# 6

# Projected Gradient

The gradient method (and its variants) described in the previous chapter can only be applied to unconstrained optimization problems. While this is a scenario with many important applications, constrained optimization problems are far more common. A general treatment of constrained optimization requires tools like Lagrangian duality, that we will explore in future chapters, so for the moment we try to answer the question: can gradient descent be extended to deal with constrained optimization? In other words, can we apply it to problems of the form:

$$\begin{aligned} &\min f(x) \\ &\quad x \in C \end{aligned}$$

at least for some special sets $C$? It turns out that the answer is positive if the set $C$ is *simple*, in the sense that projecting an arbitrary point $x \in \mathbb{R}^n$ onto it can be done efficiently. Let us start with some preliminary concepts.

## 6.1 Normal Cones

**Definition 6.1.** *Given a convex set $C$ and a point $x \in C$, the* normal cone *at $x$, denoted as $N_C(x)$, is defined as:*

$$N_C(x) = \{d \in \mathbb{R}^n \mid d^\top (y - x) \leq 0 \; \forall y \in C\}$$

Intuitively, the normal cone at a point $x \in C$ can be understood as:

- the set of directions *perpendicular* to $C$ at $x$;
- the set of vectors defining supporting hyperplanes at $x$.

A graphical representation on a two-dimensional example is given in Figure 6.1.

Normal cones are a convenient tool to describe optimality conditions in the constrained case and when the function to optimize is convex *and* differentiable. Indeed, we have the following characterization:

**Theorem 6.1.** *Given the optimization problem:*

$$\begin{aligned} &\min f(x) \\ &\quad x \in C \end{aligned}$$

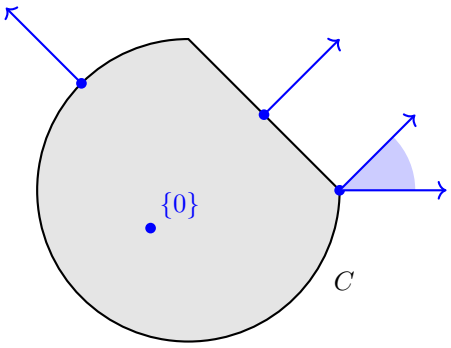


Figure 6.1: Normal cones.

*In other words, iff* $-\nabla f(x^*) \in N_C(x^*)$.

*where $f$ is convex and differentiable, and $C$ is convex, we have that $x^* \in C$ is optimal if and only if:*

$$\nabla f(x^*)^\top (y - x^*) \geq 0 \quad \forall y \in C$$

*Proof.* $\Leftarrow$) Follows trivially by convexity:

$$f(y) \geq f(x^*) + \underbrace{\nabla f(x^*)^\top (y - x^*)}_{\geq 0}$$
$$\geq f(x^*)$$

$\Rightarrow$) By contradiction, let $y \in C$ such that $\nabla f(x^*)^\top (y - x^*) < 0$. Consider the convex combination $z(t) = ty + (1-t)x^*$, for $t \in [0, 1]$: clearly, $z(t) \in C$. Now:

$$\frac{df(z(t))}{dt}\Big|_{t=0} = \nabla f(x^*)^\top (y - x^*) < 0$$

so for a small enough $t > 0$ we have $f(z(t)) < f(x^*)$, a contradiction since $x^*$ is an optimal solution.

□

Note that for $C = \mathbb{R}^n$ we recover the usual first-order condition $\nabla f(x^*) = 0$.

## 6.2 Projections

Intuitively, the projection of a point $y$ onto a set $C$ is the point in $C$ at minimal distance w.r.t. $y$. Formally, we get the following definition:

**Definition 6.2.** *Given a closed and convex set $C$, and a point $y \in \mathbb{R}^n$, we call* projection *of $y$ onto $C$ the (unique) optimal solution of the optimization problem:*

$$\min \frac{1}{2}\|y - x\|_2^2$$
$$x \in C$$

*and we denote it with $P_C(y)$.*

The projection of a point into a convex set satisfies a few properties.

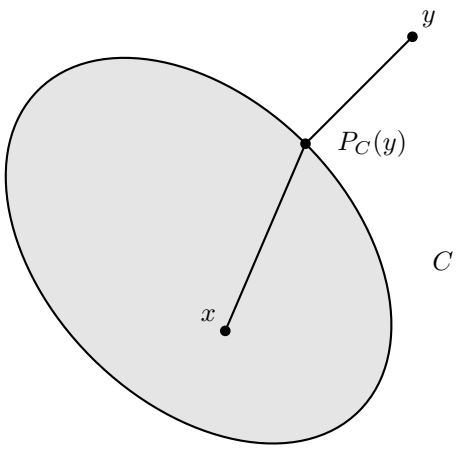


Figure 6.2: Minimum principle.

**Proposition 6.1** (P1)**.** *For any $x \in C$, we have:*

*This is actually a characterization: a point is a projection iff it satisfies this property, which goes under the name of* minimum principle.

$$(y - P_C(y))^\top (x - P_C(y)) \leq 0$$

*Proof.* This is a direct consequence of Theorem 6.1. Choosing $f(z) = \frac{1}{2}\|y-z\|_2^2$, we have $\nabla f(z) = -(y - z) = (z - y)$, so $P_C(y)$ is optimal iff:

$$(P_C(y) - y)^\top (x - P_C(y)) \geq 0 \quad \forall x \in C$$

which is exactly the condition we want to prove. □

A graphical representation is depicted in Figure 6.2.

**Proposition 6.2** (P2)**.** *Any projection is a* contraction*:*

$$\|P_C(a) - P_C(b)\| \leq \|a - b\| \quad \forall a, b \in \mathbb{R}^n$$

*Proof.* From $P1$ we have:

$$(a - P_C(a))^\top (x - P_C(a)) \leq 0 \quad \forall x \in C$$
$$(b - P_C(b))^\top (x - P_C(b)) \leq 0 \quad \forall x \in C$$

If we can choose $x = P_C(b)$ in the first inequality and $x = P_C(a)$ in the second, and add the two, we obtain:

$$[(b - a) + (P_C(a) - P_C(b))]^\top (P_C(a) - P_C(b)) \leq 0$$

We can distribute, apply the Cauchy-Schwartz inequality, and simplify to obtain our result:

$$\begin{aligned}\|P_C(a) - P_C(b)\|^2 &\leq (a - b)^\top (P_C(a) - P_C(b)) \\ &\leq \|a - b\| \cdot \|P_C(a) - P_C(b)\|\end{aligned}$$

□

## 6.3 Projected gradient

Remember that the standard gradient descent iteration is:

$$x^{k+1} = x^k - t_k \nabla f(x^k)$$

and we interpret this choice as minimizing the quadratic approximation:

$$x^{k+1} = \arg\min_{v \in \mathbb{R}^n} \{ f(x^k) + \nabla f(x^k)^\top (v - x^k) + \frac{1}{2t_k} \|v - x^k\|^2 \}$$

The intuition is to replace $v \in \mathbb{R}^n$ with $v \in C$ in the minimization above.

$$\begin{aligned}
x^{k+1} &= \arg\min_{v \in C} \left\{ f(x^k) + \nabla f(x^k)^\top (v - x^k) + \frac{1}{2t_k} \|v - x^k\|^2 \right\} \\
&= \arg\min_{v \in C} \{ \underbrace{2t_k f(x^k)}_{\text{replace by another constant}} + 2t_k \nabla f(x^k)^\top (v - x^k) + \|v - x^k\|^2 \} \\
&= \arg\min_{v \in C} \{ t_k^2 \|\nabla f(x^k)\|^2 + 2t_k \nabla f(x^k)^\top (v - x^k) + \|v - x^k\|^2 \} \\
&= \arg\min_{v \in C} \{ \|(v - x^k) + t_k \nabla f(x^k)\|^2 \} \\
&= \arg\min_{v \in C} \left\{ \frac{1}{2} \|v - (\underbrace{x^k - t_k \nabla f(x^k)}_{\text{usual } x^{k+1}})\|^2 \right\}
\end{aligned}$$

We thus obtain the rule:

$$x^{k+1} = P_C(x^k - t_k \nabla f(x^k))$$

where we first take the usual gradient step, but then project back onto $C$ before moving to the next iteration. Thus, for the whole method to be viable, we require the set $C$ to be simple enough that the projection operator can be implemented efficiently. Examples of sets for which we can compute the projection efficiently are:

*Usually, this means in $O(n)$.*

- the non-negative orthant $x \geq 0$: the projection can be computed easily as $x_j = \max\{x_j, 0\}$ for all $j = 1, \ldots, n$;
- the bounded box $l \leq x \leq u$: this is a generalization of the previous case, we just need to clip each coordinate to the corresponding bounds;
- the probability simplex $\sum_{j=1}^n x_j = 1, x \geq 0$;
- a single linear constraint $a^\top x = b$ or $a^\top \leq b$;
- norm balls $\|x\|_p \leq b$ or cones $\|x\|_p \leq y$.

## 6.4 Convergence

The fundamental tool to analyze the convergence of projected gradient descent is the so-called gradient mapping.

**Definition 6.3.** *Given a closed and convex set $C \neq \emptyset$ and a differentiable function $f$, we define the* gradient mapping *as the function $G_\alpha : \mathbb{R}^n \to \mathbb{R}^n$, parametrized by $\alpha > 0$:*

$$G_\alpha(x) = \frac{1}{\alpha}\left(x - P_C(x - \alpha \nabla f(x))\right)$$

The definition above can be intuitively understood starting from the inner part: first we compute the next point with a standard gradient approach, assuming a step size of $\alpha$, then we project back onto $C$, and finally we compute the direction we actually moved to in the projected gradient, assuming the same step size. Note that the gradient mapping above allows us to state the projected gradient method with an update rule which is formally very similar to the plain gradient descent:

$$x^{k+1} = x^k - t_k G_{t_k}(x^k)$$

so $G_{t_k}(x^k)$ takes the role of $\nabla f(x^k)$. Also, note that, for $C = \mathbb{R}^n$, we get exactly $G_\alpha(x^k) = \nabla f(x^k)$, independent of $\alpha$. This generalization of the gradient has very similar properties w.r.t. the actual gradient.

**Proposition 6.3.** *$x^* \in C$ is optimal $\Leftrightarrow G_\alpha(x^*) = 0$.*

*Proof.* $\Rightarrow$) By the normal cone theorem, we have $-\nabla f(x^*) \in N_C(x^*)$. By definition, this means that (scaling does not change anything):

$$-\alpha \nabla f(x^*)^\top (y - x^*) \leq 0 \quad \forall y \in C$$

We can add and remove $x^*$ from the above to obtain:

$$(x^* - \alpha \nabla f(x^*) + x^*)^\top (y - x^*) \leq 0 \quad \forall y \in C$$

By the minimum principle, we thus have that:

$$P_C(x^* - \alpha \nabla f(x^*)) = x^*$$

Finally:

$$\begin{aligned} G_\alpha(x^*) &= \frac{1}{\alpha}(x^* - P_C(x^* - \alpha \nabla f(x^*))) \\ &= \frac{1}{\alpha}(x^* - x^*) = 0 \end{aligned}$$

$\Leftarrow$) Since $G_\alpha(x^*) = 0$, we have that:

$$x^* = P_C\left(x^* - \alpha \nabla f(x^k)\right)$$

By the minimum principle, we also have that:

$$\left[(x^* - \alpha \nabla f(x^k)) - P_C(x^* - \alpha \nabla f(x^k))\right]^\top (y - P_C(x^* - \alpha \nabla f(x^k))) \leq 0 \quad \forall y \in C$$

Plugging $x^* = P_C(x^* - \alpha \nabla f(x^k))$ into the above we obtain:

$$\left(x^* - \alpha \nabla f(x^k) - x^*\right)^\top (y - x^*) \leq 0 \quad \forall y \in C$$

and thus $-\nabla f(x^*) \in N_C(x^*)$. By the normal cone theorem, this implies that $x^*$ is an optimal solution.

□

**Proposition 6.4.** *With the usual constant step size $t_k = 1/M$, the norm of the gradient mapping gives the guaranteed objective progress:*

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2M} \|G_{1/M}(x^k)\|^2$$

*Proof.* Let us start again by the minimum principle:

$$(y - P_C(y))^\top (z - PC_(y)) \leq 0 \quad \forall z \in C$$

In the above, we can choose $z = x^k$ and $y = x^k - \frac{1}{M} \nabla f(x^k)$. Noting that, by definition, this implies $P_C(y) = x^{k+1}$, we obtain:

$$(x^k - \frac{1}{M} \nabla f(x^k) - x^{k+1})^\top (x^k - x^{k+1}) \leq 0$$

After some reshuffling, the above can be rewritten into:

$$\begin{aligned} \nabla f(x^k)^\top (x^{k+1} - x^k) &\leq -M \| \underbrace{x^k - x^{k+1}}_{-\frac{1}{M} G_{1/M}(x^k)} \|^2 \\ &= -\frac{1}{M} \|G_{1/M}(x^k)\|^2 \end{aligned}$$

We can now apply the inequality just proved to the standard descent lemma:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \underbrace{\nabla f(x^k)^\top (x^{k+1} - x^k)}_{\leq -\frac{1}{M} \|G_{1/M}(x^k)\|^2} + \frac{M}{2} \| \underbrace{x^{k+1} - x^k}_{\frac{1}{M} G_{1/M}(x^k)} \|^2 \\ &\leq f(x^k) - \frac{1}{M} \|G_{1/M}(x^k)\|^2 + \frac{1}{2M} \|G_{1/M}(x^k)\|^2 \\ &= f(x^k) - \frac{1}{2M} \|G_{1/M}(x^k)\|^2 \end{aligned}$$

□

Given the above, we can prove the exact same convergence results as for standard gradient descent, namely an error rate of $O(1/\varepsilon)$ in the convex case and $O(\log 1/\varepsilon)$ in the strongly convex/PL case.