6 Proximal gradient methods

Motivation: constrained optimization Consider the problem of minimizing a convex function $f(x)$ on a convex set C. To do this the projected gradient descent iterates are as follows: starting from any $x_0 \in C$ proceed

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

where $P_C(x) = \operatorname{argmin}\{\|x-y\|_2 : y \in C\}$ is the Euclidean projection on C. One can adapt the convergence proof of the gradient method to show that (1) converges to $\min_{x\in C} f(x)$ at the rate $O(1/k)$. See Exercise sheet 1.

Optimization problems with a splitting structure In this lecture we consider a general class of optimization problems where the objective function $f(x)$ "splits" into two parts $f(x) = g(x) + h(x)$ where $g(x)$ 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

$$\min_{x\in\mathbb{R}^n} \quad f(x) = g(x) + h(x). \tag{2}$$

Examples:

- If h is the indicator function of a convex set C defined as
$$h(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{else} \end{cases}$$
then problem (2) is equivalent to minimizing $g(x)$ on C.
- Optimization problems of the form (2) are very common in statistics where $g(x)$ is a "data fidelity" term (e.g., $g(x) = \|Ax - b\|_2^2$ for a linear model with a squared loss) and $h(x)$ is a "regularization" term (e.g., $g(x) = \|x\|_1$ to promote sparsity).

The proximal mapping Given a convex function $h : D \to \mathbb{R}$ define the *proximal operator* associated to h by

$$\operatorname{prox}_h(x) = \operatorname*{argmin}_{u\in D} \left\{ h(u) + \frac{1}{2}\|u - x\|_2^2 \right\}.$$

The map prox_h is well defined because the function $u \mapsto h(x) + \frac{1}{2}\|u-x\|_2^2$ (for fixed x) is *strongly convex*, and thus has a unique minimum.

Remark 1. *When h is the indicator function of convex set C, then $\operatorname{prox}_h(x)$ is the Euclidean projection of x on C.*

Computing prox_h is itself a convex optimization problem. However for some "simple" functions h one can compute $\operatorname{prox}_h(x)$ analytically. (See below for some examples.) We will need the following proposition concerning $\operatorname{prox}_h(x)$:

Proposition 6.1. *We have $u = \operatorname{prox}_h(x)$ iff $x - u \in \partial h(u)$, where $\partial h(u)$ is the subdifferential of f at u.*

Proof. One can verify that u is a minimizer of a convex function F iff $0 \in \partial F(u)$. Also one can check that $\partial(F_1+F_2)(u) = \partial F_1(u) + \partial F_2(u)$ where the latter is the Minkowski addition of sets (i.e., $A + B = \{a + b : a \in A, b \in B\}$).

Applying these two facts we get: $u = \text{prox}_h(x)$ iff the zero vector is in the subdifferential of $h + \frac{1}{2}\|\cdot - x\|_2^2$. The second term is smooth and its gradient at a point u is $u - x$. Thus $u = \text{prox}_h(x)$ iff $0 \in \partial h(u) + (u - x)$ i.e., $x - u \in \partial h(u)$. □

Proximal gradient method The proximal gradient method to solve (2) proceeds as follows. Starting from any $x_0 \in \mathbb{R}^n$, iterate:

$$x_{k+1} = \text{prox}_{t_k h}\left(x_k - t_k \nabla g(x_k)\right) \tag{3}$$

where $t_k > 0$ are the step sizes. Unrolling the definition of prox this means

$$\begin{aligned} x_{k+1} &= \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ h(u) + \frac{1}{2t_k}\|x_k - t_k \nabla g(x_k) - u\|_2^2 \right\} \\ &= \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ g(x_k) + \nabla g(x_k)^T u + \frac{1}{2t_k}\|u - x_k\|_2^2 + h(u) \right\} \end{aligned}$$

The term $g(x_k) + \nabla g(x_k)^T u + \frac{1}{2t_k}\|u - x_k\|_2^2$ is a quadratic model for $g(x)$ centered at $x = x_k$. Note that when h is the indicator function of convex set C, then iterates (3) correspond to projected gradient descent (1).

Convergence proof of proximal gradient method is very similar to gradient method. We sketch the proof now.

Theorem 6.1. *Assume $g : \mathbb{R}^n \to \mathbb{R}$ is convex L-smooth (i.e., ∇g is L-Lipschitz) and h is convex. For constant step size $t_k = t \in (0, 1/L]$ the iterations of (3) satisfy $f(x_k) - f^* \leq \frac{1}{2kt}\|x_0 - x^*\|_2^2$.*

Proof. For any x, let $\tilde{x} = x - t\nabla g(x)$ and $x^+ = \text{prox}_{th}(\tilde{x})$. Using L-smoothness of g and $t \in (0, 1/L]$ we have (same as with the gradient method)

$$g(x^+) \leq g(x) + \nabla g(x)^T(x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2.$$

Now we use that $\nabla g(x) = -\frac{1}{t}(\tilde{x} - x) = -\frac{1}{t}(\tilde{x} - x^+ + x^+ - x)$, and $0 < t \leq 1/L$ to get

$$g(x^+) \leq g(x) - \frac{1}{2t}\|x^+ - x\|_2^2 + \frac{1}{t}(\tilde{x} - x^+)^T(x - x^+). \tag{4}$$

For any fixed z, convexity of g tells us that $g(x) \leq g(z) + \nabla g(x)^T(x - z)$. Thus continuing from (4) we get

$$\begin{aligned} g(x^+) - g(z) &\leq \nabla g(x)^T(x - z) - \frac{1}{2t}\|x^+ - x\|_2^2 + \frac{1}{t}(\tilde{x} - x^+)^T(x - x^+) \\ &\overset{(a)}{=} -\frac{1}{t}(\tilde{x} - x^+ + x^+ - x)^T(x - z) - \frac{1}{2t}\|x^+ - x\|_2^2 + \frac{1}{t}(\tilde{x} - x^+)^T(x - x^+) \\ &= -\frac{1}{2t}\left[\|x^+ - x\|_2^2 + 2(x^+ - x)^T(x - z)\right] + \frac{1}{t}(\tilde{x} - x^+)^T(z - x^+) \\ &\overset{(b)}{=} -\frac{1}{2t}\left[\|x^+ - z\|_2^2 - \|x - z\|_2^2\right] + \frac{1}{t}(\tilde{x} - x^+)^T(z - x^+) \end{aligned} \tag{5}$$

where in (a) we used the fact that $\nabla g(x) = -\frac{1}{t}(\tilde{x} - x) = -\frac{1}{t}(\tilde{x} - x^+ + x^+ - x)$ and in (b) we used completion of squares. Since $x^+ = \text{prox}_{th}(\tilde{x})$ we know from Proposition 6.1 that $\tilde{x} - x^+ \in t\partial h(x^+)$, i.e., $\frac{1}{t}(\tilde{x}-x^+) \in \partial h(x^+)$. It thus follows, by convexity of h, that $h(z) \geq h(x^+)+\frac{1}{t}(\tilde{x}-x^+)^T(z-x^+)$. Adding $h(x^+) - h(z)$ to each side of the inequality in (5) and using the last inequality gives us

$$f(x^+) - f(z) \leq -\frac{1}{2t}\left[\|x^+ - z\|_2^2 - \|x - z\|_2^2\right]. \tag{6}$$

Note that if we set $z = x$, inequality (6) tells us that the value of f decreases at each step, i.e., $f(x^+) < f(x)$. To finish the proof we set $z = x^*$ in (6) and use a telescoping sum (see end of proof of convergence of gradient method). □

Fast proximal gradient method There is a fast version of the proximal gradient method that converges in $O(1/k^2)$. The algorithm is very similar to what we saw in last lecture; the only difference is the proximal operator:

$$\begin{cases} y = x_k + \beta_k(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{t_k h}\left(y - t_k \nabla g(y)\right). \end{cases} \tag{7}$$

One can adapt the proof of the fast gradient method to show that (7) (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

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda\|x\|_1. \tag{8}$$

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 (8) fits (2) with $g(x) = \|Ax - b\|_2^2$ and $h(x) = \lambda\|x\|_1$. The proximal operator of $\|x\|_1$ has a closed-form expression, as follows (exercise!):

$$(\operatorname{prox}_{t\|\cdot\|_1}(x))_i = \begin{cases} x_i - t & \text{if } x_i \geq t \\ 0 & \text{if } x_i \in [-t, t] \\ x_i + t & \text{if } x_i \leq -t. \end{cases} \tag{9}$$

It is known as the *soft-thresholding* (or also *shrinkage thresholding*) operator. The proximal gradient method applied to (8) 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 (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.