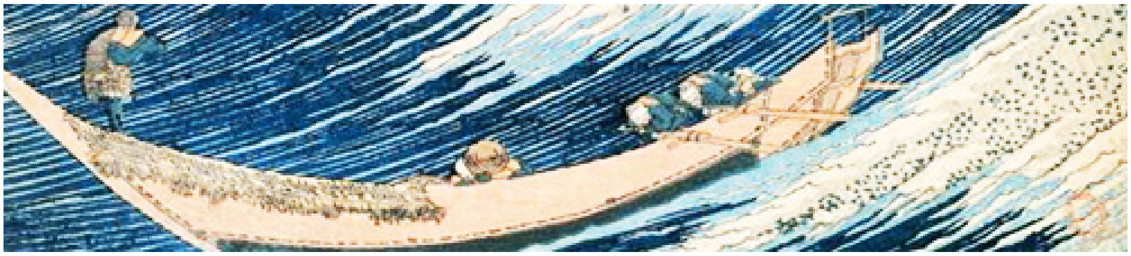
Mathematical Foundations of Data Sciences



Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
www.gpeyre.com
www.numerical-tours.com

October 19, 2017

Chapter 14

Convex Optimization

The main references for this chapter are [10, 11, 5], see also [31, 4, 3].

We consider a general convex optimization problem

$$\min_{x\in\mathcal{H}} f(x) \tag{14.1}$$

where $\mathcal{H} = \mathbb{R}^N$ is a finite dimensional Hilbertian (i.e. Euclidean) space, and try to devise "cheap" algorithms with a low computational cost per iterations. The class of algorithms considered are first order, i.e. they make use of gradient information.

14.1 Gradient Descent Methods

We have already encountered the gradient descent method informally in Section **??** for the regularization of inverse problem. We now give a detailed analysis of the method.

14.1.1 Gradient Descent

The optimization program (10.26) is a example of unconstrained convex optimization of the form (14.1) where $f : \mathcal{H} \to \mathbb{R}$ is a $\mathcal{C}^1$ function with Lipschitz gradient (so-called "smooth" function). Recall that the gradient $\nabla f : \mathcal{H} \mapsto \mathcal{H}$ of this functional (not to be confound with the discretized gradient $\nabla x \in \mathcal{H}$ of f) is defined by the following first order relation

$$f(x+r) = f(x) + \langle f,\, r\rangle_{\mathcal{H}} + O(\|r\|_{\mathcal{H}}^2)$$

where we used $O(\|r\|_{\mathcal{H}}^2)$ in place of $o(\|r\|_{\mathcal{H}})$ (for differentiable function) because we assume here f is of class $\mathcal{C}^1$ (i.e. the gradient is continuous). Section 10.4.3 shows typical examples of gradient computation.

For such a function, the gradient descent algorithm is defined as

$$x^{(\ell+1)} \overset{\text{def.}}{=} x^{(\ell)} - \tau_\ell \nabla f(x^{(\ell)}), \tag{14.2}$$

where the step size $\tau_\ell > 0$ should be small enough to guarantee convergence, but large enough for this algorithm to be fast.

One also needs to quantify the smoothness of f. This is enforced by requiring that the gradient is L-Lipschitz, i.e.

$$\forall\, (x,x') \in (\mathcal{H})^2, \quad \|\nabla f(x) - \nabla f(x')\| \leqslant L\|x - x'\|. \tag{$\mathcal{R}_L$}$$

In order to obtain fast convergence of the iterates themselve, it is needed that the function has enough "curvature" (i.e. is not too flat), which corresponds to imposing that f is μ-strongly convex

$$\forall\, (x,x'), \in (\mathcal{H})^2, \quad \langle \nabla f(x) - \nabla f(x'),\, x - x'\rangle \geqslant \mu\|x - x'\|^2. \tag{$\mathcal{S}_\mu$}$$

The following proposition express these conditions as constraints on the hessian for $\mathcal{C}^2$ functions.

Proposition 32. *Conditions* ($\mathcal{R}_L$) *and* ($\mathcal{S}_\mu$) *imply*

$$\forall\,(x,x'),\quad f(x') + \langle \nabla f(x),\, x'-x\rangle + \frac{\mu}{2}\|x-x'\|^2 \leqslant f(x) \leqslant f(x') + \langle \nabla f(x'),\, x'-x\rangle + \frac{L}{2}\|x-x'\|^2. \quad (14.3)$$

If f is of class $\mathcal{C}^2$, conditions ($\mathcal{R}_L$) *and* ($\mathcal{S}_\mu$) *are equivalent to*

$$\forall\,x,\quad \mu\mathrm{Id}_{N\times N} \preceq \partial^2 f(x) \preceq L\mathrm{Id}_{N\times N} \quad (14.4)$$

where $\partial^2 f(x) \in \mathbb{R}^{N\times N}$ is the Hessian of f, and where $\preceq$ is the natural order on symmetric matrices, i.e.

$$A \preceq B \quad\Longleftrightarrow\quad \forall\,x\in\mathcal{H},\quad \langle Au,\,u\rangle \leqslant \langle Bu,\,u\rangle.$$

Proof. We prove (14.3), using Taylor expansion with integral remain

$$f(x') - f(x) = \int_0^1 \langle \nabla f(x_t),\, x'-x\rangle \mathrm{d}t = \langle \nabla f(x),\, x'-x\rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x),\, x'-x\rangle \mathrm{d}t$$

where $x_t \overset{\text{def.}}{=} f + t(x'-x)$. Using Cauchy-Schwartz, and then the smoothness hypothesis ($\mathcal{R}_L$)

$$f(x') - f(x) \leqslant \langle \nabla f(x),\, x'-x\rangle + \int_0^1 L\|x_t - f\|\|x'-x\|\mathrm{d}t \leqslant \langle \nabla f(x),\, x'-x\rangle + L\|x'-x\|^2 \int_0^1 t\mathrm{d}t$$

which is the desired upper-bound. Using directly ($\mathcal{S}_\mu$) gives

$$f(x') - f(x) = \langle \nabla f(x),\, x'-x\rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x),\, \frac{x_t - x}{t}\rangle \mathrm{d}t \geqslant \langle \nabla f(x),\, x'-x\rangle + \mu \int_0^1 \frac{1}{t}\|x_t - x\|^2 \mathrm{d}t$$

which gives the desired result since $\|x_t - x\|^2/t = t\|x'-x\|^2$. □

The relation (14.3) shows that a smooth (resp. strongly convex) functional is bellow a quadratic tangential majorant (resp. minorant).

Condition (14.4) thus reads that the singular values of $\partial^2 f(x)$ should be contained in the interval $[\mu, L]$. The upper bound is also equivalent to $\|\partial^2 f(x)\|_{\text{op}} \leqslant L$ where $\|\cdot\|_{\text{op}}$ is the operator norm, i.e. the largest singular value. In the special case of a quadratic function $\mathcal{Q}$ of the form (10.24), $\partial^2 f(x) = A$ is constant, so that $[\mu, L]$ can be chosen to be the range of the singular values of A.

The following theorem ensure the convergence of the gradient descent with a linear speed.

Theorem 34. *If f satisfy conditions* ($\mathcal{R}_L$) *and* ($\mathcal{S}_\mu$), *assuming there exists* $(\tau_{\min}, \tau_{\max})$ *such that*

$$0 < \tau_{\min} \leqslant \tau_\ell \leqslant \tau_{\max} < \frac{2\mu}{L} \quad (14.5)$$

then there exists $0 \leqslant \rho < 1$ *such that*

$$\|x^{(\ell)} - x^\star\| \leqslant \rho^\ell \|x^{(0)} - x^\star\| \quad (14.6)$$

where $x^\star$ is the unique solution to (**??**).

Proof. Since $\nabla f(x^\star) = 0$, one has

$$x^{(\ell+1)} - x^\star = (x^{(\ell)} - x^\star) - \tau_\ell(\nabla f(x^{(\ell)}) - \nabla f(x^\star)).$$

Hence, using strong convexity and Lipschitz gradient

$$\begin{aligned}\|x^{(\ell+1)} - x^\star\|^2 &= \|x^{(\ell)} - x^\star\|^2 - 2\tau_\ell\langle x^{(\ell)} - x^\star,\, \nabla f(x^{(\ell)}) - \nabla f(x^\star)\rangle + \tau_\ell^2\|\nabla f(x^{(\ell)}) - \nabla f(x^\star)\|^2 \\ &\leqslant P(\tau_\ell)\|x^{(\ell)} - x^\star\|^2 \quad \text{where} \quad P(\tau) = 1 - 2\mu\tau + L^2\tau^2.\end{aligned}$$

Figure 14.1, left, shows visually the shape of the second order polynomial P, which shows that condition (14.11) on τ_ℓ implies

$$P(\tau_\ell)^{\frac{1}{2}} \leqslant \rho \overset{\text{def.}}{=} \max(P(\tau_{\min}), P(\tau_{\max}))^{\frac{1}{2}} < 1,$$

which shows the desired result. □

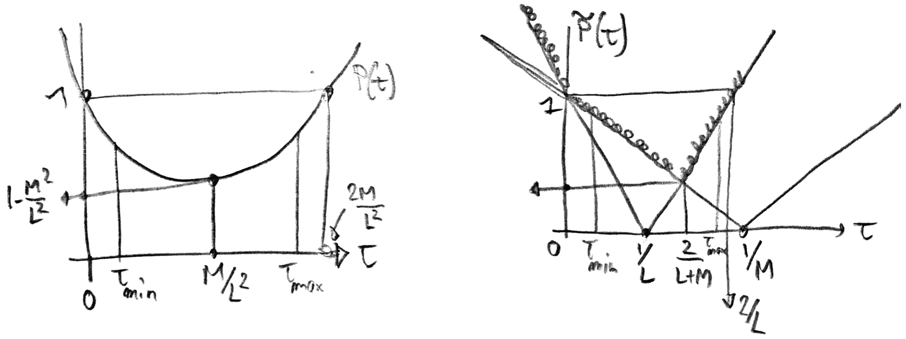


Figure 14.1: Contraction constant $P(\tau)$ and $\tilde{P}(\tau)$ for a gradient descent step in the generic case (left) and for a quadratic function (right).

The error decay rate (14.9), although it is geometrical $O(\rho^\ell)$ is called a "linear rate" in the optimization literature. It is a "global" rate because it hold for all ℓ (and not only for large enough ℓ). The best (smallest) rate ρ is obtained when choosing

$$\tau_\ell = \frac{\mu}{L^2} \quad \Longrightarrow \quad \rho = 1 - \frac{\mu^2}{L^2}. \tag{14.7}$$

In the case of a quadratic functional of the form (10.24), one can sharpen the convergence proof because the iterates are computed in closed form using matrix multiplication

$$x^{(\ell)} - x^\star = (\mathrm{Id}_N - \tau_\ell A)(x^{(0)} - x^\star)$$

which leads to the following proposition (see also Figure 14.1, right, for the corresponding contraction constant involved as a function of τ).

Proposition 33. *For $f(x) = \langle A,\, f\rangle - \langle b,\, x\rangle$ with the singular values of A upper-bounded by L, assuming there exists $(\tau_{\min}, \tau_{\max})$ such that*

$$0 < \tau_{\min} \leqslant \tau_\ell \leqslant \tilde{\tau}_{\max} < \frac{2}{L} \tag{14.8}$$

then there exists $0 \leqslant \tilde{\rho} < 1$ such that

$$\|x^{(\ell)} - x^\star\| \leqslant \tilde{\rho}^\ell \|x^{(0)} - x^\star\|. \tag{14.9}$$

If the singular values are lower bounded by μ, then the best rate $\tilde{\rho}$ is obtained for

$$\tau_\ell = \frac{2}{L+\mu} \quad \Longrightarrow \quad \tilde{\rho} \overset{\text{def.}}{=} \frac{L-\mu}{L+\mu}. \tag{14.10}$$

The maximum allowable step size $\tilde{\tau}_{\max}$ in (14.11) is much larger than $\tau_{\max}$ given in (14.11), and the optimal rate (14.10) is also much better (smaller) than the one in (14.7). In particular, if

$$\varepsilon \overset{\text{def.}}{=} M/L \ll 1$$

(which is the typical setup for ill-posed problems), then

$$\rho \sim 1 - \varepsilon^2 \quad \text{and} \quad \tilde{\rho} \sim 1 - 2\varepsilon.$$

The quantity ε in some sense reflects the inverse-conditioning of the problem. For quadratic function, it indeed corresponds exactly to the inverse of the condition number (which is the ratio of the largest to smallest singular value). The condition number is minimum and equal to 1 for orthogonal matrices.

These two results are however complementary. Indeed, if the gradient descent converges, then ultimately $x^{(\ell)}$ is close to $x^\star$, so that one can approximate up to second order $f(x) \approx f(x^\star) + \langle Af,\, f\rangle - \langle f,\, b\rangle$ with $A = \partial^2 f(x^\star)$ and $b = -\nabla f(x^\star)$. So that the "local" rate, the one obtained after a large enough of iterations, is actually driven by $\tilde\rho$ and not ρ. It is thus important to distinguish between the global rate and the local rate. In practice, descent algorithm typically have two phase: a first "slow" phase govern by the global rate, and a second "fast" phase governed by the local rate. Unfortunately, the optimal step sizes τ_ℓ are in general different for the two phase, so that optimal adaptation of step size is a difficult problems. This is why more advanced users typically use various line search strategies (to find the optimal step size at each iteration) or use second order information using quasi-Newton technics (BFGS).

The convergence result of Proposition 33 does not requires strong convexity, while Theorem 34 does. In the general non-strongly convex case, it is still possible to prove convergence, but the rate is only sub-linear, and is only on the value of f, not on the iterate $x^{(\ell)}$ themselves. Note that in this case, the solution of the minimization problem is not necessarily unique. The proof is more technical.

Theorem 35. *If f satisfy conditions $(\mathcal{R}_L)$, assuming there exists $(\tau_{\min}, \tau_{\max})$ such that*

$$0 < \tau_{\min} \leqslant \tau_\ell \leqslant \tau_{\max} < \frac{2}{L}, \tag{14.11}$$

then $x^{(\ell)}$ converges to a solution $x^\star$ of (**??**) *and there exists $C > 0$ such that*

$$f(x^{(\ell)}) - f(x^\star) \leqslant \frac{C}{\ell+1}. \tag{14.12}$$

Proof. We only prove (14.12) since the proof that $x^{(\ell)}$ converges is more technical. Note indeed that if the minimizer $x^\star$ is non-unique, then it might be the case that the iterate $x^{(\ell)}$ "cycle" while approaching the set of minimizer, but actually convexity of f prevents this kind of pathological behavior. For simplicity, we do the proof in the case $\tau_\ell = 1/L$, but it extends to the general case. The L-smoothness property imply (14.3), which reads

$$f(x^{(\ell+1)}) \leqslant f(x^{(\ell)}) + \langle \nabla f(x^{(\ell)}),\, x^{(\ell+1)} - x^{(\ell)}\rangle + \frac{L}{2}\|x^{(\ell+1)} - x^{(\ell)}\|^2.$$

Using the fact that $x^{(\ell+1)} - x^{(\ell)} = -\frac{2}{L}\nabla f(x^{(\ell)})$, one obtains

$$f(x^{(\ell+1)}) \leqslant f(x^{(\ell)}) - \|\nabla f(x^{(\ell)})\|^2 \leqslant f(x^{(\ell)}) - \frac{1}{2L}\|\nabla f(x^{(\ell)})\|^2 \tag{14.13}$$

This shows that $(f(x^{(\ell)}))_\ell$ is a decaying sequence. By convexity

$$f(x^{(\ell)}) + \langle \nabla f(x^{(\ell)}),\, x^\star - x^{(\ell)}\rangle \ \leqslant f(x^\star)$$

and plugging this in (14.13) shows

$$\begin{aligned} f(x^{(\ell+1)}) &\leqslant f(x^\star) - \langle \nabla f(x^{(\ell)}),\, x^\star - x^{(\ell)}\rangle - \frac{1}{2L}\|\nabla f(x^{(\ell)})\|^2 \\ &= f(x^\star) + \frac{L}{2}\left(\|x^{(\ell)} - x^\star\|^2 - \|x^{(\ell)} - x^\star - \frac{1}{L}\nabla f(x^{(\ell)})\|^2\right) \\ &= f(x^\star) + \frac{L}{2}\left(\|x^{(\ell)} - x^\star\|^2 - \|x^\star - x^{(\ell+1)}\|^2\right). \end{aligned}$$

Summing these inequalities for $\ell = 0, \ldots, k$, one obtains

$$\sum_{\ell=1}^{k} f(x^{(\ell+1)}) - kx^\star \leqslant \frac{L}{2}\left(\|x^{(0)} - x^\star\|^2 - \|x^{(k+1)} - x^\star\|^2\right)$$

and since $f(x^{(\ell+1)})$ is decaying $\sum_{\ell=1}^{k} f(x^{(\ell+1)}) \geqslant (k+1)f(x^{(k+1)})$, thus

$$f(x^{(k+1)}) - f(x^\star) \leqslant \frac{L\|x^{(0)} - x^\star\|^2}{2(k+1)}$$

which gives (14.12) for $C \overset{\text{def.}}{=} L\|x^{(0)} - x^\star\|^2/2$. □

14.1.2 Sub-gradient Descent

The gradient descent (14.2) cannot be applied on a non-smooth function f. One can use in place of a gradient a sub-gradient, which defines the sub-gradient descent

$$x^{(\ell+1)} \overset{\text{def.}}{=} x^{(\ell)} - \tau_\ell g^{(\ell)} \quad \text{where} \quad g^{(\ell)} \in \partial f(x^{(\ell)}). \tag{14.14}$$

The main issue with this scheme is that to ensure convergence, the iterate should go to zero. One can easily convince oneself why by looking at the iterates on a function $f(x) = |x|$.

Theorem 36. *If $\sum_\ell \tau_\ell = +\infty$ and $\sum_\ell \tau_\ell^2 < +\infty$, then $x^{(\ell)}$ converges to a minimizer of f.*

14.1.3 Projected Gradient Descent

We consider a generic constraint optimization problem as

$$\min_{x \in \mathcal{C}} f(x) \tag{14.15}$$

where $\mathcal{C} \subset \mathbb{R}^S$ is a closed convex set and $f : \mathbb{R}^S \to \mathbb{R}$ is a smooth convex function (at least of class $\mathcal{C}^1$).

The gradient descent algorithm (14.2) is generalized to solve a constrained problem using the projected gradient descent

$$x^{(\ell+1)} \overset{\text{def.}}{=} \text{Proj}_{\mathcal{C}}\left(x^{(\ell)} - \tau_\ell \nabla f(x^{(\ell)})\right), \tag{14.16}$$

where $\text{Proj}_{\mathcal{C}}$ is the orthogonal projector on $\mathcal{C}$

$$\text{Proj}_{\mathcal{C}}(x) = \underset{x' \in \mathcal{C}}{\text{argmin}} \, \|x - x'\|$$

which is always uniquely defined because $\mathcal{C}$ is closed and convex. The following proposition shows that all the convergence properties of the classical gradient descent caries over to this projected algorithm.

Theorem 37. *Theorems 34 and 35 still holds when replacing iterations* (14.2) *by* (14.16).

Proof. The proof of Theorem 34 extends because the projector is contractant, $\|\text{Proj}_{\mathcal{C}}(x) - \text{Proj}_{\mathcal{C}}(x')\| \leqslant \|x - x'\|$ so that the strict contraction properties of the gradient descent is maintained by this projection. □

The main bottleneck that often prevents to use (14.16) is that the projector is often complicated to compute. We are however lucky since for ℓ^1 mininization, one can apply in a straightforward manner this method.

14.2 Proximal Operators

For non-smooth functions f, it is not possible to perform an "explicit" gradient descent step because the gradient is not even defined. One thus needs to replace this "explicit" step by an "implicit" one, which is possible even if f is non-smooth.

14.2.1 Proximal Map

The implicit stepping of amplitude $\tau > 0$ is defined as

$$\forall\, x, \quad \text{Prox}_{\tau f}(x) \overset{\text{def.}}{=} \underset{x'}{\text{argmin}} \, \frac{1}{2}\|x - x'\|^2 + f(x'). \tag{14.17}$$

It amounts to minimize function f locally around x, in a ball of radius controlled by τ. This the involved function $\frac{1}{2}\|x - \cdot\|^2 + f$ is strongly convex, this operator $\text{Prox}_{\tau f}$ is well defined and single-valued.

When $f = \iota_{\mathcal{C}}$ is an indicator, the proximal map boils down to a projection $\text{Prox}_{\iota_{\mathcal{C}}} = \text{Proj}_{\mathcal{C}}$, it is thus in some sense a generalization of the projection to arbitrary function. And can also be interpreted as a projector on a level set of f.

Examples As simple examples, let us list

$$\mathrm{Prox}_{\frac{\tau}{2}\|\cdot\|^2}(x) = \frac{x}{1+\tau}, \quad \text{and} \quad \mathrm{Prox}_{\tau\|\cdot\|_1} = \mathcal{S}_\tau^1(x),$$

where the soft-thresholding is defined as

$$\mathcal{S}_\tau^1(x) = (S_\tau(x_i))_{i=1}^N \quad \text{where} \quad S_\tau(r) = \mathrm{sign}(r)(|r| - \lambda)_+,$$

(see also (11.4)).

Note that in some case, the proximal map of a non-convex function is well defined, for instance $\mathrm{Prox}_{\tau\|\cdot\|_0}$ is the hard thresholding associated to the threshold $\sqrt{2\tau}$, see Proposition 24.

Basic properties. If $f(x) = \sum_{k=1}^K f(x_k)$ for $x = (x_1, \ldots, x_K)$ is separable, then

$$\mathrm{Prox}_{\tau f}(x) = (\mathrm{Prox}_{\tau f_k}(x_k))_{k=1}^K.$$

The following proposition is very useful.

Proposition 34. *If $A \in \mathbb{R}^{P\times N}$ is a tight frame, i.e. $AA^* = \mathrm{Id}_P$, then*

$$\mathrm{Prox}_{f\circ A} = A^* \circ \mathrm{Prox}_f \circ A + \mathrm{Id}_N - A^* A.$$

In particular, if A is orthogonal, then $\mathrm{Prox}_{f\circ A} = A^* \circ \mathrm{Prox}_f \circ A$.

Link with sub-differential. For a set-valued map $U : \mathcal{H} \hookrightarrow \mathcal{G}$, we define the inverse set-valued map $U^{-1} : \mathcal{G} \hookrightarrow \mathcal{H}$ by

$$h \in U^{-1}(g) \quad \Longleftrightarrow \quad g \in U(h)$$

One has the following equivalence

$$z = \mathrm{Prox}_{\tau f}(x) \Leftrightarrow 0 \in z - x + \tau\partial f(z) \Leftrightarrow x \in (\mathrm{Id} + \tau\partial f)(z) \Leftrightarrow z = (\mathrm{Id} + \tau\partial f)^{-1}(x)$$

where for the last equivalence, we have replace "$\in$" by "$=$" because the proximal map is single valued. The proximal operator is hence often referred to the "resolvent" $\mathrm{Prox}_{\tau f} = (\mathrm{Id}+\tau\partial f)^{-1}$ of the maximal monotone operator ∂f.

14.2.2 Proximal Point Algorithm

One has the following equivalence

$$x^\star \in \mathrm{argmin}\, f \quad \Leftrightarrow \quad 0 \in \partial f(x^\star) \quad \Leftrightarrow \quad x^\star \in (\mathrm{Id} + \tau\partial f)(x^\star) \tag{14.18}$$

$$\Leftrightarrow \quad x^\star = (\mathrm{Id} + \tau\partial f)^{-1}(x^\star) = \mathrm{Prox}_{\tau f}(x^\star). \tag{14.19}$$

This shows that being a minimizer of f is equivalent to being a fixed point of $\mathrm{Prox}_{\tau f}$. This suggest the following fixed point iterations, which are called the proximal point algorithm

$$x^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Prox}_{\tau_\ell f}(x^{(\ell)}). \tag{14.20}$$

On contrast to the gradient descent fixed point scheme, the proximal point method is converging for any sequence of steps.

Theorem 38. *If $0 < \tau_{\min} \leqslant \tau_\ell \leqslant \gamma_{\max} < +\infty$, then $x^{(\ell)} \to x^\star$ a minimizer of f.*

This implicit step (14.20) should be compared with a gradient descent step (14.2)

$$x^{(\ell+1)} \overset{\text{def.}}{=} (\mathrm{Id} + \tau_\ell \nabla f)(x^{(\ell)}).$$

One sees that the implicit resolvent $(\mathrm{Id} - \tau_\ell \partial f)^{-1}$ replaces the explicit step $\mathrm{Id} + \tau_\ell \nabla f$. For small τ_ℓ and smooth f, they are equivalent at first order. But the implicit step is well defined even for non-smooth function, and the scheme (the proximal point) is always convergent (whereas the explicit step size should be small enough for the gradient descent to converge). This is inline with the general idea the implicit stepping (e.g. implicit Euler for integrating ODE, which is very similar to the proximal point method) is more stable. Of course, the drawback is that explicit step are very easy to implement whereas in general proximal map are hard to solve (most of the time as hard as solving the initial problem).

14.2.3 Forward-Backward

It is in general impossible to compute $\mathrm{Prox}_{\gamma f}$ so that the proximal point algorithm is not implementable. In oder to derive more practical algorithms, it is important to restrict the class of considered function, by imposing some structure on f. We consider functions of the form

$$f(x) = \mathcal{F}(x) + \mathcal{G}(x) \tag{14.21}$$

where $\mathcal{G} \in \Gamma_0(\mathcal{H})$ can be an arbitrary, but $\mathcal{F}$ needs to be smooth.

One can modify the fixe point derivation (14.18) to account for this special structure

$$\begin{aligned} x^\star \in \operatorname{argmin} \mathcal{F} + \mathcal{G} \quad &\Leftrightarrow \quad 0 \in \nabla \mathcal{F}(x^\star) + \partial \mathcal{G}(x^\star) \quad \Leftrightarrow \quad x^\star - \tau \nabla \mathcal{F}(x^\star) \in (\mathrm{Id} + \tau \partial \mathcal{G})(x^\star) \\ &\Leftrightarrow \quad x^\star = (\mathrm{Id} + \tau \partial \mathcal{G})^{-1} \circ (\mathrm{Id} - \tau \nabla \mathcal{F})(x^\star). \end{aligned}$$

This fixed point suggests the following algorithm, with the celebrated Forward-Backward

$$x^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Prox}_{\tau_\ell \mathcal{G}} \left(x^{(\ell)} - \tau_\ell \nabla \mathcal{F}(x^{(\ell)}) \right). \tag{14.22}$$

Derivation using surrogate functionals. An intuitive way to derive this algorithm, and also a way to prove its convergence, it using the concept of surrogate functional.

To derive an iterative algorithm, we modify the energy $f(x)$ to obtain a surrogate functional $f(x, x^{(\ell)})$ whose minimization corresponds to a simpler optimization problem, and define the iterations as

$$x^{(\ell+1)} \overset{\text{def.}}{=} \underset{x}{\operatorname{argmin}}\, f(x, x^{(\ell)}). \tag{14.23}$$

In order to ensure convergence, this function should satisfy the following property

$$f(x) \leqslant f(x, x') \quad \text{and} \quad f(x, x) = f(x) \tag{14.24}$$

and $f(x) - f(x, x')$ should be a smooth function. Property (14.24) guarantees that f is decaying by the iterations

$$f(x^{(\ell+1)}) \leqslant f(x^{(\ell)})$$

and it simple to check that actually all accumulation points of $(x^{(\ell)})_\ell$ are stationary points of f.

In order to derive a valid surrogate $f(x, x')$ for our functional (14.21), since we assume $\mathcal{F}$ is L-smooth (i.e. satisfies ($\mathcal{R}_L$)), let us recall the quadratic majorant (14.3)

$$\mathcal{F}(x) \leqslant \mathcal{F}(x') + \langle \nabla \mathcal{F}(x'),\, x' - x \rangle + \frac{L}{2} \|x - x'\|^2,$$

so that for $0 < \tau < \frac{1}{L}$, the function

$$f(x, x') \overset{\text{def.}}{=} \mathcal{F}(x') + \langle \nabla \mathcal{F}(x'),\, x' - x \rangle + \frac{1}{2\tau} \|x - x'\|^2 + \mathcal{G}(x) \tag{14.25}$$

satisfies the surrogate conditions (14.24). The following proposition shows that minimizing the surrogate functional corresponds to the computation of a so-called proximal operator.

Proposition 35. *The update* (14.23) *for the surrogate* (14.25) *is exactly* (14.22).

Proof. This follows from the fact that

$$\langle \nabla \mathcal{F}(x'),\, x' - x \rangle + \frac{1}{2\tau}\|x - x'\|^2 = \frac{1}{2\tau}\|x - (x' - \tau \nabla \mathcal{F}(x'))\|^2 + \text{cst}.$$

□

Convergence of FB. Although we impose $\tau < 1/L$ to ensure majorization property, one can actually show convergence under the same hypothesis as for the gradient descent, i.e. $0 < \tau < 2/L$, with the same convergence rates. This means that Theorem 37 for the projected gradient descent extend to FB.

Theorem 39. *Theorems 34 and 35 still holds when replacing iterations* (14.2) *by* (14.22).

Note furthermore that the projected gradient descent algorithm (14.16) is recovered as a special case of (14.22) when setting $J = \iota_{\mathcal{C}}$ the indicator of the constraint set, since $\text{Prox}_{\rho J} = \text{Proj}_{\mathcal{C}}$ in this case.

Of course the difficult point is to be able to compute in closed form $\text{Prox}_{\tau \mathcal{G}}$ in (14.22), and this is usually possible only for very simple function. We have already seen such an example in Section 11.3.3 for the resolution of ℓ^1-regularized inverse problems (the Lasso).

14.3 Primal-Dual Algorithms

14.3.1 Forward-backward on the Dual

Chambolle's algorithm. Chambolle in [9] detail an algorithm to minimize exactly the TV denoising problem

$$x^\star_\lambda = \underset{x \in \mathcal{H}}{\text{argmin}}\ \frac{1}{2}\|x - x'\|^2 + \lambda \|x\|_{\text{TV}}. \tag{14.26}$$

It uses a relationship between the vectorial ℓ^1 and ℓ^∞ norms

$$\|v\|_1 = \sum_{m=0}^{N-1} \|v_m\| \quad \text{and} \quad \|v\|_\infty = \max_{0 \leqslant m < N} \|v_m\|$$

where each $v_m \in \mathbb{R}^2$ and $v \in \mathbb{R}^{N \times 2}$. One has

$$\|v\|_1 = \max_{\|w\|_\infty \leqslant 1} \langle w,\, u \rangle$$

which allows one to re-write the optimization (14.26) as

$$\min_{x \in \mathcal{H}} \max_{\|w\|_\infty \leqslant 1} \frac{1}{2}\|x - x'\|^2 + \lambda \langle w,\, \nabla g \rangle.$$

Exchanging the roles of the min and the max, one proves that the solution of (14.26) is re-written as

$$x^\star_\lambda = x + \lambda\, \text{div}(w^\star) \tag{14.27}$$

where

$$w^\star \in \underset{\|w\|_\infty \leqslant 1}{\text{argmin}}\ \|f + \lambda\, \text{div}(w^\star)\|^2. \tag{14.28}$$

The convex optimization problem (14.28) computes a dual vector field $w^\star \in \mathbb{R}^{N \times 2}$, from which the denoised image is recovered using (14.27).

The dual problem (14.28) is the minimization of a quadratic functional subject to a convex ℓ^∞ constraint. It can thus be solved using for instance a projected gradient descent

$$w_m^{(k+1)} = \frac{\tilde{w}_m^{(k)}}{\max(|\tilde{w}_m^{(k)}|, 1)} \quad \text{where} \quad \tilde{w}^{(k)} = w^{(k)} + \tau \nabla(f/\lambda + \operatorname{div}(w^{(k)})).$$

If the gradient step size satisfy $0 < \tau < 1/4$, one can prove that

$$f + \lambda \operatorname{div}(w^{(k)}) \longrightarrow f_\lambda^\star \quad \text{when} \quad k \to +\infty.$$

14.3.2 Douglas-Rachford

14.3.3 Primal-Dual Splitting

Bibliography

[1] P. Alliez and C. Gotsman. Recent advances in compression of 3d meshes. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 3–26. Springer Verlag, 2005.

[2] P. Alliez, G. Ucelli, C. Gotsman, and M. Attene. Recent advances in remeshing of surfaces. In *AIM@SHAPE repport*. 2005.

[3] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[7] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Académie des Sciences*, Serie I(346):589–592, 2006.

[8] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[9] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[10] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

[11] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[12] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[13] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.

[14] Philippe G Ciarlet. Introduction à l'analyse numérique matricielle et à l'optimisation. 1982.

[15] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.

[16] P. Schroeder et al. D. Zorin. Subdivision surfaces in character animation. In *Course notes at SIGGRAPH 2000*, July 2000.

[17] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.

[18] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.

[19] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

[20] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[21] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.

[22] M. S. Floater and K. Hormann. Surface parameterization: a tutorial and survey. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 157–186. Springer Verlag, 2005.

[23] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[24] I. Guskov, W. Sweldens, and P. Schröder. Multiresolution signal processing for meshes. In Alyn Rockwood, editor, *Proceedings of the Conference on Computer Graphics (Siggraph99)*, pages 325–334. ACM Press, August8–13 1999.

[25] A. Khodakovsky, P. Schröder, and W. Sweldens. Progressive geometry compression. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH-00)*, pages 271–278, New York, July 23–28 2000. ACMPress.

[26] L. Kobbelt. $\sqrt{3}$ subdivision. In Sheila Hoffmeyer, editor, *Proc. of SIGGRAPH'00*, pages 103–112, New York, July 23–28 2000. ACMPress.

[27] M. Lounsbery, T. D. DeRose, and J. Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Trans. Graph.*, 16(1):34–73, 1997.

[28] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd edition.* Academic Press, San Diego, 2009.

[29] Stephane Mallat. *A wavelet tour of signal processing: the sparse way.* Academic press, 2008.

[30] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[31] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[32] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier.* Ellipses, 2004.

[33] Gabriel Peyré and Marco Cuturi. Computational optimal transport. 2017.

[34] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.

[35] E. Praun and H. Hoppe. Spherical parametrization and remeshing. *ACM Transactions on Graphics*, 22(3):340–349, July 2003.

[36] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[37] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

[38] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging*. Springer, 2009.

[39] P. Schröder and W. Sweldens. Spherical Wavelets: Efficiently Representing Functions on the Sphere. In *Proc. of SIGGRAPH 95*, pages 161–172, 1995.

[40] P. Schröder and W. Sweldens. Spherical wavelets: Texture processing. In P. Hanrahan and W. Purgathofer, editors, *Rendering Techniques '95*. Springer Verlag, Wien, New York, August 1995.

[41] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[42] A. Sheffer, E. Praun, and K. Rose. Mesh parameterization methods and their applications. *Found. Trends. Comput. Graph. Vis.*, 2(2):105–171, 2006.

[43] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis*. Cambridge university press, 2015.

[44] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Applied and Computation Harmonic Analysis*, 3(2):186–200, 1996.

[45] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.