Convex nonsmooth optimization
Part III: Algorithms

Barbara Pascal

LS2N, CNRS, Centrale Nantes, Nantes University, Nantes, France
barbara.pascal@cnrs.fr

http://bpascal-fr.github.io

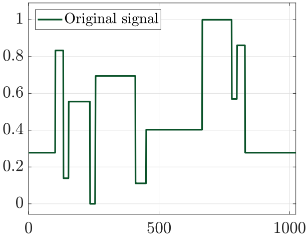
Collaboration

This course is a direct adaptation of the course built by
Jean-Christophe Pesquet (CentraleSupélec) and Nelly Pustelnik (LPENSL)

Reconstruction of a piecewise noisy signal

Ground truth

$$\overline{x} \in \mathbb{R}^N$$



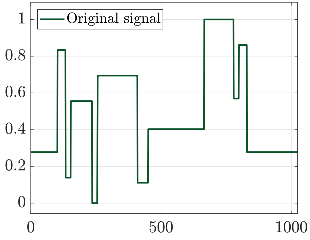
Reconstruction of a piecewise noisy signal

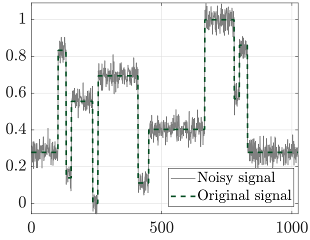
Ground truth

$$\overline{x} \in \mathbb{R}^N$$

Gaussian noise with $\sigma = 0.04$

$$y = \overline{x} + \xi \in \mathbb{R}^N$$





Purpose: recover the true signal with *sharp* transitions

Denoising by functional minimization

Regularized scheme — $\mathbf{D}$: *differential operator*, $\|\cdot\|_p$: ℓ_p*-norm*

$$\widehat{x}(y;\lambda) \in \underset{x\in\mathbb{R}^N}{\mathrm{Argmin}}\ \frac{1}{2}\|x-y\|_2^2 + \lambda\|\mathbf{D}x\|_p^p$$

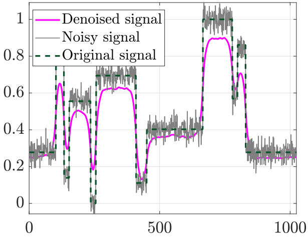
Denoising by functional minimization

Regularized scheme — $\mathbf{D}$: *differential operator*, $\|\cdot\|_p$: ℓ_p*-norm*

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x\in\mathbb{R}^N} \frac{1}{2}\|x-y\|_2^2 + \lambda\|\mathbf{D}x\|_p^p$$

Tikhonov regularizer $\|\mathbf{D}x\|_2^2$

Smooth: gradient descent



✗ fuzzy transitions

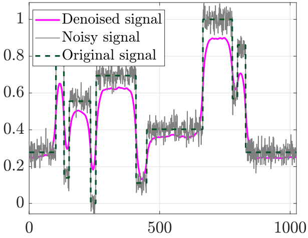
Denoising by functional minimization

Regularized scheme — $\mathbf{D}$: *differential operator*, $\|\cdot\|_p$: *ℓ_p-norm*

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x\in\mathbb{R}^N} \frac{1}{2}\|x-y\|_2^2 + \lambda\|\mathbf{D}x\|_p^p$$

Tikhonov regularizer $\|\mathbf{D}x\|_2^2$

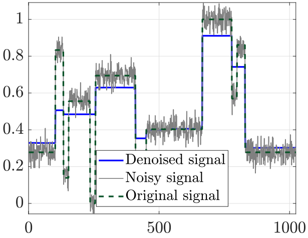
Smooth: gradient descent



✗ fuzzy transitions

Total Variation $\|\mathbf{D}x\|_1$

Nonsmooth: proximal algorithm



✓ sharp transitions

Formulation of the problem

Piecewise denoising

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|x - y\|_2^2 + \lambda \|\mathbf{D}x\|_1$$

Formulation of the problem

Piecewise denoising

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|x - y\|_2^2 + \lambda \|\mathbf{D}x\|_1$$

- *Smooth* data-fidelity $\qquad f(x) = \frac{1}{2}\|x - y\|_2^2$

Formulation of the problem

Piecewise denoising

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x\in\mathbb{R}^N} \frac{1}{2}\|x-y\|_2^2 + \lambda\|\mathbf{D}x\|_1$$

- *Smooth* data-fidelity $\quad f(x) = \frac{1}{2}\|x-y\|_2^2$
- *Non-smooth* regularizer $\quad h(\mathbf{L}x) = \lambda\|\mathbf{D}x\|_1, \quad$ with $h(z) = \lambda\|z\|_1$

Formulation of the problem

Piecewise denoising

$$\widehat{x}(y;\lambda) \in \underset{x\in\mathbb{R}^N}{\mathrm{Argmin}}\ \frac{1}{2}\|x-y\|_2^2 + \lambda\|\mathbf{D}x\|_1$$

- *Smooth* data-fidelity $\quad f(x) = \frac{1}{2}\|x-y\|_2^2$
- *Non-smooth* regularizer $\quad h(\mathbf{L}x) = \lambda\|\mathbf{D}x\|_1, \quad$ with $h(z) = \lambda\|z\|_1$

General form:

$$\widehat{x} \in \underset{x\in\mathbb{R}^N}{\mathrm{Argmin}}\ \{f(x) + h(\mathbf{L}x) = f(x) + g(x)\}$$

f *smooth*; h and $g = h(\mathbf{L}\cdot)$ *nonsmooth*.

Optimization algorithm: *Forward-Backward*

Let $\mathcal{H}$ be a Hilbert space.
Let $f \in \Gamma_0(\mathcal{H})$ be differentiable with a ν-Lipschitzian gradient where $\nu \in \,]0, +\infty[$.
Let $g \in \Gamma_0(\mathcal{H})$.

Let $\gamma \in]0, 2/\nu[$ and $\delta = \min\{1, 1/(\nu\gamma)\} + 1/2$.
Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $[0, \delta[$ such that $\sum_{n\in\mathbb{N}} \lambda_n(\delta - \lambda_n) = +\infty$.

Assume that $\operatorname{Argmin}(f+g) \neq \varnothing$. Let $x_0 \in \mathcal{H}$ and

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_n = x_n - \gamma \nabla f(x_n) \\ x_{n+1} = x_n + \lambda_n(\operatorname{prox}_{\gamma g} y_n - x_n). \end{cases}$$

Then, $(x_n)_{n\in\mathbb{N}}$ converges weakly to a minimizer of $f + g$.

Example: bounded least-squares

Observation model:

$$y = \mathbf{A}\overline{x} + \xi \in \mathbb{R}^P,$$

linear operator $\mathbf{A} \in \mathbb{R}^{P \times N}$, ξ Gaussian noise, ground truth $\overline{x} \in \mathbb{R}^N$, s.t.

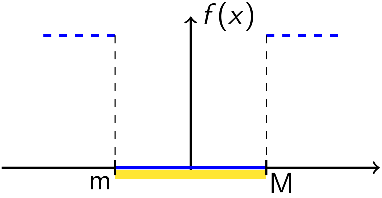
$$\forall i \in \{1, \dots, N\}, \quad m \leq \overline{x}_i \leq M$$

Bounded least-squares

$$C = \left\{ x \in \mathbb{R}^N \,|\, \forall i, \;\; x_i \in [m, M] \right\}$$

$$\widehat{x} \in \underset{x \in C}{\mathrm{Argmin}} \; \frac{1}{2} \|y - \mathbf{A}x\|_2^2$$

$$\iff \quad \widehat{x} \in \underset{x \in \mathbb{R}^N}{\mathrm{Argmin}} \; \frac{1}{2} \|y - \mathbf{A}x\|_2^2 + \iota_C(x)$$



Optimization algorithm: projected gradient

Let $\mathcal{H}$ be a Hilbert space.
Let $f \in \Gamma_0(\mathcal{H})$ be differentiable with a ν-Lipschitzian gradient where $\nu \in \,]0, +\infty[$.
Let C a nonempty closed convex subset of $\mathcal{H}$ and P_C the projection on C.

Let $\gamma \in]0, 2/\nu[$ and $\delta = \min\{1, 1/(\nu\gamma)\} + 1/2$.
Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $[0, \delta[$ such that $\sum_{n\in\mathbb{N}} \lambda_n(\delta - \lambda_n) = +\infty$.

Assume that $\mathrm{Argmin}_{x\in C}\, g(x) \neq \varnothing$. Let $x_0 \in \mathcal{H}$ and

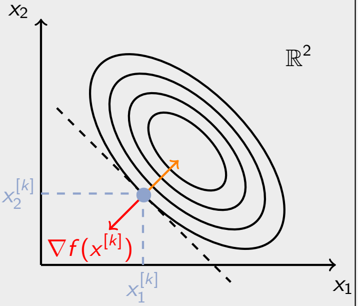
$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_n = x_n - \gamma \nabla f(x_n) \\ x_{n+1} = x_n + \lambda_n (P_C y_n - x_n). \end{cases}$$

Then, $(x_n)_{n\in\mathbb{N}}$ converges weakly to a minimizer of g over C.

Optimization algorithm: gradient descent

Let $\mathcal{H}$ be a Hilbert space.
Let $f \in \Gamma_0(\mathcal{H})$ differentiable with a ν-Lipschitz gradient, $\nu \in \left]0, +\infty\right[$.

Le $\gamma \in]0, 2/\nu[$.



Assume that $\mathrm{Argmin}\, f \neq \varnothing$. Let $x_0 \in \mathcal{H}$ and

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma \nabla f(x_n)$$

Then, $(x_n)_{n\in\mathbb{N}}$ converges weakly to a minimizer of f.

Optimization algorithm: Douglas-Rachford

Let $\mathcal{H}$ be a Hilbert space.

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$.

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_n = \mathrm{prox}_{\gamma g} x_n \\ z_n = \mathrm{prox}_{\gamma f}(2y_n - x_n) \\ x_{n+1} = x_n + \lambda_n (z_n - y_n). \end{cases}$$

Optimization algorithm: Douglas-Rachford

Let $\mathcal{H}$ be a Hilbert space.
Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$.

Let $\gamma \in \,]0, +\infty[$ and let $(\lambda_n)_{n\in\mathbb{N}}$ a sequence in $[0, 2]$ s.t. $\sum_{n\in\mathbb{N}} \lambda_n(2-\lambda_n) = +\infty$.

Assume that $\operatorname{Argmin}(f+g) \neq \varnothing$. Let $x_0 \in \mathcal{H}$ and

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_n = \operatorname{prox}_{\gamma g} x_n \\ z_n = \operatorname{prox}_{\gamma f}(2y_n - x_n) \\ x_{n+1} = x_n + \lambda_n(z_n - y_n). \end{cases}$$

The following properties are satisfied:

- $x_n \rightharpoonup \widehat{x}$
- $z_n - y_n \to 0$, $y_n \rightharpoonup \widehat{y}$, $z_n \rightharpoonup \widehat{y}$ where $\widehat{y} = \operatorname{prox}_{\gamma g}\widehat{x} \in \operatorname{Argmin}(f+g)$.

Optimization algorithm: Douglas-Rachford

Let $\mathcal{H}$ and $\mathcal{G}$ be two finite dimensional Hilbert spaces.
Let $g \in \Gamma_0(\mathcal{H})$ and $L \in \mathcal{B}(\mathcal{G}, \mathcal{H})$ s.t. L^*L is a isomorphism.

Let $\gamma \in \,]0, +\infty[$ and let $(\lambda_n)_{n\in\mathbb{N}}$ a sequence in $[0, 2]$ s.t. $\sum_{n\in\mathbb{N}} \lambda_n(2-\lambda_n) = +\infty$.

Assume that $\operatorname{Argmin}(g \circ L) \neq \varnothing$. Let $x_0 \in \mathcal{H}$, $v_0 = (L^*L)^{-1}L^*x_0$ and

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_n = \operatorname{prox}_{\gamma g} x_n \\ c_n = (L^*L)^{-1}L^*y_n \\ x_{n+1} = x_n + \lambda_n\big(L(2c_n - v_n) - y_n\big) \\ v_{n+1} = v_n + \lambda_n(c_n - v_n). \end{cases}$$

We have then $v_n \rightharpoonup \widehat{v}$ where $\widehat{v} \in \operatorname{Argmin}(g \circ L)$.

Optimization algorithm: Douglas-Rachford

Sketch of proof:

$$\underset{v \in \mathcal{G}}{\text{minimize}} \; g(Lv) \;\Leftrightarrow\; \underset{x \in \mathcal{H}}{\text{minimize}} \; \iota_E(x) + g(x)$$

where $E = \operatorname{ran} L$.
We apply Douglas-Rachford algorithm with
$f = \iota_E \Rightarrow \operatorname{prox}_{\gamma f} = P_E$ by setting

$$(\forall n \in \mathbb{N}) \qquad P_E y_n = Lc_n \text{ and } P_E x_n = Lv_n$$

where $c_n = \underset{c \in \mathcal{H}}{\operatorname{argmin}} \; \|y_n - Lc\|^2 = (L^* L)^{-1} L^* y_n$.

Optimization algorithm: Douglas-Rachford

Particular case of Douglas-Rachford algorithm:
$\mathcal{H} = \mathcal{H}_1 \times \cdots \mathcal{H}_m$ where $\mathcal{H}_1, \ldots, \mathcal{H}_m$ Hilbert spaces
$\big(\forall x = (x_1, \ldots, x_m) \in \mathcal{H}\big)$ $g(x) = \sum_{i=1}^m g_i(x_i)$
where $(\forall i \in \{1, \ldots, m\}$ $g_i \in \Gamma_0(\mathcal{H}_i)$
$L\colon v \mapsto (L_1 v, \ldots, L_m v)$ where $(\forall i \in \{1, \ldots, m\})$ $L_i \in \mathcal{B}(\mathcal{G}, \mathcal{H}_i)$.

PPXA+ algorithm

Let $(x_{0,i})_{1 \le i \le m} \in \mathcal{H}$, $v_0 = (\sum_{i=1}^m L_i^* L_i)^{-1} \sum_{i=1}^m L_i^* x_{0,i}$ and

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_{n,i} = \operatorname{prox}_{\gamma g_i} x_{n,i}, \quad i \in \{1, \ldots, m\} \\ c_n = (\sum_{i=1}^m L_i^* L_i)^{-1} \sum_{i=1}^m L_i^* y_{n,i} \\ x_{n+1,i} = x_{n,i} + \lambda_n \big(L_i(2c_n - v_n) - y_{n,i}\big), \quad i \in \{1, \ldots, m\} \\ v_{n+1} = v_n + \lambda_n (c_n - v_n). \end{cases}$$

We have then $v_n \rightharpoonup \widehat{v} \in \operatorname{Argmin} \sum_{i=1}^m g_i \circ L_i$.

Optimization algorithm: Douglas-Rachford

Particular case of Douglas-Rachford algorithm:

$\mathcal{H} = \mathcal{H}_1 \times \cdots \mathcal{H}_m$ where $\mathcal{H}_1 = \ldots = \mathcal{H}_m$ Hilbert spaces

$\big(\forall x = (x_1, \ldots, x_m) \in \mathcal{H}\big)\ g(x) = \sum_{i=1}^m g_i(x_i)$

where $(\forall i \in \{1, \ldots, m\}\ g_i \in \Gamma_0(\mathcal{H}_i)$

$L\colon v \mapsto (L_1 v, \ldots, L_m v)$ where $L_1 = \ldots = L_m = \mathrm{Id}$.

PPXA algorithm

Let $(x_{0,i})_{1 \le i \le m} \in \mathcal{H}$, $v_0 = \frac{1}{m} \sum_{i=1}^m x_{0,i}$ and

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_{n,i} = \mathrm{prox}_{\gamma g_i} x_{n,i}, \quad i \in \{1, \ldots, m\} \\ c_n = \frac{1}{m} \sum_{i=1}^m y_{n,i} \\ x_{n+1,i} = x_{n,i} + \lambda_n \big(2c_n - v_n - y_{n,i}\big), \quad i \in \{1, \ldots, m\} \\ v_{n+1} = v_n + \lambda_n (c_n - v_n). \end{cases}$$

We have then $v_n \rightharpoonup \widehat{v} \in \mathrm{Argmin} \sum_{i=1}^m g_i$.

Optimization algorithms

Forward-Backward	$f_1 + f_2$	f_1 gradient Lipschitz prox_{f_2}	[Combettes,Wajs,2005]
ISTA	$f_1 + f_2$	f_1 gradient Lipschitz $f_2 = \lambda\|\cdot\|_1$	[Daubechies et al, 2003]
Projected gradient	$f_1 + f_2$	f_1 gradient Lipschitz $f_2 = \iota_C$	
Gradient descent	$f_1 + f_2$	f_1 gradient Lipschitz $f_2 = 0$	
Douglas-Rachford	$f_1 + f_2$	prox_{f_1} prox_{f_2}	[Combettes,Pesquet, 2007]
PPXA	$\sum_i f_i$	prox_{f_i}	[Combettes,Pesquet, 2008]
PPXA+	$\sum_i f_i \circ L_i$	prox_{f_i} $(\sum_{i=1}^m L_i^* L_i)^{-1}$	[Pesquet, Pustelnik, 2012]

ISTA: Iterative Shrinkage-Thresholding Algorithm

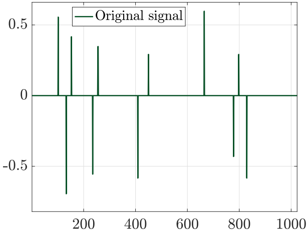
Sparse estimation

Let $y \in \mathbb{R}^N$ some noisy observation of a pulse signal and consider the estimator:

$$\widehat{x}(y;\lambda) \in \underset{x \in \mathbb{R}^N}{\mathrm{Argmin}} \ \frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_1$$

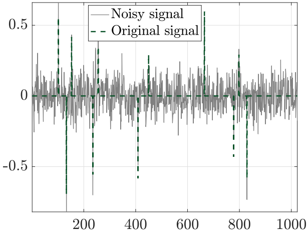
Ground truth

$\overline{x} \in \mathbb{R}^N$



Gaussian noise with $\sigma = 0.1$

$y = \overline{x} + \xi \in \mathbb{R}^N$



ISTA: Iterative Shrinkage-Thresholding Algorithm

Sparse estimation

Let $y \in \mathbb{R}^N$ some noisy observation of a pulse signal and consider the estimator:

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_1$$

Let $f(x) = \frac{1}{2}\|x-y\|_2^2$ and $g(x) = \lambda\|x\|_1$.

1. Compute the gradient of f.
2. Let $\gamma > 0$, compute the proximity operator $\mathrm{prox}_{\gamma f}$.
3. Give the expression of the proximity operator $\mathrm{prox}_{\gamma g}$.
4. Write the Forward-Backward scheme computing $\widehat{x}(y;\lambda)$.
5. Write the Douglas-Rachford scheme computing $\widehat{x}(y;\lambda)$.

ISTA: Iterative Shrinkage-Thresholding Algorithm

Sparse estimation

Let $y \in \mathbb{R}^N$ some noisy observation of a pulse signal and consider the estimator:

$$\widehat{x}(y;\lambda) \in \underset{x\in\mathbb{R}^N}{\mathrm{Argmin}}\ \frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_1$$

Let $f(x) = \frac{1}{2}\|x-y\|_2^2$ and $g(x) = \lambda\|x\|_1$.

1. F is smooth and its gradient writes

$$\nabla f(x) = x - y \in \mathbb{R}^N$$

ISTA: Iterative Shrinkage-Thresholding Algorithm

Sparse estimation

Let $y \in \mathbb{R}^N$ some noisy observation of a pulse signal and consider the estimator:

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|x - y\|_2^2 + \lambda\|x\|_1$$

Let $f(x) = \frac{1}{2}\|x - y\|_2^2$ and $g(x) = \lambda\|x\|_1$.

2. g is proper, lower-semicontinuous, convex; its proximity operator is defined as

$$\operatorname{prox}_{\gamma f}(x) = \operatorname*{argmin}_{z \in \mathbb{R}^N} \frac{1}{2}\|z - x\|_2^2 + \frac{\gamma}{2}\|z - y\|_2^2.$$

Then, $p = \operatorname{prox}_{\gamma f}(x) \iff z - x + \gamma(z - y) = 0$ and hence

$$\operatorname{prox}_{\gamma f}(x) = \frac{x + \gamma y}{1 + \gamma}.$$

ISTA: Iterative Shrinkage-Thresholding Algorithm

Sparse estimation

Let $y \in \mathbb{R}^N$ some noisy observation of a pulse signal and consider the estimator:

$$\widehat{x}(y;\lambda) \in \underset{x\in\mathbb{R}^N}{\mathrm{Argmin}}\ \frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_1$$

Let $f(x) = \dfrac{1}{2}\|x-y\|_2^2$ and $g(x) = \lambda\|x\|_1$.

3. g is proper, lower-semicontinuous, convex and separable $g(x) = \sum_{i=1}^{N} g_i(x_i)$ with $g_i(x_i) = |x_i|$ and

$$\mathrm{prox}_{\gamma g_i}(x_i) = \begin{cases} 0 & \text{if } |x_i| \leq \gamma \\ x_i - \mathrm{sgn}(x_i)\gamma & \text{otherwise.} \end{cases}$$

ISTA: Iterative Shrinkage-Thresholding Algorithm

Sparse estimation

Let $y \in \mathbb{R}^N$ some noisy observation of a pulse signal and consider the estimator:

$$\widehat{x}(y;\lambda) \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_1$$

4. The function f has a 1-Lipschitz gradient. Then, for $\gamma \in]0,2[$, $x_0 \in \mathbb{R}^N$

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_n = x_n - \gamma(x_n - y) \\ x_{n+1} = \operatorname{prox}_{\gamma\lambda\|\cdot\|_1}(y_n) \end{cases}$$

$(x_n)_{n\in\mathbb{N}}$ converges toward $\widehat{x}(y;\lambda)$.

The sequence $(\lambda_n)_{n\in\mathbb{N}}$ has been chosen constant equal to 1.

ISTA: Iterative Shrinkage-Thresholding Algorithm

Sparse estimation

Let $y \in \mathbb{R}^N$ some noisy observation of a pulse signal and consider the estimator:

$$\widehat{x}(y;\lambda) \in \underset{x\in\mathbb{R}^N}{\operatorname{Argmin}} \ \frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_1$$

For $\gamma \in]0,+\infty[$

5.

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} y_n = \operatorname{prox}_{\gamma\lambda\|\cdot\|_1} x_n \\ z_n = \dfrac{2y_n - x_n + \gamma y}{1+\gamma} \\ x_{n+1} = x_n + z_n - y_n \end{cases}$$

$(y_n)_{n\in\mathbb{N}}$ converges toward $\widehat{x}(y;\lambda)$.

ISTA: Iterative Shrinkage-Thresholding Algorithm

Standard ISTA-like algorithm to minimize $F(x) = f(x) + g(x)$

f differentiable with β-Lipschitz gradient and $\gamma \in]0, 2/\beta[$.

Forward-backward algorithm

$$x_{n+1} = \mathrm{prox}_{\gamma g}\left(x_n - \gamma \nabla f(x_n)\right).$$

Convergence rate:

$$F(x_n) - \min F = F(x_n) - F(\widehat{x}) \leq \frac{C}{n}$$

with $C > 0$ a constant depending on the characteristics of the problem.

FISTA: **Fast** Iterative Shrinkage-Thresholding Algorithm

Accelerated ISTA to minimize $F(x) = f(x) + g(x)$

f differentiable with β-Lipschitz gradient and $\gamma \in]0, 2/\beta[$.

Forward-backward algorithm with inertia

$$y_n = \operatorname{prox}_{\gamma g}\left(x_n - \gamma \nabla f(x_n)\right)$$

$$t_{n+1} = \frac{1 + \sqrt{1 + 4t_n^2}}{2}$$

$$x_{n+1} = y_n + \frac{t_n - 1}{t_{n+1}}\left(y_n - y_{n-1}\right)$$

Convergence rate:

$$F(x_n) - \min F = F(x_n) - F(\widehat{x}) \leq \frac{C}{n^2}.$$