5.4 Projected gradient methods

Both, the active set method and the interior point require the solution of a linear system in every step, which may become too costly for large-scale problems. Moreover, the active set method has the additional disadvantage for a large number of variables that at most one index can be inserted or removed in $\mathcal{W}_k$. Many applications, especially in statistical/machine learning, do not require high accuracy and a first-order method would be sufficient. The basic idea of *projected gradient methods* is to perform a gradient step and then project it to satisfy the constraints. To carry out the projection effectively requires the constraints to be sufficiently simple. It turns out that projected gradient methods are best understood and analyzed via proximal point mappings.[8]

5.4.1 Subgradients and proximal point mappings

Consider an *extended* function $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and let $\operatorname{dom} p = \{\mathbf{x} \in \mathbb{R}^n : p(x) < \infty\}$ denote the domain of p. We call the extended function p convex if $\operatorname{dom} p$ is a convex set and p is convex on $\operatorname{dom} p$ in the sense of Definition 4.18.

The role of p will be to describe a convex constraint set (see Example 5.19 below) and, as such, it will not be differentiable. The notion of subgradients is the most common way to extend the concept of gradients to non-differentiable convex functions.

Definition 5.16 *Let $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. A* subgradient *of p at $\mathbf{x}_0 \in \operatorname{dom} p$ is any vector $g \in \mathbb{R}^n$ such that*

$$p(\mathbf{x}) \geq p(\mathbf{x}_0) + \langle g, \mathbf{x} - \mathbf{x}_0 \rangle, \qquad \forall \mathbf{x} \in \operatorname{dom} p. \tag{5.43}$$

The subdifferential *$\partial p(\mathbf{x}_0)$ is the set of all subgradients of f at $\mathbf{x}_0$.*

Subgradients have a number of important and interesting properties; see Section 3.1 of [N]. In the following, we limit ourselves to the properties that are needed for the subsequent developments.

Theorem 5.17 *Let $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. $\mathbf{x}^\star \in \mathbb{R}^n$ is a global minimum of p if and only if $0 \in \partial p(\mathbf{x}^\star)$.*

Proof. This immediately follows from the definition: $0 \in \partial p(\mathbf{x}^\star)$ is equivalent to

$$p(\mathbf{x}) \geq p(\mathbf{x}^*) + \langle 0, \mathbf{x} - \mathbf{x}^* \rangle = p(\mathbf{x}^*)$$

for all $\mathbf{x} \in \operatorname{dom} p$. □

[8] Part of the material in this section is based on Recht's lecture notes available from `http://pages.cs.wisc.edu/~brecht/cs726docs/ProjectedGradientMethods.pdf`.

Lemma 5.18 *For* $\mathbf{x}, \mathbf{y} \in \operatorname{dom} p$ *let* $\mathbf{g_x} \in \partial p(\mathbf{x}), \mathbf{g_y} \in \partial p(\mathbf{y})$*. Then* $\langle \mathbf{g_x} - \mathbf{g_y}, \mathbf{x} - \mathbf{y} \rangle \geq 0$.

Proof. By definition,

$$\langle g_{\mathbf{x}}, \mathbf{y} - \mathbf{x} \rangle \leq p(\mathbf{y}) - p(\mathbf{x}), \quad \langle g_{\mathbf{y}}, \mathbf{x} - \mathbf{y} \rangle \leq p(\mathbf{x}) - p(\mathbf{y}).$$

Adding both inequalities yields $-\langle \mathbf{g_x} - \mathbf{g_y}, \mathbf{x} - \mathbf{y} \rangle \leq 0$. □

Given a convex function $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, we define the *proximity operator* or the *proximity point mapping* as

$$\operatorname{prox}_p(\mathbf{x}) := \arg\min_{\mathbf{y}} \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + p(\mathbf{y}). \tag{5.44}$$

Note that the strong convexity of the target function implies that the minimizer is uniquely defined. By Theorem 5.17, $\operatorname{prox}_p(\mathbf{x})$ is the unique vector satisfying

$$\mathbf{x} - \operatorname{prox}_p(\mathbf{x}) \in \partial p(\operatorname{prox}_p(\mathbf{x})).$$

Example 5.19 Let C be a convex set. Then the indicator function

$$i_C(\mathbf{x}) := \begin{cases} 0 & \mathbf{x} \in C, \\ \infty & \text{otherwise}, \end{cases}$$

is convex. By its definition (5.44), $\operatorname{prox}_{i_C}(\mathbf{x})$ is the point in C closest to $\mathbf{x}$.

For $p(\mathbf{x}) = \mu \|\mathbf{x}\|_1$, we have

$$\big(\operatorname{prox}_{i_C}(\mathbf{x})\big)_i = \begin{cases} x_i + \mu & \text{if } x_i < -\mu, \\ x_i - \mu & \text{if } x_i > \mu, \\ 0 & \text{otherwise.} \end{cases}$$

◇

Lemma 5.20 *Let* $q(\mathbf{x}) := \mathbf{x} - \operatorname{prox}_p(\mathbf{x})$*. Then the following relations hold:*

1. $q(\mathbf{x}) \in \partial p(\operatorname{prox}_p(\mathbf{x}))$*;*
2. $\langle \operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y}), q(\mathbf{x}) - q(\mathbf{y}) \rangle \geq 0$*;*
3. $\| \operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y})\|_2^2 + \|q(\mathbf{x}) - q(\mathbf{y})\|_2^2 \leq \|\mathbf{x} - \mathbf{y}\|_2^2$*;*
4. $\| \operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y})\|_2 = \|\mathbf{x} - \mathbf{y}\|_2$ *if and only if* $\operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y}) = \mathbf{x} - \mathbf{y}$.

Proof. 1. follows by definition. 2. follows from 1. combined with Lemma 5.18. 3. follows from

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|_2^2 &= \|(\operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y})) + (q(\mathbf{x}) - q(\mathbf{y}))\|_2^2 \\ &= \| \operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y})\|_2^2 + 2\langle \operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y}), q(\mathbf{x}) - q(\mathbf{y}) \rangle \\ &\quad + \|q(\mathbf{x}) - q(\mathbf{y})\|_2^2 \\ &\geq \| \operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y})\|_2^2 + \|q(\mathbf{x}) - q(\mathbf{y})\|_2^2. \end{aligned}$$

where we used 2. in the inequality. 4. is a direct consequence of 3. □

Lemma 5.20.3 implies that the proximity operator is *nonexpansive*:

$$\| \operatorname{prox}_p(\mathbf{x}) - \operatorname{prox}_p(\mathbf{y})\|_2^2 \le \|\mathbf{x} - \mathbf{y}\|_2^2. \tag{5.45}$$

In other words, $\operatorname{prox}_p(\mathbf{x})$ is Lipschitz continuous constant 1.

Given $\mathbf{x}_0 \in \mathbb{R}^n$, the *proximal point algorithm* is defined by the iteration

$$\mathbf{x}_{k+1} = \operatorname{prox}_p(\mathbf{x}_k), \qquad k = 0, 1, \dots . \tag{5.46}$$

Lemma 5.21 *The proximal point algorithm* (5.46) *converges to a minimizer of* p.

Proof. The nonexpansive property (5.45) implies that the sequence $\{\mathbf{x}_k\}$ is bounded and therefore has one or several limit points. Let $\bar{\mathbf{x}}$ be such a limit point and let $\mathbf{x}^\star$ be a minimizer of p, that is, $0 \in \partial p(\mathbf{x}^\star)$. Again by (5.45), we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 = \| \operatorname{prox}_p(\mathbf{x}_k) - \operatorname{prox}_p(\mathbf{x}^\star)\|_2 \le \|\mathbf{x}_k - \mathbf{x}^\star\|_2.$$

Hence

$$\|\mathbf{x}_k - \mathbf{x}^\star\|_2 \overset{k\to\infty}{\longrightarrow} \|\bar{\mathbf{x}} - \mathbf{x}^\star\|_2. \tag{5.47}$$

By continuity, $\operatorname{prox}_p(\bar{\mathbf{x}})$ is also a limit point of $\mathbf{x}_k$ and, consequently,

$$\| \operatorname{prox}_p(\bar{\mathbf{x}}) - \operatorname{prox}_p(\mathbf{x}^\star)\|_2 = \| \operatorname{prox}_p(\bar{\mathbf{x}}) - \mathbf{x}^\star\|_2 = \|\bar{\mathbf{x}} - \mathbf{x}^\star\|_2.$$

By Lemma 5.20.4, this implies $\operatorname{prox}_p(\bar{\mathbf{x}}) - \operatorname{prox}_p(\mathbf{x}^\star) = \bar{\mathbf{x}} - \mathbf{x}^\star$. Hence $\operatorname{prox}_p(\bar{\mathbf{x}}) = \bar{\mathbf{x}}$ and $0 \in \partial p(\bar{\mathbf{x}})$. This allows us to replace $\mathbf{x}^\star$ by $\bar{\mathbf{x}}$ in (5.47), leading to $\|\mathbf{x}_k - \bar{\mathbf{x}}\|_2 \overset{k\to\infty}{\longrightarrow} 0$. □

5.4.2 The projected gradient scheme

We now consider a convex optimization problem that admits a decomposition of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + p(\mathbf{x}), \tag{5.48}$$

where f is a smooth convex function, while p is an extended convex function. The idea is that p contains all the difficulty in terms of nonsmoothness but its form is sufficiently simple that it is possible to apply the proximity operator conveniently. Important examples for p are those given in Example 5.19. In principle, one can extend the gradient method to use (normalized) subgradients of $f(\mathbf{x}) + p(\mathbf{x})$, but this method can converge quite slowly if p is nonsmooth. In such cases, the following *projected gradient scheme* can be much faster:

$$\mathbf{x}_{k+1} = \operatorname{prox}_{\alpha_k p}\big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\big), \qquad k = 0, 1, \dots, \tag{5.49}$$

where $\alpha_0, \alpha_1, \dots$ is a sequence of (suitably chosen) positive step sizes.

The rationale behind algorithm (5.49) follows from the following simple result.

Lemma 5.22 *Let f, p be convex and additionally assume that f is differentiable. Then $\mathbf{x}^\star$ is a minimizer for* (5.48) *if and only if*

$$\mathbf{x}^\star = \operatorname{prox}_{\alpha p}\left(\mathbf{x}^\star - \alpha \nabla f(\mathbf{x}^\star)\right)$$

holds for all $\alpha > 0$.

Proof. By Theorem 5.17, $\mathbf{x}^\star$ is a minimizer if and only if

$$-\nabla f(\mathbf{x}^\star) \in \partial p(\mathbf{x}^\star) \quad \Leftrightarrow \quad \mathbf{x}^\star - \alpha \nabla f(\mathbf{x}^\star) - \mathbf{x}^\star \in \alpha \partial p(\mathbf{x}^\star),$$

which is equivalent to $\mathbf{x}^\star = \operatorname{prox}_{\alpha p}\left(\mathbf{x}^\star - \alpha \nabla f(\mathbf{x}^\star)\right)$. □

In the convergence analysis of (5.49) we focus on the case of an L-smooth μ-strongly convex function f. Recall that this implies

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2. \tag{5.50}$$

Since $f + p$ is strongly convex, the minimizer $\mathbf{x}^*$ of (5.48) is uniquely determined. Using Lemma 5.20, we obtain

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 &= \left\| \operatorname{prox}_{\alpha_k p}\left(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\right) - \operatorname{prox}_{\alpha_k p}\left(\mathbf{x}^\star - \alpha_k \nabla f(\mathbf{x}^\star)\right)\right\|_2 \\ &\leq \|\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) - \mathbf{x}^\star + \alpha_k \nabla f(\mathbf{x}^\star)\|_2.\end{aligned}$$

Using (5.50) one can show that

$$\|\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) - \mathbf{x}^\star + \alpha_k \nabla f(\mathbf{x}^\star)\|_2 \leq \max\{|1 - \alpha_k \mu|, |1 - \alpha_k L|\}\|\mathbf{x}_k - \mathbf{x}^\star\|_2.$$

Choosing the factor $\alpha_k \equiv \frac{2}{\mu + L}$ minimizes the first factor (compare with Theorem 4.23!), which yields

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|_2$$

with $\kappa = L/\mu$.

5.4.3 Quadratic programs with box constraints

It is illustrative to see how simple the projected gradient scheme becomes when applied to a quadratic program with box constraints on $\mathbf{x}$:

$$\begin{array}{rl}\displaystyle\min_{\mathbf{x} \in \mathbb{R}^n} & f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T G \mathbf{x} + \mathbf{x}^T \mathbf{h} \\ \text{subject to} & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u},\end{array} \tag{5.51}$$

where $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$ are vectors containing lower and upper bounds on the entries of $\mathbf{x}$. We explicitly allow for entries $-\infty$ and ∞ in $\mathbf{l}$ and $\mathbf{u}$, respectively. The admissible set is clearly convex:

$$\Omega = \{\mathbf{x} : \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}\}.$$

Thus, (5.51) is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + i_\Omega(\mathbf{x}). \tag{5.52}$$

Note that the proximity operator is given by

$$\big(\operatorname{prox}_{i_\Omega}(\mathbf{y})\big)_i := \begin{cases} l_i & \text{if } y_i < l_i, \\ y_i & \text{if } y_i \in [l_i, u_i] \\ u_i & \text{if } y_i > u_i. \end{cases}$$

Since the gradient of f at a point $\mathbf{y}$ is given by $G\mathbf{y} + \mathbf{h}$, the projected gradient scheme takes the form

$$\mathbf{x}_{k+1} = \operatorname{prox}_{i_\Omega}\big(\mathbf{x}_k - \alpha_k(G\mathbf{x}_k + \mathbf{h})\big), \qquad k = 0, 1, \dots,$$

In the special case (5.51), a good step size α_k can be determined from the so called Cauchy point $\mathbf{x}^c$, which is the first local minimizer of the piecewise quadratic function

$$t \mapsto f(\mathbf{x}(\alpha)),$$

where $x(\alpha) := \operatorname{prox}_{i_\Omega}\big(\mathbf{x}_k - \alpha(G\mathbf{x}_k + \mathbf{h})\big)$. This computation is complicated by the fact that $\mathbf{x}(\alpha)$ has kinks and hence $f(\mathbf{x}(\alpha))$ is not differentiable. We therefore need to subsequently consider subintervals on which $\mathbf{x}(\alpha)$ is smooth. Once this is done, the Cauchy point $\mathbf{x}^c$ could be readily used as the next iterate. However, it turns out to be beneficial to further optimize the non-active components of $\mathbf{x}^c$ by approximately solving the corresponding QP with a few steps of an iterative method. The technical details can be found in Section 16.7 of [NW].