10 Proximal methods

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

\[
\text{prox}_f(y) = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2}\|x - y\|^2 \right\}.
\]

Clearly the proximal operator of the indicator function \( I_C \) to projection operator.

The next proposition guarantees that \( \text{prox}_f \) 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 \( f : \mathbb{R}^n \to \mathbb{R} \). Prove that the following are equivalent: (i) \( f \) is lower-semicontinuous, (ii) \( \text{epi}(f) \) is closed, (iii) all the sublevel sets \( f^{-1}((-\infty, a]) \) are closed.

Proposition 10.1. If \( f \) is lower-semicontinuous, then \( \text{prox}_f(y) \) 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 \( g(x) \leq t \iff \langle a, x \rangle + b + (1/2)\|x - y\|^2_2 \leq t \iff \|x - (y - a)\|^2_2 \leq C \) 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

\[
x = \text{prox}_f(y) \iff 0 \in \partial f(x) + (x - y) \iff y \in x + \partial f(x).
\]

Remark 1. If \( f \) is smooth, we see that \( x = \text{prox}_f(y) \) 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

\[
\|\text{prox}_f(y_1) - \text{prox}_f(y_2)\|_2 \leq \|y_1 - y_2\|_2.
\]

To see why, let \( x_1 = \text{prox}_f(y_1) \) and \( x_2 = \text{prox}_f(y_2) \). 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

\[
\|x_1 - x_2\|^2_2 \leq \langle y_1 - y_2, x_1 - x_2 \rangle
\]

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 $t > 0$,

$$\text{prox}_{tf}(y) = \arg\min_{x \in \mathbb{R}} \{ |x| + 1/(2t)(x - y)^2 \} = S_t(y) := \begin{cases} 
  y + t & \text{if } y \leq -t \\
  0 & \text{if } |y| < t \\
  y - t & \text{if } y \geq t.
\end{cases} \quad (4)$$

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

![Figure 1](image.png)

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 $\text{prox}$ of $f$ decomposes:

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

This implies for example that the prox operator of the $\ell_1$ norm function is a componentwise soft-thresholding:

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

**EXERCISE:** Compute the proximal operators for the following functions: (i) $f(x) = (1/2)x^T Ax$ 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 consider 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

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

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:

$$x_{k+1} = \text{prox}_{t_k h}(x_k - t_k \nabla f(x_k)) \quad (6)$$
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., \( x^* = \text{prox}_{th}(x^* - t\nabla f(x^*)) \), then this means by (2) that \( x^* - t\nabla f(x^*) - x^* \in th(x^*) \), 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 \( x_{k+1} = \text{prox}_{tkh}(x_k - tk\nabla f(x_k)) \) should satisfy

\[
x_{k+1} = x_k - tk\nabla f(x_k) - tkh'(x_{k+1})
\]

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 \( \text{prox} \), we see that the iterate (6) can be written as

\[
x_{k+1} = \arg \min_{u \in \mathbb{R}^n} \left\{ h(u) + \frac{1}{2lt_k} \| x_k - tk\nabla f(x_k) - u \|_2^2 \right\} = \arg \min_{u \in \mathbb{R}^n} \left\{ f(x_k) + \langle \nabla f(x_k), u \rangle + h(u) + \frac{1}{2lt_k} \| u - x_k \|_2^2 \right\}
\]

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 \( \frac{1}{2lt_k} \| u - x_k \|_2^2 \) 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 \( mL \leq \nabla^2 f(x) \leq LI \). We have, using the fact that \( x^* \) is a fixed point of the iteration map (see second remark above)

\[
\| x^+ - x^* \|_2 = \| \text{prox}_{th}(x - t\nabla f(x)) - \text{prox}_{th}(x^* - t\nabla f(x^*)) \|_2 \\
\leq \| x - x^* - t(\nabla f(x) - \nabla f(x^*)) \|_2
\]

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 \( \| x^+ - x^* \|_2 \leq \|(I - tM)(x - x^*)\|_2 \leq \|I - tM\| \|x - x^*\|_2 \) where \( \|I - tM\| \) is the operator norm of \( I - tM \). When \( t = 2/(m + L) \) we have already seen in Lecture 3 that \( \|I - tM\| \leq (L - m)/(L + m) \).

This shows that \( \| x_k - x^* \|_2 \leq \left( \frac{L - m}{L + m} \right)^k \| x_0 - x^* \|_2 \).

- 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., \( \nabla f \) is \( L \)-Lipschitz) and \( h \) is convex. For constant step size \( t_k = t \in (0, 1/L] \) the iterations of (6) satisfy \( F(x_k) - F^* \leq \frac{1}{2tk} \| x_0 - x^* \|_2^2 \).
Proof. We start in the same way as the standard gradient method
\[
f(x^+) \leq f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{L}{2} \|x^+ - x\|^2_2.
\]
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
\[
f(x^+) \leq f(x) - \frac{1}{t} \|x - x^+\|^2_2 + \langle h'(x^+), x - x^+ \rangle + \frac{L}{2} \|x^+ - x\|^2_2
\]
\[
\leq f(x) - \frac{1}{t} \|x - x^+\|^2_2 (1 - Lt/2) + \langle h'(x^+), x - x^+ \rangle
\]
\[
= f(x) - \frac{1}{2t} \|x - x^+\|^2_2 + \langle h'(x^+), x - x^+ \rangle
\]
where in the last line we used \(t = 1/L\). Now we subtract \(f(x^+)\) from each side to get
\[
f(x^+) - f(x^*) \leq f(x) - f(x^*) - \frac{1}{2t} \|x - x^+\|^2_2 + \langle h'(x^+), x - x^+ \rangle
\]
\[
\leq \langle \nabla f(x), x - x^* \rangle - \frac{1}{2t} \|x - x^+\|^2_2 + \langle h'(x^+), x - x^+ \rangle
\]
\[
= \left( \frac{x - x^+}{t} - h'(x^+), x - x^* \right) - \frac{1}{2t} \|x - x^+\|^2_2 + \langle h'(x^+), x - x^+ \rangle
\]
\[
= \frac{1}{2t} \left( \|x - x^*\|^2_2 - \|x^+ - x^*\|^2_2 \right) + \langle h'(x^+), x - x^+ \rangle
\]
\[
\leq \frac{1}{2t} \left( \|x - x^*\|^2_2 - \|x^+ - x^*\|^2_2 \right) + h(x^*) - h(x^+)
\]
where in \(a\) we used completion of squares, and in \(b\) we used convexity of \(h\). The last inequality tells us that
\[
F(x^+) - F(x^*) \leq \frac{1}{2t} \left( \|x - x^*\|^2_2 - \|x^+ - x^*\|^2_2 \right).
\]
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:
\[
\begin{align*}
y_k &= x_k + \beta_k (x_k - x_{k-1}) \\
x_{k+1} &= \text{prox}_{t_k h}(y - t_k \nabla f(y)).
\end{align*}
\]
(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 \(\ell_1\) regularization (Lasso, compressed sensing, …) Consider the problem
\[
\min_{x \in \mathbb{R}^n} \|Ax - b\|^2_2 + \lambda \|x\|_1.
\]
(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
\[
x_{k+1} = S_M(x_k - 2tA^T(Ax_k - b))
\]
where \(S_M\) is the soft-thresholding operator (4) with parameter \(\lambda t\). The fast version is known as FISTA [BT09].

4
References
