldsymbol{\mu}^H \boldsymbol{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}]\right]\Big\} d\boldsymbol{\beta}$$

and $d_{\boldsymbol{z}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{z}} \sum_{i=1}^m \nu_i g_i(\boldsymbol{z}) - \mathbb{R}\mathrm{e}[\boldsymbol{\mu}^H \boldsymbol{z}]$. The minimum in $d_{\boldsymbol{z}}$ is tractable since the objective is convex, given that $\nu_i \geq 0$ and the g_i are convex functions. In certain cases, e.g., when g_i is a quadratic loss, $d_{\boldsymbol{z}}$ may even have a closed-form expression. On the other hand, d_X is in general a non-convex problem. When F_0 and $\boldsymbol{F}$ are normal integrands [35, Def. 14.27], this issue is addressed by exploiting the separability of the objective across $\boldsymbol{\beta}$ as shown in Proposition 1. Examples of normal integrands include functions $f(x, \boldsymbol{\beta})$ that are continuous in x for all fixed $\boldsymbol{\beta}$ and measurable in $\boldsymbol{\beta}$ for all fixed x (also known as Carathéodory) or when Ω is Borel and $f(\cdot, \boldsymbol{\beta})$ is lower semicontinuous for all fixed $\boldsymbol{\beta}$ [35]. Note that these functions can be nonlinear and need not be convex.

Proposition 1: Consider the functional optimization problem in (6) and assume that F_0 and the elements of $\boldsymbol{F}$ are normal integrands. Let $\gamma^{(0)}(\boldsymbol{\mu}, \boldsymbol{\beta}) = F_0(0, \boldsymbol{\beta}) + \mathbb{R}\mathrm{e}[\boldsymbol{\mu}^H \boldsymbol{F}(0, \boldsymbol{\beta})]$ and define

$$\gamma^o(\boldsymbol{\mu}, \boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathbb{R}\mathrm{e}\left[\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right]. \quad (7)$$

Then, for $\mathcal{S}(\boldsymbol{\mu}) = \{\boldsymbol{\beta} \in \Omega : \gamma^o(\boldsymbol{\mu}, \boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}, \boldsymbol{\beta}) - \lambda\}$,

$$d_X(\boldsymbol{\mu}) = \int_{\mathcal{S}(\boldsymbol{\mu})} [\lambda + \gamma^o(\boldsymbol{\mu}, \boldsymbol{\beta})]d\boldsymbol{\beta} + \int_{\Omega\backslash\mathcal{S}(\boldsymbol{\mu})} \gamma^{(0)}(\boldsymbol{\mu}, \boldsymbol{\beta})d\boldsymbol{\beta}. \quad (8)$$

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

Lemma 1 (Separability principle [35, Thm. 14.60]): Let $G(x, \boldsymbol{\beta})$ be a normal integrand and $\mathcal{X}$ be a decomposable space. Then,

$$\inf_{\substack{X \in \mathcal{X} \\ X(\boldsymbol{\beta}) \in \mathcal{P}}} \int_\Omega G[X(\boldsymbol{\beta}), \boldsymbol{\beta}]d\boldsymbol{\beta} = \int_\Omega \inf_{x \in \mathcal{P}} G(x, \boldsymbol{\beta})d\boldsymbol{\beta}. \quad (9)$$

Since Ω is a compact subset of $\mathbb{R}^n$ and the indicator function is lower semicontinuous [35, Ex. 14.31], the integrand in (6) is normal and we can restrict ourselves to solving the optimization individually for each $\boldsymbol{\beta}$, i.e., if X_d is a solution of (6), then

$$X_d(\boldsymbol{\beta}) \in \underset{x \in \mathcal{P}}{\arg\min}\, F_0[x, \boldsymbol{\beta}] + \lambda\mathbb{I}[x \neq 0] + \mathbb{R}\mathrm{e}\left[\boldsymbol{\mu}^H \boldsymbol{F}[x, \boldsymbol{\beta}]\right].$$
$$\text{(PII)}$$

Despite the non-convexity of the indicator function, (PII) is a scalar problem, whose solution involves a simple thresholding scheme. Indeed, only two conditions need to be checked: (i) if $X_d(\boldsymbol{\beta}) = 0$, then the indicator function vanishes and the objective of (PII) evaluates to $\gamma^{(0)}(\boldsymbol{\beta})$; (ii) if $X_d(\boldsymbol{\beta}) \neq 0$, then the indicator function is one and the objective of (PII) evaluates to $\lambda + \gamma^o(\boldsymbol{\beta})$. The value of (PII) is the minimum of these two cases, which from Lemma 1 yields the desired result in (8). ∎

Proposition 1 provides a practical way to evaluate (5), i.e., to evaluate the objective of the dual problem (P-SFP). Still, it

relies on the ability to efficiently solve (7), which may be an issue if F_0, $\boldsymbol{F}$, or $\mathcal{P}$ are non-convex. Nevertheless, (7) remains a scalar problem that can typically be solved efficiently using global optimization techniques [38] or through efficient local search procedures (see Section V).

The tractability of the dual problem (P-SFP) does not imply that it provides a solution to the original problem (P-SFP). In fact, since SFPs are not convex programs it is not immediate that (P-SFP) is worth solving at all: there is no reason to expect that the optimal value of (P-SFP) is anything more than a lower bound on the optimal value of (P-SFP) [37]. In the sequel, we proceed to show that this is not the case and that we can actually obtain a solution of (P-SFP) by solving (P-SFP).

III. STRONG DUALITY AND ITS IMPLICATIONS

Though we have argued that the dual problem of (P-SFP) is potentially tractable, we are ultimately interested in solving (P-SFP) itself. This section tackles this limitation by showing that (P-SFP) and (P-SFP) have the same values (Theorem 1). In Section IV, we show how, under mild conditions, this result allows us to efficiently find a solution for (P-SFP). Before that, however, we use strong duality to derive a relation between SFPs and L_1-norm optimization problems when their solution saturates (Section III-B).

A. Strong Duality of Sparse Functional Programs

The main result of this section is presented in the following theorem:

Theorem 1: Suppose that F_0 and $\boldsymbol{F}$ have no point masses (Dirac deltas) and that there exists a (P-SFP)-feasible pair $(X', \boldsymbol{z}')$, i.e., $X' \in \mathcal{X}$ with $X'(\boldsymbol{\beta}) \in \mathcal{P}$ a.e. and $\boldsymbol{z}' \in \mathbb{C}^p$, such that $\boldsymbol{z}' = \int_\Omega \boldsymbol{F}[X'(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta}$ and $g_i(\boldsymbol{z}') < 0$ for all $i = 1, \dots, m$. Then, strong duality holds for (P-SFP), i.e., if $P^\star$ is the optimal value of (P-SFP) and $D^\star$ is the optimal value of (P-SFP), then $P^\star = D^\star$.

Theorem 1 states that although (P-SFP) is a non-convex functional program, it has zero duality gap, suggesting that it can be solved through its tractable dual (P-SFP). A noteworthy feature of this approach is that it precludes discretization by tackling (P-SFP) directly. Discretizing (P-SFP) may not only result in NP-hard problems, but leads to high dimensional, potentially ill-conditioned problems. It is also worth noting that Theorem 1 is *non-parametric* in the sense that it makes no assumption on the existence or validity of the measurement model in (P-SFP). In particular, it does not require that the data arise from a specific model in which the parameters are sparse. This implies, for instance, that the sparsest functional linear model that fits a set of measurements can be determined regardless of whether these measurements arise from a truly sparse, linear model. This is useful in practice when sparse solutions are sought, not for epistemological reasons, but for reducing computational or measurement costs.

Proof of Theorem 1: Recall from weak duality that the dual problem is a lower bound on the value of the primal, so that $D^\star \leq P^\star$ [37]. Hence, it suffices to prove that $D^\star \geq P^\star$. To do so, denote the cost function of (P-SFP) by $f_0(X) \triangleq \int_\Omega F_0[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} + \lambda \|X\|_{L_0}$ and define the cost-constraints

set

$$\mathcal{C} = \left\{ (c, \boldsymbol{u}, \boldsymbol{k}_R, \boldsymbol{k}_I) \mid \exists (X, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{C}^p \text{ such that} \right.$$

$$X(\boldsymbol{\beta}) \in \mathcal{P} \text{ a.e.}, f_0(X) \leq c, g_i(\boldsymbol{z}) \leq [\boldsymbol{u}]_i, \qquad (10)$$

$$\left. \text{and } \int_\Omega \boldsymbol{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} - \boldsymbol{z} = \boldsymbol{k}_R + j\boldsymbol{k}_I \right\}.$$

In words, $\mathcal{C}$ describes the range of values taken by the objective and constraints of (P-SFP). Observe that (10) separates the real- and complex-valued parts of the equality constraint in (P-SFP). Hence, $\mathcal{C} \subset \mathbb{R}^{2p+m+1}$, allowing us to directly leverage classical convex geometry results. The crux of this proof is to show that $\mathcal{C}$ is a convex set even though (P-SFP) is not a convex program. We summarize this result in the following technical lemma whose proof relies on Lyapunov's convexity theorem [39]:

Lemma 2: Under the assumptions of Theorem 1, the cost-constraints set $\mathcal{C}$ in (10) is a non-empty convex set.

Proof: See Appendix A. ∎

We may then leverage the following result from convex geometry:

Proposition 2 (Supporting hyperplane theorem [40, Prop. 1.5.1]): Let $\mathcal{A} \subset \mathbb{R}^n$ be a nonempty convex set. If $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ is not in the interior of $\mathcal{A}$, then there exists a hyperplane passing through $\tilde{\boldsymbol{x}}$ such that $\mathcal{A}$ is in one of its closed halfspaces, i.e., there exists $\boldsymbol{p} \neq \boldsymbol{0}$ such that $\boldsymbol{p}^T \tilde{\boldsymbol{x}} \leq \boldsymbol{p}^T \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{A}$.

Formally, start by observing that the point $(P^\star, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$ cannot be in the interior of $\mathcal{C}$. Indeed, there would otherwise exist $\delta > 0$ such that $(P^\star - \delta, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}) \in \mathcal{C}$, violating the optimality of P. Proposition 2 therefore implies that there exists a non-zero vector $(\lambda_0, \boldsymbol{\nu}, \boldsymbol{\mu}_R, \boldsymbol{\mu}_I) \in \mathbb{R}^{2p+m+1}$ such that for all $(c, \boldsymbol{u}, \boldsymbol{k}_R, \boldsymbol{k}_I) \in \mathcal{C}$,

$$\lambda_0 c + \boldsymbol{\nu}^T \boldsymbol{u} + \boldsymbol{\mu}_R^T \boldsymbol{k}_R + \boldsymbol{\mu}_I^T \boldsymbol{k}_I \geq \lambda_0 P^\star. \qquad (11)$$

Observe that the vector defining the hyperplane uses the same notation as for the dual variables of (P-SFP) foreshadowing the fact that these hyperplanes span the values of the Lagrangian (3). From (11), we immediately obtain that $\lambda_0 \geq 0$ and $\boldsymbol{\nu} \succeq \boldsymbol{0}$. Indeed, note that $\mathcal{C}$ is unbounded above in its first $m+1$ components, i.e., if $(c, \boldsymbol{u}, \boldsymbol{k}_R, \boldsymbol{k}_I) \in \mathcal{C}$ then $(c', \boldsymbol{u}', \boldsymbol{k}_R, \boldsymbol{k}_I) \in \mathcal{C}$ for any $(c', \boldsymbol{u}') \succeq (c, \boldsymbol{u})$. Hence, if any component of λ_0 or $\boldsymbol{\nu}$ were negative, there would exist a vector in $\mathcal{C}$ that makes the left-hand side of (11) arbitrarily small, eventually violating the inequality. Let us now show that $\lambda_0 \neq 0$.

To do so, suppose $\lambda_0 = 0$. Then, (11) reduces to

$$\boldsymbol{\nu}^T \boldsymbol{u} + \boldsymbol{\mu}_R^T \boldsymbol{k}_R + \boldsymbol{\mu}_I^T \boldsymbol{k}_I \geq 0, \qquad (12)$$

for all $(c, \boldsymbol{u}, \boldsymbol{k}_R, \boldsymbol{k}_I) \in \mathcal{C}$. However, (12) leads to a contradiction because its left-hand side can always be made negative. Indeed, if $[\boldsymbol{\nu}]_i > 0$ for any i, then the hypothesis on the existence of a strictly feasible point for (P-SFP) violates (12). Explicitly, since there exists $(X', \boldsymbol{z}')$ such that $\int_\Omega \boldsymbol{F}[X'(\boldsymbol{\beta}), \boldsymbol{\beta}] d\boldsymbol{\beta} = \boldsymbol{z}'$ and $g_i(\boldsymbol{z}') < 0$ for all $i = 1, \dots, m$, then $(c_0, -\delta \mathbb{1}, \boldsymbol{0}, \boldsymbol{0}) \in \mathcal{C}$ for some c_0 and $\delta > 0$, where $\mathbb{1}$ is a vector of ones. Thus, if $[\boldsymbol{\nu}]_i > 0$ for any i, we obtain $-\delta(\boldsymbol{\nu}^T \mathbb{1}) < 0$, which violates (12).

On the other hand, if $\boldsymbol{\nu} = \boldsymbol{0}$, then (12) reduces to $\boldsymbol{\mu}_R^T \boldsymbol{k}_R + \boldsymbol{\mu}_I^T \boldsymbol{k}_I \geq 0$ which cannot holds because for $(\bar{c}, \bar{\boldsymbol{u}}, -\boldsymbol{\mu}_R, -\boldsymbol{\mu}_I) \in$

C, (12) evaluates to $-\|\boldsymbol{\mu}_R\|^2 - \|\boldsymbol{\mu}_I\|^2 < 0$. To see that this vector is indeed an element of C, simply choose any $\bar{X} \in C$ with $\bar{X}(\beta) \in \mathcal{P}$ a.e. and let $\bar{\boldsymbol{z}} = -\boldsymbol{\mu}_R - j\boldsymbol{\mu}_I - \int_\Omega \boldsymbol{F}[\bar{X}(\boldsymbol{\beta}), \boldsymbol{\beta}]d\boldsymbol{\beta}$, $[\bar{\boldsymbol{u}}]_i = g_i(\bar{\boldsymbol{z}})$, and $\bar{c} = f_0(\bar{X})$. Hence, it must be that $\lambda_0 \neq 0$.

However, for $\lambda_0 \neq 0$, (11)

$$c + \tilde{\boldsymbol{\nu}}^T \boldsymbol{u} + \tilde{\boldsymbol{\mu}}_R^T \boldsymbol{k}_R + \tilde{\boldsymbol{\mu}}_I^T \boldsymbol{k}_I \geq P,$$

where $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu}/\lambda_0$, $\tilde{\boldsymbol{\mu}}_R = \boldsymbol{\mu}_R/\lambda_0$, and $\tilde{\boldsymbol{\mu}}_I = \boldsymbol{\mu}_I/\lambda_0$, which from the definition of C implies that

$$f_0(X) + \sum_{i=1}^m \tilde{\nu}_i g_i(\boldsymbol{z}) + \tilde{\boldsymbol{\mu}}_R^T \mathbb{Re}\left[\int_\Omega \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}\right] \\ + \tilde{\boldsymbol{\mu}}_I^T \mathbb{Im}\left[\int_\Omega \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}\right] \geq P^\star, \quad (13)$$

for any (P-SFP)-feasible pair $(X, \boldsymbol{z})$. Letting $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}_R + j\tilde{\boldsymbol{\mu}}_I$, we recognize that (13) in fact bounds the value of the Lagrangian in (3) for any (P-SFP)-feasible pair $(X, \boldsymbol{z})$, i.e., $\mathcal{L}(X, \boldsymbol{z}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}}) \geq P^\star$. Taking the minimum of the left-hand side of (13) hence implies $D^\star \geq P^\star$, thus concluding the proof. ∎

B. SFPs and L_1-Norm Optimization Problems

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

$$\begin{aligned} \underset{X \in L_\infty, \boldsymbol{z} \in C^p}{\text{minimize}} \quad & \|X\|_{L_q} \\ \text{subject to} \quad & g_i(\boldsymbol{z}) \leq 0 \\ & \boldsymbol{z} = \int_\Omega \boldsymbol{F}[X(\boldsymbol{\beta}), \boldsymbol{\beta}]d\boldsymbol{\beta} \\ & |X| \leq \Gamma \text{ a.e.} \end{aligned} \quad (\text{P}_q)$$

Problem (P$_0$) [i.e., (P$_q$) with $q = 0$] is an instance of (P-SFP) without regularization ($F_0 \equiv 0$) in which $\mathcal{P}$ 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 $x^o(\boldsymbol{\mu}, \boldsymbol{\beta}) = \text{argmin}_{|x| \leq \Gamma} |x| + \mathbb{Re}[\boldsymbol{\mu}^T \boldsymbol{F}(x, \boldsymbol{\beta})]$ saturate, i.e., $x^o(\boldsymbol{\mu}, \boldsymbol{\beta}) \neq 0 \Rightarrow |x^o(\boldsymbol{\mu}, \boldsymbol{\beta})| = \Gamma$ for all $\boldsymbol{\mu} \in \mathbb{C}^p$ and $\boldsymbol{\beta} \in \Omega$. If $P_0^\star$ ($P_1^\star$) is the optimal value of (P$_q$) for $q = 0$ ($q = 1$) and Slater's condition holds, then

$$P_0^\star = \frac{P_1^\star}{\Gamma}.$$

Proof: The proof follows by relating the dual values of (P$_q$) for $q = \{0, 1\}$ and then using strong duality. Start by defining the Lagrangian of (P$_q$) as

$$\mathcal{L}(X, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \|X\|_{L_q} + \sum_i \nu_i g_i(\boldsymbol{z}) \\ + \mathbb{Re}\left[\boldsymbol{\mu}^H\left(\int_\Omega \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}\right)\right]. \quad (14)$$

For $q = 0$, Proposition 1 yields

$$d_0(\boldsymbol{\mu}, \nu) = \int_{\mathcal{S}_0(\boldsymbol{\mu})} \left\{1 + \min_{|x| \leq \Gamma} \mathbb{Re}\left[\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right]\right\} d\boldsymbol{\beta} \\ + w(\boldsymbol{\mu}, \boldsymbol{\nu}). \quad (15)$$

where

$$\mathcal{S}_0(\boldsymbol{\mu}) = \left\{\boldsymbol{\beta} \in \Omega \mid \min_{|x| \leq \Gamma} \mathbb{Re}\left[\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right] < -1\right\} \quad (16)$$

and $w(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{z}} \sum_i \nu_i g_i(\boldsymbol{z}) - \mathbb{Re}[\boldsymbol{\mu}^H \boldsymbol{z}]$. Notice that w is homogeneous, i.e., $w(\alpha\boldsymbol{\mu}, \alpha\boldsymbol{\nu}) = \alpha w(\boldsymbol{\mu}, \boldsymbol{\nu})$ for $\alpha > 0$. Proceeding similarly from (14), the dual function of (P$_1$) is

$$d_1(\boldsymbol{\mu}, \nu) = \int_\Omega \left\{\min_{|x| \leq \Gamma} |x| + \mathbb{Re}\left[\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right]\right\} d\boldsymbol{\beta} \\ + w(\boldsymbol{\mu}, \boldsymbol{\nu}). \quad (17)$$

Using the the saturation hypothesis, the integrand in (17) is nontrivial only over the set

$$\mathcal{S}_1(\boldsymbol{\mu}) = \left\{\boldsymbol{\beta} \in \Omega \mid \min_{|x| \leq \Gamma} \mathbb{Re}\left[\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right] < -\Gamma\right\}. \quad (18)$$

Hence,

$$d_1(\boldsymbol{\mu}, \nu) = \int_{\mathcal{S}_1(\boldsymbol{\mu})} \left\{\Gamma + \min_{|x| \leq \Gamma} \mathbb{Re}\left[\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right]\right\} d\boldsymbol{\beta} \\ + w(\boldsymbol{\mu}, \boldsymbol{\nu}). \quad (19)$$

To proceed, note that the dual functions in (15) and (19) are related by

$$d_0(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{1}{\Gamma} d_1(\Gamma\boldsymbol{\mu}, \Gamma\boldsymbol{\nu}). \quad (20)$$

Indeed, observe from (16) and (18) that $\mathcal{S}_1(\Gamma\boldsymbol{\mu}) = \mathcal{S}_0(\boldsymbol{\mu})$. Thus,

$$\frac{1}{\Gamma} \int_{\mathcal{S}_1(\Gamma\boldsymbol{\mu})} \left\{\Gamma + \min_{|x| \leq \Gamma} \mathbb{Re}\left[\Gamma\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right]\right\} d\boldsymbol{\beta} \\ = \int_{\mathcal{S}_0(\boldsymbol{\mu})} \left\{1 + \min_{|x| \leq \Gamma} \mathbb{Re}\left[\boldsymbol{\mu}^H \boldsymbol{F}(x, \boldsymbol{\beta})\right]\right\} d\boldsymbol{\beta}.$$

The homogeneity of w then yields (20). Immediately, it holds that if $(\boldsymbol{\mu}^o, \boldsymbol{\nu}^o)$ is a maximum of d_0, then $(\Gamma\boldsymbol{\mu}^o, \Gamma\boldsymbol{\nu}^o)$ is a maximum of d_1. To see this is the case, note from (20) that

$$\nabla d_0(\boldsymbol{\mu}^o, \boldsymbol{\nu}^o) = \boldsymbol{0} \Leftrightarrow \nabla d_1(\Gamma\boldsymbol{\mu}^o, \Gamma\boldsymbol{\nu}^o) = \boldsymbol{0},$$

so that $(\Gamma\boldsymbol{\mu}^o, \Gamma\boldsymbol{\nu}^o)$ is a critical point of d_1. Since d_1 is a concave function, $(\Gamma\boldsymbol{\mu}^o, \Gamma\boldsymbol{\nu}^o)$ must be a global maximum.

To conclude, observe that (P$_q$) has zero duality gap for both $q = 0$, due to Theorem 1, and $q = 1$, because it is a convex program. From (20) we then obtain

$$P_0^\star = \max_{\boldsymbol{\mu}, \boldsymbol{\nu} \geq 0} d_0(\boldsymbol{\mu}, \boldsymbol{\nu}) = d_0(\boldsymbol{\mu}^\star, \boldsymbol{\nu}^\star) \\ = \frac{1}{\Gamma} d_1(\Gamma\boldsymbol{\mu}^\star, \Gamma\boldsymbol{\nu}^\star) = \frac{1}{\Gamma} \max_{\boldsymbol{\mu}, \boldsymbol{\nu} \geq 0} d_1(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{P_1^\star}{\Gamma}. \quad ∎$$

Proposition 3 shows that a large class of L_0- and L_1-norm minimization problems found in functional 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
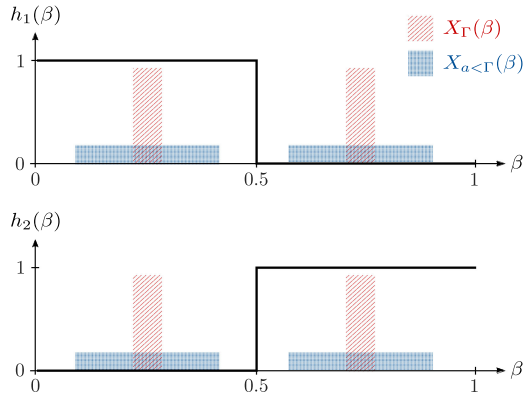
Fig. 1. Illustration of Example 1.

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 for incoherent, linear dictionaries [16], [17]. Still, Proposition 3 does not imply that the solution of the L_0- and L_1-norm problems are the same, as is the case for discrete results. In fact, though they have the same optimal value, (P_1) admits solutions with larger support (see Example 1). Although conditions exist for which the L_1-norm minimization problem with linear dictionaries yields minimum support solutions [24], [26], [27], Theorem 1 precludes the use of this relaxation for both linear and nonlinear dictionaries.

Example 1: Proposition 3 gives an equivalence between L_0- and L_1-norm minimization problems in terms of their optimal values, but not in terms of their solutions. We illustrate this point using the example depicted in Fig. 1. Let $\Omega = [0, 1]$, $g(\boldsymbol{z}) = \|\boldsymbol{y} - \boldsymbol{z}\|_2^2$, $\boldsymbol{y} = [\, y_1 \ y_2 \,]^T$ with $|y_1|, |y_2| < \Gamma/2$, and $\boldsymbol{F}(x, \beta) = \boldsymbol{h}(\beta)x$, where $\boldsymbol{h}(\beta) = [\, h'(\beta) \ 1 - h'(\beta) \,]^T$ with $h'(\beta) = \mathbb{I}(\beta \in [0, 1/2])$. Due to the form of h', it is readily seen that the optimal value of (P_1) is $P_1^\star = |y_1| + |y_2|$.

Now consider the family of functions indexed by $0 < a \leq \Gamma$

$$X_a(\beta) = a\,\mathrm{sign}(y_1)\mathbb{I}(\beta \in \mathcal{A}_1) + a\,\mathrm{sign}(y_2)\mathbb{I}(\beta \in \mathcal{A}_2), \quad (21)$$

where $\mathcal{A}_1 \subseteq [0, 1/2]$ with $\|\mathcal{A}_1\|_{L_0} = |y_1|/a$ and $\mathcal{A}_2 \subseteq [1/2, 1]$ with $\|\mathcal{A}_2\|_{L_0} = |y_2|/a$ (e.g., X_Γ and $X_{a<\Gamma}$ in Fig. 1). For all a, X_a is a solution of (P_1) [it is (P_q)-feasible with value $P_1^\star$]. However, its support is given by $\|X_a\|_{L_0} = (|y_1| + |y_2|)/a$. Thus, (P_1) admits solutions that do not have minimum support ($X_{a<\Gamma}$ in Fig. 1), whereas only X_Γ is a solution of (P_0).

IV. Solving Sparse Functional Programs

Theorem 1 from the previous section establishes duality as a fruitful approach for solving the sparse functional program (P-SFP). Indeed, the strong duality of (P-SFP) implies that

$$(X^\star, \boldsymbol{z}^\star) \in \underset{\substack{X \in \mathcal{X}, \, \boldsymbol{z} \in \mathbb{C}^p, \\ X(\boldsymbol{\beta}) \in \mathcal{P}}}{\mathrm{argmin}} \ \mathcal{L}(X, \boldsymbol{z}, \boldsymbol{\mu}^\star, \boldsymbol{\nu}^\star), \quad (22)$$

for the Lagrangian $\mathcal{L}$ in (3), where $\boldsymbol{\mu}^\star$ and $\boldsymbol{\nu}^\star$ are the solutions of (P-SFP) [37]. When this set is a singleton, the inclusion becomes equality and we recover the unique primal solution $X^\star$. This occurs when the Lagragian (3) has a single minimizer, i.e., when (PII) is a singleton. This is the case, for instance,

---

Algorithm 1: Dual Ascent for SFPs.

$\boldsymbol{\mu}^{(0)} = \boldsymbol{0}$, $\nu_i^{(0)} = 1$
for $t = 1, \ldots, T$ **do**
$\quad X_{t-1}(\boldsymbol{\beta}) = X_d(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\beta})$
$\quad \boldsymbol{z}_{t-1} = \boldsymbol{z}_d(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})$
$\quad \boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \eta_t \left[\displaystyle\int_\Omega \boldsymbol{F}[X_{t-1}(\boldsymbol{\beta}), \boldsymbol{\beta}]d\boldsymbol{\beta} - \boldsymbol{z}_{t-1} \right]$
$\quad \nu_i^{(t)} = \left[\nu_i^{(t-1)} + \eta_t g_i(\boldsymbol{z}_{t-1}) \right]_+$
$\quad P_t = d(\boldsymbol{\mu}^{(t)}, \boldsymbol{\nu}^{(t)})$
end for
$X^\star(\boldsymbol{\beta}) = X_d(\boldsymbol{\mu}^{(t^\star)}, \boldsymbol{\beta})$ for $t^\star \in \underset{1 \leq t \leq T}{\mathrm{argmax}} \, P_t$

---

when $F_0(x, \boldsymbol{\beta}) = |x|^2$, in which case $\mathcal{L}$ is strongly convex in X [37]. Since Proposition 1 allows us to solve (22), all that remains is to address the issue of solving (P-SFP) to obtain $(\boldsymbol{\mu}^\star, \boldsymbol{\nu}^\star)$.

Note that (P-SFP) is a convex program and can therefore be solved using any (stochastic) convex optimization algorithm [37], [41]. For illustration, this section introduces an algorithm based on supergradient ascent. For ease of reference, a step-by-step guide to solving SFPs is presented in Appendix B.

Recall that a supergradient of a concave function $f : \Omega \to \mathbb{R}$ at $\boldsymbol{x} \in \Omega$ is any vector $\boldsymbol{p}$ that satisfies the inequality $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \boldsymbol{p}^T(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{y} \in \Omega$. Although a supergradient may not be an ascent direction at $\boldsymbol{x}$, taking small steps in its direction decreases the distance to any maximizer of f [37].

It is straightforward to show that the constraint slacks in (3) are supergradients of the dual function d with respect to their corresponding dual variables [37]. Explicitly,

$$\boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{\mu}', \boldsymbol{\nu}') = \int_\Omega \boldsymbol{F}\left[X_d(\boldsymbol{\mu}', \boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}_d(\boldsymbol{\mu}', \boldsymbol{\nu}') \quad (23a)$$

$$\boldsymbol{p}_{\nu_i}(\boldsymbol{\mu}', \boldsymbol{\nu}') = g_i\left[\boldsymbol{z}_d(\boldsymbol{\mu}', \boldsymbol{\nu}')\right] \quad (23b)$$

are supergradients of d for the dual minimizers

$$X_d(\boldsymbol{\mu}, \cdot) \in \underset{\substack{X \in \mathcal{X} \\ X(\boldsymbol{\beta}) \in \mathcal{P}}}{\mathrm{argmin}} \int_\Omega \left\{ F_0\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] + \lambda \mathbb{I}\left[X(\boldsymbol{\beta}) \neq 0\right] \right.$$
$$\left. + \mathrm{Re}\left[\boldsymbol{\mu}^H \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right]\right] \right\} d\boldsymbol{\beta}, \quad (24a)$$

$$\boldsymbol{z}_d(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \underset{\boldsymbol{z}}{\mathrm{argmin}} \sum_i \nu_i g_i(\boldsymbol{z}) - \mathrm{Re}[\boldsymbol{\mu}^H \boldsymbol{z}]. \quad (24b)$$

Algorithm 1, with step size $\eta_t > 0$, then yields the optimal dual variables $(\boldsymbol{\mu}^\star, \boldsymbol{\nu}^\star)$ and a solution $X^\star$ of (P-SFP).

Given that the optimization problem in (24b) is convex, there are two hurdles in evaluating (23): (i) obtaining X_d involves solving the non-convex, infinite dimensional problem in (24a) and (ii) the integral in (23a) may not have an explicit form or this form is too cumbersome to be useful in practice. We have already argued that despite its non-convexity, the minimization in (24a) is tractable by exploiting separability (see Proposition 1). The resulting scalar problem often has a closed-form solution (see

Section V for examples) or can be tackled using global optimization techniques [38]. Note that though this approach does not explicitly yield the function X_d, it allows $X_d(\boldsymbol{\mu}, \boldsymbol{\beta})$ to be evaluated for any $\boldsymbol{\beta} \in \Omega$ using (PII). This is enough to numerically compute the integral in (23a). This integral [(ii)] may either be approximated numerically or done without by leveraging stochastic optimization techniques. Step-by-step descriptions of both methods are presented in Appendix B.

In the first case, we effectively solve a perturbed version of (P-SFP) and the difference between the optimal value of the original problem and that obtained numerically depends linearly on the precision of the integral computation under mild technical conditions:

Proposition 4: Suppose that
  i) the perturbation function of (P-SFP) is differentiable around the origin;
 ii) $\int_\Omega F_0(0, \boldsymbol{\beta}) d\boldsymbol{\beta} = 0$ and $\int_\Omega \boldsymbol{F}(0, \boldsymbol{\beta}) d\boldsymbol{\beta} = \boldsymbol{0}$;
iii) there exists $\alpha > 0$ such that $g_i(\alpha \mathbb{1}), g_i(-\alpha \mathbb{1}) < \infty$; and
iv) there exists a strictly feasible pair $(X^\dagger, \boldsymbol{z}^\dagger)$ (Slater's condition) for (P-SFP) such that $g_i(\boldsymbol{z}^\dagger) < -\epsilon$, for $\epsilon > 0$, and $\bar{F}_0 = \int_\Omega F_0(X^\dagger(\boldsymbol{\beta}), \boldsymbol{\beta}) d\boldsymbol{\beta} < \infty$.

If $P^\star$ is the optimal value of (P-SFP) and $P_\delta^\star$ is the value of the solution obtained by Algorithm 1 when evaluating the integral in (23a) with approximation error $0 < \delta \ll 1$, then $|P^\star - P_\delta^\star| \leq \mathcal{O}(\delta)$.

Proof: See Appendix C. ∎

In the second case, the integral in (23a) is approximated using Monte Carlo integration, i.e., by drawing a set of $\boldsymbol{\beta}_j$ independently and uniformly at random from Ω and taking

$$\hat{\boldsymbol{p}}_{\boldsymbol{\mu}} = \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{F} \left[X_d(\boldsymbol{\mu}', \boldsymbol{\beta}_j), \boldsymbol{\beta}_j \right] - \boldsymbol{z}_d(\boldsymbol{\mu}', \boldsymbol{\nu}'). \qquad (25)$$

Since Monte Carlo integration is an unbiased estimators, $\hat{\boldsymbol{p}}_{\boldsymbol{\mu}}$ is an unbiased estimate of $\boldsymbol{p}_{\boldsymbol{\mu}}$. Taking $N = 1$ in (25) is akin to performing stochastic (super)gradient ascent on the dual function d. For $N > 1$, we obtain a mini-batch type algorithm. Typical convergence guarantees hold in both cases [37], [42], [43].

Algorithm 1, though effective, may converge slowly depending on the numerical properties of the problem. Faster, problem independent convergence rates can be obtained using, for instance, second-order methods or by exploiting specific structures of SFP instances. Investigating the use and fit of these approaches to solving (P-SFP) is, however, beyond the scope of this paper.

V. APPLICATIONS

So far, we have focused on whether SFPs are tractable. In this section, we illustrate their expressiveness by using (P-SFP) to cast the problems of nonlinear spectral estimation and robust functional data classification.

A. Nonlinear Line Spectrum Estimation

The first example application of SFPs is in the context of continuous, possibly nonlinear, sparse dictionary recovery/ denoising problems. Formally, let $\boldsymbol{y} \in \mathbb{C}^p$ collect samples y_i,

$i = 1, \ldots, p$, of a signal. Our goal is to represent $\boldsymbol{y}$ using as few atoms as possible from the nonlinear dictionary

$$\mathcal{D} = \{ \boldsymbol{F}(\cdot, \boldsymbol{\beta}) : \mathbb{C} \to \mathbb{C}^p \mid \boldsymbol{\beta} \in \Omega \}. \qquad (26)$$

Explicitly, we wish to find $\{(x_k, \boldsymbol{\beta}_k)\}$ such that

$$\hat{\boldsymbol{y}} = \sum_{k=1}^{K} \boldsymbol{F}(x_k, \boldsymbol{\beta}_k) \qquad (27)$$

is close to $\boldsymbol{y}$ for some small K. Notice that, in contrast to classical dictionary recovery, the relation between the coefficients x_k and the signal $\hat{\boldsymbol{y}}$ is not necessarily linear. Moreover, Ω is an uncountable set, so that we select from a continuum of atoms as opposed to the discrete, finite case.

To make the discussion concrete, consider the problem of estimating the parameters of a small number of saturated sinusoids from samples of their superposition. This problem is found in several signal processing applications, such as telecommunication and direction of arrival (DOA) estimation, where nonlinear behaviors are common due to hardware limitations of the sources. Formally, we wish to estimate the frequencies, amplitudes, and phases of K sinusoids from the set of noisy samples

$$y_i = \sum_{k=1}^{K} \rho \left[a_k \cos(2\pi f_k t_i) \right] + n_i, \quad \text{for } i = 1, \ldots, p, \qquad (28)$$

where $f_k \in [0, 1/2]$ is the frequency and $a_k \in \mathbb{R}$ is the amplitude/phase of the k-th component; t_i is the fixed, known sampling time of the i-th sample; $\{n_i\}$ are independent and identically distributed (i.i.d.) zero-mean random variables with variance $\mathbb{E}n_i^2 = \sigma_n^2$ representing the measurement noise; and ρ is a function that models the source nonlinearity with $\rho(0) = 0$.

To pose this estimation problem as an SFP, we need an approximate continuous representation of the signal model in (28). We say approximate because the nonlinearity ρ may prevent us from finding a measurable function X such that $\int \rho[X(\varphi) \cos(2\pi\varphi t_i)] d\varphi = \rho[x \cos(2\pi f t_i)]$ for a fixed amplitude-frequency pair (x, f). Even if ρ allows it, an exact representation may involve Dirac deltas, which violates a hypothesis of Theorem 1 and prevents us from efficiently finding a solution of (P-SFP). The following proposition introduces a functional signal model that approximates (28) arbitrarily well using parameters in L_2.

Proposition 5: For fixed $a, t \in \mathbb{R}$, $f \in [0, 1/2]$, define the hyperparameter $B \in \mathbb{R}_+$ and let

$$r(B) = B \int_0^{\frac{1}{2}} \rho \left[X'(\varphi) \cos(2\pi\varphi t) \right] d\varphi. \qquad (29)$$

If $X'(\varphi) = a$ for $\varphi \in [f - B^{-1}, f + B^{-1}]$ and zero everywhere else, then $r(B) \to \rho[a \cos(2\pi f t)]$ as $B \to \infty$.

Proof: Note that (29) is equivalent to

$$r(B) = \int_0^{\frac{1}{2}} B \cdot \Pi_{f, B^{-1}}(\varphi) \rho \left[a \cos(2\pi\varphi t) \right] d\varphi$$

with $\Pi_{f,b}(\varphi) = \mathbb{I}(\varphi \in [f - b, f + b])$. The result then follows from the fact that $B \cdot \Pi_{f,B^{-1}}(\varphi)$ converges weakly to $\delta(\varphi - f)$ as $B \to \infty$, where δ is the Dirac's delta [44]. ∎

Proposition 5 allows us to cast nonlinear line spectrum estimation as the SFP

$$\underset{X \in L_2}{\text{minimize}} \quad \|X\|_{L_2} + \lambda \|X\|_{L_0}$$

$$\text{subject to} \quad \sum_{i=1}^{p} (y_i - \hat{y}_i)^2 \leq \epsilon \qquad \text{(PIII)}$$

$$\hat{y}_i = B \int_0^{\frac{1}{2}} \rho \left[X(\varphi) \cos(2\pi \varphi t_i) \right] d\varphi,$$

$$\text{for } i = 1, \ldots, p,$$

where $B > 0$ is an approximation parameter and $\epsilon > 0$ determines the solution fit. Problem (PIII) explicitly seeks the sparsest function X that fits the observations given the model in (28). The L_2-norm regularization improves robustness to noise as well as the numerical properties of the dual by adding shrinkage. Note that since $X \in L_2$, the solution $X^\star$ of (PIII) does not contain atoms and is instead a superposition of bump functions around the component frequencies f_k (see, e.g., Fig. 4). As Proposition 5 suggests, the height and width of each bump depends on the amplitude of the sinusoidal component and the choice of B. Thus, the parameter a_k from (28) can be estimated using

$$\hat{a}_k = B \int_{\mathcal{B}_k} X^\star(\varphi) d\varphi, \qquad (30)$$

where $X^\star$ is a solution of (PIII) and $\mathcal{B}_k \subset [0, 1/2]$ contains a single bump. The parameter f_k can then be estimated using the center frequency of the bump. Naturally, B should be as large as possible so that (29) is a good approximation of (28), improving the parameter estimates. Choosing B too large, however, degrades the numerical properties of the dual problem, making it harder to solve in practice. Similar trade-offs are found several methods when tuning regularization parameters, for instance, elastic net [45], [46].

Since ρ is an arbitrary function, a particular case of (PIII) performs spectral estimation with linear sources, i.e., when $\rho(z) = z$ in (28) and there is no saturation. The dual function is straightforward to evaluate in this case since the optimization problem (7) from Proposition 1 becomes a quadratic program that admits a closed-form solution. However, a myriad of classical methods such as MUSIC or atomic soft thresholding (AST) have been proposed for the linear case. MUSIC performs line spectrum estimation using the eigendecomposition of the empirical autocorrelation matrix of the measurements y_i [3]. Nevertheless, it can only be used in single snapshot applications when the signal is sampled regularly—see [3] for details—and requires that the number K of components be known *a priori*. The AST approach, on the other hand, is based on an atomic norm relaxation of the sparse estimation problem and leverages duality and spectral properties of Toeplitz matrices to preclude discretization [19], [20]. Both methods first obtain the component frequencies and then determine amplitudes and phases using
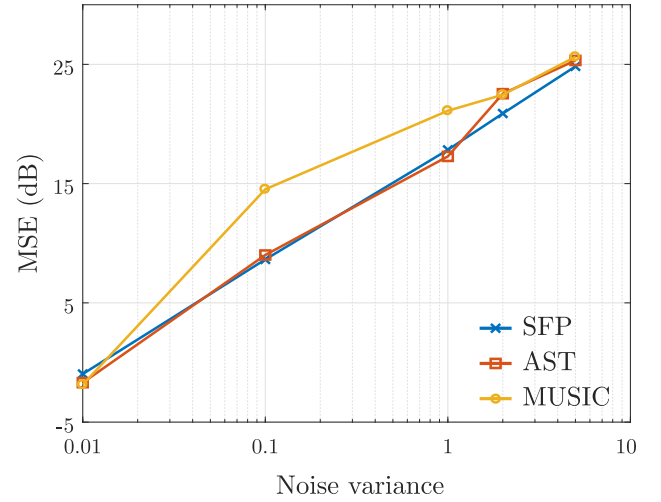

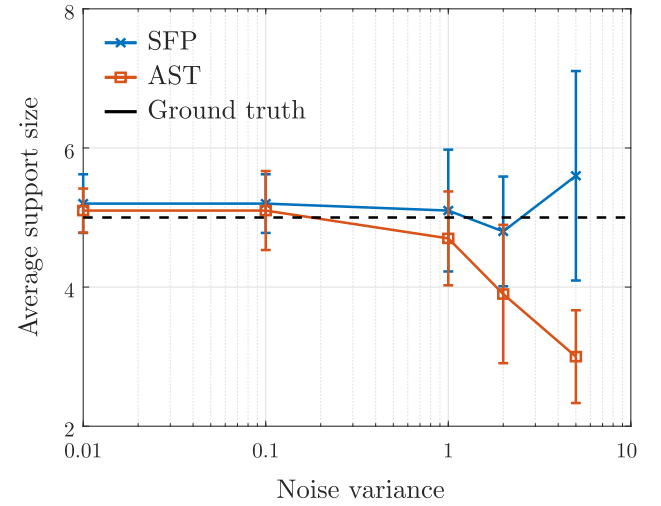Fig. 2. Reconstruction MSE for line spectral estimation of linear sources.

Fig. 3. Support size estimation for line spectral estimation of linear sources.

least squares. These different approaches are compared in Figs. 2 and 3.

These plots display the average performance over 10 realizations that used $p = 61$ samples ($t_i = -30, \ldots, 30$) of the superposition of $K = 5$ components whose the frequencies f_k were drawn uniformly at random with a minimum spacing of $4/p$ and whose amplitudes a_k were taken randomly and independently from $[0.5, 3]$. Problem (PIII) was solved using the approximate supergradient method described in Appendix B with $B = 1$, $\lambda = 5000$ for all noise levels expect $\sigma_n^2 = 5$ which used $\lambda = 6000$, and $\epsilon = p\sigma_n^2$. For the AST method, we used the optimal regularization from [20] which depends on σ_n^2. In all cases, the reconstruction MSE is evaluated as

$$\text{MSE} = \sum_{i=1}^{p} (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ denotes the samples reconstructed based on the K components with largest magnitudes obtained by each algorithm. For AST, the support is obtained from the peaks of the trigonometric polynomial defined by the dual as in [19], [20] and for SFP, from
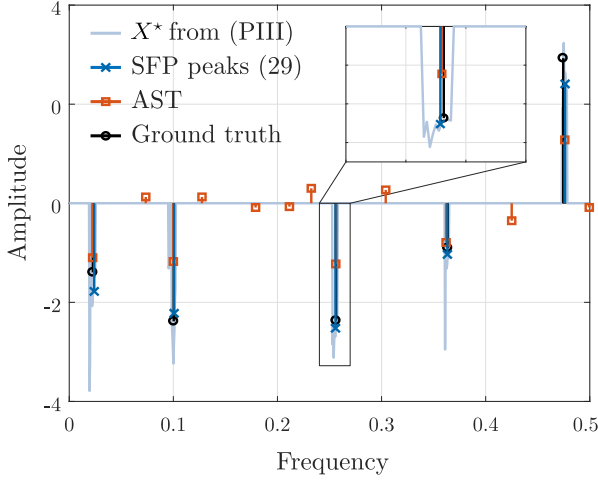
Fig. 4. Solutions obtained for line spectral estimation of saturated sources.

Fig. 5. Reconstruction MSE for line spectral estimation of saturated sources.

the center of the bumps in the solution $X^\star$ of (PIII) (as illustrated in Fig. 4).

In high SNR scenarios, all methods display similar performance. As the level of noise increases, however, the advantages of explicitly minimizing the L_0-norm instead of its convex surrogate become clearer, especially with respect to support identification. Observe in Fig. 3 that as σ_n^2 increases the number of components obtained from AST decreases considerably, despite using the optimal regularization parameter. Finally, it is worth noting that although the performances are similar, AST involves solving a semidefinite program (SDP), which becomes infeasible in practice as the number of samples p grows and has motivated the study of dimensionality reduction techniques and sampling patterns [47]. On the other hand, efficient solvers based on coordinate ascent can be leveraged to solve large-scale SFPs [37], [41].

Still, the signal reconstruction performance is similar across methods in the linear case (Fig. 2). This is not surprising given the close relation between L_0- and L_1-norm minimization (Theorem 3). In contrast, when the signals are distorted by a nonlinearity, the linear measurement model of AST and MUSIC tends to underestimate the amplitude of the components (Fig. 4). Though greedy approaches to atomic norm minimization are able to deal with nonlinear dictionaries, optimally selecting single atoms from these infinite dimensional dictionary is challenging. Exhaustive, grid-based heuristics have been proposed for specific problems without guarantees [48].

To illustrate this effect, consider the hard saturation

$$\rho(x) = \begin{cases} x, & |x| \leq r \\ r \cdot \text{sign}(x), & \text{otherwise} \end{cases}, \quad (31)$$

where $r > 0$ defines the saturation level. Though computing the dual function may seem challenging in this case due to the nonlinearity, it turns out to be tractable due to the scalar nature of the problem. Indeed, we obtain the dual minimizer from Proposition 1 by evaluating

$$\gamma^o(\boldsymbol{\mu}, \varphi) = \min_{x \in \mathbb{R}} \; x^2 + \boldsymbol{\mu}^T \rho\left[x\boldsymbol{h}(\varphi)\right],$$
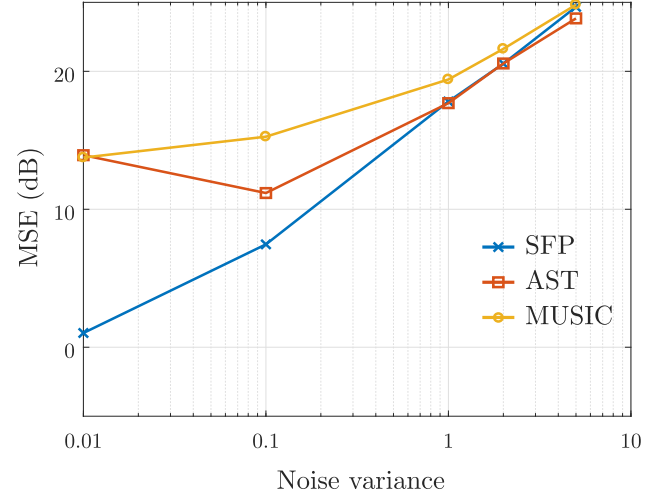
where $[\boldsymbol{h}]_i = \cos(2\pi\varphi t_i)$ for $i = 1, \ldots, p$ and the function ρ applies element-wise. Since we can determine *a priori* which of the elements will saturate, solving this non-convex problem actually reduces to finding the minimum value of p quadratic problems. Namely, assume that $\boldsymbol{h}$ is sorted such that $h_1 \leq \cdots \leq h_p$ and let $\boldsymbol{w}_i(x) = [h_1 x \; \cdots \; h_i x \; r \; \cdots \; r]^T$, where r is the saturation level from (31). For conciseness, we omit the dependence on φ. Then, $\gamma^o(\boldsymbol{\mu}) = \min_{1 \leq i \leq p} \gamma_i^o(\boldsymbol{\mu})$ for

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

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

Fig. 4 shows the solutions obtained using (PIII) and AST for ρ as in (31) with $r = 1$. We omit the results for MUSIC in this plot as its performance is similar to AST. Notice that since (PIII) takes the the nonlinear nature of the signal into account it provides more precise parameter estimates. This is evident in Fig. 5, which shows that (PIII) leads to lower reconstruction errors, especially in higher SNRs. This is expected since neither AST nor MUSIC take the nonlinear effects into account. Yet, as the noise increases and begins to dominate over mismodeling, the performance of all methods becomes similar. This effect is more pronounced here than in the linear case because the saturation limits the energy of the signal leading to even lower effective SNRs. For instance, the average SNR for $\sigma_n^2 = 2$ in Fig. 2 is 6.6 dB, whereas in Fig. 5, it is 2.05 dB.

In these experiments, the signal samples were constructed as in the linear case, but we used for (PIII) $B = 200$, $\epsilon = p\sigma_n^2$, and $\lambda = 100$ for all noise levels expect $\sigma_n^2 \in \{2, 5\}$ which used $\lambda = 80$. For the AST method, we again used the optimal regularization parameter from [20]. Better results could not be obtained by hand-tuning the regularization.

B. Robust Functional Data Analysis

Functional data analysis extends classical statistical methods to data supported on continuous domains. Since it copes with

non-uniformly sampled data and precludes registration, this tool set is especially appropriated for analyzing time series without assuming generative models, such as AR or ARMAX [49]. For concreteness, consider the functional extension of logistic regression: given a data pair (y_i, Z_i) with label $y_i \in \{0, 1\}$ and independent variable $Z_i : [0, 1] \to \mathbb{R}$, the probability that y_i is positive is modeled as

$$\Pr[y_i = 1] = \frac{1}{1 + \exp\left(-\int_0^1 Z_i(\tau)W(\tau)d\tau + b\right)}, \quad (32)$$

where $W : [0, 1] \to \mathbb{R}$ is the functional classifier parameter and b is the intercept. Although the domain of Z_i and W can be an arbitrary compact set, we use the normalized [0,1] for simplicity. Typically, some smoothness prior is assumed for W so that the statistical problem is well-posed, e.g., by using splines or imposing that W has small RKHS norm [49]. Observe that if we replace $\int_0^1 Z_i(\tau)W(\tau)d\tau$ by $\boldsymbol{w}^T\boldsymbol{z}_i$, for $\boldsymbol{w}, \boldsymbol{z}_i \in \mathbb{R}^m$, we recover the classical, finite dimensional logistic model.

As is the case with traditional (discrete) logistic regression, the classifier in (32) is sensitive to outliers. In fact, it has been shown recently that any classifier trained by minimizing a convex loss, as is the case of logistic regression or support vector machines (SVM), suffers from this issue [50], [51]. Although sparsity has been used to mitigate this drawback using convex surrogates such as the ℓ_1-norm [52], [53], these methods remain susceptible to extreme data points caused by impulsive noise or other measurement errors [51].

One approach to addressing this weakness is replacing the inner product in (32) by a robust version that reduces the influence of these extreme samples. In [50], [51], this is done by computing inner products over a subset of the data. Here, however, since (P-SFP) allows us to consider arbitrary nonlinearities in the data model, we can explicitly limit the influence of any sample by saturating the inner product in (32). Explicitly,

$$\Pr[y_i = 1] = \frac{1}{1 + \exp\left(-\int_0^1 \rho\left[Z_i(\tau)W(\tau)\right]d\tau + b\right)}, \quad (33)$$

where ρ is the saturation from (31). Notice that (33) controls the influence of any data point by using the threshold r from the saturation (31). In fact, notice that due to the saturation, the value of the inner product in (33) lies in the range $[-r, r]$. Using the negative log likelihood expression for logistic regression [45], we then formulate the following SFP for learning the robust classifier

$$\begin{aligned}
\underset{X \in L_2}{\text{minimize}} \quad & \|W\|_{L_2} + b^2 + \lambda\|W\|_{L_0} \\
\text{subject to} \quad & -\sum_{i=1}^p \log\left[1 + \exp\left((1 - 2y_i)\hat{y}_i\right)\right] \le \epsilon \\
& \hat{y}_i = \int_{\mathcal{T}} \rho\left[Z_i(\beta)W(\beta)\right]d\beta + b, \\
& \text{for } i = 1, \ldots, p,
\end{aligned} \quad \text{(PIV)}$$

for some fit parameter $\epsilon > 0$. Notice that (PIV) also allows us to fit sparse functional coefficient W by setting $\lambda > 0$. Moreover,
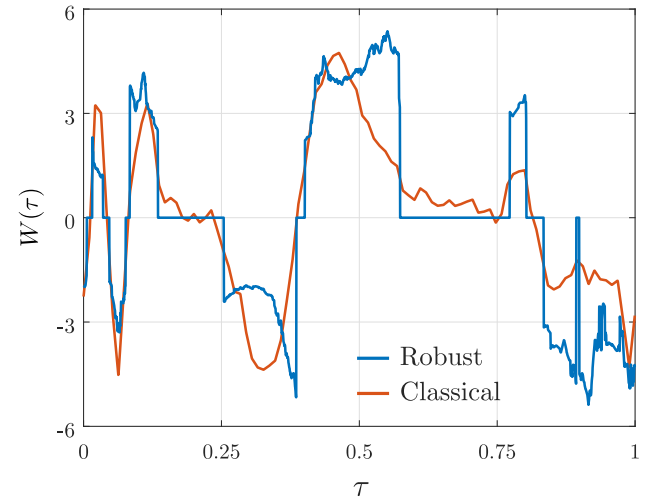

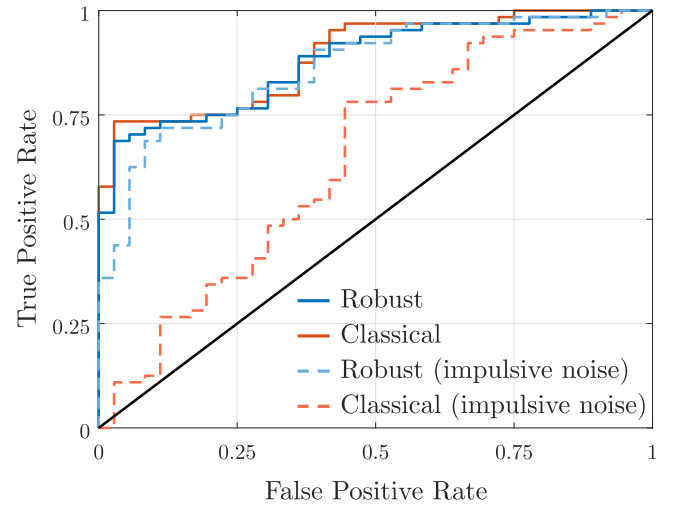Fig. 6.   Solution of functional logistic regression for ECG classification.

Fig. 7.   Receiver operating characteristic (ROC) curve for logistic classifiers in the presence of impulsive noise.

although it is written in terms of the logistic likelihood, other convex criteria such as the hinge loss could be used to obtain robust SVMs.

To illustrate the performance of the robust classifier (33), we consider the problem of identifying whether an electrocardiogram (ECG) signal comes from a healthy heart or one that suffered a myocardial infarction, i.e., a heart attack. The continuous time series Z_i are obtained by linearly interpolating a single heartbeat (see examples in Figs. 8 and 9). Other techniques, such as sinc or spline interpolation are also commonly used in functional data analysis [49]. The labels y_i indicate whether a heart is healthy (1) or not (0). The samples used in the following experiments were taken from the ECG200 dataset [54], [55], which draws from the MIT-BIH Supraventricular Arrhythmia Database [56]. To train the classical functional logistic classifier in (32), we solved (PIV) with $\lambda = 0$ and $r \to \infty$, i.e., no sparsity regularization and no saturation of the inner product. For the robust version, we used $\lambda = 10$ and $r = 4$. In both cases,
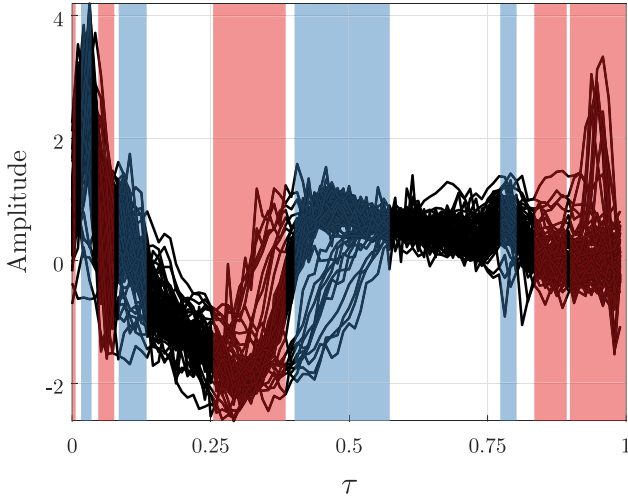
Fig. 8.   ECG of healthy heart and sparse functional coefficients (positive coefficients: blue, negative coefficients: red).
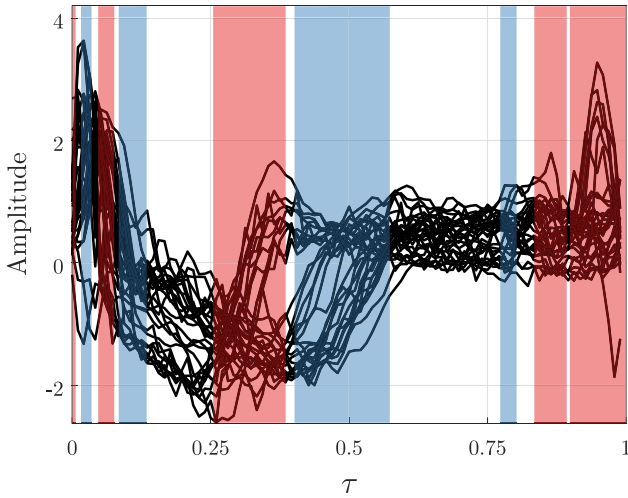

Fig. 9.   ECG of heart with myocardial infraction and sparse functional weights (positive coefficients: blue, negative coefficients: red).

the classifier was fitted with $\epsilon = -46$ using the approximate supergradient method described in Appendix B.

Notice in Fig. 6 that the value of the coefficients of the classical and robust classifiers are similar, leading to comparable performance on both training and test sets (approximately 80% accuracy). The receiver operating characteristic (ROC) curve of both classifiers on the test set is displayed in solid lines in Fig. 7. The robustness of these classifiers to outliers, on the other hand, is considerably different. To illustrate this behavior, corruption by impulsive noise was simulated by randomly adding ±20 to a random subset of 10% of the samples from each heartbeat in the test set. The resulting ROC curves are shown in dashed lines. Although the performance of the linear logistic classifier has now degraded (the test accuracy dropped to 66%), the ROC of the robust version remains unaltered due to the nonlinearity ρ in (33) limiting the effect of the corruption (test accuracy of 76%).

Additionally, the sparsity of the robust classifier parameters improves interpretability by focusing on the portions of the signal that differentiate between normal and abnormal heartbeats (Figs. 8 and 9). For instance, healthy heart signals tend to have negative values for $\tau \in [0.25, 0.4]$ and positive values for $\tau \in [0.4, 0.6]$, whereas hearts that suffered myocardial infarctions do not. On the other hand, there is no discriminant information for $\tau \in [0.6, 0.75]$ and, perhaps less intuitively, between 0.15 and 0.25.

VI. CONCLUSION

We proposed to tackle nonlinear, continuous problems involving sparsity penalties directly by solving sparse functional optimization problems. To do so, we showed that a large class of these mathematical programs have no duality gap and can therefore be solved by means of their dual problems. Duality simultaneously bypasses the infinite dimensionality and non-convexity hurdles of the original problem and enables the use of efficient algorithms to solve these non-convex functional programs. Signal processing applications (nonlinear line spectral estimation and robust functional logistic regression) were used to illustrate the expressiveness of this technique, that we foresee can be used to solve a wide variety of problems in different domains. Future work includes investigating second-order stochastic optimization algorithms to improve the convergence rate of Algorithm 1 and obtaining identifiability/recovery results for problems such as line spectral estimation. We also believe these strong duality results apply to problems beyond sparsity.

APPENDIX A
PROOF OF LEMMA 2

Let $(c, \boldsymbol{u}, \boldsymbol{k}_R, \boldsymbol{k}_I), (c', \boldsymbol{u}', \boldsymbol{k}'_R, \boldsymbol{k}'_I)$ be arbitrary points in $\mathcal{C}$ achieved for $(X, \boldsymbol{z}), (X', \boldsymbol{z}') \in \mathcal{X} \times \mathbb{C}^p$. In other words, it holds that $X(\boldsymbol{\beta}), X'(\boldsymbol{\beta}) \in \mathcal{P}$ a.e., $f_0(X) \leq c$, $f_0(X') \leq c'$, $g_i(\boldsymbol{z}) \leq [\boldsymbol{u}]_i$, $g_i(\boldsymbol{z}') \leq [\boldsymbol{u}']_i$,

$$\int_\Omega \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z} = \boldsymbol{k}_R + j\boldsymbol{k}_I \triangleq \boldsymbol{k},$$

$$\int_\Omega \boldsymbol{F}\left[X'(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}' = \boldsymbol{k}'_R + j\boldsymbol{k}'_I \triangleq \boldsymbol{k}',$$

where we have defined the shorthands $\boldsymbol{k}$ and $\boldsymbol{k}'$ for conciseness. To show that $\mathcal{C}$ is convex, it suffices to show that $\theta(c, \boldsymbol{u}, \boldsymbol{k}_R, \boldsymbol{k}_I) + (1 - \theta)(c', \boldsymbol{u}', \boldsymbol{k}'_R, \boldsymbol{k}'_I) \in \mathcal{C}$ for any $\theta \in [0, 1]$. Equivalently, we must obtain $(X_\theta, \boldsymbol{z}_\theta) \in \mathcal{X} \times \mathbb{C}^p$ such that $X_\theta(\boldsymbol{\beta}) \in \mathcal{P}$ a.e.,

$$f_0(X_\theta) \leq \theta c + (1 - \theta)c', \tag{34a}$$

$$g_i(\boldsymbol{z}_\theta) \leq [\theta \boldsymbol{u} + (1 - \theta)\boldsymbol{u}']_i, \tag{34b}$$

$$\int_\Omega \boldsymbol{F}\left[X_\theta(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}_\theta = \theta\boldsymbol{k} + (1 - \theta)\boldsymbol{k}', \tag{34c}$$

for any $0 \leq \theta \leq 1$. To do so, we will rely on the following classical theorem about the range of non-atomic vector measures:

Theorem 2 (Lyapunov's convexity theorem [39]): Let $\mathfrak{v} : \mathcal{B} \to \mathbb{C}^n$ be a vector measure over the measurable space $(\Omega, \mathcal{B})$. If $\mathfrak{v}$ is

non-atomic, then its range is convex, i.e., the set $\{\mathfrak{v}(\mathcal{A}) : \mathcal{A} \in \mathcal{B}\}$ is a convex set.

To see how Theorem 2 allows us to construct the desired X_θ, start by defining a $2(p+1) \times 1$ vector measure $\mathfrak{p}$ over $(\Omega, \mathcal{B})$ such that for every set $\mathcal{Z} \in \mathcal{B}$ we have

$$
\mathfrak{p}(\mathcal{Z}) = \begin{bmatrix} \int_{\mathcal{Z}} \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} \\ \int_{\mathcal{Z}} \boldsymbol{F}\left[X'(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} \\ \int_{\mathcal{Z}} \left[F_0\left(X(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + \lambda \mathbb{I}\left(X(\boldsymbol{\beta}) \neq 0\right)\right] d\boldsymbol{\beta} \\ \int_{\mathcal{Z}} \left[F_0\left(X'(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + \lambda \mathbb{I}\left(X'(\boldsymbol{\beta}) \neq 0\right)\right] d\boldsymbol{\beta} \end{bmatrix}. \quad (35)
$$

Notice that $\mathfrak{p}$ is a proper vector measure, so that $\mathfrak{p}(\emptyset) = \boldsymbol{0}$. Also, observe that evaluating $\mathfrak{p}$ on the whole space Ω yields

$$
\mathfrak{p}(\Omega) = \begin{bmatrix} \int_{\Omega} \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} \\ \int_{\Omega} \boldsymbol{F}\left[X'(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} \\ f_0(X) \\ f_0(X') \end{bmatrix} \Rightarrow \mathfrak{p}(\Omega) = \begin{bmatrix} \boldsymbol{k} + \boldsymbol{z} \\ \boldsymbol{k}' + \boldsymbol{z}' \\ f_0(X) \\ f_0(X') \end{bmatrix}. \quad (36)
$$

Finally, observe that since F_0 and $\boldsymbol{F}$ do not contain Dirac deltas, they induce non-atomic measures. Consequently, $\mathfrak{p}$ is non-atomic.

To proceed, use Theorem 2 to find a set $\mathcal{T}_\theta \in \mathcal{B}$ such that

$$
\mathfrak{p}(\mathcal{T}_\theta) = \theta \mathfrak{p}(\Omega) + (1 - \theta)\mathfrak{p}(\emptyset) = \theta \mathfrak{p}(\Omega) \quad (37)
$$

for $\theta \in [0, 1]$. 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

$$
\mathfrak{p}(\Omega \setminus \mathcal{T}_\theta) = (1 - \theta)\mathfrak{p}(\Omega). \quad (38)
$$

From (37) and (38), construct X_θ as

$$
X_\theta(\boldsymbol{\beta}) = \begin{cases} X(\boldsymbol{\beta}), & \text{for } \boldsymbol{\beta} \in \mathcal{T}_\theta \\ X'(\boldsymbol{\beta}), & \text{for } \boldsymbol{\beta} \in \Omega \setminus \mathcal{T}_\theta \end{cases} \quad (39)
$$

and let $\boldsymbol{z}_\theta = \theta \boldsymbol{z} + (1 - \theta)\boldsymbol{z}'$. We claim that this pair satisfies (34). It is straightforward from the fact that $\mathcal{X}$ is decomposable and that $\mathcal{P}$ is a pointwise constraint, that $X_\theta \in \mathcal{X}$ and $X_\theta(\boldsymbol{\beta}) \in \mathcal{P}$ a.e.

Let us start by showing that X_θ satisfies (34a). Evaluating f_0 at X_θ yields

$$
f_0(X_\theta) = \int_{\Omega} \left[F_0\left(X_\theta(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + \lambda \mathbb{I}\left(X_\theta(\boldsymbol{\beta}) \neq 0\right)\right] d\boldsymbol{\beta}
$$

$$
= \int_{\mathcal{T}_\theta} \left[F_0\left(X(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + \lambda \mathbb{I}\left(X(\boldsymbol{\beta}) \neq 0\right)\right] d\boldsymbol{\beta}
$$

$$
+ \int_{\Omega \setminus \mathcal{T}_\theta} \left[F_0\left(X'(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + \lambda \mathbb{I}\left(X'(\boldsymbol{\beta}) \neq 0\right)\right] d\boldsymbol{\beta}.
$$

From (35), we can write these terms using the last two rows of the vector measure $\mathfrak{p}$ as

$$
f_0(X_\theta) = [\mathfrak{p}(\mathcal{T}_\theta)]_{2p+1} + [\mathfrak{p}(\Omega \setminus \mathcal{T}_\theta)]_{2p+2}. \quad (40)
$$

Then, using (37) and (38) we obtain that

$$
f_0(X_\theta) = [\theta \mathfrak{p}(\Omega)]_{2p+1} + [(1 - \theta)\mathfrak{p}(\Omega)]_{2p+2}
$$

$$
= \theta f_0(X) + (1 - \theta)f_0(X') \leq \theta c + (1 - \theta)c'.
$$

To proceed, notice that since the g_i are convex functions, (34b) obtains immediately. Explicitly,

$$
g_i(\boldsymbol{z}_\theta) = g_i(\theta \boldsymbol{z} + (1 - \theta)\boldsymbol{z}') \leq \theta g_i(\boldsymbol{z}) + (1 - \theta)g_i(\boldsymbol{z}')
$$

$$
\leq [\theta \boldsymbol{u} + (1 - \theta)\boldsymbol{u}']_i.
$$

Finally, we can use the same machinery as in (40) to obtain (34c). Indeed,

$$
\int_{\Omega} \boldsymbol{F}\left[X_\theta(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} = \int_{\mathcal{T}_\theta} \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta}
$$

$$
+ \int_{\Omega \setminus \mathcal{T}_\theta} \boldsymbol{F}\left[X'(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta}
$$

$$
= [\mathfrak{p}(\mathcal{T}_\theta)]_{\mathcal{S}_1} + [\mathfrak{p}(\Omega \setminus \mathcal{T}_\theta)]_{\mathcal{S}_2}
$$

$$
= [\theta \mathfrak{p}(\Omega)]_{\mathcal{S}_1} + [(1 - \theta)\mathfrak{p}(\Omega)]_{\mathcal{S}_2}
$$

$$
= \theta \boldsymbol{k} + (1 - \theta)\boldsymbol{k}' + \boldsymbol{z}_\theta,
$$

where $\mathcal{S}_1 = \{1, \ldots, p\}$ and $\mathcal{S}_2 = \{p+1, \ldots, 2p\}$ select rows 1 through p and $p+1$ through $2p$, respectively, of the vector measure $\mathfrak{p}$.

To conclude, since there exists a pair $(X_\theta, \boldsymbol{z}_\theta) \in \mathcal{X} \times \mathbb{C}^p$ with $X_\theta(\boldsymbol{\beta}) \in \mathcal{P}$ a.e. and such that (34) holds for any $\theta \in [0, 1]$ and $(c, \boldsymbol{u}, \boldsymbol{k}_R, \boldsymbol{k}_I), (c', \boldsymbol{u}', \boldsymbol{k}'_R, \boldsymbol{k}'_I) \in \mathcal{C}$, the set $\mathcal{C}$ is convex. Moreover, the strictly feasible pair $(X', \boldsymbol{z}')$ from the hypotheses implies that $\mathcal{C}$ cannot be empty.

APPENDIX B
A STEP-BY-STEP GUIDE TO SOLVING SFPS

Start with a problem of the form (P-SFP). Initialize $\boldsymbol{\mu}_0$ and $\nu_{i,0} > 0$; compute $d_{\boldsymbol{z},0} = \min_{\boldsymbol{z}} \sum_i \nu_{i,0} g_i(\boldsymbol{z}) - \mathbb{R}\mathrm{e}[\boldsymbol{\mu}_0^H \boldsymbol{z}]$ and let $\boldsymbol{z}_0$ be its minimizer; evaluate

$$
\gamma_0^o(\boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathbb{R}\mathrm{e}\left[\boldsymbol{\mu}_0^H \boldsymbol{F}(x, \boldsymbol{\beta})\right] \quad (41)
$$

and let $\bar{X}_0(\boldsymbol{\beta})$ be its minimizer; define the initial solution support to be $\mathcal{S}_0 = \{\boldsymbol{\beta} \in \Omega : \gamma_0^o(\boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}_0, \boldsymbol{\beta}) - \lambda\}$ for $\gamma^{(0)}$ defined as in Proposition 1; obtain the primal solution

$$
X_0(\boldsymbol{\beta}) = \begin{cases} \bar{X}_0(\boldsymbol{\beta}), & \boldsymbol{\beta} \in \mathcal{S}_0 \\ 0, & \text{otherwise} \end{cases}
$$

and evaluate the initial dual objective using

$$
d_0 = d_{\boldsymbol{z},0} + I\left[(\lambda + \gamma_0^o(\boldsymbol{\beta})) \times \mathbb{I}(\boldsymbol{\beta} \in \mathcal{S}_0)\right]
$$

$$
+ I\left[\gamma^{(0)}(\boldsymbol{\mu}_0, \boldsymbol{\beta}) \times \mathbb{I}(\boldsymbol{\beta} \in \Omega \setminus \mathcal{S}_0)\right],
$$

where I denotes a numerical integration method. Then, proceed using one of the following solvers.

Approximate supergradient ascent: Consider a numerical integration procedure represented by $I(\cdot)$ such that

$$
\left| I(f) - \int_{\Omega} f(\boldsymbol{\beta}) d\boldsymbol{\beta} \right| \leq \delta, \quad (42)
$$

for $\delta > 0$ and assume that I applies element-wise to vectors. Let $X_0^\star = X_0$ and for $t = 1, \ldots, T$:

i) compute the supergradients

$$\boldsymbol{p}_{\boldsymbol{\mu},t-1} = I\left[\boldsymbol{F}\left(X_{t-1}(\boldsymbol{\beta}), \boldsymbol{\beta}\right)\right] - \boldsymbol{z}_{t-1}$$

$$\boldsymbol{p}_{\nu_i,t-1} = g_i\left[\boldsymbol{z}_{t-1}\right],$$

ii) update the dual variables

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \eta_t \boldsymbol{p}_{\boldsymbol{\mu},t-1}$$

$$\nu_{i,t} = \left[\nu_{i,t-1} + \eta_t g_i\left(\boldsymbol{z}_{t-1}\right)\right]_+,$$

iii) evaluate $d_{\boldsymbol{z},t} = \min_{\boldsymbol{z}} \sum_i \nu_{i,t} g_i(\boldsymbol{z}) - \mathrm{Re}[\boldsymbol{\mu}_t^H \boldsymbol{z}]$ and let $\boldsymbol{z}_t$ be its minimizer,

iv) evaluate

$$\gamma_t^o(\boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathrm{Re}\left[\boldsymbol{\mu}_t^H \boldsymbol{F}(x, \boldsymbol{\beta})\right],$$

and let $\bar{X}_t(\boldsymbol{\beta})$ be its minimizers,

v) evaluate the dual function

$$d_t = d_{\boldsymbol{z},t} + I\left[\left(\lambda + \gamma_t^o(\boldsymbol{\beta})\right) \times \mathbb{I}\left(\boldsymbol{\beta} \in \mathcal{S}_t\right)\right]$$
$$+ I\left[\gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) \times \mathbb{I}\left(\boldsymbol{\beta} \in \Omega \setminus \mathcal{S}_t\right)\right],$$

for $\mathcal{S}_t = \{\boldsymbol{\beta} \in \Omega : \gamma_t^o(\boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) - \lambda\}$, and

vi) if $d_t > d_{t-1} + 2\delta$, obtain the primal solution

$$X_t(\boldsymbol{\beta}) = \begin{cases} \bar{X}_t(\boldsymbol{\beta}), & \boldsymbol{\beta} \in \mathcal{S}_t \\ 0, & \text{otherwise} \end{cases}$$

and let $X_t^\star = X_t$. Otherwise, $X_t^\star = X_{t-1}^\star$.
The solution of (P-SFP) is given by $X_T^\star$.

Stochastic supergradient ascent: Choose the mini-batch size $N \geq 1$ and initialize the solution set $\mathcal{X}_0 = \emptyset$. For $t = 1, \ldots, T$:

i) draw $\{\boldsymbol{\beta}_j\}$, $j = 1, \ldots, N$, uniformly at random from Ω and compute the stochastic supergradients

$$\boldsymbol{p}_{\boldsymbol{\mu},t-1} = \frac{1}{N}\sum_{j=1}^N \boldsymbol{F}\left[X_{t-1}(\boldsymbol{\beta}_j), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}_{t-1}$$

$$\boldsymbol{p}_{\nu_i,t-1} = g_i\left[\boldsymbol{z}_{t-1}\right],$$

ii) update the dual variables

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \eta_t \boldsymbol{p}_{\boldsymbol{\mu},t-1}$$

$$\nu_{i,t} = \left[\nu_{i,t-1} + \eta_t g_i\left(\boldsymbol{z}_{t-1}\right)\right]_+,$$

iii) evaluate $d_{\boldsymbol{z},t} = \min_{\boldsymbol{z}} \sum_i \nu_{i,t} g_i(\boldsymbol{z}) - \mathrm{Re}[\boldsymbol{\mu}_t^H \boldsymbol{z}]$ and let $\boldsymbol{z}_t$ be its minimizer,

iv) evaluate

$$\gamma_t^o(\boldsymbol{\beta}) = \min_{x \in \mathcal{P}} F_0(x, \boldsymbol{\beta}) + \mathrm{Re}\left[\boldsymbol{\mu}_t^H \boldsymbol{F}(x, \boldsymbol{\beta})\right]$$

and let $\bar{X}_t(\boldsymbol{\beta})$ be its minimizer,

v) evaluate the dual function

$$d_t = d_{\boldsymbol{z},t} + \int_{\mathcal{S}_t} \left[\lambda + \gamma_t^o(\boldsymbol{\beta})\right] d\boldsymbol{\beta} + \int_{\Omega \setminus \mathcal{S}_t} \gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) d\boldsymbol{\beta},$$

for $\mathcal{S}_t = \{\boldsymbol{\beta} \in \Omega : \gamma_t^o(\boldsymbol{\beta}) < \gamma^{(0)}(\boldsymbol{\mu}_t, \boldsymbol{\beta}) - \lambda\}$, and

vi) if $d_t > d_{t-1}$, obtain the primal solution

$$X_{t-1}(\boldsymbol{\beta}) = \begin{cases} \bar{X}_{t-1}(\boldsymbol{\beta}), & \boldsymbol{\beta} \in \mathcal{S}_{t-1} \\ 0, & \text{otherwise} \end{cases}$$

and let $\mathcal{X}_t = \mathcal{X}_{t-1} \cup X_t$. Otherwise, $\mathcal{X}_t = \mathcal{X}_{t-1}$.
The final solution is obtained by averaging the elements of $\mathcal{X}_T$, i.e.,

$$X^\star(\boldsymbol{\beta}) = \frac{1}{|\mathcal{X}_T|} \sum_{X \in \mathcal{X}_T} X(\boldsymbol{\beta}).$$

APPENDIX C
PROOF OF PROPOSITION 4

We actually prove the following quantitative version of Proposition 4:

Proposition 6: Under the conditions of Proposition 4, $|P^\star - P_\delta^\star| \leq c\delta + o(\delta^2)$ for

$$c = \frac{\bar{F}_0 + \lambda \mathfrak{m}(\Omega)}{\alpha \epsilon} \max\left(\left|\sum_i g_i(-\alpha \mathbb{1})\right|, \left|\sum_i g_i(\alpha \mathbb{1})\right|\right), \tag{43}$$

where $P^\star$ is the optimal value of (P-SFP) and $P_\delta^\star$ is the value of the solution obtained by Algorithm 1 when evaluating the integral in the supergradient (23a) with approximation error $0 < \delta \ll 1$ [as in (42)].

Proof: Start by noticing that evaluating the integral in (23a) numerically introduces an error term in the supergradient. Explicitly, (23a) becomes

$$\tilde{\boldsymbol{g}}_{\boldsymbol{\mu}}(\boldsymbol{\mu}', \nu_i') = \int_\Omega \boldsymbol{F}\left[X_d(\boldsymbol{\mu}', \boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} - \boldsymbol{z}_d(\boldsymbol{\mu}', \nu_i') + \boldsymbol{\delta}, \tag{44}$$

where $\boldsymbol{\delta}$ is an error vector whose magnitude is bounded by δ, i.e., $|[\boldsymbol{\delta}]_i| < \delta$. Then, observe that (44) is the supergradient of the dual function of a perturbed version of (P-SFP), namely

$$\text{minimize} \quad \int_\Omega F_0\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} + \lambda \|X\|_{L_0}$$

$$\text{subject to} \quad g_i(\boldsymbol{z}) \leq 0$$

$$\boldsymbol{z} = \int_\Omega \boldsymbol{F}\left[X(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} + \boldsymbol{\delta} \tag{PV}$$

$$X \in \mathcal{X}.$$

Hence, the value $P_\delta^\star$ of the solution obtained by the using approximate supergradient in Algorithm 1 is the optimal value of (PV). We can therefore use perturbation theory to relate the values of $P_\delta^\star$ and $P^\star$.

Formally, using the fact that the perturbation function of (P-SFP) is differentiable around zero [hypothesis (i)], we obtain the Taylor expansion $P_\delta^\star = P^\star - \boldsymbol{\mu}^{\star T} \boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2^2)$, where $o(t)$ is a term such that $o(t)/t \to 0$ as $t \to 0$ [37]. Hence, using the triangle inequality and the upper bound on the elements of $|\boldsymbol{\delta}|$, we can write

$$|P^\star - P_\delta^\star| = \left|\boldsymbol{\mu}^{\star T} \boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2^2)\right| \leq \left|\boldsymbol{\mu}^{\star T} \mathbb{1}\right| \delta + o(\delta^2). \tag{45}$$

It suffices now to bound $|\boldsymbol{\mu}^{\star T} \mathbb{1}|$, which we do in two steps.

First, we obtain an upper bound on $\boldsymbol{\mu}^{\star T}\mathbb{1}$ by recalling from (4) that the dual function d is the value of a minimization problem. Thus, taking the suboptimal $X \equiv 0$ and $\boldsymbol{z} = \alpha\mathbb{1}$, $\alpha > 0$, under hypothesis (ii) yields

$$d(\boldsymbol{\mu}^{\star}, \nu_i^{\star}) \leq \sum_i \nu_i^{\star} g_i(\alpha\mathbb{1}) - \alpha\boldsymbol{\mu}^{\star T}\mathbb{1}.$$

From Theorem 1, $d(\boldsymbol{\mu}^{\star}, \nu_i^{\star}) = P^{\star} \geq 0$, which gives

$$\boldsymbol{\mu}^{\star T}\mathbb{1} \leq \frac{\sum_i \nu_i^{\star} g_i(\alpha\mathbb{1})}{\alpha}. \tag{46}$$

Proceeding in a similar manner, we derive a lower bound by taking $X \equiv 0$ and $\boldsymbol{z} = -\alpha\mathbb{1}$ in (4), leading to

$$\boldsymbol{\mu}^{\star T}\mathbb{1} \geq -\frac{\sum_i \nu_i^{\star} g_i(-\alpha\mathbb{1})}{\alpha}. \tag{47}$$

Using the Cauchy-Schwartz inequality, the bounds in (46) and (47) yield

$$\left| \boldsymbol{\mu}^{\star T}\mathbb{1} \right| \leq \frac{\|\boldsymbol{\nu}^{\star}\|_1}{\alpha} \max\left(\left| \sum_i g_i(-\alpha\mathbb{1}) \right|, \left| \sum_i g_i(\alpha\mathbb{1}) \right| \right), \tag{48}$$

where $\boldsymbol{\nu}^{\star} = [\nu_i^{\star}]$ is a vector that collects the optimal dual variables $\nu_i^{\star}$. Note that since $\nu_i^{\star} \geq 0$, we have that $|\sum_i \nu_i^{\star}| = \|\boldsymbol{\nu}^{\star}\|_1$. All that remains to evaluate (48) is to bound $\|\boldsymbol{\nu}^{\star}\|_1$ using a classical result from optimization theory.

Explicitly, consider the strictly feasible pair $(X^{\dagger}, \boldsymbol{z}^{\dagger})$ from hypothesis (iv) and recall that $g_i(\boldsymbol{z}^{\dagger}) \leq -\epsilon$ for some $\epsilon > 0$. Plugging these suboptimal values in (4) yields

$$d(\boldsymbol{\mu}^{\star}, \nu_i^{\star}) \leq \int_{\Omega} F_0\left[X^{\dagger}(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} + \lambda \left\|X^{\dagger}\right\|_{L_0} - \sum_i \nu_i^{\star}\epsilon. \tag{49}$$

Recall that $\nu_i^{\star} \geq 0$, $\epsilon > 0$, and $d(\boldsymbol{\mu}^{\star}, \nu_i^{\star}) \geq 0$ (from Theorem 1). Thus, using the fact that $\|X^{\dagger}\|_{L_0} \leq \mathfrak{m}(\Omega)$, we readily obtain from (49) that

$$\|\boldsymbol{\nu}^{\star}\|_1 \leq \frac{\int_{\Omega} F_0\left[X^{\dagger}(\boldsymbol{\beta}), \boldsymbol{\beta}\right] d\boldsymbol{\beta} + \lambda\mathfrak{m}(\Omega)}{\epsilon}. \tag{50}$$
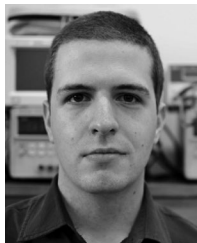
Combining (48) and (50) in (45) we obtain that $|P^{\star} - P_{\delta}^{\star}| \leq c\delta + o(\delta^2)$ for c as in (43). Furthermore, hypotheses (iii) and (iv), together with $\mathfrak{m}(\Omega) < \infty$ (since Ω is compact), imply that $c < \infty$, so that indeed $|P^{\star} - P_{\delta}^{\star}| \leq \mathcal{O}(\delta)$. ∎

REFERENCES

[1] L. F. O. Chamon, Y. C. Eldar, and A. Ribeiro, "Strong duality of sparse functional optimization," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2018, pp. 4739–4743.

[2] L. F. O. Chamon, Y. C. Eldar, and A. Ribeiro, "Sparse recovery over nonlinear dictionaries," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2019, pp. 4878–4882.

[3] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*, Englewood Cliffs, NJ, USA: Prentice-Hall, 2005.

[4] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, "Global solutions of variational models with convex regularization," *SIAM J. Imag. Sci.*, vol. 3, no. 4, pp. 1122–1145, 2010.

[5] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4735–4744, Oct. 2011.

[6] O. Bar-Ilan and Y. C. Eldar, "Sub-nyquist radar via doppler focusing," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1796–1811, Apr. 2014.

[7] D. Ma *et al.*, "Magnetic resonance fingerprinting," *Nature*, vol. 495, no. 7440, pp. 187–192, 2013.

[8] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Group-lasso on splines for spectrum cartography," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4648–4663, Oct. 2011.

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

[10] M. Unser, "Sampling—50 years after Shannon," *Proc. IEEE*, vol. 88, no. 4, pp. 569–587, 2000.

[11] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1417–1428, Jun. 2002.

[12] M. Mishali, Y. C. Eldar, and A. J. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4719–4734, Oct. 2011.

[13] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[14] G. S. Kimeldorf and G. Wahba, "A correspondence between bayesian estimation on stochastic processes and smoothing by splines," *Ann. Math. Statist.*, vol. 41, no. 2, pp. 495–502, 1970.

[15] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. Comput. Learn. Theory*, 2001, pp. 416–426.

[16] Y. C. Eldar and G. Kutyniok, Eds., *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[17] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhauser, Basel, Switzerland: Birhaüser, 2013.

[18] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[19] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7465–7490, Nov. 2013.

[20] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5987–5999, Dec. 2013.

[21] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Commun. Pure Appl. Math.*, vol. 67, no. 6, pp. 906–956, 2014.

[22] M. Cho, K.V. Mishra, J. Cai, and W. Xu, "Block iterative reweighted algorithms for super-resolution of spectrally sparse signals," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2319–2313, Dec. 2015.

[23] Z. Yang and L. Xie, "On gridless sparse methods for line spectral estimation from complete and incomplete data," *IEEE Trans. Signal Process.*, vol. 63, no. 12, pp. 3139–3153, Jun. 2015.

[24] 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, no. 4, pp. 2171–2187, Apr. 2017.

[25] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2182–2195, May 2011.

[26] B. Adcock and A. C. Hansen, "Generalized sampling and infinite-dimensional compressed sensing," *Found. Comput. Math.*, vol. 16, no. 5, pp. 1263–1323, 2016.

[27] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: A new theory for compressed sensing," *Forum Math., Sigma*, vol. 5, pp. 1–84, 2017.

[28] V. Duval and G. Peyré, "Sparse regularization on thin grids I: The Lasso," *Inverse Problems*, vol. 33, no. 5, 2017, Art. no. 055008.

[29] V. Duval and G. Peyré, "Sparse spikes super-resolution on thin grids II: The continuous basis pursuit," *Inverse Problems*, vol. 33, no. 9, 2017, Art. no. 095008.

[30] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3448–3450, Jun. 2013.

[31] A. M. Tillmann and M. E. Pfetsch, "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1248–1259, Feb. 2014.

[32] A. Natarajan and Y. Wu, "Computational complexity of certifying restricted isometry property," in *Approximation, Randomization, Combinatorial Optim. Algorithms Techn.*, 2014, pp. 371–380.

[33] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," *SIAM J. Optim.*, vol. 23, no. 3, pp. 1480–1509, 2013.

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

[35] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Berlin, Germany: Springer, 1998.

[36] A. Shapiro, "On duality theory of convex semi-infinite programming," *Optimization*, vol. 54, no. 6, pp. 535–543, 2006.

[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, MA, USA: Cambridge Univ. Press, 2004.

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

[39] J. Diestel and J. J. Uhl, Jr., *Vector Measures*. Providence, Rhode Island: AMS, 1977.

[40] D. P. Bertsekas, *Convex Optimization Theory*. Belmont, MA, USA: Athena Scientific, 2009.

[41] D. P. Bertsekas, *Convex Optimization Algorithms*. Belmont, MA, USA: Athena Scientific, 2015.

[42] A. Ruszczyński and W. Syski, "On convergence of the stochastic subgradient method with on-line stepsize rules," *J. Math. Anal. Appl.*, vol. 114, no. 2, pp. 512–527, 1986.

[43] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[44] W. Rudin, *Functional Analysis*. New York, NY, USA: McGraw-Hill, 1991.

[45] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Germany: Springer, 2009.

[46] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Berlin, Germany: Springer, 2018.

[47] M.F. da Costa and W. Dai, "Sampling patterns for off-the-grid spectral estimation," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, 2017, pp. 318–322.

[48] N. Rao, P. Shah, and S. Wright, "Forward-backward greedy algorithms for atomic norm regularization," *IEEE Trans. Signal Process.*, vol. 63, no. 21, pp. 5798–5811, Nov. 2015.

[49] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Berlin, Germany: Springer, 2005.

[50] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 774–782.

[51] J. Feng, H. Xu, S. Mannor, and S. Yan, "Robust logistic regression and classification," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 253–261.

[52] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 482–494, Jan. 2013.

[53] J. Tibshirani and C. D. Manning, "Robust logistic regression using shift parameters," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, 2014, pp. 124–129.

[54] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data Mining Knowl. Discovery*, vol. 31, no. 3, pp. 606–660, 2017.

[55] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "UEA & UCR time series classification repository: ECG200 dataset," 2017. [Online]. Available: http://www.timeseriesclassification.com/description.php?Dataset=ECG200

[56] A. L. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2002. She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. She was previously a Professor with the Department of Electrical Engineering, Technion, where she held the Edwards Chair in Engineering. She is also a Visiting Professor with MIT, a Visiting Scientist with the Broad Institute, and an Adjunct Professor with Duke University, Durham, NC, USA and was a Visiting Professor with Stanford. Her research interests include the broad areas of statistical signal processing, sampling theory and compressed sensing, learning, and optimization methods, and their applications to biology and optics. Dr. Eldar has received numerous awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), and the IEEE Kiyo Tomiyasu Award (2016). She was a Horev Fellow of the Leaders in Science and Technology program, Technion and an Alon Fellow. She received the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair at the Technion, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technions Award for Excellence in Teaching (twice). She received several best paper awards and best demo awards together with her research students and colleagues including the SIAM outstanding Paper Prize and the IET Circuits, Devices and Systems Premium Award, and was selected as one of the 50 most influential women in Israel. She was a member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education. She is a member of the Israel Academy of Sciences and Humanities (elected 2017), a EURASIP Fellow. She is the Editor-in-Chief of *Foundations and Trends in Signal Processing*, a member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. In the past, she was a Signal Processing Society Distinguished Lecturer, member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*. She was Co-Chair and Technical Co-Chair of several international conferences and workshops.

Luiz F. O. Chamon (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the University of São Paulo, São Paulo, Brazil, in 2011 and 2015, respectively. In 2009, he was an undergraduate exchange student at the master's in acoustics with the École Centrale de Lyon, Lyon, France. He is currently working toward the Ph.D. degree in electrical and systems engineering with the University of Pennsylvania (Penn), Philadelphia, PA, USA. In 2009, he was an Assistant Instructor and Consultant on nondestructive testing with IN-SACAST Formation Continue. From 2010 to 2014, he worked as a Signal Processing and Statistical Consultant on a project with EMBRAER. His research interest include signal processing, optimization, statistics, and control.

Alejandro Ribeiro received the B.Sc. degree in electrical engineering from the Universidad de la Republica Oriental del Uruguay, Montevideo, Uruguay, in 1998 and the M.Sc. and Ph.D. degrees in electrical engineering from the Department of Electrical and Computer Engineering, the University of Minnesota, Minneapolis, MN, USA, in 2005 and 2007, respectively. From 1998 to 2003, he was a member of the technical staff with Bellsouth Montevideo. After his M.Sc. and Ph.D studies, in 2008, he joined the University of Pennsylvania (Penn), Philadelphia, PA, USA, where he is currently the Rosenbluth Associate Professor with the Department of Electrical and Systems Engineering. His research interests include the applications of statistical signal processing to the study of networks and networked phenomena. His focus is on structured representations of networked data structures, graph signal processing, network optimization, robot teams, and networked control. Dr. Ribeiro received the 2014 O. Hugo Schuck Best Paper Award, and paper awards at CDC 2017, 2016 SSP Workshop, 2016 SAM Workshop, 2015 Asilomar SSC Conference, ACC 2013, ICASSP 2006, and ICASSP 2005. His teaching has been recognized with the 2017 Lindback award for distinguished teaching and the 2012 S. Reid Warren, Jr. Award presented by Penn's undergraduate student body for outstanding teaching. Dr. Ribeiro is a Fulbright scholar class of 2003 and a Penn Fellow class of 2015.