Adaptive signals recovery by convex optimization

Dmitrii Ostrovskii

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA

COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques Appliquées

Arrêté ministériel : 25 mai 2016

Présentée par

DMITRII OSTROVSKII

Thèse dirigée par

Anatoli IOUDITSKI, Professeur, Université Grenoble Alpes

et codirigée par

Laurent DESBAT, Professeur, Université Grenoble Alpes
Zaid HARCHAOUI, Assistant Professor, University of Washington

préparée au sein du **Laboratoire Jean Kuntzmann** dans l'**École Doctorale Mathématiques, Sciences et Technologies de l'Information et de l'Informatique**

Reconstruction adaptative de signaux par optimisation convexe

Thèse soutenue publiquement le **11 janvier 2018**,
devant le jury composé de :

Monsieur Olivier CAPPÉ
Directeur de recherche, CNRS Délégation Île-de-France Sud, Rapporteur
Monsieur Arnak DALALYAN
Professeur, ENSAE ParisTech, Rapporteur
Madame Céline LÉVY-LEDUC
Professeur, AgroParisTech, Examinateur
Monsieur Yuri GOLUBEV
Directeur de recherche, CNRS Délégation Provence et Corse, Examinateur
Monsieur Anatoli IOUDITSKI
Professeur, Université Grenoble Alpes, Directeur de thèse
Monsieur Laurent DESBAT
Professeur, Université Grenoble Alpes, Co-directeur de thèse
Monsieur Zaid HARCHAOUI
Assistant Professor, University of Washnigton, Co-directeur de thèse
Monsieur Gabriel PEYRÉ
Directeur de recherche, CNRS Délégation Paris, Président

Acknowledgements

My deepest gratitude goes to my advisor Anatoli Juditsky. With his keen intuition, energy, wisdom, and integrity, Anatoli has been a great source of inspiration to me ever since I met him. No less important were the heartiness and gentle encouragement with which he invariably approached me during this enterprise. Apart from helping me progress towards graduation, his attitude made me learn a great deal about patience, humility, and kindness.

Another leading role in this story has been played by Zaid Harchaoui, my thesis co-advisor and navigator in the troubled waters of the graduate school. Time and again he vouched for me while forgiving my repeating missteps, and have put a tremendous effort into crafting a better researcher from as unyielding a material as I was. It cannot be overstated how lucky I am to have had such a mentor and friend. I also thank Laurent Desbat, my other co-advisor, for supporting me during the last months before the graduation.

I am greatly indebted to Yuri Golubev, my master's advisor, for introducing me to the fascinating field of mathematical statistics, and helping me develop mathematical intuition. Yuri's favorable recommendation letter was instrumental in getting me admitted to the PhD program at Université Grenoble Alpes. I am also grateful to my referees Olivier Cappé and Arnak Dalalyan, as well as other members of my defence committee, for taking their time to read the manuscript and make suggestions for its improvement.

During the last four years, I enjoyed the company of my friends Pavel, Nikita, Valentin, Tolya, Thom, and Sergey. We shared some truly unforgettable moments, without which my life in Grenoble and Seattle would not have been as enjoyable and full.

I am grateful to my girlfriend Natasha for always being by my side during the last year.

Finally, and most importantly, words cannot express how thankful I am to my parents who have made great sacrifices to give me a better life. I am dedicating this thesis to them.

Abstract

We consider the problem of denoising a signal observed in Gaussian noise. In this problem, classical linear estimators are quasi-optimal provided that the set of possible signals is convex, compact, and known a priori. However, when the set is unspecified, designing an estimator which does not "know" the underlying structure of a signal yet has favorable theoretical guarantees of statistical performance remains a challenging problem. In this thesis, we study a new family of estimators for statistical recovery of signals with certain time-invariance properties. Such signals are characterized by their harmonic structure, which is usually unknown in practice. We propose new estimators that are capable of exploiting the unknown harmonic structure of a signal to reconstruct. We demonstrate that these estimators admit theoretical performance guarantees, in the form of oracle inequalities, in a variety of settings. We provide efficient algorithmic implementation of these estimators via first-order optimization algorithms with non-Euclidean geometry, and evaluate them on synthetic data as well as some real-world signals and images.

Résumé

Nous considérons le problème de débruitage d'un signal ou d'une image observés dans le bruit gaussien. Dans ce problème les estimateurs linéaires classiques sont quasi-optimaux quand l'ensemble des signaux, qui doit être convexe et compact, est connu a priori. Si cet ensemble n'est pas spécifié, la conception d'un estimateur adaptatif qui "ne connait pas" la structure cachée du signal reste un problème difficile. Dans cette thèse, nous étudions une nouvelle famille d'estimateurs des signaux satisfaisant certains propriétés d'invariance dans le temps. De tels signaux sont caractérisés par leur structure harmonique, qui est généralement inconnu dans la pratique. Nous proposons des nouveaux estimateurs capables d'exploiter la structure harmonique inconnue du signal à reconstruire. Nous démontrons que ces estimateurs obéissent aux divers "inégalités d'oracle," et nous proposons une implémentation algorithmique numériquement efficace de ces estimateurs basée sur des algorithmes d'optimisation de premier ordre. Nous évaluons ces estimateurs sur des données synthétiques et sur des signaux et images réelles.

Contents

1 Introduction **1**
- 1.1 General problem overview . . . 1
- 1.2 Recoverable signals and convolution-type estimators . . . 7
- 1.3 Adaptive convolution-type estimators . . . 9
- 1.4 Shift-invariance . . . 12
- 1.5 Harmonic oscillations and frequency estimation . . . 15
- 1.6 Outline of subsequent chapters . . . 20
- 1.7 Notation . . . 21
- 1.A Linear minimaxity on subspaces . . . 24

2 Uniform-fit estimators **25**
- 2.1 Problem statement . . . 25
  - 2.1.1 Recoverable signals . . . 26
  - 2.1.2 Limit of performance . . . 27
- 2.2 Adaptive estimators . . . 27
  - 2.2.1 Uniform-fit estimators . . . 28
  - 2.2.2 Bandwidth adaptation . . . 30
  - 2.2.3 Prediction setting . . . 32
- 2.3 Experiments . . . 34
- 2.4 Proofs . . . 35
  - 2.4.1 Proof of Theorem 2.1.1 . . . 35
  - 2.4.2 Proofs of Propositions 2.2.1–2.2.3 . . . 42
  - 2.4.3 Proof of Proposition 2.2.4 . . . 46
  - 2.4.4 Proof of Theorem 2.2.2 . . . 46

3 Least-squares estimators **48**
- 3.1 Problem statement . . . 49
  - 3.1.1 Adaptive estimators . . . 51
- 3.2 Theoretical results . . . 52
  - 3.2.1 Oracle inequalities for ℓ_2-loss . . . 52
  - 3.2.2 Corollaries for recoverable signals . . . 54
  - 3.2.3 Prediction setting . . . 56
- 3.3 Experiments . . . 58
- 3.4 Proofs of oracle inequalities . . . 63
  - 3.4.1 Technical tools . . . 63

3.4.2 Proof of Theorem 3.2.1 . . . 66
3.4.3 Proof of Theorem 3.2.2 . . . 72
3.4.4 Proof of Theorem 3.2.3 . . . 73
3.5 Remaining proofs . . . 75
3.A Why least-squares estimators cannot be analyzed as Lasso? . . . 76

4 Shift-invariance and recoverability **78**
4.1 General situation . . . 78
4.1.1 Reduction to the exact case . . . 78
4.1.2 Exact case . . . 79
4.2 Generalized harmonic oscillations . . . 81
4.2.1 Bounds for prediction . . . 81
4.2.2 Full recovery of harmonic oscillations . . . 82
4.3 Proofs . . . 84

5 Algorithmic implementation and complexity analysis **92**
5.1 Tools from convex optimization . . . 94
5.1.1 Proximal setup . . . 94
5.1.2 Composite minimization . . . 96
5.1.3 Composite saddle-point problems . . . 96
5.1.4 General accuracy bounds . . . 98
5.2 Efficient algorithmic implementation . . . 99
5.2.1 Computation of prox-mappings . . . 101
5.3 Theoretical complexity analysis . . . 102
5.3.1 Bounding the absolute accuracy . . . 103
5.3.2 Statistical accuracy and algorithmic complexity . . . 104
5.4 Experiments . . . 106
5.5 Remaining proofs . . . 109
5.A Adaptive stepsize policies . . . 112
5.B Online accuracy certificates . . . 112

Conclusion and perspectives **115**

Chapter 1

Introduction

1.1 General problem overview

This thesis is primarily concerned with the problem of discrete-time signal denoising. This problem has attracted significant attention of both the statistical estimation and the signal processing communities, see [IK81, Nem00, Tsy08, Was06, Hay91, Kay93] and references therein, and can be stated as follows. Let $x = (x_t)_{t\in\mathbb{Z}}$ be an unknown complex-valued signal which corresponds to a function of a continuous argument sampled over a constant interval. The goal is to estimate this signal from the noisy observations[1]

$$y_\tau := x_\tau + \sigma\zeta_\tau, \quad \tau \in \mathbb{Z}, \tag{1.1}$$

where ζ_τ is random noise with the standard complex Gaussian distribution $\mathbb{C}\mathcal{N}(0,1)$, meaning that the real and imaginary parts of ζ_τ are independent standard Gaussian random variables, and all ζ_τ are mutually independent. In order to make the discussion more substantial, let us formulate some natural requirements for possible estimators $\widehat{x}$ of x.

First, in practice we do not get to observe (1.1) on the entire $\mathbb{Z}$. Rather, we are interested in estimating x_t at some "reference points", and for every such point t we have observations in its neighborhood of the certain size $n \in \mathbb{Z}^+$ (here $\mathbb{Z}^+$ is the set of non-negative integers), that is, $y_{t+\tau}$ for $|\tau| \le n$. It is then natural to assume that the corresponding estimator only depends on these observations, and hence is *local* in the time domain.

Second, it is appealing to focus on *linear estimators* – those linearly dependent on the observations – for their simplicity in use and good interpretability. When taking into account the above locality requirement, linear estimators of the signal x_t at a fixed point t take the form

$$\widehat{x}_t^\varphi = \sum_{|\tau|\le n} \varphi_\tau y_{t-\tau}, \tag{1.2}$$

where $\varphi_\tau \in \mathbb{C}$, $|\tau| \le n$. Following the signal processing terminology, we call the vector φ in the above expression a *filter*. For what is to follow, it is useful to introduce specific classes of filters:

- *two-sided*, or *bilateral* filters:

$$\mathbb{C}_n(\mathbb{Z}) := \{\varphi \in \mathbb{C}(\mathbb{Z}) : \ \varphi_\tau = 0 \ \text{ if } \ \tau \notin [-n, n]\},$$

[1] In this chapter, we introduce non-standard notation as we use it. It will be fully presented in Section 1.7.

- *one-sided*, or *causal* filters:

$$\mathbb{C}_n^+(\mathbb{Z}) := \{\varphi \in \mathbb{C}(\mathbb{Z}) : \ \varphi_\tau = 0 \ \text{ if } \ \tau \notin [0, n]\},$$

- more generally, *predictive* and *interpolating* filters for some $h \in \mathbb{Z}$,

$$\mathbb{C}_n^h(\mathbb{Z}) := \{\varphi \in \mathbb{C}(\mathbb{Z}) : \ \varphi_\tau = 0 \ \text{ if } \ \tau \notin [h, h+n]\},$$

  where $h > 0$ corresponds to prediction, or extrapolation, with "horizon" h beyond $[t-n, t+n]$, and $h \in [-n, 0]$ to interpolation by a filter with "lobes" of unequal length.

Before we embark on the discussion of the above requirements for estimators, it would be useful to introduce a measure of comparison for them. Given an estimate $\widehat{x} = \widehat{x}(y)$, a *loss* $\ell(\widehat{x}(y), x)$ is a measure of misfit between the estimator and the signal when the observations y are fixed. In particular, when one is interested in recovering the value x_t at some "reference" point $t \in \mathbb{Z}$, one can use the *pointwise loss*

$$\ell_t(\widehat{x}(y), x) = |\widehat{x}_t(y) - x_t| \, .$$

More generally, one might be interested in estimating the signal in the n-neighborhood of a given point t for some $n \in \mathbb{Z}^+$, or even on the entire $\mathbb{Z}$. To this end, let us introduce the local seminorms on $\mathbb{C}(\mathbb{Z})$, defined, for $p \geq 1$, by

$$\|x\|_{n,p} := \left(\sum_{|\tau| \leq n} |x_\tau|^p \right)^{1/p}, \quad \|x\|_{n,\infty} := \max_{|\tau| \leq n} |x_\tau|,$$

together with their "global" counterparts, ℓ_p-norms on $\mathbb{C}(\mathbb{Z})$, defined as $\|x\|_p := \lim_{n \to \infty} \|x\|_{n,p}$ whenever the limit exists. Besides, let us define the time-shift operator Δ on $\mathbb{C}(\mathbb{Z})$ acting as

$$[\Delta x]_t = x_{t-1}.$$

Note that $\|\Delta^{-t} x\|_{n,p} = \left(\sum_{|\tau| \leq n} |x_{t+\tau}|^p \right)^{1/p}$. Then, one appropriate loss for estimating x in a neighbourhood of t would be the (normalized) *local ℓ_2-loss at* point t,

$$\frac{1}{\sqrt{2n+1}} \|\Delta^{-t}[\widehat{x}(y) - x]\|_{n,2}.$$

The local ℓ_2-loss at $t = 0$ will be called simply *the local ℓ_2-loss*. On the other hand, when estimating the signal on the entire $\mathbb{Z}$, it is appropriate to use the *global ℓ_2-loss* $\|\widehat{x}(y) - x\|_2$.

In the case where the signal is meant to be a regular sampling of some function, the requirement that an estimator is local should not appear too restrictive. On the other hand, it would be reasonable to ask why we restricted ourselves to linear estimators. Indeed, one could imagine that passing from arbitrary estimators to linear ones might bear a high price for our ability to capture the underlying structure of the signal. Indeed, all possible estimators of x_t from the observations in the n-sized neighborhood of t form an infinite-dimensional space, whereas the space of the corresponding linear estimators is an $(2n+1)$-dimensional vector space. Can we be sure that the class of estimators remains expressive enough after such a dramatic restriction? Surprisingly, the

answer to this question is positive. It turns out that under rather general assumptions, the class of all linear estimators of x_t is guaranteed to contain an estimator which is almost as good as the best possible one. In order to formulate this result rigorously, let us first remind the reader the classical decision-theoretic framework of estimation theory. For the rest of this section, we drop the underlying assumption that x corresponds to a regularly sampled function, and respectively, we do not force the locality of the estimators.

Once the loss ℓ has been chosen, the next step is to choose the *risk* – averaging of the loss, in some way or another, over the randomness of the observations. A common choice of risk, used in this chapter, is the *quadratic risk*

$$\mathrm{MSE}(\widehat{x}, x) := \left(\mathbb{E}[\ell(\widehat{x}, x)^2]\right)^{1/2}, \tag{1.3}$$

but other choices are possible, for example, an upper confidence bound on the loss. In this chapter, the quadratic risk associated to the local ℓ_2-loss will be called simply *the ℓ_2-risk*; the choice of n will be clear from the context.

It is natural for a "good" estimator to have small risk on *multiple* possible signals. Hence, it is reasonable to fix in advance a set $\mathcal{X} \subset \mathbb{C}(\mathbb{Z})$ of possible signals, usually corresponding to some prior assumptions on the sampled function, and compare estimators via the *maximal risk*,

$$\mathrm{MSE}_{\mathcal{X}}(\widehat{x}) := \sup_{x \in \mathcal{X}} \mathrm{MSE}(\widehat{x}, x).$$

As such, one aims to achieve the *minimax risk* on $\mathcal{X}$,

$$\mathrm{MSE}^*(\mathcal{X}) := \inf_{\widehat{x}} \mathrm{MSE}_{\mathcal{X}}(\widehat{x}),$$

where the infinum is taken over *all* possible estimators. Any estimator delivering the above infinum (assuming that the infinum is finite and is attained) is called a *minimax estimator*. Thus, the goal of the statistician in the decision-theoretic setting is to approach the minimax risk as closely as possible.

It was first noted by the authors of [IK84] that linear estimators are nearly optimal in a rather general situation. In our setting, their result can be stated as follows. Fix the pointwise loss ℓ_t, and consider the *linear minimax risk*

$$\mathrm{MSE}^{\mathrm{lin}}(\mathcal{X}) := \inf_{\varphi \in \mathbb{C}(\mathbb{Z})} \mathrm{MSE}_{\mathcal{X}}(\widehat{x}^{\varphi}),$$

in which the choice of estimators is restricted to linear ones. Then one has

$$\mathrm{MSE}^{\mathrm{lin}}(\mathcal{X}) \le 1.25\, \mathrm{MSE}^*(\mathcal{X}) \tag{1.4}$$

whenever $\mathcal{X}$ is *convex, compact, and centrally symmetric*. Furthermore, in [DLM90] it was found that pointwise loss in (1.4) can be replaced with the full ℓ_2-loss (and hence also with local ones), under somewhat more restrictive assumptions:

- central symmetry of $\mathcal{X}$ must be replaced with *orthosymmetry* ($\mathcal{X}$ is called orthosymmetric if it is axially symmetric with respect to each axis of some orthonormal basis);
- convexity of $\mathcal{X}$ must be replaced with convexity of its *quadratic hull* $\mathcal{X}_+^2 := \{(|x_t|^2)_{t \in \mathbb{Z}} : x \in \mathcal{X}\}$. A set $\mathcal{X}$ satisfying this assumption is called *quadratically convex*.

Moreover, this result can be extended to the sets of the form $\mathcal{X} \times \mathbb{C}^m$ [Joh11, p. 45]. Finally, under similar assumptions, a near-optimal linear estimator, as well as its maximal risk, can be computed efficiently via convex optimization, see *e.g.* [JN17]. The "prototypic" example of a quadratically convex and orthosymmetric set is an *ellipsoid*, whose quadratic hull is an intersection of a halfspace and the positive orthant. A more general class of quadratically convex sets is given by *ℓ_p-bodies* – sets of the form $\{x \in \mathbb{C}(\mathbb{Z}) : \|Ax\|_p \leq 1\}$, where $p \in [2, \infty]$, and A is a linear transformation of $\mathbb{C}(\mathbb{Z})$. Following [Joh11, Chapter IV], let us discuss these examples in more detail.

- *Subspaces.* Let $\mathcal{X} = \mathcal{S}$, a vector subspace of $\mathbb{C}(\mathbb{Z})$ of dimension s. Then, a minimax estimator on $\mathcal{X}$, with respect to the global ℓ_2-loss, is explicitly given by the Euclidean projector onto $\mathcal{S}$, and the corresponding risk is $\sigma\sqrt{s}$ for any signal from $\mathcal{X}$, as directly follows from the optimality of the maximum likelihood estimator $\widehat{x} = y$ on $\mathbb{C}^n$, see [Joh11]. This observation can be specified to the "local setting" where one only has observations (1.1) on the set $\{\tau : \tau \leq |n|\}$; equivalently, one can assume that x "lives" on $\mathbb{C}_n(\mathbb{Z})$. Formally, suppose that $x \in \mathcal{X}$, where
$$\mathcal{X} = \mathcal{S} \cap \mathbb{C}_n(\mathbb{Z}),$$
and $\mathcal{S}$ is a subspace of $\mathbb{C}(\mathbb{Z})$ of dimension $s \leq 2n+1$ (note that such $\mathcal{X}$ can be identified with an s-dimensional subspace of $\mathbb{C}^{2n+1}$). Then, the projector onto $\mathcal{X}$ is a near-minimax estimator with respect to the local ℓ_2-loss, and its ℓ_2-risk is
$$\sigma\sqrt{\frac{s}{2n+1}}.$$
Linear estimators are also near-optimal when estimating a linear functional on a subspace, see Section 1.A.

- *ℓ_p-bodies and smoothness.* Quadratically convex subsets of $\mathbb{C}(\mathbb{Z})$ include ℓ_p-bodies with $p \geq 2$, in particular, ellipsoids ($p = 2$) and hyperrectangles ($p = \infty$). These sets are naturally suited to represent smoothness of a function via its coefficients in orthonormal bases. Suppose that a function $f : [0,1] \to \mathbb{R}$ is α-times weakly differentiable for some $\alpha \geq 1$, and let $D^\alpha f$ be its weak derivative of order α, see *e.g.* [AF03]. Define L_p-norms on $[0,1]$ in the standard way,
$$\|g\|_{L_p} := \left(\int_{[0,1]} |g(u)|^p \mu(\mathrm{d}u)\right)^{1/p}, \quad p \geq 1,$$
$$\|g\|_{L_\infty} := \sup_{u \in [0,1]} |g(u)|.$$
The set of solutions for the inequality
$$\|D^\alpha f\|_{L_2} \leq L, \tag{1.5}$$
for some α and $L \geq 0$, is called a *Sobolev ball*, and denoted $\mathcal{S}_{\alpha,L}$. Consider the problem of

estimating $f \in \mathcal{S}_{\alpha,L}$ from the observations on the regular grid[2],

$$f(t/N) + \sigma\xi_t, \quad t \in \{0, ..., N-1\}, \tag{1.6}$$

using either the L_2-loss or the ℓ_2-loss on the grid. For both losses, this problem is asymptotically equivalent, see [BL96], to the problem of estimating $X \in \mathbb{C}(\mathbb{Z})$, the Fourier series of f, in the loss $\|\widehat{X} - X\|_2$ from the observations of the form (1.1),

$$Y_k = X_k + \frac{\sigma}{\sqrt{N}} Z_k, \quad k \in \mathbb{Z},$$

where $Z_k \in \mathbb{C}\mathcal{N}(0,1)$, under the assumption that X belongs to an ellipsoid of a special form called *Sobolev ellipsoid*. In this setting, there exists a near-minimax linear estimator for X, and hence there also exists an asymptotically near-minimax, as $N \to \infty$, estimator for $f \in \mathcal{S}_{\alpha,L}$, linear in the observations (1.6). One way to obtain such an estimator is by starting with a kernel estimator, and then choosing its bandwidth appropriately, see *e.g.* [Tsy08]. Similar results can be obtained for *Hölder balls* $\mathcal{H}_{\alpha,L}$, defined as the set of solutions to

$$\|D^\alpha f\|_{L_\infty} \leq L. \tag{1.7}$$

The appropriate basis in this case is given by the orthogonal wavelet transform [Mal08]. The set $\mathcal{H}_{\alpha,L}$ then corresponds to a hyperrectangle (an ℓ_∞-body) in the wavelet transform domain, see *e.g.* [Joh11, Theorem B10].

Adaptive estimation. While the above specific cases are remarkable and important, the genuine power of the results of [DLM90] is their generality. In stark contrast with the classical results of nonparametric estimation theory, one does not have to require that $\mathcal{X}$ has some specific form, such as an ellipsoid or a subspace, in order to have near-optimality of linear estimators. Instead, one has an "operational" result: whenever it happens that one can compute a linear minimax estimator on $\mathcal{X}$, that is, minimize the corresponding maximal risk over a finite-dimensional space, this estimator is "automatically" guaranteed to be nearly minimax for $\mathcal{X}$. This fact explains why linear estimators usually serve as a fundamental building block in *adaptive estimation* problems, In these problems, the signal is assumed to come from the union over a large – potentially even uncountable – *family* of sets $\{\mathcal{X}_\pi\}$, $\pi \in \Pi$. The hyperparameter π might correspond to the unknown structure of the signal – in the above examples, the support, or the smoothness information (α, L). It is assumed that each individual set $\mathcal{X}_\pi$ admits a near-optimal linear estimator $\widehat{x}^\pi$. The goal is to adapt to the unknown structure of the signal, as specified by hyperparameter π, by finding an *adaptive* estimator $\widehat{x}$ with a maximal risk on $\mathcal{X}_\pi$ similar to the maximal risk of $\widehat{x}^\pi$ *for any* π. The hope is that due to the additional special structure of the "basic" linear estimators $\{\widehat{x}^\pi\}$, one would be able to "search" over them efficiently, with respect to both statistical and computational notions of performance.

Two classical instances of adaptive estimation problems, which correspond to the examples of the set $\mathcal{X}$ considered above, are as follows.

- *Sparse recovery.* Assume that x is an s-sparse vector in $\mathbb{C}_n(\mathbb{Z})$, and $2n+1 \geq s$ (here it is again convenient to identify $\mathbb{C}_n(\mathbb{Z})$ with $\mathbb{C}^{2n+1}$). Equivalently, x comes from some

[2] In the literature, Sobolev and Hölder balls are more often defined by $\|f\|_{L_p} + \|D^\alpha f\|_{L_p} \leq L$ in order to enforce compactness, see *e.g.* [AF03]. Here we do not need this, see *e.g.* [Joh11, p. 45] for a related discussion.

s-dimensional subspace $\mathcal{X}_\pi$ of $\mathbb{C}(\mathbb{Z})$, spanned by some s vectors of the standard basis. In the associated family $\{\mathcal{X}_\pi\}$, the hyper-parameter π specifies the *support* – the choice of these vectors. Recall that since $\mathcal{X}_\pi$ is a linear subspace of $\mathbb{C}(\mathbb{Z})$, a perfect choice for the "basic" estimator $\widehat{x}^\pi$ is the projector on $\mathcal{X}_\pi$, or, equivalently, the least-squares estimator $\widehat{x}^\pi = \mathrm{argmin}_{x\in\mathcal{X}_\pi} \|y-x\|_{n,2}^2$, and its ℓ_2-risk is $\sigma\sqrt{s/(2n+1)}$. The challenge is that the "actual" $\mathcal{X}_\pi$, as given by the support of x, is unknown. The classical remedy is to penalize the quadratic criterion by the ℓ_1-norm of the signal, arriving at the optimization problem

$$\min_{x\in\mathbb{C}_n(\mathbb{Z})} \left\{\tfrac{1}{2}\|y-x\|_{n,2}^2 + \lambda\|x\|_{n,1}\right\}.$$

One can show that when the regularization parameter λ is correctly chosen, the ℓ_2-risk of an optimal solution $\widehat{x}$ of this problem matches that of the unknown estimator $\widehat{x}^\pi$ up to the factor $O(\sqrt{\log n})$, see [Joh11][3]. From the computational viewpoint, this estimator is reduced to performing componentwise *soft thresholding* of the observations,

$$\widehat{x}_k = (|y_k| - \lambda)\,\mathrm{sign}(y_k),$$

and as such, can be computed efficienly in $O(n)$. Importantly, the soft-thresholding estimator retains its nice statistical properties when generalized to the case of indirect observations,

$$y = \Psi\theta + \sigma\zeta$$

where Ψ is a linear mapping from $\mathbb{C}_p(\mathbb{Z})$ to $\mathbb{C}_n(\mathbb{Z})$, which can be understood as a *dictionary* of p "basic" signals, typically with $p \gg n$. Sparsity is now imposed on the vector of coefficients θ; indirectly, this corresponds to the assumption that the dictionary is rich enough to admit a short, or compressed, representation of the true signal (typically one works in the regime $s \ll n$). The corresponding generalization of the soft-thresholding estimator, called Lasso [Tib96], [CT07], [BRT09], is given as an optimal solution of a convex program

$$\min_{\theta\in\mathbb{C}_p(\mathbb{Z})} \left\{\|y-\Psi\theta\|_{n,2}^2 + \lambda\|\theta\|_{p,1}\right\};$$

as such, it can be efficiently computed via convex optimization. The ℓ_2-risk of Lasso is known to be within factor $O(\sqrt{\log p})$ from the ℓ_2-risk of $\widehat{\theta}^\pi$, provided that Ψ satisfies some incoherency conditions such as RIP (see [BVDG11, Chapter 6] for an extensive overview of these conditions and relations between them).

- *Adaptation to the unknown smoothness.* Here, each $\mathcal{X}_\pi$ is formed by signals $x_t = f(t/N)$, $t = 0, ..., N-1$, corresponding to regularly sampled functions from a particular smoothness class, such as a Sobolev ball $\mathcal{S}_{\alpha,L}$ or a Hölder ball $\mathcal{H}_{\alpha,L}$. The set Π of the values of hyperparameters $\pi = (\alpha, L)$ corresponds to a continuous smoothness "scale". Adaptive estimators can be constructed, in a computationally efficient way, via Lepski-type procedures [Lep91], [LMS97], [GL08], [GL11], [EP96], or other techniques such as unbiased risk estimation or the wavelet thresholding, see [Tsy08], [Joh11] and references therein.

Oracle approach. When dealing with adaptive estimation problems, it is important to realize that in certain situations, the explicit specification of the family $\{\mathcal{X}_\pi\}$ can be omitted provided

[3] We use the asymptotic notation $O(\cdot)$ and $\Omega(\cdot)$ in the usual sense; see Section 1.7 for the precise definitions.

that it is known, a priori, that every set $\mathcal{X}_\pi$ admits a near-optimal estimator. This idea underlies the *oracle approach* to adaptive estimation which is pursued in this thesis. In this approach, one does not explicitly characterize the family, but instead, assumes the existence of a linear *oracle estimator* $\widehat{x}^o$ which has uniformly small risk on all possible signals of interest. The oracle is not an actual estimator since it is allowed to depend on the unknown "structure" of the signal (which corresponds to the unknown hyperparameter in the "family-based" adaptive estimation problems). However, one can often find a good "proxy" to the oracle, the adaptive estimator $\widehat{x}$, whose risk is close to the oracle risk. In other words, one aims to find an estimator $\widehat{x}$ for which it is possible to prove an *oracle inequality*

$$\mathrm{MSE}(\widehat{x}, x) \le C\mathrm{MSE}(\widehat{x}^o, x) + \mathrm{Rem}(\widehat{x}^o, x), \tag{1.8}$$

where $C \ge 1$ is some absolute constant, and $\mathrm{Rem}(\widehat{x}^o, x)$ is a remainder term, known to be small compared to the oracle risk $\mathrm{MSE}(\widehat{x}^o, x)$ uniformly over all possible signals. An especially appealing kind of oracle inequalities are those with $C = 1$; such inequalities are called *sharp*, see [Tsy08]. In order to succeed in finding an estimator obeying an oracle inequality, one should impose additional restrictions, besides linearity, on the set of possible oracle estimators, which would allow to narrow down the search among them. In the classical approach, one explicitly specifies the sets $\mathcal{X}_\pi$, and restricts the class of possible oracles to be minimax, or nearly minimax, linear estimators for these sets. In the oracle approach, one imposes restrictions directly on the oracle, without explicitly specifying the family $\{\mathcal{X}_\pi\}$.

In the next section, we will present one possible restriction of the oracle which turns out to be relevant in the context of the original signal denoising problem – *time-invariance* of the oracle.

1.2 Recoverable signals and convolution-type estimators

Returning to the signal denoising problem, our assumption is that the oracle is *invariant* in the $O(n)$-sized neighborhood of a reference point, meaning that the same filter can be applied in the whole neighborhood, with a uniformly small pointwise risk for all the points in this neighborhood. This gives rise to the following definition, which is a simplified variant of the corresponding definitions in Chapter 2.

Definition 1.2.1 (Recoverable signals: simplified definition). *Signal $x \in \mathbb{C}(\mathbb{Z})$ is called* recoverable *at point $t \in \mathbb{Z}$ with parameters $(n, \rho) \in \mathbb{Z}^+ \times \mathbb{R}^+$, if there exists a filter $\phi^o \in \mathbb{C}_n(\mathbb{Z})$ such that the corresponding estimator $\widehat{x}^o_t = \sum_{|\tau|\le n} \phi^o_\tau y_{t-\tau}$ satisfies*

$$(\mathbb{E}[\ell_{t+\tau}(\widehat{x}^o, x)^2])^{1/2} \le \frac{\sigma\rho}{\sqrt{2n+1}}, \quad |\tau| \le 3n. \tag{1.9}$$

In other words, the linear estimator corresponding to a fixed *filter ϕ^o has uniform error $O(1/\sqrt{n})$ in an $O(n)$-sized neighborhood of t[4]. This uniform error is specified in terms of parameter ρ.*

Recall that the corresponding estimator $\widehat{x}^o_t$ is still given as a linear combination of the observations in the neightbourhood of t, cf. (1.2), but now t in that formula changes, whereas φ

[4]The reason for choosing the neighbourhood of radius $3n$ in (1.9) will become clear in Chapter 2.

must remain the same. Using the notion of *convolution* $x*y \in \mathbb{C}(\mathbb{Z})$ of two sequences $x, y \in \mathbb{C}(\mathbb{Z})$,

$$[x * y]_t := \sum_{\tau \in \mathbb{Z}} x_\tau y_{t-\tau},$$

the above fact can be conveniently expressed as $\widehat{x}^o = \varphi^o * y$. Hence, in the sequel we use the term *convolution-type estimators*. Note that despite formally referring to a signal in $\mathbb{C}(\mathbb{Z})$, Definition 1.2.1 is fully *local*: it actually concerns only the points in the $O(n)$-neighborhood of t. Another observation is that the expected square of the pointwise loss can be decomposed as the sum of the bias and the variance,

$$\mathbb{E}\ell_t^2(\widehat{x}^o, x) = [|x - \phi^o * x]_t|^2 + \sigma^2 \mathbb{E}|[\phi^o * \zeta]_t|^2;$$

Moreover, independence of the noise instances gives $\mathbb{E}|[\phi^o * \zeta]_t|^2 = \|\phi^o\|_2^2$. An immediate consequence of this observation is that the filter ϕ^o, whose existence is guaranteed by Definition 1.2.1, satisfies the following two properties:

- approximate reproduction of the signal:

$$|x - \phi^o * x|_{t+\tau} \le \frac{\sigma\rho}{\sqrt{2n+1}}, \quad |\tau| \le 3n; \tag{1.10}$$

- small ℓ_2-norm

$$\|\phi^o\|_2 \le \frac{\rho}{\sqrt{2n+1}}. \tag{1.11}$$

Moreover, these two properties clearly imply (1.9) with $\sqrt{2}\rho$ instead of ρ, so we could alternatively use them to define recoverable signals.

Classical example: smooth signals. Although the definition of recoverable signals might at this point seem somewhat artificial, in fact, the signals encountered in the classical estimation theory often satisfy this definition. The most basic example is the one of smooth signals which has already been mentioned before. The corresponding argument, found in [JN09, JN10], is as follows. Consider a function $f : [0,1] \to \mathbb{R}$, sampled over the regular grid with N points, $x_t = f(t/N), \quad t = 0, ..., N-1$, and smooth, say, in the sense of Sobolev (1.5) or Hölder (1.7). A classical choice of estimator in this situation is

$$\widehat{f}(t/N) = \frac{1}{2hN+1} \sum_{|\tau| \le hN} K\left(\frac{\tau}{hN}\right) y_{t-\tau}, \quad t \in [hN, (1-h)N], \tag{1.12}$$

or the *kernel* estimator. Here, $K : [-1,1] \to \mathbb{R}$ is a *kernel* with bounded support, which satisfies

$$\int_{-1}^{1} K(u)\mathrm{d}u = 1, \quad \int_{-1}^{1} K^2(u)\mathrm{d}u = \rho^2.$$

Note that a kernel gives rise to a filter $\phi^o \in \mathbb{C}_n(\mathbb{Z})$, $n = \lfloor hN \rfloor$, given by $\phi^o_\tau = \frac{1}{2hN+1} K\left(\frac{\tau}{hN}\right)$, so that (1.12) is exactly the corresponding linear estimator. Moreover, as follows from the conditions

above, the ℓ_2-norm of this filter satisfies

$$\|\phi^o\|_2 = O\left(\frac{\rho}{\sqrt{2n+1}}\right).$$

On the other hand, the variance of $\widehat{f}(t/N)$ is $\sigma^2\|\phi^o\|_2^2$, and if the bandwidth is chosen to balance the deterministic error $|x - \phi^o * x|_t$ with the stochastic error $|\phi^o * y|_t$, one has

$$|x - \phi^o * x|_t = O\left(\frac{\sigma\rho}{\sqrt{2n+1}}\right), \qquad t \in [hN, (1-h)N].$$

As a consequence, the signal is $(n, O(\rho))$-recoverable, by the same oracle filter, at all points of the observation domain $[0, N]$ except close to the borders.

In fact, the realm of recoverable signals is not limited to smooth signals. However, before we present a richer and more relevant in the context of this thesis class of recoverable signals, let us first explore the prospects of estimating such signals.

1.3 Adaptive convolution-type estimators

As we will discover in Chapters 2 and 3, recoverable signals admit a number of adaptive estimators. The common principle behind the construction of these estimators is as follows: one searches for an estimator of the form $\widehat{x} = \widehat{\varphi} * y$ where $\widehat{\varphi} = \widehat{\varphi}(y)$ is a data-dependent filter that "mimics", in one way or another, the unknown oracle filter. Specifically, in order to capture the property (1.10) of the oracle, one minimizes the observable counterpart of the signal reproduction error, a discrepancy term of the type $\ell(\varphi * y, y)$, where ℓ is some loss function, over some class of filters. Such class must be large enough to contain the oracle, yet not too large so that the search over it is still efficient. A readily given example of the filter class is the one consisting of all possible filters satisfying (1.11).

The described approach seems reasonable, but its implementation is not straightforward. First, it is unclear what loss to use. Second, and perhaps more importantly, the choice of a right filter class turns out to be non-trivial. As a simple illustration of these challenges, let us choose the local ℓ_2-loss, and the filter class which directly corresponds to (1.11). We then arrive at the following selection rule:

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_n(\mathbb{Z})} \left\{ \|\Delta^{-t}[y - \varphi * y]\|_{3n,2} : \quad \|\varphi\|_2 \le \frac{\rho}{\sqrt{2n+1}} \right\}.$$

However, we are not aware of any theoretical results for this procedure. As it turns out, the crucial step to obtain such results is to restrict the class of filters. It can be proved (see *e.g.* Proposition 2.2.4 and [JN09]) that the existence of ϕ^o satisfying (1.10)–(1.11) implies the existence of another filter $\varphi^o \in \mathbb{C}_{2n}(\mathbb{Z})$ with a twice larger support, which satisfies an analogue[5] of (1.10),

$$|x - \varphi^o * x|_{t+\tau} \le \frac{C\sigma\rho^2}{\sqrt{2n+1}}, \quad |\tau| \le 2n, \tag{1.13}$$

[5] In the sequel, C is an absolute constant that can recovered from the proofs, idem for C_i with integer subscript.

and the following counterpart of (1.11),

$$\|F_{2n}\left[\varphi^o\right]\|_1 \le \frac{2\rho^2}{\sqrt{4n+1}}, \tag{1.14}$$

where F_n is the Discrete Fourier transform of a sequence in the n-sized symmetric neighborhood of the origin as given by

$$[F_n x]_k = \frac{1}{\sqrt{2n+1}} \sum_{|\tau|\le n} \exp\left(\frac{2\pi i k\tau}{2n+1}\right) x_\tau \quad |k| \le n.$$

Note that if we treat ρ as a constant and ignore the shrinkage of the neighbourhood, (1.14) is stronger than (1.11). As simple as it is (the proof takes about half a page), this observation lays the ground for a whole theory of adaptive signal denoising, originating from [Nem91], [GN97] and further developed in [JN09, JN10, HJNO15, OHJN16a].

A typical result of this theory, whose proof is presented in Chapter 2, is as follows. Consider the following *uniform-fit estimator*, which tries to mimic φ^o instead of the initial oracle, and minimizes the loss given by the ℓ_∞-norm of the Discrete Fourier transform of the residual:

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi\in\mathbb{C}_{2n}(\mathbb{Z})} \left\{ \left\|F_{2n}\left[\Delta^{-t}[y-\varphi*y]\right]\right\|_{2n,\infty} : \ \|F_{2n}\left[\varphi\right]\|_1 \le \frac{\varrho}{\sqrt{4n+1}} \right\}, \tag{1.15}$$

where $\varrho = 2\rho^2$. For it, we obtain the following result (cf. Proposition 2.2.1 and [JN09]):

Theorem 1.3.1. *Let x be (n,ρ)-recoverable at t. Then, for the estimator $\widehat{x} = \widehat{\varphi} * y$ it holds*

$$(\mathbb{E}[\ell_t(\widehat{x},x)^2])^{1/2} = O\left(\frac{\sigma\rho^4\sqrt{\log(n+1)}}{\sqrt{n+1}}\right).$$

Price of adaptation. As we see, the adaptive filter admits a guarantee for the pointwise loss precisely at point t under the hypothesis that there exists an oracle filter with a small pointwise risk in the neighborhood of this point. However, we suffer an additional multiplicative factor $O(\rho^3\sqrt{\log n})$ with respect to (1.9). In Chapter 2, we will demonstrate that the factor $\Omega(\rho\sqrt{\log n})$ is unavoidable. Nonetheless, ϱ influences the error in a direct way, and we are interested both in keeping this parameter as small as possible from the beginning, that is, finding (n,ρ)-recoverable signals with small ρ, and in constructing adaptive estimators (and proving the corresponding bounds) with the adaptation price depending on ρ as mildly as possible. Chapters 3 and 4 are mainly dedicated to improving upon the results of Chapter 2 in these directions, albeit in a slightly different setting where one uses the ℓ_2-loss, and the signal is additionally assumed to be well-approximated in a shift-invariant subspace, cf. Section 1.4.

Least-squares estimators. While the ℓ_1-constraint on the filter in the Fourier domain invariably appears in all known to us constructions of adaptive convolution-type estimators [JN09, JN10, HJNO15, OHJN16a], the use of the ℓ_∞-criterion is motivated primarily by technical convenience. Indeed, the noise has Gaussian distribution, and hence it would be natural to

minimize the ℓ_2-norm of the residual, leading to the following optimization problem:

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_{2n}(\mathbb{Z})} \left\{ \left\| \Delta^{-t}[y - \varphi * y] \right\|_{2n,2} : \quad \|F_{2n}[\varphi]\|_1 \le \frac{\varrho}{\sqrt{4n+1}} \right\}. \tag{1.16}$$

Such *least-squares estimator* is investigated in Chapter 3 along with its various penalized versions. Least-squares estimators can be demonstrated to have superior performance, both for ℓ_2-loss and the pointwise loss, compared to the uniform-fit ones (1.15). In particular, when φ^o is taken as an oracle, we are able to prove *sharp* oracle inequalities for ℓ_2-loss, effectively reducing the price for adaptation to $O(\rho + \sqrt{\log n})$ for ℓ_2-loss (cf. Theorems 3.2.1–3.2.3) and to $O(\rho^2 + \rho\sqrt{\log n})$ for the pointwise loss (cf. Proposition 3.2.1). The study of these estimators led us to the in-depth exploration of one large class of recoverable signals – those belonging to shift-invariant subspaces, see Section 1.4.

Interpolation and prediction. Besides recovery with "bilateral" filters (those coming from a class $\mathbb{C}_n(\mathbb{Z})$ for some n), in the context of signal processing one can be interested in the related *one-sided filtering* problem where the filter is "unilateral", that is, belongs to $\mathbb{C}_n^+(\mathbb{Z})$. More generally, one can consider related *interpolation* and *prediction* problems which correspond to the filter classes $\mathbb{C}_n^h(\mathbb{Z})$ with $h \in \mathbb{Z}$. It turns out that the framework of recoverable signals can be easily extended to these cases, along with the constructions of adaptive convolution-type estimators. We present this extension in Chapters 2 and 3.

Bandwidth selection. Condition (1.9) is local, meaning that it only concerns $O(n)$ observations in a neighborhood of t. Ideally, we would like to select n that minimizes the oracle risk (1.9). Note that the right-hand side in (1.9) is necessarily large whenever either the stochastic or the deterministic part of the error is large, hence the "right" choice of n is the one for which the two parts are balanced. In the literature on nonparametric estimation, the task of inferring such n from data is called *(adaptive) bandwidth selection*. In Chapter 2 we resolve this task by proposing a procedure based on Lepski's method [Lep91].

Comparison with aggregation methods. It is instructive to compare our approach to adaptive estimation in the context of convolution-type estimators with those proposed within the *aggregation* community. Aggregation theory, initiated in [**?**] and [Nem00], seeks to mimic the *best* estimator in a given family of M linear or affine estimators; alternatively, one can address more ambitious tasks of *linear*, *convex*, or ℓ_2-aggregation, where the family of estimators to compete with is, correspondingly, an M-dimensional linear space, an ℓ_1-ball, or an ℓ_2-ball. In all cases, *discrete* or "sparsity-enforcing" structure of the family of initial estimators is exploited to obtain statistical convergence rates, and if the family lacks such a structure (as is the case for ℓ_2-balls), the set should first be discretized as explained below. A number of methods for constructing an adaptive estimator, called aggregation procedures within this community, have been proposed. These methods depart from constructing unbiased risk estimates for the squared ℓ_2-risk of the candidate estimators (which is always possible, when the noise is Gaussian, by Stein's method), and then rely on these risk estimates to construct the final estimator, either by simply choosing the one among the family that has the smallest risk estimate (empirical risk minimization), see [LM09], or aggregating the estimators with weights minimizing of a certain functional which depends on the risk estimates, as in ℓ_1-constrained minimization of Juditsky

and Nemirovski [JN00], exponential weighting method [LB06, RT12, DS+12, G+10], mirror averaging [JRT+08], and Q-aggregation [DRX+14, B+18]. A common result in aggregation theory is as follows: assuming that the family is finite and comprises M initial estimators, one shows a sharp oracle inequality for the squared ℓ_2-loss, with a minimax-optimal remainder term which scales as $\sigma^2 \log M/(2n+1)$. Similar results hold for infinite families, with M replaced with a suitable measure of complexity of the family such as its metric entropy. However, aggregation procedures lose they appeal, from a practical point of view, when the family of estimators is not discrete or sparsity-enforcing, for example in the case of an ℓ_2-ball or an ellipsoid. In this case, while aggregation methods can still be used to obtain statistical optimality results, doing so requires first to discretize the family by constructing a suitable ε-net (or multiple nets at different scales), see [YB99], and only aggregate estimators living in the net. Since the size of the ε-net for the required scale in the interesting cases scales exponentially in M, such aggregation procedures become intractable[6]. Note that this is exactly the situation arising in our problems of interest since in our case the oracle is known to be bounded in ℓ_2-norm, cf. (1.11). As explained above, our remedy to this problem is by passing to the new oracle which lives on the ℓ_1-ball rather than the ℓ_2-one. As discussed above, this reduction is possible thanks to the special (convolution-type) structure of the estimators that we consider. Notably, this comes at at price: we lose the possibility to obtain sharp adaptation results with respect to the *initial* oracle. Yet another restriction of the aggregation procedures is their reliance on unbiased risk estimation, and hence restriction to the use of ℓ_2-loss as a performance measure.

1.4 Shift-invariance

Previously, we demonstrated that signals sampled from smooth functions provide one example of recoverable signals. A more interesting class of such signals appears in Chapter 3 and is further discussed in detail in Chapter 4. It is given by the signals which belong to small-dimensional shift-invariant subspaces of $\mathbb{C}(\mathbb{Z})$ – invariant subspaces of the time shift operator Δ on $\mathbb{C}(\mathbb{Z})$.

Definition 1.4.1. *A subspace $\mathcal{S}$ of $\mathbb{C}(\mathbb{Z})$ is called* shift-invariant *if $\Delta\mathcal{S} \subseteq \mathcal{S}$ where operator $\Delta : \mathbb{C}(\mathbb{Z}) \to \mathbb{C}(\mathbb{Z})$ acts as $[\Delta x]_t = x_{t-1}$.*

Shift-invariant subspaces of $\mathbb{C}(\mathbb{Z})$ of finite dimension can be explicitly characterized as solution sets of ordinary difference equations. Specifically, consider a homogeneous linear difference equation on $\mathbb{C}(\mathbb{Z})$ with a polynomial operator,

$$[p(\Delta)x]_t \left[= \sum_{\tau=0}^{s} p_\tau x_{t-\tau}\right] \equiv 0, \quad t \in \mathbb{Z}. \tag{1.17}$$

Here, $p(z)$ is a polynomial $p(z) = 1 + p_1 z + ... + p_s z^s$ of degree s normalized by the condition $p(0) = 1$. A simple fact from the theory of ordinary difference equations, whose proof we provide in Chapter 4, is that the set of solutions for any such equation is a shift-invariant subspace of dimension s, and vice versa [7]. Precisely, the solution set of (1.17) is spanned by *exponential*

[6] Here we do not touch the alternative line of research on aggregation which dates back to [Kne94], where the family is alternatively assumed to be *ordered* in a certain sense which allows to circumvent discretization, see [G+10, GO14] and references therein.

[7] We did not manage to find in the literature any elementary proof of this result in the case of discrete time. The case of continuous time is discussed in [AK64], while [Lai79] and [Szé82] extend the result to Abelian groups.

polynomials specified by the roots of $p(z)$. Specifically, let for $k = 1, ..., r \le s$ the numbers z_k be the (distinct) roots of the polynomial $p(z)$ with the corresponding multiplicities m_k. Choose $\omega_k \in \mathbb{C}$ such that $z_k = e^{i\omega_k}$, then the solutions of (1.17) are given as the exponential polynomials

$$x_t = \sum_{k=1}^{r} q_k(t) e^{i\omega_k t}, \tag{1.18}$$

where $q_k(\cdot)$ are arbitrary polynomials with degrees $m_k - 1$. For instance, discrete-time polynomials of degree $s-1$ (that is, exponential polynomials with all zero frequencies) form a linear space of dimension s of solutions to (1.17) with $p(z) = (1-z)^s$. Another important example is a *harmonic oscillation* $x_t = \sum_{k=1}^{s} C_k \mathrm{e}^{i\omega_k t}$ with frequencies $\omega_k \in [0, 2\pi]$; here, $p(z)$ is the polynomial $p(z) = \prod_{k=1}^{s}(1 - e^{i\omega_k} z)$.

As demonstrated in Chapter 4, signals coming from shift-invariant subspaces admit oracle filters with small norm. For example (cf. Proposition 4.1.3), if x belongs to a shift-invariant subspace of dimension s, one can exhibit, for any $n \ge s$, an oracle filter $\phi^o \in \mathbb{C}_n(\mathbb{Z})$ which *exactly reproduces* the signal, that is,

$$x - \phi^o * x \equiv 0, \tag{1.19}$$

and for which (1.11) holds with

$$\rho(s) = \sqrt{s}. \tag{1.20}$$

In particular, x is (n, ρ)-recoverable with such ρ, so that Theorem 1.3.1 provides a guarantee

$$(\mathbb{E}[\ell_t(\widehat{x}, x)^2])^{1/2} = O\left(\frac{\sigma s^2 \sqrt{\log(n+1)}}{\sqrt{n+1}}\right)$$

for the pointwise loss. This result and its consequences merit a detailed discussion.

First, we obtained that the signals from *any* shift-invariant subspace of a bounded dimension are recoverable with a uniformly bounded ρ. As a consequence, the estimator (1.15) solves the adaptive estimation problem in which the signal is assumed to come from the family $\{\mathcal{S}_\pi\}$ of all shift-invariant subspaces whose dimension is at most s, and π is the tuple of complex "frequencies" ω_k in (1.18), counted with multiplicities.

Second, the exact signal reproduction (1.19) is not surprising, for the following two reasons.

- Due to (1.19) being satisfied for any signal from a subspace $\mathcal{S}$, the operator $\Phi^o(x) = \phi^o * x$ contains $\mathcal{S}$ in its invariant subspace, and in that it resembles the projector on $\mathcal{S}$. In fact, in the construction of Proposition 4.1.3, this filter is extracted from the projector operator, and (1.20) directly follows from the fact that the ℓ_2-risk of this estimator is $\sigma\sqrt{s/(2n+1)}$.

- On the other hand, filters satisfying (1.19) generalize the classical concept of *finite-order kernels* [Tsy08] to the case of arbitrary shift-invariant subspaces and discrete observations. A function $K : [-1, 1] \to \mathbb{R}$ is called a *kernel of order* s if it satisfies equations

$$\int_{-1}^{1} K(u)\mathrm{d}u = 1, \quad \int_{-1}^{1} u^j K(u)\mathrm{d}u = 0, \quad j = 1, ..., s.$$

  Now, suppose the usual sampling model $x_t = f(t/N)$, $t = 0, ..., N-1$, for a function

$f : [0,1] \to \mathbb{R}$, and consider the convolution of f with $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$,

$$[f * K_h](u) := \frac{1}{h}\int_{-h}^{h} K\left(\frac{v}{h}\right) f(u-v)\mathrm{d}v, \quad u \in [h, 1-h].$$

In the large sample limit, this convolution becomes the pointwise mean of the kernel estimator (1.12); hence, $f(u) - [f * K_h](u)$ is its bias. On the other hand, for any polynomial p of degree s, we have the analogue of (1.19):

$$p(u) - [p * K_h](u) = 0, \quad u \in [h, 1-h]. \tag{1.21}$$

Indeed, denoting $p(u) = \sum_{k=0}^{s} a_k u^k$, we have, for any $u \in [h, 1-h]$,

$$[p * K_h](u) = \frac{1}{h}\sum_{k=0}^{s} a_k \int_{-h}^{h} (u-v)^k K\left(\frac{v}{h}\right)\mathrm{d}v.$$

Applying the binomial formula, we further get

$$[p * K_h](u) = \frac{1}{h}\sum_{k=0}^{s} a_k \sum_{j=0}^{k} u^{k-j}(-1)^j \binom{k}{j} \int_{-h}^{h} v^j K\left(\frac{v}{h}\right)\mathrm{d}v.$$

The integral in the right-hand side is equal to h if $j = 0$ and vanishes otherwise, so we obtain (1.21).

Remark 1.4.1 (Nonparametric case: solutions to differential inequalities). It is important to note that exact recoverability (1.19) can be relaxed to its approximate counterpart. This allows to extend our results from the parametric setting (subspaces) to the nonparametric one which corresponds to *neighborhoods* of subspaces, or sampled solutions of differential inequalities, see *e.g.* [JN10] and references therein. Indeed, consider first the simplest case in which $\mathcal{S}$ is the subspace of all polynomials of degree at most s. Suppose that a function $f(\cdot) : [0,1] \to \mathbb{R}$ comes from $\mathcal{H}_{s+1,L}$, that is, satisfies (1.7) with $\alpha = s+1$. As before, let x be the sampling of such function on a regular grid with N points, $x_t = f(t/N)$, $t = 0, ..., N-1$. Now, note that whenever the bandwidth h is such that $n := \lfloor hN \rfloor \geq s$, for any $t \in [hN, (1-h)N]$ one can find a signal $x^{\mathcal{S}}$ – a regularly sampled polynomial of degree s – which satisfies

$$\left|[x - x^{\mathcal{S}}]_{t+\tau}\right| \leq C_s L h^s, \quad |\tau| \leq n, \tag{1.22}$$

where C_s is a constant which depends only on s. As such polynomial, one can simply take the sampled Taylor polynomial of $f(\cdot)$ at the point t of degree s. On the other hand, the signal $x^{\mathcal{S}}$, when extended to $\mathbb{Z}$, belongs to a shift-invariant subspace of dimension s, and as such, can be exactly reproduced by a filter $\phi^o \in \mathbb{C}_n(\mathbb{Z})$ satisfying $\|\phi^o\|_2 = O(\sqrt{s/n})$. Hence, the corresponding estimator $\widehat{x}^o_t = [\phi^o * y]_t$ satisfies

$$|[x - \widehat{x}^o]_{t+\tau}| = O\left(C_s L h^s + \frac{\sigma\sqrt{s}}{\sqrt{hN}}\right), \quad |\tau| \leq n.$$

The "proper" choice of h (which yields, in particular, the minimax rate of functions satisfying (1.7))

is the one which balances the two terms, so that the overall error becomes

$$|[x - \widehat{x}^o]_{t+\tau}| = O\left(\frac{\sigma\sqrt{s}}{\sqrt{hN}}\right), \quad |\tau| \le n.$$

As a consequence, "smooth" signals, corresponding to Hölder-smooth functions in $\mathcal{H}_{s,L}$, are recoverable with $\rho = O(\sqrt{s})$ when h is properly chosen [8]. Now, a simple argument [Nem00, Lemma 4.3.1] shows that the bound (1.22) can be extended to the case of a differential inequality

$$\|p(D)f\|_{L_\infty} \le L, \tag{1.23}$$

where $p(\cdot)$ is an *arbitrary* polynomial of degree $s+1$. In that case, $x^{\mathcal{S}}$ corresponds to a function $f^{\mathcal{S}}$ which satisfies the ordinary differential equation

$$p(D)f^{\mathcal{S}} \equiv 0,$$

that is, to an exponential polynomial. As a result, with the same choice of h as above (up to a multiplicative factor depending on s), recoverability holds for a signal corresponding to (1.23).

The concept of signals close to a shift-invariant subspace, this time in ℓ_2-norm rather than in the sense of (1.23), will emerge again in Chapter 3 when discussing least-squares estimators.

Remark 1.4.2. Note that the dependency $\rho(s) = O(\sqrt{s})$ is the best that one could possibly expect in a certain sense – namely, when one fixes a shift-invariant subspace $\mathcal{S}$, and requires that the oracle filter is *fixed* on $\mathcal{S}$, while the signal is allowed to range over $\mathcal{S}$. In this setting, going beyond $\rho(s) = O(\sqrt{s})$ would contradict the general fact that the minimax ℓ_2-risk on any subspace is $\Omega(\sigma\sqrt{s/(2n+1)})$[9]. However, when one restricts the class of filters to one-sided ones, that is, $\mathbb{C}_n^+(\mathbb{Z})$, estimation becomes a much more complicated task. In particular, when $p(z)$ is *unstable* – has roots *inside* the unit circle – the set of solutions to the equation (1.17) contains signals which grow exponentially fast. A simple calculation shows that such signals cannot be estimated consistently in a point t by any linear estimator which uses only the observations on the left of t. The situation becomes more optimistic in the specific case of *generalized harmonic oscillations*, where all ω_k in (1.18) are real, and the exponential growth of the signal is impossible. For such signals, one can point out one-sided oracle filters with polynomial bounds on the function $\rho(s)$. Specifically, in Chapter 4 we construct a filter which satisfies (1.19) together with

$$\rho(s) = O(s\sqrt{\log n}). \tag{1.24}$$

The existence of one-sided filters with smaller dependencies $\rho(s)$ remains an open question.

1.5 Harmonic oscillations and frequency estimation

The results discussed in the previous section are of interest, in particular, in connection with the classical problem of estimation of signals with *line spectra*, where the signal is assumed to be of

[8] Note that we essentially repeated the argument for recoverability of smooth signals from Section 1.2, but this time additionally requiring that the oracle exactly reproduces the signal from a shift-invariant subspace.

[9] However, note that in the precise sense of Definition 1.2.1, where the oracle is defined for a particular signal, oracles with better $\rho(s)$ *may* exist. In fact, one has $\rho = 0$ for $x \equiv 0$ which belongs to *any* shift-invariant subspace.

the form

$$x_t = \sum_{k=1}^{s} \alpha_k e^{i\omega_k t}, \quad t = 0, ..., n-1. \tag{1.25}$$

While in the context of denoising, the ultimate goal is to estimate the signal x itself from the observations (1.1), the interest of signal processing community historically was directed towards *spectral estimation* methods which aim at estimating the signal frequencies. Note that this problem is at least as hard as the estimation of the signal itself. Namely, suppose that x is correctly parametrized by its frequencies (ω_k) and amplitudes (α_k) as it is the case, for example, when $n \geq 2s$; note that otherwise the frequency estimation problem is not correctly stated. Then, any consistent estimator of the frequencies provides a consistent estimator of the amplitudes via least-squares, resulting in a consistent estimator $\widehat{x}$ of x. Moreover, if the parametrization of a harmonic oscillation (1.25) by (ω_k) and (α_k) is *regular* – that is, defines a regular parametric model [VdV98] – the asymptotic normality of the frequency estimator would yield, by the delta-method, the asymptotic normality of $\widehat{x}$.

Let us now briefly overview some techniques for frequency estimation of harmonic oscillations.

Maximum likelihood estimator. The maximum likelihood estimator (MLE) of the frequencies is given by non-linear least-squares. In the setting where $n \to \infty$, while the frequencies remain fixed, the MLE is shown to be asymptotically efficient, see [SMFS89] and [SN89]. Its main disadvantage is that it cannot be implemented efficiently because of the non-convexity of the corresponding log-likelihood. Moreover, the log-likelihood has a very sharp global maximum, and hence, computing it by a search algorithm requires accurate initialization [SMFS89]. Various heuristic approaches with a good empirical performance have been proposed, such as initialization via the Discrete Fourier transform [YPM94], [LS96], approximation of the log-likelihood [KSS86], [UT96], [BM86], maximization by the alternative projections algorithm [ZW88] and by the EM algorithm, see [SS04] and references therein. Despite that, the computation of the MLE remains a highly challenging task, see *e.g.* [SS04].

Interestingly, the MLE is related to the *periodogram* estimator, which dates back to the classical paper [Sch06], and is often used in the context of testing a signal for periodicity and estimating the base frequency, see [QT91], [QH01] and references therein. Here, the *periodogram*

$$\widehat{P}_n(\omega) = \left| \sum_{t=0}^{n-1} e^{i\omega t} y_t \right|,$$

is taken as an estimate of the power spectral density of the signal, defined as the modulus of the Discrete-Time Fourier transform,

$$P(\omega) = \left| \sum_{t \in \mathbb{Z}} e^{i\omega t} x_t \right|.$$

In the case of harmonic oscillations (1.25), $P(\omega)$ becomes a measure supported on the set of signal frequencies. This suggests a computationally cheap way of estimating the frequencies as the positions of s highest peaks of $\widehat{P}_n(\omega)$. When the search of the peaks is restricted to the Discrete Fourier transform grid, this approach turns out to be exactly equivalent to the MLE. Moreover, the statistical performance of the two estimators is similar in the general case,

see [SM05], provided that the frequencies are *well-separated*. Specifically, the minimal frequency separation in the wrap-around distance, $\delta_{\min} := \min_{i \neq j} |\omega_i - \omega_j|$, must satisfy

$$\delta_{\min} \geq \frac{2\pi\kappa}{n} \tag{1.26}$$

for some $\kappa \geq 1$; in particular, the frequencies are in the different "bins" of the Discrete Fourier transform. The condition (1.26) often arises in the literature.

Subspace methods. The classical approach unifies the so-called *subspace methods*: Pisarenko harmonic decomposition [Pis73], MUSIC [Sch79], [Sch86], ESPRIT [RK89], Cadzow's method [Cad80], [Cad88] and the Matrix Pencil algorithm [HS90], see also the monographs [SM05], [MIK00] and references therein. These methods heavily exploit the geometry of the space $\mathbb{C}^m$ corresponding to the first $m \geq s$ observations in order to estimate the frequencies of the signal. This geometry can be summarized as follows. For any $m \geq s$, the space $\mathbb{C}^m$ contains a shift-invariant subspace $\mathcal{S}$ of dimension s, spanned by the columns of the $m \times s$ Vandermonde matrix V_m which can be written as $V_m = [v_m(\omega_1) \cdots v_m(\omega_s)]$ using the so-called transfer (or imaging) function

$$v_m(\omega) = [1 \; e^{i\omega} \; \cdots \; e^{i\omega(m-1)}]^{\mathrm{T}}.$$

In the literature on subspace methods, $\mathcal{S}$ is usually called the *signal subspace*, and its orthogonal complement $\mathcal{N}$ is called the *noise subspace*. Various subspace methods differ in how they use this structure to obtain the frequency estimates. Following the recent literature [LF16], [Moi15], [AB16], these methods can be conveniently described in a unified way using the Hankel matrix operator [10] $H_{m,n}(x) : \mathbb{C}^n \to \mathbb{C}^{(n-m+1)\times m}$,

$$H_{m,n}(x) := \begin{bmatrix} x_0 & x_1 & \cdots & x_{m-1} \\ x_1 & x_2 & \cdots & x_m \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-m} & x_{n-m+1} & \cdots & x_{n-1} \end{bmatrix}.$$

The decomposition of the signal slice $x^{(t)} := x_t^{t+m-1}$,

$$\begin{aligned} x^{(t)} &= V_m \alpha^{(t)}, \\ \alpha^{(t)} &= [\alpha_1 e^{i\omega_1 t} \cdots \alpha_s e^{i\omega_s t}]^{\mathrm{T}}, \end{aligned}$$

implies the following decomposition of $H = H_{m,n}(x)$:

$$H = V_{n-m+1} \mathrm{Diag}(\alpha_1, ..., \alpha_s) V_m^{\mathrm{H}}, \tag{1.27}$$

see *e.g.* [LF16]. Hence, $\mathrm{rank}(H) = s$ provided that $s \leq m \leq n - s + 1$. Moreover, the null-space of H coincides with the null-space of V_m^{H} and with the noise subspace $\mathcal{N}$.

Some of the widely used subspace methods are presented below.

- MUSIC, standing for MUltiple SIgnal Characterization/Classification, was proposed in [Sch79]

[10] We prefer this formalism to the more classical one using covariance matrices. The reason for that is that whenever $s > 1$, the covariance matrix of the true signal has rank s only if the signal components corresponding to different frequencies have random and uncorrelated phases, see [SN89].

and independently in [BK79]. It relies on the characterization of $\mathcal{N}$ as the null-space of V_m. Denote $\{w_j\}_{j=1,...,m-s}$ the basis of $\mathcal{N}$ formed by the right singular vectors of H corresponding to the zero singular value. Noting that $\mathcal{S}$ is spanned by the transfer vectors $\{v_m(\omega_k)\}_{k=1}^s$, we can write

$$v_m^{\mathrm{H}}(\omega)w_j = 0, \quad j = 1, ..., m-s \tag{1.28}$$

for the frequencies $\omega \in \{\omega_k\}_{k=1}^s$, provided that $m \geq s+1$, that is, the noise subspace is non-degenerate. Moreover, the condition (1.28) exactly identifies the signal frequencies: as follows from the non-degeneracy of the extended Vandermonde matrix $[V_m; v_m(\omega)]$, (1.28) does not hold if ω is not one of the signal frequencies. The MUSIC algorithm exploits this observation, replacing H with its observable counterpart $H_{m,n}(y)$. Specifically, one first computes the singular vectors $\widehat{w}_j$ corresponding to the lowest $m-s$ singular values of $H_{m,n}(y)$, and then estimates the signal frequencies as the locations of s highest peaks of the function

$$\frac{v_m^{\mathrm{H}}(\omega)v_m(\omega)}{\left|\sum_{j=1}^{m-s} v_m^{\mathrm{H}}(\omega)\widehat{w}_j\right|^2},$$

called the MUSIC pseudospectrum.

- ESPRIT, standing for "Estimation of Signal Parameters via Rotational Invariance Techniques", was proposed by Kailath and coauthors in [RPK86] and [RK89]. Following the presentation of [AB16], the key property is that the matrices V_- and V_+ which comprise, correspondingly, the first and the last $m-1$ rows of the matrix $V = V_m$, are related through a diagonal rotation matrix $R = \mathrm{Diag}(e^{i\omega_1}, ..., e^{i\omega_s})$, as $V_+ = V_- R$, or equivalently,

$$R = V_-^{\dagger} V_+,$$

where

$$A^{\dagger} := (A^{\mathrm{H}}A)^{-1}A^{\mathrm{H}}$$

is the (left) pseudoinverse of A. Suppose that $m \geq s+1$, and consider the $m \times s$ matrix S formed by s top (right) singular vectors of H. Since it comprises a basis of the signal subspace, we have that $S = VP$ for some non-degenerate $s \times s$ matrix P. Hence, introducing S_- and S_+ analogously to V_- and V_+, we obtain that matrix $\Phi := S_-^{\mathrm{H}} S_+$ has eigendecomposition $\Phi = P^{-1}RP$, so that $\{\omega_k\}_{k=1}^s$ are the phases of its eigenvalues. In ESPRIT, one first forms $\widehat{S}$ from s top right singular vectors of $H(y)$, and then applies the above procedure to $\widehat{S}$ instead of S.

- The Matrix Pencil method [HS90], [HS91], [YF96], [AB16] also uses the rotation matrix R. Here, one considers H_- and H_+,

$$H_- = H_{m-1,n-1}(x), \quad H_+ = H_{m-1,n-1}(\Delta^{-1}x);$$

recall that $\Delta^{-1}x$ is the unit-lagged signal. Then, similarly to (1.27), one has

$$H_- = V_{n-m+1}\Lambda V_{m-1}^{\mathrm{H}}, \quad H_+ = V_{n-m+1}\Lambda R V_{m-1}^{\mathrm{H}}$$

where $\Lambda = \mathrm{Diag}(\alpha_1, ..., \alpha_s)$. Writing $H_+ - \lambda H_- = V_{n-m+1}\Lambda(R - \lambda I)V_{m-1}^{\mathrm{H}}$, we see that the

elements of D are exactly the solutions of the generalized eigenproblem,

$$\det(H_+ - \lambda H_-) = 0,$$

or equivalently, the eigenvalues of $H_-^\dagger H_+$. The frequencies are estimated as the phases of the generalized eigenvalues of the pair of observable counterparts of H_- and H_+.

Vast literature is devoted to the statistical analysis of subspace methods in the asymptotic regime, see [SN89], [RH89], [SS91], [YF96], [YF98], and references therein; the standard result is asymptotic efficiency of a particular method in the limit $m, n \to \infty$. Meanwhile, the only known to us non-asymptotic results can be found in [LF16] for MUSIC, and [Moi15] for the Matrix Pencil method in the deterministic setting. Although undoubtedly important in the context of frequency estimation, methods in [LF16] and [Moi15] deal with adversarial noise, and hence are of limited interest in the context of the observation model (1.1). Moreover, the adversarial assumption is essential for these works since the proofs heavily rely on some instruments from matrix perturbation theory such as the Davis-Kahan and Wedin theorems [Wed72], see also [Vu11], [CZ16] for an overview of these techniques and their applications in statistics. Another restriction in [LF16] and [Moi15] is the frequency separation assumption (1.26).

Atomic norm denoising. The idea of this approach, which was proposed in the context of compressed sensing by [CRPW12], [CFG14], and adapted for denoising in [BTR13]–[TBR13], is to treat the signal as a sparse combination of elements from the dictionary of complex sinusoid signals with frequencies on the unit circle. In contrast with the standard sparse recovery problems, this dictionary is continuous. Furthermore, discretization of the dictionary via oversampling does not help because the discretized dictionary quickly becomes highly coherent (see [DB13]). Fortunately, this difficulty can be overcome using the notion of atomic norm [CRPW12]. The corresponding algorithm, called Atomic Soft Thresholding (AST), was introduced in [BTR13] and futher studied in [TBR13]. AST is known to achieve the optimal ℓ_2-risk for denoising, in the non-asymptotic regime, and under the assumption that the frequencies are well-separated as in (1.26). However, the known bounds for AST in the case of non-separated frequencies are inferior [BTR13]. Moreover, even the asymptotic normality of AST, in the regime where s is constant, and $\delta_{\min} \to 0$ at a rate faster than $O(1/n)$, has not been established.

Differences with our approach. As we see, all the approaches to the problem of denoising harmonic oscillations (1.25) which have been presented so far solve this problem by essentially reducing it to the problem of identifying the harmonic structure of the signal. In contrast to that, our approach directly focuses on the signal denoising problem. While, when possible, efficient estimation of the frequencies guarantees fine recovery of the signal in the time domain, the reverse is not true in the general situation. In particular, in the finite-sample case, and when the frequencies are not well-separated as in (1.26), frequency estimation can be a fundamentally more difficult task than signal reconstruction, at least in the deterministic setting, see *e.g.* [Moi15, Corollary 3.2]. On the other hand, the bounds on $\rho(s)$ such as (1.24) do not require the minimal separation assumption (1.26). As a consequence, our estimators achieve, up to a logarithmic factor, the ℓ_2-risk $O(\sigma\sqrt{s^\kappa/n})$ for a moderate $\kappa \geq 1$, *without any conditions on the frequencies.*

1.6 Outline of subsequent chapters

Chapters 2 to **4** are based on [HJNO15] and [OHJN16a] (see also the full version [OHJN16b] of the latter publication) but also include some previously unpublished results. In these chapters, we describe adaptive estimators for recoverable signals, study their statistical properties, and finally, investigate the relation between recoverability and shift-invariance.

- In **Chapter 2** we focus on the uniform-fit estimator (1.15) and its counterparts, and correspondingly, on the pointwise loss. We prove theoretical upper bounds on the pointwise loss of adaptive estimators, and provide a lower bound for the price of adaptation of any estimator of a recoverable signal. We also describe the extension of our results to the case of prediction. Finally, we present a bandwidth selection technique for the estimators and prove the corresponding statistical accuracy bound.

- In **Chapter 3** we discuss the least-squares estimators, which turn out to possess better statistical properties than those based on ℓ_∞-fit. In particular, it turns out to be possible to prove sharp oracle inequalities (those with unit leading constants) for the ℓ_2-loss of such estimators. Chapter 3 is concluded by numerical experiments on synthetic data which demonstrate practical viability of our approach.

- In **Chapter 4** we study signals which belong to shift-invariant subspaces. We provide bounds for the dependency $\varrho(s)$ for these signals, where s is the subspace dimension, for two qualitatively different situations: when the oracle filter can be two-sided, and when it is required to be causal. Combining these results with oracle inequalities for least-squares estimators, we construct an estimator with state-of-the-art ℓ_2-risk of $O(\sigma\sqrt{s^3/n})$ up to a logarithmic in n factor. As a standalone contribution, we give an improved bound on $\rho(s)$ under the minimal frequency separation (1.26). Finally, we compare these results with the atomic norm approach of [BTR13] and [TBR13] in numerical experiments.

Chapter 5 follows [OH18]. In this chapter, we present the algorithms for efficient numerical solution of the optimization problems arising when constructing adaptive estimators. From the computational viewpoint, these problems belong to the class of second-order convex programs which can be solved using interior-point methods [BTN01]. However, computational complexity of interior-point methods applied to programs with *dense* matrices grows polynomially with the problem dimension, so that large problems of such type arising in speech and image processing (sample sizes up to 10^8) are well beyond the reach of these techniques. On the other hand, the optimization problems underlying the proposed adaptive estimators share some properties that make them naturally suitable to another potent class of algorithms, *first-order proximal methods*. Based on this observation, we develop a unified approach towards the numerical solution of these optimization problems, based on two specific first-order algorithms which are suited particularly well for our goals, Fast Gradient Method [NN13] and Mirror Prox [NN13, JN11b]. We then provide bounds on the computation complexity of the resulting procedures, explicitly taking into account their statistical nature. We conclude by presenting numerical experiments which demonstrate the practical usefulness of our framework.

1.7 Notation

Let us now introduce the notation which will be used throughout subsequent chapters. We allow ourselves to slightly alter it, with an explicit mention in every case, if the circumstances so require. Besides, additional notation is sometimes presented and used in the proofs.

Spaces. $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ are the sets of all integer, real, and complex numbers correspondingly. $\mathbb{Z}^+$ is the set of all non-negative integers, $\mathbb{Z}^{++}$ the set of all strictly positive integers, idem for $\mathbb{R}$. $\mathbb{R}^n$ and $\mathbb{C}^n$ for $n \in \mathbb{Z}^+$ are the real and complex coordinate spaces. $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ are the spaces of $m \times n$ matrices with real or complex elements.

Complex numbers. $\Re(z)$ and $\Im(z)$ denote, correspondingly, the real and imaginary parts of a complex number $z \in \mathbb{C}$, and $\overline{z} = \Re(z) - i\Im(z)$ denotes the complex conjugate of z.

Probability. $\mathbb{P}(E)$ denotes the probability of an event E. When described by a predicate $\mathcal{C}$, an event is written as $\{\mathcal{C}\}$. The expectation is denoted by $\mathbb{E}$, and the probability measure over which it is taken is clear from context.

Asymptotic notation. $\log(x)$ denotes the *natural* logarithm. Symbols C and c sometimes with integer subscripts, stand for absolute constants whose exact values can be recovered from the proofs. We use the "big O" notation: for two functions f, g of the same argument, $f = O(g)$ means that there exists a constant $C \geq 0$ such that $f \leq Cg$ for any possible value of the argument; $f = \Omega(g)$ is the same as $g = O(f)$. Besides, $f = \tilde{O}(g)$ means that $f \leq g$ holds up to a logarithmic factor in the common argument of f and g, and is equivalent to $g = \tilde{\Omega}(f)$.

Matrices, vectors, and signal slices. We follow the "Matlab convention" for matrices: $[A\,B]$ and $[A; B]$ denote, correspondingly, the horizontal and vertical concatenations of two matrices of compatible dimensions. Unless explicitly stated otherwise, all vectors are column vectors. We denote A^{T} the transpose of a complex-valued matrix A, and A^{H} its conjugate transpose. We denote $\overline{A}$ the conjugation of A without transposition. We denote A^{-1} the inverse of A whenever it is guaranteed to exist. We denote $\mathrm{Tr}(A)$ the trace of a matrix A, $\det(A)$ its determinant, $\|A\|_{\mathrm{F}}$ the Frobenius norm, and $\|A\|_{\mathrm{op}}$ the operator norm. Besides, in Chapter 5 we use subordinate matrix norms, $\|A\|_{\alpha\to\beta} := \max_{\|u\|_\alpha=1} \|Au\|_\beta$, where $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ are some norms in the domain and co-domain of A. We denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ the maximal and minimal eigenvalues of a Hermitian matrix A. We denote $\mathrm{Diag}(a)$ the diagonal matrix formed from a vector $a \in \mathbb{C}^n$. We denote I the identity matrix, sometimes with a subscript indicating its size. We denote $\langle\cdot,\cdot\rangle$ the Hermitian scalar product: for $a, b \in \mathbb{C}^n$, $\langle a, b\rangle = a^{\mathrm{H}}b$. Given a signal $x \in \mathbb{C}(\mathbb{Z})$ and $m, n \in \mathbb{Z}$ such that $m \leq n$, we define

$$x_m^n := [x_m; ...; x_n] \in \mathbb{C}^{m+n+1}.$$

Note that the set $\{[x]_m^n, x \in \mathbb{C}(\mathbb{Z})\}$ can be identified with the vector space $\mathbb{C}^{n-m+1}$.

Filters. Recall that $\mathbb{C}(\mathbb{Z})$ is the linear space of all two-sided complex sequences

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

Given $\varphi \in \mathbb{C}(\mathbb{Z})$ with a finite number of non-vanishing elements, and observations y as defined in (1.1), we associate with φ a linear estimate of x_t for the signal $x \in \mathbb{C}(\mathbb{Z})$ according to

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

The smallest integer n such that $\varphi_\tau = 0$ whenever $|\tau| > n$ is called the order of a *filter* φ; the estimator of order n has at most $2n+1$ non-zero entries. Note that $\widehat{x}_t$ is nothing but a kernel estimate over the grid $\mathbb{Z}$ with a finitely supported kernel φ. We distinguish the following specific classes of filters corresponding to finite-dimensional subspaces of $\mathbb{C}(\mathbb{Z})$:

- *two-sided*, or *bilateral* filters:

$$\mathbb{C}_n(\mathbb{Z}) := \{\varphi \in \mathbb{C}(\mathbb{Z}) : \ \varphi_\tau = 0 \ \text{ if } \ \tau \notin [-n, n]\},$$

- *one-sided*, or *causal* filters:

$$\mathbb{C}_n^+(\mathbb{Z}) := \{\varphi \in \mathbb{C}(\mathbb{Z}) : \ \varphi_\tau = 0 \ \text{ if } \ \tau \notin [0, n]\},$$

- more generally, "shifted" filters

$$\mathbb{C}_n^h(\mathbb{Z}) := \{\varphi \in \mathbb{C}(\mathbb{Z}) : \ \varphi_\tau = 0 \ \text{ if } \ \tau \notin [h, h+n]\},$$

  with the shift $h \in \mathbb{Z}$. Note that $h \in [-n, 0]$ encodes *bilateral* filters with lobes of different length, while $h \geq 0$ corresponds to *predictive* filters which allow for extrapolation with "horizon" h beyond the observation domain.

Note that the terminology we use here has direct signal processing counterparts: estimation with bilateral filters corresponds to *linear interpolation*, and estimation by h-predictive filters – to *linear filtering* (when $h = 0$) and *prediction* (when $h > 0$).

Convolution. It is convenient to identify a filter φ with the finite Laurent sum $\varphi(z) = \sum_j \varphi_j z^j$. Note that the convolution $\varphi * \psi$ of two filters corresponds to the product $\varphi(z)\psi(z)$, and therefore the order of $\varphi * \psi$ is at most the sum of the orders of φ and ψ. Recalling the definition of the right-shift operator Δ on $\mathbb{C}(\mathbb{Z})$,

$$[\Delta x]_t = x_{t-1},$$

and its inverse, the left-shift Δ^{-1}, $[\Delta^{-1}x]_t = x_{t+1}$, the linear estimate $[\varphi * y]_t$ with rational φ may be alternatively expressed as $[\varphi(\Delta)y]_t$.

Fourier transform. For any non-negative integer n, let Γ_n be the set of complex roots of unity of degree $2n+1$, and let $\mathbb{C}(\Gamma_n)$ be the space of all complex-valued functions on Γ_n. We define the *(symmetric and unitary) Fourier transform FT)* operator $F_n : \mathbb{C}^{2n+1}(\mathbb{Z}) \to \mathbb{C}(\Gamma_n)$,

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

Note that $\mathbb{C}(\Gamma_n)$ can be identified with $\mathbb{C}^{2n+1}$, so that FT becomes a unitary transformation of $\mathbb{C}^{2n+1}$. Moreover, the FT inversion formula holds:

$$x_\tau = (2n+1)^{-1/2} \sum_{\mu \in \Gamma_n} (F_n x)(\mu) \mu^{-\tau}, \quad |\tau| \le n.$$

Norms in the time and spectral domains. Given $p \in [1, +\infty]$ and a non-negative integer n, we introduce the semi-norms on $\mathbb{C}(\mathbb{Z})$ defined by

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

with the natural interpretation for $p = +\infty$. When such notation is unambiguous, we also use $\|\cdot\|_p$ to denote the "usual" ℓ_p-norm on $\mathbb{C}(\mathbb{Z})$ (*e.g.* for x such that $\mathrm{ord}(x) = n$, $\|x\|_p = \|x\|_{n,p}$). On the other hand, the Fourier transform allows to equip $\mathbb{C}(\mathbb{Z})$ with seminorms associated with the standard p-norms in the frequency domain:

$$\|x\|^{\mathrm{F}}_{n,p} := \|F_n x\|_p = \left(\sum_{\mu \in \Gamma_n} |(F_n x)(\mu)|^p \right)^{1/p}, \quad p \in [1, +\infty]. \tag{1.29}$$

Note that with this notation, the unitarity of F_n translates to the Parseval identity:

$$\langle x^n_{-n}, [y]^n_{-n} \rangle = \langle F_n x, F_n y \rangle, \quad \|x\|_{n,2} = \|x\|^{\mathrm{F}}_{n,2}. \tag{1.30}$$

Unilateral Fourier transform and norms. Extra notation is required to deal with predictive filters. First, we introduce the unilateral semi-norms on $\mathbb{C}(\mathbb{Z})$,

$$\|x\|^+_{n,p} := \left(\sum_{0 \le \tau \le n} |x_\tau|^p \right)^{1/p}.$$

Now, for any non-negative integer n, let Γ^+_n be the set of complex roots of unity of degree $n+1$, and let $\mathbb{C}(\Gamma^+_n)$ be the space of all complex-valued functions on Γ^+_n. We define the *(unilateral) Fourier transform (FT)* operator $F^+_n : \mathbb{C}(\mathbb{Z}) \to \mathbb{C}(\Gamma^+_n)$ as

$$(F^+_n x)(\mu) := (n+1)^{-1/2} \sum_{\tau=0}^{n} x_\tau \mu^\tau \left[= (n+1)^{-1/2} x(\mu),\ x \in \mathbb{C}^0_n(\mathbb{Z}) \right], \quad \mu \in \Gamma^+_n.$$

We denote $\|\cdot\|^{\mathrm{F}+}_{n,p}$ the spectral domain norms analogous to (1.29) but associated with the unilateral FT:

$$\|x\|^{\mathrm{F}+}_{n,p} := \left\| F^+_n x \right\|_p = \left(\sum_{\mu \in \Gamma^+_n} \left| (F^+_n x)(\mu) \right|^p \right)^{1/p}, \quad p \in [1, +\infty].$$

1.A Linear minimaxity on subspaces

Linear estimators are near-minimax for subspaces. Let us prove that in the case of the pointwise loss $\ell(\widehat{x}, x) = |x_0 - \widehat{x}_0|$ and quadratic risk (1.3),

$$\mathrm{MSE}^*(\mathcal{S}) \le 1.25\,\mathrm{MSE}^{\mathrm{lin}}(\mathcal{S})$$

for any finite-dimensional subspace $\mathcal{S}$ of $\mathbb{C}(\mathbb{Z})$, where MSE^* and $\mathrm{MSE}^{\mathrm{lin}}$ are the minimax and the linear minimax risks. Let $(\mathcal{X}_k)$, $k \in \mathbb{Z}^+$, be an increasing sequence of convex, centrally-symmetric, and compact subsets of $\mathcal{S}$ converging to $\mathcal{S}$, and consider the associated sequences of the risks $\mathrm{MSE}^*(\mathcal{X}_k)$ and $\mathrm{MSE}^{\mathrm{lin}}(\mathcal{X}_k)$. Clearly, $\mathrm{MSE}^*(\mathcal{X}_k) \le 1.25\,\mathrm{MSE}^{\mathrm{lin}}(\mathcal{X}_k)$. On the other hand, the risk of the estimator $\widehat{x}_0 = [\Pi_{\mathcal{S}}(y)]_0$, where $\Pi_{\mathcal{S}}$ is the Euclidean projector on $\mathcal{S}$, is upper-bounded by σ^2 for any $x \in \mathcal{S}$. Hence, $\mathrm{MSE}^{\mathrm{lin}}(\mathcal{S})$ is bounded. Since both sequences $\mathrm{MSE}^*(\mathcal{X}_k)$ and $\mathrm{MSE}^{\mathrm{lin}}(\mathcal{X}_k)$ are non-decreasing, they converge, and one can take the limit, obtaining the desired bound. □

Minimax linear estimator on a subspace. In fact, a linear minimax estimator of x_t on $\mathcal{S} \subset \mathbb{C}_n(\mathbb{Z})$, where $2n + 1 \ge s = \dim(\mathcal{S})$, can be computed explicitly. Indeed, assume w.l.o.g. that $t = 0$, and consider linear estimators of x_0 on a subspace $\mathcal{S} \subset \mathbb{C}_n(\mathbb{Z})$ from the observations $y_{-n}, ..., y_n$. Note that each such estimator has the form $\widehat{x}_0 = \langle \phi, y \rangle$ for some $\phi \in \mathbb{C}_n(\mathbb{Z})$. Its squared quadratic risk is $|x_0 - \langle \phi, x \rangle|^2 + \sigma^2 \|\phi\|_2^2$. It is clear that any linear minimax estimator $\widehat{x}^o = \langle \phi^o, y \rangle$ is unbiased on $\mathcal{S}$. Indeed, let $\langle e_0 - \phi, \tilde{x} \rangle \neq 0$ for some $\tilde{x}$. Then, we can drive the risk of $\widehat{x}^o$ to infinity by taking the sequence of points $x^{(k)} = k\tilde{x}$, which contradicts the fact that this risk is finite (see above). Thus, we conclude that

$$\phi^o \in \operatorname*{Argmin}_{\phi \in \mathbb{C}_n(\mathbb{Z})} \{ \|\phi\|_2 : \quad \langle \phi, x \rangle = x_0 \quad \forall x \in \mathcal{S} \}\,. \tag{1.31}$$

Introducing the vector $[e_0]_\tau = \mathbb{1}\{\tau = 0\}$ such that $x_0 = \langle e_0, x \rangle$, and identifying $\mathbb{C}_n(\mathbb{Z})$ and $\mathbb{C}^{2n+1}$, we arrive at the following characterization:

$$\phi^o \in \operatorname*{Argmin}_{\phi \in \mathbb{C}^{2n+1}} \{ \|\phi\|_2 : A\phi = Ae_0 \}\,,$$

where $A \in \mathbb{C}^{s \times (2n+1)}$ is a matrix whose rows form a basis for $\mathcal{S}_n = \mathcal{S} \cap \mathbb{C}_n(\mathbb{Z})$. Then,

$$\phi^o = A^\dagger A e_0 = \Pi_{\mathcal{S}_n}[e_0],$$

where $A^\dagger = A^{\mathrm{H}}(AA^{\mathrm{H}})^{-1}$ is the right pseudo-inverse of A, and $\Pi_{\mathcal{S}_n}$ is the projector on $\mathcal{S}_n$. □

Chapter 2

Uniform-fit estimators

In this chapter, which is based on [HJNO15], we describe adaptive estimators for recoverable signals (cf. Definition 1.2.1), and study their statistical properties. We focus on the task of pointwise estimation, and our main contribution is a study of the *uniform-fit* estimators – (1.15) and its counterparts – which turn out to be well-suited for this task. Specificaly, in the case where the risk is given by the length of the confidence interval for the true value of x_t, we show that one can adapt to a well-performing time-invariant linear oracle within a logarithmic in n factor; moreover, the adaptive estimator is given as an optimal solution of a convex program, and as such, can be computed efficiently. However, the estimator (1.15) has a drawback: to compute it, one needs an *a priori* guess on the norm of the oracle filter. Hence, we present the analogues of (1.15) which are adaptive to the unknown filter norm parameter. Overall, our contributions are as follows:

- We provide a lower bound to show that the extra logarithmic factor is unavoidable.
- We describe an extension of the approach to the prediction problem, where the goal is to predict the signal x_t based on the previous observations $(y_{t-h-\tau})_{0\leq\tau\leq n}$ for $h \geq 0$.
- We present bandwidth-adaptive versions of the proposed approaches, where the appropriate window n is tuned in a data-driven manner, using a similar scheme to the so-called Lepski's method in nonparametric estimation [Lep91].

We conclude this chapter with preliminary numerical experiments on synthetic data in order to demonstrate the viability of our approach. The proofs are deferred until Section 2.4.

2.1 Problem statement

Recall that our primal interest is in estimating a deterministic signal $x = (x_\tau)_{\tau\in\mathbb{Z}}$ from noisy observations

$$y_\tau = x_\tau + \sigma\zeta_\tau. \tag{2.1}$$

For convenience, we assume that both the signal and the noise are complex-valued, with $\zeta_\tau = \zeta_\tau^{(1)} + i\zeta_\tau^{(2)}$, where $\zeta_\tau^{(j)} \sim \mathcal{N}(0,1)$ are independent for each $\tau \in \mathbb{Z}$ and $j \in \{1,2\}$. We are first interested in pointwise estimation, that is estimating the signal x_t at a fixed time instant t. The estimation procedure shall be local in the time domain, *i.e.*we are allowed to use only

observations $(y_{t+\tau})_{|\tau|\le Cn}$ in the $O(n)$-neighborhood of t for $n \in \mathbb{Z}^+$. 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 $(y_{t-h-\tau})_{0\le\tau\le Cn}$ for $h \ge 0$.

2.1.1 Recoverable signals

Our goal in this section is to formulate our *a priori* assumptions on the signals. Recall that we introduced *recoverable signals* in Section 1.2, see Definition 1.2.1, as a class of signals amenable to pointwise estimation by a time-invariant linear estimator. Following [JN09], it will be convenient for us to use an extended version of Definition 1.2.1 which provides a separate control of the variance and bias, and also allows for different widths of the filter and the "estimation region".

Definition 2.1.1 (Recoverable signals). *Let $m \in \mathbb{Z}^+$, $n \in \mathbb{Z}^+ \cup \{+\infty\}$, $t \in \mathbb{Z}$, and $\rho, \theta \in \mathbb{R}^+$. We say that $x \in \mathbb{C}(\mathbb{Z})$* is recoverable at t with parameters (m, n, ρ, θ), *denoted as*

$$x \in \mathcal{R}^t_{m,n}(\rho, \theta)$$

if there exists a filter $\phi^o \in \mathbb{C}_m(\mathbb{Z})$ which satisfies

$$\|\phi^o\|_{m,2} \le \frac{\rho}{\sqrt{2m+1}}, \tag{2.2}$$

and

$$|x_\tau - [\phi^o * x]_\tau| \le \frac{\sigma\theta\rho}{\sqrt{2m+1}} \quad \forall \tau : |\tau - t| \le m + n. \tag{2.3}$$

Signals in the class $\mathcal{R}^0_{m,n}(\rho, \theta)$ are simply called recoverable (with parameters (m, n, ρ, θ)), *and in this case we write $\mathcal{R}_{m,n}(\rho, \theta)$. Finally, the class of signals $\mathcal{P}(\rho, \theta)$ that are $(m, \infty, \rho, \theta)$-recoverable for all $m \in \mathbb{Z}^+$, with ρ and θ fixed independently of m, is called* (ρ, θ)-parametric.

Remark 2.1.1. A direct consequence of relations (2.2)–(2.3) is that if $x \in \mathcal{R}^t_{m,n}(\theta, \rho)$, the estimate $\phi^o * y$ of x corresponding to the oracle ϕ^o satisfies the following for all $\tau : |\tau - t| \le m + n$:

$$\mathbb{E}\,|[x - \phi^o * y]_\tau|^2 \le \frac{\sigma^2(1+\theta^2)\rho^2}{2m+1}. \tag{2.4}$$

Moreover, recalling that the $(1-\alpha)$-quantile of the χ^2_2 distribution is $2\log[\alpha^{-1}]$, one has

$$\mathbb{P}\left\{|[x - \phi^o * y]_\tau| \le |x_\tau - [\phi^o * x]_\tau| + \sigma\,|[\phi^o * \zeta]_\tau| \le \frac{\sigma(\theta + \sqrt{\log[\alpha^{-1}]})\rho}{\sqrt{2m+1}}\right\} \ge 1 - \alpha. \tag{2.5}$$

Remark 2.1.2. Note that putting $n = 2m$ in (2.4), we obtain that the bound (1.9), see Definition 1.2.1, holds for recoverable signals in the sense of Definition 2.1.1, with ρ substituted by $\rho\sqrt{1+\theta^2}$. Hence, the two definitions are equivalent, up to a constant factor in the parameters, whenever m and n are of the same order, and $\theta = O(1)$. In particular, this is the case for the specific examples of signals considered in Chapter 1:

- Smooth signals are recoverable with ρ depending only on the "degree of smoothness" – the order of differentiability – by the same argument as the one from Section 1.2; as for θ, it can be thought of as a numerical constant which specifies the desired bias-variance balance;

- Exponential polynomials of degree s are recoverable with $\rho = O(\sqrt{s})$ and $\theta = 0$ as it will be shown in Chapter 4, see (1.19)–(1.20).

Remark 2.1.3. It is essential to emphasize that the "realm" of all recoverable signals enjoys a certain *algebraic structure*, being closed with respect to a number of operations: scaling, taking linear combinations, amplitude and frequency modulation [JN09]. Given some basic families of signals (such as discretized smooth functions and harmonic oscillations), one can build new classes of recoverable signals with the class parameters readily given by the calculus rules.

2.1.2 Limit of performance

The following theorem gives a lower bound on the risk of recovering a simple signal, as measured by the width of a confidence interval.

Theorem 2.1.1 ([HJNO15, Theorem 2]). *For any $n \in \mathbb{Z}^{++}$, integer $m \geq 2$, and ρ satisfying*

$$1 \leq \rho \leq m^{\beta}, \quad \beta < 1/4,$$

one can point out a family $\mathcal{F}_{m,n}(\rho)$ of signals from $\mathbb{C}(\mathbb{Z})$ such that the following holds:

(I) *for each signal $x \in \mathcal{F}_{m,n}(\rho)$ there exists a filter $\phi \in \mathbb{C}_m(\mathbb{Z})$ satisfying*

$$\|\phi\|_2 \leq \frac{\rho}{\sqrt{2m+1}}, \quad \|x - \phi * x\|_{n,\infty} = 0,$$

that is, $\mathcal{F}_{m,n}(\rho)$ is a subset of the set $\mathcal{R}_{m,n}(\rho, \theta)$ of recoverable signals at $t = 0$;

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

$$\sup_{x \in \mathcal{F}_{m,n}(\rho)} \mathbb{P}\left\{|x_0 - \widehat{x}_0| \geq C\sigma\rho^2\sqrt{\frac{(1-4\beta)\log m}{m+n}}\right\} \geq 1/8. \tag{2.6}$$

Comparing (2.6) with (2.5), we see that the multiplicative factor $\rho\sqrt{\log m}$ is the unavoidable price of adaptation to the oracle ϕ^o in the pointwise setting.

2.2 Adaptive estimators

Without loss of generality, let us consider the problem of estimation of the value x_t at time instance $t = 0$ given the observations (y_τ), $-2n \leq |\tau - t| \leq 2n$. Our approach relies on the following key assumption, cf. (1.13)–(1.14), which states the existence of a new oracle φ^o with enhanced properties.

Assumption 2.2.1. *Let $t \in \mathbb{Z}$ be fixed. We assume the existence of $\varphi^o \in \mathbb{C}_n(\mathbb{Z})$ such that*

(a) *φ^o is of small ℓ_1-norm in the frequency domain:*

$$\|\varphi^o\|^{\mathrm{F}}_{n,1} \leq \frac{\varrho}{\sqrt{2n+1}};$$

(b) *signal* $x \in \mathbb{C}(\mathbb{Z})$ *satisfies*

$$\|\Delta^{-t}[x - \varphi^o * x]\|_{n,\infty} \le \frac{\sigma\theta\varrho}{\sqrt{2n+1}},$$

in other words, the bias of φ^o *as applied to the signal* x *is uniformly bounded in the* n*-neighbourhood* $[t-n, t+n]$ *of* t*.*

Remark 2.2.1. Note that using the Parseval identity (1.30), part (a) of Assumption 2.2.1 implies

$$\|\varphi^o\|_{n,2} = \|\varphi^o\|^{\mathrm{F}}_{n,2} \le \|\varphi^o\|^{\mathrm{F}}_{n,1} \le \frac{\varrho}{\sqrt{2n+1}},$$

thus the estimator $\varphi^o * y$ of x satisfies

$$\mathbb{E}([x - \varphi^o * y]_\tau)^2 \le \frac{\sigma^2(1+\theta^2)\varrho^2}{2n+1}, \quad |\tau - t| \le n,$$

or, with probability at least $1-\alpha$,

$$|[x - \varphi^o * y]_\tau| \le |x_\tau - [\varphi^o * x]_\tau| + \sigma|[\varphi^o * \zeta]_\tau| \le \frac{\sigma(\theta + \sqrt{\ln[\alpha^{-1}]})\varrho}{\sqrt{2n+1}}.$$

In other words, we have the direct analogues of (2.4) and (2.5) in which ρ gets replaced with ϱ.

2.2.1 Uniform-fit estimators

We now consider a family of estimators of the form $\widehat{x} = \widehat{\varphi} * y$, where the adaptive filter $\widehat{\varphi}$ is given as an optimal solution of a certain optimization problem. We use the umbrella term *uniform-fit estimators*, since the crucial ingredient of these optimization problems is minimization of the ℓ_∞-norm of the Fourier-domain discrepancy. In particular, we consider the following estimators (introduced with $m \ne n$ for the purpose of referencing from other chapters and used only with $m = n$ in the rest of this chapter).

- *Constrained uniform-fit* esitmator, first proposed in [Nem91], see also [GN97] and [JN09], is given for $\bar{\varrho} \ge 0$ by

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m(\mathbb{Z})} \left\{ \left\|\Delta^{-t}[y - \varphi * y]\right\|^{\mathrm{F}}_{n,\infty} : \ \|\varphi\|^{\mathrm{F}}_{m,1} \le \frac{\bar{\varrho}}{\sqrt{2m+1}} \right\}; \qquad \textbf{(Con-UF)}$$

- *Epigraph uniform-fit* estimator [HJNO15] is given as the φ-component of an optimal solution $(\widehat{\varphi}, \widehat{r})$ of

$$\min_{\substack{r \ge 0 \\ \varphi \in \mathbb{C}_m(\mathbb{Z})}} \left\{ r : \begin{array}{l} \|\varphi\|^{\mathrm{F}}_{m,1} \le \frac{r}{\sqrt{2m+1}}, \\ \left\|\Delta^{-t}[y - \varphi * y]\right\|^{\mathrm{F}}_{n,\infty} \le \sigma(\bar{\theta} r + \bar{\Theta}(1+r)) \end{array} \right\} \qquad \textbf{(Epi-UF)}$$

for some $\bar{\theta}, \bar{\Theta} \ge 0$.

- *Penalized uniform-fit* estimator, proposed here for the first time, is given for $\lambda \geq 0$ by

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m(\mathbb{Z})} \left\{ \left\| \Delta^{-t}[y - \varphi * y] \right\|_{n,\infty}^{\mathrm{F}} + \sigma\lambda\sqrt{2m+1}\|\varphi\|_{m,1}^{\mathrm{F}} \right\}. \tag{\textbf{Pen-UF}}$$

We first bound the pointwise loss of (**Con-UF**) under Assumption 2.2.1. In what follows, we use the logarithmic factor parametrized by the confidence parameter $\alpha \in (0, 1]$:

$$\bar{\Theta}_n := 2\sqrt{\log[n+1]} + \sqrt{2\log[1/\alpha]}. \tag{2.7}$$

Proposition 2.2.1. *Suppose that Assumption 2.2.1 holds, and let $\widehat{\varphi}$ be an optimal solution to (**Con-UF**) with $m = n$ and $\bar{\varrho} \geq \varrho$. Then with probability at least $1 - \alpha$,*

$$|[x - \widehat{\varphi} * y]_t| \leq \frac{\sigma(3\bar{\Theta}_{2n} + 2\theta)\bar{\varrho}(1 + \bar{\varrho})}{\sqrt{2n+1}},$$

The constrained estimator requires the knowledge of ϱ. This requirement can be relaxed for the penalized and the epigraph estimators; instead, some bounds on θ and σ are necessary.

Proposition 2.2.2. *Suppose that Assumption 2.2.1 holds, and let $(\widehat{\varphi}, \widehat{\varrho})$ be an optimal solution to (**Epi-UF**) with $m = n$, $\bar{\theta} \geq \theta$, and $\bar{\Theta} \geq \bar{\Theta}_{2n}$. Then with probability at least $1 - \alpha$,*

$$|[x - \widehat{\varphi} * y]_t| \leq \frac{\sigma(3\bar{\Theta} + 2\bar{\theta})\varrho(1 + \varrho)}{\sqrt{2n+1}}.$$

Proposition 2.2.3. *Suppose that Assumption 2.2.1 holds with $\varrho \geq 1$, and let $\widehat{\varphi}$ be an optimal solution to (**Pen-UF**) with $m = n$ and $\lambda \geq 2\bar{\Theta}_{2n} + \theta$. Then with probability at least $1 - \alpha$,*

$$|[x - \widehat{\varphi} * y]_t| \leq \frac{5\sigma\lambda\varrho^2}{\sqrt{2n+1}}.$$

Uniform-fit estimators for recoverable signals. Given a signal satisfying Definition 2.1.1, the crucial question is under what circumstances an oracle φ^o satisfying Assumption 2.2.1 exists, or in other words, how large is parameter ϱ in terms of the recoverability parameters (ρ, θ). A partial answer to this question is provided by the following key result.

Proposition 2.2.4 (Oracle auto-convolution [HJNO15, Proposition 3]). *Let $\phi^o \in \mathbb{C}_m(\mathbb{Z})$ be an oracle filter corresponding to $x \in \mathcal{R}_{m,n}^t(\rho, \theta)$, and let $\varphi^o = \phi^o * \phi^o$. Then,*

(I) *the ℓ_1-norm of $\varphi^o \in \mathbb{C}_{2m}(\mathbb{Z})$ in the frequency domain is small:*

$$\|\varphi^o\|_{2m,1}^{\mathrm{F}} \leq \frac{2\rho^2}{\sqrt{4m+1}}.$$

(II) *φ^o reproduces x with small bias in the n-neighborhood of t, namely,*

$$|x_\tau - [\varphi^o * x]_\tau| \leq \frac{\sigma\theta\rho(1+\rho)}{\sqrt{2m+1}}, \quad |\tau - t| \leq n.$$

Let us now suppose that an oracle filter ϕ^o corresponds to a recoverable signal x at t with parameters $(m_0, 2m_0, \theta, \rho)$, $\rho \geq 1$. When choosing φ^o to be the auto-convolution of ϕ^o, $\varphi^o \in \mathbb{C}_{2m_0}(\mathbb{Z})$, we conclude by Proposition 2.2.4 that Assumption 2.2.1 holds with $n = 2m_0$ and $\varrho = 2\sqrt{2}\rho^2$. As a result, for this choice of φ^o we obtain that convolution-type estimator $\widehat{x}_t = [\widehat{\varphi} * y]_t$, where $\widehat{\varphi}$ can be any of the estimators corresponding to (**Con-UF**), (**Epi-UF**), or (**Pen-UF**) with "proper" choice of parameters, with high probability satisfies

$$|x_t - \widehat{x}_t| \leq O\left(\frac{\sigma\rho^4(1+\theta)\bar{\Theta}_n}{\sqrt{2n+1}}\right).$$

To be specific, we state the result for (**Epi-UF**) – a direct corollary of Proposition 2.2.2. Its counterparts for other two estimators are straightforward.

Theorem 2.2.1 ([**HJNO15, Theorem 5**]). *Suppose that $x \in \mathcal{R}^t_{m_0,2m_0}(\rho,\theta)$ for a given $m_0 \in \mathbb{Z}^{++}$, where $\rho \geq 1$ and $\theta > 0$. Then, estimate $\widehat{x}_t[n,y]$ of x_t, obtained as an optimal solution to (**Epi-UF**) with $n = 2m_0$, $\bar{\theta} \geq \theta$, and $\bar{\Theta} \geq \bar{\Theta}_{2n}$, satisfies with probability at least $1-\alpha$:*

$$|x_t - \widehat{x}_t[n,y]| \leq \frac{\sigma(3\bar{\Theta}_{2n} + 2\bar{\theta})\varrho(1+\varrho)}{\sqrt{2n+1}}, \quad \textit{where } \varrho = 2\sqrt{2}\rho^2.$$

Theorem 2.2.1 states that the risk of the adaptive estimator is equivalent to the risk of the ideal (oracle) filter ϕ^o up to the factor $O(\rho^3\sqrt{\log n})$. We note without proof that when given an independent copy of the observations y, this factor can be reduced to $O(\rho^2\sqrt{\log n})$[1]. On the other hand, the lower bound of Theorem 2.1.1 states that this factor – the "price for adaptation" – is $\Omega(\rho\sqrt{\log n})$. We conclude that there is a gap of $O(\rho^2)$ between the lower and the upper bound (resp. $O(\rho)$ in the case where two independent samples are available). As we will show in Chapter 3, this gap can be reduced, under additional assumptions, using least-squares estimators.

2.2.2 Bandwidth adaptation

Suppose that we are given observations (y_τ), $t - 2N \leq \tau \leq t + 2N$, and we are interested in recovering x_t. Let $\varrho(n) \geq 1$, $n \geq 1$, be monotonous non-increasing function of n: for $n' \geq n$, $\varrho(n') \leq \varrho(n)$. Now assume that for all $1 \leq n \leq n_*(x)[\leq N]$ there is a $\varphi^o(n) \in \mathbb{C}_n(\mathbb{Z})$ such that

$$\|\varphi^o(n)\|^{\mathrm{F}}_{n,1} \leq \frac{\varrho(n)}{\sqrt{2n+1}}, \quad \text{and} \quad \|\Delta^{-t}[x - \varphi^o(n) * y]\|_{n,\infty} \leq \frac{\sigma\theta\varrho(n)}{\sqrt{2n+1}} \tag{2.8}$$

for a given $\theta > 0$. What we do not know is what is $\varphi^o(n)$, and what is the best window parameter $n_*(x)$. The objective is to estimate x_t with essentially the same accuracy as if we knew the best value $n_*(x)$ of n. This can be done, using (**Con-UF**) or (**Epi-UF**) as subroutines, in the following manner. Using (**Epi-UF**) to be specific, let

$$\bar{\vartheta}_N := \sqrt{6\log[2N+1]} + \sqrt{2\log[1/\alpha]}, \tag{2.9}$$

[1] When the noise entries are i.i.d. $\mathcal{N}(0,\sigma^2)$ with known σ^2, it is always possible to obtain such copy, since $y^{(1)} = y + \sigma\eta$ and $y^{(2)} = y - \sigma\eta$ will be independent whenever η is independent of ζ and has the same distribution. One should note, however, that this will result in multiplying the noise variance by a factor of 2.

and for $\bar{\theta} \geq \theta$ define

$$\varepsilon(n) = \frac{\sigma(3\bar{\vartheta}_N + 2\bar{\theta})\varrho(n)(1+\varrho(n))}{\sqrt{2n+1}}. \tag{2.10}$$

Now, for any $n \leq N$, let $\widehat{x}_t[n,y]$ be the estimator of x_t yielded by an optimal solution $(\widehat{\varphi}(n,y), \widehat{\varrho}(n))$ to (**Epi-UF**) with the setup n, $\bar{\theta} \geq \theta$, and $\bar{\Theta} \geq \bar{\vartheta}_N$.

We say that $\widehat{x}_t[n,y]$ is *admissible*, if for all $0 \leq n' \leq n$,

$$|\widehat{x}_t[n',y] - \widehat{x}_t[n,y]| \leq \varepsilon(n') + \varepsilon(n). \tag{2.11}$$

On one hand, admissible values of n clearly exist (*e.g.* $n = 0$), and moreover, the property of a given n to be admissible is observable. On the other hand, one can deduce from the definition of $n_*(x)$, cf. the proof of Theorem 2.2.2 below, that for any $n \leq n_*(x)$ with high probability it holds

$$|x_t - \widehat{x}_t[n,y]| \leq \varepsilon(n);$$

moreover, the additional factor $\sqrt{\log N}$ in $\varepsilon(n)$ guarantees that the above holds *simultaneously* over the range of all $n \leq n_*(x)$. As a consequence, with high probability (2.11) holds for $n_*(x)$ so that $n_*(x)$ is admissible. Based on this fact, we propose a very simple bandwidth adaptation procedure outlined in Algorithm 1; in fact, it is a routine application of Lepski's method [Lep91]. We proceed by computing pointwise estimators $\widehat{x}_t[n,y]$ for all values $n \geq N$, and selecting the one corresponding to the *largest admissible value* $\widehat{n}(y)$. The error of this estimator is then guaranteed to be within a constant factor from $\varepsilon(n_*(x))$ since, due to the admissibility of $n_*(x)$, with high probability one has $\widehat{n}(y) \geq n_*(x)$, whence

$$\begin{aligned} |x_t - \widehat{x}_t[\widehat{n}(y),y]| &\leq |x_t - \widehat{x}_t[n_*(x),y]| + |\widehat{x}_t[n_*(x),y] - \widehat{x}_t[\widehat{n}(y),y]| \\ &\leq 2\varepsilon(n_*(x)) + \varepsilon(\widehat{n}(y)) \leq 3\varepsilon(n_*(x)) \end{aligned}$$

where the last inequality is due to $\varepsilon(n)$ being non-increasing in n. To summarize, we have the following theorem whose formal argument is given in Section 2.4.

Theorem 2.2.2 ([**HJNO15, Theorem 6**])**.** *Let $\bar{x}_t$ be the final estimate yielded by Algorithm 1, the setup for the adaptive estimator being $(\sigma, \varrho(\cdot), \theta, \alpha, N)$, and suppose that x satisfies (2.8) for $1 \leq n \leq n_*(x)$. Then, with probability at least $1-\alpha$ it holds*

$$|\bar{x}_t - x_t| \leq 3\varepsilon(n_*(x)) = \frac{3\sigma(3\bar{\vartheta}_N + 2\bar{\theta})\varrho(n_*(x))(1+\varrho(n_*(x)))\sigma}{\sqrt{2n_*(x)+1}}.$$

As stated by the theorem, the price to pay for not knowing the optimal bandwidth $n_*(x)$ is that logarithmic factor $\log(n_*(x))$ is replaced with $\log(N)$ in the risk bounds.

Remark 2.2.2. It is clear from the proof of Theorem 2.2.2 that the bound of the theorem remains true, up to a constant factor, if instead of the full search over $\{1,...,N\}$ in the for-loop of Algorithm 1, one iterates over the dyadic grid $\{1,2,4,...,2^{\lfloor \log_2(N) \rfloor}\}$. This allows to save tremendous amount of computation when implementing Algorithm 1 in practice.

Algorithm 1 : `Pointwise Recovery with Bandwidth Adaptation`

Input: $\theta, \sigma > 0$, $\alpha \in (0, 1]$, $N \in \mathbb{Z}^{++}$, and $\varrho(n)$, $n = 1, ..., N$

1: $\widehat{x}[0, y] := y$ and $\varepsilon(0) = \sigma\bar{\vartheta}_N$;

2: **for** $n = 1, ..., N$ **do**

3: $\widehat{x}_t[n, y] := \widehat{\varphi}(n, y) * y$,

where $\widehat{\varphi}(n, y)$ is the φ-component of an optimal solution $(\widehat{\varphi}(n, y), \widehat{\varrho}(n))$ to (**Epi-UF**) initialized with n, $\bar{\theta} \geq \theta$, and $\bar{\Theta} \geq \bar{\vartheta}_N$, see (2.9);

4: compute $\varepsilon(n)$ by (2.10);

5: **end for**

6: $\bar{x}_t := \widehat{x}_t[\widehat{n}(y), y]$ where $\widehat{n}(y)$ is the largest admissible $n \in [0, N]$;

Output: $\bar{x}_t$

2.2.3 Prediction setting

***h*-predictable signals.** The proposed approach can also be used to estimate the value of x_t when only noisy observations on the left for $\tau \leq t$ (or only on the right, for $\tau \geq t$) of t are available. Suppose that we aim at estimating the value x_t of the signal $x \in \mathbb{C}(\mathbb{Z})$ at $t \in \mathbb{Z}$, and, for the sake of definiteness, assume that past observations (y_τ), $\tau \leq t - h$ up to the moment $t - h$ are available for a given $h \in \mathbb{Z}^+$ which we refer to as *prediction horizon*. For reader's convenience, we state here some obvious counterparts of the results of Section 2.2 for the prediction setting. The proofs closely follow their counterparts in the case of estimation with bilateral filters, so we omit them.

Definition 2.2.1. *Let $m, h \in \mathbb{Z}^+$, $n \in \mathbb{Z}^+ \cup \{+\infty\}$, $t \in \mathbb{Z}$ and $\rho, \theta \in \mathbb{R}^+$ be fixed. We say that $x \in \mathbb{C}(\mathbb{Z})$ is* h-predictable *at t (or recoverable with respect to h-step prediction) with parameters (m, n, ρ, θ), denoted*

$$x \in \mathcal{R}^t_{m,n,h}(\rho, \theta),$$

if there exists an h-predictive filter *$\phi^o \in \mathbb{C}^h_m(\mathbb{Z})$ such that*

$$\|\phi^o\|_2 \leq \frac{\rho}{\sqrt{m+1}},$$

and

$$|x_\tau - [\phi^o * x]_\tau| \leq \frac{\sigma\theta\rho}{\sqrt{m+1}}, \quad \forall \tau : \ t - m - n - 3h \leq \tau \leq t.$$

Remark 2.2.3 (Interpolation). We can extend the above definition by allowing h to be negative; then, $h < 0$ corresponds to *interpolation* by a filter in the class $\mathbb{C}^{-|h|}_m(\mathbb{Z})$ *i.e.* vanishing outside $[-|h|, n - |h|]$. Such extended definition will be used in Section 3.2.3 of Chapter 3.

As in the case of bilateral estimation, we assume the existence of an oracle filter with small ℓ_1-norm of the Fourier transform.

Assumption 2.2.2. *Let $t \in \mathbb{Z}$ be fixed. There exists a filter $\varphi^o \in \mathbb{C}^h_n(\mathbb{Z})$ such that*

(a) *φ^o is of small ℓ_1-norm in the frequency domain:*[2]

$$\|\Delta^{-h}\varphi^o\|^{\mathrm{F}+}_{n,1} \le \frac{\varrho}{\sqrt{n+1}};$$

(b) *signal $x \in \mathbb{C}(\mathbb{Z})$ satisfies*

$$|x_\tau - [\varphi^o * x]_\tau| \le \frac{\sigma\theta\varrho}{\sqrt{n+1}}, \quad \forall \tau : \ t-n-h \le \tau \le t,$$

in other words, the bias of φ^o as applied to the signal x is uniformly bounded in the unilateral $(n+h)$-neighbourhood of t.

We now present the counterpart of Proposition 2.2.4 for the prediction setting, which provides an oracle φ^o satisfying Assumption 2.2.2.

Proposition 2.2.5. *For $m, n, h \in \mathbb{Z}^+$, let $\phi^o \in \mathbb{C}^h_m(\mathbb{Z})$ be an oracle for $x \in \mathcal{R}^t_{m,n,h}(\rho,\theta)$. Then, there exists a filter $\varphi^o \in \mathbb{C}^{2h}_{2m}(\mathbb{Z})$ – the auto-convolution $\varphi^o = \phi^o * \phi^o$ – which satisfies*

$$\|\Delta^{-2h}\varphi^o\|^{\mathrm{F}+}_{2m,1} \le \frac{2\rho^2}{\sqrt{2m+1}},$$

and reproduces the signal with small bias in the $(n+2h)$-unilateral neighbourhood of t:

$$|x_\tau - [\varphi^o * x]_\tau| \le \frac{\sigma\theta\rho(1+\rho)}{\sqrt{m+1}}, \quad \forall \tau : \ t-n-2h \le \tau \le t$$

We now present the adaptive epigraph prediction estimator. It has the form $\widehat{x}_t = [\widehat{\varphi} * y]_t$, where $\widehat{\varphi}$ is as follows: given setup parameters $m, n, h \in \mathbb{Z}^+$ and $\bar{\theta}, \bar{\Theta} \ge 0$, and observations

$$(y_\tau), \quad t-m-n-2h \le \tau \le t-h,$$

one computes an optimal solution $(\widehat{\varphi}; \widehat{r})$ to the optimization problem

$$\min_{\substack{r \ge 0 \\ \varphi \in \mathbb{C}^h_m(\mathbb{Z})}} \left\{ r : \begin{array}{l} \|\Delta^{-h}[\varphi]\|^{\mathrm{F}+}_{m,1} \le \frac{r}{\sqrt{m+1}}, \\ \left\|\Delta^{h+n-t}[y - \varphi * y]\right\|^{\mathrm{F}+}_{n,\infty} \le \sigma(\bar{\theta} r + \bar{\Theta}(1+r)) \end{array} \right\}. \qquad \textbf{(Epi-UF-Pred)}$$

Proposition 2.2.6. *Suppose that Assumption 2.2.2 holds, and let $(\widehat{\varphi}, \widehat{\varrho})$ be an optimal solution to* **(Epi-UF-Pred)** *with $m = n$, $\bar{\theta} \ge \theta$, and $\bar{\Theta} \ge \bar{\Theta}_{n+h}$, see (2.7). Then with probability $\ge 1-\alpha$,*

$$|x_t - [\widehat{\varphi} * y]_t| \le \frac{\sigma(3\bar{\Theta} + 2\bar{\theta})\varrho(1+\varrho)}{\sqrt{n+1}}.$$

Propositions 2.2.5 and 2.2.6 together imply the following counterpart of Theorem 2.2.1.

Theorem 2.2.3 ([HJNO15, Theorem 10]). *Suppose that x is predictable at t, specifically, $x \in \mathcal{R}^t_{m_0,2m_0,h_0}(\rho,\theta)$ for given $m_0, h_0 \in \mathbb{Z}^+$, $\rho \ge 1$, and $\theta \ge 0$. Then, estimate $\widehat{x}_t[n,h,y]$ of x_t,*

[2] Recall that F^+_n is the *unilateral* Fourier transform, and $\|\cdot\|^{\mathrm{F}}_{n,p}$ are the associated norms, see Section 1.7.

obtained as an optimal solution to (**Epi-UF-Pred**) *with parameters* $m = n = 2m_0$, $h = 2h_0$, $\bar{\theta} \geq \theta$, *and* $\bar{\Theta} \geq \bar{\Theta}_{n+h}$, *satisfies with probability at least* $1 - \alpha$*:*

$$|x_t - \widehat{x}_t[n, h, y]| \leq \frac{\sigma(3\bar{\Theta} + 2\bar{\theta})\varrho(1 + \varrho)}{\sqrt{n+1}}, \quad \textit{where } \varrho = 2\sqrt{2}\rho^2.$$

2.3 Experiments

The experiments in this section are mainly intended for illustration purposes; extensive Monte-Carlo simulations will be presented in Chapter 3.

Implementation. The optimization problems presented in Section 2.2.1 are well-structured convex optimization problems, namely, second-order cone programs. Such problems can be efficiently solved to high-accuracy using interior-point methods. In particular, in the experiments of this chapter we used the state-of-the-art solver Mosek [AA13] with Matlab interface[3].

The estimated signals are generalized harmonic oscillations – members of the parametric class $\mathcal{H}_s$ of solutions to homogeneous linear difference equations $p(\Delta)x = 0$ with a polynomial $p(\cdot)$ of degree s with the roots on the unit circle. As discussed in Sections 1.4–1.5, such signals are naturally suited to our framework, see Chapter 4 for further discussion. On the other hand, harmonic oscillations with a fixed vector $\boldsymbol{\omega}$ of the frequencies of individual harmonics span a linear subspace $\mathcal{S}$ of dimension s specified by $\boldsymbol{\omega}$ (see Section 1.4). Thus, we are dealing with the subspace recovery problem: the signal belongs to an unknown low-dimensional subspace of dimension s of $\mathbb{C}(\mathbb{Z})$, and the challenge is to achieve a nearly-parametric estimation rate without the knowledge of the underlying structure of the signal as specified by the subspace. Meanwhile, in this case, we actually know the structure of generated signals, that is, subspace $\mathcal{S}$ or polynomial $p(\cdot)$ in (1.17), we may just compute a linear oracle explicitly, as we explain below.

We compare the pointwise estimation performance of the proposed approach and the linear time-invariant oracle, referred to, respectively, as *Adaptive Recovery* and *LTI Oracle* in the figures. The experimental setup is as follows. We assume that $2n = 200$ noisy observations of the signal over the regular grid are available. In each experiment, the signal comes from a given shift-invariant subspace $\mathcal{S}$. The objective is to predict the signal at the time instants $t = n + 1, ..., 2n$ using the preceding observations, which corresponds to predictive recovery with the prediction horizon $h = 0$ in Section 2.2.3; specifically, we used the constrained predictive recovery – the constrained counterpart of (**Epi-UF-Pred**) – with $\bar{\varrho} = 2s$, where s was the order of the polynomial $p(\cdot)$, cf. Section 1.4. We compare this adaptive recovery with the *linear time-invariant oracle*, defined as an optimal solution of the following optimization problem:

$$\begin{array}{ll} \min\limits_{\phi \in \mathbb{C}_n^+(\mathbb{Z})} & \|\phi\|_{n,2}^+ \\ \text{s.t.} & x_t - [\phi * x]_t = 0, \quad t = n+1, ..., 2n, \quad \forall x \in \mathcal{S}, \end{array} \tag{2.12}$$

that is, the filter of the minimal ℓ_2-norm in the class $\mathbb{C}_n^+(\mathbb{Z})$ which exactly reproduces any signal from $\mathcal{S}$. The motivation for choosing such filter is as follows: linear minimax estimator of x_t is necessarily unbiased (see Section 1.A), and (2.12) corresponds to a *time-invariant* unbiased

[3] Recall that in Chapter 5 we describe implementation of the proposed estimators via first-order algorithms.

linear estimator with minimal pointwise risk. Note that (2.12) can be solved explicitly after expressing the constraint as a system of linear equations in ϕ.

We proceed according to the following experimental protocol. First, we generate the signal according to one of the scenarios described below. Then we perform $N = 10$ Monte-Carlo trials; in each trial, we add the noise and compute the two estimators from noisy observations. Finally, we plot the sample-based confidence bound corresponding to 2 standard deviations (or 95%-confidence) for each estimator.

- In Figure 2-1, we present results for recovery of a harmonic oscillation with one (real) component – the signal

$$x_\tau = \cos(2\pi\tau/\sqrt{5}), \quad \tau = 1, ..., 200,$$

observed with SNR $= 0.5$.

- In Figure 2-2, we present results for recovery of a harmonic oscillation with 4 real components, with frequencies sampled uniformly at random from $[0, 2\pi]$ and amplitudes sampled uniformly at random from $[0, 1]$. This signal is also observed with SNR $= 0.5$.

- The third example of signal is a polynomially-modulated oscillation,

$$x_\tau = \sum_{j=1}^{\ell} c_j \tau^{m_j - 1} e^{i\omega_j \tau}, \quad 1 \le \tau \le 200, \tag{2.13}$$

with random (ω_j), $1 \le j \le \ell$, sampled independently uniformly at random from $[0, 2\pi]$. In Figure 2-3, we present the results of the experiment for the signal (2.13) 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 $[-1, 1]$. The signal is observed with SNR $= 2$.

2.4 Proofs

2.4.1 Proof of Theorem 2.1.1

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

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

We put

$$\beta = \sigma\sqrt{\ln[\ell/\varrho]}. \tag{2.14}$$

Note that for $m \ge 2$,

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

so we have

$$\beta \ge \sigma\sqrt{(1-2\alpha)\log[(2m+1)/2]}.$$

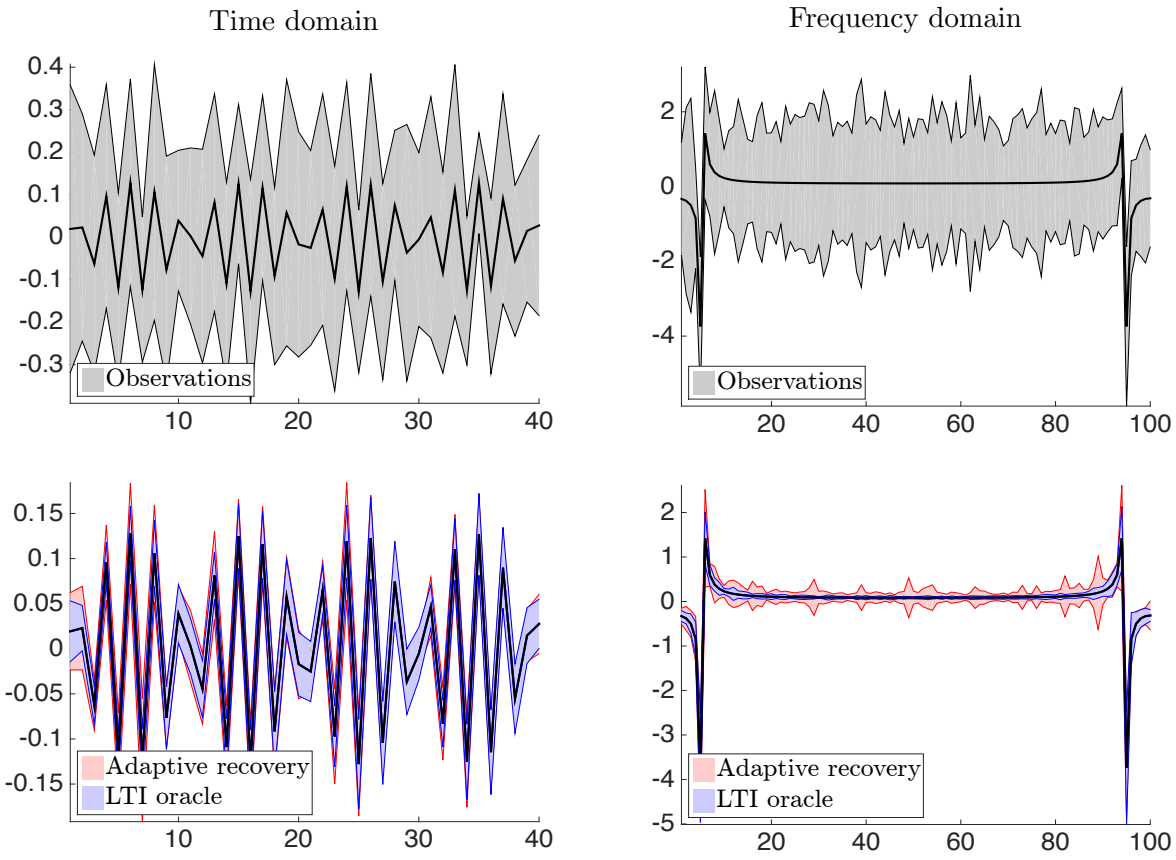

Figure 2-1. Recovery of a harmonic oscillation with one frequency (only real part shown, in natural scale), observed with SNR = 0.5. Left: the last 40 samples in the time domain; right: FT of the last 100 samples. SNR attained: 3.62 for the LTI oracle, 1.49 for the adaptive recovery.

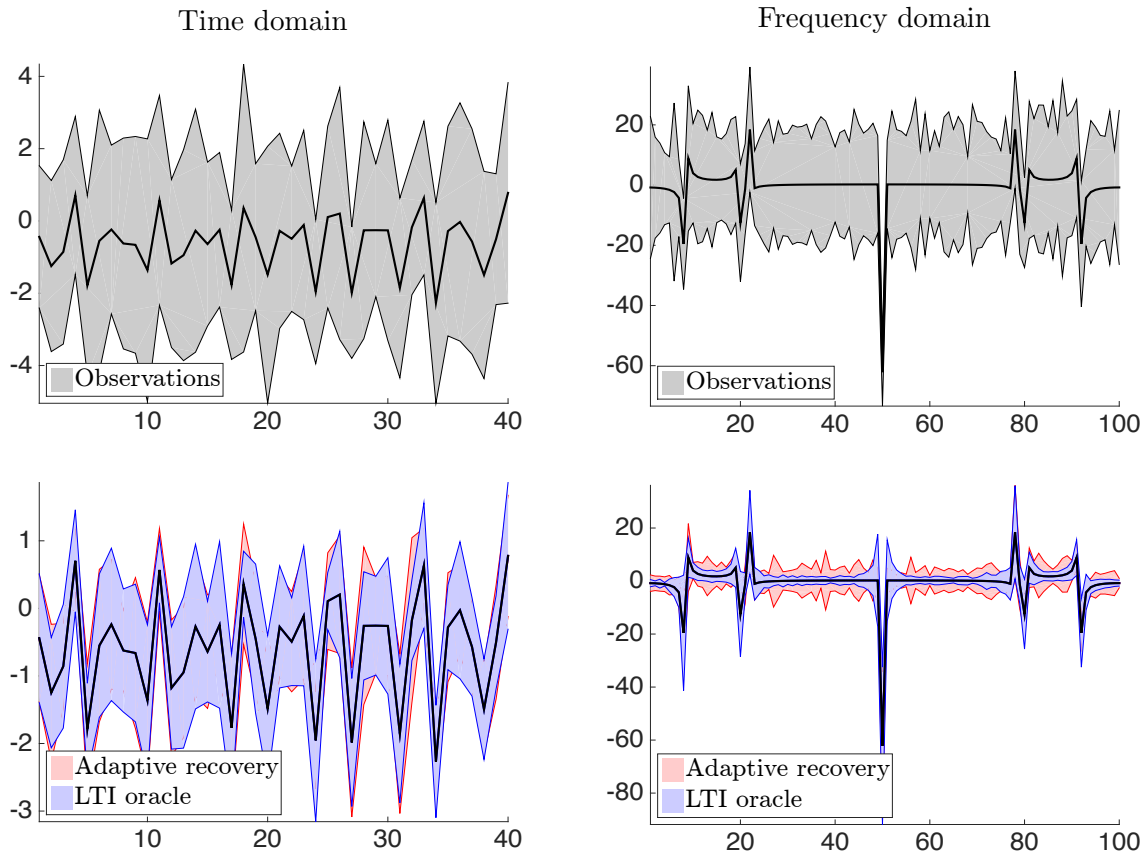

Figure 2-2. Recovery of a harmonic oscillation with 4 real frequencies (only real part shown, in natural scale) observed with SNR = 0.5. Left: the last 40 samples in the time domain; right: FT of the last 100 samples. SNR attained: 1.52 for the LTI oracle, 1.19 for the adaptive recovery.

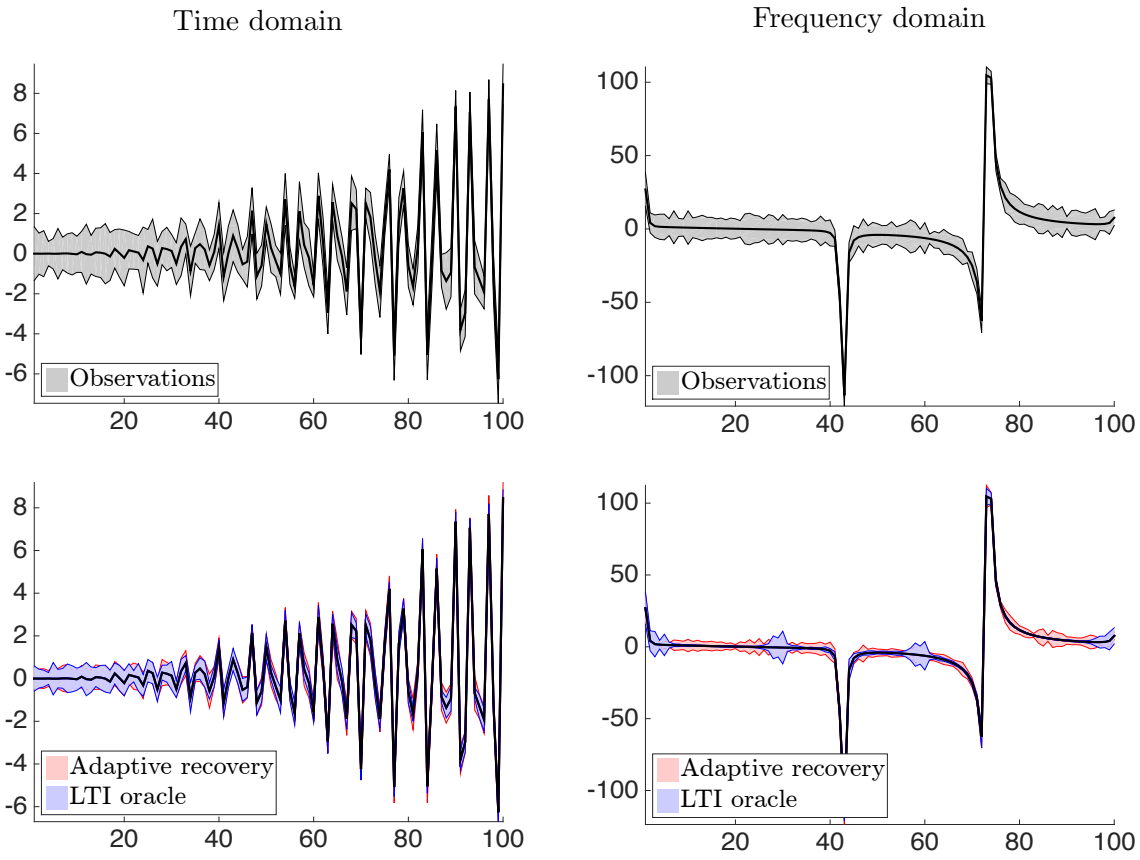

Figure 2-3. Recovery of an amplitude-modulated oscillation (only real part shown, in natural scale), observed with SNR = 2. Left: the last 100 samples in the time domain; right: FT of these samples. SNR attained: 4.40 for the LTI oracle, 3.63 for the adaptive recovery.

Now consider the following family $\mathcal{F}_{m,n}(\rho)$ of signals $x \in \mathbb{C}(\mathbb{Z})$. Divide the set Γ_m of complex roots of unity of degree $2m+1$ into ϱ buckets, each containing ℓ elements, as follows: denoting $\mu = \exp(2\pi i/(2m+1))$, the j-th bucket is

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

The family $\mathcal{F}_{m,n}(\rho)$ consists of the zero signal $x^{(0)} \equiv 0$, and all signals $x^{(r)} \in \mathbb{C}_{n'}(\mathbb{Z})$, $r = (r_1, \ldots, r_\varrho)$, $n' = m+n$, of the form

$$\begin{aligned} x_k^{(r)} &= \frac{\beta}{\sqrt{2n'+1}} \sum_{j=1}^{\varrho} \mu^{r_j k} \mathbb{1}\{|k| \le n'\} \\ &= \frac{\beta}{\sqrt{2n'+1}} \sum_{j=1}^{\varrho} \exp\left(\frac{2\pi i r_j k}{2m+1}\right) \mathbb{1}\{|k| \le n'\}, \quad r_j \in \{\ell(j-1), \ldots, \ell j - 1\}. \end{aligned}$$

Step 1^o. Let us verify the statement (I) of the theorem. The signal $x^{(0)}$ is trivially recovered, whereas for $x^{(r)}$, $r = (r_1 \ldots r_\varrho)$, consider $\phi^{(r)} = \sum_{j=1}^{\varrho} \phi^{(r_j)} \in \mathbb{C}_m(\mathbb{Z})$ with

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

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

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

Hence $x^{(r)} \in \mathcal{R}_{m,n}(\rho, 0)$ as required.

Step 2^o. We shall rely on the following lemma which is a standard route to prove lower bounds, cf. [Tsy08], [Was06]. Consider the problem of testing zero hypothesis H_0 about the distribution of an observation $z \in \mathcal{Z}$ against an alternative H_1, and recall a standard definition: any measurable function $T(z)$ with values in $\{0, 1\}$, such that $T(z) = 1$ implies that H_0 is rejected, is called a *statistical test*.

Lemma 2.4.1. *Consider the problem of testing the hypothesis $H_0 : x \equiv 0$ against $H_1 : x \in \mathcal{X}$. Let the measure $\mathbb{P}_0$ correspond to H_0, the measure $\mathbb{P}_x$ correspond to x for each $x \in \mathcal{X}$, and suppose that Radon-Nikodym derivative $\mathrm{d}\mathbb{P}_x(z)/\mathrm{d}\mathbb{P}_0(z)$ exists for each $x \in \mathcal{X}$ and $z \in \mathbb{Z}$. Finally, assume that we are given a prior distribution π of x such that $\pi(\zeta \in \mathcal{X}) \ge 1 - \varepsilon_\pi$. Then for any statistical test T satisfying*

$$\varepsilon_0(T) := \mathbb{P}_0(T = 1) \le \rho,$$

we have

$$\varepsilon_1(T) := \sup_{x \in \mathcal{X}} \mathbb{P}_x(T = 0) \ge \frac{1}{2} - \rho \mathbb{E}_0[L_\pi^2(z)] - \varepsilon_\pi, \tag{2.15}$$

where $\mathbb{E}_0$ *is the expectation over* $\mathbb{P}_0$*, and* $L_\pi(z)$ *is the Bayesian likelihood ratio:*

$$L_\pi(z) = \int L(z,x)\mathrm{d}\pi(x) = \int \frac{\mathrm{d}\mathbb{P}_x(z)}{\mathrm{d}\mathbb{P}_0(z)}\mathrm{d}\pi(x)$$

Proof of the lemma. Let $\mathcal{R} = \{z : T(z) = 1\}$ be the rejection region of T, and let $\mathcal{R}^c$ be its complement in $\mathcal{Z}$. We have $\mathbb{P}_0(\mathcal{R}) \le \rho$. Next, for any $t \ge 0$ we obtain

$$\begin{aligned}
\varepsilon_1(T) &= \sup_{x\in\mathcal{X}} \int_{\mathcal{R}^c} \mathrm{d}\mathbb{P}_x(z) \\
&\ge \int \mathrm{d}\pi(x) \int_{\mathcal{R}^c} \mathrm{d}\mathbb{P}_x(z) - \varepsilon_\pi \\
&= \int \mathrm{d}\pi(x) \int_{\mathcal{R}^c} L(x,z)\mathrm{d}\mathbb{P}_0(z) - \varepsilon_\pi \\
&= \int_{\mathcal{R}^c} L_\pi(z)\mathrm{d}\mathbb{P}_0(z) - \varepsilon_\pi \\
&\ge \int_{\mathcal{R}^c} L_\pi(z)\mathrm{d}\mathbb{P}_0(z) + t(\mathbb{P}_0(\mathcal{R}) - \rho) - \varepsilon_\pi \\
&\ge \int \min(t, L_\pi(z))\mathrm{d}\mathbb{P}_0(z) - t\rho - \varepsilon_\pi,
\end{aligned}$$

where the last identity is by Fubini's theorem. Recall that $\min(a,b) = (a+b-|a-b|)/2$, and

$$\mathbb{E}_0[L_\pi(z)] = \int \mathrm{d}\mathbb{P}_x(z)\mathrm{d}\pi(x) = 1.$$

Thus,

$$\begin{aligned}
\varepsilon_1(T) &\ge \frac{t+1}{2} - \frac{1}{2}\mathbb{E}_0[|t - L_\pi(z)|] - t\rho - \varepsilon_\pi \\
&\ge \frac{t+1}{2} - \frac{1}{2}\left(\mathbb{E}_0[(t - L_\pi(y))^2]\right)^{1/2} - t\rho - \varepsilon_\pi.
\end{aligned} \tag{2.16}$$

We have $\mathbb{E}_0[(t - L_\pi(y))^2] = t^2 - 2t + \mathbb{E}_0[L_\pi^2(y)]$, and

$$\begin{aligned}
(\mathbb{E}_0[(t - L_\pi(y))^2])^{1/2} &= (t^2 - 2t + \mathbb{E}_0[L_\pi^2(y)])^{1/2} \\
&\le t\left(1 - t^{-1} + (2t^2)^{-1}\mathbb{E}_0[L_\pi^2(y)]\right).
\end{aligned}$$

When substituting the above bound into (2.16), we obtain

$$\begin{aligned}
\varepsilon_1(T) &\ge \frac{1}{2} + \frac{t}{2}\left(t^{-1} - (2t^2)^{-1}\mathbb{E}_0[L_\pi^2(y)]\right) - t\rho - \varepsilon_\pi \\
&= 1 - (2t)^{-1}\mathbb{E}_0[L_\pi^2(y)] - t\rho - \varepsilon_\pi,
\end{aligned}$$

which gives (2.15) for $t = (2\rho)^{-1}$. □

Step 3^o. Let us consider now the following hypothesis testing problem.

Given an observation y *as in* (1.1)*, we want to test simple hypothesis* $H_0 : x = x^{(0)}$ *against the composite alternative* H_1 *stating that* x *is one of* $x^{(r)} \in \mathcal{F}_{m,n}(\rho)$*,* $r \ne 0$*.*

We are to use Lemma 2.4.1 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 $L(y,r) = \mathrm{d}\mathbb{P}_r/\mathrm{d}\mathbb{P}_0$ the likelihood ratio, where $\mathbb{P}_r$ is the normal distribution of the observation $y = x^{(r)} + \sigma\zeta$, and $\mathbb{P}_0$ is the distribution of the noise. Then,

$$\begin{aligned} L(y,r) &= \prod_{\tau=-n'}^{n'} \exp\left(\frac{1}{2\sigma^2}\left[\overline{x_\tau^{(r)}}\zeta_\tau + x_\tau^{(r)}\bar{\zeta}_\tau - |x_\tau^{(r)}|^2\right]\right) \\ &= \exp\left(\frac{1}{2\sigma^2}\left[\langle [\zeta]_{-n'}^{n'}, [x^{(r)}]_{-n'}^{n'}\rangle + \langle [x^{(r)}]_{-n'}^{n'}, [\zeta]_{-n'}^{n'}\rangle - \|x^{(r)}\|_{n',2}^2\right]\right). \end{aligned}$$

Let us denote $X^{(r)} = F_{n'}x$, and $\varsigma = F_{n'}\zeta$. Note that $X^{(r)}$ is real by construction – we have

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

On the other hand, ς_k, $0 \le k \le 2n'$, are independent standard complex-valued Gaussian random variables. Using the Parseval identity (1.30), we get

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

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, $j = 1, ..., \varrho$. Also, let $L_\pi(y) = \mathbb{E}_\pi L(y,r)$ be the likelihood expectation under the prior distribution π:

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

Clearly, $\mathbb{E}_0 L_\pi(y) = 1$ where the external expectation is over the noise distribution under H_0. Let us compute $\mathbb{E}_0[L_\pi^2(y)]$. We have $\mathbb{E}_0[L_\pi^2(y)] = \prod_{j=1}^{\varrho} I_j$ where

$$\begin{aligned} I_j &= \mathbb{E}_0\left[\left\{\frac{1}{\ell}\sum_{k=\ell(j-1)}^{\ell j-1}\exp\left(\frac{2\beta\eta_k-\beta^2}{2\sigma^2}\right)\right\}^2\right] \\ &= (1-\ell)+\frac{1}{\ell^2}\sum_{k=\ell(j-1)}^{\ell j-1}\mathbb{E}_0\exp\left(\frac{2\beta\eta_k-\beta^2}{\sigma^2}\right) \\ &= 1+\frac{1}{\ell}\left[\exp\left(\frac{\beta^2}{\sigma^2}\right)-1\right] \le \exp\left[\frac{1}{\ell}\exp\left(\frac{\beta^2}{\sigma^2}\right)\right]. \end{aligned}$$

We conclude that

$$\mathbb{E}_0[L_\pi^2(y)] \le \exp\left[\frac{\varrho}{\ell}\exp\left(\frac{\beta^2}{\sigma^2}\right)\right] \le e.$$

We now apply Lemma 2.4.1 in our setting with $\rho = 1/8$ and $\varepsilon_\pi = 0$. We conclude that, for any test T with $\varepsilon_0(T) \le 1/8$, we have

$$\varepsilon_1(T) \ge \frac{1}{2} - \frac{e}{8} > 1/8.$$

Step 4^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 $1/8$

$$|\widehat{x}_0 - x_0| \ge \frac{\beta\varrho}{2\sqrt{2n'+1}}.$$

Note that when $x^{(r)} \in \mathcal{F}_{m,n}(\rho)$, we have $x^{(r)}(0) = \frac{\beta\varrho}{\sqrt{2n'+1}}$ for $r \ne 0$, while $x_0^{(0)} = 0$. Let us consider the test $\widehat{T}$ for distinguishing between H_0 and H_1 as in the testing problem of Step 3^o as follows: $\widehat{T}$ rejects H_0 if $\widehat{x}_0 > \beta\varrho/(2\sqrt{2n'+1})$ 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 Step 3^o. We conclude that there is no estimator $\widehat{x}_0$ of x_0 using the observation y and such that with probability $\le 1/8$,

$$|\widehat{x}_0 - x_0| \ge \frac{\beta\varrho}{2\sqrt{2n'+1}} \ge \frac{\sigma\rho^2}{4}\sqrt{\frac{(1-2\alpha)\log[(2m+1)/2]}{2n'+1}} \ge \frac{\sigma\rho^2}{4}\sqrt{\frac{(1-2\alpha)\log m}{2(m+n)+1}},$$

and statement (II) of the theorem is proved. □

2.4.2 Proofs of Propositions 2.2.1–2.2.3

For the sake of clarity, we shall present the proofs for $t = 0$. Extension to arbitrary $t \in \mathbb{Z}$ is straightforward.

Control of the stochastic term. We start with the following simple fact. The random variables $|F_n\zeta(\mu)|^2$, $\mu \in \Gamma_n$, are i.i.d. and distributed according to the χ^2_2 law. Thus,

$$\mathbb{P}\left\{|F_{2n}\zeta(\mu)| \le q\right\} = 1 - e^{-q/2},$$

and

$$\mathbb{P}\left\{\max_{\mu\in\Gamma(n)} |F_n\zeta(\mu)| \le u\right\} = \left(1 - e^{-u^2/2}\right)^{2n+1},$$

so that

$$\mathbb{P}\left\{\|\zeta\|^{\mathrm{F}}_{n,\infty} \ge \varkappa\right\} = \epsilon \;\text{ for }\; \varkappa = \sqrt{-2\log[1-(1-\epsilon)^{1/(2n+1)}]}. \tag{2.17}$$

On the other hand,

$$\begin{aligned}\mathbb{P}\left\{\|\zeta\|^{\mathrm{F}}_{n,\infty} \ge \sqrt{2\ln[2n+1]} + u\right\} &\le (2n+1)\exp\left(-\frac{1}{2}(u+\sqrt{2\log[2n+1]})^2\right)\\ &\le e^{-u^2/2}.\end{aligned}$$

Recall that Δ is the left shift operator, defined for all $x \in \mathbb{C}(\mathbb{Z})$ as $[\Delta x]_\tau := [x]_{\tau-1}$, and let

$$\Theta_n(\zeta) := \max_{-n\le\tau\le n} \|\Delta^{-\tau}\zeta\|^{\mathrm{F}}_{n,\infty}.$$

Then

$$\begin{aligned}\mathbb{P}\left\{\Theta_n(\zeta) \ge 2\sqrt{\log[2n+1]} + u\right\} &\le (2n+1)^2\exp(-(u+2\sqrt{\log[2n+1]})^2/2)\\ &\le e^{-u^2/2}.\end{aligned} \tag{2.18}$$

Let Ξ_α be the subset of noise realizations such that

$$\Theta_n(\zeta) \le \bar{\Theta}_{2n} \;\; \forall \zeta \in \Xi_\alpha$$

where $\bar{\Theta}_n$ was defined in (2.7). Then by (2.18) it follows that $\mathbb{P}(\Xi_\alpha) \ge 1-\alpha$.

Proof of Proposition 2.2.1. We begin with the following decomposition (recall that $\varphi * y = \varphi(\Delta)y$):

$$\begin{aligned}|[x - \widehat{\varphi}(\Delta)y]_0| &\le \sigma|[\widehat{\varphi}*\zeta]_0| + \overbrace{|[x-\widehat{\varphi}(\Delta)x]_0|}^{\delta}\\ &\le \sigma\|\widehat{\varphi}\|^{\mathrm{F}}_{n,1}\|\zeta\|^{\mathrm{F}}_{n,\infty} + \delta\\ &\le \frac{\varrho\sigma\Theta_n(\zeta)}{\sqrt{2n+1}} + \delta.\end{aligned} \tag{2.19}$$

To obtain the second line we used that $[\widehat{\varphi} * \zeta]_0 = \langle \overline{\zeta^n_{-n}}, \widehat{\varphi}_n^{-n} \rangle$, and for the last line we used Assumption 2.2.1.a. Now we need to bound δ to get the bound of the theorem. We have

$$\begin{aligned} \delta = |[(1-\widehat{\varphi}(\Delta))x]_0| &\leq |[(1-\widehat{\varphi}(\Delta))(1-\varphi^o(\Delta))x]_0| + |[\varphi^o(\Delta)(1-\widehat{\varphi}(\Delta))x]_0| \\ &\leq (1+\|\widehat{\varphi}\|_1)\,\|(1-\varphi^o(\Delta))x\|_{n,\infty} + \|\varphi^o\|^{\mathrm{F}}_{n,1}\|(1-\widehat{\varphi}(\Delta))x\|^{\mathrm{F}}_{n,\infty}. \end{aligned}$$

Discrepancy of the oracle φ^o in the time domain can be bounded using Assumption 2.2.1.b:

$$\|(1-\varphi^o(\Delta))x\|_{n,\infty} \leq \frac{\theta\varrho\sigma}{\sqrt{2n+1}}.$$

On the other hand, using that due to $\bar{\varrho} \geq \varrho$ the oracle φ^o is feasible in (**Con-UF**), and the spectral discrepancy term can be bounded as follows:

$$\begin{aligned} \|(1-\widehat{\varphi}(\Delta))x\|^*_{n,\infty} &\leq \|(1-\widehat{\varphi}(\Delta))y\|^{\mathrm{F}}_{n,\infty} + \sigma\|(1-\widehat{\varphi}(\Delta))\zeta\|^{\mathrm{F}}_{n,\infty} \\ &\leq \|(1-\widehat{\varphi}(\Delta))y\|^{\mathrm{F}}_{n,\infty} + \sigma(1+\|\widehat{\varphi}\|_1)\Theta_n(\zeta) \\ &\leq \|(1-\varphi^o(\Delta))y\|^{\mathrm{F}}_{n,\infty} + \sigma(1+\|\widehat{\varphi}\|_1)\Theta_n(\zeta) \\ &\leq \|(1-\varphi^o(\Delta))x\|^{\mathrm{F}}_{n,\infty} + \sigma(2+\|\varphi^o\|_1+\|\widehat{\varphi}\|_1)\Theta_n(\zeta). \end{aligned} \tag{2.20}$$

Meanwhile, using Assumption 2.2.1.b, we can bound the oracle discrepancy in the Fourier domain:

$$\begin{aligned} \|(1-\varphi^o(\Delta))x\|^{\mathrm{F}}_{n,\infty} &\leq \|(1-\varphi^o(\Delta))x\|^{\mathrm{F}}_{n,2} \\ &= \|(1-\varphi^o(\Delta))x\|_{n,2} \\ &\leq \sqrt{2n+1}\,\|(1-\varphi^o(\Delta))x\|_{n,\infty} \leq \theta\varrho\sigma. \end{aligned} \tag{2.21}$$

Collecting the above, we obtain

$$\delta \leq (1+\|\widehat{\varphi}\|_1)\frac{\theta\varrho\sigma}{\sqrt{2n+1}} + \sigma\|\varphi^o\|^{\mathrm{F}}_{n,1}\left[\theta\varrho + (2+\|\varphi^o\|_1+\|\widehat{\varphi}\|_1)\Theta_n(\zeta)\right].$$

Note that $\|\varphi^o\|^{\mathrm{F}}_{n,1}$ is bounded by Assumption 2.2.1.a. It remains to bound $\|\varphi^o\|_1$ and $\|\widehat{\varphi}\|_1$:

$$\|\varphi^o\|_1 \leq \sqrt{2n+1}\|\varphi^o\|_2 \leq \sqrt{2n+1}\|\varphi^o\|^{\mathrm{F}}_{n,1} \leq \varrho, \tag{2.22}$$

and similarly $\|\widehat{\varphi}\|_1 \leq \varrho$. Finally, we arrive at

$$|[x-\widehat{\varphi}(\Delta)x]_0| \leq \frac{\varrho\sigma}{\sqrt{2n+1}}\left[\theta(1+2\bar{\varrho}) + 2(1+\bar{\varrho})\Theta_n(\zeta)\right],$$

and the bound of Proposition 2.2.1 follows using (2.19). □

Proof of Proposition 2.2.2. Inspecting the previous proof, we note that under the premise of Proposition 2.2.2, whenever $\bar{\Theta}_{2n} \geq \Theta_n(\zeta)$ – which happens with probability at least $1-\alpha$ due to 1^o – the oracle φ^o is feasible to (**Epi-UF**), hence an optimal solution $\widehat{\varphi}$ of (**Epi-UF**) satisfies

$$\|\widehat{\varphi}\|^{\mathrm{F}}_{n,1} \leq \frac{\varrho}{\sqrt{2n+1}},$$

and proceeding further as in the previous proof, we complete the proof of the proposition. □

Proof of Proposition 2.2.3. The proof goes along the same as that of Proposition 2.2.1. However, we must take into account a different condition for oracle feasibility. Denote $\widehat{\varphi}$ an optimal solution of (**Pen-UF**). Proceeding as in (2.19) and using Assumption 2.2.1, we obtain

$$\begin{aligned}
&|[x - \widehat{\varphi}(\Delta)y]_0| \\
&\leq \sigma\|\widehat{\varphi}\|_{n,1}^{\mathrm{F}}\|\zeta\|_{n,\infty}^{\mathrm{F}} + |[(1 - \widehat{\varphi}(\Delta))x]_0| \\
&\leq \sigma\|\widehat{\varphi}\|_{n,1}^{\mathrm{F}}\|\zeta\|_{n,\infty}^{\mathrm{F}} + \|\varphi^o\|_{n,1}^{\mathrm{F}}\|(1 - \widehat{\varphi}(\Delta))x\|_{n,\infty}^{\mathrm{F}} + (1 + \|\widehat{\varphi}\|_1)\|(1 - \varphi^o(\Delta))x\|_{n,\infty} \\
&\leq \sigma\|\widehat{\varphi}\|_{n,1}^{\mathrm{F}}\Theta_n(\zeta) + \frac{\varrho}{\sqrt{2n+1}}\|(1 - \widehat{\varphi}(\Delta))x\|_{n,\infty}^{\mathrm{F}} + \frac{\theta\varrho\sigma}{\sqrt{2n+1}}(1 + \|\widehat{\varphi}\|_1).
\end{aligned} \tag{2.23}$$

In what follows, let us condition on the event $\Theta_n(\zeta) \leq \bar{\Theta}_{2n}$ the probability of which is $\geq 1 - \alpha$. Feasibility of $\widehat{\varphi}$ for (**Pen-UF**) yields

$$\begin{aligned}
\|(1 - \widehat{\varphi}(\Delta))y\|_{n,\infty}^{\mathrm{F}} + \lambda\sigma\sqrt{2n+1}\|\widehat{\varphi}\|_{n,1}^{\mathrm{F}} &\leq \|(1 - \varphi^o(\Delta))y\|_{n,\infty}^{\mathrm{F}} + \lambda\sigma\sqrt{2n+1}\|\varphi^o\|_{n,1}^{\mathrm{F}} \\
&\leq \theta\varrho\sigma + (1+\varrho)\sigma\Theta_n(\zeta) + \lambda\varrho\sigma \\
&\leq (\theta + 2\Theta_n(\zeta) + \lambda)\varrho\sigma \\
&\leq 2\lambda\varrho\sigma,
\end{aligned} \tag{2.24}$$

Here first we used (2.21), (2.22), and the last line of (2.20), then that $\varrho \geq 1$, and finally in the last line we recalled the choice of λ used in the theorem. Now from (2.24) we obtain

$$\|\widehat{\varphi}\|_{n,1}^{\mathrm{F}} \leq \frac{2\varrho}{\sqrt{2n+1}} \tag{2.25}$$

and

$$1 + \|\widehat{\varphi}\|_1 \leq 1 + \sqrt{2n+1}\|\widehat{\varphi}\|_{n,1}^{\mathrm{F}} \leq 1 + 2\varrho \leq 3\varrho. \tag{2.26}$$

Further, using (2.24) and (2.26), we get

$$\begin{aligned}
\|(1 - \widehat{\varphi}(\Delta))x\|_{n,\infty}^{\mathrm{F}} &\leq \|(1 - \widehat{\varphi}(\Delta))y\|_{n,\infty}^{\mathrm{F}} + \sigma(1 + \|\widehat{\varphi}\|_1)\Theta_n(\zeta) \\
&\leq (2\lambda + 3\Theta_n(\zeta))\varrho\sigma
\end{aligned} \tag{2.27}$$

Substituting (2.25)–(2.27) into (2.23), we arrive at

$$\begin{aligned}
|[x - \widehat{\varphi}(\Delta)y]_0| &\leq \frac{(2\lambda + 5\Theta_n(\zeta) + 2\theta)\varrho^2\sigma}{\sqrt{2n+1}} \\
&\leq \frac{5\lambda\varrho^2\sigma}{\sqrt{2n+1}}.
\end{aligned}$$

□

2.4.3 Proof of Proposition 2.2.4

Let $\phi^o \in \mathbb{C}_m(\mathbb{Z})$, and let $\varphi^o \in \mathbb{C}_{2m}(\mathbb{Z})$ be given as $\varphi^o = \phi^o * \phi^o$. Then,

$$\begin{aligned}
\|\varphi^o\|_1^{\mathrm{F}} &= (4m+1)^{-1/2} \sum_{\mu\in\Gamma_{2m}} |\varphi^o(\mu)| \\
&= (4m+1)^{1/2} \sum_{\mu\in\Gamma_{2m}} \left(\frac{|\phi^o(\mu)|}{(4m+1)^{1/2}}\right)^2 \\
&= (4m+1)^{1/2} [\|\phi^o\|_{2m,2}^{\mathrm{F}}]^2 \\
&= (4m+1)^{1/2} \|\phi^o\|_{2m,2}^2 \\
&= (4m+1)^{1/2} \|\phi^o\|_{m,2}^2. \qquad (2.28)
\end{aligned}$$

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

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

Now let $\phi^o \in \mathbb{C}_m(\mathbb{Z})$ be the oracle filter from Definition 2.1.1, *i.e.* $\|\phi^o\|_2 \le \frac{\rho}{\sqrt{2m+1}}$. Then by (2.28),

$$\begin{aligned}
\|\varphi^o\|_1^{\mathrm{F}} &\le (4m+1)^{1/2} \|\phi^o\|_{m,2}^2 \\
&\le \frac{\sqrt{2}\rho^2}{\sqrt{2m+1}}.
\end{aligned}$$

Moreover, let $x \in \mathcal{R}^t_{m,n}(\rho,\theta)$. Then $\|\varphi^o\|_1 \le \sqrt{2m+1}\|\varphi^o\|_2 \le \rho$, and due to (2.3),

$$\begin{aligned}
|x_\tau - [\varphi^o * x]_\tau| &\le (1 + \|\phi^o\|_1) \max_{|s|\le m} |[x - \phi^o * x]_{\tau-s}| \\
&\le (1+\rho) \frac{\theta\rho\sigma}{\sqrt{2m+1}}
\end{aligned}$$

for all $\tau \in [t-n,\ t+n]$. □

2.4.4 Proof of Theorem 2.2.2

W.l.o.g. we focus on the case $t = 0$. Let us introduce the stochastic term

$$\vartheta_N(\zeta) := \max_{n\le N} \Theta_n(\zeta) = \max_{n\le N} \max_{|\tau|\le n} \|\Delta^{-\tau}\zeta\|_{n,\infty}^{\mathrm{F}}.$$

Proceeding as in part 1^o of the proof of Propositions 2.2.1 and 2.2.2, we can prove that the event $\vartheta_N \le \bar{\vartheta}_N$, where $\bar{\vartheta}_N$ is defined by (2.9), happens with probability $\ge 1-\alpha$. But from the proof of Proposition 2.2.2 (specifically the end of 3^o and 4^o), that event implies that simultaneously for

all $n \le N$ it holds

$$|x_0 - [\widehat{\varphi}(n, y) * y]_0| \le \varepsilon(n),$$

see (2.10). In turn, this implies that simultaneously over all pairs (n', n): $0 \le n' \le n \le n_*(x)$,

$$\left|[\widehat{\varphi}(n, y) * y]_0 - \left[[\widehat{\varphi}(n', y) * y\right]_0\right| \le \varepsilon(n) + \varepsilon(n').$$

As a consequence, $n_*(x)$ is admissible, whence $\widehat{n}(y) \ge n_*(x)$. Finally,

$$\begin{aligned} |x_0 - [\widehat{\varphi}(\widehat{n}(y), y) * y]_0| &\le |x_0 - [\widehat{\varphi}(n_*(x), y) * y]_0| + |[\widehat{\varphi}(\widehat{n}(y), y) * y]_0 - [\widehat{\varphi}(n_*(x), y) * y]_0| \\ &\le 2\varepsilon(n_*(x)) + \varepsilon(\widehat{n}(y)) \\ &\le 3\varepsilon(n_*(x)). \end{aligned}$$

Here we used first that $\widehat{n}(y)$ is admissible ($\widehat{n}(y) \ge n_*(x)$), and then that $\varepsilon(\cdot)$ is non-increasing. □

Chapter 3

Least-squares estimators

Recall that in the general signal denoising problem, the goal is to estimate a complex-valued *signal* $x \in \mathbb{C}(\mathbb{Z})$ from noisy observations

$$y_\tau = x_\tau + \sigma\zeta_\tau, \quad \tau \in \mathbb{Z}, \tag{3.1}$$

where $\zeta_\tau \sim \mathbb{C}\mathcal{N}(0,1)$ are i.i.d. standard complex-valued Gaussian random variables[1]. Moreover, we justified reduction to the *local* version of the above problem where the goal is to estimate x_t at some point $t \in \mathbb{Z}$, using observations in the neighborhood of size n of this point for some $n \in \mathbb{Z}_+$.

In Chapter 1, we motivated the focus on *linear estimators* which write as

$$\widehat{x}_t = \sum_{\tau=-n}^{n} \phi_\tau y_{t-\tau}, \quad -n \le t \le n,$$

where $\phi \in \mathbb{C}_n(\mathbb{Z})$ is called a *filter*. Linear estimators have been thoroughly studied in various forms, they are both theoretically attractive and easy to use in practice. In particular, we made the following important observation: under a rather general assumption about the set $\mathcal{X} \subset \mathbb{C}(\mathbb{Z})$ of possible signals, the *linear* minimax estimator of x is nearly *minimax* on $\mathcal{X}$, with respect to both the pointwise loss and the ℓ_2-loss. Besides, if $\mathcal{X}$ can be specified in a computationally tractable way, a near-minimax linear estimator can be efficiently computed by solving a convex optimization problem, see [JN17] and references therein. This observation prompts the following approach to filtering: given a computationally tractable set $\mathcal{X}$, one can simply compute a near-minimax linear estimator by feeding $\mathcal{X}$ to a convex optimization algorithm. The strength of this approach, however, comes at a price: the set $\mathcal{X}$ must be specified. Therefore, when one faces a recovery problem *without prior knowledge of* $\mathcal{X}$, this approach cannot be implemented.

In Chapter 2 we presented an alternative, more robust approach to denoising. Instead of requiring $\mathcal{X}$ to be specified beforehand, we assumed the existence of a *linear oracle* – a well-performing linear estimator of x_t. We proved that one can efficiently adapt to the linear oracle filter, provided that the signal x is *recoverable*, meaning that there exists a filter ϕ which recovers x uniformly well in the $O(n)$-sized neighborhood of t, and its ℓ_2-norm is bounded by $O(\rho/\sqrt{n+1})$ for a moderate $\rho \ge 1$ (see Definition 2.1.1 and its simplified version, Definition 1.2.1). The adaptive estimators have the form $\widehat{x} = y * \widehat{\varphi}$, where the adaptive filter $\widehat{\varphi}$ is computed by minimizing the ℓ_∞-norm of the discrepancy $y - y * \varphi$, in the Fourier domain, constrained or

[1] Recall that we use the "Matlab" notation $[a_1; ...; a_n]$ for column vectors; cf. Section 1.7 for all notation.

penalized by the ℓ_1-norm of the filter in the Fourier domain. Compared with the oracle linear filter, the price for adaptation (*i.e.*, the suboptimality factor) for such *uniform-fit* estimators has been proved to be $O(\rho^3\sqrt{\log n})$, with the lower bound of $\Omega(\rho\sqrt{\log n})$ for the pointwise loss.

In this chapter, we study a new family of estimators, obtained by solving a least-squares problem constrained or penalized, as before, by the ℓ_1-norm of the filter in the Fourier domain. Similarly to the previously introduced ones, the new estimators can be efficiently computed using convex optimization, furthermore, their formulation is easily amenable to first-order proximal algorithms (see Chapter 5). We prove exact oracle inequalities for the ℓ_2-loss of these estimators, and show that the price for adaptation improves upon that for the previously introduced estimators to $O(\rho^2 + \rho\sqrt{\log n})$ for the pointwise loss and to $O(\rho + \sqrt{\log n})$ for ℓ_2-loss. However, the improved bounds require an extra assumption that the the signal to recover is *approximately shift-invariant*. Namely, $x \in \mathbb{C}(\mathbb{Z})$ must be representable as the sum of two components:

- the component in a small-dimensional shift-invariant linear subspace $\mathcal{S} \subset \mathbb{C}(\mathbb{Z})$;
- the residual component which is controlled explicitly in ℓ_2-norm.

The shift-invariant subspace, as well as the decomposition itself, can be completely unknown to the statistician. As it was briefly discussed in Chapter 1, this assumption is related to the previously introduced notion of *recoverable* signals. Our analysis of this relation, which will be presented in Chapter 4, led us to the improved bounds for the problem of denoising harmonic oscillations (cf. Section 1.5). After presenting the main theoretical results in the setting of causal filtering, we revisit the interpolation and prediction settings previously presented in Chapter 2, extending the new estimators to these settings. Finally, we present numerical experiments that show the potential of the approach on synthetic and real-world signals.

3.1 Problem statement

Preliminaries. In this chapter, we will make use of convinient notation for neighborhoods of the origin in $\mathbb{Z}$: for any $n \in \mathbb{Z}^+$ and $h \in \mathbb{Z}$,

$$\mathrm{D}_n := \{-n, ..., n\}, \quad \mathrm{D}_n^+ := \{0, ..., n\}, \quad \mathrm{D}_n^- := \{-n, ..., 0\}, \quad \mathrm{D}_n^h := \{h, ..., h+n\}.$$

Our focus in this chapter is primarily on the (local) ℓ_2-loss around $t = 0$ rather than simply on the pointwise loss, and estimation is required in some neighborhood of the point of interest called the *estimation domain*. Hence, we can now make a distinction between the width n of the estimation domain, and the width m of the window of the filter used to estimate the signal on that domain. Specifically, given some $m, n \in \mathbb{Z}_+$, we will from now on assume that noisy observations of x are given on the *observation domain* D_{m+n},

$$y_\tau = x_\tau + \sigma\zeta_\tau, \quad \tau \in \mathrm{D}_{m+n}, \tag{3.2}$$

and consider the two following tasks (see Fig. 3-1):

- *full recovery*, where estimation is required on the whole observation domain D_{m+n};
- *partial recovery*, where the signal is only required to be estimated on subdomain $\mathrm{D}_n \subseteq \mathrm{D}_{m+n}$ using a filter $\varphi \in \mathbb{C}_m(\mathbb{Z})$, preferably with $n \geq cm$ for some constant $c > 0$.

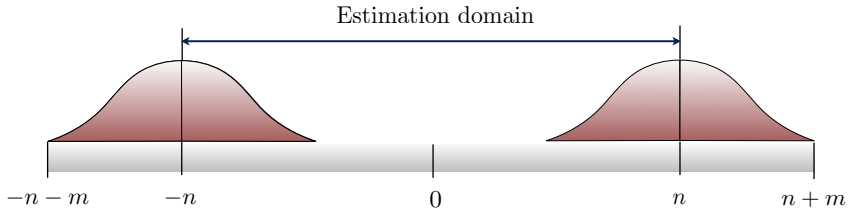

Figure 3-1. Full observation domain D_{m+n} and estimation domain D_n for a convolution-type estimator associated with a bilateral filter $\phi \in \mathbb{C}_m(\mathbb{Z})$.

In this chapter, we restrict attention to partial recovery, returning to full recovery in Chapter 4.

We now recall Definfition 2.1.1 from the previous chapter: signal x is called (m, n, ρ, θ)-*recoverable* (at point $t = 0$), denoted $x \in \mathcal{R}_{m,n}(\rho, \theta)$, if there exists a filter $\phi^o \in \mathbb{C}_m(\mathbb{Z})$ which satisfies

$$\|\phi^o\|_2 \le \frac{\rho}{\sqrt{2m+1}}, \tag{3.3}$$

and

$$|x_\tau - [\phi^o * x]_\tau| \le \frac{\sigma\theta\rho}{\sqrt{2m+1}}, \quad \tau \in \mathrm{D}_{m+n}. \tag{3.4}$$

Using that ϕ^o is non-random, we immediately bound the pointwise risk for such signals[2]:

$$\left[\mathbb{E}|x_\tau - [\phi^o * y]_\tau|^2\right]^{1/2} \le \frac{\sigma\sqrt{1+\theta^2}\rho}{\sqrt{2m+1}}, \quad \tau \in \mathrm{D}_n, \tag{3.5}$$

and, as a consequence, bound the ℓ_2-risk:

$$\left[\mathbb{E}\|x - \phi^o * y\|_{n,2}^2\right]^{1/2} \le \kappa_{m,n}\sigma\sqrt{1+\theta^2}\rho, \tag{3.6}$$

where we defined the ratio

$$\kappa_{m,n} := \sqrt{\frac{2n+1}{2m+1}}.$$

Existence of an oracle. Recall that we are interested in the task of *partial recovery*: estimate the signal on D_n from noisy observations (1.1) on D_{m+n} using a filter $\varphi \in \mathbb{C}_m(\mathbb{Z})$. Without loss of generality, assume for a moment that $m = 2m_0$ for some $m_0 \in \mathbb{Z}$, and that x is (m_0, n, ρ, θ)-recoverable. Similarly to the argument in Chapter 2, cf. Proposition 2.2.4, this would imply the existence of a filter $\varphi^o \in \mathbb{C}_m(\mathbb{Z})$ with small ℓ_1-norm and pointwise risk on $\mathrm{D}_n = \{-n, ..., n\}$. In particular,

$$\|\varphi^o\|_{m,1}^{\mathrm{F}} \le \frac{\varrho}{\sqrt{2m+1}}, \quad \varrho := 2\rho^2, \tag{3.7}$$

and

$$|x_\tau - [\varphi^o * x]_\tau| \le \frac{\sqrt{2}\sigma\theta\varrho}{\sqrt{2m+1}}, \quad \tau \in \mathrm{D}_n, \tag{3.8}$$

[2] While we can bound the risk on D_{m+n}, we avoid it since estimator $[\phi^o * y]_{-(m+n)}^{m+n}$ uses observations on D_{2m+n}.

cf. (3.3) and (3.4), which results in

$$\left[\mathbb{E}|x_\tau - [\varphi^o * y]_\tau|^2\right]^{1/2} \le \frac{\sigma\sqrt{1+2\theta^2}\varrho}{\sqrt{2m+1}}, \quad \tau \in \mathrm{D}_n, \tag{3.9}$$

and

$$\left[\mathbb{E}\|x - \varphi^o * y\|_{n,2}^2\right]^{1/2} \le \sigma\kappa_{m,n}\sqrt{1+2\theta^2}\varrho. \tag{3.10}$$

Unfortunately, this "ideal" estimator is unavailable. We now present *adaptive estimators* that are able to "mimic" the statistical properties of φ^o, as given by (3.9)-(3.10), whenever it exists.

3.1.1 Adaptive estimators

Given $m, n \in \mathbb{Z}_+$ and $\bar{\varrho} > 0$, let $\widehat{\varphi}_{\mathrm{con}}$ be defined as follows:

$$\widehat{\varphi}_{\mathrm{con}} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m(\mathbb{Z})} \left\{ \|y - \varphi * y\|_{n,2}^2 : \|\varphi\|_{m,1}^{\mathrm{F}} \le \frac{\bar{\varrho}}{\sqrt{2m+1}} \right\}, \tag{**Con-LS**}$$

We refer to $\widehat{x} = \widehat{\varphi}_{\mathrm{con}} * y$ as the *constrained (least-squares) estimator*. In the sequel, we will prove a sharp oracle inequality for this estimator, stating that the ℓ_2-loss of this estimator is comparable to the ℓ_2-loss of *any* filter φ feasible to (**Con-LS**), and in particular, to φ^o provided that $\bar{\varrho} = \varrho$. However, this requires the knowledge of ϱ or a non-trivial upper bound on it, which is not always available. Hence, we also propose alternative estimators in which ϱ is not required to be known in advance: the *ordinary penalized estimator* $\widehat{x} = \widehat{\varphi}_{\mathrm{pen}} * y$ where for some $\lambda > 0$ filter $\widehat{\varphi}_{\mathrm{pen}}$ is defined by

$$\widehat{\varphi}_{\mathrm{pen}} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m(\mathbb{Z})} \left\{ \|y - \varphi * y\|_{n,2}^2 + \sigma^2\lambda\sqrt{2m+1}\|\varphi\|_{m,1}^{\mathrm{F}} \right\}, \tag{**Pen-LS**}$$

and the *improved penalized estimator* $\widehat{x} = \widehat{\varphi}_{\mathrm{pen}^2} * y$ with the squared ℓ_1-norm penalty term:

$$\widehat{\varphi}_{\mathrm{pen}^2} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m(\mathbb{Z})} \left\{ \|y - \varphi * y\|_{n,2}^2 + \sigma^2\lambda^2(2m+1)[\|\varphi\|_{m,1}^{\mathrm{F}}]^2 \right\}. \tag{**Pen^2-LS**}$$

As we see, instead of the knowledge of ϱ, some knowledge of noise variance σ^2 is required to properly tune both estimators. In fact, we will show that both penalized estimators with properly selected λ enjoy essentially the same risk bounds as the contrained estimator with the "optimal" choice of $\bar{\varrho}$ – the one balancing the norm and bias of the oracle in the best possible way, with some differences between (**Pen-LS**) and (**Pen^2-LS**) that will be discussed later on. Hence, the practical recommendation is to use penalized estimators (preferably (**Pen^2-LS**) as discussed in Section 3.2.1) whenever it is possible, *i.e.* whenever σ^2 is known or can be estimated from data.

3.2 Theoretical results

3.2.1 Oracle inequalities for ℓ_2-loss

To analyze the adaptive least-squares estimators, we need the following assumption[3]:

Assumption 3.2.1 (Approximate shift-invariance). *$x \in \mathbb{C}(\mathbb{Z})$ admits a decomposition*

$$x = x^{\mathcal{S}} + \varepsilon.$$

Here, $x^{\mathcal{S}} \in \mathcal{S}$, where $\mathcal{S}$ is some shift-invariant linear subspace of $\mathbb{C}(\mathbb{Z})$ with $s := \dim(\mathcal{S}) \leq 2n+1$, and ε is bounded in the ℓ_2-norm: for some $\varkappa \geq 0$ one has

$$\left\|\Delta^{-\tau}\varepsilon\right\|_{n,2} \leq \varkappa\sigma, \quad \tau \in \mathrm{D}_m. \tag{3.11}$$

In other words, Assumption 3.2.1 states the existence of a shift-invariant linear subspace $\mathcal{S}$, $\Delta\mathcal{S} \subseteq \mathcal{S}$ with controlled dimension, to which the signal is close in ℓ_2-norm. Importantly, the decomposition of the signal, as well as the subspace $\mathcal{S}$, can be completely unknown. Besides, Assumption 3.2.1 merits some further remarks.

Remark 3.2.1. Letting the signal to be close, in ℓ_2-norm, to a shift-invariant subspace, instead of simply belonging to the subspace, is essential. It significantly extends the set of signals to which our theory applies, allowing to address the nonparametric setting. For example, signals which are close to discrete-time polynomials, which satisfy homogeneous linear difference equations, and hence belong to a small shift-invariant subspaces, are Sobolev-smooth functions sampled over the uniform grid [JN10]. More interesting examples will be considered in Chapter 4.

Remark 3.2.2. Assumption 3.2.1 looks similar to signal recoverability according to Definition 2.1.1 which also postulates some kind of "invariance" of the signal, claiming that there exists a time-invariant filter which reproduces the signal on a certain interval. However, the actual relationship between the two notions is rather intricate, and will be investigated in Chapter 4.

We now present *oracle inequalities* which relate the ℓ_2-loss of adaptive filter $\widehat{\varphi}$ to the loss of any feasible solution φ to the corresponding optimization problem. These inequalities, interesting for their own sake, will be used later on to obtain performance guarantees for the proposed estimators in ℓ_2-loss and the pointwise loss. We first state the result for the constrained estimator.

Theorem 3.2.1. *Let $m, n \in \mathbb{Z}^+$, and let $\widehat{\varphi}_{\mathrm{con}}$ be an optimal solution to (**Con-LS**) with $\bar{\varrho}$ such that $\bar{\varrho} \geq 1$. Suppose that Assumption 3.2.1 holds with some $(s, \varkappa)$, and let φ be any feasible solution to (**Con-LS**). Then for any $0 < \alpha \leq 1$, the following holds with probability at least $1-\alpha$: $\widehat{\varphi}_{\mathrm{con}}$ satisfies*

$$\|x - \widehat{\varphi}_{\mathrm{con}} * y\|_{n,2} \leq \|x - \varphi * y\|_{n,2} + C\sigma(\bar{\mathrm{Q}})^{1/2}, \tag{3.12}$$

where

$$\bar{\mathrm{Q}} = \bar{\mathrm{Q}}(\bar{\varrho}, s, \varkappa, \kappa_{m,n}, \alpha) := \bar{\varrho}(\kappa_{m,n}^2 + 1)\log[(m+n)/\alpha] + \bar{\varrho}\varkappa\sqrt{\log[1/\alpha]} + s. \tag{3.13}$$

The proof of Theorem 3.2.1 is provided in Section 3.4 as well as those of the other results of this section. As it is explained in Section 3.A, the analysis of (**Con-LS**) relies on the concepts different than those used in the traditional analysis of ℓ_1-penalized estimators.

[3] Recall that the lag operators Δ and Δ^{-1} on $\mathbb{C}(\mathbb{Z})$ are defined by $[\Delta x]_t = x_{t-1}$ and $[\Delta^{-1}x]_t = x_{t+1}$.

We now formulate the counterpart of Theorem 3.2.1 for the penalized estimator (**Pen-LS**).

Theorem 3.2.2. *Let $m, n \in \mathbb{Z}^{++}$, and let $\widehat{\varphi}_{\mathrm{pen}}$ be an optimal solution to* (**Pen-LS**) *with λ satisfying*

$$\lambda \geq \underline{\lambda} := 4\sqrt{2}(1 + 2.25\kappa_{m,n})^2 \log[42(m+n+1)/\alpha] \tag{3.14}$$

with α such that $0 < \alpha \leq 1$. Suppose that Assumption 3.2.1 holds with some $(s, \varkappa)$, and let φ be any feasible solution to (**Pen-LS**). *Then, the following holds with probability at least $1 - \alpha$: $\widehat{\varphi}_{\mathrm{pen}}$ satisfies*

$$\|x - \widehat{\varphi}_{\mathrm{pen}} * y\|_{n,2} \leq \|x - \varphi * y\|_{n,2} + C\sigma(\widehat{Q})^{1/2}, \tag{3.15}$$

where

$$\widehat{Q} = \widehat{Q}(\lambda, \varrho, \widehat{\varrho}, s, \varkappa, \alpha) := \varrho\lambda + (\widehat{\varrho} + 1)\varkappa\sqrt{\log[1/\alpha]} + s, \tag{3.16}$$

and

$$\varrho := \sqrt{2m+1}\|\varphi\|^{\mathrm{F}}_{m,1}, \quad \widehat{\varrho} := \sqrt{2m+1}\|\widehat{\varphi}_{\mathrm{pen}}\|^{\mathrm{F}}_{m,1}.$$

Moreover, the term depending on $\widehat{\varrho}$ can be removed from (3.16) if one sets $\lambda \geq 2\underline{\lambda}$, provided that

$$\varkappa\sqrt{2\log[16/\alpha]} \leq \underline{\lambda}. \tag{3.17}$$

Theorem 3.2.2 shows that in the "parametric regime" of small $\varkappa$, estimator (**Pen-LS**) achieves essentially the same statistical accuracy as the constrained estimator (**Con-LS**) that "knows" the right value of ϱ. However, if one cannot guarantee that the approximation error of the oracle is small as specified by (3.17), the remainder term depends on the unknown $\widehat{\varrho}$; obviously, such a bound is of limited interest. As the next theorem shows, estimator (**Pen^2-LS**) is spared from this disadvantage, at the expense of having a larger price of adaptation.

Theorem 3.2.3. *Let $m, n \in \mathbb{Z}_+$, and let $\widehat{\varphi}_{\mathrm{pen}^2}$ be an optimal solution to* (**Pen^2-LS**) *with some $\lambda > 0$. Suppose that Assumption 3.2.1 holds with some $(s, \varkappa)$, and let φ be any feasible solution to* (**Pen^2-LS**). *Then for any $0 < \alpha \leq 1$, the following holds with probability at least $1 - \alpha$: $\widehat{\varphi}_{\mathrm{pen}^2}$ satisfies*

$$\big\|x - \widehat{\varphi}_{\mathrm{pen}^2} * y\big\|_{n,2} \leq \|x - \varphi * y\|_{n,2} + \sigma(\lambda\varrho + C_1\lambda^{-1}\mathrm{Q}_1 + C_2\mathrm{Q}_2^{1/2}), \tag{3.18}$$

where $\varrho = \sqrt{2m+1}\|\varphi\|^{\mathrm{F}}_{m,1}$, and

$$\begin{aligned}
\mathrm{Q}_1 &= \mathrm{Q}_1(\varkappa, \kappa_{m,n}, \alpha) = (\kappa^2_{m,n} + 1)\log[(m+n)/\alpha] + \varkappa\sqrt{\log[1/\alpha]} + 1, \\
\mathrm{Q}_2 &= \mathrm{Q}_2(\varrho, s, \varkappa, \alpha) = \varrho\log[1/\alpha] + \varkappa\sqrt{\log[1/\alpha]} + s.
\end{aligned}$$

In particular, when choosing $\lambda = \mathrm{Q}_1^{1/2}$, we get

$$\big\|x - \widehat{\varphi}_{\mathrm{pen}^2} * y\big\|_{n,2} \leq \|x - \varphi * y\|_{n,2} + C\sigma(\varrho\mathrm{Q}_1^{1/2} + \mathrm{Q}_2^{1/2}).$$

Note that choosing $\lambda = \sqrt{C_1\mathrm{Q}_1/\varrho}$ in (**Pen^2-LS**) would result in the same remainder term (of the order $\sqrt{\varrho}$) as for the constrained estimator with "optimal" constraint parameter $\bar{\varrho} = \varrho$. Clearly, this choice cannot be implemented since parameter ϱ is unknown. Nevertheless, Theorem 3.2.3 provides us with an implementable choice[4] of λ that still results in an oracle inequality, at the expense of a larger remainder term which scales as ϱ. As a result of this suboptimal choice of λ, oracle inequality (3.18) essentially loses its sharpness when applied to a simple signal. Indeed,

in this case we can only hope that the oracle loss itself scales as ϱ, cf. (3.10). However, in the application to recoverable signals, one is interested in the bounds on the overall risk of $\widehat{\varphi}$ compared to that of φ^o, and not just the remainder term. As we are about to see in an instant, the loss of sharpness in this case is not critical in this case.

3.2.2 Corollaries for recoverable signals

Returning to the above discussion, assume that in addition to Assumption 3.2.1, one has recoverability: $x \in \mathcal{R}_{m_0,n}(\rho,\theta)$ with $m = 2m_0$. Then, it is not hard to prove the following $(1-\alpha)$-confidence bound for the ℓ_2-loss of the oracle $\varphi \in \mathbb{C}_m(\mathbb{Z})$ (cf. Section 3.5):

$$\|x - \varphi^o * y\|_{n,2} \le 4\sigma\kappa_{m,n}\rho^2\big(1+\sqrt{2}\theta+\sqrt{\log[1/\alpha]}\big). \tag{3.19}$$

When combined with Theorems 3.2.1–3.2.3, this bound implies the following result:

Corollary 3.2.1. *Assume that $x \in \mathcal{R}_{m_0,n}(\rho,\theta)$ with $\rho \ge 1/\sqrt{2}$, and let $m = 2m_0 > 0$. Moreover, suppose that Assumption 3.2.1 holds with some $(s,\varkappa)$, and let $\widehat{\varphi}_{\mathrm{con}}$ and $\widehat{\varphi}_{\mathrm{pen}^2}$ be, correspondingly, optimal solutions to (**Con-LS**) with $\bar{\varrho} = 2\rho^2$, and to (**Pen^2-LS**) with λ chosen as in the premise of Theorem 3.2.3. Then for any $0 < \alpha \le 1$, with probability at least $1-\alpha$ one has*

$$\begin{aligned}\|x - \widehat{\varphi}_{\mathrm{con}} * y\|_{n,2} \le\ & 4\sigma\kappa_{m,n}\rho^2(1+\sqrt{2}\theta+\sqrt{\log[1/\alpha]}) \\ & + C\sigma\big\{\rho\sqrt{(\kappa_{m,n}^2+1)\log[(m+n+1)/\alpha]} + \varkappa\sqrt{\log[1/\alpha]} + \sqrt{s}\big\},\end{aligned} \tag{3.20}$$

$$\begin{aligned}\big\|x - \widehat{\varphi}_{\mathrm{pen}^2} * y\big\|_{n,2} \le\ & 4\sigma\kappa_{m,n}\rho^2(1+\sqrt{2}\theta+\sqrt{\log[1/\alpha]}) \\ & + C\sigma\Big\{\rho\sqrt{(\kappa_{m,n}^2+1)\log[(m+n+1)/\alpha]} + (\varkappa+1)\sqrt{\log[1/\alpha]} + \sqrt{s}\Big\}.\end{aligned} \tag{3.21}$$

As a consequence,

$$\big[\mathbb{E}\|x - \widehat{\varphi}_{\mathrm{con}} * y\|_{n,2}^2\big]^{1/2} \le C\sigma\left\{\kappa_{m,n}\rho^2(1+\theta) + \rho\sqrt{(\kappa_{m,n}^2+1)\log(m+n+1)+\varkappa} + \sqrt{s}\right\}, \tag{3.22}$$

$$\big[\mathbb{E}\|x - \widehat{\varphi}_{\mathrm{pen}^2} * y\|_{n,2}^2\big]^{1/2} \le C\sigma\left\{\kappa_{m,n}\rho^2(1+\theta) + \rho^2\sqrt{(\kappa_{m,n}^2+1)\log(m+n+1)+\varkappa} + \sqrt{s}\right\}. \tag{3.23}$$

*Finally, the bounds (3.20) and (3.22) also hold for an optimal solution $\widehat{\varphi}_{\mathrm{pen}}$ to (**Pen-LS**) with $\lambda = \underline{\lambda}$, cf. (3.14), provided that $\varkappa$ satisfies (3.17).*

As we see, the resulting bounds for $\widehat{\varphi}_{\mathrm{con}}$ and $\widehat{\varphi}_{\mathrm{pen}^2}$ coincide up to a logarithmic factor. Moreover, as we will show later on in Chapter 4 (cf. Section 4.1.2), if x *belongs* to a shift-invariant subspace (that is, $\varkappa = 0$), then $x \in \mathcal{R}_{m,n}(\rho,\theta)$ with $\theta = 0$ and $\rho = O(\sqrt{s})$, so that the right-hand sides in (3.22) and (3.23) are both bounded from above with

$$Cs(\kappa_{m,n}\sqrt{\log(m+n)}+1).$$

[4] In contrast to that of ϱ, the knowledge of $\varkappa$ is a reasonable assumption, since in practice one usually fixes $\varkappa$ in advance to ensure the desired bias-variance ratio, see also Proposition 4.1.1 in Chapter 4.

One should realize, however, that Theorems 3.2.1–3.2.3 *per se* are not tied to the particular choice of oracle characterized by (3.7)–(3.8), and moreover, do not assume at all that the signal is simple. Hence, the oracle inequalities would remain useful in case where the bound $\rho = O(\sqrt{s})$ is unavailable – *e.g.* in the situation considered in Section 3.2.3 when dealing with one-sided filters. On the other hand, if there exists, by chance, an oracle with smaller ρ, adaptive estimators are guaranteed to be competitive against it.

Guarantees for pointwise loss. Under the premise of Corollary 3.2.1, *i.e.* when the signal is recoverable according to Definition 2.1.1 and also satisfies Assumption 3.2.1, one can also bound the *pointwise* loss of the adaptive estimators on a subdomain of D_n.

Proposition 3.2.1. *Suppose that the premise of Corollary 3.2.1 holds with $n \geq m_0$ (recall that $m = 2m_0 > 0$), and fix $0 < \alpha \leq 1$. Let $\widehat{\varphi}_{\mathrm{con}}$ be an optimal solution to* (**Con-LS**) *with $\bar{\varrho} = 2\rho^2$, and let $\widehat{\varphi}_{\mathrm{pen}^2}$ be an optimal solution to* (**Pen^2-LS**) *with λ chosen as in the premise of Theorem 3.2.3. Then, for any fixed $t \in \mathrm{D}_{n-m_0}$, the following holds with probability $\geq 1-\alpha$:*

$$|x_t - [\widehat{\varphi}_{\mathrm{con}} * y]_t| \leq \frac{C\rho\,[\text{r.h.s. of (3.20)}]}{\sqrt{2m+1}}, \quad \left|x_t - [\widehat{\varphi}_{\mathrm{pen}^2} * y]_t\right| \leq \frac{C\rho\,[\text{r.h.s. of (3.21)}]}{\sqrt{2m+1}}. \tag{3.24}$$

As a consequence,

$$[\mathbb{E}|x_t - [\widehat{\varphi}_{\mathrm{con}} * y]_t|^2]^{1/2} \leq \frac{C\rho\,[\text{r.h.s. of (3.22)}]}{\sqrt{2m+1}}, \quad [\mathbb{E}|x_t - [\widehat{\varphi}_{\mathrm{pen}^2} * y]_t|^2]^{1/2} \leq \frac{C\rho\,[\text{r.h.s. of (3.23)}]}{\sqrt{2m+1}}. \tag{3.25}$$

Finally, estimator $\widehat{\varphi}_{\mathrm{pen}}$ admits the same guarantees as $\widehat{\varphi}_{\mathrm{con}}$ under the premise of Corollary 3.2.1.

A few remarks are in order.

- We see that at the expense of additional assumption of approximate shift-invariance, least-squares estimators compare favorably with the uniform-fit estimators studied in Chapter 2. Indeed, we showed that when considering the pointwise loss, the price of adaptation for uniform-fit estimators is $O(\rho^3\sqrt{\log(m+n)})$ whenever $x \in \mathcal{R}_{m,n}(\rho,\theta)$, with a lower bound of $\Omega(\rho\sqrt{\log m})$ when $m \geq cn$. Meanwhile, least-squares estimators only suffer the factor of $O(\rho^2 + \rho\sqrt{\log(m+n)})$.
- The new estimators are also advantageous from the computational perspective. As stated by Proposition 3.2.1, one has a pointwise guarantee on the whole subdomain D_{n-m_0}, with *the same* adaptive filter, whereas in the case of uniform-fit estimators, one has to fit a new filter at any target point. Note that the computation of adaptive filters is relatively expensive as it requires to solve a convex program, whereas the computation of the estimate $\widehat{\varphi} * y$ on D_n, once $\widehat{\varphi}$ has been obtained, can be done in time $\tilde{O}(m+n)$ via Fast Fourier transform. In the bandwidth adaptation setting discussed in Section 2.2.2, where the signal lives on the "global" domain D_N and $m+n$ corresponds to the unknown window width parameter, the above observation allows to dramatically reduce the computational price of the proposed estimators, compared to the uniform-fit ones, by applying them in a "blockwise" manner. Namely, the overall domain must be divided into blocks of size

$m+n$ with $n=cm$, and then it suffices to compute only one adaptive filter for each block. Then, one can chose the optimal block size adaptively, separately for any point of D_N, via a Lepski-type procedure akin to the one described in Section 2.2.2.

3.2.3 Prediction setting

So far in this chapter, we only considered the *interpolation* setting where one estimates the signal by tuning a two-sided filter $\widehat{\varphi} \in \mathbb{C}_m(\mathbb{Z})$. Meanwhile, the cases of *filtering* $\varphi \in \mathbb{C}_m^+(\mathbb{Z})$ and *prediction* $\varphi \in \mathbb{C}_m^h(\mathbb{Z})$ can also be of interest, either on their own right, or when full recovery is required, since two-sided filters cannot be used near the borders of the observation domain. We now present an extension of the results obtained in the previous section to the prediction setting; as such, we also cover filtering since the latter corresponds to prediction with $h=0$. We remind that additional notation for the prediction setting has been introduced in Section 1.7.

Following the presentation of Section 2.2.3, we assume that one is given a *horizon* $h \in \mathbb{Z}^+$ and observations (y_τ) such that[5] $-m-n-h \le \tau \le 0$, and our first objective would be to estimate x with small ℓ_2-loss on the interval $[-n,0]$ by taking convolution of y with a predictive filter $\widehat{\varphi} \in \mathbb{C}_m^h(\mathbb{Z})$ fitted from these observations (later on, we will also see that this filter can be used as well to estimate the signal to the right of $t=0$). We consider adaptive predictive filters $\widehat{\varphi}$ that are given as optimal solutions to the following optimization problems:

$$\widehat{\varphi}_{\mathrm{con}} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m^h(\mathbb{Z})} \left\{ \left[\|\Delta^n[y-\varphi*y]\|_{n,2}^+ \right]^2 : \ \left\|\Delta^{-h}[\varphi]\right\|_{m,1}^{\mathrm{F}+} \le \frac{\bar{\varrho}}{\sqrt{m+1}} \right\}, \qquad \textbf{(Con-LS-Pred)}$$

$$\widehat{\varphi}_{\mathrm{pen}} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m^h(\mathbb{Z})} \left\{ \left[\|\Delta^n[y-\varphi*y]\|_{n,2}^+ \right]^2 + \sigma^2\lambda\sqrt{m+1}\left\|\Delta^{-h}[\varphi]\right\|_{m,1}^{\mathrm{F}+} \right\}, \qquad \textbf{(Pen-LS-Pred)}$$

$$\widehat{\varphi}_{\mathrm{pen}^2} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_m^h(\mathbb{Z})} \left\{ \left[\|\Delta^n[y-\varphi*y]\|_{n,2}^+ \right]^2 + \sigma^2\lambda^2(m+1)\left[\left\|\Delta^{-h}[\varphi]\right\|_{m,1}^{\mathrm{F}+}\right]^2 \right\}. \qquad \textbf{(Pen}^{\mathbf{2}}\textbf{-LS-Pred)}$$

A close inspection of the proofs of Theorems 3.2.1–3.2.3 shows that those results remain valid, with obvious adjustments, provided that Assumption 3.2.1 is replaced with the following one:

Assumption 3.2.2. *$x \in \mathbb{C}(\mathbb{Z})$ admits a decomposition $x = x^{\mathcal{S}} + \varepsilon$, where $x^{\mathcal{S}}$ belongs to a shift-invariant subspace $\mathcal{S} \in \mathbb{C}(\mathbb{Z})$ with dimension $s \le n+1$, and ε can be bounded via $\varkappa \ge 0$:*

$$\|\Delta^\tau[\varepsilon]\|_{n,2}^- \le \varkappa\sigma, \quad 0 \le \tau \le h+m. \tag{3.26}$$

We now recall Definition 2.2.1 of predictable signals:given parameters $m,n,h \in \mathbb{Z}^+$, $\rho \ge 1$, and $\theta \ge 0$, signal x is called (m,n,h,ρ,θ)-*predictable* at $t \in \mathbb{Z}$, denoted $x \in \mathcal{R}_{m,n,h}^t(\rho,\theta)$, if there exists a filter $\phi^o \in \mathbb{C}_m^h(\mathbb{Z})$ that satisfies

$$\|\phi^o\|_2 \le \frac{\rho}{\sqrt{m+1}},$$

and

$$|x_\tau - [\phi^o * x]_\tau| \le \frac{\sigma\theta\rho}{\sqrt{m+1}}, \quad t-m-n-3h \le \tau \le t.$$

[5] More precisely, we only need observations on two intervals $[-m-n-h,-h]$ and $[-n,0]$ with combined length at most $m+2n$. As a consequence, we do not have to pay an extra logarithmic factor in h in the risk bounds.

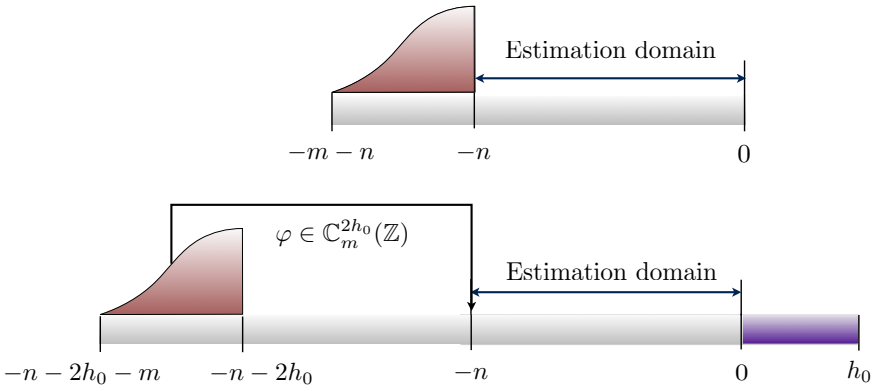

Figure 3-1. Filtering (above) and prediction with horizon $h = 2h_0$ (below). For the sake of clarity, we illustrate prediction with $h < n$ since otherwise the observation domain becomes disconnected. Signal is not observed in the violet region, yet we are able to estimate it (pointwise) in that region as guaranteed by Proposition 3.2.2.

Now, suppose that $x \in \mathcal{R}_{m_0,n,h_0}(\rho,\theta)$, and let $\varrho = 2\rho^2$, $m = 2m_0$, and $h = 2h_0$. By Proposition 2.2.5, we know that there exists a suitable oracle for the predictive estimators introduced above: a filter $\varphi^o \in \mathbb{C}_m^h(\mathbb{Z})$ with the following properties, cf. (3.7)–(3.8):

$$\|\Delta^{-h}[\varphi^o]\|_{m,1}^{\mathrm{F}+} \le \frac{\varrho}{\sqrt{m+1}}, \quad \text{and} \quad |x_\tau - [\varphi^o * x]_\tau| \le \frac{\sqrt{2}\sigma\theta\varrho}{\sqrt{m+1}}, \quad t-n-h \le \tau \le t.$$

Extensions of Corollary 3.2.1 and Proposition 3.2.1 to the prediction setting is now straightforward:

Proposition 3.2.2. *Assume that $x \in \mathcal{R}^t_{m_0,n,h_0}(\rho,\theta)$ with $t = 2h_0$ and $\rho \ge 1/\sqrt{2}$. Let $\varrho = 2\rho^2$, $m = 2m_0 \in \mathbb{Z}^{++}$, and $h = 2h_0 \in \mathbb{Z}^{++}$. Moreover, suppose that Assumption 3.2.1 holds with some $(s,\varkappa)$, and let $\widehat{\varphi}_{\mathrm{con}}$, $\widehat{\varphi}_{\mathrm{pen}}$ and $\widehat{\varphi}_{\mathrm{pen}^2}$ be, correspondingly, optimal solutions to* (**Con-LS-Pred**), (**Pen-LS-Pred**), (**Pen^2-LS-Pred**) *with parameters as in the premises of Theorems 3.2.1–3.2.3.*

1. *Then, quantities*

$$\|\Delta^n[x - \widehat{\varphi}_{\mathrm{con}} * y]\|_{n,2}^+, \quad \|\Delta^n[x - \widehat{\varphi}_{\mathrm{pen}} * y]\|_{n,2}^+, \quad \left\|\Delta^n[x - \widehat{\varphi}_{\mathrm{pen}^2} * y]\right\|_{n,2}^+,$$

   enjoy the same bounds (3.20)–(3.23), up to a multiplicative constant, as their counterparts in the interpolation setting.

2. *Moreover, whenever $n \ge m_0$, quantities*

$$|x_t - [\widehat{\varphi}_{\mathrm{con}} * y]_t|, \quad |x_t - [\widehat{\varphi}_{\mathrm{pen}} * y]_t|, \quad \left|x_t - [\widehat{\varphi}_{\mathrm{pen}^2} * y]_t\right|, \quad h_0 - n + m_0 \le t \le h_0$$

   also enjoy the same bounds (3.24)–(3.25) as their counterparts in the interpolation setting.

Pointwise extrapolation. Note that as claimed in the second part of Proposition 3.2.2, if the signal is predictable, adaptive filters $\widehat{\varphi}_{\mathrm{con}}$ and $\widehat{\varphi}_{\mathrm{pen}}$ allow to *extrapolate* it: observing

$$(y_\tau), \quad -2h_0 - m - n \le \tau \le 0,$$

one is able to estimate x_t at point $t = h_0$, by fitting a filter $\widehat{\varphi} \in \mathbb{C}_m^{2h_0}(\mathbb{Z})$ and evaluating $[\widehat{\varphi} * y]_{h_0}$.

3.3 Experiments

Here we present the results of numerical experiments with proposed adaptive estimators in several application scenarios. We compare the performance of the penalized least-squares estimator (**Pen-LS**) with that of the Lasso estimator of [BTR13] in signal and image denoising problems (as discussed in [BTR13], the latter corresponds to the discretized version of the Atomic Soft Thresholding (AST) algorithm metnioned in Section 1.5). For the full picture, we also include in the comparison the penalized uniform-fit estimator (**Pen-UF**) studied in Chapter 2.

Implementation details. We aim at *full recovery* of signals and images observed on $[-n, n]^d$ where $d = 1$ for signals and $d = 2$ for images[6]. Note that convolution-type estimators presented in Section 3.1.1 only recover the signal on the positive part of the domain, $[0, n]^d$. To obtain full recovery, one can proceed as follows.

- If performance is measured by the ℓ_2-loss of an estimator, as in these experiments, one can compute the total of 2^d one-sided filters, flipping the observations along each axis.
- If a pointwise guarantee is desired, it also suffices to compute $O(2^d)$ filters in a "blockwise manner" as follows from the second part of Proposition 3.2.2. However, the resulting estimate will have visible "tiling" artefacts, which might be practically undesirable: in audio denoising that would lead to high-frequency noise artifacts, while in image denoising one would see "tiling" artifacts. One solution is to replacing the above blockwise estimate with a "continuous" one, in which the interpolation shift h gradually changes over $[-n, 0]$ as one changes the recovery point. However, this approach leads to a tremendous amount of computation as one has to compute $\Omega(n^d)$ filters. An alternative approach is to construct a blockwise estimate with overlapping blocks, obtaining overlapping estimates, and then average over these estimates. As our experience suggests, half-overlapping blocks are largely enough to suppress the "tiling" artefacts in image denoising. This recommendation coincides with the know-how in audio denoising [YMB08].

To compute the penalized least-squares estimator, we used a version of Nesterov's Fast Gradient Method with online stepsize policy, running it for 1000 iterations. The uniform-fit estimator was computed by Composite Mirror Prox (with non-adaptive stepsize) run for 3000 iterations as the online stepsize selection policy has not been implemented in this case[7]. Further details of the implementation of these methods and their behaviour being irrelevant in the context of this dicussion, we describe them *in extenso* in Chapter 5. Finally, discussion of the discretization approach underlying the competing Lasso estimator can be found in [BTR13, Section 3.6].

Experimental setup. We use the same experimental protocol in our signal and image denoising experiments. For each level of the signal-to-noise ratio

$$\mathrm{SNR} \in \{0.25,\ 0.5,\ 1,\ 2,\ 4,\ 8,\ 16\},$$

[6] The generalization of the framework for images is straightforward, hence we omit its formal presentation.

[7] As we found out experimentally, the stepsize selection policy in the case of Nesterov's algorithm takes 2-3 step evaluations, so the total amount of computations of the two algorithms is roughly the same.

we perform $T = 40$ Monte-Carlo trials. In each trial, we generate a random signal x on a regular grid with n points, corrupted by the i.i.d. Gaussian noise of variance σ^2. The signal is normalized: $\|x\|_2 = 1$ so $\mathrm{SNR}^{-1} = \sigma\sqrt{n}$. We set the regularization parameter in each method as follows. For the least-squares estimator, we set $\lambda = 2\log[63n/\alpha]$ as suggested by (3.14) with $\kappa_{m,n} = 1$, with the value of the confidence parameter $\alpha = 0.1$. For the uniform-fit estimator, we used the value $\lambda = 4\sqrt{\log[n/\alpha]}$ as recommended by Proposition 2.2.3. Finally, for Lasso [BTR13], we use the choice $\lambda = \sigma\sqrt{2\log n}$ common in the literature on sparse recovery. We report experimental results by plotting the empirical average of the ℓ_2-loss $\|\widehat{x} - x\|_2$ over T trials (subsequently called "ℓ_2-error"), along with its 95% empirical confidence interval, versus the inverse signal-to-noise ratio SNR^{-1}.

Signal denoising. We consider denoising of a one-dimensional signal in two different scenarios, fixing $T = 40$ and $n = 100$. In the *RandomSpikes* scenario, the signal is a harmonic oscillation (1.25) with 4 components, each characterized by a spike of a random amplitude at a random position in the continuous frequency domain $[0, 2\pi]$. In the *CoherentSpikes* scenario, the same number of spikes is sampled by pairs. Spikes in each pair have the same amplitude and are separated by $0.2\pi/n$ (one tenth of a Discrete Fourier transform bin) in order to make recovery harder due to high signal coherency. Note that in both cases, the signal satisfies Assumption 3.2.1 with $\varkappa = 0$, and is recoverable with a moderate ϱ, cf. Sections 1.4–1.5 and Chapter 4. Surprisingly, we find *RandomSpikes* to be slightly harder than *CoherentSpikes* for both methods, see Figure 3-1. We find that (**Pen-LS**) outperforms Lasso for all noise levels. In turn, Lasso uniformly outperforms the uniform-fit estimator (**Pen-UF**) in scenario *RandomSpikes*, but is outperformed by the latter in scenario *CoherentSpikes* with high SNR values. The performance gains are the more significant the higher is SNR.

Image denoising. We now consider recovery of an unknown regression function f on the regular grid on $[0, 1]^2$ given the noisy observations[8]:

$$y_\tau = x_\tau + \sigma\zeta_\tau, \quad \tau \in \{0, 1, ..., n_0 - 1\}^2, \tag{3.27}$$

where $x_\tau = f(\tau/n_0)$. We fix $T = 40$, and the grid dimension $n_0 = 40$; the number of samples is $n = n_0^2$. We study three different scenarios for generating the signal in this experiment.

- *RandomSpikes-2D* and *CoherentSpikes-2D* are two-dimensional counterparts of the scenarios studied in the signal denoising experiment: the signal is a harmonic oscillation in $\mathbb{R}^2$ with 4 random frequencies and amplitudes. The minimal frequency separation in the *CoherentSpikes-2D* scenario is $0.2\pi/n_0$ in each dimension of the torus $[0, 2\pi]^2$. The results are shown in Figure 3-1. Same as in one-dimensional recovery, (**Pen-LS**) outperforms Lasso, which outperforms (**Pen-UF**), for all values of SNR, especially for high ones.

- In scenario *DimensionReduction-2D*, we investigate the problem of estimating a function with a hidden low-dimensional structure. We consider the single-index regression model:

$$f(t) = g(\theta^{\mathrm{T}} t), \quad g(\cdot) \in \mathcal{S}^{\mathrm{per}}_{\beta,1}. \tag{3.28}$$

[8] The generalization of our framework to the multi-dimensional setting is quite routine, and we omit it.

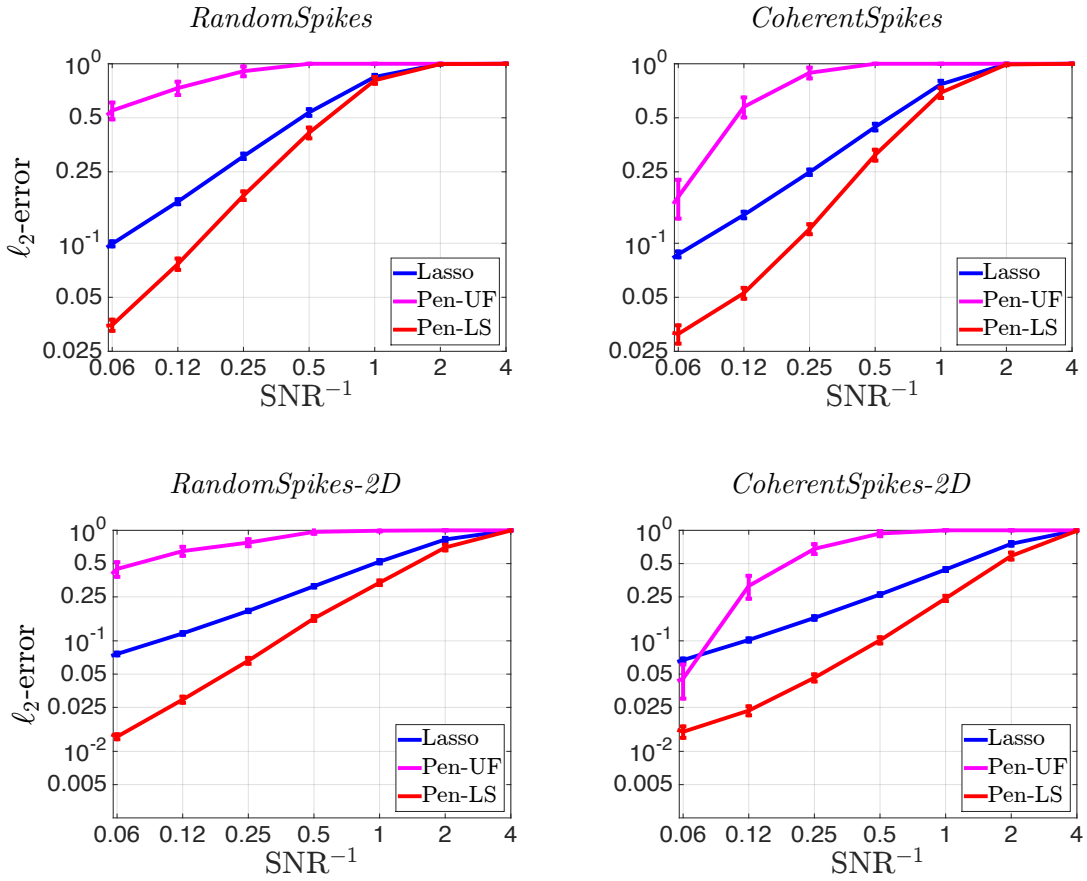

Figure 3-1. Signal and image denoising in different scenarios. Steep segments of the curves for high noise levels correspond to thresholding the observations to zero.

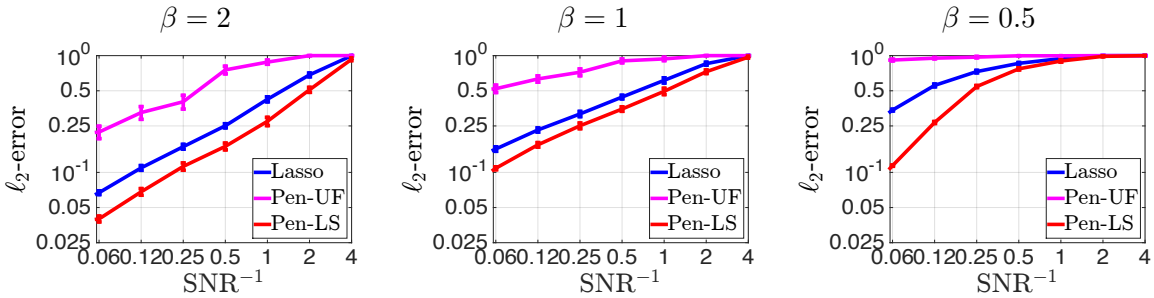

Figure 3-2. Image denoising in scenario *DimensionReduction-2D* with decreasing smoothness.

where $\mathcal{S}^{\text{per}}_{\beta,1}$ is the Sobolev-type smoothness class defined by $\|D^\beta g\|_{L_2} \le 1$, cf. (1.5), with the additional constraint that the weak derivative $D^\beta f$ is periodic on $[0,1]$. The unknown structure is formalized as the direction θ. In our experiments we sample the direction θ uniformly at random and consider different values of the smoothness index β. If it is known a priori that the regression function possesses the structure (3.28), and only the index is unknown, one can use estimators attaining "one-dimensional" rates of recovery; see e.g. [LS14] and references therein. In contrast, our estimators are not aware of the underlying structure but might still adapt to it.

As shown in Figure 3-2, our estimators remain well-behaved in this scenario despite the fact that the available theoretical bounds become pessimistic. For example, the signal (3.28) with a smooth g can be approximated by a small number of harmonic oscillations in $\mathbb{R}^2$. As follows from the proof of [JN09, Proposition 10] combined with Proposition 4.2.2, see Chapter 4, for a harmonic oscillation with s frequencies in $\mathbb{R}^d$ there exists a reproducing linear filter with $\rho(s) = \tilde{O}(s^d)$, *i.e.* the theoretical guarantee is conservative for small β.

Demonstration experiments. The purpose of these experiments is to visually illustrate the application of our approach on specific examples. As before, we compare the least-squares estimators of Section 3.1.1 with Lasso as in [BTR13]. We use the same parameter settings for the all estimators as in the experiments of the previous section.

- **Harmonic oscillations in 2-D.** In this experiment, whose results are shown in Figure 3-3, we estimate a harmonic oscillation in $\mathbb{R}^2$ with 4 random frequencies, observed with $\text{SNR} = 0.5$. The signal is normalized in ℓ_2-norm.
- **Dimension reduction.** In Figure 3-4 we present the results of denoising a single-index signal (3.28), $\text{SNR} = 1$, with the direction θ close to the diagonal $(1,1)$, for two values of the smoothness index $\beta \in \{1,2\}$. One can see that Lasso tends to oversmooth the signal.
- **Denoising textures.** In this experiment (see Figure 3-5), we apply the proposed recovery methods to denoise two images from the Original Brodatz texture database[9] observed according to (3.27). We set $\text{SNR} = 1$. Here, instead of (**Pen-LS**) we use the constrained estimator (**Con-LS**) with $\bar{\varrho} = 4$. As in the above experiments, we use the Lasso [BTR13] with $\lambda = \sigma\sqrt{2\log n}$; as before, $n = n_0^2$ is the number of pixels. In addition, for the constrained estimator we employ pointwise adaptive bandwidth selection:

  1. For every b on the dyadic grid $\{1, 2, ..., \lfloor n_0/2 \rfloor\}$, we divide the (square) image into square blocks of size b.
  2. Inside each block, we construct a pointwise estimate of the signal as discussed above.
  3. Finally, we run a pointwise bandwidth selection algorithm akin to Algorithm 1, cf. Section 2.2.2, for every point of the grid.

  We use the contrained estimator since the knowledge of ϱ is anyway necessary for bandwidth selection as it enters the "admissibility threshold" (2.10). From Figure 3-5 it can be seen that, despite comparable quality in the mean square sense, the two methods significantly differ in their local behavior. In particular, (**Con-LS**) better restores the local signal features while Lasso tends to oversmooth.

[9] http://multibandtexture.recherche.usherbrooke.ca/original_brodatz.html

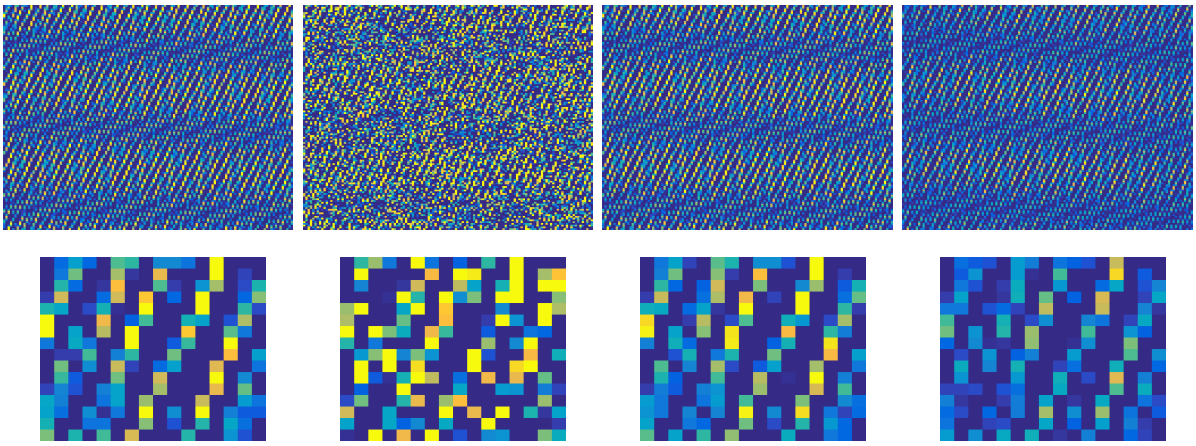

Figure 3-3. Recovery of a harmonic oscillation with 4 frequencies observed with SNR $= 0.5$, i.e. $\|y - x\|_2 \approx 2$. Second row: magnified upper left corner of the image. ℓ_2-error: 0.18 for (**Pen-LS**), 0.45 for Lasso [BTR13].

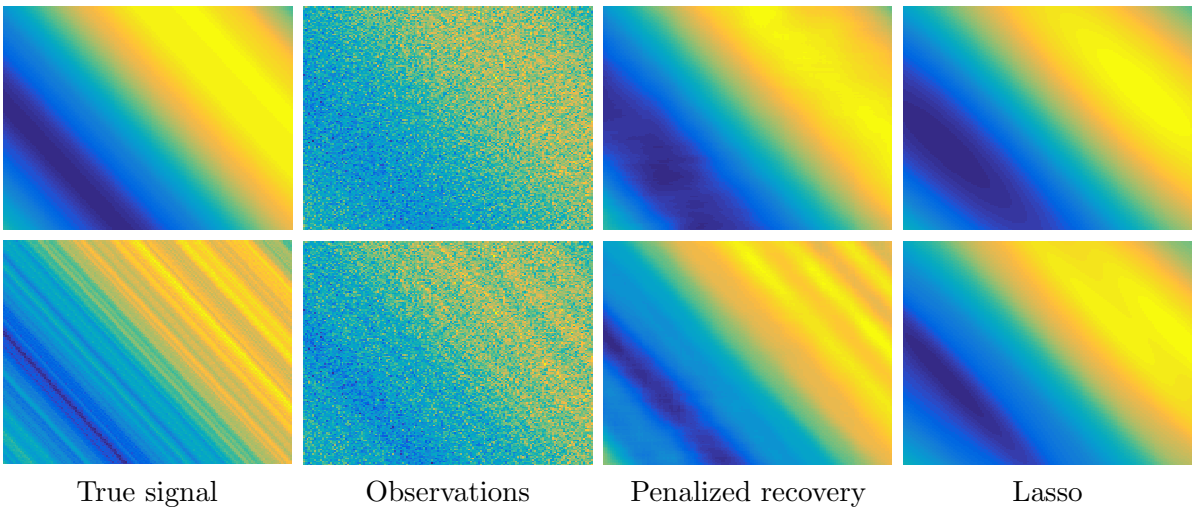

Figure 3-4. Recovery of a single-index signal (3.28) observed with SNR $= 1$, for $\beta = 2$ (1st row) and $\beta = 1$ (2nd row). ℓ_2-error, (**Pen-LS**) vs. Lasso: 0.07 vs. 0.13 for $\beta = 2$; 0.25 vs. 0.31 for $\beta = 1$.

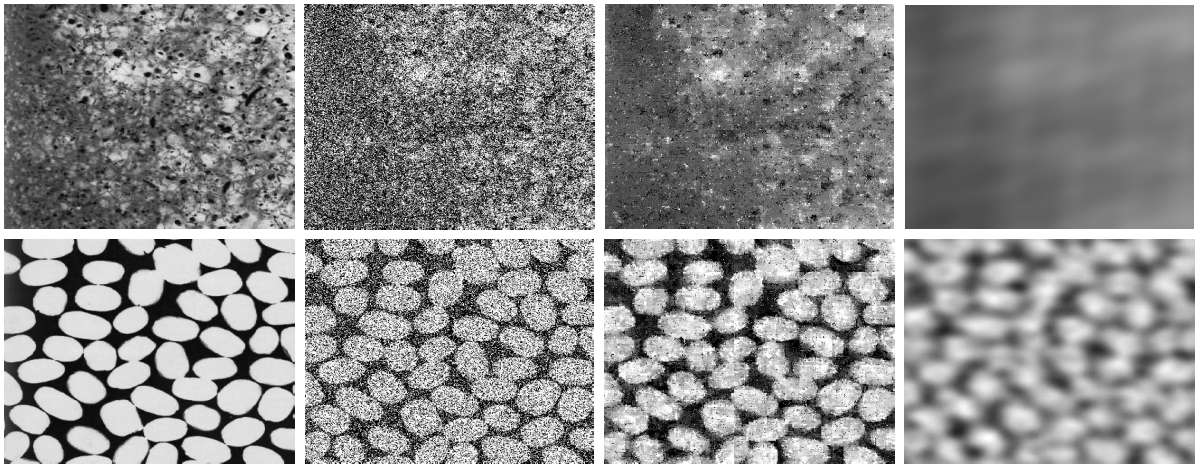

Figure 3-5. Recovery of two instances of the Original Brodatz database, cut (in half) to 320×320 and observed with SNR $= 1$. ℓ_2-error: $1.35e4$ for (**Con-LS**) vs. $1.25e4$ for Lasso in the first row (inst. D73); $1.97e4$ for (**Con-LS**) vs. $2.02e4$ for Lasso in the second row (D75).

3.4 Proofs of oracle inequalities

In this section, we prove our main results – sharp oracle inequalities for regularized least-squares estimators (Theorems 3.2.1–3.2.3). First, let us present some additional notation and technical tools to be used in the proofs.

Additional notation. $\Re(z)$ and $\Im(z)$ denote, correspondingly, the real and imaginary parts of a complex number $z \in \mathbb{C}$, and $\overline{z} = \Re(z) - i\Im(z)$ denotes the complex conjugate of z. We denote A^{T} the transpose of a complex-valued matrix A, and A^{H} its conjugate transpose. We denote $\overline{A}$ the conjugation of A without transposition. We denote A^{-1} the inverse of A whenever it is guaranteed to exist. We denote $\mathrm{Tr}(A)$ the trace of a matrix A, $\det(A)$ its determinant, $\|A\|_{\mathrm{F}}$ the Frobenius norm, and $\|A\|_{\mathrm{op}}$ the operator norm. We denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ the maximal and minimal eigenvalues of a Hermitian matrix A. We denote $\mathrm{Diag}(a)$ the diagonal matrix formed from a vector $a \in \mathbb{C}^n$. We denote I the identity matrix, sometimes with a subscript indicating its size. We denote $\langle \cdot, \cdot \rangle$ the Hermitian scalar product: for two complex vectors a, b of the same dimension, $\langle a, b \rangle = a^{\mathrm{H}} b$. We denote $\langle x, y \rangle_n := \langle [x]_{-n}^n, [y]_{-n}^n \rangle_n$ for $x, y \in \mathbb{C}(\mathbb{Z})$.

In what follows, we associate linear maps $\mathbb{C}_n(\mathbb{Z}) \to \mathbb{C}_{n'}(\mathbb{Z})$ with matrices in $\mathbb{C}^{(2n+1)\times(2n'+1)}$.

3.4.1 Technical tools

Convolution matrices. We use various matrix-vector representations of discrete convolution.

- Given $y \in \mathbb{C}(\mathbb{Z})$, we associate to it an $(2n+1) \times (2m+1)$ matrix

$$T(y) = \begin{bmatrix} y_{-n+m} & \cdots & y_{-n} & \cdots & y_{-n-m} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ y_m & \cdots & y_0 & \cdots & y_{-m} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ y_{n+m} & \cdots & y_n & \cdots & y_{n-m} \end{bmatrix}, \tag{3.29}$$

  such that $[\varphi * y]_{-n}^{n} = T(y)[\varphi]_{-m}^{m}$ for $\varphi \in \mathbb{C}_m(\mathbb{Z})$. Its squared Frobenius norm satisfies

$$\|T(y)\|_{\mathrm{F}}^2 = \sum_{\tau \in \mathrm{D}_m} \|\Delta^\tau y\|_{n,2}^2. \tag{3.30}$$

- Given $\varphi \in \mathbb{C}_m(\mathbb{Z})$, consider an $(2n+1) \times (2m+2n+1)$ matrix

$$M(\varphi) = \begin{bmatrix} \varphi_m & \cdots & \cdots & \varphi_{-m} & 0 & \cdots & \cdots & 0 \\ 0 & \varphi_m & \cdots & \cdots & \varphi_{-m} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \cdots & \cdots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \ddots & \cdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \varphi_m & \cdots & \cdots & \varphi_{-m} \end{bmatrix}, \tag{3.31}$$

  such that for $y \in \mathbb{C}(\mathbb{Z})$ one has $[\varphi * y]_{-n}^{n} = M(\varphi)[y]_{-m-n}^{m+n}$, and

$$\|M(\varphi)\|_{\mathrm{F}}^2 = (2n+1)\|\varphi\|_{m,2}^2. \tag{3.32}$$

- Given $\varphi \in \mathbb{C}_m(\mathbb{Z})$, consider the following circulant matrix of size $2m+2n+1$:

$$C(\varphi) = \begin{bmatrix} \varphi_0 & \cdots & \cdots & \varphi_{-m} & 0 & \cdots & \cdots & \cdots & 0 & \varphi_m & \cdots & \cdots & \varphi_1 \\ \varphi_1 & \varphi_0 & \cdots & \cdots & \varphi_{-m} & 0 & \cdots & \cdots & \cdots & 0 & \varphi_m & \cdots & \varphi_2 \\ \cdots & \cdots & \ddots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \ddots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \ddots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \varphi_m & \cdots & \cdots & \varphi_0 & \cdots & \cdots & \varphi_{-m} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \ddots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \ddots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \ddots & \cdots \\ \varphi_{-1} & \cdots & \cdots & \varphi_{-m} & 0 & \cdots & \cdots & \cdots & 0 & \varphi_m & \cdots & \cdots & \varphi_0 \end{bmatrix}. \tag{3.33}$$

Note that $C(\varphi)[y]_{-m-n}^{m+n}$ is the circular convolution of $[y]_{-m-n}^{m+n}$ and the zero-padded filter

$$\tilde{\varphi} := [\varphi]_{-m-n}^{m+n} = [0; ...; \varphi_{-m}; ...; \varphi_m; 0; ...; 0],$$

that is, convolution of the periodic extensions of $[y]_{-m-n}^{m+n}$ and $\tilde{\varphi}$ evaluated on D_{m+n}. Hence, by the diagonalization property of the DFT operator one has

$$C(\varphi) = \sqrt{2m+2n+1} F_{m+n}^{\mathrm{H}} \mathrm{diag}(F_{m+n}[\tilde{\varphi}]) F_{m+n}. \tag{3.34}$$

Besides, note that

$$\|C(\varphi)\|_{\mathrm{F}}^2 = (2m+2n+1)\|\varphi\|_{m,2}^2.$$

Deviation bounds for quadratic forms. We use simple probabilistic facts listed below.

- Let $\zeta \sim \mathbb{C}\mathcal{N}(0, I_n)$ be a standard complex Gaussian vector, meaning that $\zeta = \xi_1 + i\xi_2$ where ξ_1 and ξ_2 are two independent draws from $\mathcal{N}(0, I_n)$. We use a simple bound

$$\mathbb{P}\left\{\|\zeta\|_\infty \le \sqrt{2\log n + 2u}\right\} \ge 1 - e^{-u} \tag{3.35}$$

  which can be verified directly using that $|\zeta_1|_2^2 \sim \chi_2^2$.

- The following deviation bounds for $\|\zeta\|_2^2 \sim \chi_{2n}^2$ are due to [LM00, Lemma 1]:

$$\begin{aligned} &\mathbb{P}\left\{\frac{\|\zeta\|_2^2}{2} \le n + \sqrt{2nu} + u\right\} \ge 1 - e^{-u}, \\ &\mathbb{P}\left\{\frac{\|\zeta\|_2^2}{2} \ge n - \sqrt{2nu}\right\} \ge 1 - e^{-u}. \end{aligned} \tag{3.36}$$

  By simple algebra we obtain an upper bound for the norm:

$$\mathbb{P}\left\{\|\zeta\|_2 \le \sqrt{2n} + \sqrt{2u}\right\} \ge 1 - e^{-u}. \tag{3.37}$$

- Further, let K be an $n \times n$ Hermitian matrix with the vector of eigenvalues $\lambda = [\lambda_1; ...; \lambda_n]$. Then the real-valued quadratic form $\zeta^{\mathrm{H}} K \zeta$ has the same distribution as $\xi^T B \xi$, where $\xi = [\xi_1; \xi_2] \sim \mathcal{N}(0, I_{2n})$, and B is a real $2n \times 2n$ symmetric matrix with the vector of eigenvalues $[\lambda; \lambda]$. We have $\mathrm{Tr}(B) = 2\mathrm{Tr}(K)$, $\|B\|_{\mathrm{F}}^2 = 2\|K\|_{\mathrm{F}}^2$ and $\|B\| = \|K\| \le \|K\|_{\mathrm{F}}$, where $\|\cdot\|$ and $\|\cdot\|_{\mathrm{F}}$ denote, correspondingly, the spectral and Frobenius norms of a matrix. Invoking again [LM00, Lemma 1] (a close inspection of the proof shows that the assumption of positive semidefiniteness can be relaxed), we have

$$\mathbb{P}\left\{\frac{\zeta^{\mathrm{H}} K \zeta}{2} \le \mathrm{Tr}(K) + (u + \sqrt{2u})\|K\|_{\mathrm{F}}\right\} \ge 1 - e^{-u}. \tag{3.38}$$

  Further, when K is positive semidefinite, we have $\|K\|_{\mathrm{F}} \le \mathrm{Tr}(K)$, whence

$$\mathbb{P}\left\{\frac{\zeta^{\mathrm{H}} K \zeta}{2} \le \mathrm{Tr}(K)(1 + \sqrt{u})^2\right\} \ge 1 - e^{-u}. \tag{3.39}$$

Reformulation of approximate shift-invariance. The following reformulation of Assumption 3.2.1 will be convenient for our purposes.

There exists an s-dimensional vector subspace $\mathcal{S}_n$ of $\mathbb{C}^{2n+1}$ and an idempotent Hermitian $(2n+1)\times(2n+1)$ matrix $\Pi_{\mathcal{S}_n}$ of rank s – projector on $\mathcal{S}_n$ – such that

$$\left\| \left(I_{2n+1} - \Pi_{\mathcal{S}_n}\right) [\Delta^\tau x]_{-n}^n \right\|_2 \left[= \|\Delta^\tau \varepsilon\|_{n,2} \right] \leq \sigma\varkappa, \quad \tau \in \mathrm{D}_m, \tag{3.40}$$

where I_{2n+1} is the identity matrix of size $2n+1$.

3.4.2 Proof of Theorem 3.2.1

Step 1^o. Let $\varphi^o \in \mathbb{C}_m(\mathbb{Z})$ be any filter satisfying the constraint in (**Con-LS**). Then,

$$\begin{aligned} \|x - \widehat{\varphi} * y\|_{n,2}^2 &\leq \|(1-\varphi^o) * y\|_{n,2}^2 - \sigma^2\|\zeta\|_{n,2}^2 - 2\sigma\Re\langle\zeta, x - \widehat{\varphi} * y\rangle_n \\ &= \|x - \varphi^o * y\|_{n,2}^2 - 2\underbrace{\sigma\Re\langle\zeta, x - \widehat{\varphi} * y\rangle_n}_{\delta^{(1)}} + 2\underbrace{\sigma\Re\langle\zeta, x - \varphi^o * y\rangle_n}_{\delta^{(2)}}. \end{aligned} \tag{3.41}$$

Let us bound $\delta^{(1)}$. Denote for brevity $I := I_{2n+1}$, and recall that $\Pi_{\mathcal{S}_n}$ is the projector on $\mathcal{S}_n$ from (3.40). We have the following decomposition:

$$\begin{aligned} \delta^{(1)} &= \underbrace{\sigma\Re\langle[\zeta]_{-n}^n, \Pi_{\mathcal{S}_n}[x - \widehat{\varphi} * y]_{-n}^n\rangle}_{\delta_1^{(1)}} + \underbrace{\sigma\Re\langle[\zeta]_{-n}^n, (I - \Pi_{\mathcal{S}_n})[x - \widehat{\varphi} * x]_{-n}^n\rangle}_{\delta_2^{(1)}} \\ &\quad - \underbrace{\sigma^2\Re\langle[\zeta]_{-n}^n, (I - \Pi_{\mathcal{S}_n})[\widehat{\varphi} * \zeta]_{-n}^n\rangle}_{\delta_3^{(1)}} \end{aligned} \tag{3.42}$$

One can easily bound $\delta_1^{(1)}$ under the premise of the theorem:

$$\begin{aligned} \left|\delta_1^{(1)}\right| &\leq \sigma\left\|\Pi_{\mathcal{S}_n}[\zeta]_{-n}^n\right\|_2 \left\|\Pi_{\mathcal{S}_n}[x - \widehat{\varphi} * y]_{-n}^n\right\|_2 \\ &\leq \sigma\left\|\Pi_{\mathcal{S}_n}[\zeta]_{-n}^n\right\|_2 \|x - \widehat{\varphi} * y\|_{n,2}. \end{aligned}$$

Note that $\Pi_{\mathcal{S}_n}[\zeta]_{-n}^n \sim \mathbb{C}\mathcal{N}(0, I_s)$, and by (3.37) we have

$$\mathbb{P}\left\{\left\|\Pi_{\mathcal{S}_n}[\zeta]_{-n}^n\right\|_2 \geq \sqrt{2s} + \sqrt{2u}\right\} \leq e^{-u},$$

which gives the bound

$$\mathbb{P}\left\{\left|\delta_1^{(1)}\right| \leq \sigma\|x - \widehat{\varphi} * y\|_{n,2}\left(\sqrt{2s} + \sqrt{2\log\left[1/\alpha_1\right]}\right)\right\} \geq 1 - \alpha_1. \tag{3.43}$$

Step 2^o. We are to bound the second term of (3.42). To this end, note first that

$$\delta_2^{(1)} = \sigma\Re\langle[\zeta]_{-n}^n, (I - \Pi_{\mathcal{S}_n})[x]_{-n}^n\rangle - \sigma\Re\langle[\zeta]_{-n}^n, (I - \Pi_{\mathcal{S}_n})[\widehat{\varphi} * x]_{-n}^n\rangle.$$

By (3.40), $\big\|(I-\Pi_{\mathcal{S}_n})[x]_{-n}^{n}\big\|_2 \le \sigma\varkappa$, thus with probability $1-\alpha$,

$$\big|\langle[\zeta]_{-n}^{n},(I-\Pi_{\mathcal{S}_n})[x]_{-n}^{n}\rangle\big| \le \sigma\varkappa\sqrt{2\log[1/\alpha]}. \tag{3.44}$$

On the other hand, using the notation defined in (3.29), we have $[\widehat{\varphi}*x]_{-n}^{n} = T(x)[\widehat{\varphi}]_{-m}^{m}$, so that

$$\langle[\zeta]_{-n}^{n},(I-\Pi_{\mathcal{S}_n})[\widehat{\varphi}*x]_{-n}^{n}\rangle = \langle[\zeta]_{-n}^{n},(I-\Pi_{\mathcal{S}_n})T(x)[\widehat{\varphi}]_{-m}^{m}\rangle.$$

Note that $[T(x)]_\tau = [\Delta^\tau x]_{-n}^{n}$ for the columns of $T(x)$, $\tau\in\mathrm{D}_m$. By (3.40), we have

$$(I-\Pi_{\mathcal{S}_n})T(x) = T(\varepsilon),$$

and by (3.30),

$$\begin{aligned}\|(I-\Pi_{\mathcal{S}_n})T(x)\|_{\mathrm{F}}^2 = \|T(\varepsilon)\|_{\mathrm{F}}^2 &= \sum_{\tau\in\mathrm{D}_m}\|\Delta^\tau\varepsilon\|_{n,2}^2\\ &\le (2m+1)\sigma^2\varkappa^2.\end{aligned}$$

Due to (3.39) we conclude that

$$\big\|T(x)^{\mathrm{H}}(I-\Pi_{\mathcal{S}_n})[\zeta]_{-n}^{n}\big\|_2^2 \le 2(2m+1)\sigma^2\varkappa^2\big(1+\sqrt{\log[1/\alpha]}\big)^2$$

with probability at least $1-\alpha$. Since

$$\big|\big\langle[\zeta]_{-n}^{n},(I-\Pi_{\mathcal{S}_n})T(x)[\widehat{\varphi}]_{-m}^{m}\big\rangle\big| \le \frac{\bar{\varrho}}{\sqrt{2m+1}}\big\|T(x)^{\mathrm{H}}(I-\Pi_{\mathcal{S}_n})[\zeta]_{-n}^{n}\big\|_2,$$

we arrive at the bound with probability $1-\alpha$:

$$\big|\big\langle[\zeta]_{-n}^{n},(I-\Pi_{\mathcal{S}_n})T(x)[\widehat{\varphi}]_{-m}^{m}\big\rangle\big| \le \sqrt{2}\sigma\varkappa\bar{\varrho}\big(1+\sqrt{\log[1/\alpha]}\big).$$

Along with (3.44) this results in the bound

$$\mathbb{P}\left\{\big|\delta_2^{(1)}\big| \le \sqrt{2}\sigma^2\varkappa(\bar{\varrho}+1)\big(1+\sqrt{\log\left[1/\min(\alpha_2,\alpha_3)\right]}\big)\right\} \ge 1-\alpha_2-\alpha_3. \tag{3.45}$$

Step 3°. Let us rewrite $\delta_3^{(1)}$ as follows:

$$\delta_3^{(1)} \;=\; \sigma^2\Re\langle[\zeta]_{-n}^{n},(I-\Pi_{\mathcal{S}_n})M(\widehat{\varphi})[\zeta]_{-m-n}^{m+n}\rangle = \sigma^2\Re\sigma^2\langle[\zeta]_{-m-n}^{m+n},QM(\widehat{\varphi})[\zeta]_{-m-n}^{m+n}\rangle,$$

where $M(\widehat{\varphi})\in\mathbb{C}^{(2n+1)\times(2m+2n+1)}$ is defined by (3.31), and $Q\in\mathbb{C}^{(2m+2n+1)\times(2n+1)}$ is given by

$$Q = \begin{bmatrix} O_{m,2n+1}\\ I-\Pi_{\mathcal{S}_n}\\ O_{m,2n+1}\end{bmatrix}.$$

(Hereafter we denote $O_{m,n}$ the $m \times n$ zero matrix.) Now, by the definition of $\widehat{\varphi}$ and since the mapping $\varphi \mapsto M(\varphi)$ is linear,

$$\begin{aligned}\delta_3^{(1)} &= \frac{\sigma^2}{2}([\zeta]_{-m-n}^{m+n})^{\mathrm{H}}(\underbrace{QM(\widehat{\varphi}) + M(\widehat{\varphi})^{\mathrm{H}}Q^{\mathrm{H}}}_{K_1(\widehat{\varphi})})[\zeta]_{-m-n}^{m+n} \\ &\leq \frac{\sigma^2\bar{\varrho}}{2\sqrt{2m+1}} \max_{\substack{u \in \mathbb{C}_m(\mathbb{Z}),\\ \|u\|_{m,1}^{\mathrm{F}} \leq 1}} (\zeta_{-m}^{n})^{\mathrm{H}} K_1(u)[\zeta]_{-m-n}^{m+n} \\ &= \frac{\sigma^2\bar{\varrho}}{\sqrt{2m+1}} \max_{j \in \mathrm{D}_m} \max_{\theta \in [0,2\pi]} \tfrac{1}{2}([\zeta]_{-m-n}^{m+n})^{\mathrm{H}} K_1(e^{i\theta}u^j)[\zeta]_{-m-n}^{m+n},\end{aligned}$$

where $u^j \in \mathbb{C}_m(\mathbb{Z})$, and $[u^j]_{-m}^{m} = F_m^{\mathrm{H}}\mathrm{e}^j$, e^j being the discrete Dirac pulse centered at $j \in \mathbb{Z}$. Indeed, $([\zeta]_{-m-n}^{m+n})^{\mathrm{H}} K_1(u)[\zeta]_{-m-n}^{m+n}$ is clearly a convex function of the argument u as a linear function of $[\Re(u); \Im(u)]$; as such, it attains its maximum over the set

$$\mathcal{B}_{m,1} = \{u \in \mathbb{C}_m(\mathbb{Z}) : \|u\|_{m,1}^{\mathrm{F}} \leq 1\} \tag{3.46}$$

at one of the extremal points $e^{i\theta}u^j$, $\theta \in [0, 2\pi]$, of this set. It can be directly verified that

$$K_1(e^{\imath\theta}u) = K_1(u)\cos\theta + K_2(u)\sin\theta,$$

where the Hermitian matrix $K_2(u)$ is given by

$$K_2(u) = i\left(QM(u) - M(u)^{\mathrm{H}}Q^{\mathrm{H}}\right).$$

Denoting $q_l^j(\zeta) = \frac{1}{2}([\zeta]_{-m-n}^{m+n})^{\mathrm{H}} K_l(u^j)[\zeta]_{-m-n}^{m+n}$ for $l = 1, 2$, we have

$$\begin{aligned}\max_{\theta \in [0,2\pi]} \tfrac{1}{2}([\zeta]_{-m-n}^{m+n})^{\mathrm{H}} K_1(e^{\imath\theta}u^j)[\zeta]_{-m-n}^{m+n} &= \max_{\theta \in [0,2\pi]} q_1^j(\zeta)\cos\theta + q_2^j(\zeta)\sin\theta \\ &= \sqrt{|q_1^j(\zeta)|^2 + |q_2^j(\zeta)|^2} \\ &\leq \sqrt{2}\max(|q_1^j(\zeta)|, |q_2^j(\zeta)|).\end{aligned} \tag{3.47}$$

By simple algebra and using (3.32), we get for $l = 1, 2$:

$$\begin{aligned}\mathrm{Tr}[K_l(u^j)^2] &\leq 4\,\mathrm{Tr}[M(u^j)M(u^j)^{\mathrm{H}}] \\ &= 4(2n+1)\|u^j\|_{m,2}^2 \\ &\leq 4(2n+1).\end{aligned}$$

Now let us bound $\mathrm{Tr}[K_l(u)]$, $l = 1, 2$, on the set , $\mathcal{B}_{m,1}$ cf. (3.46). One can verify that for the circulant matrix $C(u)$, cf. (3.33), it holds:

$$QM(u) = RC(u),$$

where $R = QQ^{\mathrm{H}}$ is an $(2m+2n+1)\times(2m+2n+1)$ projection matrix of rank s defined by

$$R = \begin{bmatrix} O_{m,m} & O_{m,n+1} & O_{m,m} \\ O_{n+1,m} & I - \Pi_{\mathcal{S}_n} & O_{n+1,m} \\ O_{m,m} & O_{m,n+1} & O_{m,m}. \end{bmatrix}$$

Hence, denoting $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{nuc}}$ the operator and nuclear matrix norms, we can bound $\mathrm{Tr}[K_l(u)]$, $l = 1, 2$, as follows:

$$\begin{aligned} |\operatorname{Tr}[K_l(u)]| &\le 2\big|\operatorname{Tr}[RC(u)]\big| \\ &\le 2\|R\|_{\mathrm{op}}\,\|C(u)\|_{\mathrm{nuc}} \\ &\le 2\|C(u)\|_{\mathrm{nuc}} \\ &= 2\sqrt{2m+2n+1}\|\tilde{u}\|^{\mathrm{F}}_{m+n,1}, \end{aligned} \tag{3.48}$$

where in the last transition we used the Fourier diagonalization property (3.34). Recall that $u \in \mathbb{C}_m(\mathbb{Z})$, hence $F_{m+n}[u]$ is the Discrete Fourier transform of the *zero-padded filter*

$$\tilde{u} = [0; ...; 0; [u]_{-m}^{m}; 0; ...; 0] \in \mathbb{C}^{2m+2n+1}.$$

The following lemma, interesting in its own right, controls the inflation of the ℓ_1-norm of the DFT of a filter after zero padding. The proof, presented later on, relies to the fact that the normalized ℓ_1-norm of the Dirichlet kernel of order N grows not faster than $\log N$.

Lemma 3.4.1 (ℓ_1**-norm of the DFT after zero-padding).** *For any $u \in \mathbb{C}_m(\mathbb{Z})$, one has*

$$\|u\|^{\mathrm{F}}_{m+n,1} \le \|u\|^{\mathrm{F}}_{m,1}\sqrt{1+\kappa^2_{m,n}}[\log(m+n+1)+3].$$

Combining this lemma with (3.48) we arrive at

$$\big|\operatorname{Tr}[K_l(u^j)]\big| \le 2\sqrt{2m+1}(\kappa^2_{m,n}+1)(\log[2m+2n+1]+3), \quad l = 1, 2.$$

By (3.38) we conclude that for any fixed pair $(l, j) \in \{1, 2\} \times \mathrm{D}_m$, with probability $\ge 1-\alpha$,

$$|q^j_l(\zeta)| \le \big|\operatorname{Tr}[K_l(u^j)]\big| + \big\|K_l(u^j)\big\|_{\mathrm{F}}\left(1+\sqrt{\log[2/\alpha]}\right)^2.$$

With $\alpha_0 = 2(2m+1)\alpha$, by the union bound together with (3.46) and (3.47) we get

$$\begin{aligned} &\mathbb{P}\left\{\delta^{(1)}_3 \le 2\sqrt{2}\sigma^2\bar{\varrho}\left[(\kappa^2_{m,n}+1)(\log[2m+2n+1]+3) + \kappa_{m,n}\big(1+\sqrt{\log\left[4(2m+1)/\alpha_0\right]}\big)^2\right]\right\} \\ &\ge 1-\alpha_0. \end{aligned} \tag{3.49}$$

Step 4^o. Bounding $\delta^{(2)}$ is relatively easy since φ^o does not depend on the noise. We decompose

$$\delta^{(2)} = \sigma\Re\langle\zeta, x - \varphi^o * x\rangle_n - \sigma^2\Re\langle\zeta, \varphi^o * \zeta\rangle_n.$$

Note that $\Re\langle\zeta, x-\varphi^o*x\rangle_n \sim \mathcal{N}(0, \|x-\varphi^o*x\|_{n,2}^2)$, therefore, with probability $\geq 1-\alpha$,

$$\Re\langle\zeta, x-\varphi^o*x\rangle_n \leq \sqrt{2\log[1/\alpha]}\|x-\varphi^o*x\|_{n,2}. \tag{3.50}$$

On the other hand, defining

$$\varrho = \sqrt{2m+1}\|\varphi^o\|_{m,1}^{\mathrm{F}},$$

we have

$$\begin{aligned}\|x-\varphi^o*x\|_{n,2} &\leq \|x-\varphi^o*y\|_{n,2} + \sigma\|\varphi^o*\zeta\|_{n,2}\\ &\leq \|x-\varphi^o*y\|_{n,2} + \sqrt{2}\sigma\varrho\kappa_{m,n}\big(1+\sqrt{\log[1/\alpha]}\big)\end{aligned} \tag{3.51}$$

with probability $1-\alpha$. Indeed, one has

$$\|\varphi^o*\zeta\|_{n,2}^2 = \big\|M(\varphi^o)[\zeta]_{-m-n}^{m+n}\big\|_2^2,$$

where for $M(\varphi^o)$ by (3.32) we have

$$\|M(\varphi^o)\|_{\mathrm{F}}^2 = (2n+1)\|\varphi^o\|_{m,2}^2 \leq \kappa_{m,n}^2\varrho^2. \tag{3.52}$$

Using (3.39) we conclude that, with probability at least $1-\alpha$,

$$\|\varphi^o*\zeta\|_{n,2}^2 \leq 2\kappa_{m,n}^2\varrho^2\big(1+\sqrt{\log[1/\alpha]}\big)^2, \tag{3.53}$$

which implies (3.51). Using (3.50) and (3.51), we get that with probability at least $1-\alpha_4-\alpha_5$,

$$\begin{aligned}&\Re\langle\zeta, x-\varphi^o*x\rangle_n\\ &\leq \sqrt{2\log\left[1/\min(\alpha_4,\alpha_5)\right]}\left[\|x-\varphi^o*y\|_{n,2} + \sqrt{2}\sigma\varrho\kappa_{m,n}\big(1+\sqrt{\log[1/\min(\alpha_4,\alpha_5)]}\big)\right]\\ &\leq \|x-\varphi^o*y\|_{n,2}\sqrt{2\log\left[1/\min(\alpha_4,\alpha_5)\right]} + 2\sigma\varrho\kappa_{m,n}\big(1+\sqrt{\log\left[1/\min(\alpha_4,\alpha_5)\right]}\big)^2.\end{aligned} \tag{3.54}$$

Now, the (indefinite) quadratic form

$$\Re\langle\zeta, \varphi^o*\zeta\rangle_n = \tfrac{1}{2}([\zeta]_{-m-n}^{m+n})^{\mathrm{H}}K_0(\varphi^o)[\zeta]_{-m-n}^{m+n},$$

where

$$K_0(\varphi^o) = \begin{bmatrix}O_{m,2m+2n+1}\\ M(\varphi^o)\\ O_{m,2m+2n+1}\end{bmatrix} + \begin{bmatrix}O_{m,2m+2n+1}\\ M(\varphi^o)\\ O_{m,2m+2n+1}\end{bmatrix}^{\mathrm{H}},$$

whence (cf. **Step 3**^{o})

$$|\operatorname{Tr}[K_0(\varphi^o)]| \leq 2(2n+1)\,|\varphi_0^o|$$

Let us bound $|\varphi_0^o|$. Let e^0 be the discrete centered Dirac vector in $\mathbb{R}^{2m+1}$, and note that $\|F_m[e^0]\|_\infty = 1/\sqrt{2m+1}$. Then,

$$|\varphi_m^o| = |\langle[\varphi^o]_{-m}^m, e^0\rangle| \leq \|\varphi^o\|_{m,1}^{\mathrm{F}}\|F_m[e^0]\|_\infty \leq \frac{\varrho}{2m+1},$$

whence $|\operatorname{Tr}[K_0(\varphi^o)]| \le 2\kappa_{m,n}^2\varrho$. On the other hand, by (3.52),

$$\|K_0(\varphi^o)\|_{\mathrm{F}}^2 \le 4\,\|M(\varphi^o)\|_{\mathrm{F}}^2 \le 4\kappa_{m,n}^2\varrho^2.$$

Hence by (3.38),

$$\mathbb{P}\left\{-\Re\langle\zeta, \varphi^o * \zeta\rangle_n \le 2\kappa_{m,n}^2\varrho + 2\kappa_{m,n}\varrho\big(1+\sqrt{2\log[1/\alpha_6]}\big)^2\right\} \ge 1-\alpha_6. \tag{3.55}$$

Step 5^o. Let us combine the bounds obtained in the previous steps with initial bound (3.41). For any $\alpha \in (0,1]$, putting $\alpha_i = \alpha/4$ for $i = 0, 1, 6$, and $\alpha_j = \alpha/16$, $2 \le j \le 5$, by the union bound we get that with probability $\ge 1-\alpha$,

$$\begin{aligned}
\|x - \widehat{\varphi} * y\|_{n,2}^2 &\le \|x - \varphi^o * y\|_{n,2}^2 + 2\delta^{(2)} - 2\delta^{(1)}\\
\text{[by (3.54)]} &\le \|x - \varphi^o * y\|_{n,2}^2 + 2\sigma\|x - \varphi^o * y\|_{n,2}\sqrt{2\log[16/\alpha]}\\
\text{[by (3.54)–(3.55)]} &\quad + 4\sigma^2\varrho\Big[\kappa_{m,n}^2 + 2\kappa_{m,n}\big(1+\sqrt{2\log[16/\alpha]}\big)^2\Big]\\
\text{[by (3.43)]} &\quad + 2\sigma\|x - \widehat{\varphi} * y\|_{n,2}\big(\sqrt{2s} + \sqrt{2\log[16/\alpha]}\big)\\
\text{[by (3.45)]} &\quad + 2\sqrt{2}\sigma^2(\bar{\varrho}+1)\big(1+\sqrt{\log[16/\alpha]}\big)\varkappa\\
\text{[by (3.49)]} &\quad + 4\sqrt{2}\sigma^2\bar{\varrho}\Big[(\kappa_{m,n}^2+1)(\log[2m+2n+1]+3)\\
&\qquad + \kappa_{m,n}\left(1+\sqrt{\log[16(m+1)/\alpha]}\right)^2\Big]
\end{aligned} \tag{3.56}$$

Now, denote $c_\alpha := \sqrt{2\log[16/\alpha]}$ and let

$$u(\alpha) = 2\big(\sqrt{2} + c_\alpha\big), \tag{3.57}$$

$$v_1(\alpha) = 4\left[\kappa_{m,n}^2 + 2\kappa_{m,n}\left(1+c_\alpha\right)^2\right], \tag{3.58}$$

$$v_2(\alpha) = 4\sqrt{2}\Big[(\kappa_{m,n}^2+1)(\log[2m+2n+1]+3) + \kappa_{m,n}\left(1+\sqrt{\log\left[16(2m+1)/\alpha\right]}\right)^2\Big]. \tag{3.59}$$

In this notation, (3.56) becomes

$$\begin{aligned}
\|x - \widehat{\varphi} * y\|_{n,2}^2 \le \|x - \varphi^o * y\|_{n,2}^2 &+ 2\sigma(\sqrt{2s} + c_\alpha)\left(\|x - \widehat{\varphi} * y\|_{n,2} + \|x - \varphi^o * y\|_{n,2}\right)\\
&+ u(\alpha)\sigma^2(\bar{\varrho}+1)\varkappa + (v_1(\alpha)+v_2(\alpha))\sigma^2\bar{\varrho},
\end{aligned} \tag{3.60}$$

which implies, by completing the squares, that

$$\|x - \widehat{\varphi} * y\|_{n,2} \le \|x - \varphi^o * y\|_{n,2} + 2\sigma(\sqrt{2s} + c_\alpha) + \sigma\sqrt{u(\alpha)(\bar{\varrho}+1)\varkappa + (v_1(\alpha)+v_2(\alpha))\bar{\varrho}}.$$

Finally, let us simplify this bound. Note that

$$u(\alpha) \le 4c_\alpha, \tag{3.61}$$

while on the other hand,

$$\begin{aligned} v_1(\alpha) + v_2(\alpha) &\le 4\sqrt{2}(\kappa_{m,n}^2 + 1)(\log[2m + 2n + 1] + 4) + 4.5(4\sqrt{2} + 8)\kappa_{m,n} \log\left[16(2m + 1)/\alpha\right] \\ &\le 8\left(1 + 4\kappa_{m,n}\right)^2 \log\left[110(m + n + 1)/\alpha\right]. \end{aligned} \tag{3.62}$$

Hence we arrive at

$$\|x - \widehat{\varphi} * y\|_{n,2} \le \|x - \varphi^o * y\|_{n,2} + 2\sigma\left(\sqrt{\bar{\varrho} V_\alpha} + \sqrt{(\bar{\varrho} + 1)c_\alpha \varkappa} + \sqrt{2s} + c_\alpha\right), \tag{3.63}$$

where we introduced

$$V_\alpha := 2\left(1 + 4\kappa_{m,n}\right)^2 \log\left[110(m + n + 1)/\alpha\right]. \tag{3.64}$$

The bound (3.12) of the theorem follows from (3.63) after straightforward simplifications. □

3.4.3 Proof of Theorem 3.2.2

Denote $\widehat{\varrho} = \sqrt{2m+1}\|\widehat{\varphi}\|_{m,1}^{\mathrm{F}}$ and $\varrho = \sqrt{2m+1}\|\varphi^o\|_{m,1}^{\mathrm{F}}$, where $\widehat{\varphi}$ is an optimal solution to (**Pen-LS**), and φ^o is any feasible solution φ^o; otherwise, we will use the same notation as previously. Similarly to (3.41), we get

$$\|x - \widehat{\varphi} * y\|_{n,2}^2 + \lambda\sigma^2\widehat{\varrho} \le \|x - \varphi^o * y\|_{n,2}^2 - 2\delta^{(1)} + 2\delta^{(2)} + \lambda\sigma^2\varrho.$$

Thus, repeating the first four steps of the proof of Theorem 3.2.1 we obtain, cf. (3.56)–(3.59),

$$\begin{aligned} \|x - \widehat{\varphi} * y\|_{n,2}^2 + \lambda\sigma^2\widehat{\varrho} \le \|x - \varphi^o * y\|_{n,2}^2 &+ 2\sigma(\|x - \varphi^o * y\|_{n,2} + \|x - \widehat{\varphi} * y\|_{n,2})(\sqrt{2s} + c_\alpha) \\ &+ \lambda\sigma^2\varrho + v_1(\alpha)\sigma^2\varrho + u(\alpha)\sigma^2\varkappa(\widehat{\varrho} + 1) + v_2(\alpha)\sigma^2\widehat{\varrho}, \end{aligned} \tag{3.65}$$

where $u(\alpha)$, $v_1(\alpha)$, and $v_2(\alpha)$ are given by (3.57)–(3.59), and $c_\alpha := \sqrt{2\log[16/\alpha]}$. Using that $(1 + \sqrt{2x})^2 \le 4.5x$ when $x \ge 2$, one may check that as long as $m, n \ge 1$,

$$v_2(\alpha) \le 4\sqrt{2}\left(\kappa_{m,n}^2 + 4.5\kappa_{m,n} + 1\right)\log\left[42(m + n + 1)/\alpha\right] = \underline{\lambda}. \tag{3.66}$$

Hence, choosing $\lambda \ge \underline{\lambda}$, we guarantee that $v_2(\alpha)\sigma^2\widehat{\varrho} - \lambda\sigma^2\widehat{\varrho} \le 0$. On the other hand, one has

$$v_1(\alpha) \le 4(1 + \kappa_{m,n})^2(1 + c_\alpha)^2 \le 3\underline{\lambda}, \tag{3.67}$$

and one arrives at

$$\begin{aligned} \|x - \widehat{\varphi} * y\|_{n,2}^2 \le \|x - \varphi^o * y\|_{n,2}^2 &+ 2\sigma\left(\|x - \widehat{\varphi} * y\|_{n,2} + \|x - \varphi^o * y\|_{n,2}\right)\left(\sqrt{2s} + c_\alpha\right) \\ &+ u(\alpha)\sigma^2\varkappa(\widehat{\varrho} + 1) + (\lambda + 3\underline{\lambda})\sigma^2\varrho. \end{aligned}$$

Now, (3.15) follows by (3.61) and using that $\lambda \ge \underline{\lambda}$. To prove the second statement of the theorem, note that under condition (3.17), one has, by (3.61),

$$u(\alpha)\sigma^2\varkappa(\widehat{\varrho} + 1) \le 4c_\alpha\sigma^2\varkappa(\widehat{\varrho} + 1) \le \sigma^2\underline{\lambda}(\widehat{\varrho} + 1) \le \sigma^2\underline{\lambda}(\widehat{\varrho} + \varrho), \tag{3.68}$$

thus arriving at

$$\|x-\widehat{\varphi}*y\|_{n,2}^2 \le \|x-\varphi^o*y\|_{n,2}^2 + 2\sigma\left(\|x-\widehat{\varphi}*y\|_{n,2}+\|x-\varphi^o*y\|_{n,2}\right)\left(\sqrt{2s}+c_\alpha\right) + (\lambda+4\underline{\lambda})\sigma^2\varrho + (2\underline{\lambda}-\lambda)\sigma^2\widehat{\varrho}.$$

We conclude by noting that the last term in the right-hand side vanishes whenever $\lambda \ge 2\underline{\lambda}$. □

3.4.4 Proof of Theorem 3.2.3

Due to feasibility of φ^o, we have the following counterpart of (3.41):

$$\|x-\widehat{\varphi}*y\|_{n,2}^2 + \lambda^2\sigma^2\widehat{\varrho}^2 \le \|x-\varphi^o*y\|_{n,2}^2 - 2\delta^{(1)} + 2\delta^{(2)} + \lambda^2\sigma^2\varrho^2.$$

Thus, repeating **Steps 1^o–4^o** of the previous proof, we obtain a counterpart of (3.60):

$$\begin{aligned}\|x-\widehat{\varphi}*y\|_{n,2}^2 + \lambda^2\sigma^2\widehat{\varrho}^2 \le\ & \|x-\varphi^o*y\|_{n,2}^2 + 2\sigma(\|x-\varphi^o*y\|_{n,2}+\|x-\widehat{\varphi}*y\|_{n,2})(\sqrt{2s}+c_\alpha)\\ & +u(\alpha)\sigma^2\varkappa + v_1(\alpha)\sigma^2\varrho + \lambda^2\sigma^2\varrho^2 + [u(\alpha)\varkappa + v_2(\alpha)]\,\sigma^2\widehat{\varrho},\end{aligned} \tag{3.69}$$

with $u(\alpha)$, $v_1(\alpha)$, and $v_2(\alpha)$ given by (3.57)–(3.59). We now consider two cases as follows.

Case (a). First, assume that

$$\begin{aligned}\|x-\widehat{\varphi}*y\|_{n,2}^2 \ \le\ & \|x-\varphi^o*y\|_{n,2}^2 \ + \ 2\sigma(\|x-\varphi^o*y\|_{n,2}+\|x-\widehat{\varphi}*y\|_{n,2})(\sqrt{2s}+c_\alpha)\\ & + \ u(\alpha)\sigma^2\varkappa + v_1(\alpha)\sigma^2\varrho + \lambda^2\sigma^2\varrho^2.\end{aligned} \tag{3.70}$$

In this case, clearly,

$$\begin{aligned}\|x-\widehat{\varphi}*y\|_{n,2} &\le \|x-\varphi^o*y\|_{n,2} + 2\sigma\left(\sqrt{2s}+c_\alpha\right) + \sqrt{u(\alpha)\sigma^2\varkappa + v_1(\alpha)\sigma^2\varrho + \lambda^2\sigma^2\varrho^2}\\ &\le \|x-\varphi^o*y\|_{n,2} + 2\sigma(\sqrt{2s}+c_\alpha) + \sigma(\sqrt{u(\alpha)\varkappa + v_1(\alpha)\varrho} + \lambda\varrho)\end{aligned} \tag{3.71}$$

Case (b). Suppose, on the contrary, that (3.70) does not hold, we then conclude from (3.69) that

$$\widehat{\varrho} \le \lambda^{-2}(u(\alpha)\varkappa + v_2(\alpha)),$$

and

$$u(\alpha)\widehat{\varrho}\varkappa + v_2(\alpha)\widehat{\varrho} \le \lambda^{-2}(u(\alpha)\varkappa + v_2(\alpha))^2.$$

When substituting the latter bound into (3.69), we obtain

$$\|x-\widehat{\varphi}*y\|_{n,2} \le \|x-\varphi^o*y\|_{n,2} + 2\sigma(\sqrt{2s}+c_\alpha) + \sigma(\sqrt{u(\alpha)\varkappa + v_1(\alpha)\varrho} + \lambda^{-1}(u(\alpha)\varkappa + v_2(\alpha)) + \lambda\varrho),$$

which is also satisfied in **Case (a)** due to (3.71).

Finally, using (3.61), (3.62), and the bound

$$v_1(\alpha) \le 4(1+\kappa_{m,n})^2(1+c_\alpha)^2$$

which directly follows from (3.58), we get that

$$\|x - \widehat{\varphi} * y\|_{n,2} \le \|x - \varphi^o * y\|_{n,2} + \sigma(\lambda\varrho + 4\lambda^{-1}(c_\alpha\varkappa + V_\alpha)) + 2\sigma\left(\sqrt{\varrho W_\alpha} + \sqrt{c_\alpha\varkappa} + \sqrt{2s} + c_\alpha\right),$$

with V_α is given by (3.64), and $W_\alpha = (1+\kappa_{m,n})^2(1+c_\alpha)^2$. The bound (3.18) of the theorem follows by simplifying the above bound in a straightforward manner. □

Proof of Lemma 3.4.1

Let us prove that the bound of the theorem,

$$\|u\|^{\mathrm{F}}_{m+n,1} \le \sqrt{1+\kappa^2_{m,n}}[\log(m+n+1)+3],$$

holds on the unit ball

$$\mathcal{B}_{m,1} = \{u \in \mathbb{C}_m(\mathbb{Z}) : \|u\|^{\mathrm{F}}_{m,1} \le 1\};$$

then, the statement will follow by the homogeneity of the norm $\|\cdot\|^{\mathrm{F}}_{m+n,1}$. We assume that $n \ge 1$ (otherwise the statement of the lemma is trivial).

First of all, function $\|u\|^{\mathrm{F}}_{m+n,1}$ is convex on $\mathcal{B}_{m,1}$, so its maximum over this set is attained at one the extreme points $u^j \in \mathbb{C}_m(\mathbb{Z})$ which are given by $F_m[u^j]^m_{-m} = e^{i\theta}e^j$ where e^j is the discrete Dirac pulse centered at $j \in \mathbb{Z}$, and $\theta \in [0, 2\pi]$. Note that

$$u^j_\tau = \frac{1}{\sqrt{2m+1}} \exp\left[i\left(\theta + \frac{2\pi\tau j}{2m+1}\right)\right],$$

hence, for $\gamma_{m,n} := \sqrt{(2m+2n+1)(2m+1)}$ we obtain

$$\begin{aligned}\left\|u^j\right\|^{\mathrm{F}}_{m+n,1} &= \frac{1}{\gamma_{m,n}} \sum_{k\in \mathrm{D}_{m+n}} \left|\sum_{\tau\in\mathrm{D}_m} \exp\left[2\pi i\tau\left(\frac{j}{2m+1} - \frac{k}{2m+2n+1}\right)\right]\right| \\ &= \frac{1}{\gamma_{m,n}} \sum_{k\in\mathrm{D}_{m+n}} \left|\mathrm{DirKer}_m\left(\omega_{jk}\right)\right|, \quad \text{where} \quad \omega_{jk} := 2\pi\left(\frac{j}{2m+1} - \frac{k}{2m+2n+1}\right),\end{aligned}$$

where $\mathrm{DirKer}_m(\cdot)$ is the Dirichlet kernel of order m:

$$\mathrm{DirKer}_m(\omega) := \begin{cases} \dfrac{\sin\left((2m+1)\omega/2\right)}{\sin\left(\omega/2\right)}, & \omega \ne 2\pi l, \\ 2m+1, & \omega = 2\pi l. \end{cases}$$

Hence,

$$\gamma_{m,n}\|u^j\|^{\mathrm{F}}_{m+n,1} \le \max_{\theta\in[0,2\pi]} \left\{\Sigma_{m,n}(\theta) := \sum_{k\in\mathrm{D}_{m+n}} \left|\mathrm{DirKer}_m\left(\frac{2\pi k}{2m+2n+1} + \theta\right)\right|\right\}. \tag{3.72}$$

For any $\theta \in [0, 2\pi]$, the summation in (3.72) is over the θ-shifted regular $(2m+2n+1)$-grid on the unit circle. The contribution to the sum $\Sigma_{m,n}(\theta)$ of the two closest to $x = 1$ points of this

grid is at most $2(2m+1)$. On the other hand, for the remaining points, we can use the bound

$$\mathrm{DirKer}_m(\omega) \le \frac{1}{|\sin(\omega/2)|} \le \frac{\pi}{\min(\omega, 2\pi-\omega)}.$$

Finally, note that $f(\omega) = \frac{\pi}{\omega}$ decreases on $[\frac{2\pi}{2m+2n+1}, \pi]$ (recall that $n \ge 1$). These considerations result in the following estimate:

$$\Sigma_{m,n}(\theta) \le 2\left(2m+1+\sum_{k=1}^{m+n+1}\frac{2m+2n+1}{2k}\right).$$

Using the bound $H_n \le \log n + 1$ for harmonic numbers, we arrive at

$$\Sigma_{m,n}(\theta) \le 2(2m+1) + (2m+2n+1)\left[\log(m+n+1)+1\right] \le (2m+2n+1)\left[\log(m+n+1)+3\right],$$

and the lemma is proved. □

3.5 Remaining proofs

Proof of relation (3.19). From (3.8) it follows that

$$\|x - \varphi^o * x\|_{n,2} \le \sqrt{2}\kappa_{m,n}\sigma\theta\varrho.$$

On the other hand,

$$\|\varphi^o * \zeta\|_{n,2}^2 = \langle \zeta, M(\varphi^o)\zeta\rangle_n,$$

where $M(\varphi)$ is defined by (3.31). Bounding $\|\varphi^o\|_2 \le \|\varphi^o\|_{m,1}^{\mathrm{F}}$ via (3.7), and using (3.32), we obtain

$$\|M(\varphi^o)\|_{\mathrm{F}}^2 = (2n+1)\|\varphi^o\|_2^2 \le \kappa_{m,n}^2\varrho^2.$$

Deviation bound (3.39) now implies, for any $0 < \alpha \le 1$, that with probability at least $1-\alpha$,

$$\|\varphi^o * \zeta\|_{n,2} \le \sqrt{2}\kappa_{m,n}\varrho\big(1+\sqrt{\log[1/\alpha]}\big), \tag{3.73}$$

and we arrive at (3.19). □

Proof of Proposition 3.2.1. We only give the proof for the contrained estimator $\widehat{\varphi} = \widehat{\varphi}_{\mathrm{con}}$; the penalized estimator can be treated analogously. First, for $t \in \mathbb{Z}$ we decompose

$$\begin{aligned}|[x - \widehat{\varphi} * y]_t| &= |[(\phi^o + (1-\phi^o)) * (x - \widehat{\varphi} * y)]_t| \\ &\le |[\phi^o * (x - \widehat{\varphi} * y)]_t| + |[(1-\widehat{\varphi}) * (1-\phi^o) * x]_t| + \sigma|[\widehat{\varphi} * \zeta]_t| + \sigma|[\widehat{\varphi} * \phi^o * \zeta]_t| \\ &:= \delta^{(1)} + \delta^{(2)} + \delta^{(3)} + \delta^{(4)}.\end{aligned} \tag{3.74}$$

For the remainder of the proof, let $t \in \mathrm{D}_{n-m_0}$. Then, we have

$$\begin{aligned}\delta^{(1)} &\le \|\phi^o\|_2 \big\|\Delta^{-t}[x - \widehat{\varphi} * y]\big\|_{m_0,2} \\ &\le \frac{\rho}{\sqrt{m+1}}\|x - \widehat{\varphi} * y\|_{n,2}.\end{aligned}$$

Using the bound of Corollary 3.2.1 with $\bar{\varrho} = 2\rho^2$, we conclude that with probability $\geq 1 - \alpha/3$,

$$\delta^{(1)} \leq \frac{C\sigma\rho}{\sqrt{m+1}}\left[\kappa_{m,n}\rho^2(1+\theta+\sqrt{\log[1/\alpha]}) + \rho\sqrt{(\kappa_{m,n}^2+1)\log[(m+n)/\alpha] + \varkappa\sqrt{\log[1/\alpha]}} + \sqrt{s}\right].$$

It remains to make sure that the remaining terms are dominated by $\delta^{(1)}$. We get

$$\begin{aligned}\delta^{(2)} &\leq (1+\|\widehat{\varphi}\|_1)\left\|\Delta^{-t}[(1-\phi^o)*x]\right\|_{m_0,\infty} \\ &\leq (1+2\rho^2)\frac{\sigma\theta\rho}{\sqrt{m+1}} \\ &\leq \frac{C\kappa_{m,n}\rho^3\sigma\theta}{\sqrt{m+1}},\end{aligned}$$

where the last transition is due to $n \geq m_0$. Further, by the Parseval's identity,

$$\begin{aligned}\delta^{(3)} &= \sigma|\langle F_m[\widehat{\varphi}], F_m[\Delta^{-t}\zeta]\rangle| \\ &\leq \sigma\|\widehat{\varphi}\|_{m,1}^{\mathrm{F}}\|\Delta^{-t}\zeta\|_{m,\infty}^{\mathrm{F}} \\ &\leq \frac{2\rho^2}{\sqrt{m+1}}\sigma\sqrt{2\log\left[3(2m+1)(2n-2m_0+1)/\alpha\right]},\end{aligned}$$

where the last inequality, holding with probability $\geq 1-\alpha/3$, is due to (3.35). Finally, observe that with probability $\geq 1-\alpha/3$ it holds

$$\|\Delta^{-t}[\phi^o * \zeta]\|_{m,2} \leq \|\phi^o * \zeta\|_{n,2} \leq \sqrt{2}\kappa_{m,n}\rho\left(1+\sqrt{\log[3/\alpha]}\right),$$

cf. (3.73). Therefore, we have for $\delta^{(4)}$:

$$\begin{aligned}\delta^{(4)} &\leq \sigma\|\widehat{\varphi}\|_{m,2}\left\|\Delta^{-t}[\phi^o * \zeta]\right\|_{m,2} \\ &\leq \sigma\frac{2\rho^2}{\sqrt{m+1}}\sqrt{2}\kappa_{m,n}\rho\left(1+\sqrt{\log[3/\alpha]}\right) \\ &= \frac{2\sqrt{2}\kappa_{m,n}\sigma\rho^3}{\sqrt{m+1}}\left(1+\sqrt{\log[3/\alpha]}\right)\end{aligned}$$

with probability $\geq 1-\alpha/3$. Substituting the bounds for $\delta^{(k)}$, $k = 1, ..., 4$, into (3.74), we arrive at the claim. □

3.A Why least-squares estimators cannot be analyzed as Lasso?

Despite striking similarity with Lasso and Dantzig selector [Tib96, CT07, BRT09], the proposed least-squares estimators are of quite different nature. First of all, minimization in these procedures is aimed to recover a filter but not the signal itself, and this filter is not sparse unless for harmonic oscillations with frequencies on the DFT grid. Second, the equivalent of "regression matrices" involved in these procedures cannot be assumed to satisfy the usual "restricted incoherency" conditions usually imposed to prove statistical properties of "classical" ℓ_1-recoveries (see [BVDG11, Chapter 6] for a comprehensive overview of these conditions). Moreover, being constructed from

the noisy signal itself, these matrices depend on the noise, which poses some extra difficulties in the analysis of the properties of these estimators, in particular, leading to the necessity of Assumption 3.2.1. Let us briefly illustrate these difficulties.

Let $m = n$ for simplicity, and, given $y \in \mathbb{C}(\mathbb{Z})$, let $T(y)$ be the $(2n+1)\times(2n+1)$ "convolution matrix" as defined by (3.29) such that for $\varphi \in \mathbb{C}_n(\mathbb{Z})$ one can write $[\varphi * y]_0^n = T(y)[\varphi]_{-n}^n$. When denoting $f = F_n[\varphi]$, the optimization problem in (**Con-LS**) can be recast as a "standard" ℓ_1-constrained least-squares problem with respect to f:

$$\min_{f\in\mathbb{C}^{2n+1}} \left\{ \|y - A_n f\|_{n,2}^2 : \|f\|_1 \leq \frac{\bar{\varrho}}{\sqrt{2n+1}} \right\}, \tag{3.75}$$

where $A_n = T(y)F_n^{\mathrm{H}}$. Observe that $f^o = F_n[\varphi^o]$ is feasible for (3.75), so that

$$\|y - A_n\widehat{f}\|_{n,2}^2 \leq \|y - A_n f^o\|_{n,2}^2,$$

where $\widehat{f} = F_n[\widehat{\varphi}]$, and

$$\begin{aligned}
\|x - A_n\widehat{f}\|_{n,2}^2 - \|x - A_n f^o\|_{n,2}^2 &\leq 2\sigma\big(\Re\langle\zeta, x - A_n f^o\rangle_n - \Re\langle\zeta, x - A_n\widehat{f}\rangle_n\big) \\
&\leq 2\sigma\big|\langle\zeta, A_n(f^o - \widehat{f})\rangle_n\big| \\
&\leq 2\sigma\|A_n^{\mathrm{H}}[\zeta]_{-n}^n\|_\infty\|f^o - \widehat{f}\|_1 \\
&\leq 4\sigma\|A_n^{\mathrm{H}}[\zeta]_{-n}^n\|_\infty\frac{\bar{\varrho}}{\sqrt{n+1}}.
\end{aligned}$$

In the "classical" situation, where $[\zeta]_{-n}^n$ is independent of A_n (see, e.g., [JN00]), one has

$$\|A_n^{\mathrm{H}}[\zeta]_0^n\|_\infty \leq c_\alpha\sqrt{\log n}\max_j\|[A_n]_j\|_2 \leq c_\alpha\sqrt{n\log n}\max_{i,j}|A_{ij}|,$$

where c_α is a logarithmic in α^{-1} factor. This would rapidly lead to the bound (3.12). In the case we are interested in, where A_n incorporates observations $[y]_{-n}^n$ and thus depends on $[\zeta]_{-n}^n$, curbing the cross term is more involved and explicitly requires Assumption 3.2.1.

Chapter 4

Shift-invariance and recoverability

Conditions of the main results of the previous chapter – oracle inequalities for least-squares estimators – merit some discussion. The primary question is how signal recoverability and predictability, cf. Definitions 2.1.1–2.2.1, is related to Assumptions 3.2.1–3.2.2 about its proximity to a shift-invariant subspace. In this brief chapter, we present results that shed some light on this relationship.

- First, we demonstrate that signals which are *uniformly close* to shift-invariant subspaces are recoverable with moderate dependency of parameters (ρ, θ) from the approximate shift-invariance parameters $(s, \varkappa)$.
- We also show that an important subset of such signals – those which are close to solutions of homogeneous linear difference equations whose solutions are prohibited from increasing or decreasing exponentially fast – are also easy to filter and predict.
- Finally, we use the above result to tackle the denoising problem for *harmonic oscillations* without imposing any frequency separation assumption, cf. Section 1.5.

Deferring all proofs to Section 4.3, we now embark on the presentation of the results.

4.1 General situation

4.1.1 Reduction to the exact case

We start with an auxilliary result which allows us to concentrate on the case of *exact* shift-invariance, *i.e.* when the signal belongs to a shift-invariant subspace without any approximation.

Proposition 4.1.1. *Suppose that $x \in \mathcal{R}_{m,n}(\rho, \theta)$, cf. Definition 2.1.1, with $\rho \geq 1$. Then, $\tilde{x} := x + \varepsilon$ with ε satisfying*

$$\|\Delta^\tau \varepsilon\|_{m,\infty} \leq \frac{\varkappa\sigma}{\sqrt{2n+1}}, \quad -m-n \leq \tau \leq m+n, \tag{4.1}$$

belongs to $\mathcal{R}_{m,n}(\rho, \tilde{\theta})$ with

$$\tilde{\theta} = \theta + \frac{2\varkappa}{\kappa_{m,n}}.$$

Similarly, assume that $x \in \mathcal{R}_{m,n,h}(\rho,\theta)$, *cf. Definition 2.2.1, with* $\rho \geq 1$, *and suppose that*

$$\|\Delta^\tau \varepsilon\|_{m,\infty} \leq \frac{\varkappa\sigma}{\sqrt{n+1}}, \quad h \leq \tau \leq m+n+4h. \tag{4.2}$$

Then, $\tilde{x}$ *belongs to* $\mathcal{R}_{m,n,h}(\rho,\tilde{\theta})$ *with*

$$\tilde{\theta} = \theta + \frac{2\sqrt{2}\varkappa}{\kappa_{m,n}}.$$

Observe that if x is an element of a shift-invariant subspace, then $\tilde{x} = x + \varepsilon$ from the premise of Proposition 4.1.1 satisfies Assumption 3.2.1 in case of (4.1) and Assumption 3.2.2 in case of (4.2). On the other hand, quite naturally, requirements (4.1), (4.2) in this case are are stronger than the corresponding conditions (3.11), (3.26) in Assumptions 3.2.1, 3.2.2: to ensure simplicity. we need proximity to the subspace to be measured in ℓ_∞-norm, rather than in ℓ_2-norm.

Remark 4.1.1. As we are about to see, signals x belonging to shift-invariant subspaces $\mathcal{S}$ can be equivalently derived as the discretized solutions to homogeneous ordinary differential equations. Namely,

$$x_\tau = f(\tau/N), \quad \tau = 0,1,...,N-1,$$

where $f : [0,1] \to \mathbb{R}$ satisfies

$$\mathrm{p}\left(\frac{\mathrm{d}}{\mathrm{d}u}\right) f(u) = 0, \quad u \in [0,1]$$

for some polynomial $\mathrm{p}(z)$ of degree $\dim(\mathcal{S})$. As such, one can show that in case of (4.1) or (4.2), $\tilde{x}$ corresponds to a function $\tilde{f}$ which satisfies a Hölder-type condition

$$\sup_{u\in[0,1]} \left| p\left(\frac{\mathrm{d}}{\mathrm{d}u}\right) \tilde{f}(u) \right| \leq L,$$

whereas (3.11), (3.26) correspond to Sobolev-type conditions in which the sup-norm is replaced with the L_2-norm, see [Nem00, Lemma 4.3.1].

4.1.2 Exact case

As we have already mentioned in Chapter 3, shift-invariant subspaces of $\mathbb{C}(\mathbb{Z})$ are closely related to homogeneous linear difference equations with constant coefficients. In fact, shift-invariant subspaces with fixed dimension are exactly the *solution sets* of such equations with fixed order. This is formally stated by the following simple proposition whose proof we provide in Section 4.3[1].

Proposition 4.1.2. *Solution set of a difference equation*

$$[p(\Delta)x]_t \left[= \sum_{\tau=0}^{s} p_\tau x_{t-\tau}\right] \equiv 0, \quad t \in \mathbb{Z}, \tag{4.3}$$

with a polynomial $p(z) = 1 + p_1 z + ... + p_s z^s$ *is a shift-invariant subspace of* $\mathbb{C}(\mathbb{Z})$ *with* $\dim(\mathcal{S}) = s$.

[1] Analogues of Proposition 4.1.1 are known in the continuous setting [AK64]; the result for the discrete-time setting follows from those for Abelian groups, see [Lai79], [Szé82], but we are not aware of any elementary proof.

Conversely, any shift-invariant subspace $\mathcal{S}$ of $\mathbb{C}(\mathbb{Z})$ with $\dim(\mathcal{S}) = s$ is the solution set of a difference equation of the form (4.3) with $\deg(p) = s$. Moreover, such polynomial is unique if normalized by $p(0) = 1$.

Note that the set of solutions of (4.3) with a fixed polynomial $\mathrm{p}(z)$ is spanned by *exponential polynomials* determined by the roots of $\mathrm{p}(z)$. Specifically, let for $k = 1, ..., r \leq s$ the numbers z_k be the distinct roots of $\mathrm{p}(z)$ with the corresponding multiplicities m_k, and choose $\omega_k \in \mathbb{C}$ such that $z_k = e^{i\omega_k}$. Then the solutions to (4.3) can be expressed as

$$x_t = \sum_{k=1}^{r} \mathrm{q}_k(t) e^{i\omega_k t}, \tag{4.4}$$

where $\mathrm{q}_k(\cdot)$ are arbitrary polynomials with $\deg(\mathrm{q}_k) = m_k - 1$. For instance, discrete-time polynomials of degree $s-1$ satisfy (4.3) with $\mathrm{p}(z) = (1-z)^s$. Another important example is that of *harmonic oscillations*

$$x_t = \sum_{k=1}^{s} C_k \mathrm{e}^{i\omega_k t}, \quad \omega_k \in [0, 2\pi] \tag{4.5}$$

which satisfy (4.3) with $\mathrm{p}(z) = \prod_{k=1}^{s}(1 - e^{i\omega_k} z)$. The set of harmonic oscillations with fixed frequencies $\omega_1, ..., \omega_s$ and varying complex amplitudes C_k form an s-dimensional shift-invariant subspace depending on the frequencies.

Interpolation for general shift-invariant subspaces. As the next result shows, elements of *arbitrary* low-dimensional shift-invariant subspaces of $\mathbb{C}(\mathbb{Z})$ (equivalently, solutions of (4.3) exponential polynomials (4.4)) are always simple. More precisely, such signals can always be interpolated by a filter $\phi \in \mathbb{C}_m(\mathbb{Z})$ with moderate ℓ_2-norm which *exactly* reproduces the signal:

Proposition 4.1.3. *Let $\mathcal{S}$ be a shift-invariant subspace of $\mathbb{C}(\mathbb{Z})$ with dimension $s \leq m+1$, and $x \in \mathcal{S}$. Then, there exists a filter $\phi^o \in \mathbb{C}_m(\mathbb{Z})$, which only depends on $\mathcal{S}$ and not on x, such that $x_t = [\phi^o * x]_t$ for any $t \in \mathbb{Z}$, and*

$$\|\phi^o\|_2 \leq \sqrt{\frac{2s}{2m+1}}. \tag{4.6}$$

As such, for any $n \in \mathbb{Z}_+$ one has $x \in \mathcal{R}_{m,n}(\rho, 0)$ with $\rho = \sqrt{2s}$.

Note that the bound $\rho = O(\sqrt{s})$ attained by the oracle filter ϕ^o from Proposition 4.1.3 is the best one could hope for provided that ϕ^o is only allowed to depend on the subspace but not on the signal itself. Indeed, the better dependency $\rho(s)$ would contradict the minimax risk bound $O(\sigma\sqrt{s/n})$ for a subspace that holds for *all possible estimators*, not only convolution-type ones, see *e.g.* [Joh11]. One should note, however, that feasible filter φ in Theorems 3.2.1–3.2.3 is allowed to depend on x, and hence φ might have a smaller risk than as implied by Proposition 4.1.3. Then, we are guaranteed to "mimic" the statistical properties of this filter, in the sense of the results obtained in Chapter 3, whenever it exists, and the price of adaptation is guaranteed to be controlled in terms of its norm.

4.2 Generalized harmonic oscillations

4.2.1 Bounds for prediction

On the other hand, it is clear that Assumption 3.2.1 is not sufficient to imply predictability of x, *i.e.* when one is allowed to use only unilateral filters. One can see, for instance, that already the signals coming from the parametric family

$$\mathcal{X}_\alpha = \{x \in \mathbb{C}(\mathbb{Z}) : x_\tau = \beta\alpha^\tau, \beta \in \mathbb{C}\},$$

for given $\alpha : |\alpha| > 1$, which form a one-dimensional shift-invariant subspace of $\mathbb{C}(\mathbb{Z})$ defined by $(1 - \alpha\Delta)x \equiv 0$, cannot be estimated consistently at $t = 0$ using only observations on the left of t, and thus do not satisfy Definition 2.2.1. Of course, this difficulty with $\mathcal{X}_\alpha$ is due to the *instability* of solutions of a difference equation which do not remain bounded when $\tau \to +\infty$. Meanwhile, "stable" signals – decaying exponents, harmonic oscillations, and their products – *are* predictable. More generally, suppose again that x belongs to a shift-invariant subspace, or equivalently, satisfies a difference equation (4.3) with characteristic polynomial $\mathrm{p}(z)$. When $p(z)$ has at least one root z_k such that $|z_k| < 1$, the solution set of (4.3) contains signals unbounded as $\tau \to \infty$, which cannot be estimated by a "causal" filter $\phi \in \mathbb{C}^h_m(\mathbb{Z})$ with $h \geq 0$. On the other hand, when $p(z)$ has a root z_k with $|z_k| > 1$, the set of solutions to (4.3) contains signals unbounded as $\tau \to -\infty$, which cannot be estimated by any "anti-causal" filter $\phi \in \mathbb{C}^h_m(\mathbb{Z})$ with $h \leq -m$.

In view of the above, it is interesting to consider the case where all roots of $\mathrm{p}(z)$ have unit modulus. In this case, the solutions of (4.3) are exponential polynomials (4.4) with

$$\omega_1, ..., \omega_s \in [0, 2\pi];$$

we call such signals *generalized harmonic oscillations* as they are sums of polynomially-modulated complex sinusoids. This class of signals has already been studied in [JN13], where the authors showed that a harmonic oscillation with s frequencies are simple in the sense of *filtering*, *i.e.* can reproduced by a small-norm one-sided filter $\phi^o \in \mathbb{C}^+_m(\mathbb{Z})$. Namely, such signals belong to any class $\mathcal{R}_{m,n,0}(\rho, 0)$ with arbirtrary $n \in \mathbb{Z}_+$, whenever m is large enough, and $\rho = \tilde{O}(s^{3/2})$, as formally stated below.

Proposition 4.2.1 ([JN13, Lemma 6.1]). *Suppose that all roots of* $\mathrm{p}(z) = 1 + p_1 z + ... + p_s z^s$ *satisfy* $|z_k| = 1$. *Then there is an absolute constant* c *such that for any* $m \geq cs^2 \log s$, *one can construct a filter* $\phi^o \in \mathbb{C}^+_m(\mathbb{Z})$, *which only depends on* $\mathrm{p}(z)$, *such that any solution to* (4.3) *satisfies* $x_t = [\phi^o * x]_t$ *for any* $t \in \mathbb{Z}$, *and*

$$\|\phi^o\|_2 \leq C\sqrt{\frac{s^3 \log(s+1)}{m+1}}. \tag{4.7}$$

We show an improvement upon this result. Its proof uses some complex-analytical techniques, and is given in Section 3.5.

Proposition 4.2.2. *Under the premise of Proposition 4.2.1, one can replace* (4.7) *with*

$$\|\phi^o\|_2 \leq C\sqrt{\frac{s^2 \log(ms+1)}{m+1}}. \tag{4.8}$$

Moreover, when dealing with "ordinary" harmonic oscillation given by (4.5)), the bound (4.8) can be further improved under the additional condition that the frequencies $\omega_1, ..., \omega_s$ are *well-separated*, see [DB13, TBR13, CFG14]. Namely, assume that all the roots $z_k = e^{i\omega_k}$ of $\mathrm{p}(z)$ are simple, let $|x - y|$ be the wrap-around metric on $[0, 2\pi]$, and consider *the minimal frequency separation* $\delta_{\min}$ defined as

$$\delta_{\min} := \min_{1 \le j \ne k \le s} |\omega_j - \omega_k|. \tag{4.9}$$

The following result shows that whenever $\delta_{\min}$ is large enough, the bound $\rho = \tilde{O}(s)$ can be improved to $\rho = O(\sqrt{s})$.

Proposition 4.2.3. *For some $\nu > 1$, let*

$$\delta_{\min} \ge \frac{2\pi\nu}{m+1}. \tag{4.10}$$

Then there exists a filter $\phi^o \in \mathbb{C}^+_m(\mathbb{Z})$ satisfying $x_t = [\phi^o * x]_t = 0$ for any $t \in \mathbb{Z}$, and such that

$$\|\phi^o\|_2 \le \sqrt{\frac{Qs}{m+1}}, \quad \textit{where} \quad Q = \frac{\nu+1}{\nu-1}.$$

In particular, whenever $\delta_{\min} \ge 4\pi/n$, one has

$$\|\phi^o\|_2 \le \sqrt{\frac{3s}{m+1}}.$$

4.2.2 Full recovery of harmonic oscillations

To illustrate the results obtained in Section 4.1.2, let us consider the problem of *full recovery* of (ordinary) harmonic oscillations. Namely, we are asked to estimate on D_N a harmonic oscillation

$$x_t = \sum_{k=1}^{s} \alpha_k e^{i\omega_k t}, \quad t \in \mathrm{D}_N, \quad N \in \mathbb{Z}_+,$$

without the knowledge of frequencies $\omega_1, ..., \omega_s$. Since estimation is required on the entire D_N, we will measure the statistical performance of an estimator $\widehat{x}$ of x via the mean-square error[2]:

$$\mathrm{MSE}(\widehat{x}, x) := (2N+1)^{-1/2} \left[\mathbb{E}\|\widehat{x} - x\|^2_{N,2}\right]^{1/2}.$$

Note that if the frequencies were known, the ordinary least-squares estimator would satisfy

$$\sup_{\mathcal{S}_{\omega_1,...,\omega_s}} \mathrm{MSE}(\widehat{x}, x) \le C\sigma\sqrt{\frac{s}{2N+1}},$$

where $\mathcal{S}_{\omega_1,...,\omega_s}$ is the set of harmonic oscillations corresponding to the fixed tuple of frequencies $\omega_1, ..., \omega_s$. On the other hand, with unknown frequencies one has the lower bound [TBR13,

[2]The results of this section can be generalized, in a straightforward manner, for the risk measured by the width of the confidence interval for ℓ_2-loss. Here we omit this generalization in order to simplify the presentation.

Theorem 2]

$$\sup_{x\in\mathcal{S}^{(s)}} \mathrm{MSE}(\widehat{x},x) \geq c\sigma\sqrt{\frac{s\log(N+1)}{2N+1}}, \tag{4.11}$$

where $\mathcal{S}^{(s)}$ is the set of all harmonic oscillations with no more than s frequencies. Moreover, the bound (4.11) is in fact attained on a subspace with *separated* frequencies in the sence of (4.10), indicating that the general case is not harder, from the statistical viewpoint, than that of well-separated frequencies . As such, one could hope to match (4.11) by some adaptive in $\omega_1,...,\omega_s$ estimator, whether the frequencies are restricted to be well-separated or not. However, to the best of our knowledge, the only estimator known to match (4.11), called *Atomic Soft Thresholding* (AST) and studied in [BTR13, TBR13], only does so in the case of well-separated frequencies, see [TBR13, Theorem 1]. As such, the question whether the lower bound (4.11) can be matched in the general case, is still open.

A crucial step towards bridging this gap has been taken in [HJNO15] where it was suggested to use one-sided version of a uniform-fit estimator jointly with Proposition 4.2.1, exploting that (4.7) holds for (generalized) harmonic oscillations without any frequency separation assumptions. In particular, fitting a "left" uniform-fit estimator $\widehat{\varphi}\in\mathbb{C}^+_N(\mathbb{Z})$ on D^+_N, and a "right" estimator $\widehat{\varphi}\in\mathbb{C}^-_N(\mathbb{Z})$ on D^-_N, and using the bound $C\rho^3\sqrt{\log N}$ on the price of adaptation for such estimators[3], one obtains for such construction the correct in σ and N rate

$$\sup_{x\in S^{(s)}} \mathrm{MSE}(\widehat{x},x) \leq C\sigma\sqrt{\frac{s^{12}\log(N+1)}{2N+1}}.$$

As we see, the price to pay is a polynomial factor in s, and additional assumption $N\geq cs^2\log s$ needed for Proposition 4.2.1.

Using the results presented in ths paper, we immeduately improve the dependence on s in the above bound by replacing one-sided uniform-fit estimators with estimators of the form (**Con-LS**), used together with Theorem 3.2.2 and the improved bound (4.8) instead of (4.7):

$$\sup_{x\in S^{(s)}} \mathrm{MSE}(\widehat{x},x) \leq C\sigma\sqrt{\frac{s^4\log^2(N+1)}{2N+1}}. \tag{4.12}$$

Note that while this estimators requries the knowledge of s in advance (AST does not), this requirement can be circumvented by using (**Pen^2-LS-Pred**) instead of (**Con-LS-Pred**), at the expense of an additional logarithmic factor. Moreover, (4.12) can be further improved if the signal frequencies are restricted to be well-separated. In this case, using Proposition 4.2.3, the same estimator satisfies

$$\sup_{x\in S^{(s)}_{\mathrm{sep}}} \mathrm{MSE}(\widehat{x},x) \leq C\sigma\sqrt{\frac{s^2+s\log(N+1)}{2N+1}} \tag{4.13}$$

where the supremum is taken over the set of harmonic oscillations with no more than s frequencies and pairwise separation at least $4\pi/(N+1)$.

Finally, let us describe how the results obtained so far can be combined, resulting in the state-of-the-art estimator of harmonic oscillations which improves over the bound (4.12) in the

[3]This bound holds for both two-sidfed and one-sided uniform-fit estiamtors, see [HJNO15].

general case while still preserving (4.13). Namely, consider the following procedure:

1. Pick some $M \leq N$, and divide the observation domain D_N into the large central subdomain D_M and the smaller subdomains $\mathrm{D}^+ := \mathrm{D}^M_{N-M}$ and $\mathrm{D}^- := \mathrm{D}^{-N}_{N-M}$.
2. Estimate the signal on D_M with a two-sided filter $\widehat{\varphi} \in \mathbb{C}_{N-M}(\mathbb{Z})$, on D^+ with a one-sided filter $\widehat{\varphi}^+ \in \mathbb{C}^+_{M+N}(\mathbb{Z})$, and on D^- with a one-sided filter $\widehat{\varphi}^- \in \mathbb{C}^-_{M+N}(\mathbb{Z})$.
3. Choose M to minimize the total bound on MSE over D_N.

Direct calculations lead to the following choice of M:

$$C_1 s \log(N+1) \leq \frac{M+1}{N-M+1} \leq C_2 s \log(N+1),$$

and the resulting bound is

$$\sup_{x \in S^{(s)}} \mathrm{MSE}(\widehat{x}, x) \leq C\sigma \sqrt{\frac{s^3 \log(N+1) + s^2 \log^2(N+1)}{2N+1}}. \tag{4.14}$$

These calculations are provided in Section 4.3.

4.3 Proofs

Proof of Proposition 4.1.1. Let $\phi^o \in \mathbb{C}_m(\mathbb{Z})$ be an oracle for $x \in \mathcal{R}_{m,n}(\rho, \theta)$, and let us use it to estimate $\tilde{x} = x + \varepsilon$. Then,

$$|\tilde{x}_t - [\phi^o * \tilde{x}]_t| \leq |x_t - [\phi^o * x]_t| + |\varepsilon_t| + |[\phi^o * \varepsilon]_t|, \quad t \in \mathrm{D}_{m+n}.$$

The first term in the right-hand side is at most $\frac{\sigma\theta\rho}{\sqrt{2m+1}}$ since $x \in \mathcal{R}_{m,n}(\rho, \theta)$. The second term can be bounded by (4.1) directly: using that $\rho \geq 1$,

$$|\varepsilon_t| \leq \frac{\sigma\varkappa}{\sqrt{2n+1}} = \frac{\sigma\varkappa\rho}{\kappa_{m,n}\sqrt{2m+1}}, \quad t \in \mathrm{D}_{m+n}.$$

The last term is controlled by the Cauchy-Schwarz inequality: for any $\tau \in \mathrm{D}_{m+n}$,

$$|[\phi^o * \varepsilon]_\tau| \leq \|\phi^o\|_2 \, \|\Delta^{-\tau}\varepsilon\|_{m,2} \leq \frac{\sigma\varkappa\rho}{\kappa_{m,n}\sqrt{2m+1}}.$$

The proof for the case of prediction is obtained in the same manner; the only adjustments are t on which the pointwise error must be controlled and a different scaling factor

$$\sqrt{\frac{m+1}{n+1}} \leq \sqrt{\frac{4m+2}{2n+1}} = \sqrt{2}\kappa_{m,n}. \qquad \square$$

Proof of Proposition 4.1.2. As a precursory remark, note that if a finite-dimensional subspace $\mathcal{S}$ is shift-invariant, i.e. $\Delta\mathcal{S} \subseteq \mathcal{S}$, then necessary Δ is a bijection on $\mathcal{S}$, and $\Delta\mathcal{S} = \mathcal{S}$. Indeed, when restricted on $\mathcal{S}$, Δ is a linear transformation with a trivial kernel and hence a bijection.

1^o. To prove the direct statement, note that the solution set of (4.3) with $\deg(\mathrm{p}(\cdot)) = s$ is a shift-invariant subspace of $\mathbb{C}(\mathbb{Z})$ – let us call it $\mathcal{S}'$. Indeed, if $x \in \mathbb{C}(\mathbb{Z})$ satisfies (4.3), so does Δx, so $\mathcal{S}'$ is shift-invariant. To see that $\dim(\mathcal{S}') = s$, note that $x \mapsto x_1^s$ is a bijection $\mathcal{S}' \to \mathbb{C}^s$: under this map arbitrary $x_1^s \in \mathbb{C}^s$ has a unique preimage. Indeed, as soon as one fixes x_1^s, (4.3) uniquely defines the next samples $x_{s+1}, x_{s+2}, ...$ (note that $\mathrm{p}(0) \neq 0$); dividing (4.3) by Δ^s, one can retrieve the remaining samples of x since $\deg(\mathrm{p}(\cdot)) = s$ (we used that Δ is bijective on $\mathcal{S}$).

2^o. To prove the converse, first note that any polynomial $\mathrm{p}(\cdot)$ with $\deg(\mathrm{p}(\cdot)) = s$ and such that $\mathrm{p}(0) = 1$ is uniquely expressed via its roots $z_1, ..., z_s$ as

$$\mathrm{p}(z) = \prod_{k=1}^{s} (1 - z/z_k).$$

Since $\mathcal{S}$ is shift-invariant, we have $\Delta\mathcal{S} = \mathcal{S}$ as discussed above, *i.e.* Δ is a bijective linear operator on $\mathcal{S}$. Let us fix some basis $\boldsymbol{e} = [e^1; ...; e^s]$ on $\mathcal{S}$ and denote A the $s \times s$ matrix of Δ in this basis, that is, $\Delta(e^j) = \sum_{i=1}^{s} a_{ij} e^i$. Moreover, by the Jordan theorem basis $\boldsymbol{e}$ can be chosen such that A is upper-triangular. Then, any vector $x \in \mathcal{S}$ satisfies $\mathrm{q}(\Delta)x \equiv 0$, where

$$\mathrm{q}(\Delta) = \prod_{i=1}^{s} (\Delta - a_{ii}) = \det(\Delta I - A)$$

is the characteristic polynomial of A. Note that $\det A = \prod_{i=1}^{s} a_{ii} \neq 0$ since Δ is a bijection. Hence, choosing

$$\mathrm{p}(\Delta) = \frac{\mathrm{q}(\Delta)}{\det A},$$

we obtain $\prod_{i=1}^{s}(1 - \Delta c_i)x \equiv 0$ for some complex coefficients $c_i \neq 0$. This means that $\mathcal{S}$ is contained in the solution set $\mathcal{S}'$ of (4.3) with $\deg(\mathrm{p}(\cdot)) = s$ and such that $\mathrm{p}(0) = 1$. Note that by **1^o**, $\mathcal{S}'$ is also a shift-invariant subspace of dimension s, thus $\mathcal{S}$ and $\mathcal{S}'$ must coincide. Finally, uniqueness of $\mathrm{p}(\cdot)$ follows from the fact that $\mathrm{q}(\cdot)$ is a characteristic polynomial of A. □

Proof of Proposition 4.1.3. Let $\Pi_{\mathcal{S}_m}$ be the Euclidean projection on $\mathcal{S}_m := \mathcal{S} \cap \mathbb{C}^+_m(\mathbb{Z})$. Since $\dim(\mathcal{S}_m) \leq s$, one has

$$\|\Pi_{\mathcal{S}_m}\|_2^2 = \mathrm{Tr}(\Pi_{\mathcal{S}_m}) \leq s.$$

As such, there is a $j \in \{0, ..., m\}$ such that the j-th row $\pi = [\Pi_{\mathcal{S}_m}]_j$ of $\Pi_{\mathcal{S}_m}$ satisfies

$$\|\pi\|_2 \leq \sqrt{\frac{s}{m+1}} \leq \sqrt{\frac{2s}{2m+1}}.$$

On the other hand, since $\Pi_{\mathcal{S}_m}$ is the projector on $\mathcal{S}_m$, one has $x_j - \langle \pi, x_0^m \rangle = 0$ for any $x \in \mathcal{S}$. Hence, using that $\Delta\mathcal{S} = \mathcal{S}$, for any $k \in \mathbb{Z}$ we have

$$x_\tau - \langle \pi, x_{\tau-j}^{\tau-j+m} \rangle = 0, \quad \tau \in \mathbb{Z}.$$

Finally, let $\phi^o \in \mathbb{C}_m(\mathbb{Z})$ be obtained by augmenting π with zeroes in such a way that the j-th entry of π becomes the central entry of ϕ^o. Obviously, $\phi^o \in \mathbb{C}_m(\mathbb{Z})$; on the other hand,

$$\|\phi^o\|_2 \le \sqrt{\frac{2s}{2m+1}} \quad \text{and} \quad x_t - [\phi^o * x]_t = 0, \quad t \in \mathbb{Z}. \qquad \square$$

Proof of Proposition 4.2.2. Note that in order to prove the theorem, we have to exhibit a vector $q \in \mathbb{C}^{n+1}$ of small ℓ_2-norm and such that the polynomial $1 - q(z) = 1 - \left[\sum_{i=0}^n q_i z^i\right]$ is divisible by $p(z)$, i.e., that there is a polynomial $r(z)$ of degree $n - s$ such that

$$1 - q(z) = r(z)p(z).$$

Indeed, this would imply that

$$x_t - [q * x]_t = [1 - q(\Delta)]x_t = r(\Delta)p(\Delta)x_t = 0$$

due to $p(\Delta)x_t = 0$,

Our objective is to prove inequality

$$\|q\|_2 \le C's\sqrt{\frac{\log[ns]}{n}}.$$

So, let $\theta_1, ..., \theta_s$ be complex numbers of modulus 1 – the roots of the polynomial $p(z)$. Given $\delta = 1 - \epsilon \in (0, 1)$, let us set $\bar\delta = 2\delta/(1 + \delta)$, so that

$$\frac{\bar\delta}{\delta} - 1 = 1 - \bar\delta > 0. \tag{4.15}$$

Consider the function

$$\bar q(z) = \prod_{i=1}^s \frac{z - \theta_i}{\delta z - \theta_i}.$$

Note that $\bar q(\cdot)$ has no singularities in the circle

$$\mathcal{B} = \{z : |z| \le 1/\bar\delta\};$$

besides this, we have $\bar q(0) = 1$.Let $|z| = 1/\bar\delta$, so that $z = \bar\delta^{-1}w$ with $|w| = 1$. We have

$$\frac{|z - \theta_i|}{|\delta z - \theta_i|} = \frac{1}{\delta}\frac{|w - \bar\delta\theta_i|}{|w - \frac{\bar\delta}{\delta}\theta_i|}.$$

We claim that when $|w| = 1$, $|w - \bar\delta\theta_i| \le |w - \frac{\bar\delta}{\delta}\theta_i|$.

> Indeed, assuming w.l.o.g. that w is not proportional to θ_i, consider triangle Δ with the vertices $A = w$, $B = \bar\delta\theta_i$ and $C = \frac{\bar\delta}{\delta}\theta_i$. Let also $D = \theta_i$. By (4.15), the segment $\overline{AD}$ is a median in Δ, and $\angle CDA$ is $\ge \frac{\pi}{2}$ (since D is the closest to C point in the unit circle, and the latter contains A), so that $|w - \bar\delta\theta_i| \le |w - \frac{\bar\delta}{\delta}\theta_i|$.

As a consequence, we get

$$z \in \mathcal{B} \;\Rightarrow\; |\bar q(z)| \le \delta^{-s}, \tag{4.16}$$

whence also

$$|z| = 1 \;\Rightarrow\; |\bar{q}(z)| \le \delta^{-s}. \tag{4.17}$$

Now, the polynomial $p(z) = \prod_{i=1}^{s}(z - \theta_i)$ on the boundary of $\mathcal{B}$ clearly satisfies

$$|p(z)| \ge \left[\frac{1}{\bar{\delta}} - 1\right]^s = \left[\frac{1-\delta}{2\delta}\right]^s,$$

which combines with (4.16) to imply that the modulus of the holomorphic in $\mathcal{B}$ function

$$\bar{r}(z) = \left[\prod_{i=1}^{s}(\delta z - \theta_i)\right]^{-1}$$

is bounded with $\delta^{-s}\left[\frac{1-\delta}{2\delta}\right]^{-s} = \left[\frac{2}{1-\delta}\right]^s$ on the boundary of $\mathcal{B}$. It follows that the coefficients r_j of the Taylor series of $\bar{r}$ satisfy

$$|r_j| \le \left[\frac{2}{1-\delta}\right]^s \bar{\delta}^j, \quad j = 0, 1, 2, ...$$

When setting

$$q^\ell(z) = p(z)r^\ell(z), \quad r^\ell(z) = \sum_{j=1}^{\ell} r_j z^j, \tag{4.18}$$

for $|z| \le 1$, utilizing the trivial upper bound $|p(z)| \le 2^s$, we get

$$\begin{aligned} |q^\ell(z) - \bar{q}(z)| &\le |p(z)||r^\ell(z) - \bar{r}(z)| \\ &\le 2^s \left[\frac{2}{1-\delta}\right]^s \sum_{j=\ell+1}^{\infty} |r_j| \\ &\le \left[\frac{4}{1-\delta}\right]^s \frac{\bar{\delta}^{\ell+1}}{1-\bar{\delta}}. \end{aligned} \tag{4.19}$$

Note that $q^\ell(0) = p(0)r^\ell(0) = p(0)\bar{r}(0) = 1$, that q^ℓ is a polynomial of degree $\ell + s$, and that q^ℓ is divisible by $p(z)$. Besides this, on the unit circumference we have, by (4.19),

$$\begin{aligned} |q^\ell(z)| &\le |\bar{q}(z)| + \left[\frac{4}{1-\delta}\right]^s \frac{\bar{\delta}^{\ell+1}}{1-\bar{\delta}} \\ &\le \delta^{-s} + \underbrace{\left[\frac{4}{1-\delta}\right]^d \frac{\bar{\delta}^{\ell+1}}{1-\bar{\delta}}}_{R}, \end{aligned} \tag{4.20}$$

where we used (4.17). Now,

$$\bar{\delta} = \frac{2\delta}{1+\delta} = \frac{2-2\epsilon}{2-\epsilon} = \frac{1-\epsilon}{1-\epsilon/2} \le 1 - \epsilon/2 \le \mathrm{e}^{-\epsilon/2},$$

and

$$\frac{1}{1-\bar{\delta}} = \frac{1+\delta}{1-\delta} = \frac{2-\epsilon}{\epsilon} \leq \frac{2}{\epsilon}.$$

We can bound from above R:

$$R = \left[\frac{4}{1-\delta}\right]^s \frac{\bar{\delta}^{\ell+1}}{1-\bar{\delta}} \leq \frac{2^{2s+1}}{\epsilon^{s+1}} \mathrm{e}^{-\epsilon\ell/2}$$

Now, given positive integer ℓ and positive α such that

$$\frac{\alpha}{\ell} \leq \tfrac{1}{4}, \tag{4.21}$$

let $\epsilon = \frac{\alpha}{2\ell s}$. Since $0 < \epsilon \leq \frac{1}{8}$, we have $-\log(\delta) = -\log(1-\epsilon) \leq 2\epsilon = \frac{\alpha}{\ell s}$, implying that $\bar{\delta} \leq \mathrm{e}^{-\epsilon/2} = \mathrm{e}^{-\frac{\alpha}{4\ell s}}$, and

$$R \leq \left[\frac{8\ell s}{\alpha}\right]^{s+1} \exp\{-\frac{\alpha}{4s}\}.$$

Now let us put

$$\alpha = \alpha(\ell, s) = 4s(s+2)\log(8\ell s);$$

observe that this choice of α satisfies (4.21), provided that

$$\ell \geq O(1)s^2 \log(s+1) \tag{4.22}$$

with properly selected absolute constant $O(1)$. With this selection of α, we have $\alpha \geq 1$, whence

$$\begin{aligned} R\left[\frac{\alpha}{\ell}\right]^{-1} &\leq \exp\left\{-\frac{\alpha}{4s}\right\}\left[\frac{8\ell s}{\alpha}\right]^{s+1}\frac{\ell}{\alpha} \\ &\leq \exp\left\{-\frac{\alpha}{4s}\right\}[8\ell s]^{s+2} \\ &\leq \exp\{-(s+2)\log(8\ell s)\}\exp\{(s+2)\log(8\ell s)\} = 1, \end{aligned}$$

that is,

$$R \leq \frac{\alpha}{\ell} \leq \frac{1}{4}. \tag{4.23}$$

Furthermore,

$$\begin{array}{rcl} \delta^{-s} & = & \exp\{-s\log(1-\epsilon)\} \leq \exp\{2\epsilon s\} = \exp\{\frac{\alpha}{\ell}\} \leq 2, \\ \delta^{-2s} & = & \exp\{-2s\log(1-\epsilon)\} \leq \exp\{4\epsilon s\} = \exp\{\frac{2\alpha}{\ell}\} \leq 1 + \exp\{\frac{1}{2}\}\frac{2\alpha}{\ell} \leq 1 + \frac{4\alpha}{\ell}. \end{array} \tag{4.24}$$

When invoking (4.20) and utilizing (4.24) and (4.23) we get

$$\begin{aligned} \frac{1}{2\pi}\oint_{|z|=1} |q^\ell(z)|^2|dz| &\leq \delta^{-2s} + 2\delta^{-s}R + R^2 \\ &\leq 1 + 4\frac{\alpha}{\ell} + 4R + \frac{1}{4}R \\ &\leq 1 + 10\frac{\alpha}{\ell}. \end{aligned}$$

On the other hand, denoting by q_0, q_1,...,$q_{\ell+s}$ the coefficients of the polynomial q^ℓ and taking into account that $\bar{q}_0 = q^\ell(0) = 1$, we have

$$1 + \sum_{i=1}^{\ell+s} |q_i|^2 = |q_0|^2 + ... + |q_{\ell+s}|^2 = \frac{1}{2\pi} \oint_{|z|=1} |q^\ell(z)|^2 |dz| \leq 1 + 10\frac{\alpha}{\ell}. \tag{4.25}$$

We are done: when denoting $n = \ell + s$, and $q(z) = \sum_{i=1}^n q_j z^j$, we have the vector of coefficients $q = [0; q_1; ...; q_n] \in \mathbb{C}^{n+1}$ of $q(z)$ such that, by (4.25),

$$\|q\|_2^2 \leq \frac{40s(s+2)\log[8s(n-s)]}{n-s},$$

and such that the polynomial $q^\ell(z) = 1 + q(z)$ is divisible by $p(z)$ due to (4.18). □

Proof of Proposition 4.2.3. As in the proof of Proposition 4.1.3, consider the projector $\Pi_{\mathcal{S}_m}$ onto the subspace $\mathcal{S}_m$ (the restriction of $\mathcal{S}$ to coordinates $0, ..., m$), but now let $\phi^o \in \mathbb{C}_m^+(\mathbb{Z})$ correspond to the *last* row of $\Pi_{\mathcal{S}_m}$. As in the proof of Proposition 4.1.3, we see that $x_t = [\phi^o * x]_t$ for any $t \in \mathbb{Z}$, and it remains to bound $\|\phi^o\|_2$. Note that the premise of the proposition is in fact equivalent to the assumption that $\mathcal{S}_m$ is spanned by the vectors

$$\left\{ v(\omega) : [v(\omega)]_t = \frac{e^{i\omega_k t}}{\sqrt{m+1}}, \quad t \in \mathrm{D}_m^+ \right\}, \quad \omega \in \{\omega_1, ..., \omega_s\}.$$

Hence, the projector $\Pi_{\mathcal{S}_m}$ can be written as

$$\Pi_{\mathcal{S}_m} = V \left(V^{\mathrm{H}} V\right)^{-1} V^{\mathrm{H}},$$

where V is an $(m+1) \times s$ Vandermonde matrix with columns $v(\omega_k)$, $k = 1, ..., s$. Note that since $s \leq m+1$, and ω_k, $k = 1, ..., s$ are distinct, matrix V has full column rank.

Now, in order to bound $\|\phi^o\|_2$ from above, it suffices to separate $\lambda_{\min}(V^{\mathrm{H}}V)$, the minimal eigenvalue of $V^{\mathrm{H}}V$, from zero. Indeed, suppose that $\lambda_{\min}(V^{\mathrm{H}}V) > 0$, and write

$$\Pi_{\mathcal{S}_m} = UU^{\mathrm{H}},$$

where $U = [U_1 \cdots U_s]$ is the unitary normalization of V:

$$U = [U_1 \cdots U_s] = V(V^{\mathrm{H}}V)^{-1/2}, \quad U^{\mathrm{H}}U = I_s.$$

Let $u = [u_1, ..., u_s]$ be the last row of U, and v that of V. One has $\phi = uU^{\mathrm{H}} = \sum_{k=1}^s u_k [U_k]^{\mathrm{H}}$, and hence, $\|\phi^o\|_2^2 = \|u\|_2^2$. On the other hand, writing $u = v(V^{\mathrm{H}}V)^{-1/2}$, we arrive at

$$\|u\|_2^2 \leq \frac{\|v\|_2^2}{\lambda_{\min}(V^{\mathrm{H}}V)} \leq \frac{s}{(m+1)\lambda_{\min}(V^{\mathrm{H}}V)},$$

the last transition being due to the bound $\frac{1}{\sqrt{m+1}}$ on the absolute values of the elements of V. Finally, let us exploit the bound on the condition number of a Vandermonde matrix (see [Moi15]):

Lemma 4.3.1 (Theorem 2.3 in [Moi15]). *For $\delta_{\min}$ given by (4.9), we have*

$$\frac{\lambda_{\max}(V^{\mathrm{H}}V)}{\lambda_{\min}(V^{\mathrm{H}}V)} \le \left(m - \frac{2\pi}{\delta_{\min}}\right)^{-1}\left(m + \frac{2\pi}{\delta_{\min}}\right).$$

We clearly have $\|V\|_{\mathrm{op}} \ge 1$, and hence $\lambda_{\max}(V^{\mathrm{H}}V) \ge 1$. Together with (4.10), this results in

$$\frac{1}{\lambda_{\min}(V^{\mathrm{H}}V)} \le \frac{\nu+1}{\nu-1},$$

whence the necessary bound on $\|\phi^o\|_2$ follows. □

Proof of the bound (4.14) for the "composite" estimator from Section 4.2.2. Recall that

$$c_1 s\log(N+1) \le \frac{M+1}{N-M+1} \le c_2 s\log(N+1). \tag{4.26}$$

W.l.o.g. assume that $\sigma = 1$, N and M are even, $s \ge 3$, and $C_1 \ge 1$; as a result, $M \ge N/2$. Recall also the partition of the domain D_N into subdomains D_M, D^+, D^-, and let $\widehat{\varphi}$, $\widehat{\varphi}^+$, $\widehat{\varphi}^-$ be the corresponding adaptive filters for these subdomains:

- $\widehat{\varphi} \in \mathbb{C}_{N-M}(\mathbb{Z})$ is an optimal solution to (**Con-LS**) with $m = N-M$, $n = M$, and
$$\bar{\varrho} = 2(\sqrt{2s})^2 = 4s;$$
- $\widehat{\varphi}^+ \in \mathbb{C}^+_{M+N}(\mathbb{Z})$ is an optimal solution to (**Con-LS-Pred**) applied to the shifted observations $\Delta^{-N}y$ with $h = 0$, $m = N+M$, $n = N-M$, and
$$\bar{\varrho} = \bar{\varrho}^+ := 2C^2s^2\log((2N+1)s+1),$$
C being the constant in (4.8);
- $\widehat{\varphi}^- \in \mathbb{C}^-_{M+N}(\mathbb{Z})$ is an optimal solution to (**Con-LS-Pred**) applied to $\Delta^M y$ with $h = -(M+N)$ and the same m, n, and $\bar{\varrho}$ as in the previous case.

Correspondingly, let $\widehat{x}$ be defined pointwise as

$$\widehat{x}_t = \begin{cases} [\widehat{\varphi} * y]_t, & t \in \mathrm{D}_M, \\ [\widehat{\varphi}^+ * y]_t, & t \in \mathrm{D}^+, \\ [\widehat{\varphi}^- * y]_t, & t \in \mathrm{D}^-. \end{cases}$$

1^o. From Proposition 4.2.2 along with (3.7)–(3.10), we obtain the existence of $\varphi^+ \in \mathbb{C}^+_{M+N}(\mathbb{Z})$ which satisfies $x_t - [\varphi^+ * x]_t = 0$ for any $t \in \mathrm{D}^+$, and

$$\|\varphi^+\|^{\mathrm{F}+}_{M+N,1} \le \frac{\bar{\varrho}^+}{\sqrt{M+N+1}} \le \frac{C_1 s^2\log(N+1)}{\sqrt{M+N+1}},$$

implying that

$$\mathbb{E}\left\|[x - \varphi^+ * y]_M^N\right\|_2^2 \le C_2\left(\frac{N-M+1}{N+M+1}\right)s^4\log^2(N+1).$$

Let $\widehat{\varphi}^{+}$ be as defined above, and denote

$$\kappa_{+} := \sqrt{\frac{N-M+1}{N+M+1}}.$$

Then, using that $\kappa_{+} \leq 1$, Proposition 3.2.2 implies

$$\mathbb{E}\left\|[x - \widehat{\varphi}^{+} * y]_{M}^{N}\right\|_{2}^{2} \leq C_3 \log^2(N)\left(\kappa_{+}^2 s^4 + s^2\right).$$

We can repeat this argument almost verbatim for $\widehat{\varphi}^{-}$, arriving at

$$\mathbb{E}\left\|[x - \widehat{x}]_{M}^{N}\right\|_{2}^{2} + \mathbb{E}\left\|[x - \widehat{x}]_{-N}^{-M}\right\|_{2}^{2} \leq 2C_3 \log^2(N)\left(\kappa_{+}^2 s^4 + s^2\right). \tag{4.27}$$

2^o. Similarly, as follows from Proposition 4.1.3, there exists a filter $\varphi \in \mathbb{C}_{N-M}(\mathbb{Z})$ such that $x_t - [\varphi * x]_t = 0$ for any $t \in \mathrm{D}_M$, and

$$\|\varphi\|_{N-M,1}^{\mathrm{F}} \leq \frac{4s}{\sqrt{2N-2M+1}}.$$

Let $\widehat{\varphi}$ be as defined above, and denote

$$\kappa := \sqrt{\frac{M+1}{N-M+1}}.$$

Proceeding as in **1°** but this time using Theorem 3.2.1, we obtain

$$\mathbb{E}\|x - \widehat{x}\|_{M,2}^2 \leq C_4\left(s^2\kappa^2 + s\left(1+\kappa^2\right)\log N\right) \leq C_5\kappa^2\left(s^2 + s\log N\right), \tag{4.28}$$

where the last transition is due to $\kappa^2 \geq 1$.

3^o. It remains to combine (4.27) and (4.28). Doing so, and using that $M+1 \geq c(N+M+1)$, we arrive at

$$\mathbb{E}\|x - \widehat{x}\|_{N,2}^2 \leq C's^2\log^2(N) + C''\left(\frac{M+1}{N-M+1}\left(s^2 + s\log N\right) + \frac{N-M+1}{M+1}s^4\log^2 N\right).$$

The choice of M according to (4.26) minimizes the right-hand side, and we obtain (4.14). □

Chapter 5

Algorithmic implementation and complexity analysis

Recall that all adaptive estimators introduced in the previous chapters have common form $\widehat{x} = \widehat{\varphi} * y$ where $*$ is the discrete convolution operator, and filter $\widehat{\varphi}$ is an optimal solution of a certain convex optimization problem. Since the discrete convolution can be performed in quasi-linear time via the Fast Fourier Transform (FFT) algorithm, computation of these estimators is reduced to solving the underlying optimization problems. These problems belong to the general class of second-order cone problems (SOCPs), and hence can in principle be solved to high numerical accuracy in polynomial time via interior-point methods [BTN01]. However, the computational complexity of interior-point methods grows polynomially with the problem dimension, and quickly becomes prohibitive when facing 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 in the statistical context. Rather, the desired level of *optimzation accuracy* should be adjusted to the *statistical accuracy* of the exact estimator.

On the other hand, these optimization problems share some key 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 Fast Fourier Transform (FFT).

- *Favorable geometry.* After a straightforward reparametrization, the feasible set of the optimization problems corresponding to the constrained estimators becomes a ball of the complex ℓ_1-norm. Proximal mappings, see e.g. [Nes13b], for such problems with respect to both the Euclidean and non-Euclidean proximal functions with "suitable geometry", see [JN11a] and [JN11b], are easily computable. The same is true for the composite proximal mappings [BT09] corresponding to the penalized recoveries.

- *Medium accuracy is sufficient.* Although the statistical bounds of Chapters 2 and 3 were formulated for the exact solutions of the optimization problems, one can actually use approximate solutions of medium accuracy without losing much from the statistical side.

Altogether, these factors make first-order proximal algorithms the tools of choice when dealing with these problems. This led us to develop a unified framework incorporating the algorithmic implementation of the adaptive convolution-type estimators, as well as the complexity analysis of the resulting algorithmic routines.

Outline. In Section 5.1, we recall two general classes of optimization problems, *composite minimization* problems [BT09, NN13] and *composite saddle-point* problems [JN11b, NN13], and the corresponding algorithms suitable for their numerical solution. In Section 5.2, we present efficient first-order optimization algorithms based on these general approaches. In particular, we show how to compute first-order oracles in quasi-linear time in the signal length using FFT. In Section 5.3, we establish worst-case complexity bounds for the proposed algorithms that explicitly depend on the statistical quantities controlling the statistical difficulty of the problem: the signal length n, the noise variance σ^2, and the filter norm parameter ϱ. These bounds imply that

$$\tilde{O}(\mathrm{PSNR}_* + 1)$$

iterations of a suitable first-order algorithm are sufficient to match the statistical properties of an exact estimator; here, $\mathrm{PSNR}_* = \frac{1}{\sigma}\|F_{2n}[x]_{-n}^{n}\|_\infty$ is the peak signal-to-noise ratio in the Fourier domain. This gives a sharp and rigorous characterization (in the present context) of the performance of "early stopping" strategies that allow to stop an optimization algorithm much earlier than dictated by the pure optimization analysis. In Section 5.4, we conduct numerical evaluation of the proposed algorithms on simulated data to complement our theoretical analysis.

Ad-hoc notation. We use the following ad-hoc notation in this chapter. We assume throughout that the observations y_τ are given on $\{-n, ..., n\}$, and the goal is to fit a *one-sided* filter $\widehat{\varphi} \in \mathbb{C}_n^+(\mathbb{Z})$ (all results obtained in this chapter can be extended to the general situation in a straightforward manner). First of all, for brevity we introduce the *residual*

$$\mathrm{Res}_p(\varphi) := \|F_n[y - \varphi * y]\|_p, \quad p \in \{2, \infty\}. \tag{5.1}$$

We *rescale* ℓ_p-seminorms on $\mathbb{C}(\mathbb{Z})$:

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

and introduce the constraint set

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

We define the one-sided unitary Discrete Fourier Transform (DFT) operator F_n on $\mathbb{C}^{n+1}$ by

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

omitting the "+" superscript. Note that theinverse operator F_n^{-1} coincides with the conjugate transpose F_n^{H}. Abusing the notation, we occasionally shorten $F_n[x]_0^n$ to $F_n[x]$. In other words, $F_n[\cdot]$ is a map $\mathbb{C}_n^+(\mathbb{Z}) \to \mathbb{C}^{n+1}$; the adjoint map $F_n^{\mathrm{H}}[x]$ sends $F_n^{\mathrm{H}}[x]_0^n$ to $\mathbb{C}_n^+(\mathbb{Z})$ via zero-padding.

Estimators. For the simplicity of reference, we now recollect all estimators of interest. Note that we adopt a different parametrization of the penalization terms which will be used hereinafter.

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \Phi_n(\bar{\varrho})} \operatorname{Res}_\infty(\varphi). \tag{\textbf{Con-UF}}$$

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

$$(\widehat{\varphi}, \widehat{\varrho}) \in \operatorname*{Argmin}_{\substack{\varrho \geq 0 \\ \varphi \in \mathbb{C}_n^+(\mathbb{Z})}} \left\{ \varrho : \begin{array}{l} \|F_n[\varphi]\|_1 \leq \frac{\varrho}{\sqrt{n+1}}, \\ \operatorname{Res}_\infty(\varphi) \leq \sigma(\bar{\theta}\varrho + \bar{\Theta}(1+\varrho)) \end{array} \right\} \tag{\textbf{Epi-UF}}$$

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

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

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

5.1 Tools from convex optimization

In this section, we introduce the tools from the theory of first-order algorithms to be used later. We describe two general type of optimization problems, composite minimization and composite saddle-point problems, together with efficient first-order algorithms for their solution. Following [NN13], we now introduce the concept of *proximal setup* which underlies these algorithms.

5.1.1 Proximal setup

Let the *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 a Euclidean one, and a *distance-generating function* (d.-g.f.),

$$\omega(u) : U \to \mathbb{R},$$

satisfying the following properties:

- $\omega(u)$ is a continuous convex function on U;
- $\omega(u)$ admits a continuous selection $\omega'(u) \in \partial\omega(u)$ of subgradients on the set

$$U^o = \{u \in U : \partial\omega(u) \neq \emptyset\};$$

- $\omega(u)$ is strongly convex with modulus 1 with respect to $\|\cdot\|$:

$$\langle \omega'(u_1) - \omega'(u_2), u_1 - u_2 \rangle \geq \|u_1 - u_2\|^2 \quad \forall u_1, u_2 \in U^o.$$

The concept of proximal setup gives rise to several important notions:

- *ω-center* $u_\omega := \operatorname{argmin}_{u \in U} \omega(u) \in U^o$;

- *Bregman distance* $D_{u^o}(u)$, defined for $u^o \in U^o$ and $u \in U$ by the relation

$$D_{u^o}(u) = \omega(u) - \omega(u^o) - \langle \omega'(u^o), u - u^o \rangle;$$

  Note that $D_{u^o}(u) \geq \frac{1}{2}\|u - u^o\|^2$ due to the strong convexity of $\omega(\cdot)$;

- *prox-mapping* $\mathrm{Prox}_{u^o}(g) : E \to U^o$, parametrized by the *prox-center* $u^o \in U^o$, defined by

$$\mathrm{Prox}_{u^o}(g) = \operatorname*{argmin}_{u \in U} \{\langle g, \xi \rangle + D_{u^o}(u)\};$$

- *ω-radius:* let $\tilde{U}$ be a subset of U containing u_ω, then the *ω-radius* of $\tilde{U}$ is defined as

$$\Omega[\tilde{U}] := \sqrt{2\left(\max_{u \in \tilde{U}} \omega(u) - \min_{u \in U} \omega(u)\right)};$$

  the name stemming from the fact that for any $u \in \tilde{U}$, $\|u - u_\omega\| \leq \sqrt{2D_{u_\omega}(u)} \leq \Omega[\tilde{U}]$.

Blockwise proximal setups. We now describe a specific family of proximal setups which proves to be useful in our situation. 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 $\mathrm{Vec}_n : \mathbb{C}^{n+1} \to \mathbb{R}^{2(n+1)}$,

$$\mathrm{Vec}_n z = [\Re(z_0); \Im(z_0); ...; \Re(z_n); \Im(z_n)]. \tag{5.2}$$

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:

$$\|u\| := \sum_{j=0}^{m} \|u^j\|_2. \tag{5.3}$$

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

Theorem 5.1.1 ([NN13], Theorem 2.1 and Corollary 2.2)**.** *Given $E = \mathbb{R}^N$ as above, $\omega : \mathbb{R}^N \to \mathbb{R}$ defined by*

$$\omega(u) = \frac{(m+1)^{(\tilde{q}-1)(2-\tilde{q})/\tilde{q}}}{2\tilde{c}} \left[\sum_{j=0}^{m} \|u^j\|_2^{\tilde{q}}\right]^{2/\tilde{q}} \tag{5.4}$$

with parameters

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

is a d.-g. f. for any ball $U_N(R)$ of the norm (5.3) 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

$$\Omega[U_N(R)] \leq C(\sqrt{\log(m+1)} + 1)R. \tag{5.5}$$

We will use two particular cases of the construction in the premise of the above theorem.

1. Case $m = n$, $k = 2$ corresponds to the ℓ_1-norm on $\mathbb{C}^{n+1}$, and specifies the *complex ℓ_1-setup*.
2. Case $m = 0$, $k = N$ corresponds to the ℓ_2-norm on $\mathbb{C}^{n+1}$, and specifies the *ℓ_2-setup* $(\|\cdot\|_2, \frac{1}{2}\|\cdot\|_2^2)$.

For what is to follow, it is convenient to introduce the following norms on $\mathbb{R}^N$:

$$\|u\|_{\mathbb{C},p} := \|\mathrm{Vec}_n^{-1} u\|_p = \|\mathrm{Vec}_n^{\mathrm{H}} u\|_p, \quad p \geq 1. \tag{5.6}$$

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 $\mathbb{R}^N$.

5.1.2 Composite minimization

In the general problem of composite minimization, the goal is to solve the convex program

$$\min_{u \in U} \{\phi(u) = f(u) + \Psi(u)\}. \tag{5.7}$$

Here, U is a domain in E equipped with the norm $\|\cdot\|$, function $f(u)$ is convex and continuously differentiable on U, and $\Psi(x)$ 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 comprising the norm $\|\cdot\|$ and a d.-g. f. $\omega(\cdot)$ for U, let us define the *composite prox-mapping*

$$\mathrm{Prox}_{\Psi,u}(g) = \operatorname*{argmin}_{\xi \in U} \{\langle g, \xi\rangle + D_u(\xi) + \Psi(\xi)\}. \tag{5.8}$$

It is reasonable to assume that this operator is computable in near-linear time in the problem dimension which will be found to be $O(n)$. This is indeed the case when $\omega(\cdot)$ and the non-smooth part of the objective $\Psi(\cdot)$ are "quasi-separable" functions (such as the powers of the block ℓ_1/ℓ_2-norm), and Z has a structure "well-adapted" to both $\omega(\cdot)$ and $\Psi(\cdot)$. As we will see further, this is indeed the case for the problems of interest.

Fast Gradient Method. This algorithm, first proposed in [Nes13a], extends celebrated Nesterov's algorithm for smooth minimization [Nes83] and its composite version [BT09] to general proximal setups. We summarize it as Algorithm 2. One iteration of this algorithm incorporates a constant number of evaluations of the gradient of $f(\cdot)$, elementwise vector operations, and computations of the composite prox-mapping. As we will see later on, for general problems of the form (5.7), Algorithm 2 admits an $O(1/T^2)$ accuracy bound after T iterations. In addition, we present in Section 5.A a version of Algorithm 2 with adaptive stepsize selection. This algorithm admits the same theoretical guarantee as Algorithm 2, cf. Section 5.1.4.

5.1.3 Composite saddle-point problems

We also consider another class of optimization problems:

$$\inf_{u \in U} \max_{v \in V} [\phi(u, v) = f(u, v) + \Psi(u)]. \tag{5.9}$$

Here, $U \subset E_u$ and $V \subset E_v$ are convex subsets of the corresponding Euclidean spaces, and V is compact; function $f(u, v)$ is convex in u, concave in v, and differentiable on $W := U \times V$;

Algorithm 2 : `Fast Gradient Method`

Input: stepsize $\eta > 0$

$u^0 = u_\omega$

$g^0 = 0 \in E$

for $t = 0, 1, ...$ **do**

  $u_t = \mathrm{Prox}_{\eta\Psi, u_\omega}\left(\eta g^t\right)$

  $\tau_t = \frac{2(t+2)}{(t+1)(t+4)}$

  $u_{t+\frac{1}{3}} = \tau_t u_t + (1-\tau_t)u^t$

  $g_t = \frac{t+2}{2}\nabla f(u_{t+\frac{1}{3}})$

  $u_{t+\frac{2}{3}} = \mathrm{Prox}_{\eta\Psi, u_t}\left(\eta g_t\right)$

  $u^{t+1} = \tau_t u_{t+\frac{2}{3}} + (1-\tau_t)u^t$

  $g^{t+1} = \sum_{\tau=0}^{t} g_t$

end for

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 $F : W \to E = E_u \times E_v$,

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

Saddle-point problem (5.9) specifies two convex optimization problems: minimization of $\overline{\phi}(u) = \max_{v\in V} \phi(u, v)$, or the primal problem, and maximization of $\underline{\phi}(v) = \inf_{u\in U} \phi(u, v)$, the dual problem. Under some general conditions which hold in the described setting [Sio58], (5.9) possesses an optimal solution $w^* = [u^*; v^*]$, called a *saddle point*, such that the value of (5.9) 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 has to construct a 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. Folllowing [NN13], let us assume that both the dual ω-radius $\Omega[V]$ and the "effective primal radius"

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

Algorithm 3 : `Composite Mirror Prox`

Input: stepsize $\eta > 0$

1: $w_0 := [u_0; v_0] = w_\omega$

2: **for** $t = 0, 1, ...$ **do**

3: $\quad w_{t+\frac{1}{2}} = \mathrm{Prox}_{\eta\Psi, w_t}(\eta F(w_t))$

4: $\quad w_{t+1} = \mathrm{Prox}_{\eta\Psi, w_t}(\eta F(w_{t+\frac{1}{2}}))$

5: $\quad w^{t+1} := [u^{t+1}; v^{t+1}] = \frac{1}{t+1}\sum_{\tau=0}^{t} w_\tau$

6: **end for**

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

$$\begin{aligned} \|w\|^2 &= \Omega^2[V]\,\|u\|_U^2 + \Omega_*^2[U]\,\|v\|_V^2, \\ \omega(w) &= \Omega^2[V]\,\omega_U(u) + \Omega_*^2[U]\,\omega_V(v). \end{aligned} \tag{5.10}$$

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

Composite Mirror Prox. This algorithm, introduced in [NN13] and summarized here as Algorithm 3, solves the general composite saddle-point problem (5.9). As we are about to see, when applied with joint proximal setup (5.10), Algorithm 3 admits an $O(1/T)$ accuracy bound. Besides, in Section 5.A we present a simple adaptive stepsize selection rule which allows to significantly accelerate Algorithm 3 in practice while still preserving its worst-case accuracy estimate as given in Section 5.1.4.

5.1.4 General accuracy bounds

Here we recall the worst-case bounds on the *absolute accuracy* of the objective after T iterations,

$$\varepsilon(T) := \phi(u^T) - \phi(u^*).$$

(For saddle-point problems, ϕ must be replaced with $\overline{\phi}$, and $\varepsilon(T)$ is upper-bounded by the duality gap.) These bounds are applicable when solving an *arbitrary* optimization problem in one of two previously described general classes with the corresponding first-order algorithm. They are expressed in terms of the "optimization" parameters that specify as the regularity of the objective and the radius of the feasible set.

Theorem 5.1.2 ([NN13, Theorem 2.14])**.** *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 2 with

stepsize $\eta = \frac{1}{L_f}$. *Then,*

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

Theorem 5.1.3 ([NN13], Theorem 3.3)**.** *Let* $f(u,v)$ *be as in* (5.14)[1]*, 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 3 with joint setup* (5.10) *and stepsize* $\eta = \frac{\Omega[V]}{\Omega_*[U] L_F}$. *Then,*

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

5.2 Efficient algorithmic implementation

We now present efficient implementation of the algorithms described in the previous section as applied to the optimization problems corresponding to the adaptive convolution-type estimators.

Change of variables. Our first step is to pass to the Fourier domain, so that, first, convolution is represented as an efficiently computable linear operator, and second, the feasible set and the penalization term become quasi-separable in the new variables. Noting that the adjoint map of $\mathrm{Vec}_n : \mathbb{C}^{n+1} \to \mathbb{R}^{2n+2}$, cf. (5.2), is given by

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

consider the transformation

$$u = \mathrm{Vec}_n F_n[\varphi] \quad b = \mathrm{Vec}_n F_n[y] \tag{5.11}$$

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

$$Au = \mathrm{Vec}_n F_n \left[y * F_n^{\mathrm{H}}[\mathrm{Vec}_n^{\mathrm{H}} u]\right]. \tag{5.12}$$

We are about to demonstrate that all recovery procedures can be recast in one of the "canonical" forms (5.7), (5.9). Moreover, the gradient computation is then reduced to evaluating the convolution-type operator A and its adjoint $A^{\mathrm{H}} = A^{\mathrm{T}}$.

[1] For simplicity, we only state the bound for bilinear $f(u,v)$. The general case can be addressed similarly but requires a somewhat more cumbersome notation.

Problem reformulation. After the change of variables (5.11), problems (**Con-LS**), (**Pen-LS**), and (**Pen^2-LS**) take form (5.7):

$$\min_{\|u\|_{\mathbb{C},1}\le R} \left[f(u) := \tfrac{1}{2}\|Au-b\|_2^2\right] + \lambda\|u\|_{\mathbb{C},1}^r, \quad r\in\{1,2\}, \tag{5.13}$$

where $\|\cdot\|_{\mathbb{C},p}$ is defined in (5.6). In particular, (**Con-LS**) is obtained from (5.13) by setting $\lambda = 0$ and $R = \frac{\overline{\varrho}}{\sqrt{n+1}}$, and (**Pen-LS**) is obtained by setting $R = \infty$ and $r = 1$. Note that

$$\nabla f(u) = A^{\mathrm{H}}(Au-b),$$

and that $A^{\mathrm{H}} = A^{\mathrm{T}}$.

On the other hand, problems (**Con-UF**), (**Pen-UF**), and (**Con-LS**) (after taking the square root of the objective in the last case) can be recast as saddle-point problems (5.9) using that the dual norm of $\|\cdot\|_{\mathbb{C},p}$ 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 a result, (**Con-UF**), (**Pen-UF**) and (**Con-LS**) are reduced to

$$\min_{\|u\|_{\mathbb{C},1}\le R} \max_{\|v\|_{\mathbb{C},q}\le 1} [f(u,v) := \langle v, Au-b\rangle] + \lambda\|u\|_{\mathbb{C},1}, \quad q\in\{1,2\}, \tag{5.14}$$

where $q = 1$ for (**Con-UF**) and (**Pen-UF**), and $q = 2$ for (**Con-LS**) after taking square root. Note that $f(u,v)$ is bilinear, and

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

Finally, note that solving (**Epi-UF**) amounts to solving a small series of problems of the type (**Con-UF**). Indeed, for $\varrho \ge 0$ let us define $\mathrm{Opt}(\varrho)$ to be the optimal value of (**Con-UF**) with a given ϱ. Clearly, $\mathrm{Opt}(\varrho)$ is a non-increasing function on $\mathbb{R}^+$, and evaluating it at a point amounts to solving an instance of (**Con-UF**). On the other hand, the optimal value of (**Epi-UF**) can be alternatively expressed as a unique optimal solution of a one-dimensional problem

$$\min_{\varrho\ge 0} \left\{\mathrm{Opt}(\varrho) : \mathrm{Opt}(\varrho) \le \sigma(\overline{\theta}+\overline{\Theta})\varrho + \sigma\overline{\Theta}\right\}.$$

In practice, this problem can be solved by bijection, leading to linear convergence in terms of the number of evaluations of $\mathrm{Opt}(\cdot)$. Moreover, [NN13] describes an even more efficient approach based on Newton's method which takes into account additional structural properties of $\mathrm{Opt}(\cdot)$ and has guaranteed quadratic convergence.

To conclude, we see that the computation of all adaptive convolution-type estimators is reduced to at least one of the two general problem types introduced in Section 5.1. We are now in the position to apply the algorithms described in Section 5.1. One iteration of either of them is reduced to a constant number of computations of the gradient (which, in turn, is reduced to evaluating A and A^{H}) and prox-mappings. Next we derive the convolution-type operator A and its adjoint A^{H} as a product of operators that can be evaluated in $O(n\log n)$.

Evaluation of A and A^{H}. Operator A, cf. (5.12), 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

$$\psi = [[\varphi]_0^n; 0_n] \in \mathbb{C}^{2n+1}.$$

Via the DFT diagonalization property, this is expressed as

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

where operator $D_y = \mathrm{diag}(F_{2n}[y]_{-n}^n)$ on $\mathbb{C}^{2n+1}$ can be constructed in $\tilde{O}(n)$ 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 $\mathbb{C}^{2n+1}$; note that its adjoint P_n^{H} is the zero-padding operator which complements $[\varphi]_0^n$ with n trailing zeroes. Then,

$$Au = \sqrt{2n+1}\cdot \mathrm{Vec}_n F_n P_n F_{2n}^{\mathrm{H}} D_y F_{2n} P_n^{\mathrm{H}} F_n^{\mathrm{H}} \mathrm{Vec}_n^{\mathrm{H}} u, \tag{5.15}$$

where all operators in the right-hand side can be evaluated in $O(n\log n)$. Moreover, we arrive at a similar representation for A^{H} by formally taking the adjoint of (5.15).

5.2.1 Computation of prox-mappings

In fact, the composite prox-mappings corresponding to both basic proximal setups introduced in Section 5.1 can be computed in $O(n)$. In the penalized case with non-squared penalty function, this computation can be done explicitly, while in the remaining cases the prox-mappings can be computed via a root-finding algorithm. Below we describe these calculations since they are relevant in the general context of proximal algorithms applied to signal processing problems.

It suffices to consider partial proximal setups separately; the case of joint setup in saddle-point problems can then be treated using that the joint prox-mapping is separable in u and v, cf. Section 5.1.3. Recall that the possible partial setups $(\|\cdot\|, \omega(\cdot))$ comprise the ℓ_2-setup with $\|\cdot\| = \|\cdot\|_{\mathbb{C},2} = \|\cdot\|_2$ and the (complex) ℓ_1-setup with $\|\cdot\| = \|\cdot\|_{\mathbb{C},1}$; in both cases, $\omega(\cdot)$ is given by (5.4). Computing $\mathrm{Prox}_{\frac{1}{L}\Psi,u}(g)$, cf. (5.8), amounts to solving

$$\min_{\xi\in\mathbb{R}^N} \left\{\xi^{\mathrm{T}}(g-\omega'(u)) + \omega(\xi) : \|\xi\|_{\mathbb{C},q} \le R\right\}, \quad q \in \{1,2\} \tag{5.16}$$

in the constrained case, and

$$\min_{\xi\in\mathbb{R}^N} \left\{\xi^{\mathrm{T}}(g-\omega'(u)) + \omega(\xi) + \frac{\lambda}{L}\|\xi\|_{\mathbb{C},1}^q\right\}, \quad q \in \{1,2\} \tag{5.17}$$

in the penalized case. In the constrained case with ℓ_2-setup, the task is reduced to the Euclidean projection onto the ℓ_2-ball if $q=2$, and onto the ℓ_1-ball if $q=1$; the latter can be done (exactly) in $\tilde{O}(N)$ via the algorithm from [DSSSC08] (for that, one first solves (5.16) for the complex phases corresponding to the pairs of components of ξ). The constrained case with ℓ_1-setup is reduced to the penalized case by passing to the Langrangian dual problem. Evaluation of the dual function amounts to solving a problem equivalent to (5.17) with $q=1$, whence (5.16) can be solved by a simple root-finding procedure if one is able to solve (5.17). As for (5.17), below we show how to solve it explicitly when $q=1$, and reduce it to one-dimensional root search, so

that it can be solved in $O(n)$ to numerical tolerance, when $q = 2$. Indeed, (5.17) can be recast in terms of the complex variable $\zeta = \mathrm{Vec}_n^{\mathrm{H}}\xi$:

$$\min_{\zeta\in\mathbb{C}^{n+1}} \left\{ \langle \zeta, z\rangle + \underline{\omega}(\zeta) + \frac{\lambda}{L}\|\zeta\|_1^q \right\}, \tag{5.18}$$

where $z = \mathrm{Vec}_n^{\mathrm{H}}(g - \omega'(u))$, and $\underline{\omega}(\zeta) = \omega(\xi)$, cf. (5.4), whence

$$\underline{\omega}(\zeta) = \frac{C(m,\tilde{q},\tilde{\gamma})\|\zeta\|_{\tilde{q}}^2}{2}, \tag{5.19}$$

with $C(m,\tilde{q},\tilde{\gamma}) = \frac{1}{\tilde{\gamma}}(m+1)^{(\tilde{q}-1)(2-\tilde{q})/\tilde{q}}$. Now, (5.18) can be minimized first with respect to the complex arguments, and then to the absolute values of the components of ζ. Denoting ζ^* a (unique) optimal solution of (5.18), the first minimization results in $\zeta_j^* = -\frac{z_j}{|z_j|}|\zeta_j^*|$, $0 \le j \le n$, and it remains to compute the absolute values $|\zeta_j^*|$.

Case $q = 1$. The first-order optimality condition implies

$$C(m,\tilde{q},\tilde{\gamma})\|\zeta^*\|_{\tilde{q}}^{2-\tilde{q}}|\zeta_j^*|^{\tilde{q}-1} + \frac{\lambda}{L}\mathbb{1}\{|\zeta_j^*| > 0\} = |z_j|. \tag{5.20}$$

Denoting $\tilde{p} = \frac{\tilde{q}}{\tilde{q}-1}$, and using the soft-thresholding operator

$$\mathrm{Soft}_M(x) = (|x| - M)_+ \,\mathrm{sign}(x),$$

we obtain the explicit solution:

$$\zeta_j^* = \frac{1}{C(m,\tilde{q},\tilde{\gamma})}\left(\frac{\theta_j}{\|\theta\|_{\tilde{p}}^{2-\tilde{q}}}\right)^{\tilde{p}/\tilde{q}}, \quad \theta_j = \mathrm{Soft}_{\lambda/L}(z_j).$$

In the case of ℓ_2-setup this reduces to $\zeta_j^* = \mathrm{Soft}_{\lambda/L}(z_j)$.

Case $q = 2$. Instead of (5.20), we arrive at

$$C(m,\tilde{q},\tilde{\gamma})\|\zeta^*\|_{\tilde{q}}^{2-\tilde{q}}|\zeta_j^*|^{\tilde{q}-1} + \frac{2\lambda\|\zeta^*\|_1}{L}\mathbb{1}\{|\zeta_j^*| > 0\} = |z_j|, \tag{5.21}$$

which we cannot solve explicitly. However, note that a counterpart of (5.21), in which $\|\zeta^*\|_1$ is replaced with parameter $t \ge 0$, can be solved explicitly similarly to (5.20). Let $\zeta^*(t)$ denote the corresponding solution for a fixed t, which can be obtained in time $O(n)$. Clearly, $\|\zeta^*(t)\|_1$ is a non-decreasing function on $\mathbb{R}^+$. Hence, (5.21) can be solved, up to numerical tolerance, by any one-dimensional root search procedure, in $O(1)$ evaluations of $\zeta^*(t)$.

5.3 Theoretical complexity analysis

Recall that in Section 5.1.4 we presented the worst-case bounds on the *absolute accuracy* of the objective when solving the general composite minimization and saddle-point problems with the

corresponding first-order algorithm, cf. Theorems 5.1.2–5.1.3. These bounds are expressed in terms of the "optimization" parameters that specify as the regularity of the objective and the radius of the feasible set. Our first theoretical contribution, presented in the next section, is put these results into the context of statistical estimation, expressing them via purely "statistical" quantities: the norm of the exact estimator and the Fourier-domain ℓ_∞-norm of the observations.

5.3.1 Bounding the absolute accuracy

First, note that in the case where the partial domain (for u or v) is an $\|\cdot\|_{\mathbb{C},2}$-norm ball, we will use the ℓ_2-setup in that variable (then the domain coincides with $U_N(R)$, cf. (5.5), while if the domain is an $\|\cdot\|_{\mathbb{C},1}$-norm ball, we can choose between ℓ_1 or ℓ_2-setups (in the latter case, the domain is contained in $U_N(R)$). Radii $\Omega[V], \Omega_*[U]$ can then be bounded as follows, cf. (5.5):

$$\Omega[V] = \tilde{O}(1), \quad \Omega_*[U] = \tilde{O}\left(\varrho/\sqrt{n+1}\right), \tag{5.22}$$

where

$$\varrho = \sqrt{n+1}\|F_n[\widehat{\varphi}]\|_1 \tag{5.23}$$

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

Another observation concerns the Lipschitz constants. Let

$$[q_u; q_v] \in \{2,1\} \times \{2,1\}, \tag{5.24}$$

depending on whether one uses the Euclidean setup ($q = 2$) or the complex ℓ_1-setup ($q = 1$) in each variable u, v; besides, let $p_v = \frac{q_v}{q_v - 1} \in \{2, \infty\}$. 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 express the Lipshitz constants L_f, L_F in terms of the subordinate norms $\|\mathcal{A}\|_{\alpha\to\beta} := \sup_{\|\psi\|_\alpha = 1} \|\mathcal{A}\psi\|_\beta$ as follows:

$$\begin{aligned} \|\mathcal{A}\|_{1\to 2}^2 \leq L_f = \|\mathcal{A}\|_{q_u \to 2}^2 \leq \|\mathcal{A}\|_{2\to 2}^2, \\ \|\mathcal{A}\|_{1\to\infty} \leq L_F = \|\mathcal{A}\|_{q_u \to p_v} \leq \|\mathcal{A}\|_{2\to 2}. \end{aligned} \tag{5.25}$$

The operator norm $\|\mathcal{A}\|_{2\to 2}$ corresponds to the partial ℓ_2-setups (in both variables), and the norm $\|\mathcal{A}\|_{1\to\infty}$ to the complex ℓ_1-setups in both variables; note that $\|\mathcal{A}\|_{1\to\infty} \leq \|\mathcal{A}\|_{1\to 2}$. Now, as we prove in Section 5.5, $\|\mathcal{A}\|_{2\to 2}$ itself can be bounded as follows:

$$\|\mathcal{A}\|_{2\to 2} \leq \sqrt{2n+1} \cdot \|F_{2n}[y]_{-n}^n\|_\infty. \tag{5.26}$$

Together with (5.22), relation 5.26 implies the following result:

Proposition 5.3.1. *Solving* (**Con-LS**), (**Pen-LS**), *or* (**Pen^2-LS**) *by Algorithm 2 with proximal setup as described above, one has*

$$\varepsilon(T) = \tilde{O}\left(\frac{\varrho^2 \|F_{2n}[y]_{-n}^n\|_\infty^2}{T^2}\right). \tag{5.27}$$

Similarly, solving (**Con-LS**) *(with square root of the objective),* (**Con-UF**)*, or* (**Pen-UF**) *by Algorithm 3, one has*

$$\varepsilon(T) = \tilde{O}\left(\frac{\varrho\|F_{2n}[y]_{-n}^{n}\|_{\infty}}{T}\right). \tag{5.28}$$

Note that Proposition 5.3.1 gives the same upper bound on the accuracy $\varepsilon(T)$ irrespectively of the proximal setup chosen among the ones described in the premise of (5.22). 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 can be as small as $\|\mathcal{A}\|_{1\to 2}$ or even $\|\mathcal{A}\|_{1\to\infty}$ when one uses the "geometry-adapted" proximal setup in which partial ℓ_1-setups are used for the variables "measured" in $\|\cdot\|_{\mathbb{C},1}$-norm. For a general linear operator $\mathcal{A}$ on $\mathbb{C}^{n+1}$ the gaps between $\|\mathcal{A}\|_{2\to 2}$ and $\|\mathcal{A}\|_{1\to 2}$ or $\|\mathcal{A}\|_{1\to\infty}$ can be as large as $\sqrt{n+1}$ or $n+1$, and hence one might expect Proposition 5.3.1 to be suboptimal for the "geometry-adapted" setup. However, as can be inferred from the representation (5.15), operator $\mathcal{A}$ has a special "almost diagonal" structure, and it is unlikely that its different subordinate norms scale differently with n. This intuition can be made precise:

Proposition 5.3.2. *Assume that $\sigma = 0$, and $x \in \mathbb{C}(\mathbb{Z})$ is $(n+1)$-periodic: $x_\tau = x_{\tau-n-1}$ for $\tau \in \mathbb{Z}$. Then, one has*

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

The proof of this proposition is given in Section 5.5.

5.3.2 Statistical accuracy and algorithmic complexity

In this section, we characterize the *statistical accuracy* of adaptive recovery procedures. It can be informally defined as the accuracy ε_* of minimizing the residual, 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 following result for the pointwise loss of approximate uniform-fit estimators follows easily when recalling the proofs of Propositions 2.2.1 and 2.2.3; its proof, omitted here, can be found in [OH18][2].

Proposition 5.3.3. *An ε_*-accurate solution $\tilde{\varphi}$ to* (**Con-UF**) *or* (**Pen-UF**) *with parameters as specified in the premises of Propositions 2.2.1 and 2.2.3, in both cases with $\varepsilon_* = O(\sigma\varrho)$, with probability $\geq 1-\delta$ satisfies the same bounds on the pointwise loss, cf. Propositions 2.2.1 and 2.2.3, as the exact solution of the corresponding optimization problem (up to a multiplicative constant).*

The next proposition controls the ℓ_2-loss for least-squares estimators (note again a different parametrization of the regularization parameter compared to the previous chapters).

Proposition 5.3.4. *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**)*,* (**Pen-LS**)*, or* (**Pen^2-LS**) *with parameters as specified in the premises of Theorems 3.2.1–3.2.3, in all cases with $\varepsilon_* = O(\sigma^2\varrho^2)$, with probability $\geq 1-\delta$ satisfies the same bounds on the ℓ_2-loss as those for the exact solution of the corresponding optimization problem (up to a multiplicative constant).*

[2] While we do not establish similar result for the epigraph estimator (**Epi-UF**), we conjecture that it must hold whenever the inner computations of (**Con-UF**), cf. Section 5.2, are done with accuracy $\varepsilon_* = O(\sigma\varrho)$ for each ϱ.

As in the case of uniform-fit estimators, the above proposition can be proved by looking to the proof of the corresponding results for the exact estimator and generalizing them by allowing some error in the optimization objective. Since the loss decomposition for least-squares estimators is significantly more involved, in Section 5.5 we give the proof for the constrained estimator.

Complexity bound. A direct consequence of Propositions 5.3.3–5.3.4 and Proposition 5.3.1 is as follows: the number of iterations T_* of a suitable first-order algorithm – Algorithm 2 for the least-squares estimators and Algorithm 3 for the uniform-fit ones – after which the statistical accuracy ε_* is guaranteed, with high probability satisfies the following bound:

$$T_* = \tilde{O}\left(\|F_{2n}[y]_{-n}^{n}\|_\infty/\sigma\right). \tag{5.29}$$

Noting that the ℓ_∞-norm of $F_n[\zeta] \sim \mathbb{C}\mathcal{N}(0, I_{n+1})$ behaves as $C\sqrt{\log n}$ with high probability, the above complexity estimate can be expressed as

$$T_* = \tilde{O}\left(\mathrm{PSNR}_* + 1\right) \tag{5.30}$$

where $\mathrm{PSNR}_* = \|F_{2n}[y]_{-n}^{n}\|_\infty/\sigma$ is the peak signal-to-noise ratio in the Fourier domain. Finally, if the signal is known to be sparse in the Fourier domain, that is, it belongs to a shift-invariant subspace $\mathcal{S}$ spanned by s complex exponentials $e^{i\omega_k \tau}$ with frequencies on the DFT grid, $\omega_k \in \left\{\frac{2\pi j}{n+1},\ j \in \mathbb{Z}\right\}$, we can write

$$\mathrm{PSNR}_* \le \mathrm{SNR}\sqrt{s} \tag{5.31}$$

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

Comparison of Algorithms 2 and 3 for (Con-LS). Note that Algorithm 3 can also be used to solve (**Con-LS**) by working instead with the non-squared residual $\mathrm{Res}_2(\varphi) = \|F_n[y - \varphi * y]\|_2$ and passing to the equivalent saddle-point problem. However, this approach fails to fully capture the structure of the objective of (**Con-LS**), and results in suboptimal performance (see also Section 5.4). In particular, (5.29) does not hold in this case. Indeed, in order to guarantee accuracy $O(\sigma^2\varrho^2)$ in the squared residual, cf. Theorem 5.3.4, it is not sufficient to achieve accuracy $O(\sigma\varrho)$ in the non-squared residual. Rather, sufficient accuracy in the non-squared residual in that case is $O\left(\frac{\sigma^2\varrho^2}{\mathrm{Res}_2(\widehat{\varphi})}\right)$, which can be much less than $\sigma\varrho$ since the optimal residual $\mathrm{Res}_2(\widehat{\varphi})$ can be as large as $\sigma\sqrt{n+1}$ (see the proof of Theorem 3.2.1).

On the other hand, Algorithm 2 does achieve (5.29) thanks to one curious property, *fast* $O(T^{-2})$ *convergence* for (**Con-LS**) with *non-squared residual*. Indeed, one has the following bound for the accuracy of (**Con-LS**) with non-squared residual due to (5.27) and the difference of squares formula:

$$\varepsilon(T) = \tilde{O}\left(\frac{\varrho\|F_{2n}[y]_{-n}^{n}\|_\infty}{T}\right)\min\left(1, \frac{T_{\mathrm{fast}}}{T}\right), \tag{5.32}$$

where

$$T_{\mathrm{fast}} = \frac{\varrho\|F_{2n}[y]_{-n}^{n}\|_\infty}{\mathrm{Res}_2(\widehat{\varphi})}, \tag{5.33}$$

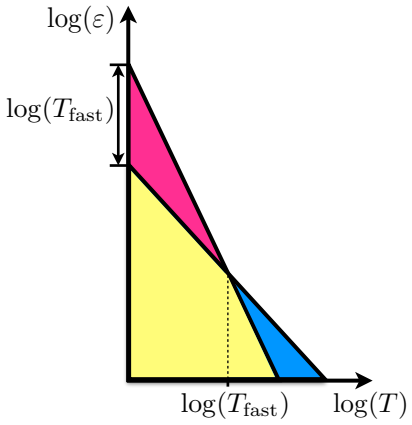

Figure 5-1. "Phase transition" for Algorithm 2. Note the unit and the double slope which correspond to different bounds in (5.32).

see Fig. 5-1. In other words, after T_{fast} iterations, $O(T^{-2})$ convergence activates for Algorithm 2, and it begins to outperform Algorithm 3. Moreover, this effect is instrumental in achieving (5.29) since usually $T_{\text{fast}} \ll T_*$, cf. (5.33) and (5.29). It is an interesting question whether this observation can be exploited in a more general context, for example, when minimizing a finite sum of the ℓ_2-norms of different affine transforms.

5.4 Experiments

In this series of experiments, our goal is to demonstrate the effectiveness of the approach, and illustrate the theoretical results of Section 5.3. We estimate signals coming from an unknown shift-invariant subspace $\mathcal{S}$ using the experimental protocol presented in Chapter 3. First, a random signal $[x_0; ...; x_n]$ with $n = 100$ is generated according to one of the scenarios described below (s and m are parameters of the scenario)[3]. 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 of this process 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.* [TBR13].

In the above scenarios, we use theoretically recommended value $\bar{\varrho} = 2\dim(\mathcal{S})$ as suggested by Proposition 4.1.3 for two-sided filters $\varphi \in \mathbb{C}_n(\mathbb{Z})$ (note that $\dim(\mathcal{S}) = s$ in *Random-s* and $\dim(\mathcal{S}) = 2s$ in *Coherent-s*).

Proof-of-concept experiment. In this experiment, we study estimator (**Con-UF**) in scenarios *Random*-16 and *Coherent*-8. We run a version of Algorithm 3 with adaptive stepsize,

[3] These scenarios repeat those in Chapter 3; their description is given for simplicity of reference.

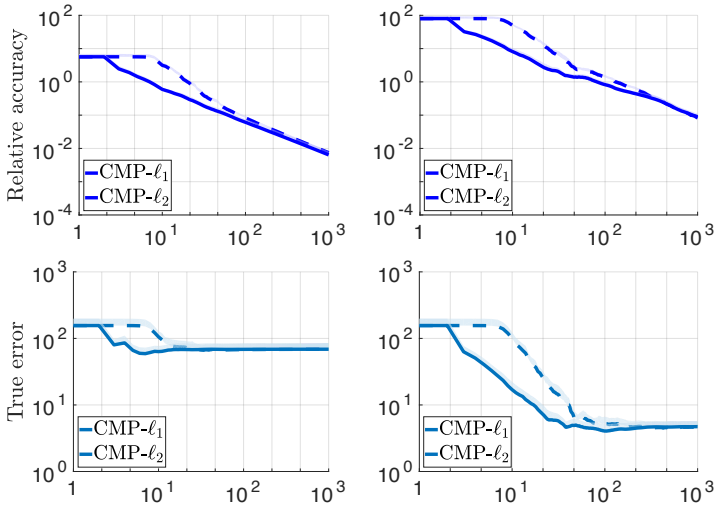

Figure 5-1. Relative accuracy and true ℓ_∞-loss $\|F_n[x - \tilde{\varphi}(T) * y]\|_\infty$ vs. iteration for the approximate solution of (**Con-UF**) by Algorithm 3 in scenario *Coherent*-8 with SNR $= 1$ (left) and SNR $= 16$ (right).

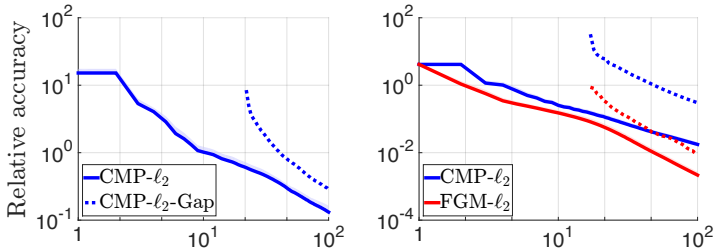

Figure 5-2. Relative accuracy vs. iteration for (**Con-UF**), left, and (**Con-LS**) with non-squared residual, right, in scenario *Coherent*-4 with SNR $= 4$. Dotted: accuracy certificate, cf. Section 5.B and [NOR10].

see [NN13], plotting the relative accuracy of the corresponding approximate solution $\tilde{\varphi}(T)$, that is, $\varepsilon(T)$ normalized by the optimal value of the residual $\|F_n[y - \widehat{\varphi} * y]\|_\infty$, versus the iteration count 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 (see Section 5.3): the full ℓ_2-setup composed from the partial ℓ_2-setups, and the "geometry-adapted" setup composed from the partial ℓ_1-setups. To obtain a proxy for $\widehat{\varphi}$, we reformulate (**Con-UF**) as a second-order cone problem, and run the MOSEK interior-point solver [AA13] with CVX interface [GB14]; 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 Figure 5-1, can be summarized as follows. First, we see that the higher is SNR, the harder is the optimization task, as predicted by (5.28). 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

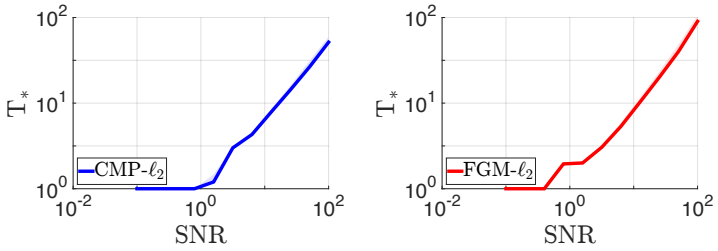

Figure 5-3. The iteration at which the statistical accuracy ε_* is attained for (**Con-UF**), left, and (**Con-LS**), right, in scenario *Random*-4.

in earlier iterations; however, the performances of the two setups match in later iterations as predicted by our theory.

Most importantly, we see that no more than a hundred iterations yield relative accuracy of 100%, resulting in approximate solution $\tilde{\varphi}$ with ℓ_∞-residual at most twice larger than the optimal one. From the analysis of uniform-fit estimators in Chapter 2, we know that the optimal value of the residual is upper-bounded with $O(\sigma\varrho)$, so that the conditions of Theorem 5.3.3 are satisfied for $\tilde{\varphi}$. We conclude that approximate solution $\tilde{\varphi}$ allows for the same theoretical bound as the exact solution $\widehat{\varphi}$, and one can predict that further optimization is redundant, and might even lead to some overfitting. This prediction is confirmed empirically: the true error curves begin to plateau no later than at $T = 10^2$.

Convergence and accuracy certificates. In this experiment, we study convergence of Algorithm 2 and Algorithm 3. We work in the same setting as previously, but this time also study estimator (**Con-LS**) for which we compare the recommended approach (Algorithm 2) and the alternative in which Algorithm 3 is applied to the version of (**Con-LS**) with non-squared residual. The results of this experiment are shown in Fig. 5-2. From the log-log plots, one can clearly see $O(T^{-1})$ convergence of Algorithm 3 for (**Con-UF**), and $O(T^{-2})$ convergence of Algorithm 2 for (**Con-LS**). Moreover, as predicted by (5.32), we see an elbow in the convergence plot of Algorithm 2 for (**Con-LS**) with non-squared residual. In addition to relative accuracy, we plot an upper bound on it obtained via the accuracy certificate technique, cf. [NOR10] and Section 5.B. Such bounds can be used to stop the algorithms one a guarantee for the desired accuracy has been obtained, and our experiment reveals them to be quite accurate in practice.

Statistical complexity bound. In this experiment (see Figure 5-3), we illustrate the affine dependency of the statistical complexity T_* from SNR predicted by our theory, see (5.30) and (5.31); note that although the signal in *Random* is not sparse on the DFT grid, its DFT is likely to have only a few large spikes which would suffice for (5.31). For various SNR values, we generate a signal in scenario *Random*-4, and define the first iteration at which $\varepsilon(T)$ crosses level $\sigma\varrho$ for (**Con-UF**) solved with Algorithm 3, and $\sigma^2\varrho^2$ for (**Con-LS**) with Algorithm 2. We see that the log-log curves plateau for low SNR and have unit tangent for high SNR, confirming our predictions.

Statistical performance with early stopping. In this experiment, we present additional scenario *Modulated-s-m*, in which the signal is a sum of sinusoids with polynomial modulation:

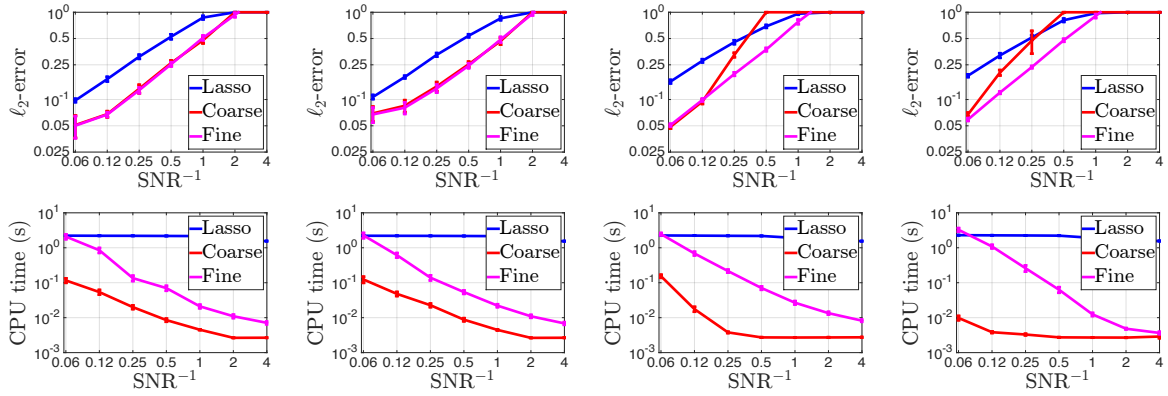

Figure 5-4. ℓ_2-loss and CPU time spent to compute estimators φ^{coarse}, φ^{fine}, and Lasso.

$x_t = \sum_{k=1}^{s} p_k(t) e^{i\omega_k t}$, where $p_k(\cdot)$ are iid polynomials of degree r with iid coefficients distributed as $\mathbb{C}\mathcal{N}(0,1)$; note that in this case, $\dim(\mathcal{S}) = 2s(m+1)$. Our goal is to study how the early stopping of an algorithm upon reaching accuracy ε_* (using an accuracy certificate) affects the statistical performance of the resulting estimator. For that, we generate signals in scenarios *Random*-4, *Coherent*-2, *Modulated*-4-2 (quadratic modulation), and *Modulated*-4-4 (quartic modulation), with different SNR, and compare three estimators: approximate solution φ^{coarse} to (**Con-LS**) with guaranteed accuracy $\varepsilon_* = \sigma^2\varrho^2$, near-optimal solution φ^{fine} with guaranteed accuracy $0.01\varepsilon_*$, and the Lasso estimator, with the standard choice of parameters as described in [BTR13], which we compute by running 3000 iterations of the FISTA algorithm [BT09]; note that the optimization problem in the latter case is unconstrained, and we do not have an accuracy certificate. We plot the scaled ℓ_2-loss of an estimator and the CPU time spent to compute it[4].

The results are shown in Fig. 5-4. We observe that φ^{coarse} has almost the same performance as φ^{fine} while being computed 1-2 orders of magnitude faster on average; both significantly outperform Lasso in all scenarios.

5.5 Remaining proofs

Proof of relation (5.26). Note that $\mathcal{A}$ can be expressed as follows, cf. (5.15):

$$\mathcal{A} = \sqrt{2n+1} \cdot F_n P_n F_{2n}^{\mathrm{H}} D_y F_{2n} P_n^{\mathrm{H}} F_n^{\mathrm{H}}. \tag{5.34}$$

By Young's inequality, for any $\psi \in \mathbb{C}^{n+1}$ we get

$$\begin{aligned}\frac{1}{2n+1}\|\mathcal{A}\psi\|_2^2 &\le \left\|D_y F_{2n} P_n^{\mathrm{H}} F_n^{\mathrm{H}}\psi\right\|_2^2 \\ &\le \left\|F_{2n}[y]_{-n}^{n}\right\|_\infty^2 \left\|F_{2n} P_n^{\mathrm{H}} F_n^{\mathrm{H}}\psi\right\|_2^2 \\ &\le \left\|F_{2n}[y]_{-n}^{n}\right\|_\infty^2 \|\psi\|_2^2,\end{aligned}$$

where we used that P_n is non-expansive. □

[4] We used MacBook Pro 2013 with 2.4 GHz Intel Core i5 CPU and 8GB of RAM.

Proof of Proposition 5.3.2. Consider the uniform grid on the unit circle

$$U_n = \left\{ \exp\left(\frac{2\pi i j}{n+1}\right) \right\}_{j=0}^{n},$$

and the twice finer grid

$$U_N = \left\{ \exp\left(\frac{2\pi i j}{N+1}\right) \right\}_{j=0}^{N}, \quad N = 2n+1.$$

Note that U_N is the union of U_n and the shifted grid

$$\tilde{U}_n = \left\{ u\, e^{i\theta},\ u \in U_n \right\}, \quad \theta = \frac{2\pi}{N+1};$$

note that $\tilde{U}_n$ and U_n do not overlap. One can check that for any $n \in \mathbb{Z}_+$ and $x \in \mathbb{C}_n(\mathbb{Z})$, the components of $F_n[x]_0^n$ form the set

$$\left\{ \frac{x(\nu)}{\sqrt{n+1}} \right\}_{\nu \in U_n},$$

where $x(\cdot)$ is the Taylor series corresponding to x:

$$x(\nu) := \sum_{\tau \in \mathbb{Z}} x_\tau \nu^\tau.$$

Now, let x be as in the premise of the theorem, and let $x^{(n)} \in \mathbb{C}_n(\mathbb{Z})$ be such that $x_\tau^{(n)} = x_\tau$ if $0 \le \tau \le n$ and $x_\tau^{(n)} = 0$ otherwise. Similarly, let us introduce $x^{(N)}$ as x restricted on $\mathbb{C}_N(\mathbb{Z})$. Then one can check that for any $\nu \in U_N$,

$$x^{(N)}(\nu) = \begin{cases} 2x^{(n)}(\nu), & \nu \in U_n, \\ 0, & \nu \in \tilde{U}_n. \end{cases} \tag{5.35}$$

In particular, this implies that

$$\|F_N[x]_0^N\|_\infty = \sqrt{2}\|F_n[x]_0^n\|_\infty. \tag{5.36}$$

Now, for any $\varphi \in \mathbb{C}_n(\mathbb{Z})$, let $\phi \in \mathbb{C}_N(\mathbb{Z})$ be its $n+1$-periodic extension, defined by

$$[\phi]_0^N = [[\varphi]_0^n; [\varphi]_0^n].$$

One can directly check that for x as in the premise of the theorem, the circular convolution of $[\phi]_0^N$ and $[x]_0^N$ is simply a one-fold repetition of $2[\varphi * x]_0^n$. Hence, using the Fourier diagonalization property together with (5.36) applied for $[\varphi * x]_0^n$ instead of x_0^n, we obtain

$$\sqrt{N+1}\,\|F_N[x] \odot F_N[\phi]\|_\infty = 2\sqrt{2}\,\|F_n[x * \varphi]\|_\infty \tag{5.37}$$

where $a \odot b$ is the elementwise product of $a, b \in \mathbb{C}^{n+1}$.

Finally, note that since $\sigma = 0$, and, as such, $x = y$ a.s., for any $\psi \in \mathbb{C}^{n+1}$ one has:

$$\mathcal{A}\psi = F_n[x * \varphi], \quad \text{where } \varphi = F_n^{\mathrm{H}}[\psi] \in \mathbb{C}_n(\mathbb{Z}).$$

Hence, using (5.37) with such φ, we arrive at

$$\begin{aligned}\|\mathcal{A}\psi\|_\infty &= \|F_n[x * \varphi]\|_\infty \\ &= \frac{\sqrt{n+1}}{2}\|F_N[x] \odot F_N[\phi]\|_\infty && \text{[by (5.37)]} \\ &= \sqrt{n+1}\|F_n[x] \odot \psi\|_\infty. && \text{[by (5.35)]}\end{aligned}$$

The claim now follows by maximizing the right-hand side in $\psi \in \mathbb{C}^{n+1} : \|\psi\|_1 \le 1$. □

Statistical accuracy for least-squares estimators. Let us first summarize the original argument for the exact optimal solution $\widehat{\varphi}$ to (**Con-LS**) with $\bar{\varrho} = \varrho$, see the proof of Theorem 3.2.1. Introducing the scaled Hermitian dot product for $\varphi, \psi \in \mathbb{C}_n(\mathbb{Z})$, $\langle \varphi, \psi \rangle_n = \frac{1}{n+1}\sum_{\tau=0}^{n} \overline{\varphi}_\tau \psi_\tau$, the squared ℓ_2-loss can be decomposed as follows:

$$\begin{aligned}\|x - \widehat{\varphi} * y\|_{n,2}^2 &= \|y - \widehat{\varphi} * y\|_{n,2}^2 - \sigma^2\|\zeta\|_{n,2}^2 - 2\sigma\langle \zeta, x - \widehat{\varphi} * y\rangle_n \\ &\le \|y - \varphi^o * y\|_{n,2}^2 - \sigma^2\|\zeta\|_{n,2}^2 - 2\sigma\langle \zeta, x - \widehat{\varphi} * y\rangle_n \\ &= \|x - \varphi^o * y\|_{n,2}^2 + 2\sigma\langle \zeta, x - \varphi^o * y\rangle_n - 2\sigma\langle \zeta, x - \widehat{\varphi} * y\rangle_n, \end{aligned} \tag{5.38}$$

where the inequality is due to feasibility of φ^o in (**Con-LS**). The dominating term in the right-hand side is the first one (corresponding to the squared oracle loss): we know that under the recoverability assumption, with probability $\ge 1 - \delta$ one has

$$\|x - \varphi^o * y\|_{n,2}^2 + 2\sigma\langle \zeta, x - \varphi^o * y\rangle_n \le \frac{C_0\sigma^2\varrho^2 \log\left(\frac{n+1}{\delta}\right)}{n+1}. \tag{5.39}$$

Then, last term in the right-hand side of (5.38) can be bounded as follows:

$$2\sigma|\langle \zeta, x - \widehat{\varphi} * y\rangle_n| \le \frac{C_1\sigma\left(\sqrt{s} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)}{\sqrt{n+1}}\|x - \widehat{\varphi} * y\|_{n,2} + \frac{C_2\sigma^2\varrho\left(1 + \log\left(\frac{n+1}{\delta}\right)\right)}{n+1}. \tag{5.40}$$

Collecting (5.38)-(5.40) and solving the resulting quadratic inequality, one could bound the scaled ℓ_2-loss of $\widehat{\varphi}$:

$$\|x - \widehat{\varphi} * y\|_{n,2} \le \frac{C\sigma}{\sqrt{n+1}}\left(\sqrt{s} + \varrho\sqrt{\log\left(\frac{n+1}{\delta}\right)}\right). \tag{5.41}$$

Now, note that in the proof of (5.40) we did not use *optimality* of $\widehat{\varphi}$; instead, the argument relied only on the following two facts:

(i) $x \in \mathcal{S}$ where $\mathcal{S}$ is a shift-invariant subspace of $\mathbb{C}(\mathbb{Z})$ with $\dim(\mathcal{S}) = s$;

(ii) one has a bound on the Fourier-domain ℓ_1-norm of $\widehat{\varphi}$: $\|F_n[\widehat{\varphi}]\|_1 \le \frac{\varrho}{\sqrt{n+1}}$.

Hence, it is now evident that an ε-accurate solution $\tilde{\varphi}$ to (**Con-LS**) with $\varepsilon = O(\sigma^2\varrho^2)$ still

satisfies (5.41). Indeed, the error decomposition (5.38) must now be replaced with

$$\|x - \tilde{\varphi} * y\|_{n,2}^2 \leq \|x - \varphi^o * y\|_{n,2}^2 + 2\sigma\langle\zeta, x - \varphi^o * y\rangle_n - 2\sigma\langle\zeta, x - \tilde{\varphi} * y\rangle_n + \frac{\varepsilon}{n+1}. \tag{5.42}$$

Then, (5.39) does not depend on $\tilde{\varphi}$, and hence is preserved. The term $\frac{\varepsilon}{n+1}$ enters additively and allows for the same upper bound as (5.39). Finally, (5.40) is preserved when replacing $\widehat{\varphi}$ with $\tilde{\varphi}$ since (i) and (ii) remain true. □

5.A Adaptive stepsize policies

Fast Gradient Method. Note that Algorithm 2 requires the knowledge of the Lipschitz constant L_f. However, it is not hard to come up with an adaptive stepsize selection policy that preserves the complexity estimate up to a constant. The corresponding algorithm is outlined as Algorithm 4. It can be shown that as long as one sets $\underline{L} \leq 2L_f$, the bound of Theorem 5.1.2 is preserved (up to an absolute constant), whereas the total number of inner loop iterations per an outer loop iteration scales as $\log(2L_f/\underline{L})$. Moreover, instead of the "aggressive" updates $L_t := \underline{L}$ in line 6 of Algorithm 4, one can consider "lazy" updates $L_t = L_{t-1}$, $t \geq 1$. Then, the complexity in terms of the number of outer loop iterations is again preserved, whereas the average number of inner loop iterations per an outer loop iteration is at most 2 as follows from the analysis of [NP06].

Composite Mirror Prox. Algorithm 3 can also be modified to incorporate adaptive stepsize selection while still allowing for the convergence guarantee stated in Theorem 5.1.3. In fact, one can replace the constant steps $\gamma \equiv L$ with stepsizes that satisfy $\gamma_\tau \geq 1/L$ as well as the following:

$$\gamma_\tau \left\langle \nabla F(w_{\tau+1/2}), w_{\tau+1/2} - w_{\tau+1} \right\rangle - D_{w_\tau}(w_{\tau+1}) \leq 0. \tag{5.43}$$

Note that condition (5.43) holds, for example, if $\gamma_\tau \equiv 1/L$. One can then employ an adaptive stepsize strategy similar to those for the Fast Gradient Method, for example, increasing the stepsize by a constant factor until (5.43) is violated, and using the last "admissible" value of γ_τ.

5.B Online accuracy certificates

The guarantees on the accuracy of optimization algorithms presented in Section 5.1 have a common shortcoming. They are "offline" and worst-case, stated once and for all, for the worst possible problem instance. Neither do they get improved in the course of computation, nor become more optimistic when facing an "easy" problem instance of the class. However, in some situations, online and "opportunistic" bounds on the accuracy are available. Following the terminology introduced in [NOR10], such bounds are called *accuracy certificates*. They can be used for the early stopping of the algorithm upon reaching desired accuracy ε). One situation in which accuracy certificates are available is saddle-point minimization (via a first-order algorithm) in the case where the domains are bounded and admit an efficiently computable *linear maximization oracle*. The latter means that the optimization problems $\max_{u \in U}\langle a, u\rangle$, $\max_{v \in V}\langle b, v\rangle$ can be efficiently solved for any a, b. An example of such domains is the unit ball of a norm $\|\cdot\|$ for which

Algorithm 4 : `Fast Gradient Method with Adaptive Stepsize Policy`

$u^0 = u_\omega$

$g^0 = 0 \in E$

for $t = 0, 1, ...$ **do**

  $s = 0$

  repeat

    $L_t = 2^s \underline{L}$

    $\eta_t = 1/L_t$

    $u_t = \mathrm{Prox}_{\eta_t \Psi, u_\omega}\left(\eta_t g^t\right)$

    $\tau_t = \frac{2(t+2)}{(t+1)(t+4)}$

    $u_{t+\frac{1}{3}} = \tau_t u_t + (1-\tau_t) u^t$

    $g_t = \frac{t+2}{2} \nabla f(u_{t+\frac{1}{3}})$

    $u_{t+\frac{2}{3}} = \mathrm{Prox}_{\eta_t \Psi, u_t}\left(\eta_t g_t\right)$

    $u^{t+1} = \tau_t u_{t+\frac{2}{3}} + (1-\tau_t) u^t$

    $\delta_t = \frac{L_t}{2} \left\| u^{t+1} - u_{t+\frac{1}{3}} \right\|^2 + \left\langle \nabla f(u_{t+\frac{1}{3}}), u^{t+1} - u_{t+\frac{1}{3}} \right\rangle + f(u_{t+\frac{1}{3}}) - f(u^{t+1})$

    $s = s + 1$

  until $\delta_t > 0$

  $g^{t+1} = \sum_{\tau=0}^{t} g_t$

end for

the dual norm $\|\cdot\|_*$ is efficiently computable. Let us now demonstrate how an accuracy certificate can be computed in this situation (see [NOR10, HJN15] for a more detailed exposition).

A *certificate* is simply a sequence $\lambda^t = (\lambda_\tau^t)_{\tau=1}^t$ of positive weights such that $\sum_{\tau=1}^t \lambda_\tau^t = 1$. Consider the λ^t-average of the iterates z_τ obtained by the algorithm,

$$z^t = [u^t, v^t] = \sum_{\tau=1}^{t} \lambda_\tau^t z_\tau.$$

A trivial example of certificate corresponds to the constant stepsize, and amounts to simple averaging. However, one might consider other choices of certificate, for which theoretical complexity bounds are preserved – for example, it might be practically reasonable to average only the last portion of the iterates, a strategy called "suffix averaging" [RSS12]. The point is that any certificate implies a non-trivial (and easily computable) upper bound on the accuracy of the corresponding candidate solution z^t. Indeed, the duality gap of a composite saddle-point

problem can be bounded as follows:

$$\begin{aligned}\overline{\phi}(u^t) - \underline{\phi}(v^t) &= \overline{\phi}(u^t) - \phi(u^t, v^t) + \phi(u^t, v^t) - \underline{\phi}(v^t) \\ &= \max_{v \in V}[\phi(u^t, v) - \phi(u^t, v^t)] - \min_{u \in U}[\phi(u, v^t) - \phi(u^t, v^t)] \\ &\leq \max_{v \in V}[\phi(u^t, v) - \phi(u^t, v^t)] + \max_{u \in U}[\phi(u^t, v^t) - \phi(u, v^t)].\end{aligned}$$

Now, using concavity of f in v, we have

$$\phi(u^t, v) - \phi(u^t, v^t) = f(u^t, v) - f(u^t, v^t) \leq \sum_{\tau=1}^{t} \lambda_\tau^t \langle F_v(z_\tau), v^t - v \rangle.$$

On the other hand, by convexity of f and Ψ in u,

$$\phi(u^t, v^t) - \phi(u, v^t) = f(u^t, v^t) - f(u, v^t) + \Psi(u^t) - \Psi(u) \leq \sum_{\tau=1}^{t} \lambda_\tau^t \langle F_u(z_\tau) + h(u_\tau), u^t - u \rangle$$

where $h(u_\tau)$ is a subgradient of $\Psi(\cdot)$ at u_τ. Combining the above facts, we get that

$$\overline{\phi}(u^t) - \underline{\phi}(v^t) \leq \max_{u \in U}[-F_u^t - h^t] + \max_{v \in V}[-F_v^t] + \sum_{\tau=1}^{t} \lambda_\tau^t \left[\langle F_u(z_\tau) + h(u_\tau), u^t \rangle + \langle F_v(z_\tau), v^t \rangle \right], \tag{5.44}$$

where

$$F_u^t = \sum_{\tau=1}^{t} \lambda_\tau^t F_u(z_\tau), \quad F_v^t = \sum_{\tau=1}^{t} \lambda_\tau^t F_v(z_\tau), \quad \text{and} \quad h^t = \sum_{\tau=1}^{t} \lambda_\tau^t h(u_\tau).$$

Note that the corresponding averages can often be recomputed in linear time in the dimension of the problem, and then upper bound (5.44) can be efficiently maintained. For example, this is the case when λ^t corresponds to a fixed sequence $\gamma_1, \gamma_2, ...,$

$$\lambda_\tau^t = \frac{\gamma_\tau}{\sum_{\tau' \leq t} \gamma_{\tau'}}, \quad \tau \leq t.$$

Note also that any bound on the duality gap implies bounds on the *relative* accuracy for the primal and the dual problem provided that $\underline{\phi}(v^t)$ (and hence the optimal value $\phi(u^*, v^*)$) is strictly positive (we used this fact in our experiments, see Section 5.4). Indeed, let $\varepsilon(t)$ be an upper bound on the duality gap (*e.g.* such as (5.44)), and hence also on the primal accuracy:

$$\overline{\phi}(u^t) - \phi(u^*, v^*) \leq \overline{\phi}(u^t) - \underline{\phi}(v^t) \leq \varepsilon(t).$$

Then, since $\phi(u^*, v^*) \geq \underline{\phi}(v^t) > 0$, we arrive at

$$\frac{\overline{\phi}(u^t) - \phi(u^*, v^*)}{\phi(u^*, v^*)} \leq \frac{\varepsilon(t)}{\underline{\phi}(v^t)}.$$

A similar bound can be obtained for the relative accuracy of the dual problem.

Conclusion and perspectives

In this thesis, we studied the problem of adaptive denoising in discrete time. Namely, a discrete-time signal is observed in Gaussian noise in a neighbourhood of some point t, and the goal is to estimate the signal at that point or in its neighborhood. Under quite general assumptions, one can prove the existence of a linear estimator with near-optimal statistical performance. In the classical formulation of the problem, the set of possible signals $\mathcal{X}$ is known, and hence a near-optimal linear estimator can be computed in advance. Instead, we study the adaptive formulation of the problem, where $\mathcal{X}$ is unknown, and a near-optimal linear estimator, called an *oracle*, cannot be explicitly computed. Instead, we aim at constructing an *adaptive estimator*, which only relies upon available observations, and has statistical risk close to the risk of the oracle. We assume that the oracle estimator takes the convolution form $\widehat{x}^o = \varphi^o * y$, where y is the vector of observations, and vector φ^o is called an *oracle filter*. The oracle filter depends on the hidden structure of the signal, and hence is unknown, but assumed to have a small pointwise risk. Specifically, one assumes that in $\Omega(n)$-neighborhood of the reference point, the pointwise risk of an oracle filter with length n is bounded by a constant times $\rho/\sqrt{n}$, where parameter $\rho \geq 1$ measures the complexity of the problem. The adaptive estimation problem is then reduced to constructing an adaptive filter $\widehat{\varphi} = \widehat{\varphi}(y)$ such that the risk of the corresponding convolution estimator is close to the risk of $\widehat{x}^o$. We studied two families of such estimators.

First, we explored *uniform-fit* estimators, in which one minimizes the ℓ_∞-norm of the discrepancy term $y - y * \varphi$ in the Fourier domain, constrained or penalized with the ℓ_1-norm of the Fourier transform of the filter. The pointwise risk of such estimators was shown to be within factor $O(\rho^3\sqrt{\log n})$ from the oracle risk, with a lower bound of $\Omega(\rho\sqrt{\log n})$ for that factor. The gap between the two bounds grows with the complexity parameter ρ. To ensure that this parameter is as small as possible, one has to solve an auxiliary task of *adaptive bandwidth selection*, that is, find the size n of the neighborhood for which the optimal bias-variance trade-off is achieved. We solve this task using Lepski's method.

Next, we introduced *least-squares* estimators, in which the ℓ_∞-norm of the Fourier-domain discrepancy term was replaced with the ℓ_2-norm. The price of adaptation for the estimators of this type is significantly reduced, specifically, to $O(\rho^2 + \rho\sqrt{\log n})$ in the case of pointwise risk, and to $O(\rho + \sqrt{\log n})$ in the case of ℓ_2-risk. Yet this improvement comes at a price: one needs an extra assumption that the signal belongs to a shift-invariant subspace of a small dimension or, more generally, is sufficiently close to such a subspace. The new assumption implies bounds on the complexity parameter ρ as a function of the subspace dimension s. We then focus on the task of tightening such bounds for various classes of filters, which results in better estimators for signals from shift-invariant subspaces, including harmonic oscillations. As a byproduct of our results, we bridge the statistical gap for the problem of estimating a general harmonic oscillation, without frequency separation conditions.

Finally, we proposed efficient algorithms for the numerical solution of the optimization problems that correspond to the proposed estimators. While these algorithms are based on the known first-order proximal methods from convex optimization, we introduced their analysis which takes into account the particular structure of the optimization problems of interest. Namely, we qualified the numerical complexity of these problems from the "statistical" point of view, describing the accuracy in the objective that is sufficient for the corresponding approximate estimator to admit similar statistical risk bounds as its exact counterpart. When combined with the technique of accuracy certificates, these results allow to equip our algorithms with a reliable stopping criterion and significantly accelerate computations. These findings were additionally confirmed by numerical experiments with synthetic data.

We now outline some possible directions of future research.

Indirect observations. Instead of signal denoising, one can consider a more general *noisy deconvolution* problem:

$$y = a * x + \sigma\zeta,$$

where the filter $a \in \mathbb{C}_m(\mathbb{Z})$ corresponds to a linear time-invariant observation operator. Problems of this kind arise in fluorescence microscopy [BBP15, Wat09] and exoplanet detection [FHL$^+$15, KLSH17], as well as in the general context of statistical inverse problems [MR96, CGP$^+$02, Joh11]. One can show that when x is a harmonic oscillation, the key quantity defining the complexity of the estimation problem is the signal distortion by the observation operator, as measured by the modulus of the z-transform of a on the Fourier-domain support of x: $\lambda := \min_{1\le k\le s} |a(e^{i\omega k})|$. Unfortunately, in the case where the harmonic structure of the signal is unknown, λ is unknown as well, and we cannot directly extend our adaptive recovery approach to this case. One way to address this problem would be as follows: first construct recoveries for a fixed value λ, and then select λ via a Lepski-type procedure – just as we did for bandwidth selection (cf. Section 2.2.2). However, the joint selection task is conceptually more difficult, since the resulting set of estimators is not totally ordered, and the regular Lepski method cannot be applied. In this connection, the generalized selection rule of Lepski and Goldenschluger for convolution-type estimators, see [GL11], looks perspective since it does not require the family of estimators to be ordered.

Signal recovery on graphs. Many signal processing problems involve signals living in the nodes of a graph. For instance, in computer graphics and vision, 3-D objects are modeled as manifolds endowed with properties such as color or texture; one way to work with these objects is by replacing them with graphs arising as their triangulations. Other examples include the models of social networks [L$^+$09], gene expression data [XOX02], and dynamic models in neuroscience [SFDSB15]. In all these applications, one has to deal with multiple time-varying processes in the nodes of a large graph, the edges of which specify the correlation between the processes. Exploiting the underlying low-rank structure is often vital in this context, and we would like to generalize our techniques to deal with this problem. One possible way towards such a generalization would be through a general framework of Fourier analysis on graphs recently introduced by [SM13] where one replaces the time-shift operator Δ with the weighted adjacency matrix of the graph.

Bibliography

[AA13] 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/.

[AB16] C. Aubel and H. Bölcskei. Deterministic performance analysis of subspace methods for cisoid parameter estimation. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 1551–1555. IEEE, 2016.

[AF03] R. A. Adams and J. Fournier. *Sobolev spaces*, volume 140. Academic press, 2003.

[AK64] P. M. Anselone and J. Korevaar. Translation invariant subspaces of finite dimension. *Proceedings of the American Mathematical Society*, 15(5):747–752, 1964.

[B+18] P. C. Bellec et al. Optimal bounds for aggregation of affine estimators. *The Annals of Statistics*, 46(1):30–59, 2018.

[BBP15] K. Bissantz, N. Bissantz, and K. Proksch. *Monitoring of significant changes over time in fluorescence microscopy imaging of living cells*. Universitätsbibliothek Dortmund, 2015.

[BK79] G. Bienvenu and L. Kopp. Principle de la goniometrie passive adaptive. In *Proc. 7ème Colloque GRESIT*, page 106, 1979.

[BL96] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, 1996.

[BM86] Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1081–1089, 1986.

[BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. SIAM, 2001.

[BTR13] B. Bhaskar, G. Tang, and B. Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Trans. Signal Processing*, 61(23):5987–5999, 2013.

[BVDG11] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[Cad80] J. Cadzow. High performance spectral estimation – a new ARMA method. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):524–529, 1980.

[Cad88] J. Cadzow. Signal enhancement — a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):49–62, 1988.

[CFG14] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

[CGP+02] L. Cavalier, G. K. Golubev, D. Picard, A. B. Tsybakov, et al. Oracle inequalities for inverse problems. *The Annals of Statistics*, 30(3):843–874, 2002.

[CRPW12] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[CT07] E. J. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 12 2007.

[CZ16] T. Cai and A. Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *arXiv:1605.00353*, 2016.

[DB13] M. F. Duarte and R. G. Baraniuk. Spectral compressive sensing. *Appl. & Comput. Harmon. Anal.*, 35(1):111–129, 2013.

[DLM90] D. L. Donoho, R. C. Liu, and B. MacGibbon. Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, pages 1416–1437, 1990.

[DRX+14] D. Dai, P. Rigollet, L. Xia, T. Zhang, et al. Aggregation of affine estimators. *Electronic Journal of Statistics*, 8(1):302–327, 2014.

[DS+12] A. Dalalyan, J. Salmon, et al. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

[DSSSC08] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 272–279, New York, NY, USA, 2008. ACM.

[EP96] S. Efromovich and M. Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statistica Sinica*, pages 925–942, 1996.

[FHL+15] D. A. Fischer, A. W. Howard, G. P. Laughlin, B. Macintosh, S. Mahadevan, J. Sahlmann, and J. C. Yee. Exoplanet detection techniques. *arXiv:1505.06869*, 2015.

[G+10] Yu. Golubev et al. On universal oracle inequalities related to high-dimensional linear models. *The Annals of Statistics*, 38(5):2751–2780, 2010.

[GB14] M. Grant and S. Boyd. *The* `CVX` *Users Guide. Release 2.1*, 2014. http://web.cvxr.com/cvx/doc/CVX.pdf.

[GL08] A. Goldenshluger and O. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 11 2008.

[GL11] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

[GN97] A. Goldenshluger and A. Nemirovski. Adaptive de-noising of signals satisfying differential inequalities. *IEEE Transactions on Information Theory*, 43(3):872–889, 1997.

[GO14] Yu. Golubev and D. Ostrovski. Concentration inequalities for the exponential weighting method. *Mathematical Methods of Statistics*, 23(1):20–37, 2014.

[Hay91] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 1991.

[HJN15] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

[HJNO15] Z. Harchaoui, A. Juditsky, A. Nemirovski, and D. Ostrovsky. Adaptive recovery of signals by convex optimization. In *Proceedings of The 28th Conference on Learning Theory (COLT) 2015, Paris, France, July 3-6, 2015*, pages 929–955, 2015.

[HS90] Y. Hua and T. K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(5):814–824, 1990.

[HS91] Y. Hua and T. K. Sarkar. On SVD for estimating generalized eigenvalues of singular matrix pencil in noise. In *Circuits and Systems, 1991., IEEE International Sympoisum on*, pages 2780–2783. IEEE, 1991.

[IK81] I. Ibragimov and R. Khasminskii. *Statistical estimation. Asymptotic Theory*, volume 16 of *Applications of Mathematics*. Springer, 1981.

[IK84] I. Ibragimov and R. Khasminskii. Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theor. Probab. & Appl.*, 29(1):1–32, 1984.

[JN00] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28:681–712, 2000.

[JN09] A. Juditsky and A. Nemirovski. Nonparametric denoising of signals with unknown local structure, I: Oracle inequalities. *Appl. & Comput. Harmon. Anal.*, 27(2):157–179, 2009.

[JN10] A. Juditsky and A. Nemirovski. Nonparametric denoising of signals with unknown local structure, II: Nonparametric function recovery. *Appl. & Comput. Harmon. Anal.*, 29(3):354–367, 2010.

[JN11a] A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[JN11b] A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, II: Utilizing problem structure. *Optimization for Machine Learning*, pages 149–183, 2011.

[JN13] A. Juditsky and A. Nemirovski. On detecting harmonic oscillations. *Bernoulli*, 23(2):1134–1165, 2013.

[JN17] A. Juditsky and A. Nemirovski. Near-optimality of linear recovery from indirect observations. *arXiv:1704.00835*, 2017.

[Joh11] I. Johnstone. *Gaussian estimation: sequence and multiresolution models*. Unpublished manuscript, 2011.

[JRT+08] A. Juditsky, P. Rigollet, A. B. Tsybakov, et al. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[Kay93] S. M. Kay. *Fundamentals of statistical signal processing*. Prentice Hall, 1993.

[KLSH17] T. H. Kim, K. M. Lee, B. Schölkopf, and M. Hirsch. Online video deblurring via dynamic temporal blending network. In *IEEE International Conference on Computer Vision (ICCV 2017)*, 2017.

[Kne94] A. Kneip. Ordered linear smoothers. *The Annals of Statistics*, pages 835–866, 1994.

[KSS86] R. Kumaresan, L. Scharf, and A. Shaw. An algorithm for pole-zero modeling and spectral analysis. *IEEE transactions on acoustics, speech, and signal processing*, 34(3):637–640, 1986.

[L+09] D. Lazer et al. Life in the network: the coming age of computational social science. *Science*, 323(5915), 2009.

[Lai79] P. G. Laird. On characterizations of exponential polynomials. *Pacific Journal of Mathematics*, 80(2):503–507, 1979.

[LB06] G. Leung and A. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

[Lep91] O. Lepski. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

[LF16] W. Liao and A. Fannjiang. Music for single-snapshot spectral estimation: Stability and super-resolution. *Applied and Computational Harmonic Analysis*, 40(1):33–67, 2016.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[LM09] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613, 2009.

[LMS97] O. Lepski, E. Mammen, and V. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, pages 929–947, 1997.

[LS96] J. Li and P. Stoica. Efficient mixed-spectrum estimation with applications to target feature extraction. *IEEE transactions on signal processing*, 44(2):281–295, 1996.

[LS14] O. Lepski and N. Serdyukova. Adaptive estimation under single-index constraint in a regression model. *Ann. Statist.*, 42(1):1–28, 2014.

[Mal08] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 2008.

[MIK00] D. G. Manolakis, V. K. Ingle, and S. M. Kogon. *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*. McGraw-Hill Boston, 2000.

[Moi15] A. Moitra. Super-resolution, extremal functions and the condition number of Vandermonde matrices. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 821–830. ACM, 2015.

[MR96] B. A. Mair and F. H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM Journal on Applied Mathematics*, 56(5):1424–1444, 1996.

[Nem91] A. Nemirovski. *On non-parametric estimation of functions satisfying differential inequalities*. Math. Sciences Research Inst., 1991.

[Nem00] 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.

[Nes83] Yu. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[Nes13a] Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[Nes13b] Yu. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[NN13] Yu. Nesterov and A. Nemirovski. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 5 2013.

[NOR10] A. Nemirovski, S. Onn, and U. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

[NP06] Yu. Nesterov and B. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[OH18] D. Ostrovskii and Z. Harchaoui. Efficient first-order algorithms for adaptive signal denoising. *arXiv:1803.11262*, March 2018.

[OHJN16a] D Ostrovsky, Z Harchaoui, A Juditsky, and A Nemirovski. Structure-blind signal recovery. In *Advances in Neural Information Processing Systems*, pages 4817–4825, 2016.

[OHJN16b] D. Ostrovsky, Z. Harchaoui, A. Juditsky, and A. Nemirovski. Structure-blind signal recovery. *arXiv:1607.05712v2*, October 2016.

[Pis73] V. F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical Journal International*, 33(3):347–366, 1973.

[QH01] B. G. Quinn and E. J. Hannan. *The estimation and tracking of frequency*, volume 9. Cambridge University Press, 2001.

[QT91] B. G. Quinn and P. J. Thomson. Estimating the frequency of a periodic function. *Biometrika*, 78(1):65–74, 1991.

[RH89] B. D. Rao and K. V. Hari. Performance analysis of root-MUSIC. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):1939–1949, 1989.

[RK89] R. Roy and T. Kailath. ESPRIT – estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on acoustics, speech, and signal processing*, 37(7):984–995, 1989.

[RPK86] R. Roy, A. Paulraj, and T. Kailath. ESPRIT – a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE transactions on acoustics, speech, and signal processing*, 34(5):1340–1342, 1986.

[RSS12] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 449–456, 2012.

[RT12] P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, pages 558–575, 2012.

[Sch06] A. Schuster. On the periodicities of sunspots. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 206:69–100, 1906.

[Sch79] R. Schmidt. Multiple emitter location and signal parameter estimation. In *Proc. RDAC Spectrum Estimation Workshop*, pages 243–256, 1979.

[Sch86] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

[SFDSB15] M. Schwemmer, A. Fairhall, S. Denéve, and E. Shea-Brown. Constructing precisely computing networks with biophysical spiking neurons. *The Journal of Neuroscience*, 35(28):10112–10134, 2015.

[Sio58] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

[SM05] P. Stoica and R. L. Moses. *Spectral analysis of signals*. Prentice Hall, 2005.

[SM13] A. Sandryhaila and J. M. F. Moura. Discrete signal processing on graphs. *IEEE transactions on signal processing*, 61(7):1644–1656, 2013.

[SMFS89] P. Stoica, R. L. Moses, B. Friedlander, and T. Soderstrom. Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):378–392, 1989.

[SN89] P. Stoica and A. Nehorai. MUSIC, maximum likelihood, and Cramer-Rao bound. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(5):720–741, 1989.

[SS91] P. Stoica and T. Soderstrom. Statistical analysis of music and subspace rotation estimates of sinusoidal frequencies. *IEEE Transactions on Signal Processing*, 39(8):1836–1847, 1991.

[SS04] P. Stoica and Y. Selen. Cyclic minimizers, majorization techniques, and the expectation-maximization algorithm: a refresher. *IEEE Signal Processing Magazine*, 21(1):112–114, 2004.

[Szé82] L. Székelyhidi. Note on exponential polynomials. *Pacific Journal of Mathematics*, 103(2):583–587, 1982.

[TBR13] G. Tang, B. 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.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 58(1):267–288, 1996.

[Tsy08] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

[UT96] S. Umesh and D. W. Tufts. Estimation of parameters of exponentially damped sinusoids using fast maximum likelihood estimation with application to NMR spectroscopy data. *IEEE Transactions on Signal Processing*, 44(9):2245–2259, 1996.

[VdV98] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

[Vu11] V. Vu. Singular vectors under random perturbation. *Random Structures & Algorithms*, 39(4):526–538, 2011.

[Was06] L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

[Wat09] J. C. Waters. Accuracy and precision in quantitative fluorescence microscopy. *The Journal of Cell Biology*, 185(7):1135–1148, 2009.

[Wed72] P.-A. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[XOX02] Y. Xu, V. Olman, and D. Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, 2002.

[YB99] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[YF96] N. Yuen and B. Friedlander. Asymptotic performance analysis of ESPRIT, higher order ESPRIT, and virtual ESPRIT algorithms. *IEEE Transactions on Signal Processing*, 44(10):2537–2550, 1996.

[YF98] N. Yuen and B. Friedlander. Performance analysis of higher order ESPRIT for localization of near-field sources. *IEEE transactions on Signal Processing*, 46(3):709–719, 1998.

[YMB08] G. Yu, S. Mallat, and E. Bacry. Audio denoising by time-frequency block thresholding. *IEEE Transactions on signal processing*, 56(5):1830–1839, 2008.

[YPM94] C. J. Ying, L. C. Potter, and R. L. Moses. On model order determination for complex exponential signals: Performance of an FFT-initialized ML algorithm. In *Proc. of IEEE SP Workshop on SSAP*, pages 43–46, 1994.

[ZW88] I. Ziskind and M. Wax. Maximum likelihood localization of multiple sources by alternating projection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(10):1553–1560, 1988.