JMLR: Workshop and Conference Proceedings vol 40:1–27, 2015

Adaptive Recovery of Signals by Convex Optimization

Zaid Harchaoui ZAID.HARCHAOUI@INRIA.FR
NYU, Inria

Anatoli Juditsky JUDITSKY@IMAG.FR
Univ. Grenoble Alpes, LJK

Arkadi Nemirovski NEMIROVS@ISYE.GATECH.EDU
Georgia Institute of Technology

Dmitry Ostrovsky DMITRII.OSTROVSKII@IMAG.FR
Univ. Grenoble Alpes, LJK

Abstract

We present a theoretical framework for adaptive estimation and prediction of signals of unknown structure in the presence of noise. The framework allows to address two intertwined challenges: (i) designing optimal statistical estimators; (ii) designing efficient numerical algorithms. In particular, we establish oracle inequalities for the performance of adaptive procedures, which rely upon convex optimization and thus can be efficiently implemented. As an application of the proposed approach, we consider denoising of harmonic oscillations.

Keywords: Nonparametric statistics, adaptive estimation, statistical applications of convex optimization, line spectral denoising

1. Introduction

We consider the problem of estimating a deterministic signal $x = (x_\tau)_{\tau\in\mathbb{Z}}$ from noisy observations

$$y_\tau = x_\tau + \sigma\zeta_\tau. \tag{1}$$

For convenience, we assume that both the signal and the noise are complex-valued, with $\zeta_\tau = \zeta_\tau^{(1)} + \imath\zeta_\tau^{(2)}$, where $\zeta_\tau^{(i)} \sim \mathcal{N}(0,1)$ are independent for each $\tau \in \mathbb{Z}$ and $i \in \{1,2\}$. We are first interested in pointwise estimation, that is estimating the signal x_t at the time instant t. The estimation procedure shall be local in the time domain, i.e. we are allowed to use only observations $(y_{t+\tau})_{|\tau|\le T}$ in the T-neighborhood of t for $T \in \mathbb{Z}_+$. We are also interested in a related prediction problem, where the goal is to predict the signal x_t based on the previous noisy observations $(y_{t-h-\tau})_{0\le\tau\le T}$ for $h \ge 0$.

This problem is "classical" in statistical estimation and signal processing, and as such, has received much attention. An abundant and comprehensive statistical literature considered various versions of the problem of pointwise estimation of an unknown signal x from noisy observations; see *e.g.* monographs (Ibragimov and Hasminskii, 1981; Nemirovski, 2000; Tsybakov, 2008; Wasserman, 2006) and references therein. In the traditional approach to the problem, the signal is assumed to belong to a specific function class $\mathcal{X}$, *e.g.* Hölder smoothness classes, ellipsoids corresponding to Sobolev classes, and more general Besov bodies in the context of wavelet analysis; an interested reader may refer, in particular, to (Ibragimov and Khasminskii, 1980; Pinsker, 1980; Stone, 1980; Golubev, 1987; Donoho et al., 1995; Donoho and Johnstone, 1998). The key quantity for analyzing

the problem is the worst possible expected loss that can be achieved for a signal living in $\mathcal{X}$ – the minimax risk. Explicit formulas of upper and lower bounds for the risk as functions of n, where n is the overall number of observations at hand, when $n \to \infty$, give a touchstone to check the optimality of a candidate estimation method. Indeed, if the lower and the upper bounds coincide within a constant multiplicative factor $O(1)$ or, ideally, within a factor $1 + o(1)$, the estimation methods underlying the upper bounds are treated as optimal, and the estimation problem is considered essentially solved. Unfortunately, such *descriptive* approach has crucial limitations since it requires quite restrictive structural assumptions on $\mathcal{X}$ to hold. Unless the function class of the signal is easy to describe and parameterize, the descriptive approach is difficult to apply.

Linear estimators, also commonly referred to as linear filters in signal processing, offer an alternative approach (Ibragimov and Khasminskii, 1984; Donoho and Low, 1992; Donoho, 1994). Such estimators are in most cases simple to use, and also often enjoy minimax optimality in a wide range of situations. Assume that the deterministic signal x is *a priori* known to belong to the set $\mathcal{X}$. We are interested in optimal estimation of, say, x_t in the minimax risk sense

$$R_{\mathcal{X},t}(T) = \inf_{\widehat{x}_t} \sup_{x \in \mathcal{X}} E\, \ell(x_t, \widehat{x}_t),$$

where $\ell(\cdot, \cdot)$ is a loss function (*e.g.*, squared or absolute, or indicator loss), and the infimum is taken over all estimates $\widehat{x}_t$ based on the observations at hand $(y_{t+\tau})_{-T \le \tau \le T}$. Now, consider all linear estimates of the form $\widehat{x}_t^\phi = \sum_{|\tau| \le T} \phi_\tau y_{t-\tau}$ for some vector $\phi = \phi^{(T)} = (\phi_\tau)_{-T \le \tau \le T}$, and the corresponding minimax risk

$$R^{\mathrm{lin}}_{\mathcal{X},t}(T) = \inf_{\phi} \sup_{x \in \mathcal{X}} E\, \ell(x_t, \widehat{x}_t^\phi).$$

Then $R^{\mathrm{lin}}_{\mathcal{X},t}(T) \le c R_{\mathcal{X},t}(T)$, where c is a moderate absolute constant (*e.g.*, $c \le 1.25$ when $\ell(\cdot)$ is a quadratic loss), *if only $\mathcal{X}$ is convex, compact and centrally symmetric*.

Moreover, if $\mathcal{X}$ is computationally tractable,[1] that is, if $\mathcal{X}$ is a convex set given by a finite system of inequalities $\{p_i(x) \le 0\}_{i=1,\dots,m}$ where $p_i(\cdot)$ are convex polynomials, then the minimax linear estimator and the corresponding nearly minimax risk can be efficiently computed by a convex optimization algorithm; see (Juditsky and Nemirovski, 2009b). This motivates an *operational* approach to the problem: receiving a computationally tractable $\mathcal{X}$ on an input, one computes the corresponding minimax linear estimator and the nearly minimax risk. The strength of the operational approach is clear: since computational tractability is, generally speaking, a much looser restriction than the ones imposed by the descriptive approach, we can now handle much more general sets $\mathcal{X}$. Yet, this strength comes with a price: we can no longer provide explicit expressions of the minimax risk. Instead, the minimax risk is now given as an optimal value of a convex optimization problem.

In this paper, we go a step further, by allowing $\mathcal{X}$ to be unspecified and/or not computationally tractable. The proposed approach builds on recent works where similar settings were also investigated (Juditsky and Nemirovski, 2009a, 2010). Assuming the *a priori* knowledge that $x \in \mathcal{X}$, the class of *linear* estimates proves to be essentially as rich as the class of *all* estimates under minimal restrictions on $\mathcal{X}$. Suppose now that $\mathcal{X}$ is not specified nor computationally tractable, meaning that we cannot compute the minimax linear estimate anymore. Instead, we know *a priori* that such $\mathcal{X}$ exists, and that it is "nice", *i.e.* there exists a well-performing (we specify this assumption formally in the next section) linear estimator of x_t – a *linear oracle*. Note that it is reasonable to assume

1. For rigorous exposition of computational tractability and complexity issues in convex optimization one may refer *e.g.* to Ben-Tal and Nemirovski (2001).

such prior information: the *linearity* of an oracle is not a limitation since it is "automatic" if only we know that the hypothetic $\mathcal{X}$ is convex, compact and centrally symmetric. We shall attempt to answer the following question, which is central to the paper.

> *Can we adapt to the linear oracle, namely, find a "proxy" to the minimax linear estimate with similar performance using only the observations (y_τ) around t?*

The main contribution of the paper is a step towards this direction, suggesting a partial answer to this question. Considering the case where the risk of estimation is the length of the confidence interval for the true value of x_t, we show that if a well-performing linear oracle is *invariant* in the $O(T)$-sized neighbourhood of t, then one can successfully adapt to it within a logarithmic in T factor. The adaptive procedure in question can be efficiently implemented through convex optimization, which finds the "proxy" estimate with high probability. We also provide a lower bound to show that the extra factor is, in fact, unavoidable. Besides, we present an interesting extension of the approach to the prediction problem, where the goal is to predict the signal x_t based on the previous noisy observations $(y_{t-h-\tau})_{0\le\tau\le T}$ for $h \ge 0$. We present *adaptive* versions of the proposed approaches, where the appropriate window T is tuned in a data-driven manner, using a similar scheme to the so-called Lepski adaptation in nonparametric estimation (Lepskii, 1991; Birgé, 2001). As an application of the obtained results, we consider the problem of denoising harmonic oscillations – solutions of an (unknown) homogeneous linear difference equation, a straightforward generalization of the problem of denoising line spectral signals (Bhaskar et al., 2013; Tang et al., 2013).

2. Problem statement

We first introduce some important notation used throughout the paper. For the reader's convenience, it is also presented, *in extenso*, in Appendix B.5.

Linear estimators. Let $\mathbb{C}(\mathbb{Z})$ be the linear space of all two-sided complex sequences, that is, $x = \{x_\tau \in \mathbb{C}, \quad \tau \in \mathbb{Z}\}$. An element $q \in \mathbb{C}(\mathbb{Z})$ with a finite number of non-vanishing elements will be called *rational*. Given a rational $q \in \mathbb{C}(\mathbb{Z})$ and observation y, as defined in (1), we associate with q a linear estimation of the t-th component x_t of the signal $x \in \mathbb{C}(\mathbb{Z})$, $t \in \mathbb{Z}$, according to

$$\widehat{x}_t = [q * y]_t := \sum_{\tau\in\mathbb{Z}} q_\tau y_{t-\tau}.$$

The smallest integer T such that $q_\tau = 0$ whenever $|\tau| > T$ is called the *order* of the *estimator* q (denoted $\mathrm{ord}(q)$); the estimator of order T has at most $2T+1$ non-zero entries. Note that $\widehat{x}_t$ is nothing but a kernel estimate over the discrete grid $\mathbb{Z}$ with a finitely supported kernel q.

We consider the following classification of linear estimators of order T:

- bilateral estimator $\phi \in \mathbb{C}_T(\mathbb{Z}) = \{q \in \mathbb{C}(\mathbb{Z}) : \mathrm{ord}(q) \le T\}$; in other words, in order to build the estimation $[\phi * y]_t$ of x_t one is allowed to use the bilateral observations y_τ, $t - T \le \tau \le t + T$.
- h-predictive causal estimator $\phi \in \mathbb{C}_T^h(\mathbb{Z}) = \{q \in \mathbb{C}(\mathbb{Z}) : q_\tau = 0 \text{ if } \tau \notin [h, T + h]\}$ for given $h, T \ge 0$; the estimation $[\phi * y]_t$ of x_t is based on observations y_τ, $t - h - T \le \tau \le t - h$ "on the left" of t.

Note that the terminology we use here has direct signal processing counterparts: what we refer to as bilateral estimation corresponds to linear interpolation; h-predictive estimation – to linear filtering (when $h = 0$) and prediction (when $h > 0$).

It is convenient to identify an estimator q with the finite Laurent sum $q(z) = \sum_j q_j z^j$. Note that the convolution $p * q$ of two estimators corresponds to the product $p(z)q(z)$, and therefore $\mathrm{ord}(p * q) \leq \mathrm{ord}(p) + \mathrm{ord}(q)$. If we denote Δ the right-shift operator on $\mathbb{C}(\mathbb{Z})$, $[\Delta x]_t = x_{t-1}$ (and its inverse – the right-shift Δ^{-1}, $[\Delta^{-1}x]_t = x_{t+1}$), the linear estimation $[q * y]_t$ with rational q may be alternatively written as $[q(\Delta)y]_t$.

Fourier transform. For any nonnegative integer T, let Γ_T be the set of complex roots of unity of degree $2T + 1$, and let $\mathbb{C}(\Gamma_T)$ be the space of all complex-valued functions on Γ_T. We define the *(symmetric and unitary) Fourier transform (FT)* operator $F_T : \mathbb{C}(\mathbb{Z}) \to \mathbb{C}(\Gamma_T)$ as

$$(F_T x)(\mu) := (2T+1)^{-1/2} \sum_{|\tau| \leq T} x_\tau \mu^\tau \left[= (2T+1)^{-1/2} x(\mu),\ x \in \mathbb{C}_T(\mathbb{Z}) \right], \quad \mu \in \Gamma_T.$$

Note that the FT inversion formula holds: $x_\tau = (2T+1)^{-1/2} \sum_{\mu \in \Gamma_T} (F_T x)(\mu) \mu^{-\tau}$ for $|\tau| \leq T$.

Spectral domain norms. Given $p \in [1, +\infty]$ and a non-negative integer T, we introduce the semi-norms on $\mathbb{C}(\mathbb{Z})$ defined by $\|x\|_{T,p} := (\sum_{|\tau| \leq T} |x_\tau|^p)^{1/p}$, with the natural interpretation for $p = +\infty$. When such notation is unambiguous, we also use $\|\cdot\|_p$ to denote the "usual" ℓ_p-norm of a finite-dimensional argument (*e.g.*for x such that $\mathrm{ord}(x) = T$, $\|x\|_p = \|x\|_{T,p}$).

The Fourier transform allows to equip $\mathbb{C}(\mathbb{Z})$ with the semi-norms associated with the standard p-norms in the frequency domain:

$$\|x\|^*_{T,p} := \|F_T x\|_p = \left(\sum_{\mu \in \Gamma_T} |(F_T x)(\mu)|^p \right)^{1/p}, \quad p \in [1, +\infty]. \tag{2}$$

Note that for ℓ_2-norms we have Parseval's inequality (see Appendix B.5 for details):

$$\langle x, y \rangle_T = \langle F_T x, F_T y \rangle, \quad \|x\|_{T,2} = \|x\|^*_{T,2}.$$

2.1. Simple signals: definition

Our goal in this section is to formulate our *a priori* assumptions on the signals x. We highlight the class of signals for which efficient pointwise estimation is possible.

Definition 1
(a) Let $T \in \mathbb{Z}_+$ and $\rho \in \mathbb{R}_+$ be fixed. We say that a linear estimator $\phi \in \mathbb{C}_T(\mathbb{Z})$ is simple with parameter $\rho \geq 1$ *if*

$$\|\phi\|_{T,2} \quad \leq \quad \rho(2T+1)^{-1/2}. \tag{3}$$

(b) Let $L \in \mathbb{Z}_+ \cup \{+\infty\}$, $t \in \mathbb{Z}$, and $\theta \in \mathbb{R}_+$. We say that $x \in \mathbb{C}(\mathbb{Z})$ is simple at t with parameters (L, T, θ, ρ)*, with notation $x \in \mathcal{S}^t_{L,T}(\theta, \rho)$, if there exists a ρ-simple estimator $\phi \in \mathbb{C}_T(\mathbb{Z})$ (*a linear oracle*) such that*

$$|x_\tau - [\phi * x]_\tau| \quad \leq \quad \sigma\theta\rho(2T+1)^{-1/2}, \tag{4}$$

for all $t-L \le \tau \le t+L$. *The class of signals which are simple with parameters* $(\infty, T, \rho, \theta)$ *for all* $T \ge 0$ *(with* ρ *and* θ *fixed independently of* T*) is called* (θ, ρ)*-parametric and denoted* $x \in \mathcal{P}(\theta, \rho)$.

A direct consequence of relations (3) and (4) is that if a signal x is simple at t with parameters (L, T, θ, ρ) then the estimation $[\phi * y]_\tau$ of x_τ with the oracle ϕ satisfies for all $\tau \in [t-L, t+L]$

$$E\left|[x - \phi * y]_\tau\right|^2 \le \frac{(2+\theta^2)\rho^2\sigma^2}{2T+1};$$

or, with probability $\ge 1-\alpha$,

$$|[x - \phi * y]_\tau| \le |x_\tau - [\phi * x]_\tau| + \sigma|[\phi * \zeta]_\tau| \le \frac{(\theta+\kappa)\,\rho\sigma}{\sqrt{2T+1}},$$

where $\kappa^2 = 2\ln[\alpha^{-1}]$ is the $(1-\alpha)$-quantile of the χ^2_2 distribution.

Remark. Smooth signals in the classical theory of nonparametric estimation are a basic example of simple signals. Specifically, consider the problem of estimating a smooth function $f:\ [0,1] \to \mathbb{R}$ from noisy observations

$$y_i = f\left(i/n\right) + \sigma\xi_i, \quad i = 1, ..., n, \tag{5}$$

where (ξ_i) is a sequence of i.i.d. Gaussian random variables, $\xi_1 \sim \mathcal{N}(0,1)$. The classical *kernel estimator* $\widehat{f}_t$ of $f(t)$ with the bandwidth $h \le \min\{t, 1-t\}$ in this problem is

$$\widehat{f}(t) = \sum_{i=1}^{n} K_h\left(t - \frac{i}{n}\right) y_i,$$

where $K_h(t) = \frac{1}{nh}K\left(t/h\right)$, and $K(t):\ [0,1] \to \mathbb{R}$ is a two-sided kernel such that $\int_0^1 K(t)dt = 1$ and $\int_0^1 K^2(t)dt = \rho^2 < \infty$. Let $x_i = f(i/n)$, $i = 1, ..., n$, be the discretization of f, and let $T = nh$. Then, the kernel estimator above for $T+1 \le i \le n-T$ can be rewritten as $\widehat{x}_i = \widehat{f}(i/n) = (\phi * y)_i$, with the linear estimator $\phi \in \mathbb{C}_T(\mathbb{Z})$, $\phi_k = (2T+1)^{-1}K(k/(2T+1))$, $k = -T, ..., T$. Note that the ℓ_2-norm of the ϕ satisfies: $\|\phi\|_2 = (2T+1)^{-1}\left(\sum_{k=-T}^{T} K^2\left(\frac{k}{2T+1}\right)\right)^{1/2}$ and converges to $\rho/(\sqrt{2T+1})$ as $T \to \infty$ for continuous kernels K. Recall that if the kernel K and the bandwidth h are "properly chosen", the kernel estimator attains $\left[E_f(\widehat{x}_i - x_i)^2\right]^{1/2} \le C\sigma/\sqrt{nh} = C\sigma T^{-1/2}$ and under appropriate conditions is minimax in the model (5); see *e.g.* Stone (1980). We conclude that the signals which are discretized version of regular functions on the n-regular grid are simple, for properly chosen T and L. Another important example of simple signals will be treated in Sec. 4.

Remark. It is essential here to emphasize that the family of simple signals enjoys an *algebraic structure with a calculus*. The class of simple signals is closed with respect to a set of operations such as scaling, taking linear combinations, amplitude and frequency modulation (Juditsky and Nemirovski, 2009a). Given some basic families of signals (like harmonic oscillations, smooth signals, etc.) we can build new simple signal classes with the class parameters readily given by the calculus rules.

2.2. Efficient recovery of simple signals: limits of performance

Theorem 2 *For any $L, T \in \mathbb{Z}_{++}$, $T \geq 2$, and $1 \leq \rho \leq T^{\beta}$, $\beta < 1/4$, one can point out a family $\mathcal{F}_{L,T}(\rho)$ of signals from $\mathbb{C}(\mathbb{Z})$ such that*

(i) for each signal $x \in \mathcal{F}_{L,T}(\rho)$ there is an estimator $\phi \in \mathbb{C}_T(\mathbb{Z})$ with the norm $\|\phi\|_2 \leq \frac{\rho}{\sqrt{2T+1}}$ and such that $\|\phi * x - x\|_{L,\infty} = 0$ (in other words, $F_{L,T}(\rho)$ is a subset of the set $\mathcal{S}^0_{L,T}(0,\rho)$ of simple signals at $t = 0$);

(ii) there is an absolute constant $c_1 > 0$ such that for any estimate $\widehat{x}_0$ of x_0 from observations (y_τ), $-\infty \leq \tau \leq \infty$, it holds

$$\sup_{x \in \mathcal{F}(T,\rho)} P\left(|\widehat{x}_0 - x_0| \geq c_1 \sigma \rho^2 \sqrt{\frac{(1-4\beta)\ln T}{L+T}}\right) \geq 1/8.$$

3. Adaptive recovery procedures

We first outline the principal ingredients and some basic facts that underline the proposed approach, before presenting the main theoretical results (see Appendix B for all the proofs). Let us consider the problem of estimation of the value x_0 of the signal at the instant $t = 0$ given observations (y_τ), $-2T \leq \tau \leq 2T$.

Backbone assumption. Let us assume from now on that *there is in the nature an (oracle) estimator* $\varphi \in \mathbb{C}_T(\mathbb{Z})$ satisfying for some $\nu > 0$

$$\|\varphi\|^*_{T,1} \leq \nu, \tag{6}$$

and such that for all $\tau \in [t - T, t + T]$ and $\vartheta \geq 0$,

$$|x_\tau - [\varphi * x]_\tau| \leq \vartheta. \tag{7}$$

The above assumption about φ has several important consequences, which we gather in the following facts (see Appendix B.4 for the proofs). To streamline the notation, we denote $O_T(1)$ the logarithmic in T multipliers which do not depend on other problem parameters.

Facts.

$$\text{[Small bias in frequency dom.]} \quad \|x - \varphi * x\|^*_{T,\infty} = \|F_T(x - \varphi * x)\|_\infty \leq O_T(1)\sqrt{T}\vartheta.$$

$$\text{[Bounded residuals]} \quad \|y - \varphi * y\|^*_{T,\infty} \leq O_T(1)\sqrt{T}(\sigma\nu + \vartheta). \qquad \text{(w.h.p.)}$$

Idea. The construction scheme of the adaptive recovery can be informally summarized as follows:

> *assuming that there exists a recovery ϕ satisfying (6) and (7), in order to "mimic" φ, given the data ν and ϑ, one can search for an estimator, say $\widehat{\varphi}$, with the* ℓ_1-norm of the Fourier transform $F_T\widehat{\varphi}$ bounded by ν *and with the* frequency-domain residual $F_T(y - \widehat{\varphi} * y)$ bounded by
>
> $$O_T(1)\sqrt{T}(\sigma\nu + \vartheta).$$

Now, let $\widehat{x}_0 = [\widehat{\varphi} * y]_0$ be an adaptive linear recovery of x_0, with the estimator $\widehat{\varphi} \in \mathbb{C}_T(\mathbb{Z})$ "adjusted to the observed signal y" with the aforementioned properties, namely, such that

$$\begin{array}{lrcl} (a) & \|\widehat{\varphi}\|^*_{T,1} & \leq & \nu, \\ (b) & \|y - \widehat{\varphi} * y\|^*_{T,\infty} & \leq & O_T(1)\sqrt{T}(\sigma\nu + \vartheta); \end{array} \tag{8}$$

note that such $\widehat{\varphi}$ exists with high probability – by our assumption, φ satisfies the premises of (8).

Sketch of analysis. Let us see how one can bound the estimation error $\widehat{x}_0 = [\widehat{\varphi} * y]_0$ of x_0. We start with the standard bias-variance decomposition of the estimation error:

$$x_0 - \widehat{x}_0 = x_0 - [\widehat{\varphi} * y]_0 = [(1 - \widehat{\varphi}) * x]_0 - \sigma[\widehat{\varphi} * \zeta]_0 \tag{9}$$

where we use the notation $1 * s = s$ for the identity operator.

There is a simple way to bound the stochastic component $[\widehat{\varphi}*\zeta]_0$ of the error when $\widehat{\varphi}$ (depending on y and thus on ζ) satisfies (8.a). Observe that due to the Parseval identity (we denote ζ_{-T}^{T} the vector with components ζ_τ, $-T \le \tau \le T$)

$$[\widehat{\varphi} * \zeta]_0 = \left\langle \widehat{\varphi}, \overline{\zeta_{-T}^{T}} \right\rangle = \left\langle F_T\widehat{\varphi}, \overline{F_T\zeta} \right\rangle,$$

so that

$$|[\widehat{\varphi} * \zeta]_0| \le \|F_T\widehat{\varphi}\|_1 \|\overline{F_T\zeta}\|_\infty = \|\widehat{\varphi}\|_{T,1}^* \|\zeta\|_{T,\infty}^*.$$

Using that $\|\zeta\|_{T,\infty}^* \le O_T(1)$, we have

$$|\sigma[\widehat{\varphi} * \zeta]_0| \le O_T(1)\sigma\nu \tag{10}$$

with overwhelming probability. The control of the first (bias) term of (9) is more involved; we proceed as follows. Using φ we decompose

$$[(1 - \widehat{\varphi}) * x]_0 = [\varphi * (1 - \widehat{\varphi}) * x]_0 + [(1 - \varphi) * (1 - \widehat{\varphi}) * x]_0. \tag{11}$$

Due to the commutativity of the convolution,

$$\begin{array}{rcl} |[(1 - \varphi) * (1 - \widehat{\varphi}) * x]_0| & = & |[(1 - \widehat{\varphi}) * (1 - \varphi) * x]_0| \\ & \le & \|1 + \widehat{\varphi}\|_{T,1} \|(1 - \varphi) * x\|_{T,\infty} \le O_T(1)(1 + \sqrt{T}\nu)\vartheta \end{array}$$

by (7). On the other hand, when moving to the frequency domain (Fourier domain), we write

$$\begin{array}{rcl} |[\varphi * (1 - \widehat{\varphi}) * x]_0| & = & \left\langle \varphi, \overline{[(1 - \widehat{\varphi}) * x]_{-T}^{T}} \right\rangle \le \left\langle F_T\varphi, \overline{F_T((1 - \widehat{\varphi}) * x)} \right\rangle \\ & \le & \|\varphi\|_{T,1}^* \|1 - \widehat{\varphi}) * x\|_{T,\infty}^* \le \nu \|1 - \widehat{\varphi}) * x\|_{T,\infty}^* \end{array}$$

However, the frequency domain bias of the adaptive estimator is bounded with high probability, due to (8.b):

$$\|(1 - \widehat{\varphi}) * x\|_{T,\infty}^* \le \|(1 - \widehat{\varphi}) * y\|_{T,\infty}^* + \sigma\|(1 - \widehat{\varphi}) * \zeta\|_{T,\infty}^* \le O_T(1)\sqrt{T}(\sigma\nu + \vartheta),$$

and

$$|[\varphi * (1 - \widehat{\varphi}) * x]_0| \le O_T(1)\sqrt{T}\nu(\sigma\nu + \vartheta).$$

Finally, when substituting the obtained bounds into (11) we get

$$|[(1 - \widehat{\varphi}) * x]_0| \le O_T(1)\sqrt{T}\nu(\sigma\nu + \vartheta),$$

which, along with (10), leads with high probability to

$$|x_0 - \widehat{x}_0| \le O_T(1)(1 + \sqrt{T}\nu)(\sigma\nu + \vartheta). \tag{12}$$

Role of the oracle estimator. Assuming the existence of an oracle estimator φ satisfying (6) and (7), we are able to find an (adaptive) estimator whose estimation error with high probability does not exceed $O_T(1)(1+\sqrt{T}\nu)(\sigma\nu+\vartheta)$. Note that the role of the oracle estimator φ is *operational* in the above construction. It allows an easy control of the stochastic term of the error of adaptive estimation (cf. (10)). Furthermore, it also serves as a mirror between resp. the frequency domain and the time domain (cf. (11)). Hence, the error of the adaptive estimation is small if its frequency-domain residual is small. Therefore, the crucial question in the proposed approach is *when an estimator satisfying (6) and (7) exists*?

Auto-convolution of simple estimators. A partial answer to this question is provided by the following proposition.

Proposition 3 *Let $\phi \in \mathbb{C}(\mathbb{Z})$ be a simple estimator corresponding to $x \in \mathcal{S}^t_{L,T}(\theta,\rho)$. $L > T$, and let $\varphi = \phi * \phi$ with $\mathrm{ord}(\phi) \le T$. Then*

(i) the ℓ_1-norm of φ in the frequency domain is small: $\|\varphi\|^*_{2T,1} \le \sqrt{2}\rho^2(2T+1)^{-1/2}$.

(ii) φ reproduces x with small bias in the $(L-T)$-neighborhood of t, namely,

$$\|x_\tau - [\varphi * x]_\tau\| \le \sigma\theta\rho(1+\rho)(2T+1)^{-1/2}, \quad \forall |t-\tau| \le L-T.$$

Let us suppose that the filter ϕ is simple, corresponding to the simple signal x with parameters $(3T, T, \theta, \rho)$. When choosing φ to be the auto-convolution of ϕ, $\varphi \in \mathbb{C}_{2T}(\mathbb{Z})$, by Proposition 3 we get $\nu \le O(1)\frac{\rho^2}{\sqrt{2T+1}}$, and $\vartheta \le \frac{\sigma\theta\rho(1+\rho)}{\sqrt{2T+1}}$. As a result, for this choice of φ, we obtain from (12) that the estimator $\widehat{x}_0 = [\widehat{\varphi} * y]_0$ satisfies with high probability

$$|\widehat{x}_0 - x_0| \le O_T(1)\frac{\sigma\rho^4(1+\theta)}{\sqrt{2T+1}}.$$

3.1. Accuracy bounds for adaptive recovery

We start with the following assumption:

Assumption 1 Let $t \in \mathbb{Z}$ be fixed. We assume the existence of the "oracle" linear estimator $\phi \in \mathbb{C}_T(\mathbb{Z})$ such that

(a) ϕ is of small ℓ_1-norm in the frequency domain: $\|\phi\|^*_{T,1} \le \varrho(2T+1)^{-1/2}$;

(b) signal $x \in \mathbb{C}(\mathbb{Z})$ satisfies[2] $\|\Delta^{-t}(x - \phi * x)\|_{T,\infty} \le \vartheta$,
in other words, the bias of ϕ as applied to the signal x is uniformly bounded in the T-neighbourhood $[t-T, t+T]$ of t.

Note that Assumption 1.a implies (by the Parseval equality) that $\|\phi\|_{T,2} = \|\phi\|^*_{T,2} \le \|\phi\|^*_{T,1} \le \varrho(2T+1)^{-1/2}$, thus the estimator $[\phi * y]_\tau$ of x_τ, $t - T \le \tau \le t + T$, satisfies

$$E([x - \phi * y]_\tau)^2 \le \vartheta^2 + \frac{2\varrho^2\sigma^2}{2T+1},$$

or, with probability $\ge 1-\alpha$,

$$|[x - \phi * y]_\tau| \le |x_\tau - [\phi * x]_\tau| + \sigma|[\phi * \zeta]_\tau| \le \vartheta + \frac{\varrho\sigma\sqrt{2\ln[\alpha^{-1}]}}{\sqrt{2T+1}}.$$

2. Recall $\left[\Delta^{-t}x\right]_0 = x_t$, so that $\Delta^{-t}\left(z_{-T}^{T}\right) = z_{t-T}^{t+T}$.

Algorithm A1. Given setup parameters $T \in \mathbb{Z}_{++}$, $\vartheta > 0$ and $\kappa > 0$, and observations $(y_\tau)_{t-2T \le \tau \le t+2T}$, find an optimal solution $(\widehat{\varphi}; \widehat{r})$ of the optimization problem

$$\begin{array}{ll} \min r, & \text{subject to} \\ & \|\Delta^{-t}(y - \varphi * y)\|^*_{T,\infty} \le \sigma(1+r)\kappa + \vartheta\sqrt{(2T+1)}, \\ & \|\varphi\|^*_{T,1} \le r(2T+1)^{-1/2}, \quad \varphi \in \mathbb{C}_T(\mathbb{Z}). \end{array} \tag{13}$$

Then build the estimate $\widehat{x}[T, y]$ of x_t according to $\widehat{x}[T, y] = [\widehat{\varphi} * y]_t$.

Remark. The optimization problem (13) is a Second-Order Conic Programming problem (SOCP), that can be solved to high-accuracy in $O(T^3)$ in time; see Appendix A.1 for details.

Proposition 4 *Suppose that Assumption 1 holds, and that Algorithm A1 is initialized with ϑ and $\kappa = 2\sqrt{\ln[2T+1]} + u$, $u > 0$. Then with probability at least $1 - e^{-u^2/2}$ one has*

$$|x_t - \widehat{x}[T, y]| \quad \le \quad (1 + 2\varrho)\left[(2\sqrt{\ln[2T+1]} + u)\frac{\sigma(1+\varrho)}{\sqrt{2T+1}} + \vartheta\right]. \tag{14}$$

Theorem 5 *Suppose that the signal x is T_*-simple, specifically, $x \in \mathcal{S}^t_{3T_*,T_*}(\theta, \rho)$ for a given $T_* \in \mathbb{Z}_{++}$, $\rho \ge 1$, $\theta > 0$. Then the estimate $\widehat{x}[2T_*, y]$ of x_t, yielded by Algorithm A1 with setup parameters $T = 2T_*$, $\vartheta = \frac{\sigma\theta\rho(1+\rho)}{\sqrt{2T_*+1}}$, and $u > 0$, satisfies with probability $\ge 1 - e^{-u^2/2}$:*

$$|x_t - \widehat{x}[2T_*, y]| \le (1 + 4\rho^2)\left[(2\sqrt{\ln[4T_*+1]} + u)\frac{\sigma(1+2\rho^2)}{\sqrt{4T_*+1}} + \frac{\sigma\theta\rho(1+\rho)}{\sqrt{2T_*+1}}\right]$$

Remark. Theorem 5 states that the risk of the adaptive recovery is equivalent to the risk of the ideal (oracle) estimation ϕ up to the factor $C\rho^3\sqrt{\ln T_*}$ where C is an absolute constant. This factor can be reduced to $\rho^2\sqrt{\ln T_*}$ using resampling by "duplicating" the observation sample,[3]

On the other hand, lower bound of Theorem 2 states that this factor – the "price for adaptation" – cannot be less than $c\rho\sqrt{\ln T_*}$. Note that there is a gap of $O(\rho^2)$ between the lower and the upper bound ($O(\rho)$ for the "bootstrapped" algorithm). For the time being, we do not know how to "fill in" this gap – which of these estimates of the price for adaptation is correct (if any).

Adaptation with respect to T. Suppose that we are given observations (y_τ), $t - 2T_\infty \le \tau \le t + 2T_\infty$, and we are interested in recovering x_t. Let $\varrho(T) \ge 1$, $T \ge 1$, be monotonous non-increasing function of T: for $T' \ge T$, $\varrho(T') \le \varrho(T)$. Now assume that for all $1 \le T \le T_*(x)(\le T_\infty)$ there is $\phi_T \in \mathbb{C}_T(\mathbb{Z})$ such that

$$\|\phi_T\|^*_{T,1} \le \frac{\varrho(T)}{\sqrt{2T+1}}, \quad \text{and} \quad \|[x - [\phi_T * y]_\tau\| \le \frac{\theta\sigma\varrho(T)}{\sqrt{2T+1}} \tag{15}$$

for a given $\theta > 0$. What we do not know is what is ϕ_T, and what is the best window parameter $T_*(x)$. The objective is to estimate x_t with essentially the same accuracy as if we knew the best value $T_*(x)$ of T.

3. Note that in the Gaussian setting it is always possible if the noise variance is known exactly: given a complex-valued standard Gaussian vector $\eta \in \mathbb{C}(\mathbb{Z})$, one can "split" y into two independent observations $y^{(1)} = y + \sigma\eta$ and $y^{(2)} = y - \sigma\eta$, albeit, with doubled noise variance.

Algorithm A2: we are given $T_\infty \in \mathbb{Z}_{++}$, $\theta, \sigma, u > 0$, and $\varrho(T)$, $T = 1, ..., T_\infty$. Set $\widehat{x}[0, y] = y$ and $\varepsilon(0) = \sigma\left(\sqrt{6\ln[2T_\infty + 1]} + u\right)$.

For every $T = 1, ..., T_\infty$ compute the estimation $\widehat{x}[T, y]$ of s_0 by Algorithm A1 with the setup T and $\vartheta = \vartheta(T) := \frac{\theta\sigma\varrho(T)}{\sqrt{2T+1}}$; set

$$\varepsilon(T) = \sigma(1 + 2\varrho(T))(1 + \widehat{r}(T))\frac{\sqrt{6\ln[2T_\infty + 1]} + u}{\sqrt{2T + 1}} + (1 + \varrho(T) + \widehat{r}(T))\vartheta(T)$$

where $\widehat{r}(T)$ is the r-component of the optimal solution to (13) corresponding to the setup for the current T.

We say that T, $0 \le T \le T_\infty$, is *admissible* if for all $0 \le T' < T$

$$|\widehat{x}[T', y] - \widehat{x}[T, y]| \le \varepsilon(T') + \varepsilon(T),$$

Note that admissible values of T exist, *e.g.*, $T = 0$, and that the property of a given T to be admissible is observable.

Set $\overline{x}_t = \widehat{x}[T(y), y]$ where $T(y)$ is the largest admissible T

Theorem 6 *Let $\overline{x}_t$ be the adaptive estimate yielded by Algorithm A2, the setup for the adaptive estimator being $(\sigma, \theta, \rho(\cdot), u, T_\infty)$. Suppose that the signal x satisfies (15). Under this assumption, we have with probability $\ge 1 - e^{-u^2/2}$*

$$|\overline{x}_t - x_t| \le \frac{3(1 + 2\varrho(T_*(x)))(1 + \varrho(T_*(x)))\sigma}{\sqrt{2T_*(x) + 1}}\left[\sqrt{6\ln[2T_\infty + 1]} + u + \theta\right].$$

3.2. Predictive estimation

h-predictable signals. The proposed approach can also be used to predict the value of x_t when only noisy observations on the left for $\tau \le t$ (or only on the right, for $\tau \ge t$) of t are available. Suppose that we aim at estimating the value x_t of the signal $x \in \mathbb{C}(\mathbb{Z})$ at $t \in Z$, and, for the sake of definiteness, assume that only the past observations (y_τ), $\tau \le t - h$ up to the moment $t - h$ are available for some given $h \in \mathbb{Z}_+$ (we refer to h as prediction horizon). For the reader's convenience, we describe in this section the counterpart results of the ones for estimation presented in Sec. 3, adapted to the prediction setting.

Unilateral Fourier transform. It is convenient to define here the unilateral counterpart of the Fourier transform and frequency domain semi-norms for the prediction setting. For any nonnegative integer T, let Γ_T^+ be the set of complex roots of unity of degree $T + 1$, and let $\mathbb{C}(\Gamma_T^+)$ be the space of all complex-valued functions on Γ_T^+. We define the *(unilateral) Fourier transform (FT)* operator $F_T^+ : \mathbb{C}(\mathbb{Z}) \to \mathbb{C}(\Gamma_T^+)$ as $(F_T^+ x)(\mu) := (T + 1)^{-1/2} \sum_{\tau=0}^{T} x_\tau \mu^\tau \left[= (T + 1)^{-1/2} x(\mu),\ x \in \mathbb{C}_T^0(\mathbb{Z})\right]$ with $\mu \in \Gamma_T^+$. With some notational abuse, in this section we denote $\|\cdot\|_{T,p}^*$, $p \in [1, \infty]$, spectral domain norms analogous to (2) but associated with the unilateral FT (see Appendix B.5).

Definition 7 *Let $L \in \mathbb{Z}_+ \cup \{+\infty\}$, $T, h \in \mathbb{Z}_+$, $t \in \mathbb{Z}$ and $\rho, \theta \in \mathbb{R}_+$ be fixed. We say that $x \in \mathbb{C}(\mathbb{Z})$ is h-predictable at t (or simple with respect to h-step prediction) with parameters (L, T, θ, ρ), with notation $x \in \mathcal{S}_{L,T}^{t,h}(\theta, \rho)$, if there exists an h-*predictive estimator $\phi \in \mathbb{C}_T^h(\mathbb{Z})$ *such that*

$$\|\phi\|_2 \le \rho(T + 1)^{-1/2},$$

and for all $t - L \le \tau \le t$,

$$|x_\tau - [\phi * x]_\tau| \le \sigma\theta\rho(T+1)^{-1/2}.$$

Note that signals which are simple at t with respect to h-step prediction generally are not simple in the sense of Definition 1: a simple h-predictive ϕ may not reproduce the signal x with small bias for $\tau > 0$. We now present the following straightforward counterpart of Proposition 3.

Proposition 8 *Let for* $T, h \in \mathbb{Z}_+$, $L \ge T + h$, $\phi \in \mathbb{C}_T^h(\mathbb{Z})$ *be an oracle for* $x \in \mathcal{S}_{L,T}^{t,h}(\theta,\rho)$. *Then there exists an estimator* $\varphi \in \mathbb{C}_{2T}^{2h}(\mathbb{Z})$ *(for instance, the autoconvolution* $\varphi = \phi * \phi \in \mathbb{C}_{2T}^{2h}(\mathbb{Z})$*) which satisfies:*

$$\|\Delta^{-2h}\varphi\|_{2T,1}^* \le \sqrt{2}\rho^2(T+1)^{-1/2}$$

and reproduces the signal with small bias in the $L-(T+h)$*-unilateral neighbourhood of* t*:*

$$|x_\tau - [\varphi * x]_\tau| \le \sigma\theta\rho(1+\rho)(T+1)^{-1/2}, \quad \forall\, \tau,\ t - L + T + h \le \tau \le t$$

Predictive recovery. As in the case of bilateral estimation, we assume the existence of an oracle estimate with a small ℓ_1-norm in the frequency domain.

Assumption 2 Let $t \in \mathbb{Z}$ be fixed. There exists an "oracle" linear estimator $\phi \in \mathbb{C}_T^h(\mathbb{Z})$ such that

(a) ϕ is of small ℓ_1-norm in the frequency domain: $\|\Delta^{-h}\phi\|_{T,1}^* \le \varrho(T+1)^{-1/2}$;

(b) the signal $x \in \mathbb{C}(\mathbb{Z})$ satisfies $|x_\tau - [\phi * x]_\tau| \le \vartheta, \quad t - T - h \le \tau \le t$,
in other words, the bias of ϕ as applied to the signal x is uniformly bounded in the unilateral $(T+h)$-neighbourhood of t.

We may now formulate the prediction procedure.

Algorithm A3. Given setup parameters $T, h \in \mathbb{Z}_+$, $\vartheta > 0$, and $\kappa > 0$, and observations $(y_\tau)_{t-2T-2h \le \tau \le t-h}$, find an optimal solution $(\widehat{\varphi}; \widehat{r})$ of the optimization problem

$$\begin{array}{ll} \min r, & \text{subject to} \\ & \|\Delta^{-(t-h-T)}(y - \varphi * y)\|_{T,\infty}^* \le \sigma(1+r)\kappa + \vartheta\sqrt{T+1}, \\ & \|\Delta^{-h}\varphi\|_{T,1}^* \le r(T+1)^{-1/2}, \ \ \varphi \in \mathbb{C}_T^h(\mathbb{Z}_+). \end{array}$$

Then build the estimate $\widehat{x}_h[T,y]$ of x_t according to $\widehat{x}_h[T,y] = [\widehat{\varphi} * y]_t$.

Proposition 9 *Suppose that Assumption 2 holds, and that Algorithm A3 is initialized with* ϑ *and* $\kappa = 2\sqrt{\ln[T+h+1]} + u$, $u > 0$. *Then with probability at least* $1 - e^{-u^2/2}$ *one has*

$$|x_t - \widehat{x}_h[T,y]| \quad \le \quad (1+2\varrho)\left[(2\sqrt{\ln[T+h+1]} + u)\frac{\sigma(1+\varrho)}{\sqrt{T+1}} + \vartheta\right].$$

Propositions 8 and 9 together imply the following counterpart of Theorem 5:

Theorem 10 *Suppose that the signal* x *is* h_**-predictable at* t, *specifically,* $x \in \mathcal{S}_{t,3T_*+2h_*,T_*}^{t,h_*}(\theta,\rho)$ *for given* $T_*, h_* \in \mathbb{Z}_+$. *Then the estimate* $\widehat{x}_{2h_*}[2T_*, y]$ *of* x_t, *yielded by Algorithm A3 with setup parameters* $T = 2T_*$, $h = 2h_*$, $\vartheta = \frac{\sigma\theta\rho(1+\rho)}{\sqrt{T_*+1}}$, *and* $u > 0$, *satisfies with probability* $\ge 1 - e^{-u^2/2}$*:*

$$|x_t - \widehat{x}_{2h_*}[2T_*, y]| \le (1+4\rho^2)\left[(2\sqrt{\ln[2T_*+2h_*+1]} + u)\frac{\sigma(1+2\rho^2)}{\sqrt{2T_*+1}} + \frac{\sigma\theta\rho(1+\rho)}{\sqrt{T_*+1}}\right].$$

4. Denoising of harmonic oscillations: discussion

Smooth signals can be seen as a particular case of signals which are solutions to difference equations $p(\Delta)x_t = \xi_t$, where $p(z)$ is a polynomial of finite degree, and $\xi \in \mathbb{C}(\mathbb{Z})$ is a bounded perturbation. In fact, as shown in (Juditsky and Nemirovski, 2009a, Proposition 10), solutions of difference equations are simple, in the sense of Definition 1, for *any* finite-degree polynomial $p(\cdot)$. In particular, the set of solutions of all homogeneous linear difference equations with $\deg(p(\cdot)) = m$ is contained in the parametric class $\mathcal{P}(0,\rho)$ with $\rho \le \rho(m) = 2^{3m/2}\sqrt{2m-1}$.

A much better bound for the constant ρ is available for the parametric class $\mathcal{H}_d(\omega)$ of *harmonic oscillations* – solutions to homogeneous difference equations $p(\Delta)x = 0$ with polynomials p with roots on the unit circle. The class $\mathcal{H}_d(\omega)$ with $\omega = \{\omega_1, ..., \omega_d\} \in \mathbb{R}^d$ is comprised of all complex-valued sequences of the form

$$x_\tau = \sum_{k=1}^{d} p_k(\tau) e^{\imath\omega_k\tau}, \tag{16}$$

with algebraic polynomials p_k of degree $m_k - 1$ where m_k is the multiplicity, $\mod 2\pi$, of ω_k in ω. In particular, for all $\omega \in \mathbb{R}^d$ signals $x \in H_d(\omega)$ are simple for all $T \ge T(d)[= O(d^2\ln(2d))]$ with parameter $\rho = 3ed^{3/2}\sqrt{\ln[2d]}$, cf. (Juditsky and Nemirovski, 2013, Lemma 6.1).

The problem of recovery of signals of the form (16) can be seen then as an extension to the line spectral estimation problem in signal processing; see (Kay, 1993; Haykin, 1996; Bhaskar et al., 2013; Tang et al., 2013) and references therein. In this problem, the signal x to recover is a sum of d sinusoids with unknown frequencies and complex amplitudes (i.e. (16) with all distinct frequencies), observed according to (1) over the grid $\{1, 2, ..., n\}$. Usually one is interested in recovering x over the same grid in the ℓ_2-norm. The challenge is then to achieve a nearly-parametric estimation rate not knowing of the underlying structure of the signal.

The state-of-the-art algorithm for line spectral estimation is Atomic Soft Thresholding (AST), analyzed in (Bhaskar et al., 2013; Tang et al., 2013), which is an instance of the Lasso method Tibshirani (1996), for the dictionary of all harmonic oscillations with frequencies in $[0, 2\pi]$. In Tang et al. (2013), the authors show that, when the frequencies of individual oscillations are $O(1/n)$-separated, AST achieves a nearly parametric quadratic risk

$$\mathbb{E}\|\widehat{x} - x\|_2^2/n = O\left(d \cdot \sigma^2 \ln[n]/n\right) \tag{17}$$

with a matching lower bound. In the general case of non-separated frequencies, the bound $\mathbb{E}\|\widehat{x} - x\|_2^2/n = O\left(\|x\|_{\mathcal{A}} \cdot \sigma\sqrt{\ln[n]/n}\right)$ is proved in Bhaskar et al. (2013); here, $\|\cdot\|_{\mathcal{A}}$ stands for the atomic norm with respect to the dictionary Chandrasekaran et al. (2012). The question is whether "fast" rates (17) are still achievable in the case where the frequencies are not $O(1/n)$-separated.

Our approach can be seen as a step towards a more complete answer to this question. Indeed, applying our estimation at each point of the grid with the window $T = \lfloor n/2 \rfloor$, we achieve, using the bound $\rho(d) = 3ed^{3/2}\sqrt{\ln[2d]}$, the quadratic risk

$$\mathbb{E}\|\widehat{x} - x\|_2^2/n = O(d^6 \ln^2 d \cdot \sigma^2 \ln[n]/n).$$

An apparent drawback of this bound is, of course, its high-degree polynomial dependency on d. However, we believe that the overly pessimistic factor ρ^4 in Theorems 5, 10 can be improved, and the bound $\rho(d)$ can be tightened.

Acknowledgments

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 MSR-Inria joint centre, and the Moore-Sloan Data Science Environment at NYU.

References

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

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

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

L. Birgé. *An alternative point of view on Lepski's method*, volume 36 of *Lecture Notes-Monograph Series*, pages 113–133. JSTOR, 2001.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012. ISSN 1615-3375. doi: 10.1007/s10208-012-9135-7. URL http://dx.doi.org/10.1007/s10208-012-9135-7.

D. L. Donoho. Statistical estimation and optimal recovery. *Ann. Statist.*, 22(1):238–270, 03 1994.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3): 879–921, 06 1998. doi: 10.1214/aos/1024691081. URL http://dx.doi.org/10.1214/aos/1024691081.

D. L. Donoho and M. G. Low. Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.*, 20(2):944–970, 06 1992.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 301–369, 1995.

G. K. Golubev. Adaptive asymptotically minimax estimators of smooth signals. *Problemy Peredachi Informatsii*, 23(1):57–67, 1987.

S. Haykin. *Adaptive Filter Theory (3rd Ed.)*. Prentice-Hall, Inc., 1996.

I. Ibragimov and R. Khasminskii. On nonparametric estimation of regression. *Soviet Math. Dokl.*, 21:810–814, 1980.

I. Ibragimov and R. Khasminskii. Nonparametric estimation of the value of a linear functional in gaussian white noise. *Theor. Probab. & Appl.*, 29:1–32, 1984.

I. A. Ibragimov and R. Z. Hasminskii. *Statistical estimation. Asymptotic Theory*, volume 16 of *Applications of Mathematics*. Springer, 1981.

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

A. Juditsky and A. Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, pages 2278–2300, 2009b.

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

A. Juditsky and A. Nemirovski. On detecting harmonic oscillations. *Bernoulli Journal*, 2013. URL http://www.bernoulli-society.org/index.php/publications/bernoulli-journal/bernoulli-journal-papers.

S. M. Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

O. V. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

A. Nemirovski. Topics in non-parametric statistics. *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXVIII-1998*, 28:85, 2000.

M. S. Pinsker. Optimal filtering of square-integrable signals in gaussian noise. *Problemy Peredachi Informatsii*, 16(2):52–68, 1980.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

G. Tang, B.N. Bhaskar, and B. Recht. Near minimax line spectral estimation. In *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pages 1–6. IEEE, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

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

Appendix A. Experiments on synthetic data

We now present the behaviour of the proposed approach, as compared to competing approaches, on several synthetic data examples. The experiments presented here are mainly intended for illustration purposes; extensive experiments are beyond the scope of this paper.

A.1. Implementation

The optimization problem (13) is a well-structured convex optimization problem, namely a Second-Order Conic Programming problem (SOCP) and be solved to high-accuracy using the state-of-the-art polynomial time Interior-Point Methods (IPMs). All variables in (13) are complex, with both the number of variables and constraints being $O(T)$. Note that (13) is a dense problem since the constraint matrix implementing the Fourier transform is dense. To solve the problem, we used the state-of-the-art solver Mosek (Andersen and Andersen, 2013) with Matlab interface.

A.2. Harmonic oscillations

The signals to recover are (generalized) harmonic oscillations – members of the parametric class $\mathcal{H}_d(\omega)$ of solutions to the homogeneous linear difference equation $p^{(\omega)}(\Delta)x = 0$ with a given polynomial $p^{(\omega)}(\cdot)$ of degree d with the roots on the unit circle (see Sec. 4). Such signals are indeed elements of a convex compact symmetric set $\mathcal{X}^{(\omega)}$, which is actually a linear subspace of dimension d described by the vector ω of the frequencies of individual harmonics. Thus, we are dealing with a sparse recovery problem: the signal lies in an unknown low-dimensional subspace of dimension d of $\mathbb{C}^n$, $d \ll n$, and the challenge here is to achieve a nearly-parametric estimation rate without the knowledge of the underlying structure of the signal. Meanwhile, since we actually *know* the structure of generated signals, that is, subspace $\mathcal{X}^{(\omega)}$ or polynomial $p^{(\omega)}(\cdot)$, we may use e.g. the machinery from Donoho (1994) to construct the minimax linear oracle. While the oracle estimator "knows" the polynomial, the proposed recovery is "blindfolded" and should adapt to that unknown structure. On the optimization side, finding the minimax linear oracle amounts to solving an SOCP; see Donoho (1994); Juditsky and Nemirovski (2009b) for details.

We compare the pointwise estimation performance of the proposed approach and the minimax linear oracle (referred to, respectively, as *Adaptive Recovery* and *Linear Minimax Recovery* in all figures).

The estimation setting is as follows. We assume that $n = 200$ noisy observations of the signal over the regular grid are available. The noise level is chosen to achieve the signal-to-noise ratio (SNR) of -3 dB. The objective here is to predict the signal at the time instants $t = 101, ..., 200$ using the preceding observations, which corresponds to predictive recovery with the prediction horizon $h = 0$ in Sec. 3.2.

In Figure 1, we present results for the recovery of a signal that consists of one harmonic oscillation – the signal $s_\tau = \cos\left(\pi\tau/\sqrt{5}\right)$, $\tau = 1, ..., 200$. In Figure 2, we present results for the recovery of a signal that consists of a sum of 5 harmonic oscillations, where the frequencies were randomly picked uniformly in $[0, 2\pi]$ and amplitudes similarly.

The third example of signal consists of polynomially-modulated oscillations

$$x_\tau = \sum_{j=1}^{\ell} c_j \tau^{m_j - 1} e^{\imath\omega_j\tau}, \quad 0 \le \tau \le n = 200, \tag{18}$$

with random $(\omega_j)_{1\le j\le \ell}$, independent and uniformly distributed over $[0, 2\pi]$.

In Figure 3, we present the results of the experiment for the signal (18) generated as follows. We fix parameters $\ell = 3$ and $m_j = 3$, $1 \le j \le \ell$, and then draw ℓ i.i.d. random frequencies ω_j uniformly on $[0, 2\pi]$, with random coefficients $c_1, \dots c_\ell$ which are independent and uniformly distributed on $[-1, 1]$.

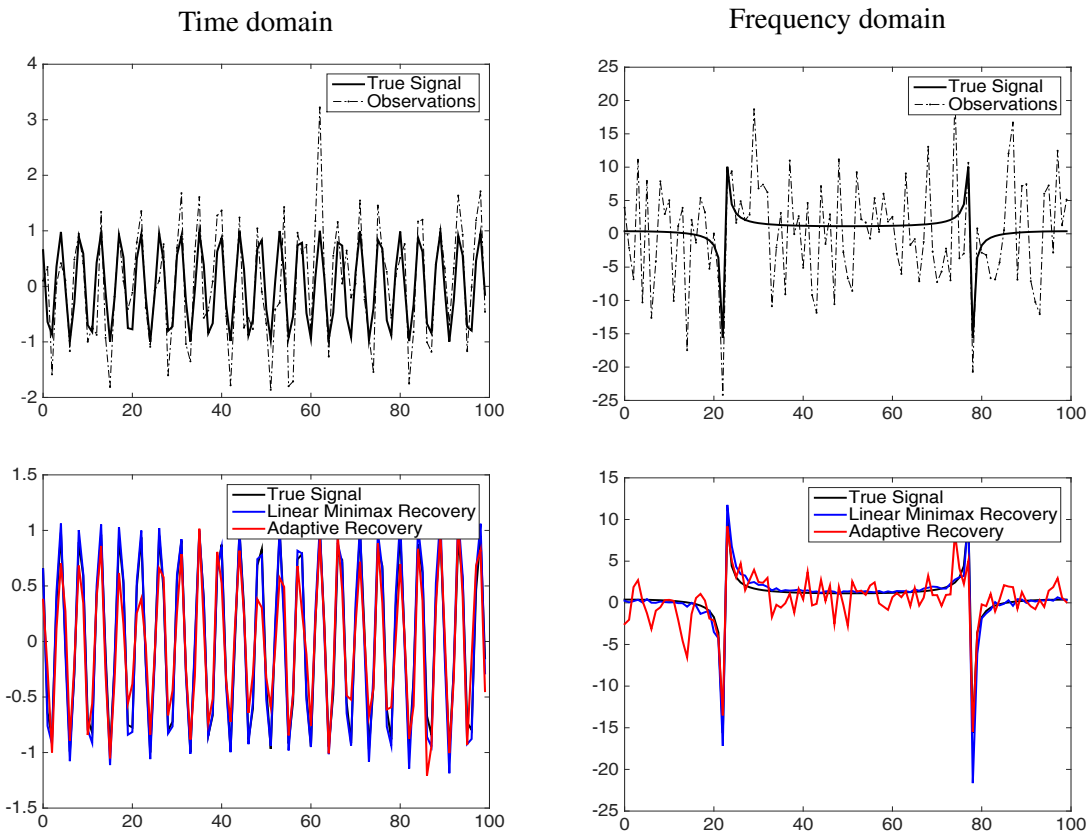

Figure 1: Recovery of one harmonic oscillation. On the left: last 100 samples of the signal, observation and recoveries in time domain, on the right: real part of corresponding signals in the frequency domain.

Appendix B. Proofs

We give here all the proofs of the theoretical results stated in the main part of the paper. In Appendix B.5, we gather the notation used both for the estimation and the prediction settings, for the reader's convenience.

B.1. Proof of Theorem 2

Let $\alpha < 1/2$, $\rho = (2T+1)^{\alpha/2}$, and $m = \lfloor \rho^2 \rfloor$. We set $\ell = \lfloor (2T+1)/m \rfloor$. Note that

$$(2T+1)^{1-\alpha} - 1 \le (2T+1)/m - 1 \le \ell \le (2T+1)/m.$$

We put

$$\beta = \sigma\sqrt{\ln[\ell/m]}. \tag{19}$$

Note that for $T \ge 2$

$$\frac{\ell}{m} \ge \frac{1}{m}\left(\frac{2T+1}{m} - 1\right) \ge \frac{2T+1}{2m^2}$$

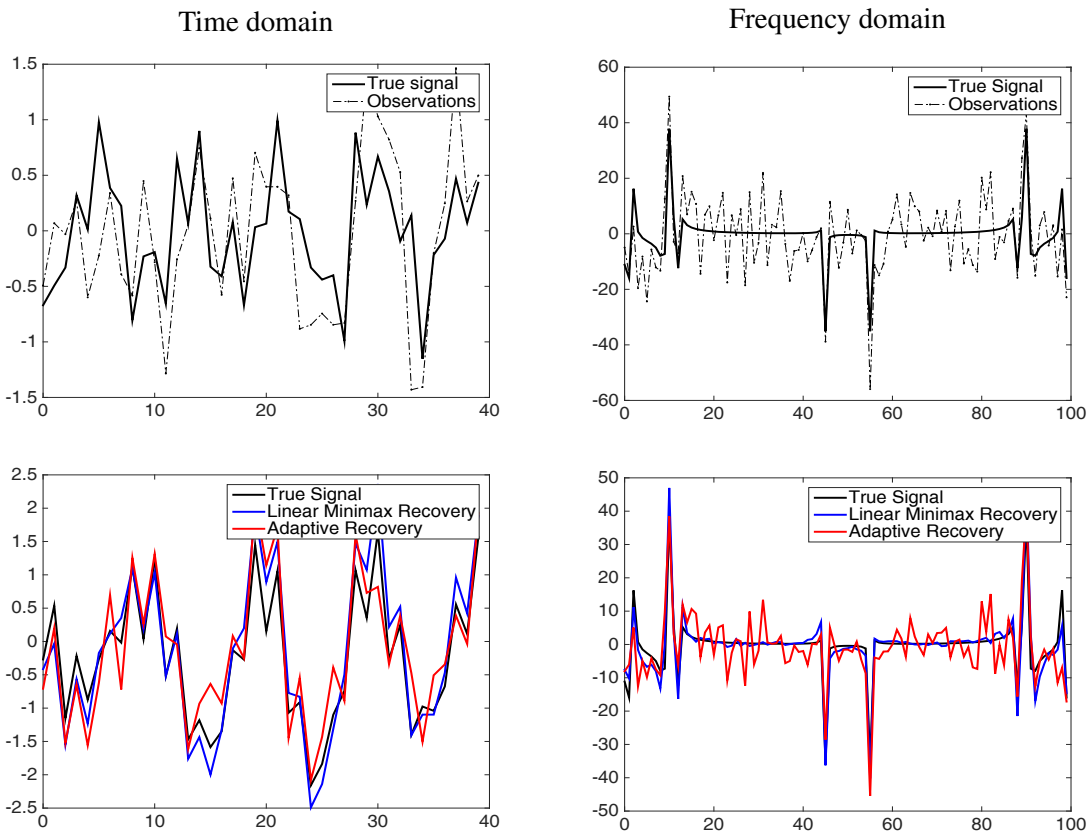

Figure 2: Recovery of a sum of 5 oscillations with random frequencies. Last 40 samples of the signal, observation and recoveries are presented in the left pane; the right pane shows the real component of the last 100 signal samples in the frequency domain.

so we have

$$\beta \geq \sigma\sqrt{(1-2\alpha)\ln[(2T+1)/2]}.$$

Now consider the following family $\mathcal{F}_{L,T}(\rho)$ of signals $x \in \mathbb{C}(\mathbb{Z})$. Let us divide the set Γ_T of complex roots of unity of degree $2T+1$ into m buckets, with ℓ elements each, as follows. Denoting $\mu = \exp(2\pi\imath/(2T+1))$, the j-th bucket is

$$\Gamma_T^{(j)} = \{\mu^{r_j} |\, r_j = \ell(j-1), \ldots, \ell j - 1\}, \quad j = 1, \ldots, m.$$

The family $\mathcal{F}_{L,T}(\rho)$ consists of the zero signal $x^{(0)} \equiv 0$, and all signals $x^{(r)} \in \mathbb{C}_{T'}(\mathbb{Z})$, $r = (r_1, \ldots, r_m)$, $T' = L+T$, of the form

$$x_k^{(r)} = \frac{\beta}{\sqrt{2T'+1}} \sum_{j=1}^{m} \mu^{r_j k}\, \mathbb{1}\{|k| \leq T'\} = \frac{\beta}{\sqrt{2T'+1}} \sum_{j=1}^{m} \exp\left(\frac{2\pi\imath r_j k}{2T+1}\right) \mathbb{1}\{|k| \leq T'\},$$
$$r_j \in \{\ell(j-1), \ldots, \ell j - 1\}.$$

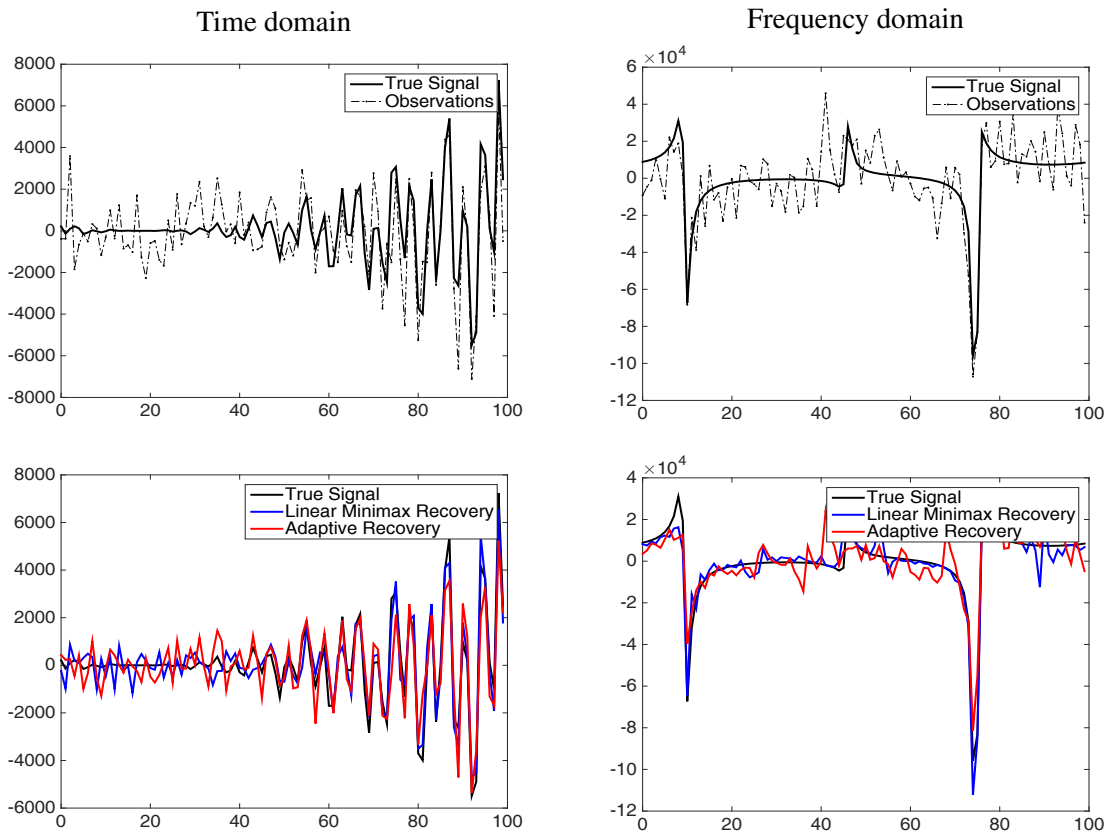

Figure 3: Recovery of amplitude-modulated oscillations. Last 100 samples of the signal, observation and recoveries are presented in the left pane (real part); the right pane shows the Fourier transform of the last 100 signal samples (real part).

1^0. Let us verify the statement (i) of the theorem. The signal $x^{(0)}$ is trivially recovered, whereas for $x^{(r)}$, $r = (r_1 \ \dots \ r_m)$, consider $\phi^{(r)} = \sum_{j=1}^m \phi^{(r_j)} \in \mathbb{C}_T(\mathbb{Z})$ with

$$\phi_k^{(r_j)} = \frac{1}{2T+1} \exp\left(\frac{2\pi \imath r_j k}{2T+1}\right) \mathbb{1}\left\{|k| \le T\right\}.$$

It is straightforward to check that the convolution $\phi^{(r)} * x^{(r)}$ exactly reproduces $x_k^{(r)}$ for $|k| \le L$. On the other hand, due to the orthogonality of $\phi^{(r_j)}$ for different r_j,

$$\left\|\phi^{(r)}\right\|_2^2 = \sum_{j=1}^m \left\|\phi^{(r_j)}\right\|_2^2 = \frac{m}{2T+1}.$$

Hence $x^{(r)} \in \mathcal{S}^0_{L,T}(0,\rho)$ as required.

2^o. We shall rely on the following lemma, which is a standard route to prove lower bounds (Tsybakov, 2008; Wasserman, 2006).

Lemma 11 *Consider the problem of testing the hypothesis $H_0 : x \equiv 0$ against $H_1 : x \in \mathcal{X}$. Assume that $\zeta \sim \pi$ is such that $\pi(\zeta \in \mathcal{X}) \geq 1 - \epsilon_\pi$. Then for any test $\mathcal{T}$ (a measurable function of observation z with values in $\{0, 1\}$ such that $\mathcal{T}(z) = 1$ implies that the zero hypothesis is rejected) satisfying*

$$\varepsilon_0(\mathcal{T}) := P_0(\mathcal{T} = 1) \leq \rho,$$

we have

$$\varepsilon_1(\mathcal{T}) := \sup_{x \in \mathcal{X}} P_x(\mathcal{T} = 0) \geq \tfrac{1}{2} - \rho E_0[L_\pi^2(y)] - \epsilon_\pi, \tag{20}$$

were $L_\pi(\cdot)$ is the Bayesian likelihood ratio:

$$L_\pi(z) = \int L(z, x) d\pi(x) = \int \frac{dP_x(z)}{dP_0(z)} d\pi(x).$$

Proof of the lemma. Let $R = \{z : \mathcal{T}(z) = 1\}$ be the rejection region of $\mathcal{T}$. We have: $P_0(R) \leq \rho$. Then for any $t \geq 0$,

$$\begin{aligned}
\varepsilon_1(\mathcal{T}) &= \sup_{x \in \mathcal{X}} \int_{R^c} dP_x(z) \geq \int_{R^c} dP_x(z) d\pi(x) - \epsilon_\pi \\
&= \int_{R^c} L(x, z) dP_0(z) d\pi(x) - \epsilon_\pi = \int_{R^c} L_\pi(z) dP_0(z) - \epsilon_\pi \\
&\geq \int_{R^c} L_\pi(z) dP_0(z) + t(P_0(A) - \rho) - \epsilon_\pi \geq \int \min[t, L_\pi(z)] dP_0(z) - t\rho - \epsilon_\pi.
\end{aligned}$$

Recall that $\min[a, b] = \frac{a+b}{2} - \frac{|a-b|}{2}$, and $E_0 L_\pi(y) = \int dP_x(z) d\pi(x) = 1$. Thus

$$\varepsilon_1(\mathcal{T}) \geq \frac{t+1}{2} - \tfrac{1}{2} E_0\{|t - L_\pi(y)|\} - t\rho - \epsilon_\pi \geq \frac{t+1}{2} - \tfrac{1}{2}[E_0\{(t - L_\pi(y))^2\}]^{1/2} - t\rho - \epsilon_\pi. \tag{21}$$

We have $E_0\{(t - L_\pi(y))^2\} = t^2 - 2t + E_0\{L_\pi(y)^2\}$, and

$$[E_0\{(t - L_\pi(y))^2\}]^{1/2} = [t^2 - 2t + E_0\{L_\pi(y)^2\}]^{1/2} \leq t\left(1 - t^{-1} + (2t^2)^{-1} E_0\{L_\pi(y)^2\}\right).$$

When substituting the above bound into (21), we obtain

$$\varepsilon_1(\mathcal{T}) \geq \tfrac{1}{2} + \frac{t}{2}\left(t^{-1} - (2t^2)^{-1} E_0\{L_\pi(y)^2\}\right) - t\rho - \epsilon_\pi = 1 - (2t)^{-1} E_0\{L_\pi(y)^2\} - t\rho - \epsilon_\pi,$$

which gives (20) for $t = (2\rho)^{-1}$. ∎

Let us consider now the following hypothesis testing problem.

> *Given an observation y, as in (1), we want to test simple hypotheses $H_0 : x = x^{(0)}$ against the composite alternative H_1 saying that x is one of $x^{(r)} \in \mathcal{F}_{L,T}(\rho)$, with $r \neq 0$.*

We are to use Lemma 11 to prove that one cannot decide between the hypotheses H_0 and H_1 with the probabilities of errors of the first and second type simultaneously not exceeding $1/8$. Let us denote

$L(y,r) = dP_r/dP_0$ the likelihood ratio, where P_r is the normal distribution of the observation $y = x^{(r)} + \sigma\zeta$ and P_0 is the distribution of the noise. We have

$$\begin{aligned} L(y,r) &= \prod_{\tau=-T'}^{\tau=T'} \exp\left(\frac{1}{2\sigma^2}(\overline{x^{(r)}}_\tau \zeta_\tau + x^{(r)}_\tau \bar{\zeta}_\tau - |x^{(r)}_\tau|^2)\right) \\ &= \exp\left(\frac{1}{2\sigma^2}(\langle \zeta, x^{(r)}\rangle + \overline{\langle \zeta, x^{(r)}\rangle} - \|x^{(r)}\|^2_{T',2})\right). \end{aligned}$$

Let us denote $X^{(r)} = F_{T'}x$, and $\varsigma = F_{T'}\zeta$. Note that $X^{(r)}$ is real by construction – we have

$$X^{(r)}_k = \begin{cases} \beta & \text{if } k = r_j, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, ς_k, $k = 0, ..., 2T'$ are independent standard complex-valued Gaussian random variables. Using the Parseval identity, we get

$$\begin{aligned} L(y,r) &= \exp\left(\frac{1}{2\sigma^2}(\langle \varsigma, X^{(r)}\rangle + \overline{\langle \varsigma, X^{(r)}\rangle} - \|X^{(r)}\|^2_2)\right) \\ &= \exp\left(\sum_{j=1}^m \frac{2\beta\eta_{r_j} - \beta^2}{2\sigma^2}\right) = \prod_{j=1}^m \exp\left(\frac{2\beta\eta_{r_j} - \beta^2}{2\sigma^2}\right), \end{aligned}$$

where $\eta_k = \Re(\varsigma_k)$.

Let now r be a random vector and let π be the distribution of r which corresponds to independent and uniformly distributed over $\{\ell(j-1), ..., \ell j - 1\}$ components r_j, $j = 1, ..., m$. Let now $L_\pi(y)E_\pi L(y,r)$ be the likelihood expectation under the a priori distribution π:

$$L_\pi(y) = \prod_{j=1}^m \frac{1}{\ell} \sum_{k=\ell(j-1)}^{\ell j-1} \exp\left(\frac{2\beta\eta_k - \beta^2}{2\sigma^2}\right).$$

Clearly, $E_0 L_\pi(y) = 1$ where the external expectation is over the noise distribution (under H_0). Let us compute $\mathbb{E}L^2_\pi(y)$. We have $E_0(L^2_\pi(y))^2 = \prod_{j=1}^m I_j$ where

$$\begin{aligned} I_j &= E_0\left[\frac{1}{\ell}\sum_{k=\ell(j-1)}^{\ell j-1} \exp\left(\frac{2\beta\eta_k - \beta^2}{2\sigma^2}\right)\right]^2 = (1-\ell) + \frac{1}{\ell^2}\sum_{k=\ell(j-1)}^{\ell j-1} E_0 \exp\left(\frac{2\beta\eta_k - \beta^2}{\sigma^2}\right) \\ &= 1 + \frac{1}{\ell}\left[\exp\left(\frac{\beta^2}{\sigma^2}\right) - 1\right] \le \exp\left(\frac{1}{\ell} e^{\frac{\beta^2}{\sigma^2}}\right). \end{aligned}$$

We conclude that

$$E_0(L^2_\pi(y))^2 \le \exp\left(\frac{m}{\ell} e^{\frac{\beta^2}{\sigma^2}}\right) \le e.$$

We now apply Lemma 11 in our setting with $\rho = 1/8$ and $\epsilon_\pi = 0$. We conclude that, for any test $\mathcal{T}$ with $\epsilon_0(\mathcal{T}) \le 1/8$, we have $\epsilon_1(\mathcal{T}) \ge \frac{1}{2} - \frac{e}{8} > 1/8$.

3^o. Now assume that there is an estimator $\widehat{x}_0$ of x_0 using the observations y, and such that with probability not exceeding $1/8$

$$|\widehat{x}_0 - x_0| \geq \frac{\beta m}{2\sqrt{2T'+1}}.$$

Note that when $x^{(r)} \in \mathcal{F}_{L,T}(\rho)$, we have $x^{(r)}(0) = \frac{\beta m}{\sqrt{2T'+1}}$ for $r \neq 0$, while $x_0^{(0)} = 0$. Let us consider the test $\widehat{\mathcal{T}}$ for distinguishing between H_0 and H_1 as in the testing problem of step 2^o as follows: $\widehat{\mathcal{T}}$ rejects H_0 if $\widehat{x}_0 > \beta m/(2\sqrt{2T'+1})$ and accepts it otherwise. Clearly, the worst probability of error such a test would be bounded by $1/8$, which is impossible, as previously seen in 2^o. We conclude that there is no estimation $\widehat{x}_0$ of x_0 using observation y and such that with probability $\leq 1/8$

$$|\widehat{x}_0 - x_0| \geq \frac{\beta m}{2\sqrt{2T'+1}} \geq \frac{\sigma\rho^2}{4}\sqrt{\frac{(1-2\alpha)\ln[(2T+1)/2]}{2T'+1}} \geq \frac{\sigma\rho^2}{4}\sqrt{\frac{(1-2\alpha)\ln T}{2(T+L)+1}}.$$

■

B.2. Proof of Proposition 4

For the sake of clarity, we shall present the proof for $t = 0$. Extension to the case of arbitrary $t \in \mathbb{Z}$ is straightforward. 1^o. We start with the following simple fact. The random variables $|F_T\zeta(\mu)|^2$, $\mu \in \Gamma_T$, are independent and identically distributed according to the χ^2_2 distribution. Thus,

$$P(|F_{2T}\zeta(\mu)| \leq q) = 1 - e^{-q/2},$$

and

$$P\left(\max_{\mu\in\Gamma(T)} |F_T\zeta(\mu)| \leq u\right) = \left(1 - e^{-u^2/2}\right)^{2T+1},$$

so that

$$P\left(\|\zeta\|^*_{T,\infty} \geq \varkappa\right) = \epsilon \ \text{ for } \ \varkappa = \sqrt{-2\ln[1-(1-\epsilon)^{1/(2T+1)}]}. \tag{22}$$

On the other hand,

$$P\left(\|\zeta\|^*_{T,\infty} \geq \sqrt{2\ln[2T+1]} + u\right) \leq (2T+1)\exp(-(u+\sqrt{2\ln[2T+1]})^2/2) \leq e^{-u^2/2}.$$

Let Δ be the left shift operator, defined for all $x \in \mathbb{C}(\mathbb{Z})$ as

$$[\Delta x]_\tau := [x]_{\tau-1}.$$

Then

$$\begin{aligned} P\left(\max_{-T\leq\tau\leq T} \|\Delta^{-\tau}\zeta\|^*_{T,\infty} \geq 2\sqrt{\ln[2T+1]} + u\right) &\leq (2T+1)^2\exp(-(u+2\sqrt{\ln[2T+1]})^2/2) \\ &\leq e^{-u^2/2}. \end{aligned} \tag{23}$$

Now, let $\Xi(u)$ be the subset of noise realizations such that

$$\max_{-T\leq\tau\leq T} \|\Delta^{-\tau}\zeta\|^*_{T,\infty} \leq 2\sqrt{\ln[2T+1]} + u, \ \ \forall\zeta \in \Xi(u);$$

by (23), $P(\Xi(u)) \geq 1 - e^{-u^2/2}$.
2^o. The second fact is that, with high probability, the oracle estimator ϕ can be computed as a solution of the convex optimization problem (13) with $\widehat{r} \leq \varrho$.

Indeed, we have

$$\|\zeta - \phi * \zeta\|^*_{T,\infty} \leq \|\zeta\|^*_{T,\infty} + \|\phi * \zeta\|^*_{T,\infty} \leq \|\zeta\|^*_{T,\infty} + \sum_{\tau=-T}^{T} |[\phi]_\tau| \, \|\Delta^{-\tau}\zeta\|^*_{T,\infty}.$$

We conclude that

$$\begin{aligned} \|\zeta - \phi * \zeta\|^*_{T,\infty} &\leq \|\zeta\|^*_{T,\infty} + \|\phi\|_{T,1} \max_{-T\leq\tau\leq T} \|\Delta^{-\tau}\zeta\|^*_{T,\infty} \\ &\leq (1 + \sqrt{2T+1}\|\phi\|_{T,2})(2\sqrt{\ln[2T+1]} + u) \\ &\leq (1+\varrho)(2\sqrt{\ln[2T+1]} + u) \end{aligned} \tag{24}$$

for all $\zeta \in \Xi(u)$. On the other hand, by Assumption 1.a, we have

$$\|x - \phi * x\|^*_{T,\infty} \leq \|x - \phi * x\|_{T,\infty}\sqrt{2T+1} \leq \vartheta\sqrt{2T+1}.$$

Along with (24) the latter inequality results in

$$\|y - \phi * y\|^*_{T,\infty} \leq \|x - \phi * x\|^*_{T,\infty} + \sigma\|\zeta - \phi * \zeta\|^*_{T,\infty} \leq \sigma(1+\varrho)(2\sqrt{\ln[2T+1]} + u) + \vartheta\sqrt{2T+1},$$

which implies the feasibility of ϕ for all $\zeta \in \Xi(u)$.
3^o. Let $\widehat{\varphi}$ be the φ-component of an optimal solution of (13). Using the bound (24), we have for all $\zeta \in \Xi(u)$

$$\begin{aligned} \|x - \widehat{\varphi} * x\|^*_{T,\infty} &\leq \|y - \widehat{\varphi} * y\|^*_{T,\infty} + \sigma\|\zeta - \widehat{\varphi} * \zeta\|^*_{T,\infty} \\ &\leq \|y - \widehat{\varphi} * y\|^*_{T,\infty} + \sigma\|\zeta\|^*_{T,\infty} + \sigma\|\widehat{\varphi}\|_{T,1} \max_{-T\leq\tau\leq T} \|\Delta^{-\tau}\zeta\|^*_{T,\infty} \\ &\leq \sigma(1+\widehat{r})(2\sqrt{\ln[2T+1]} + u) + \vartheta\sqrt{2T+1} + \sigma\|\zeta\|^*_{T,\infty} \\ &\quad + \sigma\,\widehat{r} \max_{-T\leq\tau\leq T} \|\Delta^{-\tau}\zeta\|^*_{T,\infty} \\ &\leq 2\sigma(1+\widehat{r})(2\sqrt{\ln[2T+1]} + u) + \vartheta\sqrt{2T+1}. \end{aligned} \tag{25}$$

4^o. We can now complete the proof of the proposition. We have for $\zeta \in \Xi(u)$,

$$|[\widehat{\varphi} * \zeta]_0| \leq \|\widehat{\varphi}\|^*_{T,1}\|\zeta\|^*_{T,\infty} \leq \frac{\widehat{r}}{\sqrt{2T+1}}(2\sqrt{\ln[2T+1]} + u). \tag{26}$$

Further, using notation $1 * x = x$ for the identity operator, we write

$$x - \widehat{\varphi} * x = (1 - \phi + \phi) * (x - \widehat{\varphi} * x) = (1 - \phi) * (1 - \widehat{\varphi}) * x + \phi * (1 - \widehat{\varphi}) * x = \delta^{(1)} + \delta^{(2)}.$$

Observe that

$$|[\delta^{(1)}]_0| = |[(1 - \widehat{\varphi}) * (1 - \phi) * x]_0| \leq \|1 + \widehat{\varphi}\|_{T,1}\|(1 - \phi) * x\|_{T,\infty} \leq (1 + \widehat{r})\vartheta. \tag{27}$$

On the other hand, using the bound (25), we obtain for $\delta^{(2)}$:

$$\begin{aligned}|[\delta^{(2)}]_0| &= |[\phi*(1-\widehat{\varphi})*x]_0| \le \|\phi\|^*_{T,1}\|(1-\widehat{\varphi})*x\|^*_{T,\infty}\\ &\le \frac{\varrho}{\sqrt{2T+1}}[2\sigma(1+\widehat{r})(2\sqrt{\ln[2T+1]}+u)+\vartheta\sqrt{2T+1}]\\ &= 2\varrho\sigma(1+\widehat{r})\frac{2\sqrt{\ln[2T+1]}+u}{\sqrt{2T+1}}+\varrho\vartheta. \qquad (28)\end{aligned}$$

Summing up the results of (26)– (28), we finally get

$$\begin{aligned}|x_0-[\widehat{\varphi}*y]_0| &= |x_0-[\widehat{\varphi}*x]_0|+|[\sigma\widehat{\varphi}*\zeta]_0|\\ &\le 2\varrho\sigma(1+\widehat{r})\frac{2\sqrt{\ln[2T+1]}+u}{\sqrt{2T+1}}+\varrho\vartheta+(1+\widehat{r})\vartheta+\frac{\widehat{r}\sigma}{\sqrt{2T+1}}(2\sqrt{\ln[2T+1]}+u)\\ &< (1+2\varrho)(1+\widehat{r})\sigma\frac{2\sqrt{\ln[2T+1]}+u}{\sqrt{2T+1}}+(1+\varrho+\widehat{r})\vartheta.\end{aligned}$$

Then (14) follows due to $\widehat{r}\le\varrho$. ∎

B.3. Proof of Theorem 6

We focus on the case $t=0$. Let $\Xi_\infty(u)$ be the subset if realization of the noise ζ such that

$$\max_{0\le T\le T_\infty}\max_{-T\le\tau\le T}\|\Delta^{-\tau}\zeta\|^*_{T,\infty}\le\sqrt{6\ln[2T_\infty+1]}+u,\quad\forall\zeta\in\Xi(u);$$

Then, $P(\Xi_\infty(u))\ge 1-e^{-u^2/2}$ (cf (23)).

We clearly have $|y-x_0|\le\epsilon(0)$ on $\Xi_\infty(u)$. Further, as a corollary to Proposition 4, we have that for $\zeta\in\Xi_\infty(u)$, all the estimates $\widehat{x}[T,y]$ satisfy $|\widehat{x}[T,y]-x_0|\le\varepsilon(T)$ for $0\le T\le T_*(x)$. Thus, if $0\le T'<T\le T_*(x)$,

$$|\widehat{x}[T',y]-\widehat{x}[T,y]|\le|\widehat{x}[T',y]-x_0|+|\widehat{x}[T,y]-x_0|\le\varepsilon(T')+\varepsilon(T),$$

and we conclude that $T(y)\ge T_*(x)$. On the other hand, if $T(y)>T_*(x)$,

$$|\overline{x}_0-x_0|=|\widehat{x}[T(y),y]-x_0|\le|\widehat{x}[T_*(x),y]-\widehat{x}[T(y),y]|+|\widehat{x}[T_*(x),y]-x_0|\le\varepsilon(T(y))+2\varepsilon(T_*(x)).$$

Along with the bound $\widehat{r}(T))\le\varrho(T)$ (cf. 2^o of the proof of Proposition 4) and because $\varrho(T(y))\le\varrho(T_*(x))$ when $T(y)>T_*(x)$ we finally get that

$$\begin{aligned}|\overline{x}_0-x_0| &\le 3(1+2\varrho(T_*(x)))\left[\sigma(1+\varrho(T_*(x)))\frac{\sqrt{6\ln[2T_\infty+1]}+u}{\sqrt{2T_*(x)+1}}+\vartheta(T_*(x))\right]\\ &\le \frac{3(1+2\varrho(T_*(x)))(1+\varrho(T_*(x)))\sigma}{\sqrt{2T_*(x)+1}}\left[\sqrt{6\ln[2T_\infty+1]}+u+\theta\right].\end{aligned}$$

∎

B.4. Facts and technical results

We give here the proof of the facts claimed in Sec. 3.

B.4.1. FACTS

$$\begin{aligned}\text{[Small bias in frequency dom.]}\quad & \|x-\varphi*x\|^*_{T,\infty}=\|F_T(x-\varphi*x)\|_\infty\le O_T(1)\sqrt{T}\vartheta.\\ \text{[Bounded residuals]}\quad & \|y-\varphi*y\|^*_{T,\infty}\le O_T(1)\sqrt{T}(\sigma\nu+\vartheta). \qquad \text{(w.h.p.)}\end{aligned}$$

Proof The "small bias bound" (7) implies that the bias $F_T(x-\varphi*x)$ in the frequency domain is also small:

$$\|x-\varphi*x\|^*_{T,\infty}=\|F_T(x-\varphi*x)\|_\infty\le\sqrt{2T+1}\|x-\varphi*x\|_{T,\infty}\le\sqrt{2T+1}\vartheta,$$

and the first claim is proved. Moreover, relation (6) allows to bound the amplitude of the Fourier transform $F_T(\zeta-\varphi*\zeta)$:

$$\begin{aligned}\|\zeta-\varphi*\zeta\|^*_{T,\infty}\le\|\zeta\|^*_{T,\infty}+\|\varphi*\zeta\|^*_{T,\infty}&\le(\|\varphi\|_1+1)\cdot\max_{-T\le\tau\le T}\|\Delta^{-\tau}\zeta\|^*_{T,\infty}\\ &\le(\sqrt{2T+1}\nu+1)\cdot\max_{-T\le\tau\le T}\|\Delta^{-\tau}\zeta\|^*_{T,\infty}.\end{aligned}$$

Recall that, because the Fourier transform is unitary, each of random vectors $F_T\Delta^{-\tau}\zeta$, $-T\le\tau\le T$, has i.i.d. standard complex-valued Gaussian components. Therefore, $|F_T\Delta^{-\tau}\zeta(\mu)|^2$ is distributed as χ^2_2 for each $-T\le\tau\le T$ and $\mu\in\Gamma_T$, hence $|F_T\Delta^{-\tau}\zeta(\mu)|\le u$ with probability at least $1-e^{-u^2/2}$. As a result, we have

$$\|\zeta-\varphi*\zeta\|^*_{T,\infty}\le(\sqrt{2T+1}\nu+1)\cdot(2\sqrt{\ln[2T+1]}+u)$$

with probability $\ge 1-e^{-u^2/2}$. Along with the bias bound, the latter inequality implies that the Fourier transform $F_T(y-\varphi*y)$ of the residual is also bounded with high probability ($\ge 1-e^{-u^2/2}$):

$$\begin{aligned}\|y-\varphi*y\|^*_{T,\infty}&\le\sigma\|\zeta-\varphi*\zeta\|^*_{T,\infty}+\|x-\varphi*x\|^*_{T,\infty}\\ &\le\sigma(\sqrt{2T+1}\nu+1)(2\sqrt{\ln[2T+1]}+u)+\sqrt{2T+1}\vartheta.\end{aligned}$$

For instance, when choosing $u=O(1)\sqrt{\ln[2T+1]}$, we obtain with probability at least $1-T^{-1}$:

$$\|y-\varphi*y\|^*_{2T,\infty}\le O_T(1)\sqrt{T}(\sigma\nu+\vartheta).$$

■

B.4.2. AUTO-CONVOLUTION OF SIMPLE ESTIMATORS

We give here the proof of Proposition 3.

Proof Let $\phi\in\mathbb{C}_T(\mathbb{Z})$, and let $\varphi\in\mathbb{C}_{2T}(\mathbb{Z})$ satisfy $\varphi=\phi*\phi$. Then

$$\begin{aligned}\|\varphi\|^*_1&=(4T+1)^{-1/2}\sum_{\mu\in\Gamma_{2T}}|\varphi(\mu)|=(4T+1)^{1/2}\sum_{\mu\in\Gamma_{2T}}\left(\frac{|\phi(\mu)|}{(4T+1)^{1/2}}\right)^2\\ &=(4T+1)^{1/2}\|\phi\|^{*2}_{2T,2}=(4T+1)^{1/2}\|\phi\|^2_{2T,2}=(4T+1)^{1/2}\|\phi\|^2_{T,2}. \qquad (29)\end{aligned}$$

Further, due to $1-\phi*\phi=(1+\phi)*(1-\phi)$, for all $x\in\mathbb{C}(\mathbb{Z})$ one has for all $\tau\in\mathbb{Z}$:

$$\begin{aligned}|x_\tau-[\varphi*x]_\tau| &= |[(1+\phi)*(1-\phi)*x]_\tau| = \left|\sum_{|s|\le T}[1+\phi]_s\,[x-\phi*x]_{\tau-s}\right| \\ &\le \|1+\phi\|_1\max_{|s|\le T}|[x-\phi*x]_{\tau-s}|\end{aligned}$$

Assume now that $\phi\in\mathbb{C}_T(\mathbb{Z})$ is simple in the sense of Definition 1, i.e. $\|\phi\|_2\le\rho(2T+1)^{-1/2}$. Then by (29)

$$\|\varphi\|_1^*\le(4T+1)^{1/2}\|\phi\|_{T,2}^2\le\sqrt{2}\frac{\rho^2}{\sqrt{2T+1}}.$$

Furthermore, let $x\in\mathcal{S}^t_{L,T}(\theta,\rho)$. Then $\|\phi\|_1\le\sqrt{2T+1}\|\phi\|_2\le\rho$, and due to (4)

$$|x_\tau-[\varphi*x]_\tau|\le\|1+\phi\|_1\max_{|s|\le T}|[x-\phi*x]_{\tau-s}|\le(1+\rho)\frac{\sigma\theta\rho}{\sqrt{2T+1}},$$

for all $t-L+T\le\tau\le t+L-T$. ■

B.5. Notation

We gather here *in extenso* the notation used throughout the paper.

Linear estimators. Let $\mathbb{C}(\mathbb{Z})$ be the linear space of all two-sided complex sequences

$$x=\{x_\tau\in\mathbb{C},\quad\tau\in\mathbb{Z}\}.$$

An element $q\in\mathbb{C}(\mathbb{Z})$ with finite number of non-vanishing elements will be called *rational*. Given a rational $q\in\mathbb{C}(\mathbb{Z})$ and observation y, as defined in (1), we associate with q a linear estimation of the t-th component x_t of the signal $x\in\mathbb{C}(\mathbb{Z})$, $t\in\mathbb{Z}$, according to

$$\widehat{x}_t=[q*y]_t:=\sum_{\tau\in\mathbb{Z}}q_\tau y_{t-\tau}.$$

The smallest integer T such that $q_\tau=0$ whenever $|\tau|>T$ is called the *order* of the *estimator* q (denoted $\mathrm{ord}(q)$); the estimator of order T has at most $2T+1$ non-zero entries. Note that $\widehat{x}_t$ is nothing but a kernel estimate over the discrete grid $\mathbb{Z}$ with a finitely supported kernel q.

We consider the following classification of linear estimators of order T:

- bilateral estimator $\phi\in\mathbb{C}_T(\mathbb{Z})=\{q\in\mathbb{C}(\mathbb{Z}):\mathrm{ord}(q)\le T\}$; in other words, in order to build the estimation $[\phi*y]_t$ of x_t one is allowed to use the bilateral observations y_τ, $t-T\le\tau\le t+T$.
- h-predictive causal estimator $\phi\in\mathbb{C}^h_T(\mathbb{Z})=\{q\in\mathbb{C}(\mathbb{Z}):q_\tau=0\text{ if }\tau\notin[h,T+h]\}$ for given $h,T\ge0$; the estimation $[\phi*y]_t$ of x_t is based on observations y_τ, $t-h-T\le\tau\le t-h$ "on the left" of t.

Note that the terminology we use here has obvious signal processing counterparts: what we refer to as bilateral estimation corresponds to linear interpolation, h-predictive estimation – to linear filtering (when $h = 0$) and prediction.

It is convenient to identify an estimator q with the finite Laurent sum $q(z) = \sum_j q_j z^j$. Note that the convolution $p * q$ of two estimators corresponds to the product $p(z)q(z)$, and therefore $\mathrm{ord}(p * q) \le \mathrm{ord}(p) + \mathrm{ord}(q)$. If we denote Δ the right-shift operator on $\mathbb{C}(\mathbb{Z})$, $[\Delta x]_t = x_{t-1}$ (and its inverse – the right-shift Δ^{-1}, $[\Delta^{-1}x]_t = x_{t+1}$), the linear estimation $[q * y]_t$ with rational q may be alternatively written as $[q(\Delta)y]_t$.

Fourier transform. For any nonnegative integer T, let Γ_T be the set of complex roots of unity of degree $2T + 1$, and let $\mathbb{C}(\Gamma_T)$ be the space of all complex-valued functions on Γ_T. We define the *(symmetric and unitary) Fourier transform (FT)* operator $F_T : \mathbb{C}(\mathbb{Z}) \to \mathbb{C}(\Gamma_T)$ as

$$(F_T x)(\mu) := (2T+1)^{-1/2} \sum_{|\tau| \le T} x_\tau \mu^\tau \left[= (2T+1)^{-1/2} x(\mu),\ x \in \mathbb{C}_T(\mathbb{Z}) \right], \quad \mu \in \Gamma_T.$$

We also have for $|\tau| \le T$ (inverse FT):

$$x_\tau = (2T+1)^{-1/2} \sum_{\mu \in \Gamma_T} (F_T x)(\mu) \mu^{-\tau}.$$

Spectral domain norms. Given $p \in [1, +\infty]$ and a non-negative integer T, we introduce the following semi-norms on $\mathbb{C}(\mathbb{Z})$:

$$\|x\|_{T,p} := \left(\sum_{|\tau| \le T} |x_\tau|^p \right)^{1/p}$$

with the standard interpretation for $p = +\infty$. We use also the norms $\|x\|_p = \lim_{T \to +\infty} \|x\|_{T,p}$ with values in $\mathbb{R}_+ \cup \{+\infty\}$. When such notation is unambiguous, we also use $\|\cdot\|_p$ to denote the "usual" ℓ_p-norm of a finite-dimensional arguments (e.g., for x such that $\mathrm{ord}(x) = T$, $\|x\|_p = \|x\|_{T,p}$, etc.).

The Fourier transform allows to equip $\mathbb{C}(\mathbb{Z})$ with the semi-norms associated with the standard p-norms in the frequency domain:

$$\|x\|^*_{T,p} := \|F_T x\|_p = \left(\sum_{\mu \in \Gamma_T} |(F_T x)(\mu)|^p \right)^{1/p}, \quad p \in [1, +\infty].$$

It is straightforward that, according to our definition, F_T is unitary with respect to the natural Hermitian inner product $\langle x, y \rangle_T := \sum_{|\tau| \le T} x_\tau \bar{y}_\tau$, i.e.

$$\langle x, y \rangle_T = \langle F_T x, F_T y \rangle,$$

where $\langle F_T x, F_T y \rangle = \sum_{\mu \in \Gamma_T} (F_T x)(\overline{F_T y})(\mu)$ by definition. In particular, Parseval's identity holds:

$$\|x\|_{T,2} = \|x\|^*_{T,2}.$$

Unilateral Fourier transform. For any nonnegative integer T, let Γ_T^+ be the set of complex roots of unity of degree $T+1$, and let $\mathbb{C}(\Gamma_T^+)$ be the space of all complex-valued functions on Γ_T^+. We define the *(unilateral) Fourier transform (FT)* operator $F_T^+ : \mathbb{C}(\mathbb{Z}) \to \mathbb{C}(\Gamma_T^+)$ as

$$(F_T^+ x)(\mu) := (T+1)^{-1/2} \sum_{\tau=0}^{T} x_\tau \mu^\tau \left[= (T+1)^{-1/2} x(\mu),\ x \in \mathbb{C}_T^0(\mathbb{Z}) \right], \quad \mu \in \Gamma_T^+.$$

Spectral domain norms for prediction. With a slight abuse of notation, in Sec. 3.2 we denote $\|\cdot\|_{T,p}^*$, $p \in [1,\infty]$ spectral domain norms analogous to (2) but associated with the unilateral FT:

$$\|x\|_{T,p}^* := \left\| F_T^+ x \right\|_p = \left(\sum_{\mu \in \Gamma_T^+} \left| (F_T^+ x)(\mu) \right|^p \right)^{1/p}, \quad p \in [1, +\infty].$$