10 Proximal methods

Proximal operator The proximal mapping is a "functional" generalization of the projection mapping. Given a convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, the proximal mapping associated to f is

$$\mathbf{prox}_f(y) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2}\|x - y\|_2^2 \right\}. \tag{1}$$

Clearly the proximal operator of the indicator function I_C of a closed convex set is precisely the projection operator.

The next proposition guarantees that $\mathbf{prox}_f$ is well-defined under mild conditions on f. A function f is *lower-semicontinuous* (lsc) if $f(x) \leq \liminf_{i\to\infty} f(x_i)$ for any sequence (x_i) converging to x.

EXERCISE: Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$. Prove that the following are equivalent: (i) f is lower-semicontinuous, (ii) $\mathbf{epi}(f)$ is closed, (iii) all the sublevel sets $f^{-1}((-\infty, a])$ are closed.

Proposition 10.1. *If f is lower-semicontinuous, then $\mathbf{prox}_f(y)$ is well-defined for all $y \in \mathbb{R}^n$.*

Proof. Let $g(x) = f(x) + (1/2)\|x - y\|_2^2$. Since g is strongly convex, any minimizer is necessarily unique. It remains to show that a minimizer exists. First note that g is bounded below: since f is convex it can be lower bounded by an affine function $f(x) \geq \langle a, x\rangle + b$, and so $g(x) \geq \langle a, x\rangle + b + (1/2)\|x-y\|_2^2 \geq \min_{x\in\mathbb{R}^n}\{\langle a, x\rangle + b + (1/2)\|x-y\|_2^2\} = c > -\infty$. Also note that the sublevel sets of g are all bounded since $g(x) \leq t \implies \langle a, x\rangle + b + (1/2)\|x-y\|_2^2 \leq t \iff \|x-(y-a)\|_2^2 \leq C$ for some constant $C > 0$. Now let (x_i) be a sequence so that $g(x_i) \downarrow \inf_{x\in\mathbb{R}^n} g(x)$. The sequence (x_i) lives in the sublevel set $\{x : g(x) \leq g(x_1)\}$ which is closed and bounded. Thus we can extract from (x_i) a converging subsequence, that converges to some x. Since g is lower semicontinuous we have $g(x) \leq \liminf_i g(x_i) = \inf g$, and so x is a minimizer of g. □

Note that

$$x = \mathbf{prox}_f(y) \iff 0 \in \partial f(x) + (x - y) \iff y \in x + \partial f(x). \tag{2}$$

Remark 1. *If f is smooth, we see that $x = \mathbf{prox}_f(y)$ is a solution to the nonlinear equation $x + \nabla f(x) = y$, i.e., it satisfies $x = (I + \nabla f)^{-1}(y)$.*

Just like with the projection, one can prove that the proximal map is nonexpansive, i.e., that

$$\|\mathbf{prox}_f(y_1) - \mathbf{prox}_f(y_2)\|_2 \leq \|y_1 - y_2\|_2.$$

To see why, let $x_1 = \mathbf{prox}_f(y_1)$ and $x_2 = \mathbf{prox}_f(y_2)$. Then $y_1 - x_1 \in \partial f(x_1)$, and so we can write:

$$f(x_2) \geq f(x_1) + \langle y_1 - x_1, x_2 - x_1\rangle.$$

Similarly, from $y_2 - x_2 \in \partial f(x_2)$, we get

$$f(x_1) \geq f(x_2) + \langle y_2 - x_2, x_1 - x_2\rangle.$$

Summing the two inequalities, we get $0 \geq \langle x_1 - y_1 + y_2 - x_2, x_1 - x_2\rangle$ which corresponds to

$$\|x_1 - x_2\|_2^2 \leq \langle y_1 - y_2, x_1 - x_2\rangle \tag{3}$$

and which, by Cauchy-Schwarz implies $\|x_1 - x_2\|_2 \leq \|y_1 - y_2\|_2$ as desired.

Example Let $f(x) = |x|$ defined on $\mathbb{R}$. Then one can verify (exercise!) that for any $t > 0$,

$$\mathbf{prox}_{tf}(y) = \operatorname*{argmin}_{x\in\mathbb{R}} \left\{|x| + 1/(2t)(x-y)^2\right\} = S_t(y) := \begin{cases} y+t & \text{if } y \leq -t \\ 0 & \text{if } |y| < t \\ y-t & \text{if } y \geq t. \end{cases} \tag{4}$$

This function is known as *soft-thresholding*. See Figure 1.

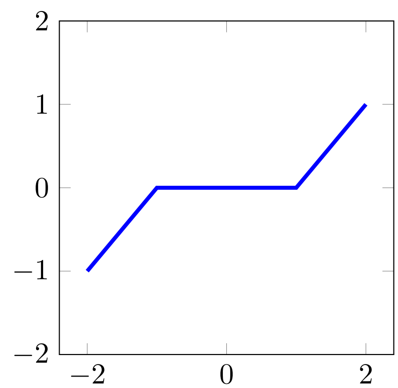

Figure 1: The soft-thresholding function (4) for $t = 1$.

Observe that if $f(x) = \sum_{i=1}^n f_i(x_i)$, then the **prox** of f decomposes:

$$(\mathbf{prox}_f(y))_i = \mathbf{prox}_{f_i}(y_i).$$

This implies for example that the prox operator of the ℓ_1 norm function is a componentwise soft-thresholding:

$$\mathbf{prox}_{t\|\cdot\|_1}(y) = [S_t(y_i)]_{1\leq i\leq n}$$

EXERCISE: Compute the proximal operators for the following functions: (i) $f(x) = (1/2)x^TAx$ where A is symmetric positive definite; (ii) $f(x) = -\sum_{i=1}^n \log x_i$ for $x \in \mathbb{R}^n_{++}$.

Proximal gradient methods We onsider a general class of optimization problems where the objective function $F(x)$ "splits" into two parts $F(x) = f(x) + h(x)$ where $f(x)$ is convex, smooth and L-Lipschitz, and $h(x)$ is convex nonsmooth but "simple" (in a way that will be clear later). So we want to solve

$$\min_{x\in\mathbb{R}^n} F(x) = f(x) + h(x). \tag{5}$$

Examples:

- Clearly if $h = I_C$ is the indicator function of a convex set C then problem (5) is equivalent to minimizing $f(x)$ on C.
- Optimization problems of the form (5) are very common in statistics where $f(x)$ is a "data fidelity" term (e.g., $f(x) = \|Ax - b\|_2^2$ for a linear model with a squared loss) and $h(x)$ is a "regularization" term (e.g., $h(x) = \|x\|_1$ to promote sparsity).

The proximal gradient method to solve (5) proceeds as follows. Starting from any $x_0 \in \mathbb{R}^n$, iterate:

$$x_{k+1} = \mathbf{prox}_{t_k h}\left(x_k - t_k\nabla f(x_k)\right) \tag{6}$$

where $t_k > 0$ are the step sizes.
Remarks:

- When h is the indicator function of convex set C, then iterates (6) correspond to projected gradient descent.

- If x^* is a fixed point of (6), i.e., $x^* = \mathbf{prox}_{th}(x^* - t\nabla f(x^*))$, then this means by (2) that $x^* - t\nabla f(x^*) - x^* \in t\partial h(x^*)$, i.e., $0 \in \partial(f+h)(x^*)$ which implies that x^* is a minimizer of $F(x) = f(x) + h(x)$, as desired.

- From (2) we know that $x_{k+1} = \mathbf{prox}_{t_k h}(x_k - t_k \nabla f(x_k))$ should satisfy

$$x_{k+1} = x_k - t_k \nabla f(x_k) - t_k h'(x_{k+1}) \tag{7}$$

  for some $h'(x_{k+1}) \in \partial h(x_{k+1})$. The main difference with a standard (sub)gradient method applied to $f+h$ is that we have $h'(x_{k+1})$ on the right-hand side, and not $h'(x_k)$. [cf. backward Euler vs. forward Euler for the discretization of ODEs. In fact, the proximal gradient method is also known as the forward-backward method.]

- Using the definition of **prox**, we see that the iterate (6) can be written as

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ h(u) + \frac{1}{2t_k}\|x_k - t_k\nabla f(x_k) - u\|_2^2 \right\} \\ &= \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ f(x_k) + \langle \nabla f(x_k), u\rangle + h(u) + \frac{1}{2t_k}\|u - x_k\|_2^2 \right\} \end{aligned}$$

  The term $f(x_k) + \langle \nabla f(x_k), u\rangle + h(u)$ is a local approximation of the cost function $f+h$ around x_k. The term $\frac{1}{2t_k}\|u - x_k\|_2^2$ ensures that we only trust this approximation close to x_k.

The convergence proof of the proximal gradient method is very similar to gradient method. We consider the two cases where f is m-strongly convex and L-smooth, and the case where f is simply L-smooth.

• f strongly convex. We assume here that f is twice differentiable, and that $mI \preceq \nabla^2 f(x) \preceq LI$. We have, using the fact that x^* is a fixed point of the iteration map (see second remark above)

$$\begin{aligned} \|x^+ - x^*\|_2 &= \|\,\mathbf{prox}_{th}(x - t\nabla f(x)) - \mathbf{prox}_{th}(x^* - t\nabla f(x^*))\|_2 \\ &\le \|x - x^* - t(\nabla f(x) - \nabla f(x^*))\|_2 \end{aligned}$$

where in the second line we used the fact that the proximal operator is nonexpansive. Now we have

$$\nabla f(x) - \nabla f(x^*) = \nabla f(x^*) + \int_0^1 \nabla^2 f(x^* + \alpha(x - x^*))(x - x^*)d\alpha = M(x - x^*)$$

where $M = \int_0^1 \nabla^2 f(x^* + \alpha(x - x^*))d\alpha$ is a symmetric matrix whose eigenvalues all lie in $[m, L]$. Thus we get $\|x^+ - x^*\|_2 \le \|(I - tM)(x - x^*)\|_2 \le \|I - tM\|\|x - x^*\|_2$ where $\|I - tM\|$ is the operator norm of $I - tM$. When $t = 2/(m+L)$ we have already seen in Lecture 3 that $\|I - tM\| \le (L-m)/(L+m)$. This shows that $\|x_k - x^*\|_2 \le \left(\frac{L-m}{L+m}\right)^k \|x_0 - x^*\|_2$.

• We now sketch the proof, in the case where f is just L-smooth.

Theorem 10.1. *Let $F = f + h$, and assume $f : \mathbb{R}^n \to \mathbb{R}$ is convex L-smooth (i.e., ∇f is L-Lipschitz) and h is convex. For constant step size $t_k = t \in (0, 1/L]$ the iterations of (6) satisfy $F(x_k) - F^* \le \frac{1}{2kt}\|x_0 - x^*\|_2^2$.*

Proof. We start in the same way as the standard gradient method

$$f(x^+) \le f(x) + \left\langle \nabla f(x), x^+ - x \right\rangle + \frac{L}{2}\|x^+ - x\|_2^2.$$

From (7) we know that we can write $x^+ = x - t\nabla f(x) - th'(x^+)$ where $h'(x^+) \in \partial h(x^+)$. Thus plugging $\nabla f(x) = \frac{1}{t}(x - x^+) - h'(x^+)$ we get

$$\begin{aligned} f(x^+) &\le f(x) - \frac{1}{t}\|x - x^+\|_2^2 + \left\langle h'(x^+), x - x^+ \right\rangle + \frac{L}{2}\|x^+ - x\|_2^2 \\ &\le f(x) - \frac{1}{t}\|x - x^+\|_2^2(1 - Lt/2) + \left\langle h'(x^+), x - x^+ \right\rangle \\ &= f(x) - \frac{1}{2t}\|x - x^+\|_2^2 + \left\langle h'(x^+), x - x^+ \right\rangle \end{aligned}$$

where in the last line we used $t = 1/L$. Now we substract $f(x^*)$ from each side to get

$$\begin{aligned} f(x^+) - f(x^*) &\le f(x) - f(x^*) - \frac{1}{2t}\|x - x^+\|_2^2 + \left\langle h'(x^+), x - x^+ \right\rangle \\ &\le \langle \nabla f(x), x - x^* \rangle - \frac{1}{2t}\|x - x^+\|_2^2 + \left\langle h'(x^+), x - x^+ \right\rangle \\ &= \left\langle \frac{x - x^+}{t} - h'(x^+), x - x^* \right\rangle - \frac{1}{2t}\|x - x^+\|_2^2 + \left\langle h'(x^+), x - x^+ \right\rangle \\ &\overset{(a)}{=} \frac{1}{2t}[\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2] + \left\langle h'(x^+), x^* - x^+ \right\rangle \\ &\overset{(b)}{\le} \frac{1}{2t}[\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2] + h(x^*) - h(x^+) \end{aligned}$$

where in (a) we used completion of squares, and in (b) we used convexity of h. The last inequality tells us that

$$F(x^+) - F(x^*) \le \frac{1}{2t}[\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2].$$

The rest of the proof is straightforward. □

Fast proximal gradient method There is a fast version of the proximal gradient method that converges in $O(1/k^2)$. The algorithm takes the form:

$$\begin{cases} y = x_k + \beta_k(x_k - x_{k-1}) \\ x_{k+1} = \mathbf{prox}_{t_k h}\left(y - t_k\nabla f(y)\right). \end{cases} \tag{8}$$

One can adapt the proof of the fast gradient method to show that (8) (with e.g., $\beta_k = (k-1)/(k+2)$) has a convergence rate of $O(1/k^2)$.

Regression with ℓ_1 regularization (Lasso, compressed sensing, ...) Consider the problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda\|x\|_1. \tag{9}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The $\|x\|_1$ term in the objective promotes sparsity in the solution x^*. Problem (9) fits (5) with $f(x) = \|Ax - b\|_2^2$ and $h(x) = \lambda\|x\|_1$. We saw that the proximal operator of h is the soft-thresholding operator. The proximal gradient method applied to (9) is called the *iterative shrinkage thresholding algorithm (ISTA)* and takes the form

$$x_{k+1} = S_{\lambda t}(x_k - 2tA^T(Ax_k - b))$$

where $S_{\lambda t}$ is the soft-thresholding operator (4) with parameter λt. The fast version is known as FISTA [BT09].

References

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009. 4

[PB14] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.