Adaptive filtering by convex optimization

Anatoli Juditsky*

joint research with Z. Harchaoui‡, A. Nemirovski† and D. Ostrovski*
*University J. Fourier, ‡New Your University, †ISyE, Georgia Tech, Atlanta

MODE, March 25, 2016
Problem statement

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

$$y_\tau = x_\tau + \sigma \xi_\tau, \, \tau \in T,$$

where $\xi$ is the (complex-valued) white noise, $\xi_\tau \sim \mathcal{N}(0, \frac{1}{2} I_2)$.
Problem statement

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

$$y_\tau = x_\tau + \sigma \xi_\tau, \quad \tau \in T,$$

where $\xi$ is the (complex-valued) white noise, $\xi_\tau \sim \mathcal{N}(0, \frac{1}{2} I_2)$.

Optimal recovery:
Assume we want to estimate the value $x_t$ of the signal at $t \in T$.

Theorem [Ibragimov, Khas’mins’ki, 1984, Donoho, 1994, etc., reform.]
Let $\mathcal{X} \subset \mathbb{C}^T$ be a convex compact and centrally symmetric set. Then for a variety of loss functions, the minimax, over $x \in \mathcal{X}$, risk of recovering $x_t$ from noisy observations (1) is attained, within factor 1.2..., by a linear in $y$ estimate, readily given along with its risk, by the solution to convex optimization problem [...]
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).
Optimal recovery

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 class $\mathcal{X}$. Is it possible to “mimic” the oracle linear estimator $\varphi^*$, i.e. to construct an adaptive estimator (which only use observations) of comparable accuracy?*
Problem reformulation

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

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

(2)

Our objective may be either

- **filtering** – estimation of \( x_{2T} \) (or \( x_{-2T} \)),
- **interpolation** – estimation of \( x_t, -2T < t < 2T \) (e.g., \( x_0 \))
- **prediction** – estimation of \( x_{2T+k} \), (or \( x_{-2T-k} \)) for some \( k \in \mathbb{N}_+ \).

We assume that the oracle estimator \( \varphi^* \) has bounded support – can be represented as a “linear filter” of length \( \leq T + 1 \). For instance, when estimating \( x_t, -T/2 \leq t \leq T/2 \),

\[
x_t^* = \sum_{\tau=-T/2}^{T/2} \varphi^*_\tau y_{t-\tau} = [\varphi^* * y]_t.
\]
For the sake of simplicity, let us assume that we want to estimate $x_0$. We say that $x \in \mathbb{C}_{-T}^T$ if $x$ vanishes outside the interval $[-T, T]$.

We say that signal $x$ is simple at $t = 0$ if there exists a (oracle) filter $\varphi^* \in \mathbb{C}_{-T/2}^{T/2}$, satisfying

- (small variance condition) $\|\varphi^*\|_2 \leq \frac{\rho}{\sqrt{T}}$,
- (small bias condition) for some $\theta > 0$ and all $-\frac{3T}{2} \leq \tau \leq \frac{3T}{2}$,

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

More generally, for $x$ which is simple at $t$, there exists $\varphi^*$ of length $T$ and a neighborhood of size $O(T)$ of $t$ where $\varphi^* * x$ reproduces $x$ with “small bias”.
Basic assumption

For the sake of simplicity, let us assume that we want to estimate $x_0$. We say that $x \in \mathbb{C}_{-T}^T$ if $x$ vanishes outside the interval $[-T, T]$.

We say that signal $x$ is simple at $t = 0$ if there exists a (oracle) filter $\varphi^* \in \mathbb{C}_{-T/2}^{T/2}$, satisfying

- (small variance condition) $\|\varphi^*\|_2 \leq \frac{\rho}{\sqrt{T}}$,
- (small bias condition) for some $\theta > 0$ and all $-\frac{3T}{2} \leq \tau \leq \frac{3T}{2}$,

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

More generally, for $x$ which is simple at $t$, there exists $\varphi^*$ of length $T$ and a neighborhood of size $O(T)$ of $t$ where $\varphi^* \ast x$ reproduces $x$ with "small bias".

As a result, a simple at $t = 0$ signal $x$ can be "well recovered" from $y$ uniformly over $-\frac{3T}{2} \leq \tau \leq \frac{3T}{2}$:

$$\mathbb{E}|x_{\tau} - [\varphi^* \ast (x + \sigma\xi)]_{\tau}|^2 = \sigma^2\mathbb{E}|[\varphi^* \ast \xi]_{\tau}|^2 + |x_{\tau} - [\varphi^* \ast x]_{\tau}|^2$$

$$= \frac{\sigma^2 \rho^2}{T} + \frac{\theta^2 \sigma^2 \rho^2}{T} = O(1) \frac{\sigma^2 \rho^2}{T}. $$
Classical example

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

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

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

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

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

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

Note that the $\ell_2$-norm of $\phi$ 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 $O(1)\rho/\sqrt{T}$.
Less classical example

Suppose that \( f : [0, 1] \rightarrow \mathbb{C} \) can be locally, when \( x - h \leq x \leq x + h \), well-approximated by an exponential polynomial:

\[
p(x) = \sum_{k=1}^{K} c_k x^{r_k} e^{i\omega_k x}
\]

with fixed frequencies \( \omega_k \in \mathbb{C} \).

An exponential polynomial, \( K = 2 \)

Note that for any \( T = 2nh > 2K \) there exists a kernel \( K_h^* \), depending on the frequencies \( \omega_k \), of the norm \( O_K(1)/\sqrt{T} \) which exactly reproduces \( p \).
Less classical examples

When applied in the problem of estimation of $f$, kernel $K_h^*$, with properly chosen $h$, recovers $f(x)$ with the “parametric rate” [J., Nemirovski, 2009, 2013]

$$O_K(1) \frac{\sigma^2}{nh} = O_K(1) \frac{\sigma^2}{T}.$$ 

Furthermore,

- The class of simple signals is quite rich, it contains, for instance, signals $x_\tau \in \mathbb{C}$ which are close to solutions to homogeneous difference equations:

$$\sum_{k=1}^{K} w_k x_{\tau-k} = 0, \quad w \in \mathbb{C}^K.$$ 

- This class allows for a calculus: linear combinations, modulations, liftings, “tensor products” of simple signals are also simple.

- More examples in multi-dimensional case [J., Nemirovski, 2009] ...
Problem reformulation

Question:
under these conditions, is it possible to design an “adaptive estimation” \( \hat{x}_0 = [\hat{\varphi} * y]_0 \) of \( x_0 \) which only relies upon observations \( y \in \mathbb{C}_{-2T}^{2T} \), and such that

\[
\left[ E|\hat{x}_0 - x_0|^2 \right]^{1/2} \asymp \frac{\sigma \rho}{\sqrt{T}}
\]
Problem reformulation

Question:
under these conditions, is it possible to design an “adaptive estimation” \( \hat{x}_0 = [\hat{\varphi} * y]_0 \) of \( x_0 \) which only relies upon observations \( y \in \mathbb{C}^{2T} \), and such that

\[
\left[ \mathbb{E} |\hat{x}_0 - x_0|^2 \right]^{1/2} \approx \frac{\sigma \rho}{\sqrt{T}}
\]

Theorem 1 [lower bound].
For any \( \rho \geq 1 \), positive \( \sigma \) and \( T \in \mathbb{N} \) large enough, one can point out a family \( \mathcal{F}_T^\rho \) of real signals on \([-2T, 2T]\) such that

- for each signal \( s \in \mathcal{F}_T^\rho \) there exists a filter \( \varphi^* \in \mathbb{R}^{T/2} \) with \( ||\varphi^*||_2 = \frac{\rho}{\sqrt{T+1}} \), such that

\[
\max_{-3T/2 \leq \tau \leq 3T/2} \left[ \mathbb{E} ((\varphi^* * y)_\tau - x_\tau)^2 \right]^{1/2} = \frac{\sigma \rho}{\sqrt{T + 1}};
\]

- there is \( c_0 > 0 \) such that for any estimate \( \hat{x}_0 \) of \( x_0 \) from observations (1) it holds

\[
\sup_{x \in \mathcal{F}_T^\rho} \left[ \mathbb{E} (\hat{x}_0 - x_0)^2 \right]^{1/2} \geq c_0 \frac{\sigma \rho}{\sqrt{T + 1}} \rho \sqrt{\log(T + 1)}.
\]
Main result

**Theorem 2 [upper bound].**

Assume that \( x \) is simple at zero with known parameters \( \rho \) and \( \theta \). Then there is an estimate \( \hat{x}_0(y) \) of \( x_0 \) such that

\[
\left[ \mathbb{E} |\hat{x}_0(y) - x_0|^2 \right]^{1/2} \leq c \frac{\sigma \rho}{\sqrt{T}} \left[ \theta + \sqrt{\log(T + 1)} \right] \rho^2.
\]

Furthermore, one has with probability \( 1 - \varepsilon \),

\[
|\hat{x}_0(y) - x_0| \leq c \frac{\sigma \rho}{\sqrt{T}} \left[ \theta + \sqrt{\log \left( \frac{T + 1}{\varepsilon} \right)} \right] \rho^2.
\]
Naive approach – Empirical Risk minimization:
For a signal $x \in \mathbb{C}^Z$, $L \in \mathbb{N}_+$, and $1 \leq p \leq \infty$, let us denote

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

Define $\hat{\varphi}$ as an optimal solution to

$$\min_{\varphi \in \mathbb{C}^{T+1}} \left\{ \|y - \varphi \ast y\|_{3T/2,2}^2 : \|\varphi\|_2 \leq \frac{\rho}{\sqrt{T}} \right\}.$$
Constructing the adaptive filter 1

Naive approach – Empirical Risk minimization:
For a signal $x \in \mathbb{C}^Z$, $L \in \mathbb{N}_+$, and $1 \leq p \leq \infty$, let us denote

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

Define $\hat{\varphi}$ as an optimal solution to

$$\min_{\varphi \in \mathbb{C}^{T+1}} \left\{ \|y - \varphi \ast y\|_{3T/2,2}^2 : \|\varphi\|_2 \leq \frac{\rho}{\sqrt{T}} \right\}.$$ 

Note that $\varphi^*$ is feasible, so that

$$\|y - \hat{\varphi} \ast y\|_{3T/2,2}^2 \leq \|y - \varphi^* \ast y\|_{3T/2,2}^2 = O_P(1) + \sigma^2 \|\xi\|_{3T/2,2}^2.$$ 

Therefore,

$$\|x - \hat{\varphi} \ast y\|_{3T/2,2}^2 = \|y - \hat{\varphi} \ast y\|_{3T/2,2}^2 - \sigma^2 \|\xi\|_{3T/2,2}^2 - 2\sigma \langle \xi, x - \hat{\varphi} \ast y \rangle_{3T/2}$$

$$= O_P(1) + 2\sigma^2 \langle \xi, \hat{\varphi} \ast \xi \rangle_{3T/2} - 2\sigma \langle \xi, x - \hat{\varphi} \ast x \rangle_{3T/2}. \quad \text{\scriptsize \textcolor{red}{O_P(\sqrt{T})}}$$
For $x \in \mathbb{C}^\mathbb{Z}$, let $F_T(x)$ be the \textit{Discrete Fourier Transform (DFT)} of $[x]_{-T}^T$. We denote $\|x\|_{T,p}^* = \|F_T x\|_p$.

**Lemma**

Suppose that $\varphi^* \in \mathbb{C}_{-T/2}^T/2$ satisfies $\|\varphi^*\|_2 \leq \frac{\rho}{\sqrt{T}}$. Let also

$$\psi^* := (\varphi^* \ast \varphi^*) \in \mathbb{C}_{-T}^T.$$

Then $\psi^*$ it holds

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

- moreover, if $x$ is simple at 0 then for $\tau : -T \leq \tau \leq T$, $|x_\tau - [\psi^* \ast x]_\tau| \leq \frac{2\sigma_\theta \rho^2}{\sqrt{T}}$. 
Constructing the adaptive filter 2

Let \( \hat{\psi} \in \mathbb{C}_T \) be an optimal solution of the following problem:

\[
\min_{\psi \in \mathbb{C}_T} \left\{ \| y - \psi \ast y \|_{T,2} : \|\psi\|_{T,1}^* \leq \frac{\sqrt{2\rho^2}}{\sqrt{T}} \right\}. \tag{P_1}
\]

Then, as before, by the feasibility of \( \psi^* \)

\[
\| y - \hat{\psi} \ast y \|_{T,2} \leq \| y - \psi^* \ast y \|_{T,2}.
\]

- We have now better control of the cross-term \( \langle \xi, \hat{\psi} \ast \xi \rangle_T \)
  (“almost” the max of a convex function over a convex polyhedron):

  \[
  \langle \xi, \hat{\psi} \ast \xi \rangle_T \leq \max_{\|\psi\|_{1}^* \leq c^2 \sqrt{2/T}} \langle \xi, \psi \ast \xi \rangle_T = O_P(\log T).
  \]

- ...
Constructing the adaptive filter 2

Let \( \hat{\psi} \in \mathbb{C}^T \) be an optimal solution of the following problem:

\[
\min_{\psi \in \mathbb{C}^T} \left\{ \| y - \psi \ast y \|_{T,2} : \| \psi \|_{T,1} \leq \frac{\sqrt{2}\rho^2}{\sqrt{T}} \right\}. \tag{P_1}
\]

Then, as before, by the feasibility of \( \psi^* \)

\[
\| y - \hat{\psi} \ast y \|_{T,2} \leq \| y - \psi^* \ast y \|_{T,2}.
\]

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

\[
\langle \xi, \hat{\psi} \ast \xi \rangle_T \leq \max_{\| \psi \|_{1} \leq c^2 \sqrt{2/T}} \langle \xi, \psi \ast \xi \rangle_T = O_P (\log T).
\]

- ...

- We finally get

\[
E \| x - [\hat{\psi} \ast y]_T \|_{2,T}^{1/2} \leq C\sigma \rho (1 + \theta) \left[ \rho \sqrt{\log T} \right]
\]

and

\[
E \| x_0 - [\hat{\psi} \ast y]_0 \| \leq \frac{C\sigma \rho (1 + \theta)}{\sqrt{T}} \left[ \rho^2 \sqrt{\log T} \right]
\]
A variant

Let $\hat{\psi}$ be an optimal solution to

$$\min_{\psi \in \mathbb{C}^{2T+1}} \left\{ \|y - \psi^* y\|_{T,\infty}^* : \|\psi\|_{T,1}^* \leq \frac{\sqrt{2}\rho^2}{\sqrt{T}} \right\} \quad (P_2)$$

Theorem 3 [upper bound]

Consider the estimation $\hat{x}_0(y) = [\hat{\psi}^* y]_0$ of $x_0$. Then

$$\mathbb{E} \left[ |x_0(y) - \hat{x}_0|^2 \right]^{1/2} \leq c \frac{\sigma \rho}{\sqrt{T}} \left[ \rho^3 \sqrt{\log[T]} + \theta \right] ,$$

and, with probability $1 - \varepsilon$,

$$|\hat{x}_0(y) - x_0| \leq c \frac{\sigma \rho}{\sqrt{T}} \left[ \rho^3 \sqrt{\log[T/\varepsilon]} + \theta \right] .$$
A summary

• Let \((x_\tau)\) admit, for some \(T\), the estimate \(x^*_\tau = [\varphi^* \ast y]_\tau\) with “bandwidth” \(T\) (i.e., with \(\varphi^* \in \mathbb{C}_{-T/2}^{T/2}\)) such that

\[
\max_{\tau:|\tau-t| \leq 3T/2} \mathbb{E}\left\{|x_\tau - x^*_\tau|^2\right\} \leq \kappa^2 := \frac{\sigma^2 \mu^2}{T + 1}
\]

for some known \(\mu \geq 1\).

• Our objective is, assuming that \(T\) and \(\mu\) are known, to recover \(x_t\) from observations \([y]_{t-2T}^{t+2T}\) nearly as well as if we were using our hypothetic estimate \(x^*_t\).
A summary

• Let \((x_\tau)\) admit, for some \(T\), the estimate \(x^*_\tau = [\varphi^* * y]_\tau\) with “bandwidth” \(T\) (i.e., with \(\varphi^* \in \mathbb{C}_{-T/2}^T\)) such that

\[
\max_{\tau:|\tau-t| \leq 3T/2} \mathbb{E} \left\{ |x_\tau - x^*_\tau|^2 \right\} \leq \kappa^2 := \frac{\sigma^2 \mu^2}{T + 1} \tag{3}
\]

for some known \(\mu \geq 1\).

• Our objective is, assuming that \(T\) and \(\mu\) are known, to recover \(x_t\) from observations \([y]_{t+2T}^{t+2T}\) nearly as well as if we were using our hypothetic estimate \(x^*_t\).

• By (3), \(|\varphi^*|_2 \leq \frac{\mu}{\sqrt{T+1}}\) and \(x\) is simple. When applying Theorem 2 or 3 with \(\rho = \mu, \theta = 1\), we conclude that the MSE of recovery \(\widehat{x}_t = [\widehat{\psi} * y]_t\) is bounded, respectively, by

\[
O(1)\mu^2 \sqrt{\log(T)} \kappa \quad \text{or} \quad O(1)\mu^3 \sqrt{\log(T)} \kappa.
\]

when using \((P_1)\) \quad when using \((P_2)\)
Adaptation to $\rho$ and $T$

In “practical applications” values of the parameter $\rho$ and of the bandwidth $T$ are unknown.

- The algorithms can be modified to be adaptive with respect to $\rho$. For instance, $(P_2)$ can be replaced with the “norm minimization” problem

$$\min_{\psi, r} \left\{ r : \| y - \psi \ast y \|^*_{T, \infty} \leq 2\sigma(1 + r)\sqrt{\log[T + 1]}, \quad \|\psi\|^*_{T, 1} \leq r(2T + 1)^{-1/2}. \right\} \quad (P'_2)$$

Instead of constrained problems, we can consider their penalized versions. For instance, $(P_1)$ can be replaced with

$$\min_{\psi} \left\{ \| y - \psi \ast y \|^2_{T, 2} + \kappa\sigma^2 \sqrt{2T + 1}\|\psi\|^*_{T, 1} \right\}. \quad (P''_1)$$

with penalty $\kappa = \kappa_0 \log(T)$.

...
Adaptation to $\rho$ and $T$

In “practical applications” values of the parameter $\rho$ and of the bandwidth $T$ are unknown.

- The algorithms can be modified to be adaptive with respect to $\rho$. For instance, $(P_2)$ can be replaced with the “norm minimization” problem

$$\min_{\psi, r} \left\{ r : \|y - \psi \ast y\|_{T, \infty}^* \leq 2\sigma(1 + r)\sqrt{\log[T + 1]}, \right. \\
\left. \|\psi\|_{T, 1}^* \leq r(2T + 1)^{-1/2}. \right\} \quad (P'_{2})$$

Instead of constrained problems, we can consider their penalized versions. For instance, $(P_1)$ can be replaced with

$$\min_{\psi} \left\{ \|y - \psi \ast y\|_{T, 2}^2 + \kappa \sigma^2 \sqrt{2T + 1}\|\psi\|_{T, 1}^* \right\}. \quad (P_{1''})$$

with penalty $\kappa = \kappa_0 \log(T)$.

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

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

• For each point $t$ of the grid we need
  
  1. choose a set of bandwidths $\{T_0 = 0, T_1 = 1, T_2 = 2, ..., T_K = n\}$,
  2. for each bandwidth $T_k$ compute an approximate solution $\hat{\psi}_{T_k, t}$ to $(P_1)$ (or $(P_2)$, $(P'_2)$,...)
  3. compute estimations $\hat{x}_t[T_k] = [\hat{\psi}_{T_k, t} * y]_t$ and aggregate them using Lepski’s algorithm.

• To reduce the numerical cost, instead of proceeding point-wise, one can use block-wise update of filters...
When applying the proposed approach to “practical” recovery of a signal or an image

• For each point $t$ of the grid we need
  1. choose a set of bandwidths $\{ T_0 = 0, \; T_1 = 1, \; T_2 = 2, \ldots, \; T_K = n \}$,
  2. for each bandwidth $T_k$ compute an approximate solution $\hat{\psi}_{T_k,t}$ to $(P_1)$ (or $(P_2)$, $(P_2')$,...)
  3. compute estimations $\hat{x}_t[T_k] = [\hat{\psi}_{T_k,t} * y]_t$ and aggregate them using Lepski’s algorithm.

• To reduce the numerical cost, instead of proceeding point-wise, one can use block-wise update of filters...

One needs to solve repeatedly problems $(P_1)$ of the kind (or alike)

$$\text{Opt} = \min_{\psi \in \mathcal{C}_{-T}^T} \left\{ f(\psi) = \| y - y * \psi \|^*_T, \; \| \psi \|^*_T,1 \leq r \right\}, \; r > 0, \; p \in \{2, \infty\}. \quad (P_*)$$
Choosing the optimization tool

Note that \((P_*)\) can be rewritten as a bilinear saddle-point problem: indeed, its objective,

\[
    f(\psi) = \max_{u \in \mathbb{C}^{2T+1}} \{ \langle u, F_T(y - y \ast \psi) \rangle, \|u\|_q \leq 1 \},
\]

where \( \frac{1}{p} + \frac{1}{q} = 1 \).

When denoting \(z = F_T(\psi)\),

\[
    \text{Opt} = \min_{\psi \in \mathbb{C}^{2T+1}} \max_{u \in \mathbb{C}^{2T+1}} \{ \langle u, Az \rangle + \langle u, b \rangle : \|u\|_q \leq 1, \|z\|_1 \leq r \},
\]

\((P_*)\)

where \(q \in \{1, 2\}\), \(b = F_T(y)\), and \(A\) is as follows:

\[
    Az = F_T \left[ y \ast F_T^{-1}(z) \right]
    = F_T \left[ F_{2T}^{-1} \left\{ F_{2T} \left[ 0_T; y; 0_T \right] \ast F_{2T} \left[ 0_{2T}; F_T^{-1}(z); 0_{2T} \right] \right\} \right]
\]

(here \([x; 0_T] \) stands for the concatenation with zero vector of length \(T\) and \(\ast\) is the Hadamard element-wise product).
• $(P_*)$ is a bilinear saddle-point problem with domains which are balls of either $\ell_2$- or $\ell_2/\ell_1$-norm.

• Problems should be solved to (relatively) low accuracy – a solution $\hat{z}$ of accuracy

$$\epsilon(\hat{z}) := f(\hat{z}) - \text{Opt} \leq \frac{1}{4}\text{Opt}$$

will be largely sufficient.

• Objective gradients can be computed in $O(n\log n)$ operations using the FFT.
Choosing the optimization tool 2

• \((P_*)\) is a bilinear saddle-point problem with domains which are balls of either \(\ell_2\) or \(\ell_2/\ell_1\)-norm.

• Problems should be solved to (relatively) low accuracy — a solution \(\hat{z}\) of accuracy

\[
\epsilon(\hat{z}) := f(\hat{z}) - \text{Opt} \leq \frac{1}{4} \text{Opt}
\]

will be largely sufficient.

• Objective gradients can be computed in \(O(n \log n)\) operations using the FFT.

Under the premise, proximal First Order algorithms appear to be methods of choice.
• $1/\epsilon$ complexity estimates (or even $1/\sqrt{\epsilon}$ under “favorable circumstances”).
• Accuracy certificates are available “at no cost”.
• 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

- \(1/\epsilon\) complexity estimates (or even \(1/\sqrt{\epsilon}\) under “favorable circumstances”).
- Accuracy certificates are available “at no cost”.
- 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 saddle-point problems (Mirror-Prox [Nemirovski, 2003], Dual Extrapolation [Nesterov, 2003], etc)

- Smoothing [Nesterov, 2003]:
  replace \(f(z) = \max_{\|u\|_q \leq 1} \langle u, Az \rangle\) with its “Nesterov’s smoothing”:

\[
    f_{\gamma}(z) = \max_{\|u\|_q \leq 1} \{ \langle u, Az \rangle + \gamma \vartheta(u) \},
\]

where \(\vartheta\) is 1-strongly convex with respect to \(\| \cdot \|_q\)-norm; then apply to \(f_{\gamma}\) 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 $\ell_2$-norm minimization: instead of smoothing one can minimize the squared norm. In this case, accelerate algorithm exhibits $1/\sqrt{\epsilon}$ complexity for $\epsilon \ll \text{Opt}$.

• 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).

• However, smoothing implementation (in its “basic form”) requires to fix from the start the regularisation parameter $\gamma \asymp 1/\epsilon$, what results in curbed convergence rates.

Extra-gradient algorithms:

• allows for easily implementable Euclidean and non-Euclidean prox and adaptive stepsize strategies;

• can be seen as “online adjustment” of the regularization $\gamma$.

• On the other hand, no simple “warm start” strategy is available in this case.
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 $\ell_2$-norm minimization: instead of smoothing one can minimize the squared norm. In this case, accelerate algorithm exhibits $1/\sqrt{\epsilon}$ complexity for $\epsilon \ll \text{Opt}$.
- 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).
- However, smoothing implementation (in its “basic form”) requires to fix from the start the regularisation parameter $\gamma \simeq 1/\epsilon$, what results in curbed convergence rates.

Extra-gradient algorithms:

- allows for easily implementable Euclidean and non-Euclidean prox and adaptive stepsize strategies;
- can be seen as “online adjustment” of the regularization $\gamma$.
- On the other hand, no simple “warm start” strategy is available in this case.
Comparing the contenders: experiments

$\ell_2$-norm minimization. Filter length $T = 200$, modulated 2nd order polynomial. Left plot – absolute error, right plot – relative error as a function of iteration count.
Simulation experiment: adaptive recovery

Comparison with Atomic Soft Thresholding (AST), a.k.a. spectral Lasso by [Bhaskar et al., 2013, Tang et al., 2013]
Modulated 4th order polynomial, SNR=1. AST over-sampling factor $\kappa = 4.$
Modulated 4th order polynomial, SNR=1. AST over-sampling factor $\kappa = 4$. 

Simulation experiment: adaptive recovery
Simulation experiments: sum of harmonic oscillations

Sum of 4 oscillations. AST over-sampling factor $\kappa = 4$. 
Sum of harmonic oscillations: zoomed image

![Ground truth](image1)

![Observations](image2)

Filtering recovery, MSE=8.9972

Lasso recovery, MSE=66.5866

Sum of 4 oscillations. AST over-sampling factor $\kappa = 4$. 
Simulation experiments: Brodatz picture

Brodatz D75 picture, SNR=1. AST over-sampling factor $\kappa = 4$. $\text{MISE}_{\text{Adapt}} = 3.2748e+03$, $\text{MISE}_{\text{AST}} = 3.2514e+03$. 