Chapter 6
Duality and primal–dual algorithms

Nelly Pustelnik

Chapter objectives.

- Understand why forward–backward is not enough for criteria of the form $f(x) + g(\mathrm{L}x)$: the proximity operator of $g \circ \mathrm{L}$ is in general not explicit.
- Master the convex conjugate, the Fenchel–Young inequality and the Moreau decomposition, which allow us to compute prox_{g^*} from prox_g.
- Derive the Fenchel–Rockafellar dual problem, the weak and strong duality results and the Kuhn–Tucker optimality conditions.
- Solve the dual problem by forward–backward (dual FB), and the primal and dual problems jointly by primal–dual algorithms (Chambolle–Pock, Condat–Vũ), which only use L, $\mathrm{L}^\top$ and the proximity operators of f and g.

Contents

1 Motivation: when $\mathrm{prox}_{g\circ L}$ is not explicit **3**
1.1 The problem . . . 3
1.2 Why forward–backward is not enough . . . 3

2 The convex conjugate **4**
2.1 Definition and geometric interpretation . . . 4
2.2 Basic properties . . . 5
2.3 Examples . . . 6
2.4 Moreau decomposition . . . 6
2.5 Strong convexity and smoothness are dual . . . 7

3 Fenchel–Rockafellar duality **8**
3.1 Primal and dual problems . . . 8
3.2 Weak and strong duality . . . 8
3.3 Optimality conditions . . . 9

4 Dual forward–backward and denoising problem **9**
4.1 Dual of TV denoising . . . 9
4.2 Algorithm . . . 9

5 Primal–dual algorithms **10**
5.1 Saddle-point formulation and Kuhn–Tucker conditions . . . 10
5.2 From the Kuhn–Tucker conditions to fixed-point equations . . . 11
5.3 The Chambolle–Pock algorithm . . . 11
5.4 Adding a smooth term: the Condat–Vũ algorithm . . . 12

6 Augmented Lagrangian and ADMM **12**
6.1 From the Lagrangian to the augmented Lagrangian . . . 13
6.2 The ADMM algorithm . . . 13
6.3 Summary . . . 14

7 Primal–dual algorithms **14**
7.1 A monotone inclusion in the product space . . . 14
7.2 The Chambolle–Pock algorithm . . . 15
7.3 Adding a smooth term: the Condat–Vũ algorithm . . . 16

8 Augmented Lagrangian and ADMM **16**

9 Summary **17**

Notation and reminders (Chapters 4 and 5). $\mathcal{H}$ and $\mathcal{G}$ are real Hilbert spaces (in the applications, $\mathcal{H} = \mathbb{R}^N$ and $\mathcal{G} = \mathbb{R}^M$), and $\mathrm{L} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ is a bounded linear operator with adjoint $\mathrm{L}^\top$ (the transpose for matrices). 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 characterization of the prox: $p = \operatorname{prox}_{\gamma f}(x) \iff x - p \in \gamma\, \partial f(p)$;
- the convergence of forward–backward for $f + g$, with f β-smooth and $\gamma \in \left]0, \frac{2}{\beta}\right[$ (Chapter 5, Theorem on the convergence of forward–backward);
- the Krasnosel'skiĭ–Mann theorem: iterating an averaged operator with a nonempty set of fixed points converges to a fixed point.

For $\lambda > 0$ and a norm $\|\cdot\|$, $\mathcal{B}_{\|\cdot\|}(0, \lambda)$ denotes the closed ball of radius λ. The clipping operator is $\operatorname{clip}_\lambda(u) = \left(\max\{-\lambda, \min\{u_m, \lambda\}\}\right)_m$, i.e. the projection onto $\mathcal{B}_{\|\cdot\|_\infty}(0, \lambda)$.

1 Motivation: when $\operatorname{prox}_{g \circ \mathrm{L}}$ is not explicit

1.1 The problem

Many restoration problems take the form

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

where $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathrm{L} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ maps the signal to a domain where the prior information is easy to express. Typical examples, with $z = \mathrm{A}\bar{x} + \varepsilon$ the observations:

- **Total variation (TV).** L is a discrete gradient, $\mathrm{L}x = (\nabla_1 x, \nabla_2 x) \in \mathbb{R}^N \times \mathbb{R}^N$, and $g = \lambda \|\cdot\|_{2,1}$ (isotropic TV) or $g = \lambda \|\cdot\|_1$ (anisotropic TV):

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; \frac{1}{2}\|\mathrm{A}x - z\|^2 + \lambda\, \mathrm{TV}(x), \qquad \mathrm{TV}(x) = \sum_{n=1}^{N} \sqrt{(\nabla_1 x)_n^2 + (\nabla_2 x)_n^2}.$$

- **Analysis sparsity in a redundant frame.** $\mathrm{L} \in \mathbb{R}^{K \times N}$ is the analysis operator of a wavelet frame (undecimated wavelets, dual-tree) and $g = \lambda \|\cdot\|_1$.
- **Several priors.** $\sum_{i=1}^m g_i(\mathrm{L}_i x)$, e.g. TV plus wavelet sparsity plus a range constraint, which can be written as $g(\mathrm{L}x)$ with $\mathrm{L} = [\mathrm{L}_1; \ldots; \mathrm{L}_m]$ and $g(v_1, \ldots, v_m) = \sum_i g_i(v_i)$.

1.2 Why forward–backward is not enough

Applying forward–backward (Chapter 5) to (1) with the smooth part f and the non-smooth part $g \circ \mathrm{L}$ gives

$$x^{[k+1]} = \operatorname{prox}_{\gamma g \circ \mathrm{L}}\left(x^{[k]} - \gamma \nabla f(x^{[k]})\right),$$

which requires the proximity operator of $f + g \circ \mathrm{L}$. specific cases.

Proposition 1.1 – (Prox of a composition with a semi-orthogonal operator [8])

Let $g \in \Gamma_0(\mathcal{G})$ and $\mathrm{L} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ be such that $\mathrm{L}\mathrm{L}^\top = \nu \mathrm{Id}$ for some $\nu > 0$. Then $g \circ \mathrm{L} \in \Gamma_0(\mathcal{H})$ and

$$\operatorname{prox}_{g \circ \mathrm{L}} = \mathrm{Id} + \frac{1}{\nu}\mathrm{L}^\top \circ (\operatorname{prox}_{\nu g} - \mathrm{Id}) \circ \mathrm{L}.$$

Proof. Let $x \in \mathcal{H}$, $v = \operatorname{prox}_{\nu g}(\mathrm{L}x)$ and $p = x + \frac{1}{\nu}\mathrm{L}^\top(v - \mathrm{L}x)$. Since $\mathrm{L}\mathrm{L}^\top = \nu\mathrm{Id}$, $\mathrm{L}p = \mathrm{L}x + (v - \mathrm{L}x) = v$. By the characterization of $\operatorname{prox}_{\nu g}$, $\mathrm{L}x - v \in \nu\,\partial g(v)$, so that

$$x - p = \frac{1}{\nu}\mathrm{L}^\top(\mathrm{L}x - v) \in \mathrm{L}^\top \partial g(\mathrm{L}p) \subset \partial(g \circ \mathrm{L})(p),$$

where the last inclusion follows from the definition of the subdifferential (if $w \in \partial g(\mathrm{L}p)$, then $g(\mathrm{L}y) \geq g(\mathrm{L}p) + \langle w \,|\, \mathrm{L}(y - p)\rangle$ for every y). Hence $p = \operatorname{prox}_{g\circ\mathrm{L}}(x)$. ■

This covers orthonormal bases ($\nu = 1$). It does *not* cover:

- finite differences ($\mathrm{L}\mathrm{L}^\top$ is a discrete Laplacian, not a multiple of Id): there is no closed form for $\operatorname{prox}_{\mathrm{TV}}$;
- redundant tight frames, for which $\mathrm{L}^\top\mathrm{L} = \nu\mathrm{Id}$ but $\mathrm{L}\mathrm{L}^\top = \nu P_{\operatorname{ran}\mathrm{L}} \neq \nu\mathrm{Id}$ (see the chapter on wavelets).

Computing $\operatorname{prox}_{\gamma g\circ\mathrm{L}}(y)$ amounts to solving the *denoising* problem

$$\underset{x\in\mathcal{H}}{\text{minimize}} \;\; \frac{1}{2}\|x - y\|^2 + \gamma\, g(\mathrm{L}x), \tag{2}$$

which is itself of the form (1). An elegant option is to change the point of view: *dualize* the problem.

2 The convex conjugate

2.1 Definition and geometric interpretation

Definition 2.1 – (Conjugate function)

Let $f : \mathcal{H} \to]-\infty, +\infty]$. Its *conjugate* (or Legendre–Fenchel transform) is

$$f^* : \mathcal{H} \to [-\infty, +\infty] : \; u \mapsto \sup_{x\in\mathcal{H}} \; \langle x \,|\, u\rangle - f(x).$$

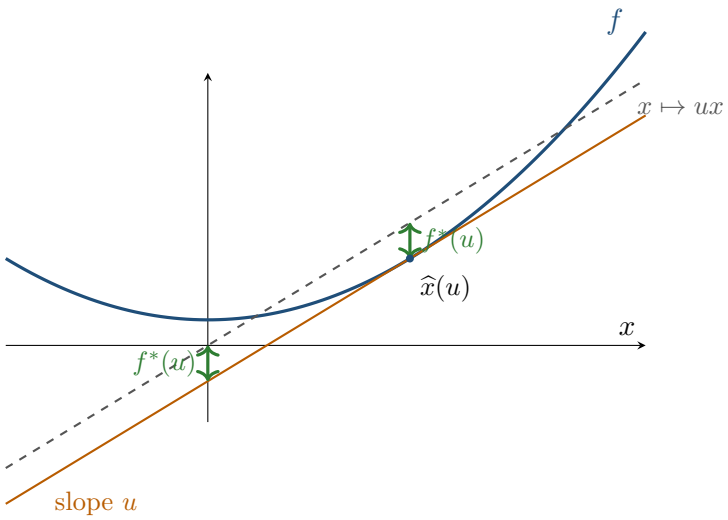


Figure 1: Geometric interpretation of $f^*(u)$ for $f(x) = \frac{1}{2}x^2 + 0.3$ and $u = 1.2$: $f^*(u)$ is the largest gap between the line $x \mapsto ux$ and the graph of f (attained at $\widehat{x}(u)$, where $u = \nabla f(\widehat{x}(u))$), and $-f^*(u)$ is the intercept of the supporting line of slope u. Here $f^*(u) = \frac{1}{2}u^2 - 0.3 = 0.42$.

2.2 Basic properties

Proposition 2.2 – (First properties)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be proper.

1. f^* is convex and lower semicontinuous (even when f is not).
2. **Fenchel–Young inequality**: for every $(x, u) \in \mathcal{H}^2$,
$$f(x) + f^*(u) \geq \langle x \,|\, u \rangle .$$
3. For every $(x, u) \in \mathcal{H}^2$,
$$f(x) + f^*(u) = \langle x \,|\, u \rangle \iff u \in \partial f(x).$$
4. If $f \leq h$, then $f^* \geq h^*$.

Proof. 1. See [1, Prop. 13.13]; see also [19, Thm. 12.2].

2. See [1, Prop. 13.15]; see also [19, §12].
3. See [1, Prop. 16.10]; see also [19, Thm. 23.5].
4. See [1, Prop. 13.16].

■

Theorem 2.3 – (Fenchel–Moreau)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be proper. Then f is convex and lower semicontinuous if and only if $f^{**} = f$. In particular, if $f \in \Gamma_0(\mathcal{H})$, then $f^* \in \Gamma_0(\mathcal{H})$.

See [1, Chap. 13].

Corollary 2.4 – (Inversion of the subdifferential)

Let $f \in \Gamma_0(\mathcal{H})$ and $(x, u) \in \mathcal{H}^2$. Then
$$u \in \partial f(x) \iff x \in \partial f^*(u).$$
In other words, $\partial f^* = (\partial f)^{-1}$.

Proof. Apply the equality case of Fenchel–Young to f and to f^*, using $f^{**} = f$. ■

2.3 Examples

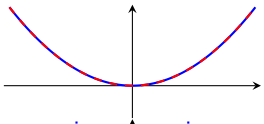
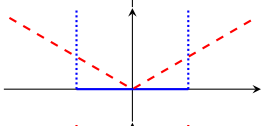
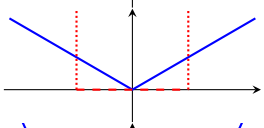
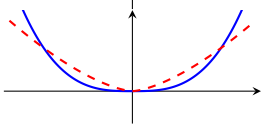
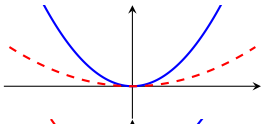
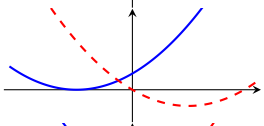
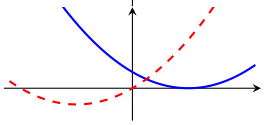
$f(x)$	$f^*(u)$	f and f^*
$\frac{1}{2}\|x\|^2$	$\frac{1}{2}\|u\|^2$	
$\iota_C(x)$, C closed convex	$\sigma_C(u) = \sup_{x\in C}\langle x \mid u\rangle$	
$\lambda\|x\|_1$	$\iota_{\mathcal{B}_{\|\cdot\|_\infty}(0,\lambda)}(u)$	
$\frac{1}{q}\|x\|^q$, $q>1$	$\frac{1}{q^*}\|u\|^{q^*}$, $\frac{1}{q}+\frac{1}{q^*}=1$	
$\alpha f(x)$, $\alpha>0$	$\alpha f^*(u/\alpha)$	
$f(x)+\langle x \mid a\rangle + b$	$f^*(u-a)-b$	
$f(x-z)$	$f^*(u)+\langle u \mid z\rangle$	
$\sum_i f_i(x_i)$ (separable)	$\sum_i f_i^*(u_i)$	(multivariate)

Table 1: Some conjugate pairs and calculus rules, with f (solid blue) and f^* (dashed red) in dimension one. Illustrations use $C=[-1,1]$, $\lambda=1$, $q=3$; the calculus rules are illustrated with $f=\frac{1}{2}|\cdot|^2$, $\alpha=2$, $a=1$, $b=\frac{1}{2}$, $z=1$. Dotted vertical lines indicate that the function is $+\infty$ outside the segment.

Remark See also [20, Table 1] for more examples

2.4 Moreau decomposition

The following identity is the workhorse of primal–dual algorithms: the proximity operator of f^* is as cheap as that of f.

Theorem 2.5 – (Moreau decomposition)

Let $f\in\Gamma_0(\mathcal{H})$ and $\gamma>0$. For every $x\in\mathcal{H}$,

$$\operatorname{prox}_{\gamma f^*}(x) = x - \gamma\,\operatorname{prox}_{\gamma^{-1}f}(\gamma^{-1}x).$$

In particular ($\gamma=1$), $x=\operatorname{prox}_f(x)+\operatorname{prox}_{f^*}(x)$.

Proof. Let $p=\operatorname{prox}_{\gamma^{-1}f}(\gamma^{-1}x)$, characterized by $\gamma^{-1}x-p\in\gamma^{-1}\partial f(p)$, i.e. $x-\gamma p\in\partial f(p)$. By Corollary 2.4, this is equivalent to $p\in\partial f^*(x-\gamma p)$, i.e. $x-(x-\gamma p)\in\gamma\,\partial f^*(x-\gamma p)$, which

characterizes $x - \gamma p = \operatorname{prox}_{\gamma f^*}(x)$. ■

Example 2.1 (ℓ_1-norm)**.** For $f = \lambda \|\cdot\|_1$, $f^* = \iota_{\mathcal{B}_{\|\cdot\|_\infty}(0,\lambda)}$ and $\operatorname{prox}_{\sigma f^*} = \operatorname{clip}_\lambda$ for every $\sigma > 0$ (a projection does not depend on the step size). Moreau's identity with $\sigma = 1$ reads $x = \operatorname{soft}_\lambda(x) + \operatorname{clip}_\lambda(x)$: soft thresholding is "what remains after clipping".

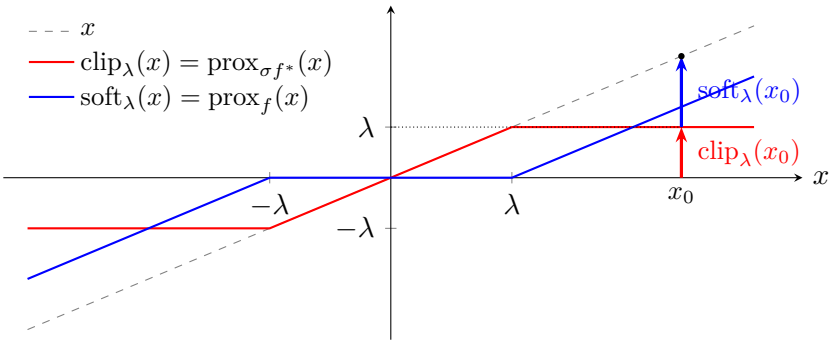


Figure 2: Moreau's identity for $f = \lambda \|\cdot\|_1$ (here $\lambda = 1$, one coordinate): $x = \operatorname{soft}_\lambda(x) + \operatorname{clip}_\lambda(x)$. The clipping $\operatorname{clip}_\lambda = \operatorname{prox}_{\sigma f^*}$ keeps the part of x inside $[-\lambda, \lambda]$, and soft thresholding keeps what exceeds it. At x_0, the two vertical arrows stack up to the identity line.

2.5 Strong convexity and smoothness are dual

Proposition 2.6 – (Strong convexity and gradient-Lipschitz)

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

1. **Duality between strong convexity and smoothness.** Let $\rho > 0$. Then f is ρ-strongly convex if and only if f^* is differentiable on $\mathcal{H}$ with a $\frac{1}{\rho}$-Lipschitz gradient.
2. **Gradient of the Moreau envelope is a prox of the conjugate.** Let $\gamma > 0$. The Moreau envelope ${}^\gamma f = \left(f^* + \frac{\gamma}{2}\|\cdot\|^2\right)^*$ is differentiable on $\mathcal{H}$ with a $\frac{1}{\gamma}$-Lipschitz gradient given by
$$\nabla {}^\gamma f(x) = \frac{1}{\gamma}\left(x - \operatorname{prox}_{\gamma f}(x)\right) = \operatorname{prox}_{\gamma^{-1} f^*}\left(\frac{x}{\gamma}\right).$$
In particular, for $\gamma = 1$, $\nabla {}^1 f = \operatorname{prox}_{f^*}$.

Proof. 1. See [1, Thm. 18.15] in Hilbert spaces; see also [21, Prop. 12.60] in finite dimension.

2. The differentiability of ${}^\gamma f$, the Lipschitz constant of its gradient and the expression $\nabla {}^\gamma f = \gamma^{-1}(\mathrm{Id} - \operatorname{prox}_{\gamma f})$ are given in [1, Prop. 12.30]; see also [22]. The second expression follows from Moreau's identity $x = \operatorname{prox}_{\gamma f}(x) + \gamma \operatorname{prox}_{\gamma^{-1} f^*}(x/\gamma)$ [1, Thm. 14.3].

■

3 Fenchel–Rockafellar duality

3.1 Primal and dual problems

Definition 3.1 – (Fenchel–Rockafellar dual problem)

Let $f : \mathcal{H} \to]-\infty, +\infty]$, $g : \mathcal{G} \to]-\infty, +\infty]$ and $\mathrm{L} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.
The *primal problem* is

$$\text{(P)} \qquad \mu = \inf_{x \in \mathcal{H}} \; f(x) + g(\mathrm{L}x), \tag{3}$$

and its *dual problem* is

$$\text{(D)} \qquad \mu^* = \inf_{u \in \mathcal{G}} \; f^*(-\mathrm{L}^\top u) + g^*(u). \tag{4}$$

Where does the dual come from? The Lagrangian. If $g \in \Gamma_0(\mathcal{G})$, Fenchel–Moreau gives $g(\mathrm{L}x) = g^{**}(\mathrm{L}x) = \sup_u \langle \mathrm{L}x \,|\, u \rangle - g^*(u)$, so that

$$\inf_x \; f(x) + g(\mathrm{L}x) = \inf_x \sup_u \; \mathcal{L}(x, u) := f(x) + \langle \mathrm{L}x \,|\, u \rangle - g^*(u). \tag{5}$$

$\mathcal{L}$ is convex in x and concave in u. Exchanging inf and sup,

$$\sup_u \inf_x \; \mathcal{L}(x, u) = \sup_u \Big(- \sup_x \big(\big\langle x \,\big|\, -\mathrm{L}^\top u \big\rangle - f(x) \big) - g^*(u) \Big) \tag{6}$$

$$= \sup_u \; -f^*(-\mathrm{L}^\top u) - g^*(u) = -\mu^*. \tag{7}$$

The primal problem is $\inf \sup \mathcal{L}$, the dual problem is $\sup \inf \mathcal{L}$ (up to a sign), and the variable $u \in \mathcal{G}$ is the *dual variable*, living in the dual space where g acts.

3.2 Weak and strong duality

Proposition 3.2 – (Weak duality)

Let f and g be proper. Then $\mu \geq -\mu^*$. If $\mu \in \mathbb{R}$, the nonnegative quantity $\mu + \mu^*$ is called the *duality gap*.

Proof. By the Fenchel–Young inequality applied to f at $(x, -\mathrm{L}^\top u)$ and to g at $(\mathrm{L}x, u)$,

$$f(x) + g(\mathrm{L}x) + f^*(-\mathrm{L}^\top u) + g^*(u) \geq \big\langle x \,\big|\, -\mathrm{L}^\top u \big\rangle + \langle \mathrm{L}x \,|\, u \rangle = 0.$$

Taking the infimum over x and over u yields $\mu + \mu^* \geq 0$. ∎

Theorem 3.3 – (Strong duality [1, Chap. 15])

Let $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathrm{L} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. If

$$\operatorname{int}(\operatorname{dom} g) \cap \mathrm{L}(\operatorname{dom} f) \neq \varnothing \quad \text{or} \quad \operatorname{dom} g \cap \operatorname{int}(\mathrm{L}(\operatorname{dom} f)) \neq \varnothing,$$

then

$$\mu = \inf_{x \in \mathcal{H}} \; f(x) + g(\mathrm{L}x) = -\min_{u \in \mathcal{G}} \; f^*(-\mathrm{L}^\top u) + g^*(u) = -\mu^*,$$

i.e. the duality gap is zero and the dual problem has a solution.

3.3 Optimality conditions

Theorem 3.4 – (Kuhn–Tucker conditions [1, Chap. 19])

Let $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathrm{L} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

1. If $(\widehat{x}, \widehat{u}) \in \mathcal{H} \times \mathcal{G}$ satisfies
$$-\mathrm{L}^\top \widehat{u} \in \partial f(\widehat{x}) \qquad \text{and} \qquad \mathrm{L}\widehat{x} \in \partial g^*(\widehat{u}), \tag{8}$$
then $\widehat{x}$ solves (P), $\widehat{u}$ solves (D), and $\mu = -\mu^*$. Such a pair is called a *Kuhn–Tucker point*.
2. Conversely, if $\widehat{x}$ satisfies $0 \in \partial f(\widehat{x}) + \mathrm{L}^\top \partial g(\mathrm{L}\widehat{x})$ (which holds for every solution of (P) under the qualification condition of Theorem 3.3), then there exists $\widehat{u}$ such that $(\widehat{x}, \widehat{u})$ is a Kuhn–Tucker point.

Proof. 1. By Corollary 2.4, the second condition reads $\widehat{u} \in \partial g(\mathrm{L}\widehat{x})$. Then the two Fenchel–Young inequalities used in the proof of Proposition 3.2 are equalities, so $f(\widehat{x}) + g(\mathrm{L}\widehat{x}) + f^*(-\mathrm{L}^\top \widehat{u}) + g^*(\widehat{u}) = 0$, i.e. the gap vanishes at $(\widehat{x}, \widehat{u})$. Since $\mu \geq -\mu^*$, both are optimal.

2. If $0 \in \partial f(\widehat{x}) + \mathrm{L}^\top \partial g(\mathrm{L}\widehat{x})$, there is $\widehat{u} \in \partial g(\mathrm{L}\widehat{x})$ with $-\mathrm{L}^\top \widehat{u} \in \partial f(\widehat{x})$, and Corollary 2.4 gives $\mathrm{L}\widehat{x} \in \partial g^*(\widehat{u})$. ■

Remark (Saddle points) Conditions (8) say exactly that $0 \in \partial_x \mathcal{L}(\widehat{x}, \widehat{u})$ and $0 \in \partial_u(-\mathcal{L})(\widehat{x}, \widehat{u})$, i.e. $(\widehat{x}, \widehat{u})$ is a *saddle point* of the Lagrangian:
$$(\forall (x, u) \in \mathcal{H} \times \mathcal{G}) \qquad \mathcal{L}(\widehat{x}, u) \leq \mathcal{L}(\widehat{x}, \widehat{u}) \leq \mathcal{L}(x, \widehat{u}).$$
Primal–dual algorithms look for such saddle points.

4 Dual forward–backward and denoising problem

4.1 Dual of TV denoising

For $g = \lambda \|\cdot\|_{2,1}$ and L the discrete gradient, the dual of $\min_x \frac{1}{2}\|x - z\|^2 + \lambda\,\mathrm{TV}(x)$ is the *constrained least-squares problem*
$$\underset{u \in \mathbb{R}^{N \times 2}}{\text{minimize}} \ \frac{1}{2}\left\|z - \mathrm{L}^\top u\right\|^2 \quad \text{s.t.} \quad \|u_n\|_2 \leq \lambda \ \ (\forall n),$$
and $\widehat{x} = z - \mathrm{L}^\top \widehat{u}$. The non-smooth, non-separable primal problem has become a smooth problem with a simple separable constraint [4].

4.2 Algorithm

The dual problem is the minimization of $h(u) + g^*(u)$ with
$$h(u) = \frac{1}{2}\left\|z - \mathrm{L}^\top u\right\|^2, \qquad \nabla h(u) = -\mathrm{L}(z - \mathrm{L}^\top u),$$
and ∇h is $\|\mathrm{L}\|^2$-Lipschitz. Forward–backward can thus be applied to the dual problem, with the proximity operator of g^* (computed from that of g by Theorem 2.5).

Algorithm 4.1 – (Dual forward–backward)

Problem: $\widehat{x} = \arg\min_x \frac{1}{2}\|x - z\|^2 + g(\mathrm{L}x)$.
Input: $u^{[0]} \in \mathcal{G}$, a step size $\tau \in \left]0, \frac{2}{\|\mathrm{L}\|^2}\right[$.
For $k = 0, 1, \ldots$:
$x^{[k]} = z - \mathrm{L}^\top u^{[k]}$
$u^{[k+1]} = \mathrm{prox}_{\tau g^*}(u^{[k]} + \tau \mathrm{L} x^{[k]})$

Theorem 4.2 – (Convergence of dual FB [7])

Assume that the qualification condition of Theorem 3.3 holds. Then $(u^{[k]})_{k\in\mathbb{N}}$ converges (weakly) to a solution $\widehat{u}$ of the dual problem and $(x^{[k]})_{k\in\mathbb{N}}$ converges strongly to the primal solution $\widehat{x} = z - \mathrm{L}^\top \widehat{u}$.

Sketch of proof. The convergence of $(u^{[k]})$ is the convergence of forward–backward (Chapter 5) applied to $h + g^*$, since ∇h is $\|\mathrm{L}\|^2$-Lipschitz and $\tau < 2/\|\mathrm{L}\|^2$. In finite dimension, the weak convergence of $(u^{[k]})$ suffices: $x^{[k]} = z - \mathrm{L}^\top u^{[k]} \to z - \mathrm{L}^\top \widehat{u}$ by continuity. ■

5 Primal–dual algorithms

Primal–dual algorithms update x and u *jointly*, without any specific function f, by searching for a Kuhn–Tucker point. We consider

$$\underset{x\in\mathcal{H}}{\text{minimize}} \;\; f_1(x) + f_2(x) + g(\mathrm{L}x),$$

with $f_1 \in \Gamma_0(\mathcal{H})$, $f_2 \in \Gamma_0(\mathcal{H})$ convex and with β-Lipschitz gradient, $g \in \Gamma_0(\mathcal{G})$ and $\mathrm{L} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

5.1 Saddle-point formulation and Kuhn–Tucker conditions

Since $g \in \Gamma_0(\mathcal{G})$, the Moreau–Fenchel theorem gives $g = g^{**}$, i.e. $g(\mathrm{L}x) = \sup_{u\in\mathcal{G}} \langle \mathrm{L}x \,|\, u\rangle - g^*(u)$. The considered minimization problem can therefore be rewritten as the search for a saddle point of the *Lagrangian*

$$(\widehat{x}, \widehat{u}) \in \underset{x\in\mathcal{H}}{\mathrm{Argmin}} \; \max_{u\in\mathcal{G}} \; \mathcal{L}(x, u) := f_1(x) + f_2(x) + \langle \mathrm{L}x \,|\, u\rangle - g^*(u). \tag{9}$$

Minimizing $\mathcal{L}(\cdot, u)$ in x yields (minus) the dual objective $(f_1 + f_2)^*(-\mathrm{L}^\top u) + g^*(u)$, and maximizing $\mathcal{L}(x, \cdot)$ in u yields the primal objective. Writing Fermat's rule on $x \mapsto \mathcal{L}(x, \widehat{u})$ and on $u \mapsto -\mathcal{L}(\widehat{x}, u)$ gives the optimality conditions.

Proposition 5.1 – (Kuhn–Tucker conditions)

Assume that $\mathrm{dom}\, g \cap \mathrm{L}(\mathrm{dom}\, f_1) \neq \varnothing$. Then $(\widehat{x}, \widehat{u}) \in \mathcal{H} \times \mathcal{G}$ is a Kuhn–Tucker point of (9), i.e. $\widehat{x}$ solves the primal problem and $\widehat{u}$ the dual problem, if and only if

$$\begin{cases} 0 \in \partial f_1(\widehat{x}) + \nabla f_2(\widehat{x}) + \mathrm{L}^\top \widehat{u}, \\ 0 \in -\mathrm{L}\widehat{x} + \partial g^*(\widehat{u}). \end{cases} \tag{10}$$

This is Theorem 3.4 applied to $f = f_1 + f_2$, using $\partial(f_1 + f_2) = \partial f_1 + \nabla f_2$.

5.2 From the Kuhn–Tucker conditions to fixed-point equations

Let $\tau > 0$ and $\sigma > 0$. Multiplying the first line of (10) by τ and the second one by σ,

$$\begin{cases} -\tau(\mathrm{L}^\top \widehat{u} + \nabla f_2(\widehat{x})) \in \tau\, \partial f_1(\widehat{x}), \\ \sigma \mathrm{L}\widehat{x} \in \sigma\, \partial g^*(\widehat{u}). \end{cases}$$

We now add $\widehat{x} - \widehat{x} = 0$ to the left-hand side of the first inclusion and $\widehat{u} - \widehat{u} = 0$ to the second one; in the second inclusion we moreover write $\widehat{x} = 2\widehat{x} - \widehat{x}$ (this seemingly innocuous choice is what makes the resulting algorithm convergent, see Proposition **??**):

$$\begin{cases} \underbrace{\widehat{x} - \tau(\mathrm{L}^\top \widehat{u} + \nabla f_2(\widehat{x}))}_{=:\, v} - \underbrace{\widehat{x}}_{=:\, p} \in \tau\, \partial f_1(\underbrace{\widehat{x}}_{p}), \\ \underbrace{\widehat{u} + \sigma \mathrm{L}(2\widehat{x} - \widehat{x})}_{=:\, v'} - \underbrace{\widehat{u}}_{=:\, p'} \in \sigma\, \partial g^*(\underbrace{\widehat{u}}_{p'}). \end{cases}$$

Recalling the characterization of the proximity operator, $v - p \in \gamma\, \partial\psi(p) \Leftrightarrow p = \mathrm{prox}_{\gamma\psi}(v)$, we obtain the *fixed-point equations*

$$\begin{cases} \widehat{x} = \mathrm{prox}_{\tau f_1}\Big(\widehat{x} - \tau(\nabla f_2(\widehat{x}) + \mathrm{L}^\top \widehat{u})\Big), \\ \widehat{u} = \mathrm{prox}_{\sigma g^*}(\widehat{u} + \sigma \mathrm{L}(2\widehat{x} - \widehat{x})). \end{cases} \tag{11}$$

The associated fixed-point iteration, where the two occurrences of $\widehat{x}$ in $2\widehat{x} - \widehat{x}$ are replaced by the new and the old primal iterates, is the Condat–Vũ algorithm (Section 7.3); with $f_2 = 0$ it is the Chambolle–Pock algorithm.

5.3 The Chambolle–Pock algorithm

Algorithm 5.2 – (Chambolle–Pock [5])

Problem: $\min_x f(x) + g(\mathrm{L}x)$, $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$.
Input: $x^{[0]} \in \mathcal{H}$, $u^{[0]} \in \mathcal{G}$, $\tau > 0$, $\sigma > 0$ with $\sigma\tau \|\mathrm{L}\|^2 < 1$.
For $k = 0, 1, \ldots$:

$$\begin{aligned} x^{[k+1]} &= \mathrm{prox}_{\tau f}(x^{[k]} - \tau \mathrm{L}^\top u^{[k]}) \\ \bar{x}^{[k+1]} &= 2x^{[k+1]} - x^{[k]} \\ u^{[k+1]} &= \mathrm{prox}_{\sigma g^*}(u^{[k]} + \sigma \mathrm{L}\bar{x}^{[k+1]}) \end{aligned}$$

Theorem 5.3 – (Convergence of Chambolle–Pock [5])

Assume that $\mathcal{H}$ and $\mathcal{G}$ are finite dimensional, that the set of Kuhn–Tucker points is nonempty and that $\sigma\tau \|\mathrm{L}\|^2 < 1$. Then $(x^{[k]}, u^{[k]})$ converges to a Kuhn–Tucker point $(\widehat{x}, \widehat{u})$; in particular $x^{[k]}$ converges to $\widehat{x}$, a solution of (P), and $u^{[k]}$ converges to uh, a solution of (D).

Remark

- The step sizes only depend on $\|\mathrm{L}\|$, not on f or g; the ratio σ/τ is a free parameter that balances the primal and dual progress, and has a strong practical impact.

Acceleration under strong convexity. If f is ρ-strongly convex, the step sizes can be adapted along the iterations to obtain an $O(1/k^2)$ rate on $\left\|x^{[k]} - \widehat{x}\right\|^2$.

Algorithm 5.4 – (Accelerated Chambolle–Pock [5])

Input: $x^{[0]} = \bar{x}^{[0]} \in \mathcal{H}$, $u^{[0]} \in \mathcal{G}$, $\tau_0, \sigma_0 > 0$ with $\sigma_0\tau_0 \|\mathrm{L}\|^2 \leq 1$.
For $k = 0, 1, \ldots$:

$$\begin{aligned}
&u^{[k+1]} = \mathrm{prox}_{\sigma_k g^*}\big(u^{[k]} + \sigma_k \mathrm{L}\bar{x}^{[k]}\big)\\
&x^{[k+1]} = \mathrm{prox}_{\tau_k f}\big(x^{[k]} - \tau_k \mathrm{L}^\top u^{[k+1]}\big)\\
&\theta_k = 1/\sqrt{1+2\rho\tau_k}, \quad \tau_{k+1} = \theta_k\tau_k, \quad \sigma_{k+1} = \sigma_k/\theta_k\\
&\bar{x}^{[k+1]} = x^{[k+1]} + \theta_k\big(x^{[k+1]} - x^{[k]}\big)
\end{aligned}$$

The product $\sigma_k\tau_k$ stays constant while the primal step decreases like $1/k$: the algorithm becomes "more and more primal–implicit", as for FISTA in the strongly convex case.

5.4 Adding a smooth term: the Condat–Vũ algorithm

For the three-term problem (**??**), one would like to handle f_2 by a gradient step, as in forward–backward, instead of computing its proximity operator (which, for $f_2 = \frac{1}{2}\|\mathrm{A}\cdot - z\|^2$, requires inverting $\mathrm{Id} + \tau\mathrm{A}^\top\mathrm{A}$). This is exactly what the fixed-point equations (11) suggest.

Algorithm 5.5 – (Condat–Vũ [10, 18])

Problem: $\min_x f_1(x) + f_2(x) + g(\mathrm{L}x)$, $f_1 \in \Gamma_0(\mathcal{H})$, f_2 convex and β-smooth, $g \in \Gamma_0(\mathcal{G})$.
Input: $x^{[0]} \in \mathcal{H}$, $u^{[0]} \in \mathcal{G}$, $\tau > 0$, $\sigma > 0$ with

$$\frac{1}{\tau} - \sigma\|\mathrm{L}\|^2 > \frac{\beta}{2},$$

and relaxation parameters $\lambda_k \in]0, \delta[$, where $\delta = 2 - \frac{\beta}{2}\big(\frac{1}{\tau} - \sigma\|\mathrm{L}\|^2\big)^{-1} \in]1, 2]$.
For $k = 0, 1, \ldots$:

$$\begin{aligned}
&p^{[k]} = \mathrm{prox}_{\tau f_1}\Big(x^{[k]} - \tau\big(\nabla f_2(x^{[k]}) + \mathrm{L}^\top u^{[k]}\big)\Big)\\
&q^{[k]} = \mathrm{prox}_{\sigma g^*}\big(u^{[k]} + \sigma\mathrm{L}(2p^{[k]} - x^{[k]})\big)\\
&(x^{[k+1]}, u^{[k+1]}) = (x^{[k]}, u^{[k]}) + \lambda_k\big((p^{[k]}, q^{[k]}) - (x^{[k]}, u^{[k]})\big)
\end{aligned}$$

With $\lambda_k \equiv 1$, the iteration simply reads

$$\begin{aligned}
&x^{[k+1]} = \mathrm{prox}_{\tau f_1}\Big(x^{[k]} - \tau\big(\nabla f_2(x^{[k]}) + \mathrm{L}^\top u^{[k]}\big)\Big)\\
&u^{[k+1]} = \mathrm{prox}_{\sigma g^*}\big(u^{[k]} + \sigma\mathrm{L}(2x^{[k+1]} - x^{[k]})\big).
\end{aligned}$$

Special cases.

- $f_2 = 0$ ($\beta = 0$), $\lambda_k \equiv 1$: the condition becomes $\sigma\tau\|\mathrm{L}\|^2 < 1$ and we recover Chambolle–Pock.
- $g = 0$ (take $u^{[k]} \equiv 0$): the condition becomes $\tau < 2/\beta$ and we recover the (relaxed) forward–backward algorithm on $f_1 + f_2$, $x^{[k+1]} = x^{[k]} + \lambda_k\big(\mathrm{prox}_{\tau f_1}(x^{[k]} - \tau\nabla f_2(x^{[k]})) - x^{[k]}\big)$.
- $f_2 = 0$, $\lambda_k \equiv 1$, $\mathrm{L} = \mathrm{Id}$ and $\sigma = 1/\tau$ (limit case): Douglas–Rachford.

6 Augmented Lagrangian and ADMM

6.1 From the Lagrangian to the augmented Lagrangian

Another classical route introduces the splitting variable $y = \mathrm{L}x$:

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ f(x)+g(\mathrm{L}x) \quad\Longleftrightarrow\quad \underset{(x,y)\in\mathcal{H}\times\mathcal{G},\ \mathrm{L}x=y}{\text{minimize}}\ f(x)+g(y).$$

The equality constraint is handled with the Lagrange function

$$\mathcal{L}(x,y,v) = f(x)+g(y)+\langle v \,|\, \mathrm{L}x-y\rangle\,,$$

where $v\in\mathcal{G}$ is the Lagrange multiplier, and one looks for a saddle point $(\widehat{x},,)$ (minimum in (x,y), maximum in v). A natural idea is to minimize alternately in x and y and then increase $\mathcal{L}$ in v:

$$x^{[k+1]} \in \operatorname*{Argmin}_x \mathcal{L}(x, y^{[k]}, v^{[k]}), \tag{12}$$

$$y^{[k+1]} \in \operatorname*{Argmin}_y \mathcal{L}(x^{[k+1]}, y, v^{[k]}), \tag{13}$$

$$v^{[k+1]} \text{ s.t. } \mathcal{L}(x^{[k+1]}, y^{[k+1]}, v^{[k+1]}) \geq \mathcal{L}(x^{[k+1]}, y^{[k+1]}, v^{[k]}). \tag{14}$$

However, *convergence is not guaranteed in general*: since $\mathcal{L}$ is only affine in x and y through the coupling term, the subproblems may have no solution or many. The remedy is to add a quadratic penalty on the constraint, which is zero at every feasible point. For $\gamma>0$, the *augmented Lagrangian* is

$$\mathcal{L}_\gamma(x,y,w) = f(x)+g(y)+\gamma\,\langle w \,|\, \mathrm{L}x-y\rangle + \frac{\gamma}{2}\,\|\mathrm{L}x-y\|^2\,,$$

where the (scaled) multiplier w is related to v by $v=\gamma w$.

6.2 The ADMM algorithm

Minimizing $\mathcal{L}_\gamma$ alternately in x and y, then performing a gradient ascent step on the multiplier with step size $1/\gamma$, gives

$$\begin{cases} x^{[k+1]} \in \operatorname{Argmin}_x\ f(x)+\gamma\left\langle w^{[k]} \,\middle|\, \mathrm{L}x-y^{[k]}\right\rangle + \frac{\gamma}{2}\left\|\mathrm{L}x-y^{[k]}\right\|^2, \\ y^{[k+1]} \in \operatorname{Argmin}_y\ g(y)+\gamma\left\langle w^{[k]} \,\middle|\, \mathrm{L}x^{[k+1]}-y\right\rangle + \frac{\gamma}{2}\left\|\mathrm{L}x^{[k+1]}-y\right\|^2, \\ w^{[k+1]} = w^{[k]} + \frac{1}{\gamma}\nabla_w\mathcal{L}_\gamma(x^{[k+1]}, y^{[k+1]}, w^{[k]}). \end{cases}$$

Completing the squares, $\gamma\,\langle w\,|\,a\rangle + \frac{\gamma}{2}\,\|a\|^2 = \frac{\gamma}{2}\,\|a+w\|^2 - \frac{\gamma}{2}\,\|w\|^2$, and using $\nabla_w\mathcal{L}_\gamma = \gamma(\mathrm{L}x-y)$, this is the *alternating direction method of multipliers*.

Algorithm 6.1 – (ADMM [14, 12])

Problem: $\min_x f(x)+g(\mathrm{L}x)$, $f\in\Gamma_0(\mathcal{H})$, $g\in\Gamma_0(\mathcal{G})$, $\mathrm{L}^\top\mathrm{L}$ invertible.
Input: $y^{[0]}, w^{[0]}\in\mathcal{G}$, $\gamma>0$.
For $k=0,1,\ldots$:

$$\begin{aligned} x^{[k+1]} &= \arg\min_x\ \tfrac{1}{\gamma}f(x)+\tfrac{1}{2}\left\|\mathrm{L}x-y^{[k]}+w^{[k]}\right\|^2 \\ s^{[k+1]} &= \mathrm{L}x^{[k+1]} \\ y^{[k+1]} &= \operatorname{prox}_{\gamma^{-1}g}(s^{[k+1]}+w^{[k]}) \\ w^{[k+1]} &= w^{[k]}+s^{[k+1]}-y^{[k+1]} \end{aligned}$$

The x-subproblem is strongly convex (hence has a unique solution) as soon as $\mathrm{L}^\top\mathrm{L}$ is invertible, i.e. L is injective. The multiplier $w^{[k]}$ accumulates the constraint residuals $\mathrm{L}x^{[k]}-y^{[k]}$.

Theorem 6.2 – (Convergence of ADMM [14, 12])

Assume that $\mathcal{H}$ and $\mathcal{G}$ are finite dimensional, that $\mathrm{L}^\top\mathrm{L}$ is invertible, that $\operatorname{dom} g \cap \mathrm{L}(\operatorname{dom} f) \neq \varnothing$ (or a similar qualification condition holds) and that $\operatorname{Argmin}(f + g \circ \mathrm{L}) \neq \varnothing$. Then, for every $\gamma > 0$,

- $x^{[k]} \to \widehat{x}$, where $\widehat{x} \in \operatorname{Argmin}(f + g \circ \mathrm{L})$;
- $\gamma w^{[k]} \to \widehat{u}$, where $\widehat{u} \in \operatorname{Argmin}\big(f^* \circ (-\mathrm{L}^\top) + g^*\big)$ is a solution of the dual problem.

6.3 Summary

Algorithm	Problem	Required operations	Reference
Forward–backward	$f_1 + f_2$	∇f_2 (β-Lipschitz), $\operatorname{prox}_{f_1}$	[9]
ISTA	$f_2 + \lambda \lVert\cdot\rVert_1$	∇f_2, soft thresholding	[23]
Douglas–Rachford	$f + g$	prox_f, prox_g	[8]
Dual FB	$f + \frac{1}{2}\lVert\cdot - z\rVert^2 + g \circ \mathrm{L}$	prox_f, prox_g, L, $\mathrm{L}^\top$	[24]
ADMM	$f + g \circ \mathrm{L}$	$(\mathrm{L}^\top\mathrm{L} + \gamma^{-1}\partial f)^{-1}$-type step, prox_g	[14, 12]
Chambolle–Pock	$f + g \circ \mathrm{L}$	prox_f, prox_g, L, $\mathrm{L}^\top$	[5]
Condat–Vũ	$f_1 + f_2 + g \circ \mathrm{L}$	$\operatorname{prox}_{f_1}$, ∇f_2, prox_g, L, $\mathrm{L}^\top$	[10, 18]

Table 2: Proximal algorithms seen in this chapter.

References

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

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

[3] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems", *IEEE Trans. Image Process.*, 18(11):2419–2434, 2009.

[4] A. Chambolle, "An algorithm for total variation minimization and applications", *J. Math. Imaging Vis.*, 20(1–2):89–97, 2004.

[5] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging", *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

[6] P. Chen, J. Huang and X. Zhang, "A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration", *Inverse Problems*, 29(2):025011, 2013.

[7] P. L. Combettes, Đ. Dũng and B. C. Vũ, "Dualization of signal recovery problems", *Set-Valued Var. Anal.*, 18(3–4):373–404, 2010.

[8] P. L. Combettes and J.-C. Pesquet, "A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery", *IEEE J. Sel. Topics Signal Process.*, 1(4):564–574, 2007.

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

[10] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms", *J. Optim. Theory Appl.*, 158(2):460–479, 2013.

[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] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators", *Math. Program.*, 55:293–318, 1992.

[13] D. Gabay, "Applications of the method of multipliers to variational inequalities", in *Augmented Lagrangian Methods*, M. Fortin and R. Glowinski (eds.), North-Holland, 1983, pp. 299–331.

[14] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation", *Comput. Math. Appl.*, 2(1):17–40, 1976.

[15] B. He and X. Yuan, "Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective", *SIAM J. Imaging Sciences*, 5(1):119–149, 2012.

[16] I. Loris and C. Verhoeven, "On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty", *Inverse Problems*, 27(12):125007, 2011.

[17] J.-C. Pesquet and N. Pustelnik, "A parallel inertial proximal optimization method", *Pacific J. Optim.*, 8(2):273–305, 2012.

[18] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators", *Adv. Comput. Math.*, 38(3):667–681, 2013.

[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[20] N. Komodakis and J.-C. Pesquet, "Playing with duality: An overview of recent primal–dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, 2015.

[21] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[22] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299, 1965.

[23] I. Daubechies, M. Defrise and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[24] P. L. Combettes, Đ. Dũng and B. C. Vũ, "Dualization of signal recovery problems," *Set-Valued and Variational Analysis*, vol. 18, no. 3–4, pp. 373–404, 2010.