10 Proximal 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

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + h(x). \tag{1}$$

Examples:

- Clearly if $h = I_C$ is the indicator function of a convex set C then problem (1) is equivalent to minimizing $f(x)$ on C.
- Optimization problems of the form (1) 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).

Proximal gradient method The proximal gradient method to solve (1) proceeds as follows. Starting from any $x_0 \in \mathbb{R}^n$, iterate:

$$x_{k+1} = \mathbf{prox}_{t_k h}\left(x_k - t_k \nabla f(x_k)\right) \tag{2}$$

where $t_k > 0$ are the step sizes. Recall that

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

and that

$$x = \mathbf{prox}_h(y) \iff 0 \in \partial h(x) + (x - y). \tag{3}$$

Remarks:

- When h is the indicator function of convex set C, then iterates (2) correspond to projected gradient descent.
- If x^* is a fixed point of (2), i.e., $x^* = \mathbf{prox}_{th}(x^* - t\nabla f(x^*))$, then this means by (3) that $x^* - t\nabla f(x^*) - x^* \in t\partial h(x^*)$, i.e., $0 \in \nabla f(x^*) + \partial h(x^*)$. Assuming $\textbf{int dom}\, f \cap \textbf{int dom}\, h \neq \emptyset$, this is equivalent to $0 \in \partial(f + h)(x^*)$ which implies that x^* is a minimizer of $F(x) = f(x) + h(x)$, as desired.
- From (3) we know that $x_{k+1} = \mathbf{prox}_{t_k h}(x_k - t_k \nabla f(x_k))$ should satisfy

$$x_{k+1} = x_k - t_k \nabla f(x_k) - t_k h'(x_{k+1}) \tag{4}$$

  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 (2) can be written as

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

  The term $f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + h(x)$ is a local approximation of the cost function $f+h$ around x_k. The term $\frac{1}{2t_k}\|x-x_k\|_2^2$ 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.

- When f is strongly convex, we can prove the following.

Theorem 10.1. *Let $F = f+h$ and assume $f:\mathbb{R}^n\to\mathbb{R}$ is m-strongly convex and L-smooth, and h is convex. For constant step size $t_k = 2/(m+L)$ the iterations of (2) $\|x_k - x^*\|_2 \leq (\frac{L-m}{L+m})^k\|x_0-x^*\|_2$.*

Proof. 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)

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

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

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

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 10.2. *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 (2) satisfy $F(x_k) - F^* \leq \frac{1}{2kt}\|x_0 - x^*\|_2^2$.*

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

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

From (4) 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

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

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:

$$\begin{cases} y = x_k + \beta_k(x_k - x_{k-1}) \\ x_{k+1} = \mathbf{prox}_{t_k h}\left(y - t_k \nabla f(y)\right). \end{cases} \tag{5}$$

One can adapt the proof of the fast gradient method to show that (5) (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

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda \|x\|_1. \tag{6}$$

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 (6) fits (1) 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 (6) 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 as seen in Lecture 9, 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. 3

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