Proximal Methods

So far, in our consideration of optimization algorithm, we dealt with functions that are continuously differentiable at least once. This allowed e.g. to use the gradient (and, perhaps, the Hessian) to perform optimization. In what follows, we will dedicate some attention to the minimization of convex non-smooth (that is, non-differentiable) functions. Such function arise in too many important applications such as compressed sensing and sparse data modeling.

1 Subgradients and subdifferential

Recall that for a $\mathcal{C}^2$ function f we could write the first-order Taylor expansion with an exact second-order remainder,

$$f(\mathbf{x}+\mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} + \frac{1}{2}\mathbf{d}^{\mathrm{T}}\nabla^2 f(\mathbf{y})\mathbf{d}$$

where $\mathbf{y}$ is some point. If f is convex, its Hessian is positive semi-definite at every point, leading to what we called the *gradient inequality*:

$$f(\mathbf{x}+\mathbf{d}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}$$

for every $\mathbf{d}$, which geometrically means that the graph of a convex function lies entirely above the tangent plane at $\mathbf{x}$ (above all its tangents in general, since the latter holds for every $\mathbf{x}$). In other words, the first-order approximation of f is a global underestimator of f. One can show that the vector $\mathbf{g} = \nabla f(\mathbf{x})$ is the only vector for which

$$f(\mathbf{x}+\mathbf{d}) \geq f(\mathbf{x}) + \mathbf{g}^{\mathrm{T}}\mathbf{d}$$

holds for every $\mathbf{d}$.

If a function f (not necessarily convex) is non-differentiable at $\mathbf{x}$, the above argument is not valid anymore as $\nabla f(\mathbf{x})$ does not exist. However, some vectors $\mathbf{g}$ may still satisfy

$$f(\mathbf{x}+\mathbf{d}) \geq f(\mathbf{x}) + \mathbf{g}^{\mathrm{T}}\mathbf{d}$$

for every $\mathbf{d}$. Such vectors are called the *subgradients* of f at $\mathbf{x}$. Subgradients form affine global underestimators of f. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, it has at least one subgradient at every point; if it is furthermore differentiable, $\nabla f(\mathbf{x})$ is a subgradient of f at $\mathbf{x}$.

For example, if f_1, f_2 are convex and $\mathcal{C}^1$ and $f = \max\{f_1, f_2\}$ is their pointwise maximum, then at every $\mathbf{x}$ such that $f_1(\mathbf{x}) > f_2(\mathbf{x})$, f has a unique subgradient $\mathbf{g} = \nabla f_1(\mathbf{x})$; at every $\mathbf{x}$ such that $f_2(\mathbf{x}) > f_1(\mathbf{x})$, f has a unique subgradient $\mathbf{g} = \nabla f_2(\mathbf{x})$; and at every $\mathbf{x}$ such that $f_1(\mathbf{x}) = f_2(\mathbf{x})$, the subgradients form a line segment $(1-\lambda)\nabla f_1(\mathbf{x}) + \lambda \nabla f_2(\mathbf{x})$, $\lambda \in [0,1]$.

Subdifferential The set of all subgradients at $\mathbf{x}$,

$$\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^{\mathrm{T}}(\mathbf{y} - \mathbf{x})\, \forall \mathbf{y} \in \operatorname{dom} f\}$$

is called the *sub-differential* of f at $\mathbf{x}$. As we already mentioned, $\partial f(\mathbf{x}) = \{\mathbf{g}\}$ iff f is differentiable at $\mathbf{x}$, with $\mathbf{g} = \nabla f(\mathbf{x})$. It can be furthemore shown that $\partial f(\mathbf{x})$ is a closed convex set.

The following simple properties define a calculus of subdifferentials:

1. *Scaling:* $\partial(af) = a\partial f$ for every $a > 0$
2. *Addition:* $\partial(f + g) = \partial f + \partial g$
3. *Affine coordinate transformation:* if $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$, then $\partial g(\mathbf{x}) = \mathbf{A}^{\mathrm{T}} \partial f(\mathbf{A}\mathbf{x} + \mathbf{b})$

Theorem 1 (Optimality conditions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then, $\mathbf{x}^*$ is a (global) minimizer iff $\mathbf{0} \in \partial f(\mathbf{x}^*)$.*

Proof. Follows directly from the inequality

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{g}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}^*)$$

holding by definition for every $\mathbf{g} \in \partial f(\mathbf{x}^*)$. $\mathbf{0} \in \partial f(\mathbf{x}^*)$ iff $f(\mathbf{y}) \geq f(\mathbf{x}^*)$. □

A similar generalization exists for the KKT conditions.

2 Proximal operator

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ be a convex function. The *proximal operator* (a.k.a. *proximity map*) is defined as

$$\pi_f(\mathbf{x}) = \arg\min_{\mathbf{y}} f(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

In particular, when f is the indicator function of a closed convex set C,

$$f(\mathbf{x}) = I_C(\mathbf{x}) = \begin{cases} 0 & : \mathbf{x} \in C \\ \infty & : \mathbf{x} \notin C, \end{cases}$$

the proximal operator reduces to orthogonal projection onto C,

$$\pi_{I_C}(\mathbf{x}) = \arg\min_{\mathbf{y}} I_C(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 = \arg\min_{\mathbf{y} \in C} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

Proximal operators can be thought of as generalized projections and they do obey various properties obeyed by projections (e.g., they are non-expanding).

The following simple properties are easy to prove:

1. *Separability:* if $f(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) + \psi(\mathbf{y})$, then
$$\pi_f(\mathbf{x}, \mathbf{y}) = (\pi_\varphi(\mathbf{x}), \pi_\psi(\mathbf{y})).$$
In particular, if $f(\mathbf{x}) = f_1(x_1) + \cdots + f_n(x_n)$, then
$$\pi_f(\mathbf{x}) = (\pi_{f_1}(x_1), \ldots, \pi_{f_n}(x_n))^{\mathrm{T}}.$$

2. *Scaling:* for $a > 0$
$$\pi_{af+b}(\mathbf{x}) = \pi_{af}(\mathbf{x})$$

3. *Coordinate scaling:* for $f(\mathbf{x}) = \varphi(a\mathbf{x} + \mathbf{b})$ with $a \neq 0$
$$\pi_f(\mathbf{x}) = \frac{1}{a}\left(\pi_{a^2\varphi}(a\mathbf{x} + \mathbf{b}) - \mathbf{b}\right)$$

4. *Orthogonal coordinate transformation:* for $f(\mathbf{x}) = \varphi(\mathbf{Q}\mathbf{x})$ with $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}$,
$$\pi_f(\mathbf{x}) = \mathbf{Q}^{\mathrm{T}}\pi_\varphi(\mathbf{Q}\mathbf{x})$$

5. *Affine addition:* for $f(\mathbf{x}) = \varphi(\mathbf{x}) + \mathbf{a}^{\mathrm{T}}\mathbf{x} + b$,
$$\pi_f(\mathbf{x}) = \pi_\varphi(\mathbf{x} - \mathbf{a})$$

3 Resolvent of the subdifferential operator

We can interpret the subdifferential operator ∂f of a convex function f on $\mathbb{R}^n$ as a point-to-set mapping or a relation on $\mathbb{R}^n$ mapping a point $\mathbf{x} \in \mathbb{R}^n$ to the set $\partial f(\mathbf{x})$. When f is differentiable, the map becomes a point-to-point map. The mapping $(\mathbf{I} + \lambda\partial f)^{-1}$, where $\mathbf{I}$ stands for the identity operator, is called the *resolvent* of the operator ∂f with parameter $\lambda > 0$. Note that all the operations in $(\mathbf{I} + \lambda\partial f)^{-1}$ are operations on relations; nevertheless, the resolvent is a function thought ∂f itself is not.

Theorem 2 (Resolvent ofthe subdifferential operator)**.**
$$\mathrm{prox}_{\lambda f} = (\mathbf{I} + \lambda\partial f)^{-1}$$

Proof. By definition, if $\mathbf{z} \in (\mathbf{I} + \lambda\partial f)^{-1}(\mathbf{x})$, then
$$\mathbf{x} \in (\mathbf{I} + \lambda\partial f)(\mathbf{z}) = \mathbf{z} + \lambda\partial f(\mathbf{z}).$$
Re-arranging the terms,
$$\mathbf{0} \in \partial f(\mathbf{z}) + \frac{1}{\lambda}(\mathbf{z} - \mathbf{x}) = \partial_{\mathbf{z}}\left(\lambda f(\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2\right).$$
This is the necessary and sufficient condition for
$$\mathbf{z} = \arg\min_{\mathbf{y}} \lambda f(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 = \pi_{\lambda f}(\mathbf{x}),$$
which shows that $\mathbf{z} \in (\mathbf{I} + \lambda\partial f)^{-1}(\mathbf{x})$ iff $\mathbf{z} = \pi_{\lambda f}(\mathbf{x})$ and, in particualr, that the latter resolvent is single-valued. □

4 Connection to gradient and Newton steps

Let us now assume that f is twice differentiable at $\mathbf{x}$ with $\nabla^2 f(\mathbf{x}) \succ 0$. Then, $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ and from first-order Taylor expansion

$$\pi_{\lambda f}(\mathbf{x}) = (\mathbf{I} + \lambda \nabla f)^1(\mathbf{x}) = \mathbf{x} - \lambda \nabla f(\mathbf{x}) + \mathcal{O}(\lambda^2).$$

In other words, for $\lambda \to 0$, the proximal operator converges to a gradient step with step size λ. Let us denote the gradient and the Hessian at $\mathbf{x}$ by $\mathbf{g} = \nabla f(\mathbf{x})$ and $\mathbf{H} = \nabla^2 f(\mathbf{x})$, respectively. Let us also denote the first- and second-order approximations of f as

$$\hat{f}_1(\mathbf{y}) = f(\mathbf{x}) + \mathbf{g}^{\mathrm{T}}(\mathbf{y} - \mathbf{x})$$

and

$$\hat{f}_2(\mathbf{y}) = f(\mathbf{x}) + \mathbf{g}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^{\mathrm{T}}\mathbf{H}(\mathbf{y} - \mathbf{x}),$$

respectively.

The proximal operator of $\hat{f}_1$ yields

$$\begin{aligned} \pi_{\lambda \hat{f}_1}(\mathbf{x}) &= \arg\min_{\mathbf{y}} \lambda f(\mathbf{x}) + \lambda \mathbf{g}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \\ &= \arg\min_{\mathbf{y}} \lambda \mathbf{g}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2. \end{aligned}$$

Demanding

$$\mathbf{0} = \nabla_{\mathbf{y}} \left(\lambda \mathbf{g}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \right) = \lambda \mathbf{g} + \mathbf{y} - \mathbf{x}$$

yields $\mathbf{y} = \pi_{\lambda \hat{f}_1}(\mathbf{x}) = \mathbf{x} - \lambda \mathbf{g}$, which is again a gradient step with stepsize λ.

Similarly, the proximal operator of $\hat{f}_2$ yields

$$\pi_{\lambda \hat{f}_2}(\mathbf{x}) = \mathbf{x} - \left(\mathbf{H} + \frac{1}{\lambda}\mathbf{I} \right)^{-1} \mathbf{g},$$

which is a modified Newton step.

5 Fixed points

Theorem 3 (Fixed point)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. Then*

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \iff \mathbf{x}^* = \pi_f(\mathbf{x}^*).$$

Proof. We prove the theorem for a subdifferentiable function, though the result is more general. Let $\mathbf{x}^*$ be a global minimizer of f, implying that for every $\mathbf{x}$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$. Then, for every $\mathbf{x}$,

$$f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 \geq f(\mathbf{x}^*) = f(\mathbf{x}^*) + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}^*\|_2^2 = \min_{\mathbf{x}} f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2,$$

implying that

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 = \pi_f(\mathbf{x}^*).$$

To prove the converse, recall that

$$\mathbf{y} = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 = \pi_f(\mathbf{z}).$$

iff

$$\mathbf{0} \in \partial\left(f(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{z}\|_2^2\right) = \partial f(\mathbf{y}) + (\mathbf{y} - \mathbf{z}).$$

Substituting $\mathbf{y} = \mathbf{z} = \mathbf{x}^*$ yields $\mathbf{0} \in \partial f(\mathbf{x}^*)$, meaning that $\mathbf{x}^* = \pi_f(\mathbf{x}^*)$ is a minimizer of f. □

A central result in functional analysis is the Banach fixed point theorem stating that if an operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is a *contraction*, i.e., there exists some $c < 1$ such that for every $\mathbf{x}, \mathbf{y}$, $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq c\|\mathbf{x} - \mathbf{y}\|$, then T has a unique fixed point $\mathbf{x}^* = T(\mathbf{x}^*)$. Furthermore, a sequence $\{\mathbf{x}_0, \mathbf{x}_1 = T(\mathbf{x}_0), \mathbf{x}_2 = T(\mathbf{x}_1), \ldots, , \mathbf{x}_{n+1} = T(\mathbf{x}_n), \ldots\}$ starting with any $\mathbf{x}_0$ converges to $\mathbf{x}_n \to \mathbf{x}^*$. The above iterative process is known as *fixed point iteration*, as it suggests how to find the fixed point.

Unfortunately, proximal operators are not contractions. While they are non-expanding (like projections), this alone does not guarantee convergence of $\pi^n(\mathbf{x}_0)$ to a fixed point (think of a rotation operator for example). However, proximal operators are *firmly non-expanding*, meaning that for every $\mathbf{x}, \mathbf{y}$

$$\|\pi_f(\mathbf{x}) - \pi_f(\mathbf{y})\|_2^2 \leq (\mathbf{x} - \mathbf{y})^{\mathrm{T}}(\pi_f(\mathbf{x}) - \pi_f(\mathbf{y})).$$

This property appears sufficient (the so-called Krasnoselskii-Mann theorem) to show that the iteration $\mathbf{x}_{k+1} = \pi_f(\mathbf{x}_k)$ converges to a fixed point of π_f.

6 Proximal gradient

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex function, $f \in \mathcal{C}^1$. We are interested in solving

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}).$$

The following iteration is known as the *proximal gradient method*:

$$\mathbf{x}_{k+1} = \pi_{\lambda_k g}(\mathbf{x}_k - \lambda_k \nabla f(\mathbf{x}_k)),$$

where $\lambda_k > 0$ is a step size.

When ∇f is Lipschitz continuous with a constant L, proximal gradient converges with the rate $\mathcal{O}(1/k)$ when a fixed step size $\lambda_k = \lambda \in (0, 1/L]$ is used. A simple modification allows these methods to converge faster, as $\mathcal{O}(1/k^2)$:

$$\begin{aligned} \mathbf{y}_{k+1} &= \mathbf{x}_k + \omega_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \\ \mathbf{x}_{k+1} &= \pi_{\lambda_k g}(\mathbf{y}_{k+1} - \lambda_k \nabla f(\mathbf{y}_{k+1})), \end{aligned}$$

where $\omega_k = k/(k+3)$.

When $g = I_{\mathcal{C}}$, the proximal gradient step becomes a regular projected gradient; when $f = 0$, it reduces to proximal minimization (our previous fixed point algorithm); when $g = 0$ it reduces to gradient descent.

7 Alternating direction method of multipliers

Let us again consider the unconstrained problem with a split objective

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}),$$

but now we split $\mathbf{x}$ into two variables, $\mathbf{x}$ and $\mathbf{z}$ and add a consensus constraint,

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + g(\mathbf{z}) \text{ s.t. } \mathbf{x} = \mathbf{z}.$$

The corresponding augmented Lagrangian is

$$L(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^{\mathrm{T}}(\mathbf{x} - \mathbf{z}) + \frac{p}{2}\|\mathbf{x} - \mathbf{z}\|_2^2,$$

with $p > 0$ and $\mathbf{y}$ playing the role of Lagrange multipliers. We can write the following augmented Lagrangian iteration

$$\begin{aligned} \mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}} L(\mathbf{x}, \mathbf{z}_k, \mathbf{y}_k) = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \mathbf{y}_k^{\mathrm{T}}\mathbf{x} + \frac{p}{2}\|\mathbf{x} - \mathbf{z}_k\|_2^2 \\ \mathbf{z}_{k+1} &= \arg\min_{\mathbf{z}} L(\mathbf{x}_{k+1}, \mathbf{z}, \mathbf{y}_k) = \arg\min_{\mathbf{z}} g(\mathbf{z}) - \mathbf{y}_k^{\mathrm{T}}\mathbf{z} + \frac{p}{2}\|\mathbf{x}_{k+1} - \mathbf{z}\|_2^2 \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + p\,(\mathbf{x}_{k+1} - \mathbf{z}_{k+1}), \end{aligned}$$

where L is minimized w.r.t. $\mathbf{x}$ and $\mathbf{z}$ independently. Denoting $\mathbf{u}_k = \mathbf{y}_k/p$ and pushing the linear terms into the quadratic term yields the following iteration

$$\begin{aligned} \mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{p}{2}\|\mathbf{x} - (\mathbf{z}_k - \mathbf{u}_k)\|_2^2 = \pi_{f/p}(bbz_k - \mathbf{u}_k) \\ \mathbf{z}_{k+1} &= \arg\min_{\mathbf{z}} g(\mathbf{z}) + \frac{p}{2}\|\mathbf{z} - (\mathbf{x}_{k+1} + \mathbf{u}_k)\|_2^2 = \pi_{g/p}(bbx_{k+1} + \mathbf{u}_k) \\ \mathbf{u}_{k+1} &= \mathbf{u}_k + \mathbf{x}_{k+1} - \mathbf{z}_{k+1} \end{aligned}$$

known as the *alternating direction method of multipliers* (ADMM, a.k.a. Douglas-Rachford or Bregman splitting).