# Adaptive Signal Denoising by Convex Optimization

Dmitrii Ostrovskii

Université Grenoble Alpes

Anatoli Juditsky
Univ. Grenoble Alpes

Arkadi Nemirovski
Georgia Tech

Zaid Harchaoui
Univ. of Washington

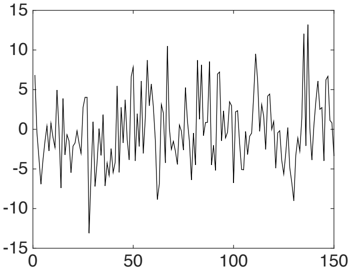
# Discrete-time signal denoising

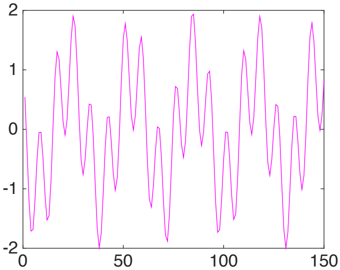
## The Problem

Recover **signal** $x = [x_{-n}; \ldots; x_n] \in \mathbb{C}_{n} = \mathbb{C}^{2n+1}$ from noisy observation

$$y_\tau = x_\tau + \sigma\xi_\tau, \quad -n \le \tau \le n,$$

$\xi_t$ are i.i.d. standard complex Gaussian, and $x_t = f(t)$ for $f : \mathbb{R} \to \mathbb{C}$.



# Discrete-time signal denoising

## The Problem

Recover **signal** $x = [x_{-n}; ...; x_n] \in \mathbb{C}_n = \mathbb{C}^{2n+1}$ from noisy observation

$$y_\tau = x_\tau + \sigma\xi_\tau, \quad -n \leq \tau \leq n,$$

$\xi_t$ are i.i.d. standard complex Gaussian, and $x_t = f(t)$ for $f : \mathbb{R} \to \mathbb{C}$.

- **Assumption:** signal has a shift-invariant structure.
- **Adaptive denoising:** the structure is unknown.

# Discrete-time signal denoising

## The Problem

Recover **signal** $x = [x_{-n}; \dots; x_n] \in \mathbb{C}_n = \mathbb{C}^{2n+1}$ from noisy observation

$$y_\tau = x_\tau + \sigma\xi_\tau, \quad -n \le \tau \le n,$$

$\xi_t$ are i.i.d. standard complex Gaussian, and $x_t = f(t)$ for $f : \mathbb{R} \to \mathbb{C}$.

- **Assumption:** signal has a shift-invariant structure.
- **Adaptive denoising:** the structure is unknown.

Example: harmonic oscillation with $s \ll n$ unknown frequencies:

$$x_\tau = \sum_{k=1}^{s} C_k e^{i\omega_k \tau}, \quad \omega_k \in [0, 2\pi[.$$

# Performance measure

Define the empirical $\ell_2$-norm: $\|x\|_{n,2} = \left(\frac{1}{2n+1}\sum_{-n\le\tau\le n}|x_\tau|^2\right)^{\frac{1}{2}}$.
**Quadratic Risk:**

$$R_n(\widehat{x}, x) = \left[\mathbf{E}\|\widehat{x} - x\|_{n,2}^2\right]^{\frac{1}{2}}.$$

# Performance measure

Define the empirical $\ell_2$-norm: $\|x\|_{n,2} = \left(\frac{1}{2n+1}\sum_{-n\le\tau\le n}|x_\tau|^2\right)^{\frac{1}{2}}$.
**Quadratic Risk:**

$$R_n(\widehat{x}, x) = \left[\mathbf{E}\|\widehat{x} - x\|_{n,2}^2\right]^{\frac{1}{2}}.$$

**Minimax Approach:** minimize the maximal risk on a given $\mathcal{X} \subset \mathbb{C}_n$:

$$\mathrm{Risk}^*(\mathcal{X}) = \inf_{\widehat{x}} \left\{ \bar{R}_{\mathcal{X}}(\widehat{x}) := \sup_{x\in\mathcal{X}} R_n(\widehat{x}, x) \right\}.$$

# Linear estimators

Define the minimax risk and the linear minimax risk:

$$\mathrm{Risk}^*(\mathcal{X}) = \inf_{\widehat{x}} \left\{ \bar{R}_{\mathcal{X}}(\widehat{x}) \right\}; \quad \mathrm{Risk}^{\mathsf{lin}}(\mathcal{X}) = \inf_{\widehat{x}=\Phi y} \left\{ \bar{R}_{\mathcal{X}}(\widehat{x}) \right\},$$

# Linear estimators

Define the minimax risk and the linear minimax risk:

$$\mathrm{Risk}^*(\mathcal{X}) = \inf_{\widehat{x}} \left\{ \bar{R}_{\mathcal{X}}(\widehat{x}) \right\}; \quad \mathrm{Risk}^{\mathsf{lin}}(\mathcal{X}) = \inf_{\widehat{x} = \Phi y} \left\{ \bar{R}_{\mathcal{X}}(\widehat{x}) \right\},$$

**Near-Optimality of linear estimators [Donoho '90]**

Whenever $\mathcal{X}$ is compact, ortho-symmetric, and quadratically convex,

$$\mathrm{Risk}^*(\mathcal{X}) \le 1.25 \cdot \mathrm{Risk}^{\mathsf{lin}}(\mathcal{X}).$$

- **Subspace**: $\mathrm{Risk}^{\mathsf{lin}}(\mathcal{S}) \approx \sigma \sqrt{\frac{\dim(\mathcal{S})}{n}}$; $\quad \widehat{x}^{\mathsf{lin}}$ is a **projector** on $\mathcal{S}$.

# Linear estimators

Define the minimax risk and the linear minimax risk:

$$\mathrm{Risk}^*(\mathcal{X}) = \inf_{\widehat{x}} \left\{ \bar{R}_{\mathcal{X}}(\widehat{x}) \right\}; \quad \mathrm{Risk}^{\mathsf{lin}}(\mathcal{X}) = \inf_{\widehat{x}=\Phi y} \left\{ \bar{R}_{\mathcal{X}}(\widehat{x}) \right\},$$

## Near-Optimality of linear estimators [Donoho '90]

Whenever $\mathcal{X}$ is compact, ortho-symmetric, and quadratically convex,

$$\mathrm{Risk}^*(\mathcal{X}) \le 1.25 \cdot \mathrm{Risk}^{\mathsf{lin}}(\mathcal{X}).$$

- **Subspace**: $\mathrm{Risk}^{\mathsf{lin}}(\mathcal{S}) \approx \sigma \sqrt{\frac{\dim(\mathcal{S})}{n}}$; $\widehat{x}^{\mathsf{lin}}$ is a **projector** on $\mathcal{S}$.

Near-optimal linear estimator can be efficiently computed [Juditsky & Nemirovski '16].

$\Rightarrow$ Linear estimators are "good" in a general situation.

# Adaptive estimation

If $\mathcal{X}$ is "good" but unknown, a good linear estimator $\widehat{x}^o$ still exists.

## Adaptive estimation task

Knowing that there exists an "oracle" $\widehat{x}^o$ – a linear estimator with a small risk $R_n(\widehat{x}, x)$ – **"mimic it":** construct $\widehat{x} = \widehat{x}(y)$ satisfying

$$R_n(\widehat{x}(y), x) \leq P \cdot R_n(\widehat{x}^o, x),$$

with the smallest possible price for adaptation $P$.

# Adaptive estimation

If $\mathcal{X}$ is "good" but unknown, a good linear estimator $\widehat{x}^o$ still exists.

**Adaptive estimation task**

Knowing that there exists an "oracle" $\widehat{x}^o$ – a linear estimator with a small risk $R_n(\widehat{x}, x)$ – **"mimic it":** construct $\widehat{x} = \widehat{x}(y)$ satisfying

$$R_n(\widehat{x}(y), x) \le P \cdot R_n(\widehat{x}^o, x),$$

with the smallest possible price for adaptation $P$.

- **Idea:** minimize an observable criterion over *linear* estimators.

# Adaptive estimation

If $\mathcal{X}$ is "good" but unknown, a good linear estimator $\widehat{x}^o$ still exists.

## Adaptive estimation task

Knowing that there exists an "oracle" $\widehat{x}^o$ – a linear estimator with a small risk $R_n(\widehat{x}, x)$ – **"mimic it":** construct $\widehat{x} = \widehat{x}(y)$ satisfying

$$R_n(\widehat{x}(y), x) \le P \cdot R_n(\widehat{x}^o, x),$$

with the smallest possible price for adaptation $P$.

- **Idea:** minimize an observable criterion over *linear* estimators.
- The class of *all* linear estimators is **too large**.
  - Risk of $\widehat{x}^o(y) = \mathrm{proj}_x(y)$ is only $\frac{\sigma}{\sqrt{n}}$ but we cannot hope to find this estimator.

$\Rightarrow$ **Regularize** using prior information.

# Linear filtering

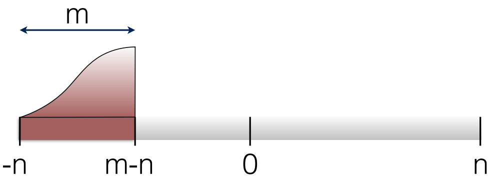
"Positive-time domain": $\mathbb{C}_n^+ = \mathbb{C}^{n+1}$, with $\|x\|_{n,2}$ and $R_n(\widehat{x}, x)$ correspondingly modified.

Consider time-invariant linear estimators.

- Linear filtering with a "left" **filter** $\varphi \in \mathbb{C}_m^+$ for some $m \le n$:

$$\widehat{x}_t = [\varphi * y]_t := \sum_{\tau=0}^{m} \varphi_\tau y_{t-\tau},$$

where $*$ is discrete convolution, and $-n + m \le t \le n$.



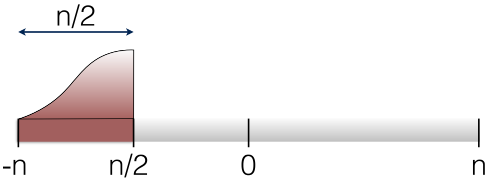
# Main assumption

We assume the existence of a linear filter with a small pointwise error.

## Assumption

$x$ is $\rho$-recoverable: there exists $\phi^o \in \mathbb{C}^+_{n/2}$ which satisfies

$$\left(\mathbf{E}\left|x_t - [\phi^o * y]_t\right|^2\right)^{1/2} \leq \frac{\sigma\rho}{\sqrt{n+1}}, \quad t \in [-n/2, n].$$



# Main assumption

We assume the existence of a linear filter with a small pointwise error.

**Assumption**

$x$ is $\rho$-recoverable: there exists $\phi^o \in \mathbb{C}^+_{n/2}$ which satisfies

$$\left(\mathbf{E}\left|x_t - [\phi^o * y]_t\right|^2\right)^{1/2} \le \frac{\sigma\rho}{\sqrt{n+1}}, \quad t \in [-n/2, n].$$

**Consequencies:**

- Small quadratic risk: $\widehat{x}^o = \phi^o * y$ satisfies

$$R_n(\widehat{x}^o, x) \le \frac{\sigma\rho}{\sqrt{n+1}}.$$

- Bias-variance decomposition:

$$\mathbf{E}|x_t - [\phi^o * y]_t|^2 = \mathbf{E}|x_t - [\phi^o * x]_t|^2 + \sigma^2 \mathbf{E}|[\phi^o * \xi]_t|^2.$$

# Main assumption

We assume the existence of a linear filter with a small pointwise error.

**Assumption**

$x$ is $\rho$-recoverable: there exists $\phi^o \in \mathbb{C}^+_{n/2}$ which satisfies

$$\left(\mathbf{E}\left|x_t - [\phi^o * y]_t\right|^2\right)^{1/2} \le \frac{\sigma\rho}{\sqrt{n+1}}, \quad t \in [-n/2, n].$$

**Consequencies:**

- Small quadratic risk: $\widehat{x}^o = \phi^o * y$ satisfies

$$R_n(\widehat{x}^o, x) \le \frac{\sigma\rho}{\sqrt{n+1}}.$$

- Bias-variance decomposition $\Rightarrow$

$$\|x - \phi^o * x\|_{n,2} \le \frac{\sigma\rho}{\sqrt{n+1}}, \quad \|\phi^o\|_2 \le \frac{\rho}{\sqrt{n+1}}.$$

# Adaptive filtering

$\mathcal{F}_n : \mathbb{C}_n^+ \to \mathbb{C}_n^+$ – unitary Discrete Fourier transform (DFT) operator.

**Estimator:** $\widehat{x} = \widehat{\varphi} * y$ where $\widehat{\varphi}$ is an optimal solution to

$$\text{minimize}_{\varphi \in \mathbb{C}_n^+} \quad \|y - \varphi * y\|_{n,2}^2 \quad \text{subject to} \;\; \|\mathcal{F}_n[\varphi]\|_1 \le \frac{r}{\sqrt{n+1}}.$$

- Efficiently computable by 1st-order cvx optimization algorithms: Fast Gradient Method [Nesterov & Nemirovski '13].

# Adaptive filtering

$\mathcal{F}_n : \mathbb{C}_n^+ \to \mathbb{C}_n^+$ – unitary Discrete Fourier transform (DFT) operator.

**Estimator:** $\widehat{x} = \widehat{\varphi} * y$ where $\widehat{\varphi}$ is an optimal solution to

$$\text{minimize}_{\varphi \in \mathbb{C}_n^+} \quad \|y - \varphi * y\|_{n,2}^2 \quad \text{subject to} \quad \|\mathcal{F}_n[\varphi]\|_1 \le \frac{r}{\sqrt{n+1}}.$$

- Efficiently computable by 1st-order cvx optimization algorithms: Fast Gradient Method [Nesterov & Nemirovski '13].

**Theorem.** If $r \ge \rho^2$, adaptive recovery $\widehat{x} = \widehat{\varphi} * y$ satisfies (w.h.p.)

$$\|x - \widehat{x}\|_{n,2} \lesssim \frac{\sigma(r + \sqrt{r \log n})}{\sqrt{n+1}}.$$

- Price of adaptation $\mathcal{O}(\rho + \sqrt{\log n})$.

# Sketch of analysis

- $\phi^o \in \mathbb{C}^+_{n/2}$ satisfies $\|\phi^o\|_2 \le \frac{\rho}{\sqrt{n+1}}$ and has small $\|x - \phi^o * y\|^+_{n,2}$.

# Sketch of analysis

- $\phi^o \in \mathbb{C}^+_{n/2}$ satisfies $\|\phi^o\|_2 \le \frac{\rho}{\sqrt{n+1}}$ and has small $\|x - \phi^o * y\|^+_{n,2}$.
- Search for $\phi^o$:

$$\widehat{\phi} \in \operatorname*{Argmin}_{\phi \in \mathbb{C}^+_{n/2}} \left\{ \|y - \phi * y\|^2_{n,2} : \ \|\phi\|_2 \le \frac{\rho}{\sqrt{n+1}} \right\}.$$

- $\phi^o$ is **feasible**, so that

$$\|y - \widehat{\phi} * y\|^2_{n,2} \le \|y - \phi^o * y\|^2_{n,2}.$$

By "simple algebra":

$$\|x - \widehat{\phi} * y\|^2_{n,2} = \|x - \phi^o * y\|^2_{n,2} + 2\sigma^2 \langle \xi, \widehat{\phi} * \xi \rangle_n + [\ldots]$$

**Fail:** cannot control the cross-term $\langle \xi, \widehat{\phi} * \xi \rangle_n$.

# Sketch of analysis, continued

## Key Fact

The **auto-convolution** $\varphi^o := \phi^o * \phi^o \in \mathbb{C}_n^+$ satisfies

$$\|\mathcal{F}_n[\varphi^o]\|_1 \le \frac{\rho^2}{\sqrt{n+1}}, \quad \|x - \varphi^o * y\|_{n,2} \le \frac{\sigma\rho^2}{\sqrt{n+1}}.$$

# Sketch of analysis, continued

**Key Fact**

The **auto-convolution** $\varphi^o := \phi^o * \phi^o \in \mathbb{C}_n^+$ satisfies

$$\|\mathcal{F}_n[\varphi^o]\|_1 \le \frac{\rho^2}{\sqrt{n+1}}, \quad \|x - \varphi^o * y\|_{n,2} \le \frac{\sigma\rho^2}{\sqrt{n+1}}.$$

- Search for $\varphi^o$:

$$\widehat{\varphi} \in \operatorname*{Argmin}_{\varphi \in \mathbb{C}_n^+} \left\{ \|y - \varphi * y\|_{n,2}^2 : \ \|\mathcal{F}_n[\varphi^o]\|_1 \le \frac{\rho^2}{\sqrt{n+1}} \right\}.$$

- As before,

$$\|x - \widehat{\varphi} * y\|_{n,2}^2 = \|x - \varphi^o * y\|_{n,2}^2 + 2\sigma^2 \langle \xi, \widehat{\varphi} * \xi \rangle_n + [\dots].$$

- **By Parseval's theorem:**

$$\langle \xi, \widehat{\varphi} * \xi \rangle_n \lesssim \|\mathcal{F}_n[\widehat{\varphi}]\|_1 \|\mathcal{F}_n[\xi]\|_\infty^2 \lesssim \frac{\rho^2 \log(n+1)}{\sqrt{n+1}}. \qquad \square$$

# Application: harmonic oscillations

Oscillation with $s$ unknown frequencies: $x_\tau = \sum_{k=1}^{s} C_k e^{i\omega_k \tau}$.

# Application: harmonic oscillations

Oscillation with $s$ unknown frequencies: $x_\tau = \sum_{k=1}^{s} C_k e^{i\omega_k \tau}$.

- State of the art: Atomic Soft Thresholding [Tang *et al.* '12]:

$$R_n(\widehat{x}, x) \lesssim \frac{\sigma\sqrt{s\log(n+1)}}{\sqrt{n+1}},$$

  optimal, but only if frequencies are separated by a DFT bin $\frac{2\pi}{\sqrt{n}}$.

# Application: harmonic oscillations

Oscillation with $s$ unknown frequencies: $x_\tau = \sum_{k=1}^{s} C_k e^{i\omega_k \tau}$.

- State of the art: Atomic Soft Thresholding [Tang *et al.* '12]:

$$R_n(\widehat{x}, x) \lesssim \frac{\sigma\sqrt{s\log(n+1)}}{\sqrt{n+1}},$$

optimal, but only if frequencies are separated by a DFT bin $\frac{2\pi}{\sqrt{n}}$.

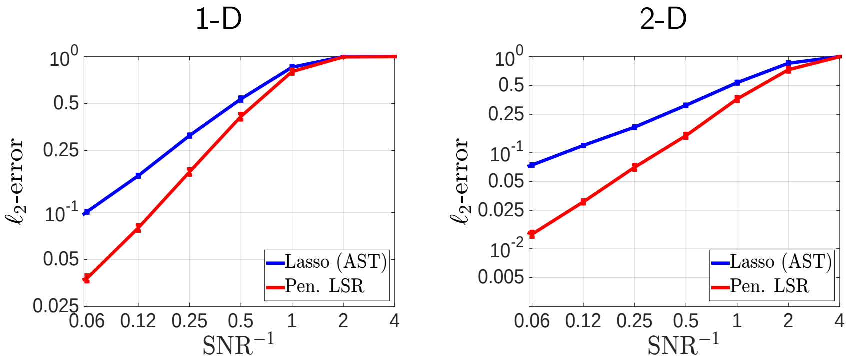
**Theorem.** Oscillation with $s$ frequencies is $\rho$-recoverable with

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

- Direct consequence: without any separation assumptions,

$$R_n(\widehat{x}, x) \lesssim \frac{\sigma s^2 \log n}{\sqrt{n+1}}.$$

Dependency on $s$ can be improved to $s^{3/2}$ (in preparation).




Denoising of random harmonic oscillations with 4 frequencies, $n = 100$ (95%-c.i.). Comparison with Atomic Soft Thresholding [Tang *et al.* '12].

# Conclusion

- We construct an adaptive estimator for time-invariant signals.
- Main idea: adaptation to a well-performing linear oracle.
- As a consequence, we solve an open problem of denoising harmonic oscillations with non-separated frequencies.

**Thank you for your attention!**

## Publications


- *Structure-Blind Signal Recovery.* NIPS 2016 (full: arXiv:1607.05712).
- *Adaptive Signal Recovery by Convex Optimization.* COLT 2015.