5 Proximal and Bundle methods

Going beyond first order can be highly beneficial in nonsmooth optimization in order not to rely on the rather loose local information brought by the subgradient.

In the previous section, we investigated first-order methods that could be seen as iterations of the type

$$\begin{aligned} & x_{k+1} = x_k - \gamma\ v_k \text{ with } v_k \in \partial F(x_k) \\ \Leftrightarrow\ & x_{k+1} = \operatorname{argmin}_u \left\{ \frac{1}{2}\|u - (x_k - \gamma\ v_k)\|_2^2 \right\} \\ \Leftrightarrow\ & x_{k+1} = \operatorname{argmin}_u \left\{ \langle -\gamma\ v_k ; x_k - u\rangle + \frac{1}{2}\|u - x_k\|_2^2 \right\} \\ \Leftrightarrow\ & x_{k+1} = \operatorname{argmin}_u \left\{ F(x_k) + \langle v_k ; u - x_k\rangle + \frac{1}{2\gamma}\|u - x_k\|_2^2 \right\}. \end{aligned}$$

Recalling that by definition (see Lecture 4)

$$\partial F(x_k) = \{ v : F(u) \geq F(x_k) + \langle v, u - x_k\rangle \text{ for all } u \in \mathbb{R}^n \},$$

a subgradient step can be seen as:

$$x_{k+1} = \operatorname{argmin}_u \left\{ \underbrace{F(x_k) + \langle v_k ; u - x_k\rangle}_{\text{(a)}} + \underbrace{\frac{1}{2\gamma}\|u - x_k\|_2^2}_{\text{(b)}} \right\}.$$

where:

(a) is a linear/first-order model that under-approximates F;

(b) is a quadratic recall/regularization/stabilization term controlled by γ.

In this chapter, we will investigate algorithms that minimize stabilized approximations of the function.

5.1 The Proximity Operator

A central tool to tackle nonsmooth functions is the *proximity operator*, introduced by [3], and denoted $\mathbf{prox}_{\gamma F}$ for a step-size $\gamma > 0$ and a nonsmooth function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$; it is defined as the set-valued mapping

$$\mathbf{prox}_{\gamma F}(y) := \operatorname{argmin}_{u \in \mathbb{R}^n} \Big\{ \underbrace{F(u) + \frac{1}{2\gamma}\|u - y\|^2}_{:= \rho_y(u)} \Big\}.$$

In the same flavor as for the gradient step, if one takes a proximal step, i.e.,

$$x = \mathbf{prox}_{\gamma F}(y)$$

for some $\gamma > 0$, the definition directly gives us

$$F(x) \leq F(y) - \frac{1}{2\gamma}\|x - y\|^2 \tag{1}$$

which mirrors the descent inequality of a gradient step on a smooth function.[1]

With this respect, the proximity operator provides a alternative to the use of subgradients or nonsmooth gradients since they are not able to provide descent inequalities such as (1). However, this comes to the cost of having to solve a minimization subproblem, which in turn question about the existence and uniqueness of the subproblem solutions.

[1] Actually, this link can be made formal since a proximal step is equivalent to a gradient step on the *Moreau envelope* defined for all $y \in \mathbb{R}^n$ as $e_\gamma F(y) = \inf_{u \in \mathbb{R}^n} \rho_y(u)$ [3, 4].

5.1.1 Properties

First, for convex functions the proximity operator exists and is unique.

Theorem 1. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then $\mathbf{prox}_{\gamma F}(y)$ is a singleton for any $\gamma > 0$ and any $y \in \mathbb{R}^n$.*

Proof. ($\star$) Since F is convex, ρ_y is $\frac{1}{\gamma}$-strongly convex. Then, we have the existence and uniqueness of the minimizers of $\rho_y(u)$ for any u, which means that $\mathbf{prox}_{\gamma F}(y)$ is well-defined and unique. □

In addition, we have that

$$x = \mathbf{prox}_{\gamma F}(y) \;\Leftrightarrow\; x = y - \gamma v \text{ with } v \in F(x)$$

and thus the proximal operator can be seen as an *implicit* subgradient descent step.

Proposition 2. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then the following propositions are equivalent:*

i) $x = \mathbf{prox}_{\gamma F}(y)$;

ii) $(y - x)/\gamma \in \partial F(x)$;

iii) $F(u) \geq F(x) + \langle (y - x)/\gamma, u - x\rangle$ for any $u \in \mathbb{R}^n$.

Proof. This follow directly from Fermat's rule and the definition of a convex subgradient. □

5.1.2 Convergence of the proximal point algorithm

Now, let us investigate the proximal point algorithm:

$$x_{k+1} = \mathbf{prox}_{\gamma F}(x_k) \tag{Proximal Point}$$

The first thing to notice is that the fixed points of this algorithm correspond to the minimizers of F.

Corollary 3. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function, then $x^\star$ is a minimizer of F if and only if $x^\star = \mathbf{prox}_{\gamma F}(x^\star)$ (for any $\gamma > 0$).*

Proof. From Proposition 2, we have that $x^\star = \mathbf{prox}_{\gamma F}(x^\star)$ if and only if $0 \in \partial F(x^\star)$ which is equivalent to $x^\star$ being a minimizer of F since it is convex. □

Now, we can analyze the convergence of our proximal point method.

Theorem 4. *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex lower semi-continuous proper function. Then, the Proximal Point method with $\gamma > 0$ verifies $F(x_{k+1}) \leq F(x_k)$ and*

- *(convergence) $x_k \to x^\star$ for some minimizer $x^\star$ of F;*
- *(rate) $F(x_k) - F(x^\star) \leq \frac{\|x^\star - x_0\|^2}{2\gamma k}$.*

Proof. We left the iterates convergence proof as an exercise, its reasoning is exactly the same as the one for the gradient algorithm. For the rate, since $x_{k+1} = \mathbf{prox}_{\gamma F}(x_k)$,

$$F(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 \leq F(x_k)$$

and thus $F(x_{k+1}) \leq F(x_k)$.

Since $x_{k+1} = \mathbf{prox}_{\gamma F}(x_k)$, it is the minimum of the $1/\gamma$-strongly convex function ρ_{x_k},[2] thus

$$F(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 \leq F(x^\star) + \frac{1}{2\gamma}\|x^\star - x_k\|^2 - \frac{1}{2\gamma}\|x_{k+1} - x^\star\|^2$$

[2] If $x^\star$ is the minimizer of a μ-strongly convex function F, then $F(x^\star) \leq F(u) - \frac{\mu}{2}\|u - x^\star\|^2$.

and by summing this inequality from $t = 0, .., k-1$, we get

$$\sum_{t=0}^{k-1} (F(x_{t+1}) - F(x^\star)) \leq \frac{1}{2\gamma} \sum_{t=0}^{k-1} (\|x^\star - x_t\|^2 - \|x_{t+1} - x^\star\|^2) - \sum_{t=0}^{k-1} \frac{1}{2\gamma} \|x_{t+1} - x_t\|^2$$
$$\leq \frac{1}{2\gamma} \|x^\star - x_0\|^2.$$

Now, since $F(x_{k+1}) \leq F(x_k)$, we get that

$$k\left(F(x_k) - F(x^\star)\right) \leq \sum_{t=0}^{k-1} (F(x_t) - F(x^\star)) \leq \frac{1}{2\gamma} \|x^\star - x_0\|^2$$

which gives the result. □

5.1.3 Examples of closed form expressions

Example 5 (Squared norm). For $F(x) = \frac{1}{2}\|x\|^2$, the proximity operator can be computed explicitly. Since $\rho_y : u \mapsto F(u) + \frac{1}{2\gamma}\|u - y\|^2$ is strongly convex, there is a unique minimizer x and it verifies $\nabla \rho_y(x) = 0$. Thus $x + \frac{1}{\gamma}(x - y) = 0$ which implies $x = y/(1+\gamma)$:

$$\mathbf{prox}_{\gamma \frac{1}{2}\|\cdot\|^2}(y) = \frac{y}{1+\gamma}.$$

Example 6 (Projection). In optimization, it is useful to define the *indicator* of set $C \subset \mathbb{R}^n$ as the function $\iota_C : \mathbb{R}^n \to \overline{\mathbb{R}}$ such that[3]

$$\iota_C(x) = \left\{ \begin{array}{ll} 0 & \text{if } x \in C \\ +\infty & \text{else} \end{array} \right. .$$

By noticing that the inner minimization in $\mathbb{R}^n$ amounts to minimizing only over C since otherwise the inner value is $+\infty$, this exactly gives the definition of the projection operator. Thus, for $C \subset \mathbb{R}^n$ a non-empty closed convex set and any $\gamma > 0$,

$$\mathbf{prox}_{\gamma \iota_C}(y) = \mathrm{proj}_C(y).$$

Note that the stepsize does not play any role here.

Example 7 (Absolute value). The proximity operator of the absolute value admits a closed form expression: for $y \in \mathbb{R}$ and $\gamma > 0$,

$$\mathbf{prox}_{\gamma|\cdot|}(y) = \begin{cases} y + \gamma & \text{if } y < -\gamma \\ 0 & \text{if } -\gamma \leq y \leq \gamma \\ y - \gamma & \text{if } y > \gamma \end{cases}$$

A very useful calculus rule for the proximity operator is that if F is separable:[4]

$$F(x_1, x_2, .., x_m) = \sum_{i=1}^{m} F_i(x_i),$$

then the proximity operator of F can be obtained from those of the (F_i):

$$\mathbf{prox}_{\gamma F}(y_1, y_2, .., y_m) = \begin{bmatrix} \mathbf{prox}_{\gamma F_1}(y_1) \\ \mathbf{prox}_{\gamma F_2}(y_2) \\ \vdots \\ \mathbf{prox}_{\gamma F_m}(y_m) \end{bmatrix}.$$

[3] This is different from the indicator I_A in probability which is equal to 1 if A is true and 0 elsewhere.

[4] By coordinates, or blocks of coordinates.

Example 8 (ℓ_1 norm). The ℓ_1-norm is defined on $\mathbb{R}^n$ as $\|x\|_1 = \sum_{i=1}^n |x_i|$. Using this separability, the proximity operator at $y \in \mathbb{R}^n$ and $\gamma > 0$,

$$\mathbf{prox}_{\gamma\|\cdot\|_1}(y) = \begin{bmatrix} \mathbf{prox}_{\gamma|\cdot|}(y_1) \\ \mathbf{prox}_{\gamma|\cdot|}(y_2) \\ \vdots \\ \mathbf{prox}_{\gamma|\cdot|}(y_m) \end{bmatrix}.$$

For more examples, see [1, Chap. 6] and the website proximity-operator.net.

5.1.4 When no closed form is available

Computing the proximity operator amounts to solving a new problem at each iteration. However, since the problem is made strongly convex (or better conditioned), this new subproblem may be easier to solve. However, this leads to a bi-level implementation. In the following section, we see an intermediate approach that allows not to compute a minimizer of the full function but rather of a piecewise linear approximation.

5.2 Bundle methods

A whole class of method is based on approximating the function using a *bundle* of past information and minimizing this approximation to provide a new point of query for enriching our approximation.

5.2.1 Cutting planes

Essentially, (sub)-gradient methods use only once the (sub)gradient information. In other words, the model of F used at iteration k is simply

$$\check{F} : u \mapsto F(x_k) + \langle v_k; u - x_k\rangle.$$

An alternative is to use all the information before k, ie.

$$F_\ell = F(x_\ell) \text{ and } v_\ell \in \partial F(x_\ell) \text{ for } \ell = 0, .., k$$

to form a *cutting plane* model

$$\check{F}_k : u \mapsto \max_{\ell=0,..,k} \{F_\ell + \langle v_\ell; u - x_\ell\rangle\}.$$

The function $\check{F}_k$ is

- convex and piecewise linear since this is a maximum of linear functions
- always below F: $\check{F}_k(x) \leq F(x)$ for all x
- increasing with k: $\check{F}_k(x) \leq \check{F}_{k+1}(x)$ for all x

Thus, minimizing $\check{F}_k$ is equivalent to solving the linear problem

$$\begin{aligned} \min_{x,t} \quad & t \\ \text{s.t.} \quad & F_\ell + \langle v_\ell; x - x_\ell\rangle \leq t \text{ for } \ell = 0, .., k \end{aligned}$$

however, this problem may be unbounded, so we need to add a compact convex constraint.

This leads to the cutting planes method

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_{u\in C} \check{F}_k(u) \\ F_{k+1} &= F(x_{k+1}) \\ v_{k+1} &\in \partial F(x_{k+1}) \end{aligned} \qquad \text{(Cutting planes)}$$

where at each iteration a constrained linear cutting plane problem has to be solved.

Theorem 9. *Let F be a convex proper lower semi-continuous function and let C be a compact convex set. Assume that F is M Lipschitz and that $\|\partial F(x)\| \leq M$ for all $x \in C$. Then, for any $\texttt{tol} > 0$, after a finite number k of iterations*

$$F(x_{k+1}) \leq \check{F}_k(x_{k+1}) + \texttt{tol}$$

and then,

$$F(x_{k+1}) \leq \min_C F + \texttt{tol}.$$

The theorem thus grants convergence to a minimizer up to some tolerance and also provides a way to stop the algorithm.

Proof. Part 1: We know that $\check{F}_k(x_{k+1}) \leq \min_C F$ since x_{k+1} is the minimizer of the lower model.

Assume for contradiction that $\check{F}_k(x_{k+1}) \leq \min_C F - b$ with $b > 0$ for all k. Since C is compact, we can extract a converging subsequence $(x_{k(\ell)})$ from (x_k). Take ℓ large enough so that $\|x_{k(\ell+1)} - x_{k(\ell)}\| \leq b/(2M)$. Then,

$$\begin{aligned}
\min_C F - b &\geq \check{F}_{k(\ell+1)-1}(x_{k(\ell+1)}) \\
&\geq \check{F}_{k(\ell)}(x_{k(\ell+1)}) \quad \text{(since } \check{F}_k \text{ is non-decreasing)} \\
&\geq F_{k(\ell)} + \langle v_{k(\ell)}; x_{k(\ell+1)} - x_{k(\ell)} \rangle \\
&\geq F_{k(\ell)} - M\|x_{k(\ell+1)} - x_{k(\ell)}\| \\
&\geq \min_C F - b/2
\end{aligned}$$

which is a contradiction thus $\check{F}_k(x_{k+1}) \to \min_C F$ (as the sequence is non-decreasing).
Part 2: Now, $F(x_k) \geq \min_C F$ by definition.

Similarly, assume for contradiction that $F(x_k) \geq \min_C F + b$ with $b > 0$ for all k and use the same subsequence as above. Then,

$$\begin{aligned}
\min_C F + b &\leq F(x_{k(\ell)}) \\
&= \check{F}_{k(\ell)}(x_{k(\ell)}) \\
&= \check{F}_{k(\ell)}(x_{k(\ell)}) - \check{F}_{k(\ell)}(x_{k(\ell+1)}) + \check{F}_{k(\ell)}(x_{k(\ell+1)}) \\
&\leq M\|x_{k(\ell)} - x_{k(\ell+1)}\| + \check{F}_{k(\ell+1)-1}(x_{k(\ell+1)}) \text{ (since } \check{F}_{k(\ell)} \leq \check{F}_{k(\ell+1)-1} \text{)} \\
&\leq b/2 + \min_C F \text{ (since } \check{F}_k(x_{k+1}) \leq \min_C F \text{ for all } k\text{)}
\end{aligned}$$

which is again a contradiction thus $F(x_{k(\ell)}) \to \ \liminf F(x_k) = \min_C F$.
Part 3: Finally, putting the two parts above together, we know that for any $\texttt{tol} > 0$, by taking ℓ sufficiently large, we have

$$\check{F}_{k(\ell)-1}(x_{k(\ell)}) \geq \min_C F - \frac{\texttt{tol}}{2} \qquad \text{and} \qquad F(x_{k(\ell)}) \leq \min_C F + \frac{\texttt{tol}}{2}$$

and thus

$$F(x_{k(\ell)}) \leq \min_C F + \frac{\texttt{tol}}{2} \leq \check{F}_{k(\ell)-1}(x_{k(\ell)}) + \texttt{tol}$$

which is our stopping criterion. Thus, the algorithm stops in finite time.

Now, we have

$$\begin{aligned}
F(x_{k(\ell)}) &\leq \check{F}_{k(\ell)-1}(x_{k(\ell)}) + \texttt{tol} \\
&\leq \check{F}_{k(\ell)-1}(x) + \texttt{tol} \text{ for any } x \in C \text{ (by definition of } x_{k(\ell)}) \\
&\leq F(x) + \texttt{tol} \text{ for any } x \in C \text{ (since } \check{F}_{k(\ell)-1} \text{ is a lower-model)}
\end{aligned}$$

which means that $F(x_{k(\ell)}) \leq \min_C F + \texttt{tol}$, which is our result. □

Even though they offer rather good convergence properties and can be very efficient if the function is V-shaped or polyhedral, the cutting planes methods also suffers from numerical instability, increasing computational complexity with the iterations, and can be particularly bad at minimizing some functions.

For instance, consider the function $F(x) = x^2/2$ in one dimension.[5] Then, if $x_0 = 1$, $x_1 = -\varepsilon < 0$, then $x_2 = (1-\varepsilon)/2$. This means that the closest x_1 to the solution, the further x_2. This is a typical *instability* behavior. It is also *not a descent* method since the functional value can increase at each iteration.

The knowledge of the past values (F_ℓ, x_ℓ, v_ℓ) is called a *bundle* of information and gives its name to the general class of algorithms using them. We will now see how these methods can be stabilized.

5.2.2 Proximal bundle

Now, we give an example of method that uses bundles of information but features:

- a quadratic recall term to take care of the instability behavior;
- a descent test to have a descent method.

The *proximal bundle* method can be written as

$$\begin{aligned}
x_{k+1} &= \operatorname{argmin}_{u\in C} \check{F}_k(u) + \frac{1}{2\gamma}\|u - \hat{x}_k\|^2 \\
\delta_{k+1} &= F(\hat{x}_k) - \check{F}_k(x_{k+1}) - \frac{1}{2\gamma}\|x_{k+1} - \hat{x}_k\|^2 \\
F_{k+1} &= F(x_{k+1}) \\
v_{k+1} &\in \partial F(x_{k+1}) \\
\hat{x}_{k+1} &= \begin{cases} x_{k+1} & \text{if } F(x_{k+1}) \le F(\hat{x}_k) - \kappa\delta_{k+1} \quad \text{(serious step)} \\ \hat{x}_k & \text{otherwise} \quad \text{(null step)} \end{cases}
\end{aligned} \tag{Proximal Bundle}$$

First, notice that since

$$\begin{aligned}
-\delta_{k+1} &= \check{F}_k(x_{k+1}) + \frac{1}{2\gamma}\|x_{k+1} - \hat{x}_k\|^2 - F(\hat{x}_k) \\
&\le \check{F}_k(\hat{x}_k) + \frac{1}{2\gamma}\|\hat{x}_k - \hat{x}_k\|^2 - F(\hat{x}_k) = \check{F}_k(\hat{x}_k) - F(\hat{x}_k) \le 0,
\end{aligned}$$

we indeed have $\delta_{k+1} \ge 0$.

Using the same techniques as before, one can prove that i) there cannot be infinitely many consecutive null steps; and ii) the sequence $(\hat{x}_k)$ minimizes F on C.

Theorem 10. *Let F be a convex proper lower semi-continuous function and let C be a closed convex set. Assume that F is M Lipschitz and that $\|\partial F(x)\| \le M$ for all $x \in C$. Then, for any $\gamma > 0$, $\kappa \in (0,1)$, and* `tol` > 0*, after a finite number k of iterations*

$$F(\hat{x}_k) \le \min_C F + \texttt{tol}.$$

Exercises

Exercise 1. Show that the proximity operator of $h : x \mapsto |x-1|$ is given for some stepsize γ as

$$\mathbf{prox}_{\gamma h}(y) = \begin{cases} y+\gamma & \text{if } y < 1-\gamma \\ 1 & \text{if } y \in [1-\gamma, 1+\gamma] \\ y-\gamma & \text{if } y \in [1-\gamma, 1+\gamma] \end{cases}$$

Exercise 2. Compute the proximity operator of the $\mathbb{R} \to \mathbb{R}$ function

$$F(x) = \sum_{i=1}^{n} |x - i|$$

for some positive integer n.

[5] This example is taken from [2, Chap. XV.1.1].

Exercise 3. Let us consider the $\mathbb{R} \to \mathbb{R}$ function

$$F(x) = \frac{a}{2}x^2 + |x-1|$$

for some $a > 0$.

We wish to use a modified cutting-plane method to minimize F. Given $F_\ell = F(x_\ell)$ and $v_\ell \in \partial F(x_\ell)$ for $\ell = 1, .., k$, we form the function

$$\check{F}_k : u \mapsto \max_{\ell=1,..,k} \left\{ F_\ell + \langle v_\ell, u - x_\ell \rangle + \frac{a}{2}(u - x_\ell)^2 \right\}$$

and wish to apply the algorithm

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_{u \in \mathbb{R}} \check{F}_k(u) \\ F_{k+1} &= F(x_{k+1}) \ , \qquad v_{k+1} \in \partial F(x_{k+1}) \end{aligned}$$

1. Show that $\check{F}_k$ is i) convex and ii) always below F (*hint: show first that F is a-strongly convex*).
2. Do you think the algorithm converges (no formal proof is expected, only high level arguments)? What are the most striking differences between this and the standard cutting-planes method?

References

[1] Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

[2] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms.* Springer Verlag, Heidelberg, 1993. Two volumes.

[3] Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

[4] Kôsaku Yosida. *Functional analysis*, volume 123. springer, 1988.