Adaptive signal denoising by convex optimization

Dmitry Ostrovsky
Université Grenoble Alpes

University of Göttingen
July 3, 2017

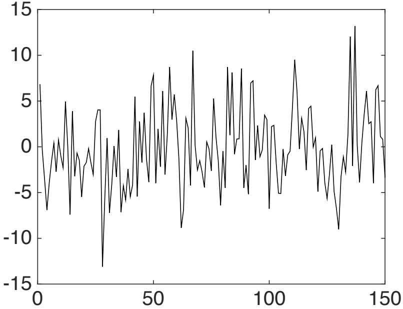
Ultimate goal

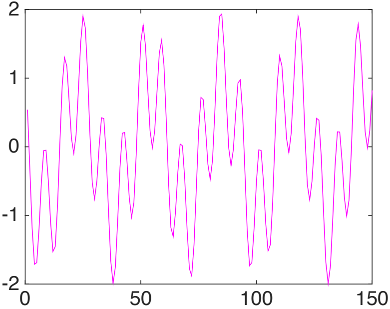
Recover a harmonic oscillation with $s \ll n$ frequencies:

$$x_t = \sum_{k=1}^{s} C_k e^{\imath \omega_k t}, \, t = 0, \ldots, n,$$

where $\{\omega_1, \ldots, \omega_s\} \subseteq [0, 2\pi)$ are unknown, from noisy observations

$$y_t = x_t + \sigma \xi_t, \quad \xi_t \sim \mathcal{N}(0, 1).$$





Ultimate goal

Recover a harmonic oscillation with $s \ll n$ frequencies:

$$x_t = \sum_{k=1}^{s} C_k e^{\imath \omega_k t}, \quad t = 0, \ldots, n,$$

where $\{\omega_1, \ldots, \omega_s\} \subseteq [0, 2\pi)$ are unknown, from noisy observations

$$y_t = x_t + \sigma \xi_t, \quad \xi_t \sim \mathcal{N}(0, 1).$$

State of the art: Atomic Soft Thresholding (Tang et al., 2012) achieves the optimal risk

$$\frac{\sigma^2 s \log(n)}{n}$$

if freqs are $\mathcal{O}(1/n)$-separated.

:(But without separation assumption, only **slow rate** $\mathcal{O}(1/\sqrt{n})$.

:) We achieve a near-optimal rate without separation assumption:

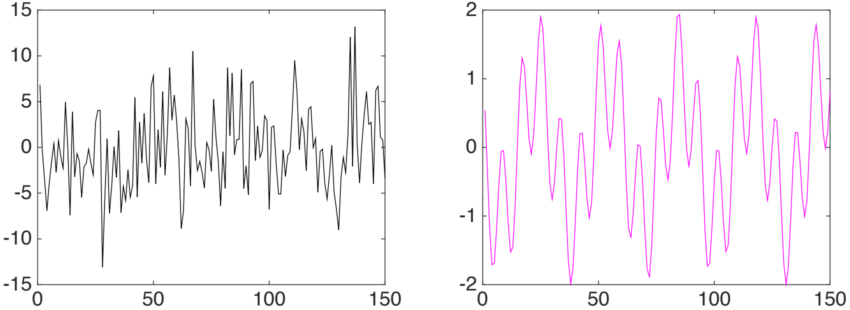
$$\frac{\sigma^2 s^4 \log^2(n)}{n}.$$

Preliminaries

Goal: recover discrete signal $x \in \mathbb{R}^n$ from a noisy observation

$$y_t = x_t + \sigma\xi_t, \quad t = 1, ..., n.$$

$\xi = (\xi_t)_{t=1}^n$ is standard Gaussian, and $x_t = f(t)$ for some $f : \mathbb{R} \to \mathbb{R}$.



Preliminaries

Goal: recover discrete signal $x \in \mathbb{R}^n$ from a noisy observation

$$y_t = x_t + \sigma\xi_t, \quad t = 0, ..., n,$$

$\xi = (\xi_t)_{t=1}^n$ is standard Gaussian, and $x_t = f(t)$ for some $f : \mathbb{R} \to \mathbb{R}$.

- Quadratic risk:

$$R(\widehat{x}, x) := \frac{1}{n}\mathbb{E}\left[\|\widehat{x} - x\|_2^2\right].$$

- We expect $R(\widehat{x}, x) = \mathcal{O}(\sigma^2/n)$.
- Linear estimators: $\widehat{x} = \Phi(y)$ for some linear operator Φ.

Example: recovery from a subspace

Recovery of the mean: suppose $x_t \equiv \mu$ for some $\mu \in \mathbb{R}$.

- Estimate μ from n repeated observations $\Rightarrow$ empirical mean:

$$\widehat{x} \equiv \frac{1}{n} \sum_{t=1}^{n} y_t .$$

Linear estimator, and $R(\widehat{x}, x) = \sigma^2/n$.

Example: recovery from a subspace

Recovery of the mean: suppose $x_t \equiv \mu$ for some $\mu \in \mathbb{R}$.

- Estimate μ from n repeated observations $\Rightarrow$ empirical mean:

$$\widehat{x} \equiv \frac{1}{n} \sum_{t=1}^{n} y_t.$$

Linear estimator, and $R(\widehat{x}, x) = \sigma^2/n$.

Equivalently, $x \in \mathcal{S}$, 1-d subspace spanned by all-ones vector.

- $\widehat{x} = \mathbf{proj}_{\mathcal{S}}(y)$, and $R(\widehat{x}, x) = \sigma^2/n$ since $\mathbf{proj}_{\mathcal{S}}(\sigma\xi) \sim \mathcal{N}(0, \sigma^2)$.

Example: recovery from a subspace

Recovery of the mean: suppose $x_t \equiv \mu$ for some $\mu \in \mathbb{R}$.

- Estimate μ from n repeated observations $\Rightarrow$ empirical mean:

$$\widehat{x} \equiv \frac{1}{n} \sum_{t=1}^{n} y_t .$$

Linear estimator, and $R(\widehat{x}, x) = \sigma^2/n$.

Equivalently, $x \in \mathcal{S}$, 1-d subspace spanned by all-ones vector.

- $\widehat{x} = \mathbf{proj}_{\mathcal{S}}(y)$, and $R(\widehat{x}, x) = \sigma^2/n$ since $\mathbf{proj}_{\mathcal{S}}(\sigma\xi) \sim \mathcal{N}(0, \sigma^2)$.

Works for **any subspace**! Suppose $x \in \mathcal{S}$ of dimension s.

- As before, take $\widehat{x} = \mathbf{proj}_{\mathcal{S}}(y)$, then

$$R(\widehat{x}, x) = \frac{\sigma^2 s}{n}.$$

Optimal risk up to a constant!

Optimality of linear estimators

When $x \in \mathcal{S}$, there exists a linear $\widehat{x}_{\mathcal{S}}$ with a near-optimal risk.
$\widehat{x}_{\mathcal{S}}$ is easy to construct if $\mathcal{S}$ is known.

Optimality of linear estimators

When $x \in \mathcal{S}$, there exists a linear $\widehat{x}_{\mathcal{S}}$ with a near-optimal risk.
$\widehat{x}_{\mathcal{S}}$ is easy to construct if $\mathcal{S}$ is known.

For any $\mathcal{X} \subseteq \mathbb{R}^n$, define the minimax risk and the linear minimax risk:

$$\bar{R}(\mathcal{X}) := \inf_{\widehat{x}} \sup_{x \in \mathcal{X}} R(\widehat{x}, x) \quad \leq \quad \bar{R}^{\mathrm{lin}}(\mathcal{X}) := \inf_{\widehat{x} = \Phi(y)} \sup_{x \in \mathcal{X}} R(\widehat{x}, x).$$

Optimality of linear estimators

When $x \in \mathcal{S}$, there exists a linear $\widehat{x}_{\mathcal{S}}$ with a near-optimal risk.
$\widehat{x}_{\mathcal{S}}$ is easy to construct if $\mathcal{S}$ is known.

For any $\mathcal{X} \subseteq \mathbb{R}^n$, define the minimax risk and the linear minimax risk:

$$\bar{R}(\mathcal{X}) := \inf_{\widehat{x}} \sup_{x \in \mathcal{X}} R(\widehat{x}, x) \quad \leq \quad \bar{R}^{\text{lin}}(\mathcal{X}) := \inf_{\widehat{x} = \Phi(y)} \sup_{x \in \mathcal{X}} R(\widehat{x}, x).$$

When $\mathcal{X}$ is a subspace, $\bar{R}^{\text{lin}}(\mathcal{X}) \leq c\bar{R}(\mathcal{X}) \Rightarrow$

we can search for a near-optimal estimator $\widehat{x}^o$ among the linear ones!

- Donoho (1990): the above holds with $c = 1.2$ for quadratically convex and orthosymmetric sets, for example, ellipsoids.
- Juditsky & Nemirovski (2016): if $\mathcal{X}$ is known,

$\widehat{x}^o$ can be computed by convex optimization!

Adaptive estimation

If "good" $\mathcal{X}$ is unknown, $\widehat{x}^o$ still exists, but not accessible directly.

- For example, $x \in \{\mathcal{X}_\alpha\}$, large family of "good" sets (subspaces).

Question: *Is it possible to "mimick" $\widehat{x}^o$, i.e. construct an adaptive estimator $\widehat{x} = \widehat{x}(y)$ with a comparable risk?*

Adaptive estimation

If "good" $\mathcal{X}$ is unknown, $\widehat{x}^o$ still exists, but not accessible directly.

- For example, $x \in \{\mathcal{X}_\alpha\}$, large family of "good" sets (subspaces).

Question: *Is it possible to "mimick" $\widehat{x}^o$, i.e. construct an adaptive estimator $\widehat{x} = \widehat{x}(y)$ with a comparable risk?*

- Adaptive estimator $\widehat{x}$ approaches $R(\widehat{x}^o, x)$ without knowing x:

$$R(\widehat{x}, x) \approx R(\widehat{x}^o, x).$$

- We hope to find such $\widehat{x}$ by a data-driven (and efficient) search over a class of linear estimators.

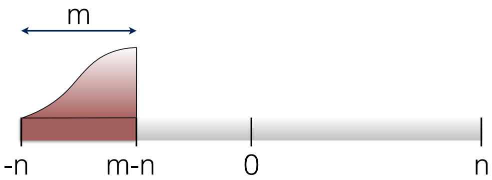
Filters

In signal processing, we usually assume time-invariance of some kind. Recall that we estimate the signal on the regular grid:

$$y_t = x_t + \sigma\xi_t, \quad t \in \{-n, ..., 0, ..., n\}.$$

- Consider time-invariant linear estimators: convolution of y with a filter $\varphi \in B_m = \{$ "vanish outside $[0, m]$ for some $m \leq n$" $\}$:

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



Filters

In signal processing, we usually assume time-invariance of some kind. Recall that we estimate the signal on the regular grid:

$$y_t = x_t + \sigma\xi_t, \quad t \in \{-n, \ldots, 0, \ldots, n\}.$$

- Consider time-invariant linear estimators: convolution of y with a filter $\varphi \in B_m = \{$ "vanish outside $[0, m]$ for some $m \leq n$" $\}$:

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

- **Goal:** recovery on $[0, n]$ via previous observations, with the risk

$$R_n(\varphi, x) := \frac{1}{n}\mathbb{E}\left[\left\| [x - \varphi * y]_0^n \right\|_2^2\right],$$

where $[x]_a^b = [x_a, \ldots, x_b]$.

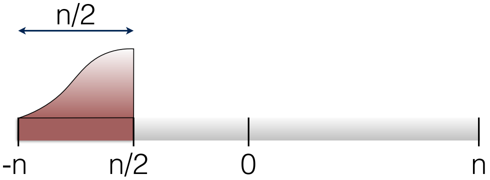
Main assumption: LTI recoverability

We assume that the class of linear filtering estimators is powerful.

Definition. x is ϱ-recoverable if there exists a $\phi^o \in B_{n/2}$ satisfying

$$R_n(\phi^o, x) \le \frac{\sigma^2 \varrho}{n}.$$

Adaptive signal denoising: find $\widehat{\varphi} = \widehat{\varphi}(y)$ s.t. $R_n(\widehat{\varphi}, x) \approx R_n(\phi^o, x)$.



Main assumption: LTI recoverability

We assume that the class of linear filtering estimators is powerful.

Definition. x is ϱ-recoverable if there exists a $\phi^o \in B_{n/2}$ satisfying

$$R_n(\phi^o, x) \leq \frac{\sigma^2 \varrho}{n}.$$

Adaptive signal denoising: find $\widehat{\varphi} = \widehat{\varphi}(y)$ s.t. $R_n(\widehat{\varphi}, x) \approx R_n(\phi^o, x)$.

Bias-variance decomposition:

$$\frac{1}{n}\mathbb{E}\left[\| [x - \phi^o * y]_0^n \|_2^2\right] = \frac{1}{n}\| [x - \phi^o * x]_0^n \|_2^2 + \frac{\sigma^2}{n}\mathbb{E}\left[\| [\phi^o * \xi]_0^n \|_2^2\right]$$

- reproduction of the signal: $\frac{1}{n}\left[\| [x - \phi^o * x]_0^n \|_2^2\right] \leq \frac{\sigma^2 \varrho}{n}$,
- small ℓ_2-norm of the oracle: $\|\phi^o\|_2^2 \leq \frac{\varrho}{n}$.

Adaptive estimator

Let $\mathcal{F}$ be the Discrete Fourier transform operator on $[0, n]$:

$$\mathcal{F}_{j\tau} = \frac{1}{\sqrt{n+1}} \exp\left(\frac{2\pi i j \tau}{n+1}\right), \quad 0 \le j, \tau \le n.$$

We propose an adaptive estimator: $\widehat{x} = \widehat{\varphi} * y$, where $\widehat{\varphi} \in B_n$ is

$$\widehat{\varphi} \in \operatorname*{argmin}_{\varphi \in B_n} \left\{ \underbrace{\|[y - \varphi * y]_0^n\|_2^2}_{\text{sample analogue of } R_n(\phi^o, x)} \quad : \quad \underbrace{\|\mathcal{F}\varphi\|_1 \le \varrho/\sqrt{n}}_{\text{regularization of the filter}} \right\}$$

Compare with the spectral Lasso:

$$\widehat{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \|[y - x]_0^n\|_2^2 \quad : \quad \|\mathcal{F}x\|_1 \le \|\mathcal{F}x^o\|_1 \right\}.$$

- No sparsity. The “dictionary matrix” Y s.t. $\varphi * y = Y(\mathcal{F}\varphi)$ is not RIP and scales differently with σ. Standard techniques fail.

Statistical bound

Recall ϱ-recoverability of x: there exists a $\phi^o \in B_{n/2}$ such that

$$R_n(\phi^o, x) \leq \frac{\sigma^2 \varrho}{n}.$$

Statistical bound

Recall ϱ-recoverability of x: there exists a $\phi^o \in B_{n/2}$ such that

$$R_n(\phi^o, x) \le \frac{\sigma^2 \varrho}{n}.$$

Theorem (Main Result)
If x is ϱ-recoverable, the filter $\widehat{\varphi}$ satisfies

$$R_n(\widehat{\varphi}, x) \le \frac{\sigma^2 \varrho}{n}(\varrho + \log n).$$

(actually a bound w.h.p.)

Price of adaptation is $\varrho \Rightarrow$ we would like ϱ to be as small as possible.

Statistical bound: naive approach

- There exists a $\phi^o \in B_{n/2}$ for which $\|\phi^o\|_2^2 \le \frac{\varrho}{n}$, $R_n(\phi^o, x) \le \frac{\sigma^2 \varrho}{n}$.
- Suppose that ϱ is known, and search for ϕ^o:

$$\widehat{\phi} \in \underset{\phi \in B_{n/2}}{\operatorname{argmin}} \left\{ \frac{1}{n} \|[y - \phi * y]_0^n\|_2^2 : \ \|\phi\|_2^2 \le \frac{\varrho}{n} \right\}.$$

- ϕ^o is feasible, so that

$$\frac{1}{n}\|y - \widehat{\phi} * y\|_2^2 \le \frac{1}{n}\|y - \phi^o * y\|_2^2 = R_n(\phi^o, x) + \frac{\sigma^2}{n}\|\xi\|_2^2 + \langle \ldots \rangle.$$

- OK at this step: $Q_n(\phi^o, x)$ is small, $\sigma^2\|\xi\|_2^2$ subtracted. **But:**

$$\frac{1}{n}\|x - \widehat{\phi} * y\|_2^2 = \underbrace{\frac{1}{n}\|y - \widehat{\phi} * y\|_2^2 - \frac{\sigma^2}{n}\|\xi\|_2^2}_{R_n(\phi^o, x)} + \langle \ldots \rangle + \frac{2\sigma^2}{n}\langle \xi, \widehat{\phi} * \xi \rangle.$$

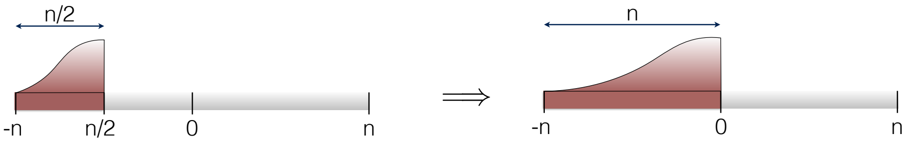
ℓ_2-constraint too weak to control $\langle \xi, \widehat{\phi} * \xi \rangle$ because $\widehat{\phi}$ is random.

Statistical bound: key insight

- There exists a $\phi^o \in B_{n/2}$ for which $\|\phi^o\|_2^2 \le \frac{\varrho}{n}$, $R_n(\phi^o, x) \le \frac{\sigma^2\varrho}{n}$.
- Instead of ϕ^o, let's mimick $\varphi^o := (\phi^o * \phi^o) \in B_n$. Can show:

$$\|\mathcal{F}\varphi^o\|_1^2 \le \frac{\varrho^2}{n},$$

$$R_n(\varphi^o, x) \le \frac{\sigma^2\varrho^2}{n}.$$



Statistical bound: key insight

- There exists a $\phi^o \in B_{n/2}$ for which $\|\phi^o\|_2^2 \le \frac{\varrho}{n}$, $R_n(\phi^o, x) \le \frac{\sigma^2 \varrho}{n}$.
- Instead of ϕ^o, let's mimick $\varphi^o := (\phi^o * \phi^o) \in B_n$. Can show:

$$\|\mathcal{F}\varphi^o\|_1^2 \le \frac{\varrho^2}{n},$$

$$R_n(\varphi^o, x) \le \frac{\sigma^2 \varrho^2}{n}.$$

- Pay an extra ϱ, but obtain a bound on the ℓ_1-norm (in Fourier).
- Problem term $\langle \xi, \widehat{\varphi} * \xi \rangle$: uniform bound + extreme points.
- Adaptive estimator $\widehat{\varphi}$ can be formulated as

$$\widehat{\varphi} \in \underset{\varphi \in B_n}{\operatorname{argmin}} \left\{ \frac{1}{n} \|[y - \varphi * y]_0^n\|_2^2 : \ \|\mathcal{F}\varphi\|_1 \le \frac{\varrho}{\sqrt{n}} \right\}$$

or the penalized problem (useful when ϱ is unknown).

Time-invariant subspace assumption

Definition. Subspace $\mathcal{S}$ of the space of sequences $(\ldots, x_{-1}, x_0, x_1, \ldots)$ is called time-invariant if it is preserved under $x_t \mapsto x_{t-1}$.

Time-invariant subspace assumption

Definition. Subspace $\mathcal{S}$ of the space of sequences $(\ldots, x_{-1}, x_0, x_1, \ldots)$ is called time-invariant if it is preserved under $x_t \mapsto x_{t-1}$.

Time-Invariant Subspace Assumption (TISA): x belongs to some time-invariant subspace of dimension $s \le n$.

TISA $\Leftrightarrow$ exp. polynomials. x satisfying TISA is an exponential polynomial of order s, with frequencies depending on $\mathcal{S}$.

Time-invariant subspace assumption

Definition. Subspace $\mathcal{S}$ of the space of sequences $(\dots, x_{-1}, x_0, x_1, \dots)$ is called time-invariant if it is preserved under $x_t \mapsto x_{t-1}$.

Time-Invariant Subspace Assumption (TISA): x belongs to some time-invariant subspace of dimension $s \le n$.

TISA $\Leftrightarrow$ exp. polynomials. x satisfying TISA is an exponential polynomial of order s, with frequencies depending on $\mathcal{S}$.

- **Example:** harmonic oscillation

$$x_t = \sum_{k=1}^{s} C_k e^{\imath \omega_k t}, \quad \tau \in \mathbb{Z}.$$

Time-invariant subspace assumption (cont.)

Theorem

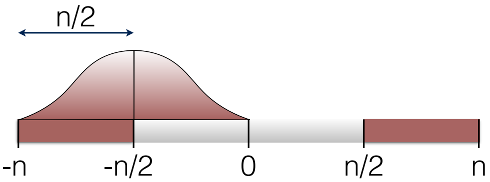
Let x satisfy TISA with some $s \leq n$.
Then, x is ϱ-recoverable with $\varrho = s^2 \log n$.

Time-invariant subspace assumption (cont.)

Theorem

Let x satisfy TISA with some $s \leq n$.
Then, x is ϱ-recoverable with $\varrho = s^2 \log n$.

Lower bound: $\varrho(s) = s$. Achievable if we allow for bilateral filters:



Time-invariant subspace assumption (cont.)

Theorem

Let x satisfy TISA with some $s \leq n$.
Then, x is ϱ-recoverable with $\varrho = s^2 \log n$.

Lower bound: $\varrho(s) = s$. Achievable if we allow for bilateral filters:

Theorem

Let x satisfy TISA with some $s \leq n$.
Then, x is ϱ-recoverable, with respect to bilateral oracle, with $\varrho = s$.

Denoising harmonic oscillations

Goal: recover x on $[-n, n]$ when frequencies are unknown:

$$x_\tau = \sum_{k=1}^{s} C_k e^{\imath \omega_k \tau},$$

Atomic Soft Thresholding (Tang & Recht, 2012):

$$R_n \leq \frac{\sigma^2 s \log n}{n}$$

if frequencies are separated, but **slow rate** $\mathcal{O}(1/\sqrt{n})$ if not.

Adaptive filtering:

$$R_n \leq \frac{\sigma^2 s^4 \log^2 n}{n}$$

without any separation assumptions. s^4 improves to s^2:

- in the separated case via Beurling's majorant (Moitra, 2014).
- in the central zone $[-n/2, n/2]$ via bilateral filters.

Optimization problem

For some $r > 0$, we want to solve:

$$\mathrm{Opt} = \min_{\varphi \in \mathbb{C}^n} \left\{ f(\varphi) = \|y - y * \varphi\|_2^2 : \ \|\mathcal{F}_n \varphi\|_1 \le r \right\}. \qquad (P)$$

- Well-structured feasible set – ℓ_2/ℓ_1-norm ball, prox in $\mathcal{O}(n \log n)$.
- First-order oracle can be computed in $O(n \log n)$.
- Low-accuracy solutions: sufficient to find a solution $\tilde{\varphi}$ satisfying

$$\varepsilon(\tilde{\varphi}) := f(\tilde{\varphi}) - \mathrm{Opt} \lesssim \frac{1}{n} \mathrm{Opt}.$$

$\Rightarrow$ proximal gradient methods.

Change of variables

$$\mathrm{Opt} = \min_{\varphi \in \mathbb{C}^n} \left\{ f(\varphi) = \|y - y * \varphi\|_2^2 : \ \|\mathcal{F}_n \varphi\|_1 \le r \right\}. \qquad (P)$$

$u := \frac{\mathcal{F}_n(\varphi)}{r} \Rightarrow$ feasible set is the unit ball of the (complex) ℓ_1-norm.

$$\begin{aligned} y * \varphi &= y * \mathcal{F}_n^{-1}(r u) \\ &= \mathcal{F}_n^{-1} \left\{ \mathcal{F}_{3n} [y; 0_n] \bullet \mathcal{F}_{3n} \left[0_{2n}; \mathcal{F}_n^{-1}(r u) \right] \right\} = \mathcal{A} u, \end{aligned}$$

where $[x; 0_n]$ is the concatenation with the zero vector of length n, and $\bullet$ is the element-wise product. Computed in $\mathcal{O}(n \log n)$ by FFT.

$$\begin{aligned} f(\varphi) = F(u) &= \|y\|_2^2 - \langle y, \mathcal{A} u \rangle - \langle \mathcal{A} u, y \rangle + \langle u, \mathcal{A}^T \mathcal{A} u \rangle, \\ \nabla F(u) &= 2(-\mathcal{A}^T y + \mathcal{A}^T \mathcal{A} u) \end{aligned}$$

(everything is complex-valued, hiding some conjugates).

Proximal mapping

So, now (P) is reformulated as a well-structured optimization problem

$$\mathrm{Opt} = \min_{u \in \mathbb{C}^n} \left\{ F(u) : \ \|u\|_1 \le 1 \right\}, \qquad (P')$$

where we can compute $F(u)$ and $\nabla F(u)$ in $\mathcal{O}(n \log n)$.
We also must be able to compute the proximal mapping:

$$\mathrm{prox}_u(g) := \operatorname*{argmin}_{\|v\|_1 \le 1} \left\{ \langle g, v \rangle + D_u(v) \right\},$$

where

$$D_u(v) := \omega(v) - \omega(u) - \langle \nabla \omega(u), v - u \rangle$$

is the Bregman divergence, and $\omega(u)$ is a "good" proximal function: smooth, 1-strongly convex, with computable prox, and with a small

$$R^2 = \max_{\|u\|_1 \le 1} \omega(u).$$

Proximal functions

Euclidean prox:

$$\omega(u) = \frac{1}{2}\|u\|_2^2 \quad \Rightarrow \quad D_u(v) = \frac{1}{2}\|v - u\|_2^2.$$

Corresponding prox is Euclidean projection on the complex ℓ_1-ball.

- Computable in $\mathcal{O}(n \log n)$, $R^2 = \mathcal{O}(1)$.
- Smoothness measured in ℓ_2-norm.

"Suitable" prox:

$$\omega(u) = \gamma\|u\|_p^p, \quad p = 1 + \frac{1}{\ln n}, \quad \gamma = \frac{e \ln n}{p}.$$

- Computable in $\mathcal{O}(n \log n)$, $R^2 = \mathcal{O}(\log n)$.
- Smoothness measured in ℓ_q-norm, $q \approx \log n \Rightarrow \| \cdot \|_q \leq C\| \cdot \|_\infty$.

Solving the optimization problem

Let L be the Lipschitz constant of $\nabla F(u)$ (precomputed from data).

Fast Gradient Method (Nesterov & Nemirovski, 2013)
Initialization: $u_0 = \mathbf{0}$; $G_0 = \mathbf{0}$.
For $t = 0, 1, \ldots$ **do**

(a) $w_t = \mathrm{prox}_{\mathbf{0}}\left(\frac{G_t}{L}\right)$.

(b) $\tau_t := \frac{2(t+2)}{(t+1)(t+4)}$.

(c) $v_{t+1} := \tau_t w_t + (1-\tau_t)u_t$

(d) $\widehat{v}_{t+1} := \mathrm{prox}_{w_t}\left(\frac{t+2}{2}\frac{\nabla F(v_{t+1})}{L}\right)$.

(e) $u_{t+1} := \tau_t \widehat{v}_{t+1} + (1-\tau_t)u_t, \quad G_{t+1} := G_t + \frac{t+2}{2}\nabla F(v_{t+1})$

Similar to Fast Gradient Descent. Convergence guarantee:

$$F(u_t) - F^* \lesssim \frac{LR^2}{t^2}$$

Experiments

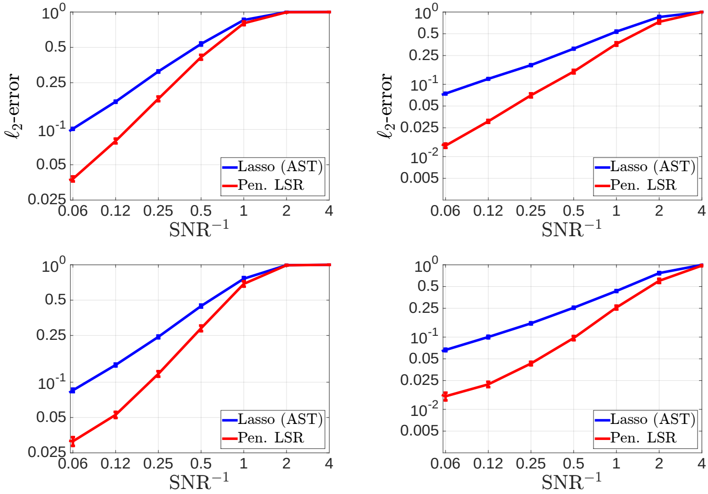
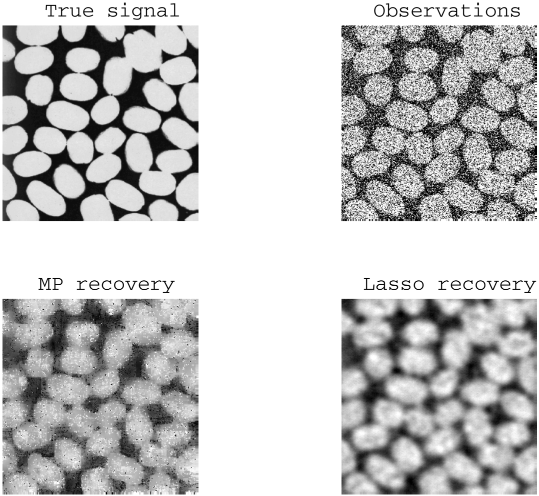


Figure: Signal and image denoising in different scenarios, 1-d (left) and 2-d (right).

Demonstration



Brodatz D75, SNR=1. Similar MSE, but Lasso tends to over-smooth.

Conclusion

We give an efficiently computable and statistically near-optimal construction of adaptive estimator for time-invariant signals.

Main idea: adaptation to the well-performing linear estimator.

As a consequence, we get fast rates of denoising harmonic oscillations without the frequency separation assumption.

Thank you for your attention!

Acknowledgements

Collaborators

Zaid Harchaoui
University of Washington

Anatoli Juditsky
Univ. Grenoble Alpes

Arkadi Nemirovski
Georgia Tech

Publications

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

Adaptive estimation: classical example

Suppose x is s-sparse, i.e. comes from $\mathcal{S}$ spanned by $\{e_{i_1}, \ldots, e_{i_s}\}$.

- Linear oracle $\widehat{x}^o = \mathrm{proj}_{\mathcal{S}}(y)$:

$$Q(\widehat{x}^o, x) = \frac{\sigma^2 s}{n}$$

- Soft-thresholding estimator (Lasso):

$$\widehat{x} = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \|x - y\|_2^2 + \lambda \|x\|_1 \right\}. \tag{1}$$

  If λ is well-chosen, $\widehat{x}$ is adaptive: not knowing $\mathcal{S}$, it satisfies

$$Q(\widehat{x}, x) \leq Q(\widehat{x}^o, x) \log(n),$$

- $\widehat{x}$ is non-linear but "looks" like a linear estimator, and can be computed by searching over linear estimators!
- Indeed, (1) is separable, and we can write $\widehat{x} = \widehat{\varphi} \cdot y$, where

$$\widehat{\varphi} = \underset{\varphi \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ f_y(\varphi) := \|y - y \cdot \varphi\|_2^2 + \lambda \|y \cdot \varphi\|_1 \right\}.$$

Better complexity estimate

After k iterations of FGM, we have for (P^2):

$$f^2(\varphi_k) \leq \mathrm{Opt}^2 + \frac{LR^2}{k^2}.$$

We get $\mathcal{O}(k^{-1})$ error for the initial problem (P):

$$f(\varphi_k) \leq \mathrm{Opt} + \frac{\sqrt{L}R}{k}.$$

Additional structure: since $\mathrm{Opt} \geq 0$,

$$f^2(\varphi_k) - \mathrm{Opt}^2 = (f(\varphi_k) - \mathrm{Opt})(f(\varphi_k) + \mathrm{Opt}) \geq 2\mathrm{Opt}(f(\varphi_k) - \mathrm{Opt}),$$

and we get an "optimistic" $\mathcal{O}(k^{-2})$ error provided that $\mathrm{Opt} > 0$:

$$f(\varphi_k) - \mathrm{Opt} \leq \frac{LR^2}{2\mathrm{Opt}\, k^2}$$