Chapter 5
Forward–backward algorithm

Nelly Pustelnik

Chapter objectives.

- Minimize a composite criterion $f + g$, with f smooth and g proximable, by combining a gradient step on f and a proximal step on g.
- Read an optimization algorithm as a *fixed-point iteration* and know the two convergence tools: the Banach–Picard theorem (contractions) and the Krasnosel'skiĭ–Mann theorem (averaged operators).
- Prove the convergence of the iterates of forward–backward for $\gamma \in \left]0, \frac{2}{\beta}\right[$, and its linear convergence under strong convexity; recover gradient descent as a special case.
- Know how to accelerate the algorithm: Nesterov inertia (FISTA) and variable metric (preconditioning).

Contents

1 Motivation **2**

2 The forward–backward algorithm **3**

3 Convergence of fixed-point iterations **4**

3.1 Strong and weak convergence 4

3.2 Contractive and averaged operators 4

3.3 Banach–Picard theorem 5

3.4 Iterating an averaged operator 5

3.5 Composition of averaged operators 6

4 Convergence of the forward–backward iterates **6**

4.1 The forward step 6

4.2 The backwards step 8

4.3 Convergence theorem 8

5 Linear convergence **9**

5.1 When f is strongly convex 9

5.2 When g is strongly convex 10

6 Rate of forward–backward on the values **10**

7 Nesterov acceleration **11**

7.1 The FISTA algorithm 11

Notation and reminders (Chapters 3 and 4). $\mathcal{H}$ is a real Hilbert space with inner product $\langle \cdot \,|\, \cdot \rangle$; in the applications, $\mathcal{H} = \mathbb{R}^N$. From Chapter 4, for $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$, we use:

- Fermat's rule: $\widehat{x} \in \operatorname{Argmin} f \iff 0 \in \partial f(\widehat{x})$;
- the monotonicity of ∂f: $\langle u - v \,|\, x - y \rangle \geq 0$ for $u \in \partial f(x)$, $v \in \partial f(y)$;
- the characterization of the prox: $p = \operatorname{prox}_{\gamma f}(x) \iff x - p \in \gamma\, \partial f(p)$;

For an operator $T : \mathcal{H} \to \mathcal{H}$, $\operatorname{Fix} T = \{x \in \mathcal{H} \mid Tx = x\}$ denotes its set of fixed points.
For a set-valued operator $\Phi : \mathcal{H} \to 2^{\mathcal{H}}$, $\operatorname{zer} \Phi = \{x \in \mathcal{H} \mid 0 \in \Phi x\}$ its set of zeros.

1 Motivation

Most problems met in Chapters 2 and 4 have a *composite* structure:

$$\underset{x \in \mathcal{H}}{\text{minimize}} \;\; F(x) = f(x) + g(x), \tag{1}$$

where f is a smooth data-fidelity term and g a non-smooth term encoding prior knowledge or constraints. For instance, to restore an image $\bar{x}$ from degraded observations $z = \mathrm{A}\bar{x} + \varepsilon$, we aim to minimize

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; \frac{1}{2} \left\| \mathrm{A}x - z \right\|^2 + \lambda \left\| Wx \right\|_1$$

or

$$\underset{x \in C}{\text{minimize}} \;\; \frac{1}{2} \left\| \mathrm{A}x - z \right\|^2,$$

with W an orthonormal wavelet transform and $C = [0, 255]^N$, that is, $g = \iota_C$.

None of the tools seen so far is satisfactory for solving (1):

- **gradient descent** (Chapter 3) requires F to be differentiable, which fails as soon as $g \neq 0$;
- the **subgradient method** (Chapter 4, § 5.3) applies, but it is not a descent method, requires vanishing step sizes and never produces exactly sparse iterates;
- the **proximal point algorithm** (Chapter 4, § 5.1) converges for any step size, but requires the prox of the *whole* function $F = f + g$, which is rarely computable.

The forward–backward algorithm exploits the additive structure of (1): f is handled *explicitly* by a gradient step, and g *implicitly* by its proximity operator. The names come from the discretization of the gradient flow $\dot{x} \in -\nabla f(x) - \partial g(x)$: forward (explicit) Euler for f, backward (implicit) Euler for g.

The fixed-point viewpoint. All the algorithms of this chapter are built in the same way: find an operator $T : \mathcal{H} \to \mathcal{H}$ such that

$$\operatorname{Argmin} F = \operatorname{zer} \Phi = \operatorname{Fix} T,$$

then iterate $x^{[k+1]} = Tx^{[k]}$. Two examples:

- *gradient descent*: if F is convex and differentiable, $\Phi = \nabla F$ and $T = \mathrm{Id} - \gamma \nabla F$;
- *proximal point*: if $F \in \Gamma_0(\mathcal{H})$, $\Phi = \partial F$ and $T = \operatorname{prox}_{\gamma F}$, since

$$\begin{aligned} \widehat{x} \in \operatorname{Argmin} F &\iff 0 \in \partial F(\widehat{x}) \\ &\iff \widehat{x} - \widehat{x} \in \gamma \partial F(\widehat{x}) \\ &\iff \widehat{x} = \operatorname{prox}_{\gamma F}(\widehat{x}). \end{aligned}$$

Two questions then arise: which operator T for $F = f + g$ (Section 2), and when does the iteration $x^{[k+1]} = Tx^{[k]}$ converge (Section 3)?

2 The forward–backward algorithm

Throughout the chapter, $f \in \Gamma_0(\mathcal{H}) \cap \mathcal{C}_\beta^{1,1}(\mathcal{H})$, $g \in \Gamma_0(\mathcal{H})$, and $F = f + g$. For a step size $\gamma > 0$, we set

$$T_\gamma = \operatorname{prox}_{\gamma g} \circ (\mathrm{Id} - \gamma \nabla f). \tag{2}$$

Algorithm 2.1 – (Forward–backward)

Input: $x^{[0]} \in \mathcal{H}$, a step size $\gamma \in \left]0, \frac{2}{\beta}\right[$
For $k = 0, 1, \ldots$:
$\quad x^{[k+1]} = \operatorname{prox}_{\gamma g}(x^{[k]} - \gamma \nabla f(x^{[k]}))$

Remark (Majorize–minimize interpretation)
By the characterization of the prox, the iteration also reads

$$\begin{aligned} x^{[k+1]} &= \operatorname*{arg\,min}_{x \in \mathcal{H}} \ \frac{1}{2\gamma} \left\| x - (x^{[k]} - \gamma \nabla f(x^{[k]})) \right\|^2 + g(x). \\ &= \operatorname*{arg\,min}_{x \in \mathcal{H}} \ \underbrace{f(x^{[k]}) + \left\langle \nabla f(x^{[k]}) \,\middle|\, x - x^{[k]} \right\rangle + \frac{1}{2\gamma} \left\| x - x^{[k]} \right\|^2 + g(x)}_{\widetilde{F}(x)}. \end{aligned}$$

The smooth part f is replaced by its linearization at $x^{[k]}$ plus a quadratic term, while g is kept as is. For $\gamma \leq 1/\beta$, the descent lemma (cf. Chapter 3) shows that $\widetilde{F}$ is a majorant of F that is exact at $x^{[k]}$.

Special cases.

- $g = 0$: $\operatorname{prox}_{\gamma g} = \mathrm{Id}$, this is **gradient descent** with constant step size.
- $f = 0$: the **proximal point algorithm**.
- $g = \iota_C$: $x^{[k+1]} = P_C(x^{[k]} - \gamma \nabla f(x^{[k]}))$, the **projected gradient** algorithm [12].
- $g = \lambda \|\cdot\|_1$: the prox is a soft thresholding, this is **ISTA** [11].

Proposition 2.2 – (Fixed points)

For every $\gamma > 0$, $\operatorname{Fix} T_\gamma = \operatorname{zer}(\nabla f + \partial g) = \operatorname{Argmin} F$.

Proof. By the characterization of the prox,

$$\begin{aligned} x = \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) &\iff x - \gamma \nabla f(x) - x \in \gamma \partial g(x) && (3) \\ &\iff 0 \in \nabla f(x) + \partial g(x). && (4) \end{aligned}$$

Since $\operatorname{dom} f = \mathcal{H}$, the sum rule gives $\nabla f + \partial g = \partial F$, and we conclude with Fermat's rule. ∎

Proposition 2.3 – (Sufficient decrease)

For $\gamma \in \left]0, \frac{2}{\beta}\right[$, the iterates of Algorithm 2.1 satisfy

$$F(x^{[k+1]}) \leq F(x^{[k]}) - \left(\frac{1}{\gamma} - \frac{\beta}{2}\right) \left\|x^{[k+1]} - x^{[k]}\right\|^2.$$

In particular, $\left(F(x^{[k]})\right)_{k\in\mathbb{N}}$ is nonincreasing: forward–backward is a descent method.

Proof. By the characterization of the prox,

$$\frac{1}{\gamma}(x^{[k]} - x^{[k+1]}) - \nabla f(x^{[k]}) \in \partial g(x^{[k+1]}),$$

hence, by definition of subdifferential

$$g(x^{[k]}) \geq g(x^{[k+1]}) - \left\langle \frac{1}{\gamma}(x^{[k]} - x^{[k+1]}) - \nabla f(x^{[k]}) \,\middle|\, x^{[k]} - x^{[k+1]} \right\rangle \tag{5}$$

$$\geq g(x^{[k+1]}) + \frac{1}{\gamma}\left\|x^{[k]} - x^{[k+1]}\right\|^2 - \left\langle \nabla f(x^{[k]}) \,\middle|\, x^{[k]} - x^{[k+1]} \right\rangle. \tag{6}$$

The descent lemmagives $f(x^{[k+1]}) \leq f(x^{[k]}) + \left\langle \nabla f(x^{[k]}) \,\middle|\, x^{[k+1]} - x^{[k]} \right\rangle + \frac{\beta}{2}\left\|x^{[k+1]} - x^{[k]}\right\|^2$.
Summing the two inequalities yields the result. ∎

3 Convergence of fixed-point iterations

3.1 Strong and weak convergence

Let $(x^{[k]})_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ and $\widehat{x} \in \mathcal{H}$.

- $(x^{[k]})_{k\in\mathbb{N}}$ *converges strongly* to $\widehat{x}$, denoted $x^{[k]} \to \widehat{x}$, if $\lim_{k\to\infty} \left\|x^{[k]} - \widehat{x}\right\| = 0$.
- $(x^{[k]})_{k\in\mathbb{N}}$ *converges weakly* to $\widehat{x}$, denoted $x^{[k]} \rightharpoonup \widehat{x}$, if $(\forall u \in \mathcal{H})\ \lim_{k\to\infty} \left\langle u \,\middle|\, x^{[k]} - \widehat{x} \right\rangle = 0$.

In a finite-dimensional space, the two notions coincide.

3.2 Contractive and averaged operators

Definition 3.1 – (Lipschitz and averaged operators)

Let $T : \mathcal{H} \to \mathcal{H}$.

- T is ω-*Lipschitz* ($\omega > 0$) if $(\forall x, u \in \mathcal{H})\ \|Tx - Tu\| \leq \omega \|x - u\|$.

  It is a *contraction* if $\omega < 1$.

  It is *nonexpansive* if $\omega = 1$.
- T is μ-*averaged* for some $\mu \in]0, 1]$ if, for all $x, u \in \mathcal{H}$,

$$\|Tx - Tu\|^2 \leq \|x - u\|^2 - \frac{1-\mu}{\mu}\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)u\|^2. \tag{7}$$

  It is *firmly nonexpansive* if $\mu = \frac{1}{2}$.

Remark One shows that T is μ-averaged if and only if $T = (1-\mu)\mathrm{Id} + \mu R$ for some nonexpansive operator R: T is an average of the identity and of a nonexpansive operator, hence the name.

Proposition 3.2 – (prox_g is firmly nonexpansive)

Let $g \in \Gamma_0(\mathcal{H})$. Then, prox_g is $\frac{1}{2}$-averaged.

Proof. Let $x, y \in \mathcal{H}$, $p = \mathrm{prox}_g\, x$ and $q = \mathrm{prox}_g\, y$.

By Chapter 4 (Proposition on non-expansiveness), $\|p-q\|^2 \leq \langle p-q \,|\, x-y\rangle$. Let us recall why: $p = \mathrm{prox}_g\, x$ is characterized by $x - p \in \partial g(p)$, and similarly $y - q \in \partial g(q)$. By monotonicity of ∂g,

$$\langle (x-p)-(y-q) \,|\, p-q\rangle \geq 0, \quad \text{that is} \quad \langle x-y \,|\, p-q\rangle \geq \|p-q\|^2 .$$

With $a = x-y$ and $b = p-q$, the identity $\|a-b\|^2 = \|a\|^2 - 2\langle a \,|\, b\rangle + \|b\|^2$ gives

$$\|(x-p)-(y-q)\|^2 = \|x-y\|^2 - 2\langle x-y \,|\, p-q\rangle + \|p-q\|^2 .$$

By Step 1, $-2\langle x-y \,|\, p-q\rangle \leq -2\|p-q\|^2$, hence

$$\|(x-p)-(y-q)\|^2 \leq \|x-y\|^2 - \|p-q\|^2 ,$$

which is the announced inequality.

Set $T = \mathrm{prox}_g$. Then $p-q = Tx - Ty$ and $(x-p)-(y-q) = (\mathrm{Id}-T)x - (\mathrm{Id}-T)y$, so the inequality reads

$$\|Tx-Ty\|^2 \leq \|x-y\|^2 - \|(\mathrm{Id}-T)x-(\mathrm{Id}-T)y\|^2 .$$

This is exactly (7) with $\frac{1-\mu}{\mu} = 1$, that is $\mu = \frac{1}{2}$. ∎

Remark The inequality is symmetric in T and $\mathrm{Id}-T$: the operator $\mathrm{Id} - \mathrm{prox}_g = \mathrm{prox}_{g^*}$ (Moreau decomposition) is also firmly nonexpansive.

3.3 Banach–Picard theorem

Theorem 3.3 – (Banach–Picard [17, 18])

Let $\omega \in [0,1[$ and $T : \mathcal{H} \to \mathcal{H}$ be ω-Lipschitz. Then T has a unique fixed point $\widehat{x}$ and, for every $x^{[0]} \in \mathcal{H}$, the sequence $x^{[k+1]} = Tx^{[k]}$ satisfies

$$(\forall k \in \mathbb{N}) \qquad \left\|x^{[k]} - \widehat{x}\right\| \leq \omega^k \left\|x^{[0]} - \widehat{x}\right\| .$$

Hence $(x^{[k]})_{k\in\mathbb{N}}$ converges strongly to $\widehat{x}$, with linear rate ω.

See also [23, Theorem 1.50].

3.4 Iterating an averaged operator

For a merely nonexpansive operator, the iteration may fail to converge: think of $T = -\mathrm{Id}$, or of a rotation. Averagedness is enough to restore convergence, but without a rate.

Theorem 3.4 – (Krasnosel'skiĭ–Mann [19, 20])

Let $\mu \in]0,1[$ and $T : \mathcal{H} \to \mathcal{H}$ be μ-averaged with $\operatorname{Fix} T \neq \varnothing$. For every $x^{[0]} \in \mathcal{H}$, the sequence $x^{[k+1]} = Tx^{[k]}$ converges weakly to a point of $\operatorname{Fix} T$.

This statement is in [23, Proposition 5.16]; the weak convergence argument in Hilbert spaces goes back to [21]. See also [22] for the use of averaged operators in splitting algorithms.

3.5 Composition of averaged operators

Proposition 3.5 – (Composition [10])

Let $\mu_1, \mu_2 \in]0,1[$, T_1 be μ_1-averaged and T_2 be μ_2-averaged. Then $T_1 \circ T_2$ is μ-averaged with

$$\mu = \frac{\mu_1 + \mu_2 - 2\mu_1\mu_2}{1 - \mu_1\mu_2} \in]0,1[.$$

4 Convergence of the forward–backward iterates

By Proposition 2.2, to apply Theorem 3.4 it is enough to show that $T_\gamma = \operatorname{prox}_{\gamma g} \circ (\operatorname{Id} - \gamma \nabla f)$ is averaged. We will first study, the forward and backward step is separatly.

4.1 The forward step

Lemma 4.1 – (Descent lemma)

If f is β-smooth, then for all $x, y \in \mathcal{H}$,

$$f(y) \leq f(x) + \langle \nabla f(x) \,|\, y - x \rangle + \frac{\beta}{2} \|y - x\|^2 .$$

Definition 4.2 – (Cocoercivity)

Let $\theta > 0$. An operator $M : \mathcal{H} \to \mathcal{H}$ is *θ-cocoercive* if

$$(\forall x, y \in \mathcal{H}) \qquad \langle x - y \,|\, Mx - My \rangle \geq \theta \|Mx - My\|^2 .$$

Remark By Cauchy–Schwarz, a θ-cocoercive operator is $\frac{1}{\theta}$-Lipschitz. The converse is false in general (a rotation by $\pi/2$ is 1-Lipschitz but $\langle x \,|\, Mx \rangle = 0$), but it holds for gradients of convex functions.

Theorem 4.3 – (Baillon–Haddad [1])

If f is convex and gradient β-Lipschitz, then ∇f is $\frac{1}{\beta}$-cocoercive.

Proof. Fix $x \in \mathcal{H}$ and define $h : \mathcal{H} \to \mathbb{R}$ such that $h(y) = f(y) - \langle \nabla f(x) \,|\, y \rangle$. h is convex, as the sum of the convex function f and of a linear function. h is differentiable with $\nabla h(y) = \nabla f(y) - \nabla f(x)$. Hence, for every $y, z \in \mathcal{H}$, $\|\nabla h(y) - \nabla h(z)\| = \|\nabla f(y) - \nabla f(z)\| \leq \beta \|y - z\|$: h has gradient β-Lipschitz, and the descent lemma applies to h. Finally, $\nabla h(x) = 0$ and since h is convex, the gradient inequality gives, for every $z \in \mathcal{H}$, $h(z) \geq h(x) + \langle \nabla h(x) \,|\, z - x \rangle = h(x)$. Hence x is a (global) minimizer of h.

Fix $y \in \mathcal{H}$ and set $z = y - \frac{1}{\beta}\nabla h(y)$. The descent lemma applied to h between y and z gives

$$\begin{aligned} h(z) &\leq h(y) + \langle \nabla h(y) \,|\, z - y \rangle + \frac{\beta}{2} \|z - y\|^2 \\ &= h(y) - \frac{1}{\beta} \|\nabla h(y)\|^2 + \frac{\beta}{2} \cdot \frac{1}{\beta^2} \|\nabla h(y)\|^2 \\ &= h(y) - \frac{1}{2\beta} \|\nabla h(y)\|^2 . \end{aligned}$$

Since x minimizes h, $h(x) \leq h(z)$, hence

$$h(x) \leq h(y) - \frac{1}{2\beta} \|\nabla h(y)\|^2 . \tag{8}$$

We have $h(x) = f(x) - \langle \nabla f(x) \,|\, x \rangle$, $h(y) = f(y) - \langle \nabla f(x) \,|\, y \rangle$ and $\nabla h(y) = \nabla f(y) - \nabla f(x)$. Substituting in (8),

$$f(x) - \langle \nabla f(x) \,|\, x \rangle \leq f(y) - \langle \nabla f(x) \,|\, y \rangle - \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|^2 ,$$

that is, for every $x, y \in \mathcal{H}$,

$$f(y) \geq f(x) + \langle \nabla f(x) \,|\, y - x \rangle + \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|^2 . \tag{9}$$

Since (9) holds for all x, y, we may exchange their roles:

$$f(x) \geq f(y) + \langle \nabla f(y) \,|\, x - y \rangle + \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2 .$$

Adding these two inequalities, the terms $f(x)$ and $f(y)$ cancel:

$$0 \geq \langle \nabla f(x) \,|\, y - x \rangle + \langle \nabla f(y) \,|\, x - y \rangle + \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2 .$$

Since $\langle \nabla f(x) \,|\, y - x \rangle + \langle \nabla f(y) \,|\, x - y \rangle = - \langle \nabla f(x) - \nabla f(y) \,|\, x - y \rangle$, we obtain

$$\langle \nabla f(x) - \nabla f(y) \,|\, x - y \rangle \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2 .$$

■

Proposition 4.4 – (Averagedness of the gradient step)

If M is θ-cocoercive and $\gamma \in]0, 2\theta[$, then $\mathrm{Id} - \gamma M$ is $\frac{\gamma}{2\theta}$-averaged. In particular, under the standing assumptions, $\mathrm{Id} - \gamma \nabla f$ is $\frac{\gamma\beta}{2}$-averaged for every $\gamma \in \left]0, \frac{2}{\beta}\right[$.

Proof. Let $T = \mathrm{Id} - \gamma M$, let $x, y \in \mathcal{H}$, and set $u = x - y$ and $v = Mx - My$, so that $Tx - Ty = u - \gamma v$. By the identity $\|a - b\|^2 = \|a\|^2 - 2\langle a \,|\, b \rangle + \|b\|^2$,

$$\|Tx - Ty\|^2 = \|u - \gamma v\|^2 = \|u\|^2 - 2\gamma \langle u \,|\, v \rangle + \gamma^2 \|v\|^2 .$$

Since M is θ-cocoercive, $\langle u \,|\, v \rangle \geq \theta \|v\|^2$. As $\gamma > 0$, this gives $-2\gamma \langle u \,|\, v \rangle \leq -2\gamma\theta \|v\|^2$, hence

$$\|Tx - Ty\|^2 \leq \|u\|^2 + (\gamma^2 - 2\gamma\theta) \|v\|^2 = \|x - y\|^2 - \gamma(2\theta - \gamma) \|Mx - My\|^2 .$$

Note that $2\theta - \gamma > 0$, so the last term is indeed a nonpositive correction.

Since $\mathrm{Id} - T = \gamma M$, we have $(\mathrm{Id} - T)x - (\mathrm{Id} - T)y = \gamma(Mx - My)$, so that $\|Mx - My\|^2 = \frac{1}{\gamma^2}\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2$. Substituting,

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{2\theta - \gamma}{\gamma}\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 .$$

This is exactly (7) provided that

$$\frac{1-\mu}{\mu} = \frac{2\theta - \gamma}{\gamma} \iff \frac{1}{\mu} - 1 = \frac{2\theta}{\gamma} - 1 \iff \mu = \frac{\gamma}{2\theta}.$$

Since $\gamma \in]0, 2\theta[$, we have $\mu \in]0, 1[$, hence T is $\frac{\gamma}{2\theta}$-averaged.

If f is convex and β-smooth, Theorem 4.3 asserts that ∇f is $\frac{1}{\beta}$-cocoercive. Applying the above with $M = \nabla f$ and $\theta = \frac{1}{\beta}$, for every $\gamma \in \left]0, \frac{2}{\beta}\right[$ the operator $\mathrm{Id} - \gamma\nabla f$ is $\frac{\gamma\beta}{2}$-averaged. ■

Corollary 4.5 – (Gradient descent)

Let $f : \mathcal{H} \to \mathbb{R}$ be convex and β-smooth with $\operatorname{Argmin} f \neq \varnothing$, and $\gamma \in \left]0, \frac{2}{\beta}\right[$. Then $x^{[k+1]} = x^{[k]} - \gamma\nabla f(x^{[k]})$ converges weakly (strongly in finite dimension) to a minimizer of f.

Proof. Take $g = 0$ in Theorem 4.7. Directly: $\mathrm{Id} - \gamma\nabla f$ is $\frac{\gamma\beta}{2}$-averaged (Proposition 4.4) and $\operatorname{Fix}(\mathrm{Id} - \gamma\nabla f) = \{x \mid \nabla f(x) = 0\} = \operatorname{Argmin} f$. ■

4.2 The backwards step

Corollary 4.6 – (Proximal point algorithm)

Let $g \in \Gamma_0(\mathcal{H})$ with $\operatorname{Argmin} g \neq \varnothing$. For every $\gamma > 0$, the sequence $x^{[k+1]} = \operatorname{prox}_{\gamma g}(x^{[k]})$ converges weakly to a minimizer of g.

Proof. $\operatorname{prox}_{\gamma g}$ is $\frac{1}{2}$-averaged and $\operatorname{Fix}\operatorname{prox}_{\gamma g} = \operatorname{Argmin} g$ (Chapter 4); apply Theorem 3.4. ■

This answers the question left open in Chapter 4: the implicit scheme converges for *any* step size.

4.3 Convergence theorem

Theorem 4.7 – (Convergence of forward–backward [9])

Assume $\operatorname{Argmin} F \neq \varnothing$ and $\gamma \in \left]0, \frac{2}{\beta}\right[$. Then the sequence $(x^{[k]})_{k\in\mathbb{N}}$ generated by Algorithm 2.1 converges weakly (strongly in finite dimension) to a minimizer $\widehat{x}$ of F, and $\left(F(x^{[k]})\right)_{k\in\mathbb{N}}$ is nonincreasing and converges to $\min F$.

Proof. By Proposition 4.4, $\mathrm{Id} - \gamma\nabla f$ is μ_2-averaged with $\mu_2 = \frac{\gamma\beta}{2} \in]0, 1[$, and by **??**, $\operatorname{prox}_{\gamma g}$ is μ_1-averaged with $\mu_1 = \frac{1}{2}$. By Proposition 3.5, T_γ is μ-averaged with

$$\mu = \frac{\frac{1}{2} + \mu_2 - \mu_2}{1 - \frac{\mu_2}{2}} = \frac{2}{4 - \gamma\beta} \in \left]\tfrac{1}{2}, 1\right[.$$

Since $\operatorname{Fix} T_\gamma = \operatorname{Argmin} F \neq \varnothing$ (Proposition 2.2), Theorem 3.4 gives the convergence of the iterates. The monotonicity of the values is Proposition 2.3; their convergence to $\min F$, with a rate, is Theorem 6.2 below for $\gamma \leq 1/\beta$ (see [9] for the general case). ■

> **Remark** The admissible step sizes $\gamma < 2/\beta$ are exactly those of gradient descent: the proximal step does not restrict the step size, which only depends on the regularity of f.

5 Linear convergence

Theorem 4.7 gives no rate on the iterates. Under strong convexity, T_γ becomes a contraction, and the Banach–Picard theorem gives a linear rate.

5.1 When f is strongly convex

Lemma 5.1

If f is ρ-strongly convex and β-smooth (hence $\rho \leq \beta$), then $\varphi = f - \frac{\rho}{2}\|\cdot\|^2$ is convex and gradient $(\beta-\rho)$-Lipschitz.

Proof. The convexity of φ is the definition of strong convexity. Moreover, $\frac{\beta-\rho}{2}\|\cdot\|^2 - \varphi = \frac{\beta}{2}\|\cdot\|^2 - f$ is convex, since its gradient $\beta\mathrm{Id} - \nabla f$ is monotone by Cauchy–Schwarz. For a convex φ, the convexity of $\frac{L}{2}\|\cdot\|^2 - \varphi$ is equivalent to $\nabla\varphi$ being L-Lipschitz [14, Th. 2.1.5], which we admit. ■

Theorem 5.2 – (Contraction of the gradient step [5])

Assume that f is ρ-strongly convex and $\gamma \in \left]0, \frac{2}{\beta}\right[$. Then $\mathrm{Id} - \gamma\nabla f$, and hence T_γ, is $\omega(\gamma)$-Lipschitz with

$$\omega(\gamma) = \max\{|1-\gamma\rho|, |1-\gamma\beta|\} \in]0,1[.$$

Consequently F has a unique minimizer $\widehat{x}$ and $\left\|x^{[k]} - \widehat{x}\right\| \leq \omega(\gamma)^k \left\|x^{[0]} - \widehat{x}\right\|$.

Proof. With φ as in Lemma 5.1, $\mathrm{Id} - \gamma\nabla f = (1-\gamma\rho)\mathrm{Id} - \gamma\nabla\varphi$. Let $x, y \in \mathcal{H}$, $u = x - y$, $v = \nabla\varphi(x) - \nabla\varphi(y)$ and $L = \beta - \rho \geq 0$. By Theorem 4.3, $\nabla\varphi$ is $\frac{1}{L}$-cocoercive, that is $\|v\|^2 \leq L\langle u \,|\, v\rangle$; in particular $\langle u \,|\, v\rangle \geq 0$ and, by Cauchy–Schwarz, $\langle u \,|\, v\rangle \leq L\|u\|^2$. Expanding and using the cocoercivity,

$$\begin{aligned}
\|(1-\gamma\rho)u - \gamma v\|^2 &= (1-\gamma\rho)^2\|u\|^2 - 2\gamma(1-\gamma\rho)\langle u \,|\, v\rangle + \gamma^2\|v\|^2 \\
&\leq (1-\gamma\rho)^2\|u\|^2 + \gamma(\gamma L - 2(1-\gamma\rho))\langle u \,|\, v\rangle \\
&= (1-\gamma\rho)^2\|u\|^2 + \gamma(\gamma(\beta+\rho) - 2)\langle u \,|\, v\rangle .
\end{aligned}$$

Since $0 \leq \langle u \,|\, v\rangle \leq L\|u\|^2$, we obtain

$$\|(1-\gamma\rho)u - \gamma v\|^2 \leq \Big((1-\gamma\rho)^2 + \gamma L \max\{0, \gamma(\beta+\rho) - 2\}\Big)\|u\|^2 .$$

If $\gamma(\beta+\rho) \leq 2$, the bound is $(1-\gamma\rho)^2\|u\|^2$. Otherwise, it equals

$$(1 - 2\gamma\rho + \gamma^2\rho^2 + \gamma^2(\beta^2 - \rho^2) - 2\gamma(\beta-\rho))\|u\|^2 = (1-\gamma\beta)^2\|u\|^2 .$$

In both cases, $\|(\mathrm{Id} - \gamma\nabla f)x - (\mathrm{Id} - \gamma\nabla f)y\| \leq \omega(\gamma)\|x - y\|$, and $\omega(\gamma) < 1$ since $0 < \gamma\rho \leq \gamma\beta < 2$. As $\operatorname{prox}_{\gamma g}$ is nonexpansive, T_γ is $\omega(\gamma)$-Lipschitz, and Theorem 3.3 applies with $\operatorname{Fix} T_\gamma = \operatorname{Argmin} F$. ■

5.2 When g is strongly convex

Theorem 5.3 – (Contraction of the proximal step [5])

Assume that g is ρ-strongly convex and $\gamma \in \left]0, \frac{2}{\beta}\right]$. Then T_γ is $\frac{1}{1+\gamma\rho}$-Lipschitz, and

$$\left\|x^{[k]} - \widehat{x}\right\| \leq \left(\frac{1}{1+\gamma\rho}\right)^k \left\|x^{[0]} - \widehat{x}\right\|.$$

The best rate is obtained for $\gamma = 2/\beta$ and equals $\frac{\beta}{\beta+2\rho}$.

Proof. If g is ρ-strongly convex, ∂g is *strongly monotone*: summing $g(y) \geq g(x) + \langle u \,|\, y - x\rangle + \frac{\rho}{2}\|y-x\|^2$ and the symmetric inequality, $\langle u - v \,|\, x - y\rangle \geq \rho\|x-y\|^2$ for $u \in \partial g(x)$, $v \in \partial g(y)$. Let $p_i = \operatorname{prox}_{\gamma g}(x_i)$, $i \in \{1,2\}$. Since $x_i - p_i \in \gamma\partial g(p_i)$, $\langle (x_1 - p_1) - (x_2 - p_2) \,|\, p_1 - p_2\rangle \geq \gamma\rho\|p_1 - p_2\|^2$, that is, $\langle x_1 - x_2 \,|\, p_1 - p_2\rangle \geq (1+\gamma\rho)\|p_1 - p_2\|^2$. By Cauchy–Schwarz, $\operatorname{prox}_{\gamma g}$ is $\frac{1}{1+\gamma\rho}$-Lipschitz. For $\gamma \leq 2/\beta$, $\mathrm{Id} - \gamma\nabla f$ is nonexpansive (Proposition 4.4, extended to $\gamma = 2\theta$ where $\mu = 1$), and we conclude by composition and Theorem 3.3. ■

6 Rate of forward–backward on the values

Without strong convexity, a rate can still be obtained on the values of F. The key tool is the following inequality.

Lemma 6.1 – (Fundamental inequality [3])

Let $\gamma \in \left]0, \frac{1}{\beta}\right]$, $y \in \mathcal{H}$ and $p = \operatorname{prox}_{\gamma g}(y - \gamma\nabla f(y))$. For every $x \in \mathcal{H}$,

$$F(x) - F(p) \geq \frac{1}{2\gamma}\Big(\|p - x\|^2 - \|y - x\|^2\Big).$$

Proof. By the descent lemma with $\beta \leq 1/\gamma$, by convexity of f, and since $u = \frac{1}{\gamma}(y-p) - \nabla f(y) \in \partial g(p)$:

$$\begin{aligned}
f(p) &\leq f(y) + \langle \nabla f(y) \,|\, p - y\rangle + \tfrac{1}{2\gamma}\|p - y\|^2, \\
f(x) &\geq f(y) + \langle \nabla f(y) \,|\, x - y\rangle, \\
g(x) &\geq g(p) + \langle u \,|\, x - p\rangle.
\end{aligned}$$

Combining the three inequalities, yields

$$\begin{aligned}
F(x) - F(p) &= (f(x) - f(p)) + (g(x) - g(p)) \\
&\geq \langle \nabla f(y) \,|\, x - y\rangle - \langle \nabla f(y) \,|\, p - y\rangle - \frac{1}{2\gamma}\|p - y\|^2 + \langle u \,|\, x - p\rangle \\
&= \langle \nabla f(y) + u \,|\, x - p\rangle - \frac{1}{2\gamma}\|p - y\|^2.
\end{aligned}$$

By definition of u, $\nabla f(y) + u = \frac{1}{\gamma}(y - p)$. Hence

$$F(x) - F(p) \geq \frac{1}{\gamma}\langle y - p \,|\, x - p\rangle - \frac{1}{2\gamma}\|y - p\|^2.$$

■

Theorem 6.2 – (Sublinear rate of forward–backward)

Assume $\operatorname{Argmin} F \neq \varnothing$ and $\gamma \in \left]0, \frac{1}{\beta}\right]$. For every $k \geq 1$ and every $\widehat{x} \in \operatorname{Argmin} F$,

$$F(x^{[k]}) - \min F \leq \frac{\left\|x^{[0]} - \widehat{x}\right\|^2}{2\gamma k}.$$

Proof. Lemma 6.1 with $y = x^{[k]}$, $p = x^{[k+1]}$ and $x = \widehat{x}$ gives $2\gamma(F(x^{[k+1]}) - \min F) \leq \left\|x^{[k]} - \widehat{x}\right\|^2 - \left\|x^{[k+1]} - \widehat{x}\right\|^2$. Summing for $k = 0, \ldots, K-1$, the right-hand side telescopes: $2\gamma \sum_{k=1}^{K}(F(x^{[k]}) - \min F) \leq \left\|x^{[0]} - \widehat{x}\right\|^2$. Since $(F(x^{[k]}))_k$ is nonincreasing (Proposition 2.3), the left-hand side is at least $2\gamma K(F(x^{[K]}) - \min F)$. ∎

With $\gamma = 1/\beta$, forward–backward thus has an $O(1/k)$ rate on the values, like gradient descent.

7 Nesterov acceleration

7.1 The FISTA algorithm

Nesterov [13] showed that an *inertial* (or *momentum*) step, which extrapolates in the direction of the last displacement, improves this rate at almost no extra cost. Beck and Teboulle [3] extended it to forward–backward.

Algorithm 7.1 – (FISTA)

Input: $x^{[0]} \in \mathcal{H}$, $\gamma \in \left]0, \frac{1}{\beta}\right]$.
$y^{[0]} = x^{[0]}$, $t_0 = 1$
For $k = 0, 1, \ldots$:

$$x^{[k+1]} = \operatorname{prox}_{\gamma g}(y^{[k]} - \gamma \nabla f(y^{[k]}))$$
$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$
$$y^{[k+1]} = x^{[k+1]} + \frac{t_k - 1}{t_{k+1}}(x^{[k+1]} - x^{[k]})$$

The forward–backward step is applied at the extrapolated point $y^{[k]}$, not at $x^{[k]}$. The inertial coefficient $\frac{t_k - 1}{t_{k+1}}$ equals 0 at the first iteration and tends to 1.

Theorem 7.2 – (Rate of FISTA [3])

Assume $\operatorname{Argmin} F \neq \varnothing$ and $\gamma \in \left]0, \frac{1}{\beta}\right]$. The iterates of FISTA satisfy, for every $k \geq 1$ and every $\widehat{x} \in \operatorname{Argmin} F$,

$$F(x^{[k]}) - \min F \leq \frac{2\left\|x^{[0]} - \widehat{x}\right\|^2}{\gamma(k+1)^2}.$$

Sketch of proof. The sequence (t_k) satisfies $t_{k+1}^2 - t_{k+1} = t_k^2$ and $t_{k+1} \geq t_k + \frac{1}{2}$, hence $t_k \geq \frac{k+2}{2}$. Apply Lemma 6.1 with $y = y^{[k]}$ and $p = x^{[k+1]}$, once with $x = x^{[k]}$ and once with $x = \widehat{x}$, and combine the two inequalities with weights $t_k - 1$ and 1. The relation $t_{k+1}^2 - t_{k+1} = t_k^2$ then shows

that the energy

$$E_k = 2\gamma\, t_{k-1}^2 (F(x^{[k]}) - \min F) + \left\| t_{k-1} x^{[k]} - (t_{k-1}-1) x^{[k-1]} - \widehat{x} \right\|^2, \qquad k \geq 1,$$

is nonincreasing, and Lemma 6.1 at the first iteration gives $E_1 \leq \left\| x^{[0]} - \widehat{x} \right\|^2$. Hence $F(x^{[k]}) - \min F \leq \frac{\left\| x^{[0]} - \widehat{x} \right\|^2}{2\gamma t_{k-1}^2} \leq \frac{2\left\| x^{[0]} - \widehat{x} \right\|^2}{\gamma (k+1)^2}$. ■

Remark

- The $O(1/k^2)$ rate is optimal: no first-order method can do better, in the worst case, on the class of convex β-smooth functions [14].
- Unlike forward–backward, FISTA is not a descent method: $\left(F(x^{[k]})\right)_k$ is not necessarily nonincreasing, and oscillations are often observed. Restart strategies, which reset t_k to 1, reduce them [15].
- The convergence of the iterates $(x^{[k]})$ is not known for the above choice of t_k. Chambolle and Dossal [6] showed that with $t_k = \frac{k+a-1}{a}$ and $a > 2$, the $O(1/k^2)$ rate is preserved and the iterates converge (weakly) to a minimizer.
- If f is ρ-strongly convex, one uses the constant inertial coefficient $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$: the linear rate becomes $1 - \frac{1}{\sqrt{\kappa}}$ instead of $1 - \frac{1}{\kappa}$, and the number of iterations drops from $O(\kappa \log(1/\varepsilon))$ to $O(\sqrt{\kappa} \log(1/\varepsilon))$.

References

[1] J.-B. Baillon and G. Haddad, "Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones", *Israel J. Math.*, 26:137–150, 1977.

[2] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed., Springer, 2017.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems", *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

[4] S. Becker and J. Fadili, "A quasi-Newton proximal splitting method", *Advances in Neural Information Processing Systems*, 2012.

[5] L. M. Briceño-Arias and N. Pustelnik, "Theoretical and numerical comparison of first order algorithms for cocoercive equations and smooth convex optimization", *Signal Processing*, 2023.

[6] A. Chambolle and C. Dossal, "On the convergence of the iterates of the 'fast iterative shrinkage/thresholding algorithm"', *J. Optim. Theory Appl.*, 166(3):968–982, 2015.

[7] E. Chouzenoux, J.-C. Pesquet and A. Repetti, "Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function", *J. Optim. Theory Appl.*, 162(1):107–132, 2014.

[8] P. L. Combettes and B. C. Vũ, "Variable metric forward–backward splitting with applications to monotone inclusions in duality", *Optimization*, 63(9):1289–1318, 2014.

[9] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting", *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[10] P. L. Combettes and I. Yamada, "Compositions and convex combinations of averaged nonexpansive operators", *J. Math. Anal. Appl.*, 425(1):55–70, 2015.

[11] I. Daubechies, M. Defrise and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint", *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[12] E. S. Levitin and B. T. Polyak, "Constrained minimization methods", *USSR Comput. Math. Math. Phys.*, 6(5):1–50, 1966.

[13] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$", *Soviet Math. Dokl.*, 27:372–376, 1983.

[14] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, 2004.

[15] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes", *Found. Comput. Math.*, 15(3):715–732, 2015.

[16] Z. Opial, "Weak convergence of the sequence of successive approximations for nonexpansive mappings", *Bull. Amer. Math. Soc.*, 73:591–597, 1967.

[17] É. Picard, Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives, *J. Math. Pures Appl.*, vol. 6, pp. 145–210, 1890.

[18] S. Banach, Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales, *Fund. Math.*, vol. 3, pp. 133–181, 1922.

[19] W. R. Mann, Mean value methods in iteration, *Proc. Amer. Math. Soc.*, vol. 4, no. 3, pp. 506–510, 1953.

[20] M. A. Krasnosel'skiĭ, Two remarks on the method of successive approximations, *Uspekhi Mat. Nauk*, vol. 10, no. 1, pp. 123–127, 1955 (in Russian).

[21] Z. Opial, Weak convergence of the sequence of successive approximations for nonexpansive mappings, *Bull. Amer. Math. Soc.*, vol. 73, no. 4, pp. 591–597, 1967.

[22] P. L. Combettes, Solving monotone inclusions via compositions of nonexpansive averaged operators, *Optimization*, vol. 53, no. 5–6, pp. 475–504, 2004.

[23] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed., Springer, Cham, 2017.