Adaptive filtering by convex optimization

Dmitry Ostrovsky*

joint research with A. Juditsky*, Z. Harchaoui† and A. Nemirovski‡
*University J. Fourier, †University of Washington, ‡Georgia Tech

Nov. 9, 2016
Paris

Problem Formulation

Motivation: linear estimators

We want to recover an unknown signal $x \in \mathbb{C}^{\mathcal{T}}$, $\mathcal{T}$ being a regular grid in $\mathbb{R}^d$, given noisy observations

$$y_\tau = x_\tau + \sigma\xi_\tau,\ \tau \in \mathcal{T}, \tag{1}$$

where ξ_τ's are the i.i.d. $\mathbb{C}\mathcal{N}(0,1)$.

Motivation: linear estimators

We want to recover an unknown signal $x \in \mathbb{C}^{\mathcal{T}}$, $\mathcal{T}$ being a regular grid in $\mathbb{R}^d$, given noisy observations

$$y_\tau = x_\tau + \sigma\xi_\tau,\ \tau \in \mathcal{T}, \tag{1}$$

where ξ_τ's are the i.i.d. $\mathbb{C}\mathcal{N}(0,1)$.

We may want to estimate the whole signal x, or the value x_t at some point $t \in \mathcal{T}$.

- There is a very general result on the optimality of linear estimators.

Motivation: linear estimators

We want to recover an unknown signal $x \in \mathbb{C}^{\mathcal{T}}$, $\mathcal{T}$ being a regular grid in $\mathbb{R}^d$, given noisy observations

$$y_\tau = x_\tau + \sigma\xi_\tau,\ \tau \in \mathcal{T}, \tag{1}$$

where ξ_τ's are the i.i.d. $\mathbb{C}\mathcal{N}(0,1)$.

We may want to estimate the whole signal x, or the value x_t at some point $t \in \mathcal{T}$.

- There is a very general result on the optimality of linear estimators.

Theorem [Ibragimov and Khas'minski 1984, Donoho 1990 & 1994]
Let $\mathcal{X} \subset \mathbb{C}^{\mathcal{T}}$ be compact, symmetric, and convex. The minimax, over $x \in \mathcal{X}$, risk of recovering x_t from (1) is attained, within factor 1.25, by a linear in y estimate, readily given, along with its risk, by a certain convex optimization problem.

Motivation: linear estimators (cont.)

In other words, if we are given a convex compact (and symmetric) set $\mathcal{X}$ of signals (e.g., set of signals satisfying some regularity constraints) then a properly selected linear estimator

$$x_t^* = \sum_{\tau \in \mathcal{T}} \varphi_\tau^* y_\tau, \quad \varphi^* \in \mathbb{C}^{\mathcal{T}},$$

is (quasi-) optimal on the class of all possible estimators.

- Computing the linear minimax estimator is "easy" for well-structured sets of signals (e.g., sets which can be described using `CVX`).

Motivation: linear estimators (cont.)

In other words, if we are given a convex compact (and symmetric) set $\mathcal{X}$ of signals (e.g., set of signals satisfying some regularity constraints) then a properly selected linear estimator

$$x_t^* = \sum_{\tau \in \mathcal{T}} \varphi_\tau^* y_\tau, \quad \varphi^* \in \mathbb{C}^{\mathcal{T}},$$

is (quasi-) optimal on the class of all possible estimators.

- Computing the linear minimax estimator is "easy" for well-structured sets of signals (e.g., sets which can be described using `CVX`).

Question:

Suppose that we do not know the set $\mathcal{X}$. Is it possible to "mimic" the oracle linear estimator φ^*, i.e. to construct an adaptive estimator (which only uses observations) of comparable accuracy?

Problem formulation

For the sake of simplicity, consider 1d situation, where the signal to recover $x \in \mathbb{C}^{\mathbb{Z}}$, and we are given $n = 4T + 1$ observations

$$y_\tau = x_\tau + \sigma\xi_\tau, \quad -2T < \tau < 2T,$$

Our objective may be either

- *filtering* – estimation of x_{2T} (or x_{-2T}),
- *interpolation* – estimation of x_t, $|t| < 2T$ (e.g., x_0),
- *prediction* – estimation of x_{2T+h}, (or x_{-2T-h}) for some "horizon" $h > 0$.

We will consider time-invariant oracle estimators with bounded support, represented as "linear filters" φ^* vanishing outside $[-L, L]$ for some $L \leq 2T$. As such, one can estimate x_t on $|t| \leq 2T - L$:

$$x_t^* = \sum_{|\tau| \leq L} \varphi_\tau^* y_{t-\tau} = [\varphi^* * y]_t.$$

We write $\varphi \in \mathbb{C}_L$ if φ vanishes outside $[-L, L]$.

Recoverable signals

For the sake of simplicity, let us assume that we want to estimate x_0.

Definition. We say that signal x is (T, ρ)-recoverable at $t = 0$ if there exists (an oracle) filter $\varphi^* \in \mathbb{C}_{T/2}$ which has a small uniform error of recovering x_τ on $[\frac{-3T}{2}, \frac{3T}{2}]$:

$$\mathbf{E}^{1/2}|x_\tau - [\varphi^* * y]_\tau|^2 \le \frac{\sigma\rho}{\sqrt{T}}, \quad |\tau| \le \frac{3T}{2}.$$

Recoverable signals

For the sake of simplicity, let us assume that we want to estimate x_0.

Definition. We say that signal x is (T, ρ)-recoverable at $t = 0$ if there exists (an oracle) filter $\varphi^* \in \mathbb{C}_{T/2}$ which has a small uniform error of recovering x_τ on $[\frac{-3T}{2}, \frac{3T}{2}]$:

$$\mathbf{E}^{1/2}|x_\tau - [\varphi^* * y]_\tau|^2 \le \frac{\sigma\rho}{\sqrt{T}}, \quad |\tau| \le \frac{3T}{2}.$$

This has two consequences:

- *small ℓ_2-norm of the oracle:* $\|\varphi^*\|_2 \le \frac{\rho}{\sqrt{T}}$;
- *reproduction on* $|\tau| \le \frac{3T}{2}$*:* $|x_\tau - [\varphi^* * x]_\tau| \le \frac{\sigma\rho}{\sqrt{T}}$.

Moreover, if the two conditions above are met, the signal is $(T, \sqrt{2}\rho)$-recoverable. Indeed,

$$\begin{aligned}\mathbf{E}|x_\tau - [\varphi^* * (x + \sigma\xi)]_\tau|^2 &= |x_\tau - [\varphi^* * x]_\tau|^2 + \sigma^2\mathbf{E}|[\varphi^* * \xi]_\tau|^2 \\ &= |x_\tau - [\varphi^* * x]_\tau|^2 + \sigma^2\|\varphi^*\|_2^2.\end{aligned}$$

Recoverable signals

For the sake of simplicity, let us assume that we want to estimate x_0.

Definition. We say that signal x is (T, ρ)-recoverable at $t = 0$ if there exists (an oracle) filter $\varphi^* \in \mathbb{C}_{T/2}$ which has a small uniform error of recovering x_τ on $[\frac{-3T}{2}, \frac{3T}{2}]$:

$$\mathbf{E}^{1/2}|x_\tau - [\varphi^* * y]_\tau|^2 \le \frac{\sigma\rho}{\sqrt{T}}, \quad |\tau| \le \frac{3T}{2}.$$

This has two consequences:

- *small ℓ_2-norm of the oracle:* $\|\varphi^*\|_2 \le \frac{\rho}{\sqrt{T}}$;
- *reproduction on* $|\tau| \le \frac{3T}{2}$*:* $|x_\tau - [\varphi^* * x]_\tau| \le \frac{\sigma\rho}{\sqrt{T}}$.

Moreover, if the two conditions above are met, the signal is $(T, \sqrt{2}\rho)$-recoverable. Indeed,

$$\begin{aligned}\mathbf{E}|x_\tau - [\varphi^* * (x + \sigma\xi)]_\tau|^2 &= |x_\tau - [\varphi^* * x]_\tau|^2 + \sigma^2\mathbf{E}|[\varphi^* * \xi]_\tau|^2 \\ &= |x_\tau - [\varphi^* * x]_\tau|^2 + \sigma^2\|\varphi^*\|_2^2.\end{aligned}$$

More generally, for x which (T, ρ)-recoverable at t, there exists φ^* of length $O(T)$ and an $O(T)$-neighborhood of t where $\varphi^* * x$ reproduces x with a small error.

Classical example

Consider the problem of estimating a smooth function $f : [0,1] \to \mathbb{R}$ from noisy observations

$$y_i = f(i/n) + \sigma\xi_i, \quad i = 1, ..., n, \ \xi \sim \mathcal{N}(0, I_n).$$

The classical kernel estimator $\widehat{f}_t$ of $f(t)$ with bandwidth h is

$$\widehat{f}(t) = \sum_{i=1}^{n} \frac{1}{2nh} K\left(\frac{t - i/n}{h}\right) y_i,$$

and $K(t) : [-1,1] \to \mathbb{R}$ is a kernel such that

$$\int_{-1}^{1} K(t)dt = 1, \quad \int_{-1}^{1} K^2(t)dt = \rho^2 < \infty.$$

Let $x_\tau = f(\tau/n)$, $\tau = 1, ..., n$, and let $T = [2nh]$. Then, the kernel estimator above can be rewritten for $T/2 + 1 \le t \le n - T/2$ as

$$\widehat{x}_t = \widehat{f}(t/n) = (\phi * y)_t, \quad \phi_\tau = \frac{1}{T} K\left(\frac{\tau}{T/2}\right), \ \tau = -T/2, ..., T/2.$$

Note that for big enough T the ℓ_2-norm of ϕ satisfies $\|\phi\|_2 \sim \rho/\sqrt{T}$, and if the kernel K and the bandwidth h are "properly chosen", the bias of the estimator is also $c\rho/\sqrt{T}$.

Shift-Invariance Assumption

A set $\mathcal{S} \subseteq \mathbb{C}^{\mathbb{Z}}$ is called shift-invariant if it is preserved under the time shift $x_t \mapsto x_{t-1}$.

Shift-Invariance Assumption (SIA): x is close to an (unknown) shift-invariant subspace $\mathcal{S} \subset \mathbb{C}^{\mathbb{Z}}$ of dimension $s = \dim(\mathcal{S}) \leq T$. Specifically, $x = x^{\mathcal{S}} + \delta$, where $x^{\mathcal{S}} \in \mathcal{S}$, and

$$\left\| [\delta_{-2T};\, \delta_{-2T+1};\, \ldots;\, \delta_{2T}] \right\|_2 \leq \kappa\sigma$$

Shift-Invariance Assumption

A set $\mathcal{S} \subseteq \mathbb{C}^{\mathbb{Z}}$ is called shift-invariant if it is preserved under the time shift $x_t \mapsto x_{t-1}$.

Shift-Invariance Assumption (SIA): x is close to an (unknown) shift-invariant subspace $\mathcal{S} \subset \mathbb{C}^{\mathbb{Z}}$ of dimension $s = \dim(\mathcal{S}) \leq T$. Specifically, $x = x^{\mathcal{S}} + \delta$, where $x^{\mathcal{S}} \in \mathcal{S}$, and

$$\left\| [\delta_{-2T};\, \delta_{-2T+1};\, \ldots;\, \delta_{2T}] \right\|_2 \leq \kappa\sigma$$

SIA $\Leftrightarrow$ exp. polynomials. x satisfying SIA approximately solves some homogeneous linear difference equation $\sum_{\tau=0}^{s} w_\tau x_{t-\tau} \equiv 0$, whose exact solutions are exponential polynomials.

Shift-Invariance Assumption

A set $\mathcal{S} \subseteq \mathbb{C}^{\mathbb{Z}}$ is called shift-invariant if it is preserved under the time shift $x_t \mapsto x_{t-1}$.

Shift-Invariance Assumption (SIA): x is close to an (unknown) shift-invariant subspace $\mathcal{S} \subset \mathbb{C}^{\mathbb{Z}}$ of dimension $s = \dim(\mathcal{S}) \leq T$. Specifically, $x = x^{\mathcal{S}} + \delta$, where $x^{\mathcal{S}} \in \mathcal{S}$, and

$$\left\| [\delta_{-2T};\, \delta_{-2T+1};\, \ldots;\, \delta_{2T}] \right\|_2 \leq \kappa\sigma$$

SIA $\Leftrightarrow$ exp. polynomials. x satisfying SIA approximately solves some homogeneous linear difference equation $\sum_{\tau=0}^{s} w_\tau x_{t-\tau} \equiv 0$, whose exact solutions are exponential polynomials.

SIA $\Rightarrow$ recoverable. x satisfying SIA is (T, ρ)-recoverable at $t = 0$ with $\rho = (1+\kappa)\sqrt{s}$. That is, there exists $\varphi^* \in \mathbb{C}_{T/2}$ such that $\|\varphi^*\|_2 \leq \sqrt{\frac{s}{T}}$, and on $|\tau| \leq \frac{3T}{2}$ one has

$$|x_\tau - [\varphi^* * x]_\tau| \leq \frac{\kappa\sigma\sqrt{s}}{\sqrt{T}}.$$

Shift-Invariance Assumption

A set $\mathcal{S} \subseteq \mathbb{C}^{\mathbb{Z}}$ is called shift-invariant if it is preserved under the time shift $x_t \mapsto x_{t-1}$.

Shift-Invariance Assumption (SIA): x is close to an (unknown) shift-invariant subspace $\mathcal{S} \subset \mathbb{C}^{\mathbb{Z}}$ of dimension $s = \dim(\mathcal{S}) \leq T$. Specifically, $x = x^{\mathcal{S}} + \delta$, where $x^{\mathcal{S}} \in \mathcal{S}$, and

$$\left\| [\delta_{-2T};\, \delta_{-2T+1};\, \ldots;\, \delta_{2T}] \right\|_2 \leq \kappa\sigma$$

SIA $\Leftrightarrow$ exp. polynomials. x satisfying SIA approximately solves some homogeneous linear difference equation $\sum_{\tau=0}^{s} w_\tau x_{t-\tau} \equiv 0$, whose exact solutions are exponential polynomials.

SIA $\Rightarrow$ recoverable. x satisfying SIA is (T, ρ)-recoverable at $t = 0$ with $\rho = (1+\kappa)\sqrt{s}$. That is, there exists $\varphi^* \in \mathbb{C}_{T/2}$ such that $\|\varphi^*\|_2 \leq \sqrt{\frac{s}{T}}$, and on $|\tau| \leq \frac{3T}{2}$ one has

$$|x_\tau - [\varphi^* * x]_\tau| \leq \frac{\kappa\sigma\sqrt{s}}{\sqrt{T}}.$$

Proof. φ^* can be easily constructed from the projector on $\mathcal{S}$. Then $x^{\mathcal{S}} - \varphi^* * x^{\mathcal{S}} = 0$, and $[\delta - \varphi^* * \delta]_\tau$ is controlled by Cauchy-Schwartz. □

Results

Main result

We saw that under SIA there exists an unknown $\varphi^* \in \mathbb{C}_{T/2}$ with $\|\varphi^*\|_2 \leq \frac{\rho}{\sqrt{T}}$ such that

$$\mathbf{E}^{1/2}|x_\tau - [\varphi^* * x]_\tau|^2 \leq \frac{\sigma\rho}{\sqrt{T}}, \quad |\tau| \leq \frac{3T}{2},$$

for a moderate $\rho = (1+\kappa)\sqrt{s}$.

Question:

Recall that our goal is to estimate x_0. Is it possible, under SIA, to design an estimator $\widehat{x}_0 = [\widehat{\varphi} * y]_0$ of x_0, which only relies on the observations $[y_{-2T}, ..., y_{2T}]$, and such that

$$\mathbf{E}^{1/2}|\widehat{x}_0 - x_0|^2 \asymp \frac{\sigma\rho}{\sqrt{T}}, \qquad \rho := (1+\kappa)\sqrt{s} \qquad (?)$$

Main result

We saw that under SIA there exists an unknown $\varphi^* \in \mathbb{C}_{T/2}$ with $\|\varphi^*\|_2 \leq \frac{\rho}{\sqrt{T}}$ such that

$$\mathbf{E}^{1/2}|x_\tau - [\varphi^* * x]_\tau|^2 \leq \frac{\sigma\rho}{\sqrt{T}}, \quad |\tau| \leq \frac{3T}{2},$$

for a moderate $\rho = (1+\kappa)\sqrt{s}$.

Question:

Recall that our goal is to estimate x_0. Is it possible, under SIA, to design an estimator $\widehat{x}_0 = [\widehat{\varphi} * y]_0$ of x_0, which only relies on the observations $[y_{-2T}, ..., y_{2T}]$, and such that

$$\mathbf{E}^{1/2}|\widehat{x}_0 - x_0|^2 \asymp \frac{\sigma\rho}{\sqrt{T}}, \qquad \rho := (1+\kappa)\sqrt{s} \qquad (?)$$

Theorem 1

Under SIA, there is an estimator $\widehat{x}_0$ of x_0, given by $\widehat{x}_0 = [\widehat{\varphi}(y) * y]_0$, such that

$$|\widehat{x}_0 - x_0| \leq \frac{C\sigma}{\sqrt{T}}\left(\rho^3 + \rho^2\sqrt{\log T}\right) = \frac{C\sigma\rho}{\sqrt{T}}\left(\rho^2 + \rho\sqrt{\log T}\right)$$

with exponentially high probability.

Constructing the adaptive filter: naive approach

For a signal $x \in \mathbb{C}^{\mathbb{Z}}$, $L \in \mathbb{N}$, and $1 \leq p \leq \infty$, denote

$$\|x\|_{L,p} := \left\| [x]_{-L}^{L} \right\|_p.$$

Empirical Risk Minimization: define $\widehat{\varphi}$ as an optimal solution to

$$\min_{\varphi \in \mathbb{C}_T} \left\{ \|y - \varphi * y\|_{3T/2,2}^2 : \ \|\varphi\|_2 \leq \frac{\rho}{\sqrt{T}} \right\}.$$

Constructing the adaptive filter: naive approach

For a signal $x \in \mathbb{C}^{\mathbb{Z}}$, $L \in \mathbb{N}$, and $1 \le p \le \infty$, denote

$$\|x\|_{L,p} := \left\| [x]_{-L}^{L} \right\|_p.$$

Empirical Risk Minimization: define $\widehat{\varphi}$ as an optimal solution to

$$\min_{\varphi \in \mathbb{C}_T} \left\{ \|y - \varphi * y\|_{3T/2,2}^2 : \ \|\varphi\|_2 \le \frac{\rho}{\sqrt{T}} \right\}.$$

Note that φ^* is feasible, so that

$$\|y - \widehat{\varphi} * y\|_{3T/2,2}^2 \le \|y - \varphi^* * y\|_{3T/2,2}^2 = \mathcal{O}_{\mathbb{P}}(\sigma^2\rho^2) + \sigma^2 \|\xi\|_{3T/2,2}^2.$$

Therefore,

$$\begin{aligned} \|x - \widehat{\varphi} * y\|_{3T/2,2}^2 &= \|y - \widehat{\varphi} * y\|_{3T/2,2}^2 - \sigma^2 \|\xi\|_{3T/2,2}^2 - 2\sigma \langle \xi, x - \widehat{\varphi} * y \rangle_{3T/2} \\ &= \mathcal{O}_{\mathbb{P}}(\sigma^2\rho^2) + \underbrace{2\sigma^2 \langle \xi, \widehat{\varphi} * \xi \rangle_{3T/2}}_{\asymp \sqrt{T}} - 2\sigma \langle \xi, x - \widehat{\varphi} * x \rangle_{3T/2}. \end{aligned}$$

Auto-convolution

For $x \in \mathbb{C}^{\mathbb{Z}}$, let $F_T(x)$ be the Discrete Fourier Transform of $[x]_{-T}^{T}$.
We denote $\|x\|_{T,p}^{*} = \|F_T x\|_p$.

Lemma

Suppose that the x is (T, ρ)-filtered by an oracle $\varphi^* \in \mathbb{C}_{T/2}$ with $\|\varphi^*\| \leq \sqrt{\frac{s}{T}}$. Define

$$\varphi^o := (\varphi^* * \varphi^*) \in \mathbb{C}_T.$$

Then for φ^o it holds

$$\|\varphi^o\|_2 = \|\varphi^o\|_{T,2}^{*} \leq \|\varphi^o\|_{T,1}^{*} \leq \frac{\sqrt{2}\rho^2}{\sqrt{T}};$$

$$|x_\tau - [\varphi^o * y]_\tau| \leq \frac{c\sigma\rho^2}{\sqrt{T}}, \quad |\tau| \leq T.$$

Constructing the adaptive filter: correct approach

Let $\widehat{\varphi} \in \mathbb{C}_T$ be an optimal solution of the following problem:

$$\min_{\varphi \in \mathbb{C}_T} \left\{ \|y - \varphi * y\|_{T,2} : \ \|\varphi\|^*_{T,1} \leq \frac{\sqrt{2}\rho^2}{\sqrt{T}} \right\}. \qquad \text{(Con-}\ell_2\text{)}$$

Then, as before, by the feasibility of φ^o

$$\|y - \widehat{\varphi} * y\|^2_{T,2} \leq \|y - \varphi^o * y\|^2_{T,2} \leq \mathcal{O}_{\mathbb{P}}(\sigma^2\rho^4) + \sigma^2\|\xi\|^2_{T,2}$$

- We have now better control over the cross-term $\langle \xi, \widehat{\varphi} * \xi \rangle_T$ ("almost" the max of a convex function over a convex polyhedron):

$$\langle \xi, \widehat{\varphi} * \xi \rangle_T \leq \max_{\|\varphi\|^*_{T,1} \leq \rho^2\sqrt{2/T}} \langle \xi, \varphi * \xi \rangle_T = \mathcal{O}_{\mathbb{P}}\big(\rho^2 \log T\big).$$

- We also use SIA directly to curb the remaining terms $\langle ... \rangle$

Constructing the adaptive filter: correct approach

Let $\widehat{\varphi} \in \mathbb{C}_T$ be an optimal solution of the following problem:

$$\min_{\varphi \in \mathbb{C}_T} \left\{ \|y - \varphi * y\|_{T,2} : \ \|\varphi\|_{T,1}^* \leq \frac{\sqrt{2}\rho^2}{\sqrt{T}} \right\}. \qquad \text{(Con-}\ell_2\text{)}$$

Then, as before, by the feasibility of φ^o

$$\|y - \widehat{\varphi} * y\|_{T,2}^2 \leq \|y - \varphi^o * y\|_{T,2}^2 \leq \mathcal{O}_{\mathbb{P}}(\sigma^2\rho^4) + \sigma^2\|\xi\|_{T,2}^2$$

- We have now better control over the cross-term $\langle \xi, \widehat{\varphi} * \xi \rangle_T$ ("almost" the max of a convex function over a convex polyhedron):

$$\langle \xi, \widehat{\varphi} * \xi \rangle_T \leq \max_{\|\varphi\|_{T,1}^* \leq \rho^2\sqrt{2/T}} \langle \xi, \varphi * \xi \rangle_T = \mathcal{O}_{\mathbb{P}}\big(\rho^2 \log T\big).$$

- We also use SIA directly to curb the remaining terms $\langle ... \rangle$
- We finally get

$$\|x - [\widehat{\varphi} * y]\|_{T,2} \leq \mathcal{O}_{\mathbb{P}}\left(\sigma\rho^2 + \rho\sqrt{\log T}\right),$$

and then

$$|x_0 - [\widehat{\varphi} * y]_0| \leq \mathcal{O}_{\mathbb{P}}\left(\frac{\sigma\rho^3 + \rho^2\sqrt{\log T}}{\sqrt{T}}\right).$$

Alternative

Let $\widehat{\varphi}$ be an optimal solution to

$$\min_{\varphi \in \mathbb{C}_T} \left\{ \|y - \varphi * y\|^*_{T,\infty} : \ \|\varphi\|^*_{T,1} \leq \frac{\sqrt{2}\rho^2}{\sqrt{T}} \right\} \qquad \text{(Con-}\ell^*_\infty\text{)}$$

Theorem 2

Suppose that x is (T, ρ)-recoverable. For $\widehat{x}_0 = [\widehat{\varphi} * y]_0$ with $\widehat{\varphi}$ given by (Con-ℓ^*_∞) it holds w.h.p.

$$|\widehat{x}_0 - x_0| \leq \frac{C\sigma\rho^4\sqrt{\log T}}{\sqrt{T}}.$$

Remarks:

- Higher price of adaptation than for (Con-ℓ_2): $\rho^3\sqrt{\log T}$ versus $\rho^2 + \rho\sqrt{\log T}$.
- SIA is not needed, recoverability is enough.

Summary

- Let (x_τ) satisfy SIA with some known s and $\kappa(T)$, and let the "bandwidth" T be chosen such that $\kappa(T) \lesssim 1$.
  Then there exists an estimate $x^*_\tau = [\varphi^* * y]_\tau$ with bandwidth $T/2$ such that

$$\max_{\tau:|\tau-t|\le 3T/2} \mathbf{E}^{1/2} |x_\tau - x^*_\tau|^2 \le \frac{\sigma\rho}{\sqrt{T}}$$

  for some known moderate $\rho \ge 1$.
- Our objective is, assuming that T and ρ are known, to recover x_t from observations $[y]_{t-2T}^{t+2T}$ nearly as well as if we were using this hypothetic estimate x^*_t.

Summary

- Let (x_τ) satisfy SIA with some known s and $\kappa(T)$, and let the "bandwidth" T be chosen such that $\kappa(T) \lesssim 1$.
  Then there exists an estimate $x^*_\tau = [\varphi^* * y]_\tau$ with bandwidth $T/2$ such that

$$\max_{\tau:|\tau-t|\le 3T/2} \mathbf{E}^{1/2} |x_\tau - x^*_\tau|^2 \le \frac{\sigma\rho}{\sqrt{T}}$$

  for some known moderate $\rho \ge 1$.

- Our objective is, assuming that T and ρ are known, to recover x_t from observations $[y]_{t-2T}^{t+2T}$ nearly as well as if we were using this hypothetic estimate x^*_t.

- This can be achieved by solving either (Con-ℓ_2) or (Con-ℓ^*_∞). The error of recovery $\widehat{x}_t = [\widehat{\varphi} * y]_t$ is bounded, respectively, by

$$\underbrace{\mathcal{O}_{\mathbb{P}}\left(\frac{\sigma\rho}{\sqrt{T}}\left(\rho^2 + \rho\sqrt{\log T}\right)\right)}_{\text{when using (Con-}\ell_2\text{)}} \quad \text{or} \quad \underbrace{\mathcal{O}_{\mathbb{P}}\left(\frac{\sigma\rho}{\sqrt{T}}\left(\rho^3\sqrt{\log T}\right)\right)}_{\text{when using (Con-}\ell^*_\infty\text{)}}$$

Adaptation to ρ and T

In "practical applications", values of parameter ρ and of the bandwidth T are unknown.

- Instead of the constrained estimators, we can use more practical penalized ones:

$$\min_{\varphi\in\mathbb{C}_T}\left\{\|y-\varphi*y\|_{T,2}^2+\mu^2\sigma^2\sqrt{T}\|\varphi\|_{T,1}^*\right\}, \tag{Pen-ℓ_2^2}$$

$$\min_{\varphi\in\mathbb{C}_T}\left\{\|y-\varphi*y\|_{T,2}+\mu^2\sigma\|\varphi\|_{T,1}^*\right\}, \tag{Pen-ℓ_2}$$

$$\min_{\varphi\in\mathbb{C}_T}\left\{\|y-\varphi*y\|_{T,\infty}^*+\mu\sigma\sqrt{T}\|\varphi\|_{T,1}^*\right\}. \tag{Pen-ℓ_∞^*}$$

With $\mu\asymp\sqrt{\log T}$ all the bounds remain valid; penalized estimators are pivotal to ρ.

Adaptation to ρ and T

In "practical applications", values of parameter ρ and of the bandwidth T are unknown.

- Instead of the constrained estimators, we can use more practical penalized ones:

$$\min_{\varphi\in\mathbb{C}_T}\left\{\|y-\varphi*y\|_{T,2}^2+\mu^2\sigma^2\sqrt{T}\|\varphi\|_{T,1}^*\right\}, \tag{Pen-ℓ_2^2}$$

$$\min_{\varphi\in\mathbb{C}_T}\left\{\|y-\varphi*y\|_{T,2}+\mu^2\sigma\|\varphi\|_{T,1}^*\right\}, \tag{Pen-ℓ_2}$$

$$\min_{\varphi\in\mathbb{C}_T}\left\{\|y-\varphi*y\|_{T,\infty}^*+\mu\sigma\sqrt{T}\|\varphi\|_{T,1}^*\right\}. \tag{Pen-ℓ_∞^*}$$

  With $\mu\asymp\sqrt{\log T}$ all the bounds remain valid; penalized estimators are pivotal to ρ.

- To choose a proper T, we can use Lepski's algorithm, which amounts to compare estimators computed for various values of T.

Operational summary

When applying the proposed approach to "practical" recovery of a signal or an image

- For each point t of the grid:
  1. choose a set of bandwidths, e.g. $\{ T_0 = 1,\ T_1 = 2,\ ...,\ T_K = 2^K \}$;
  2. for each bandwidth T_k compute an approximate solution $\widehat{\varphi}_{T_k,t}$ to (Con-ℓ_2), ...;
  3. compute estimations $\widehat{x}_t[T_k] = [\widehat{\varphi}_{T_k,t} * y]_t$ and choose the "best" among them via Lepski's algorithm.
- To reduce the numerical cost, instead of proceeding point-wise, one can use block-wise update of filters, using the ℓ_2-bound of Theorem 1.

Optimization Tools

Reduction to a bilinear saddle-point problem 1

One needs to solve repeatedly problems (Con-ℓ_2) of the kind (or alike):

$$\mathrm{Opt} = \min_{\varphi \in \mathbb{C}_T} \left\{ f(\varphi) = \|y - y * \varphi\|^*_{T,p} : \ \|\varphi\|^*_{T,1} \leq r \right\}, \ r > 0, \ p \in \{2, \infty\}. \qquad (P)$$

Note that (P) can be rewritten as a bilinear saddle-point problem: indeed, its objective is

$$f(\varphi) = \max_{v \in \mathbb{C}^{2T+1}} \{ \langle F_T(y - y * \varphi), v \rangle, \ \|v\|_q \leq 1 \},$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

When denoting $u = F_T(\varphi)/r$,

$$\mathrm{Opt} = \min_{u \in \mathcal{B}_{T,1}} \max_{v \in \mathcal{B}_{T,q}} \langle \mathcal{A}u + b, v \rangle \qquad (SP)$$

where $q \in \{1, 2\}$, $b = F_T(y)$, $\mathcal{B}_{T,q}$ is the unit ball of the complex ℓ_q-norm in $\mathbb{C}^{2T+1}$, and

$$\begin{aligned} \mathcal{A}u &= F_T \left[y * F_T^{-1}(ru) \right] \\ &= F_T \left[F_{3T}^{-1} \left\{ F_{3T} \left[0_T; y; 0_T \right] \bullet F_{3T} \left[0_{2T}; F_T^{-1}(ru); 0_{2T} \right] \right\} \right] \end{aligned}$$

(here $[x; 0_T]$ stands for the concatenation with zero vector of length T and $\bullet$ is the Hadamard element-wise product). This is by the standard trick of representing the usual convolution via the circular one.

Reduction to a bilinear saddle-point problem 2

- (SP) is a bilinear saddle-point problem on the balls of either ℓ_2- or ℓ_2/ℓ_1-norm.
- Problems should be solved to (relatively) low accuracy – a solution $\widehat{u}$ of accuracy

$$\varepsilon(\widehat{u}) := f(\widehat{u}) - \mathrm{Opt} \leq \delta \mathrm{Opt}$$

  will be largely sufficient with $\delta \asymp \frac{1}{\sqrt{T}}$ for $q = 2$ and $\delta = \frac{1}{2}$ for $q = 1$.
- Objective gradients can be computed in $O(n \log n)$ operations using the FFT.

Reduction to a bilinear saddle-point problem 2

- (SP) is a bilinear saddle-point problem on the balls of either ℓ_2- or ℓ_2/ℓ_1-norm.
- Problems should be solved to (relatively) low accuracy – a solution $\widehat{u}$ of accuracy

$$\varepsilon(\widehat{u}) := f(\widehat{u}) - \mathrm{Opt} \leq \delta \mathrm{Opt}$$

  will be largely sufficient with $\delta \asymp \frac{1}{\sqrt{T}}$ for $q = 2$ and $\delta = \frac{1}{2}$ for $q = 1$.
- Objective gradients can be computed in $O(n \log n)$ operations using the FFT.

Under this premise, proximal first-order algorithms appear to be methods of choice.

Proximal algorithms for bilinear saddle-point optimization

- $1/\varepsilon$ complexity estimates (or even $1/\sqrt{\varepsilon}$ under "favorable circumstances").
- Accuracy certificates are available.
- Favorable geometry of the problem domain – simple $O(n)$ proximal computation.
- Fully profit from fast gradient computation – $O(n \log n)$ cost per iteration.

Proximal algorithms for bilinear saddle-point optimization

- $1/\varepsilon$ complexity estimates (or even $1/\sqrt{\varepsilon}$ under "favorable circumstances").
- Accuracy certificates are available.
- Favorable geometry of the problem domain – simple $O(n)$ proximal computation.
- Fully profit from fast gradient computation – $O(n \log n)$ cost per iteration.

We have a choice of at least 2 efficient techniques:

- Extra-gradient algorithms for smooth saddle-point problems (Mirror-Prox [Nemirovski, 2003], Dual Extrapolation [Nesterov, 2003], etc)
- Smoothing [Nesterov, 2003]:
  replace $f(u) = \max_{\|v\|_q \le 1} \langle v, \mathcal{A}u \rangle$ with its "Nesterov's smoothing":

$$f_\gamma(u) = \max_{\|v\|_q \le 1} \{\langle v, \mathcal{A}u \rangle + \gamma\omega(v)\},$$

  where $\omega(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|_q$-norm; then apply to f_γ Nesterov's accelerated algorithm for smooth optimization.

Comparing the contenders: theory

Nesterov accelerated algorithm:

- allows for easily implementable Euclidean and non-Euclidean prox and adaptive stepsize strategies;
- receives a "special mention" in the case of ℓ_2-norm minimization: instead of smoothing one can minimize the squared norm.
  - In this case, accelerated algorithm exhibits $1/\sqrt{\varepsilon}$ complexity for $\varepsilon \ll \mathrm{Opt}$.
- allows for the easily implementable warm start: the theoretical accuracy estimate depends on the initial distance to the optimum (though not on the sub-optimality of the initial solution).
  - Initializing by the optimal solution of the previous pixel/block brings 5-10 fold acceleration in practice!
- However, smoothing implementation (in its "basic form") requires to fix from the start the regularisation parameter $\gamma \asymp 1/\varepsilon$, resulting in curbed convergence rates.

Comparing the contenders: theory

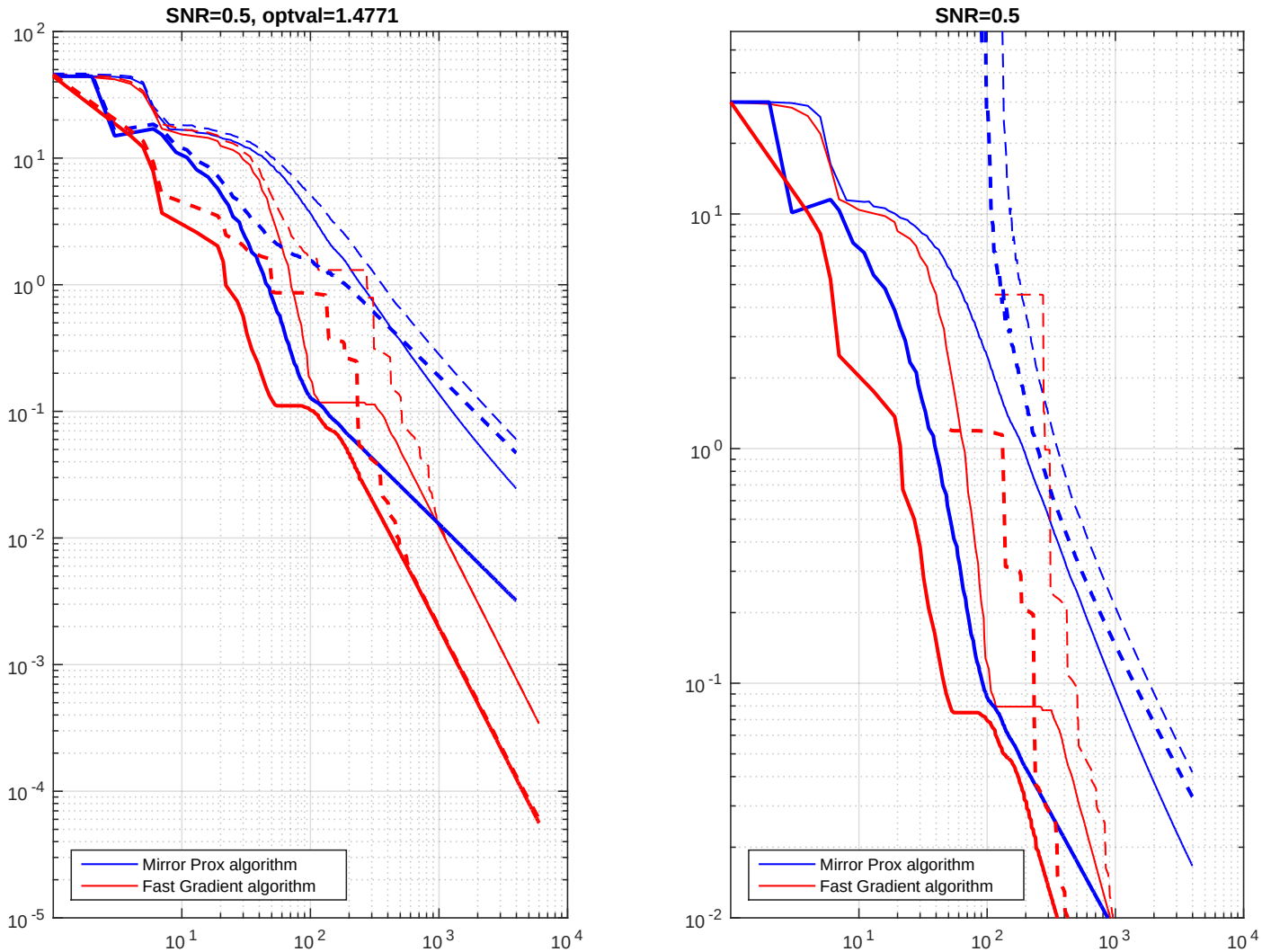
Nesterov accelerated algorithm:

- allows for easily implementable Euclidean and non-Euclidean prox and adaptive stepsize strategies;
- receives a "special mention" in the case of ℓ_2-norm minimization: instead of smoothing one can minimize the squared norm.
  - In this case, accelerated algorithm exhibits $1/\sqrt{\varepsilon}$ complexity for $\varepsilon \ll \mathrm{Opt}$.
- allows for the easily implementable warm start: the theoretical accuracy estimate depends on the initial distance to the optimum (though not on the sub-optimality of the initial solution).
  - Initializing by the optimal solution of the previous pixel/block brings 5-10 fold acceleration in practice!
- However, smoothing implementation (in its "basic form") requires to fix from the start the regularisation parameter $\gamma \asymp 1/\varepsilon$, resulting in curbed convergence rates.

Extra-gradient algorithms:

- allow for easily implementable Euclidean and non-Euclidean prox and adaptive stepsize strategies;
- can be seen as "online adjustment" of the regularization γ.
- on the other hand, no simple "warm start" strategy is available in this case.

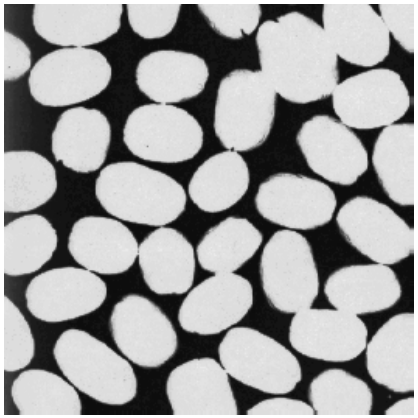
Comparing the contenders: experiments



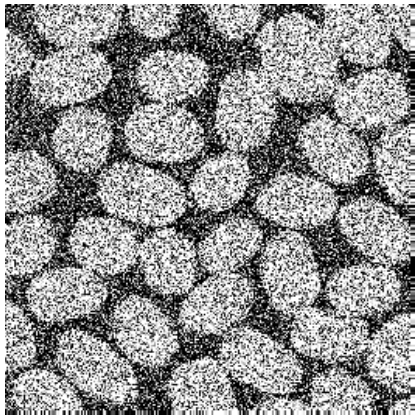
ℓ_2-norm minimization. Filter length $T = 200$, modulated 2nd order polynomial.
Left plot – absolute error, right plot – relative error as a function of iteration count.

Demonstration: Brodatz picture

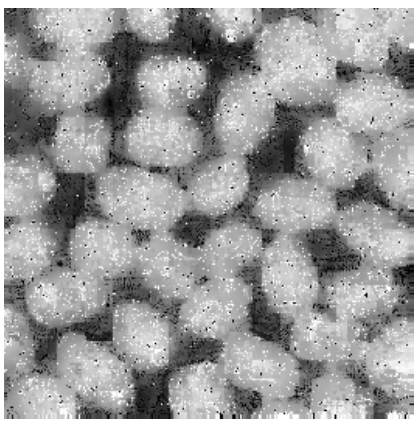
True signal



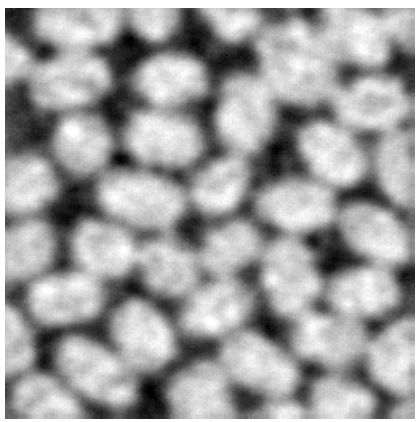
Observations



MP recovery



Lasso recovery



Brodatz D75 picture, SNR=1. AST over-sampling factor 4.
MISE_{Adapt}=3.2748e+03, MISE_{AST}=3.2514e+03.