Efficient First-Order Algorithms for Adaptive Signal Denoising

Dmitrii Ostrovskii [1] **Zaid Harchaoui** [2]

Abstract

We consider the problem of discrete-time signal denoising, focusing on a specific family of non-linear convolution-type estimators. Each such estimator is associated with a time-invariant filter which is obtained adaptively, by solving a certain convex optimization problem. Adaptive convolution-type estimators were demonstrated to have favorable statistical properties, see (Juditsky & Nemirovski, 2009; 2010; Harchaoui et al., 2015b; Ostrovsky et al., 2016). Our first contribution is an efficient algorithmic implementation of these estimators via the known first-order proximal algorithms. Our second contribution is a computational complexity analysis of the proposed procedures, which takes into account their statistical nature and the related notion of statistical accuracy. The proposed procedures and their analysis are illustrated on a simulated data benchmark.

1. Introduction

We consider the problem of discrete-time signal denoising. The goal is to estimate a discrete-time complex signal (x_τ) observed in complex Gaussian noise of level σ on $[-n, n]$:

$$y_\tau := x_\tau + \sigma\zeta_\tau, \quad \tau = -n, ..., n. \tag{1}$$

Here, ζ_τ are i.i.d. random variables with standard complex Gaussian distribution $\mathbb{C}\mathcal{N}(0, 1)$, that is, $\mathrm{Re}(\zeta_\tau)$ and $\mathrm{Im}(\zeta_\tau)$ are independent standard Gaussian random variables.

Signal denoising is a classical problem in statistical estimation and signal processing; see (Ibragimov & Khasminskii, 1981; Nemirovski, 2000; Tsybakov, 2008; Wasserman, 2006; Haykin, 1991; Kay, 1993). The conventional approach is to assume that x comes from a known set $\mathcal{X}$ with a simple structure that can be exploited to build the estimator. For example, one might consider signals belonging to linear subspaces $\mathcal{S}$ of signals whose spectral representation, as given by the Discrete Fourier or Discrete Wavelet transform, comes from a linearly transformed ℓ_p-ball, see (Tsybakov, 2008; Johnstone, 2011). In all these cases, estimators with near-optimal statistical performance can be found in explicit form, and correspond to linear functionals of the observations y – hence the name *linear estimators*.

We focus here on a family of *non-linear* estimators with larger applicability and strong theoretical guarantees, in particular when the structure of the signal is unknown beforehand, as studied in (Nemirovski, 1992; Juditsky & Nemirovski, 2009; 2010; Harchaoui et al., 2015b; Ostrovsky et al., 2016). Assuming for convenience that one must estimate x_t on $[0, n]$ from observations (1), these estimators can be expressed as

$$\widehat{x}_t^{\varphi} = [\varphi * y]_t := \sum_{\tau\in\mathbb{Z}} \varphi_\tau y_{t-\tau} \quad 0 \le t \le n; \tag{2}$$

here φ is called a *filter* and is supported on $[0, n]$ which we write as $\varphi \in \mathbb{C}_n(\mathbb{Z})$, and $*$ is the (non-circular) discrete convolution. For estimators in this family, the filter is then obtained as an optimal solution to some convex optimization problem. For instance, the *Penalized Least-Squares* estimator (Ostrovsky et al., 2016) is defined by

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi\in\mathbb{C}_n(\mathbb{Z})} \tfrac{1}{2}\|F_n[y - \varphi * y]\|_2^2 + \lambda\|F_n[\varphi]\|_1, \tag{3}$$

where F_n is the Discrete Fourier transform (DFT) on $\mathbb{C}^{n+1}$, and $\|\cdot\|_p$ is the ℓ_p-norm on $\mathbb{C}^{n+1}$. We shall give a summary of the various estimators of the family in the end of this section. Optimization problems associated to all of them rest upon a common principle – minimization of the residual $\|F_n[y - \varphi * y]\|_p$, with $p \in \{2, \infty\}$, regularized via the ℓ_1-norm of the DFT of the filter.

The statistical properties of adaptive convolution-type estimators have been extensively studied. In particular, such estimators were shown to be nearly minimax-optimal, with respect to the pointwise loss and ℓ_2-loss, for signals belonging to arbitrary, and unknown, shift-invariant linear subspaces of $\mathbb{C}(\mathbb{Z})$ with bounded dimension, or sufficiently close to such subspaces as measured by the local ℓ_p-norms, see (Nemirovski, 1992; Juditsky & Nemirovski, 2009; 2010; Harchaoui et al., 2015b; Ostrovsky et al., 2016). We give a

[1]SIERRA Project-Team, INRIA Paris, Paris, France [2]Department of Statistics, University of Washington, Seattle, USA. Correspondence to: <dmitrii.ostrovskii@inria.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

summary of statistical properties of convolution-type estimators in the supplementary material.

However, the question of the algorithmic implementation of such estimators remains largely unexplored; in fact, we are not aware of any publicly available implementation of these estimators. Our goal here is to close this gap. Note that problems similar to (3) belong to the general class of second-order cone problems, and hence can in principle be solved to high numerical accuracy in polynomial time via interior-point methods (Ben-Tal & Nemirovski, 2001). However, the computational complexity of interior-point methods grows polynomially with the problem dimension, and becomes prohibitive in signal and image denoising problems (for example, in image denoising this number is proportional to the number of pixels which might be as large as 10^8). Furthermore, it is unclear whether high-accuracy solutions are necessary when the optimization problem is solved with the goal of obtaining a statistical estimator. In such cases, the level of accuracy sought, or the amount of computations performed, should rather be *adjusted* to the statistical performance of the exact estimator itself. While these matters have previously been investigated in the context of linear regression (Pilanci & Wainwright, 2016) and sparse recovery (Bruer et al., 2015), our work studies them in the context of convolution-type estimators.

Notably, (3) and its counterparts have favorable properties:

- *Easily accessible first-order information.* The objective value and gradient at a given point can be computed in time $O(n \log n)$ via a series of Fast Fourier Transforms (FFT) and elementwise vector operations.
- *Simple geometry.* After a straightforward reparametrization, one is left with ℓ_1-norm penalty or ℓ_1-ball as a feasible set in the constrained formulation. Prox-mappings for such problems, with respect to both the Euclidean and the "ℓ_1-adapted" distance-generating functions, can be computed efficiently.
- *Medium accuracy is sufficient.* We show that approximate solutions with specified (medium) accuracy preserve the statistical performance of the exact solutions.

All these properties make first-order optimization algorithms the tools of choice to deal with (3) and similar problems.

Outline. In Section 2, we recall two general classes of optimization problems, *composite minimization* (Beck & Teboulle, 2009; Nesterov & Nemirovski, 2013) and *composite saddle-point* problems (Juditsky & Nemirovski, 2011; Nesterov & Nemirovski, 2013), and the first-order optimization algorithms suitable for their numerical solution. In Section 3, we show how to recast the optimization problems related to convolution-type estimators in one of the above general forms. We then describe how to compute first-order oracles in the resulting problems efficiently using FFT. In Section 4, we establish problem-specific worst-case complexity bounds for the proposed first-order algorithms. These bounds are expressed in terms of the quantities that control the statistical difficulty of the signal recovery problem: signal length n, noise variance σ^2, and parameter r corresponding to the ℓ_1-norm of the Discrete Fourier tranform of the optimal solution. A remarkable consequence of these bounds is that just $\tilde{O}(\text{PSNR}+1)$ iterations of a suitable first-order algorithm are sufficient to match the statistical properties of an exact estimator; here $\text{PSNR} := \|F_{2n}[x]_{-n}^{n}\|_\infty/\sigma$ is the peak signal-to-noise ratio in the Fourier domain. This gives a rigorous characterization (in the present context) of the performance of "early stopping" strategies that allow to stop an optimization algorithm much earlier than dictated purely by the optimization analysis. In Section 5, we present numerical experiments on simulated data which complement our theoretical analysis[1].

Notation. We denote $\mathbb{C}(\mathbb{Z})$ the space of all complex-valued signals on $\mathbb{Z}$, or, simply, the space of all two-sided complex sequences. We call $\mathbb{C}_n(\mathbb{Z})$ the finite-dimensional subspace of $\mathbb{C}(\mathbb{Z})$ consisting of signals supported on $[0, n]$:

$$\mathbb{C}_n(\mathbb{Z}) = \{(x_\tau) \in \mathbb{C}(\mathbb{Z}) : x_\tau = 0 \ \text{ whenever } \tau \notin [0, n]\};$$

its counterpart $\mathbb{C}_n^{\pm}(\mathbb{Z})$ consists of all signals supported on $[-n, n]$. The unknown signal is assumed to come from one of such subspaces, which corresponds to a finite signal length. Note that signals from $\mathbb{C}(\mathbb{Z})$ can be naturally mapped to column vectors by means of the index-restriction operator $[\cdot]_m^n$, defined for any $m, n \in \mathbb{Z}$ such that $m \le n$ as

$$[x]_m^n \in \mathbb{C}^{n-m+1}.$$

In particular, $[\cdot]_0^n$ and $[\cdot]_{-n}^n$ define one-to-one mappings $\mathbb{C}_n(\mathbb{Z}) \to \mathbb{C}^{n+1}$ and $\mathbb{C}_n^{\pm}(\mathbb{Z}) \to \mathbb{C}^{2n+1}$. For convenience, column-vectors in $\mathbb{C}^{n+1}$ and $\mathbb{C}^{2n+1}$ will be indexed starting from zero. We define the scaled ℓ_p-seminorms on $\mathbb{C}(\mathbb{Z})$:

$$\|x\|_{n,p} := \frac{\|[x]_0^n\|_p}{(n+1)^{1/p}} = \left(\frac{1}{n+1}\sum_{\tau=0}^{n}|x_\tau|^p\right)^{1/p}, \ p \ge 1.$$

We use the "Matlab notation" for matrix concatenation: $[A; B]$ is the vertical, and $[A, B]$ the horizontal concatenation of two matrices with compatible dimensions. We introduce the unitary Discrete Fourier Transform (DFT) operator F_n on $\mathbb{C}^{n+1}$, defined by

$$[F_n x]_k = \frac{1}{\sqrt{n+1}}\sum_{t=0}^{n} x_t \exp\left(\frac{2\pi i k t}{n+1}\right), \quad 0 \le k \le n.$$

[1] The code reproducing all our experiments is available online at `https://github.com/ostrodmit/AlgoRec`.

The unitarity of F_n implies that its inverse F_n^{-1} coincides with its conjugate transpose $F_n^{\rm H}$. Slightly abusing the notation, we will occasionally shorten $F_n[x]_0^n$ to $F_n[x]$. In other words, $F_n[\cdot]$ is a map $\mathbb{C}_n(\mathbb{Z}) \to \mathbb{C}^{n+1}$, and the adjoint map $F_n^{\rm H}[x]$ simply sends $F_n^{\rm H}[x]_0^n$ to $\mathbb{C}_n(\mathbb{Z})$ via zero-padding. We use the "Big-O" notation: for two non-negative functions f, g on the same domain, $g = O(f)$ means that there is a generic constant $C \geq 0$ such that $g \leq Cf$ for any admissible value of the argument; $g = \tilde{O}(f)$ means that C is replaced with $C(\log^\kappa(n) + 1)$ for some $\kappa > 0$; hereinafter $\log(\cdot)$ is the natural logarithm, and C is a generic constant.

Estimators. We now summarize all the estimators that are of interest in this paper. For brevity, we use the notation

$$\mathrm{Res}_p(\varphi) := \|F_n[y - \varphi * y]\|_p. \tag{4}$$

- *Constrained Uniform-Fit estimator,* given for $\overline{r} \geq 0$ by

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \Phi_n(\overline{r})} \mathrm{Res}_\infty(\varphi), \tag{Con-UF}$$

$$\Phi_n(\overline{r}) := \left\{ \varphi \in \mathbb{C}_n(\mathbb{Z}) : \|F_n[\varphi]\|_1 \leq \frac{\overline{r}}{\sqrt{n+1}} \right\};$$

- *Constrained Least-Squares estimator:*

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \Phi_n(\overline{r})} \tfrac{1}{2}\mathrm{Res}_2^2(\varphi); \tag{Con-LS}$$

- *Penalized Uniform-Fit estimator:*

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_n(\mathbb{Z})} \mathrm{Res}_\infty(\varphi) + \lambda\|F_n[\varphi]\|_1; \tag{Pen-UF}$$

- *Penalized Least-Squares estimator:*

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_n(\mathbb{Z})} \tfrac{1}{2}\mathrm{Res}_2^2(\varphi) + \lambda\|F_n[\varphi]\|_1. \tag{Pen-LS}$$

We also consider (Con-LS*) and (Pen-LS*) – counterparts of (Con-LS) and (Pen-LS) in which $\frac{1}{2}\mathrm{Res}_2^2(\varphi)$ is replaced with non-squared residual $\mathrm{Res}_2(\varphi)$. Note that (Con-LS*) is equivalent to (Con-LS), *i.e.* results in the same estimator; however, this does *not* hold for (Pen-LS*) and (Pen-LS).

2. Tools from Convex Optimization

In this section, we recall the tools from first-order convex optimization to be used later. We describe two general types of optimization problems, *composite minimization* and *composite saddle-point* problems, together with efficient first-order algorithms for their solution. Following (Nesterov & Nemirovski, 2013), we begin by introducing the concept of *proximal setup* which underlies these algorithms.

2.1. Proximal Setup

Let a *domain* U be a closed convex set in a Euclidean space E. A *proximal setup* for U is given by a norm $\|\cdot\|$ on E (not necessarily Euclidean), and a *distance-generating function* (d.-g. f.) $\omega(u) : U \to \mathbb{R}$, such that $\omega(u)$ is continuous and convex on U, admits a continuous selection $\omega'(u) \in \partial\omega(u)$ of subgradients on the set $\{u \in U : \partial\omega(u) \neq \emptyset\}$, and is 1-strongly convex with respect to $\|\cdot\|$.

The concept of proximal setup gives rise to several notions (see (Nesterov & Nemirovski, 2013) for a detailed exposition): the ω-center u_ω, the Bregman divergence $D_u(\cdot)$, the ω-radius $\Omega[\cdot]$ and the prox-mapping $\mathrm{Prox}_u(\cdot)$ defined as

$$\mathrm{Prox}_u(g) = \operatorname*{argmin}_{\xi \in U} \{\langle g, \xi\rangle + D_u(\xi)\} .$$

Blockwise Proximal Setups. We now describe a specific family of proximal setups which proves to be useful for our purposes. Let $E = \mathbb{R}^N$ with $N = 2(n+1)$; note that we can identify this space with $\mathbb{C}^{n+1}$ via (Hermitian) vectorization map $\mathrm{Vec}_n : \mathbb{C}^{n+1} \to \mathbb{R}^{2(n+1)}$,

$$\mathrm{Vec}_n z = [\mathrm{Re}(z_0); \mathrm{Im}(z_0); ...; \mathrm{Re}(z_n); \mathrm{Im}(z_n)]. \tag{5}$$

Now, supposing that $N = k(m+1)$ for some non-negative integers m, k, let us split $u = [u^0; ...; u^m] \in \mathbb{R}^N$ into $m+1$ blocks of size k, and equip $\mathbb{R}^N$ with the group ℓ_1/ℓ_2-norm:

$$\|u\| := \sum_{j=0}^m \|u^j\|_2. \tag{6}$$

We also define the balls $U_N(R) := \{u \in \mathbb{R}^N : \|u\| \leq R\}$.

Theorem 2.1 ((Nesterov & Nemirovski, 2013))**.** *Given $E = \mathbb{R}^N$ as above, $\omega : \mathbb{R}^N \to \mathbb{R}$ defined by*

$$\omega(u) = \frac{(m+1)^{(\tilde{q}-1)(2-\tilde{q})/\tilde{q}}}{2\tilde{c}} \left[\sum_{j=0}^m \|u^j\|_2^{\tilde{q}}\right]^{2/\tilde{q}} \tag{7}$$

with $(\tilde{q}, \tilde{c}) = \begin{cases} \left(2, \frac{1}{m+1}\right), & m \leq 1, \\ \left(1 + \frac{1}{\log(m+1)}, \frac{1}{e\log(m+1)}\right), & m \geq 2, \end{cases}$

is a d.-g. f. for any ball $U_N(R)$ of the norm (6) *with ω-center $u_\omega = 0$. Moreover, for some constant C and any $R \geq 0$ and $m, k \in \mathbb{Z}_+$, ω-radius of $U_N(R)$ is bounded as*

$$\Omega[U_N(R)] \leq C(\sqrt{\log(m+1)} + 1)R. \tag{8}$$

We will use two particular cases of the above construction.

(i) Case $m = n$, $k = 2$ corresponds to the ℓ_1-norm on $\mathbb{C}^{n+1}$, and specifies the *complex ℓ_1-setup*.

Algorithm 1 Fast Gradient Method

Input: stepsize $\eta > 0$
$u^0 = u_\omega$
$g^0 = 0 \in E$
for $t = 0, 1, ...$ **do**
$u_t = \mathrm{Prox}_{\eta\Psi, u_\omega}(\eta g^t)$
$\tau_t = \frac{2(t+2)}{(t+1)(t+4)}$
$u_{t+\frac{1}{3}} = \tau_t u_t + (1-\tau_t)u^t$
$g_t = \frac{t+2}{2}\nabla f(u_{t+\frac{1}{3}})$
$u_{t+\frac{2}{3}} = \mathrm{Prox}_{\eta\Psi, u_t}(\eta g_t)$
$u^{t+1} = \tau_t u_{t+\frac{2}{3}} + (1-\tau_t)u^t$
$g^{t+1} = \sum_{\tau=0}^{t} g_t$
end for

(ii) Case $m = 0$, $k = N$ corresponds to the ℓ_2-norm on $\mathbb{C}^{n+1}$, and specifies the ℓ_2-*setup* $(\|\cdot\|_2, \frac{1}{2}\|\cdot\|_2^2)$.

To work with them, we introduce specific norms on $\mathbb{R}^N$:

$$\|u\|_{\mathbb{C},p} := \|\mathrm{Vec}_n^{-1} u\|_p = \|\mathrm{Vec}_n^{\mathsf{H}} u\|_p, \quad p \geq 1. \tag{9}$$

Note that $\|\cdot\|_{\mathbb{C},1}$ gives the norm $\|\cdot\|$ in the complex ℓ_1-setup, while $\|\cdot\|_{\mathbb{C},2}$ coincides with the standard ℓ_2-norm on $\mathbb{R}^N$.

2.2. Composite Minimization Problems

The general *composite minimization* problem has the form

$$\min_{u\in U} \{\phi(u) = f(u) + \Psi(u)\}. \tag{10}$$

Here, U is a domain in E equipped with $\|\cdot\|$, $f(u)$ is convex and continuously differentiable on U, and $\Psi(u)$ is convex, lower-semicontinuous, finite on the relative interior of U, and can be non-smooth. Assuming that U is equipped with a proximal setup $(\|\cdot\|, \omega(\cdot))$, let us define the *composite prox-mapping*, see (Beck & Teboulle, 2009), as follows:

$$\mathrm{Prox}_{\Psi,u}(g) = \operatorname*{argmin}_{\xi\in U} \{\langle g, \xi\rangle + D_u(\xi) + \Psi(\xi)\}. \tag{11}$$

Fast Gradient Method. Fast Gradient Method (FGM), summarized as Algorithm 1, was introduced in (Nesterov, 2013) as an extension of the celebrated Nesterov algorithm for smooth minimization (Nesterov, 1983) to the case of constrained problems with non-Euclidean proximal setups. It is guaranteed to find an approximate solution of (10) with $O(1/T^2)$ accuracy after T iterations. We defer the rigorous statement of this accuracy bound to Sec. 4.

2.3. Composite Saddle-Point Problems

We also consider general *composite saddle-point* problems:

$$\inf_{u\in U} \max_{v\in V} [\phi(u,v) = f(u,v) + \Psi(u)]. \tag{12}$$

Here, $U \subset E_u$ and $V \subset E_v$ are domains in the corresponding Euclidean spaces E_u, E_v, and in addition V is compact; function $f(u,v)$ is convex in u, concave in v, and differentiable on $W := U \times V$; function $\Psi(u)$ is convex, lower-semicontinuous, can be non-smooth, and is such that $\mathrm{Prox}_{\Psi,u}(g)$ is easily computable. We can associate with f a smooth vector field $F : W \to E_u \times E_v$, given by

$$F([u;v]) = [\nabla_u f(u,v); -\nabla_v f(u,v)].$$

Saddle-point problem (12) specifies two convex optimization problems: that of minimization of $\overline{\phi}(u) = \max_{v\in V} \phi(u,v)$, or the primal problem, and that of maximization of $\underline{\phi}(v) = \inf_{u\in U} \phi(u,v)$, or the dual problem. Under the general conditions which hold in the described setting, see *e.g.* (Sion, 1958), (12) possesses an optimal solution $w^* = [u^*; v^*]$, called a *saddle point*, such that the value of (12) is $\phi(u^*, v^*) = \overline{\phi}(u^*) = \underline{\phi}(v^*)$, and u^*, v^* are optimal solutions to the primal and dual problems. The quality of a candidate solution $w = [u; v]$ can be evaluated via the *duality gap* – the sum of the primal and dual accuracies:

$$\overline{\phi}(u) - \underline{\phi}(v) = [\overline{\phi}(u) - \overline{\phi}(u^*)] + [\underline{\phi}(v^*) - \underline{\phi}(v)].$$

Constructing the Joint Setup. When having a saddle-point problem at hand, one usually begins with "partial" proximal setups $(\|\cdot\|_U, \omega_U)$ for $U \subseteq E_u$, and $(\|\cdot\|_V, \omega_V)$ for $V \subset E_v$, and must construct a "joint" proximal setup on W. Let us introduce the segment $U_* = [u^*, u_\omega]$, where u_ω is the u-component of the ω-center w_ω of W. Moreover, folllowing (Nesterov & Nemirovski, 2013), let us assume that the dual ω-radius $\Omega[V]$ and the "effective" primal ω-radius, defined as

$$\Omega_*[U] := \min(\Omega[U], \Omega[U_*]),$$

are known (note that $\Omega[U]$ can be infinite but $\Omega_*[U]$ cannot). We can then construct a proximal setup

$$\begin{aligned} \|w\|^2 &= \Omega^2[V]\,\|u\|_U^2 + \Omega_*^2[U]\,\|v\|_V^2, \\ \omega(w) &= \Omega^2[V]\,\omega_U(u) + \Omega_*^2[U]\,\omega_V(v). \end{aligned} \tag{13}$$

Note that the corresponding joint prox-mapping is reduced to the prox-mappings for the primal and dual setups.

Composite Mirror Prox. Composite Mirror Prox (CMP), introduced in (Nesterov & Nemirovski, 2013) and summarized here as Algorithm 2, solves the general composite saddle-point problem (12). When applied with proximal setup (13), this algorithm admits an $O(1/T)$ accuracy bound after T iterations; the formal statement is deferred to Sec. 4.

3. Algorithmic Implementation

Change of Variables. When working with convolution-type estimators, our first step is to transfer the problem to

Algorithm 2 Composite Mirror Prox

Input: stepsize $\eta > 0$
$w_0 := [u_0; v_0] = w_\omega$
for $t = 0, 1, ...$ **do**
$w_{t+\frac{1}{2}} = \text{Prox}_{\eta\Psi, w_t}(\eta F(w_t))$
$w_{t+1} = \text{Prox}_{\eta\Psi, w_t}(\eta F(w_{t+\frac{1}{2}}))$
$w^{t+1} := [u^{t+1}; v^{t+1}] = \frac{1}{t+1}\sum_{\tau=0}^{t} w_\tau$
end for

the Fourier domain, so that the feasible set and the penalization term become quasi-separable. Namely, noting that the adjoint map of $\text{Vec}_n : \mathbb{C}^{n+1} \to \mathbb{R}^{2n+2}$, cf. (5), is given by

$$\text{Vec}_n^{\text{H}} u = [u_0;\, u_2;\, ...;\, u_{2n}] + i[u_1;\, u_3;\, ...;\, u_{2n+1}],$$

consider the transformation

$$u = \text{Vec}_n F_n[\varphi] \quad b = \text{Vec}_n F_n[y] \tag{14}$$

Note that $\varphi = F_n^{\text{H}}[\text{Vec}_n^{\text{H}} u] \in \mathbb{C}_n(\mathbb{Z})$, and hence

$$\|F_n[y - y * \varphi]\|_2^2 = \|Au - b\|_2^2,$$

where $A : \mathbb{R}^{2n+2} \to \mathbb{R}^{2n+2}$ is defined by

$$Au = \text{Vec}_n F_n \left[y * F_n^{\text{H}}[\text{Vec}_n^{\text{H}} u]\right]. \tag{15}$$

We are about to see that all recovery procedures can indeed be cast into one of the "canonical" forms (10), (12). Moreover, the gradient computation is then reduced to evaluating the convolution-type operator A and its adjoint $A^{\text{H}} = A^{\text{T}}$.

Problem Reformulation. After the change of variables (14), problems (Con-LS) and (Pen-LS) take form (10):

$$\min_{\|u\|_{\mathbb{C},1} \le R} \left[f(u) := \tfrac{1}{2}\|Au - b\|_2^2\right] + \lambda\|u\|_{\mathbb{C},1}, \tag{16}$$

where $\|\cdot\|_{\mathbb{C},p}$ is defined in (9). In particular, (Con-LS) is obtained from (16) by setting $\lambda = 0$ and $R = \frac{\overline{r}}{\sqrt{n+1}}$, and (Pen-LS) is obtained by setting $R = \infty$. Note that

$$\nabla f(u) = A^{\text{T}}(Au - b).$$

On the other hand, problems (Con-UF), (Pen-UF), and (Con-LS*) can be recast as saddle-point problems (12). Indeed, the dual norm to $\|\cdot\|_{\mathbb{C},p}$ is $\|\cdot\|_{\mathbb{C},q}$ with $q = \frac{p}{p-1}$, whence

$$\|F_n[y - y * \varphi]\|_p = \|Au - b\|_{\mathbb{C},p} = \max_{\|v\|_{\mathbb{C},q} \le 1} \langle v, Au - b\rangle;$$

as such, (Con-UF), (Pen-UF) and (Con-LS*) are reduced to a saddle-point problem

$$\min_{\|u\|_{\mathbb{C},1} \le R} \max_{\|v\|_{\mathbb{C},q} \le 1} [f(u,v) := \langle v, Au - b\rangle] + \lambda\|u\|_{\mathbb{C},1}, \tag{17}$$

where $q = 1$ for (Con-UF) and (Pen-UF), and $q = 2$ in case of (Con-LS*). Note that $f(u,v)$ is bilinear, and one has

$$[\nabla_u f(u,v); \nabla_v f(u,v)] = [A^{\text{T}} v; Au - b].$$

We are now in the position to apply the algorithms described in Sec. 2. One iteration of either of them is reduced to a few computations of the gradient (which, in turn, is reduced to evaluating A and A^{T}) and prox-mappings. We now show how to evaluate operators A and A^{T} in time $O(n \log n)$.

Evaluation of Au and $A^{\text{T}}v$. Operator A, cf. (15), can be evaluated in time $O(n \log n)$ via FFT. The key fact is that the convolution $[y * \varphi]_0^n$ is contained in the first $n+1$ coordinates of the *circular* convolution of $[y]_{-n}^n$ with a zero-padded filter $\psi = [[\varphi]_0^n; 0_n] \in \mathbb{C}^{2n+1}$. Using the DFT diagonalization property, this fact can be expressed as

$$[y * \varphi]_t = \sqrt{2n+1}\,[F_{2n}^{\text{H}} D_y F_{2n} \psi]_t, \quad 0 \le t \le n,$$

where operator $D_y = \text{diag}(F_{2n}[y]_{-n}^n)$ on $\mathbb{C}^{2n+1}$ can be constructed in $O(n \log n)$ by FFT, and evaluated in $O(n)$. Let $P_n : \mathbb{C}^{2n+1} \to \mathbb{C}^{n+1}$ project to the first $n+1$ coordinates of $\mathbb{C}^{2n+1}$; its adjoint P_n^{H} is the zero-padding operator which complements $[\varphi]_0^n$ with n trailing zeroes. Then,

$$Au = \sqrt{2n+1} \cdot \text{Vec}_n F_n P_n F_{2n}^{\text{H}} D_y F_{2n} P_n^{\text{H}} F_n^{\text{H}} \text{Vec}_n^{\text{H}} u, \tag{18}$$

where all operators in the right-hand side can be evaluated in $O(n \log n)$. Operator $A^{\text{T}} = A^{\text{H}}$ can be treated in the same manner by taking the adjoint of (18).

Computation of Prox-Mappings. It is worth mentioning that the composite prox-mappings in all cases of interest can be computed in time $O(n)$; in some cases it can be done explicitly, and in others via a root-finding algorithm. These computations are described in the supplementary material.

4. Theoretical Analysis

4.1. Bounds on Absolute Accuracy

We first recall from (Nesterov & Nemirovski, 2013) the worst-case bounds on the *absolute accuracy* in objective, defined as $\varepsilon(t) := \phi(u^t) - \phi(u^*)$ for composite minimization problems, and $\overline{\varepsilon}(t) := \overline{\phi}(u^t) - \overline{\phi}(u^*)$ for saddle-point problems. These bounds, summarized in Theorems 4.1–4.2 below, are applicable when solving *arbitrary* problems of the types (10), (12) with the suitable first-order algorithm, and are expressed in terms of the "optimization" parameters that specify the regularity of the objective and the ω-radius.

Theorem 4.1. *Suppose that f has L_f-Lipschitz gradient:*

$$\|\nabla f(u) - \nabla f(u')\|_* \le L_f\|u - u'\| \quad \forall u, u' \in U$$

where $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$, and let u^T be generated by T iterations of Algorithm 1 with stepsize $\eta = \frac{1}{L_f}$. Then,

$$\varepsilon(T) = O\left(\frac{L_f \Omega_*^2[U]}{T^2}\right).$$

Theorem 4.2. *Let $f(u,v)$ be as in* (17)[2]*, and assume that vector field F is L_F-Lipschitz on $W = U \times V$:*

$$\|F(w) - F(w')\|_* \le L_F \|w - w'\| \quad \forall w, w' \in W.$$

Let $w^T = [u^T; v^T]$ be generated by T iterations of Algorithm 2 with joint setup (13) *and $\eta = \frac{\Omega[V]}{\Omega_*[U] L_F}$. Then,*

$$\overline{\varepsilon}(T) = O\left(\frac{L_F \Omega_*[U] \Omega[V]}{T}\right).$$

Our next goal is to translate these bounds into the language of "statistical" parameters such as the norm of exact estimator and the peak signal-to-noise ratio in the Fourier domain, cf. Sec. 1. Let us make a couple of observations beforehand.

The first observation concerns the proximal setups to be used, and allows to control the ω-radii. If the partial domain (for u or v) is an $\|\cdot\|_{\mathbb{C},2}$-norm ball, we will naturally use the ℓ_2-setup in that variable. If the domain is an $\|\cdot\|_{\mathbb{C},1}$-norm ball, we will consider choosing between the ℓ_1-setup which is "adapted" to the geometry of the problem, see (Nesterov & Nemirovski, 2013), or the ℓ_2-setup due to its simplicity in use. Note that in all these cases, the partial domains either coincide with or are contained in the balls $U_N(1), U_N(R)$ of the corresponding norms, cf. (8), whence ω-radii $\Omega[V], \Omega_*[U]$ can be bounded as follows:

$$\Omega[V] = \tilde{O}(1), \quad \Omega_*[U] = \tilde{O}\left(r/\sqrt{n+1}\right), \tag{19}$$

where

$$r = \sqrt{n+1}\|F_n[\widehat{\varphi}]\|_1 \tag{20}$$

is the scaled norm of an optimal solution (note that $\overline{r} \ge r$).

The second observation concerns the Lipschitz constants L_f, L_F in the chosen setups. It is convenient to define parameters q_u, q_v that take values in $\{2, 1\}$ depending on the partial setup used in the corresponding variable; besides, let $p_u = \frac{q_u}{q_u - 1}$ and $p_v = \frac{q_v}{q_v - 1}$. Introducing the complex counterpart of A, operator $\mathcal{A} : \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$ given by

$$\mathcal{A}[\varphi]_0^n = F_n[y * F_n^{\mathrm{H}}[\varphi]_0^n] \quad \Leftrightarrow \quad A = \mathrm{Vec}_n \circ \mathcal{A} \circ \mathrm{Vec}_n^{\mathrm{H}},$$

we can conveniently express Lipshitz constants L_f, L_F in terms of operator norms $\|\mathcal{A}\|_{\alpha\to\beta} := \sup_{\|\psi\|_\alpha = 1} \|\mathcal{A}\psi\|_\beta$:

$$\begin{aligned} \|\mathcal{A}\|_{1\to 2}^2 \le L_f = \|\mathcal{A}\|_{q_u \to 2}^2 \le \|\mathcal{A}\|_{2\to 2}^2, \\ \|\mathcal{A}\|_{1\to\infty} \le L_F = \|\mathcal{A}\|_{q_u \to p_v} \le \|\mathcal{A}\|_{2\to 2}. \end{aligned} \tag{21}$$

Now, the norm $\|\mathcal{A}\|_{2\to 2}$ itself can be bounded as follows:

[2]For simplicity, we only state the bound for bilinear $f(u,v)$.

Lemma 4.1. *One has*

$$\|\mathcal{A}\|_{2\to 2} \le \sqrt{2n+1} \cdot \|F_{2n}[y]_{-n}^n\|_\infty.$$

The proof of this lemma appears in the supplementary material. Together with (19), Lemma 4.1 results in the following

Proposition 4.1. *Solving* (Con-LS) *or* (Pen-LS) *by Algorithm 1 with proximal setup as described above, one has*

$$\varepsilon(T) = \tilde{O}\left(\frac{r^2 \|F_{2n}[y]_{-n}^n\|_\infty^2}{T^2}\right). \tag{22}$$

Similarly, solving (Con-UF)*,* (Pen-UF)*,* (Con-LS*)*, or* (Pen-LS*) *by Algorithm 2 with proximal setup as described above,*

$$\overline{\varepsilon}(T) = \tilde{O}\left(\frac{r \|F_{2n}[y]_{-n}^n\|_\infty}{T}\right). \tag{23}$$

Discussion: comparison of setups. Note that Proposition 4.1 gives the same upper bound on the accuracy $\varepsilon(T)$ irrespectively of the chosen proximal setup. This is because we used the operator norm $\|\mathcal{A}\|_{2\to 2}$ as an upper bound for L_f and $\sqrt{L_F}$ while these quantities are in fact equal to $\|\mathcal{A}\|_{1\to 2}$ or $\|\mathcal{A}\|_{1\to\infty} \le \|\mathcal{A}\|_{1\to 2}$ when one uses the "geometry-adapted" ℓ_1-setup in at least one of the variables. For a general linear operator $\mathcal{A}$ on $\mathbb{C}^{n+1}$ the gaps between $\|\mathcal{A}\|_{2\to 2}$ and the latter norms can be as large as $\sqrt{n+1}$ or $n+1$, hence one might expect the bound of Proposition 4.1 to be loose. However, intuitively $\mathcal{A}$ is "almost" a diagonal operator – it would as such is we worked with the *circular* convolution. Hence, we can expect its various $\|\cdot\|_{q\to p}$ norms in (21) to be mutually close (in the case $\mathcal{A} = \mathrm{diag}(a)$ they all coincide with $\|a\|_\infty$). This heuristic observation can be made precise:

Proposition 4.2. *Assume that $\sigma = 0$, and $x \in \mathbb{C}(\mathbb{Z})$ is $(n+1)$-periodic: $x_\tau = x_{\tau-n-1}, \tau \in \mathbb{Z}$. Then, one has*

$$\|\mathcal{A}\|_{1\to\infty} = \sqrt{n+1}\|F_n[x]\|_\infty.$$

4.2. Statistical Accuracy and Complexity Bounds

In this section, we first characterize the *statistical accuracy* of adaptive recovery procedures, defined as the absolute accuracy ε_* sufficient for the corresponding approximate estimator $\tilde{\varphi}$ to admit the same, up to a constant factor, theoretical risk bound as the exact estimator $\widehat{\varphi}$. The exact meaning of "risk bound" here depends on the estimator in consideration: for uniform-fit estimators it is the bound on the pointwise loss that was proved in (Harchaoui et al., 2015b), and for least-squares estimators it is the bound on the ℓ_2-loss proved in (Ostrovsky et al., 2016). The next two results state that statistical accuracy, defined in this sense, can be chosen as σr for uniform-fit procedures, and $\sigma^2 r^2$ for least-squares procedures. The arguments, provided in the supplementary material, closely follow those in (Harchaoui et al., 2015b) and (Ostrovsky et al., 2016).

Theorem 4.3. *An ε_*-accurate solution $\tilde{\varphi}$ to* (Con-UF) *with $\overline{r} = r$, or to* (Pen-UF) *with $\lambda = 16\sigma\sqrt{(n+1)\left(1+\log\left(\frac{n+1}{\delta}\right)\right)}$, in both cases with $\varepsilon_* = O(\sigma r)$, with prob. $\geq 1-\delta$ satisfies*

$$|x_n - [\tilde{\varphi} * y]_n| \leq \frac{C\sigma r^2\sqrt{1+\log\left(\frac{n+1}{\delta}\right)}}{\sqrt{n+1}}. \tag{24}$$

While Theorem 4.3 controls the pointwise loss for uniform-fit estimators, the next theorem controls the ℓ_2-loss for least-squares estimators. To state it, we recall that a linear subspace $\mathcal{S}$ of $\mathbb{C}(\mathbb{Z})$ is called *shift-invariant* if it is an invariant subspace of the lag operator Δ: $[\Delta x]_\tau = x_{\tau-1}$ on $\mathbb{C}(\mathbb{Z})$.

Theorem 4.4. *Assume that x belongs to a shift-invariant subspace $\mathcal{S}$ with $\dim(\mathcal{S}) \leq n$. Then, an ε_*-accurate solution $\tilde{\varphi}$ to* (Con-LS) *with $\overline{r} = r$ or to* (Pen-LS) *with $\lambda = 8\sqrt{2}\sigma^2\sqrt{n+1}\left(2+\log\left(\frac{8(n+1)}{\delta}\right)\right)$, in all cases with $\varepsilon_* = O(\sigma^2 r^2)$, with prob. $\geq 1-\delta$ satisfies*

$$\|x - \tilde{\varphi} * y\|_{n,2} \leq \frac{C\sigma\left(r\sqrt{1+\log\left(\frac{n+1}{\delta}\right)} + \sqrt{\dim(\mathcal{S})}\right)}{\sqrt{n+1}}. \tag{25}$$

Complexity Bound. Combining Theorems 4.3–4.4 with Proposition 4.1, we arrive at the following conclusion: for both classes of estimators, the number of iterations T_* of the suitable first-order algorithm (Algorithm 1 for the least-squares estimators and Algorithm 2 for the uniform-fit ones) that guarantees accuracy ε_*, with high probability satisfies

$$T_* = \tilde{O}\left(\|F_{2n}[y]_{-n}^n\|_\infty/\sigma\right) = \tilde{O}\left(\text{PSNR}+1\right). \tag{26}$$

Here, $\text{PSNR} := \|F_{2n}[x]_{-n}^n\|_\infty/\sigma$ is the peak signal-to-noise ratio in the Fourier domain, and we used the unitary invariance of the complex Gaussian distribution. Moreover, if it is known that the signal is sparse in the Fourier domain, that is, $\mathcal{S}$ is spanned by s complex exponentials $e^{i\omega_k\tau}$ with frequencies on the grid, $\omega_k \in \left\{\frac{2\pi j}{n+1},\ j\in\mathbb{Z}\right\}$, we can write

$$\text{PSNR} = O(\text{SNR}\sqrt{s}) \tag{27}$$

where $\text{SNR} = \|x\|_{n,2}/\sigma$ is the usual signal-to-noise ratio.

Discussion: different ways of solving (Con-LS)**.** Note that Algorithm 2 can be used to solve problems (Con-LS*) and (Pen-LS*) with non-squared residual by reducing them to (composite) saddle-point problems as shown in Sec. 3. Hence, when solving (Con-LS) we have two alternatives: either to solve it directly with Algorithm 1, or to solve instead the equivalent problem (Con-LS*) with Algorithm 2. Note that the complexity bound (26) only holds when Algorithm 1, and we can guess that this way of treating (Con-LS) is more beneficial. Indeed, whenever the optimal residual $\text{Res}_2(\widehat{\varphi})$ is strictly positive, attaining accuracy σ^2r^2 for (Con-LS) is equivalent to attaining accuracy $\varepsilon_{**} = \frac{\sigma^2r^2}{\text{Res}_2(\widehat{\varphi})}$, rather than $\varepsilon_* = \sigma r$, for (Con-LS*), where $\text{Res}_2(\widehat{\varphi})$ is the optimal residual. Using Proposition 4.1, the number of iterations of Algorithm 2 to guarantee that is

$$T_{**} = \frac{\text{Res}_2(\widehat{\varphi})}{\sigma}\cdot O(\text{PSNR}+1).$$

Potentially, this is much worse than (26) since $\text{Res}_2(\widehat{\varphi})$ is expected to scale as the ℓ_2-norm of the noise, *i.e.* $\sigma\sqrt{n+1}$.

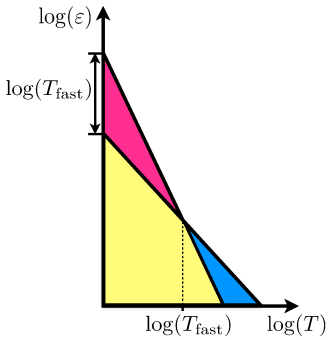

Figure 1: "Phase transition" for Algorithm 1. The different slopes correspond to (29) and (30).

One curious property of Algorithm 1 in the present context is its fast $O(1/T^2)$ convergence in terms of the objective of (Con-LS*). This fact, although surprizing at a first glance since the objective of (Con-LS*) is *non-smooth*, has a simple explanation. Note that in case of (Con-LS), (22) becomes

$$\text{Res}_2^2(\tilde{\varphi}) - \text{Res}_2^2(\widehat{\varphi}) = \tilde{O}\left(\frac{r^2\|F_{2n}[y]_{-n}^n\|_\infty^2}{T^2}\right). \tag{28}$$

Dividing by $\text{Res}_2(\tilde{\varphi}) + \text{Res}_2(\widehat{\varphi}) \geq 2\text{Res}_2(\widehat{\varphi})$, we obtain

$$\text{Res}_2(\tilde{\varphi}) - \text{Res}_2(\widehat{\varphi}) = \tilde{O}\left(\frac{r^2\|F_{2n}[y]_{-n}^n\|_\infty^2}{\text{Res}_2(\widehat{\varphi})T^2}\right), \tag{29}$$

i.e. $O(1/T^2)$ convergence for (Con-LS*) if $\text{Res}_2(\widehat{\varphi}) > 0$. Moreover, this bound is crucial to achieve (26), since (26) is exactly what is required for the right-hand side of (29) to be upper-bounded by $\varepsilon_{**} = \frac{\sigma^2r^2}{\text{Res}_2(\widehat{\varphi})}$.

Finally, note that for small T, the $O(1/T^2)$ bound (29) is dominated by the $O(1/T)$ bound

$$\text{Res}_2(\tilde{\varphi}) - \text{Res}_2(\widehat{\varphi}) = \tilde{O}\left(\frac{r\|F_{2n}[y]_{-n}^n\|_\infty}{T}\right), \tag{30}$$

which is obtained from (28) by putting $\text{Res}_2(\widehat{\varphi})$ into the right-hand side and taking the square root. Hence, we can expect to see the faster $O(1/T^2)$ convergence after

$$T_{\text{fast}} = \frac{r\|F_{2n}[y]_{-n}^n\|_\infty}{\text{Res}_2(\widehat{\varphi})} \tag{31}$$

iterations of Algorithm 1, as graphically shown in Fig. 1.

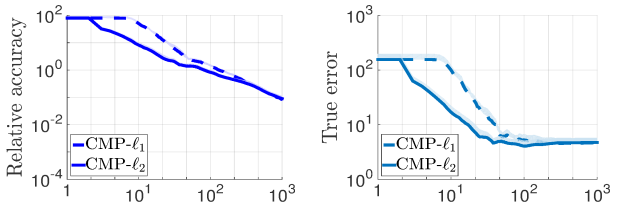

Figure 2: Relative accuracy, left, and ℓ_∞-loss $\|F_n[x - \tilde{\varphi}(T) * y]\|_\infty$, right, vs. iteration for approximate solutions to (Con-UF) by Algorithm 2 in *Coherent*-8 with SNR $= 16$.

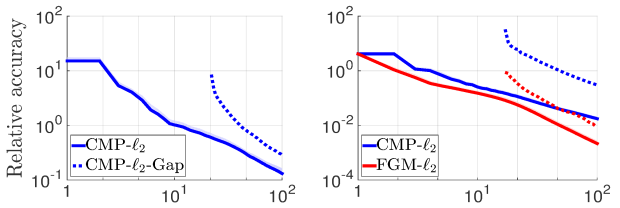

Figure 3: Relative accuracy vs. iteration for (Con-UF), left, and (Con-LS*), right, in scenario *Coherent*-4 with SNR $= 4$. Dotted: accuracy certificates, see (Nemirovski et al., 2010).

5. Experiments

In this series of experiments, our goal is to demonstrate the effectiveness of the approach and illustrate the theoretical results of Sec. 4. We estimate signals coming from an unknown shift-invariant subspace $\mathcal{S}$, implementing the following experimental protocol. First, a random signal $[x_0; ...; x_n]$ with $n = 100$ is generated according to one of the scenarios described below (s is a parameter in both scenarios). Then, x is normalized so that $\|[x]_0^n\|_2 = 1$, and corrupted by i.i.d. Gaussian noise with a chosen level of SNR $= (\sigma\sqrt{n})^{-1}$. A number of independent trials is performed to ensure the statistical significance of the results.

- In scenario *Random-s*, the signal is a harmonic oscillation with s frequencies: $x_t = \sum_{k=1}^{s} a_k e^{i\omega_k t}$. The frequencies are sampled uniformly at random on $[0, 2\pi[$, and the amplitudes uniformly on $[0, 1]$.
- In scenario *Coherent-s*, we sample s *pairs* of close frequencies. Frequencies in each pair have the same amplitude and are separated only by $\frac{0.2\pi}{n} - 0.1$ DFT bin – so that the signal violates the usual frequency separation conditions, see *e.g.* (Tang et al., 2013).

For constrained estimator we set $\overline{r} = 2\dim(\mathcal{S})$ as suggested in (Ostrovsky et al., 2016) for two-sided filters. Note that $\dim(\mathcal{S}) = s$ in *Random-s* and $\dim(\mathcal{S}) = 2s$ in *Coherent-s*.

Proof-of-Concept. In this experiment, we study estimator (Con-UF) in scenarios *Random*-16 and *Coherent*-8. We run a version of CMP (Algorithm 2) with adaptive stepsize, see (Nesterov & Nemirovski, 2013), plotting the relative accuracy of the corresponding approximate solution $\tilde{\varphi}(T)$, that is, $\varepsilon(T)$ normalized by the optimal value of the residual $\mathrm{Res}_\infty(\widehat{\varphi})$, versus T. We also trace the true estimation error as measured by the ℓ_∞-loss in the Fourier domain, $\|F_n[x - \tilde{\varphi}(T) * y]\|_\infty$. Two joint proximal setups are considered: the full ℓ_2-setup composed from the partial ℓ_2-setups, and the full ℓ_1-setup composed from the partial ℓ_1-setups. To obtain a proxy for $\widehat{\varphi}$, we recast (Con-UF) as a second-order cone problem, and run the MOSEK interior-point solver (Andersen & Andersen, 2013); note that this method is only available for small-sized problems. We show upper 95%-confidence bounds for the convergence curves.

The results of this experiment, shown in Fig. 2, can be summarized as follows. First, we see that the complexity of the optimization task grows with SNR as predicted by (23). Second, provided that the number of frequencies is the same, there is no significant difference between scenarios *Random* and *Coherent* for the computational performance of our algorithms (albeit we find *Coherent* to be slightly harder, and we only show the results for this scenario here). We also find, somewhat unexpectedly, that the ℓ_2-setup outperforms the "geometry-adapted" setup in earlier iterations; however, the performances of the two setups match in later iterations.

Overall, we find that the first 100 iterations result in 100% relative accuracy, *i.e.*. $\mathrm{Res}_\infty(\tilde{\varphi}) \le 2\mathrm{Res}_\infty(\widehat{\varphi})$. In fact, from the analysis of uniform-fit estimators in the proof of Theorem 4.3 we can derive the bound $\mathrm{Res}_\infty(\widehat{\varphi}) = \tilde{O}(\sigma r)$, implying that the conditions of Theorem 4.3 are met for $\tilde{\varphi}$. As such, we can predict that further optimization is redundant. This is empirically confirmed: the true error begins to plateau after no more than 100 iterations.

Convergence and Accuracy Certificates. Here we illustrate the convergence of FGM (Algorithm 1) and CMP (Algorithm 2), including the case of (Con-LS*) where both algorithms can be applied and thus compared. We work in the same setting as previously, but this time also study (Con-LS*) for which we compare the recommended approach via Algorithm 1 and the alternative approach via Algorithm 2 as discussed in Sec. 4.2. The results are shown in Fig. 3. We empirically observe $O(1/T)$ convergence of Algorithm 2 when solving (Con-UF), as well as $O(1/T^2)$ convergence of Algorithm 1 when solving (Con-LS*), after a certain threshold as predicted by (29)–(31). In addition to accuracy curves, we plot upper bounds on them obtained via the technique of accuracy certificates, see (Nemirovski et al., 2010) and the supplementary material. Such bounds can be used for early stopping of the algorithms once the desired accuracy has been attained.

Additional experiments are presented in the supplementary material.

Acknowledgements

The authors would like to thank Anatoli Juditsky for fruitful discussions. This work was supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025), the project Titan (CNRS-Mastodons), the project MACARON (ANR-14-CE23-0003-01), the NSF TRIPODS Award (CCF-1740551), the program "Learning in Machines and Brains" of CIFAR, and a Criteo Faculty Research Award.

References

Andersen, E. and Andersen, K. *The MOSEK optimization toolbox for MATLAB manual. Version 7.0*, 2013. `http://docs.mosek.com/7.0/toolbox/`.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Ben-Tal, A. and Nemirovski, A. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. SIAM, 2001.

Bhaskar, B., Tang, G., and Recht, B. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

Bruer, J., Tropp, J., Cevher, V., and Becker, S. Designing statistical estimators that balance sample size, risk, and computational cost. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):612–624, 2015.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pp. 272–279, 2008.

Harchaoui, Z., Juditsky, A., and Nemirovski, A. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015a.

Harchaoui, Z., Juditsky, A., Nemirovski, A., and Ostrovsky, D. Adaptive recovery of signals by convex optimization. In *Proceedings of the 28th Conference on Learning Theory (COLT '15)*, pp. 929–955, 2015b.

Haykin, S. *Adaptive Filter Theory*. Prentice Hall, 1991.

Ibragimov, I. and Khasminskii, R. *Statistical estimation. Asymptotic Theory*. Springer, 1981.

Johnstone, I. *Gaussian estimation: sequence and multiresolution models*. Unpublished manuscript, 2011.

Juditsky, A. and Nemirovski, A. Nonparametric denoising of signals with unknown local structure, I: Oracle inequalities. *Applied and Computational Harmonic Analysis*, 27 (2):157–179, 2009.

Juditsky, A. and Nemirovski, A. Nonparametric denoising of signals with unknown local structure, II: Nonparametric function recovery. *Applied and Computational Harmonic Analysis*, 29(3):354–367, 2010.

Juditsky, A. and Nemirovski, A. First-order methods for nonsmooth convex large-scale optimization, II: Utilizing problem structure. *Optimization for Machine Learning*, 30(9):149–183, 2011.

Kay, S. *Fundamentals of statistical signal processing*. Prentice Hall, 1993.

Nemirovski, A. On non-parametric estimation of functions satisfying differential inequalities. *Advances in Soviet Mathematics*, 12:7–43, 1992.

Nemirovski, A. Topics in non-parametric statistics. *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXVIII-1998*, 28:87–285, 2000.

Nemirovski, A., Onn, S., and Rothblum, U. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

Nesterov, Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Nesterov, Y. Gradient methods for minimizing composite objective functions. *Mathematical Programming*, 140(1): 125–161, 2013.

Nesterov, Y. and Nemirovski, A. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica*, 22(5): 509–575, 2013.

Ostrovsky, D., Harchaoui, Z., Juditsky, A., and Nemirovski, A. Structure-blind signal recovery. *arXiv:1607.05712v2*.

Ostrovsky, D., Harchaoui, Z., Juditsky, A., and Nemirovski, A. Structure-blind signal recovery. In *Advances in Neural Information Processing Systems*, pp. 4817–4825, 2016.

Pilanci, M. and Wainwright, M. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879, 2016.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML '12)*, pp. 449–456, 2012.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Tang, G., Bhaskar, B., and Recht, B. Near-minimax line spectral estimation. In *Proceedings of the 47th Annual Conference on Information Sciences and Systems (CISS '13)*, pp. 1–6, 2013.

Tsybakov, A. *Introduction to Nonparametric Estimation*. Springer, 2008.

Wasserman, L. *All of Nonparametric Statistics*. Springer, 2006.